Proceedings of Spatial Accuracy 2016

Jean-Stéphane Bailly, Daniel A. Griffith, Didier Josselin

Proceedings of Spatial Accuracy 2016

Editors:
Jean-Stéphane Bailly, Daniel Griffith & Didier Josselin

[ISBN: 978-2-9105-4510-5]

organized by

INRA SCIENCE & IMPACT · LISAH · AgroParisTech

ISARA International Spatial Accuracy Research Association
spatial-accuracy.org

ESPACE · cnrs · Université d'Avignon et des Pays de Vaucluse · Montpellier SupAgro

• Spatial uncertainty in knowledge-based systems

• Spatial accuracy quantification in mapping

• Spatial inference under uncertainty

• Scale effects in spatial accuracy

• Uncertainty in 3D geodata

• Spatial uncertainties simulation and propagation

• Spatial model sensitivity analyses

• Poster Session

Spatial uncertainty in knowledge-based systems

Is spatial information in ICT data reliable?

Maxime Lenormand[1,*], **Thomas Louail**[2,3], **Marc Barthelemy**[4,5], **José J. Ramasco**[2]

[1]Irstea, UMR TETIS, 500 rue François Breton, FR-34093 Montpellier, France
[2]IFISC (CSIC-UIB), Campus UIB, E-07122 Palma de Mallorca, Spain
[3]Géographie-Cités (CNRS - Paris 1 - Paris 7), 13 rue du four, FR-75006 Paris, France
[4]Institut de Physique Théorique, CEA-CNRS (URA 2306), F-91191, Gif-sur-Yvette, France
[5]CAMS, EHESS-CNRS (UMR 8557), 190-198 avenue de France, FR-75013 Paris, France

*Corresponding author: maxime.lenormand@irstea.fr

Abstract

While an increasing number of human activities are studied using data produced by individuals' ICT devices, there have been relatively few contributions investigating the robustness of results against fluctuations of data characteristics. In particular, when ICT data contain spatial information, they represent an invaluable new source for analyzing urban phenomena. Here, we present a stability analysis of higher-level information extracted from mobile phone metadata passively produced during an entire year by 9 million individuals in Senegal. We focus on two specific information-retrieval tasks: (a) the identification of land use in the region of Dakar by analyzing the temporal rhythms of the communication activity; (b) the identification of home and work locations of anonymized individuals, allowing for the construction of the Origin-Destination (OD) matrices for commuting flows. Our analysis reveals that the spatial distributions of land use computed for different samples are remarkably robust, with on average 80% of shared surface area between the different spatial partitions. The OD matrix is less stable with a share of about 70% of commuters in common when considering all types of flows. Better results can be obtained at larger levels of aggregation. These different results confirm that ICT data are mostly a very useful source for the spatial analysis of urban systems, but that their reliability should be tested more thoroughly.

I INTRODUCTION

Massive amounts of geolocated data are passively and continuously generated by individuals when they use their mobile and connected devices: smart phones, credit cards, GPSs, RFIDs or remote sensing devices. This deluge of digital footprints is growing at an extremely fast pace and represents an unprecedented opportunity for researchers, to address quantitatively challenging problems, in the hope of unveiling new insights on the dynamics of human societies. Many fields are concerned by the development of new techniques to handle these vast datasets, and range from applied mathematics, physics, to computer science, with plenty of applications to a variety of disciplines such as medicine, public health and social sciences for example.

Although *big data* have the advantage of large samples sizes (millions of observations), and high spatio-temporal resolution, they also raise new challenging issues. Some are technical and related to the storage, management and processing of these data (Kaisler et al., 2013), and others are methodological, such as the statistical validity of analysis performed on such data. For example, in the case of mobile phone data, researchers have often no control on the data collection that is usually made for other purposes. In many cases various hidden biases can affect the spatial behavior of anonymized individuals. Observing the world through the lenses

of data generated by information and communications technologies (ICT) may therefore lead to possible distortions, and possibly to erroneous conclusions (Lewis, 2015). It is thus crucial to perform statistical tests and to develop methods in order to assess the robustness of the results obtained with ICT data. In the research community that studies human mobility (in particular in cities, and related urban dynamics Ratti et al. (2006); Louail et al. (2014); Calabrese et al. (2015); Louail et al. (2015), efforts in this sense have been made in recent years, notably by comparing the results obtained with different data sources (Tizzoni et al., 2014; Lenormand et al., 2014; Deville et al., 2014; Alexander et al., 2015; Toole et al., 2015). However, the robustness of results to sample selection has, to our knowledge, never been studied yet.

In the following we present two examples of such uncertainty analysis on results obtained with mobile phone data recorded in Senegal in 2013 (de Montjoye et al., 2014). We concentrate on two information-retrieval tasks: first, we evaluate the uncertainty when inferring land use from the rhythms of human activity (Soto and Frías-Martínez, 2011; Frías-Martínez et al., 2012; Toole et al., 2012; Pei et al., 2014; Lenormand et al., 2015), and second, we study the uncertainty when identifying individuals' most frequented locations (Ahas et al., 2010; Isaacman et al., 2011; Lenormand et al., 2014; Toole et al., 2015). We conclude by mentioning possible future steps to clearly assess the relevance of various ICT data sources for studying different phenomena.

II STUDY AREA AND DATA DESCRIPTION

We focus here on the region of Dakar, Senegal. The mobile phone data consists in call detail records (CDR) of phone calls and short messages exchanged by more than 9 million of anonymized Orange's customers. They were collected in Senegal in 2013, and were released to research teams in the framework of the 2014 Orange Data for Development challenge (de Montjoye et al., 2014). We will use for our study the second dataset (SET2) that was made available by Orange. It contains fine-grained location data on a rolling 2-week basis at the individual level. For each of the 25 two-week periods, a sample of about $300,000$ mobile phone users is selected at the country scale. Whenever one of these individuals uses his/her mobile phone during the two-week period, the time and his/her position (at the level of serving cell tower) is recorded. This information can be used to study human activity and mobility patterns in the region of Dakar that is here divided into 457 spatial subunits. The partition is the Voronoi tessellation constructed from the location of antennas in the city chosen as nodes. Each Voronoi cell approximates thus the activity zone served by the antenna located at its center (see Figure 1a).

III INFERRING LAND USE FROM MOBILE PHONE ACTIVITY

Geolocalized ICT data have been widely used to infer land use from human activity (Soto and Frías-Martínez, 2011; Frías-Martínez et al., 2012; Toole et al., 2012; Pei et al., 2014; Lenormand et al., 2015). The basic idea is to divide the region of interest into zones, then extract a temporal signal of activity for each of these zones, and finally cluster together zones that display similar signals. Each of these clusters corresponds to a certain type of activity (*Residential*, *Commercial*, ...). We use here the functional approach proposed in Lenormand et al. (2015). The method takes as input, for each cell, a signal composed of 168 points ($24\text{h}\times 7\text{days}$), each value corresponding to the number of users located in this cell, at this hour of the day and this day of the week. These signals are then normalized by the total hourly activity, in order to subtract trends introduced by circadian rhythms. A Pearson correlation matrix between cells is then computed. Two spatial units whose activity is strongly correlated in time will have a high

positive correlation value. This similarity matrix can be represented by a undirected weighted network, which is then clustered using the Infomap community detection algorithm (Rosvall and Bergstrom, 2008). This method has the advantage to be non-parametric (the number of clusters is not fixed *a priori*).

In order to extract temporal signals of activity in the region of Dakar, we first need to estimate the number of people in each zone, per hour, for each of the 350 days of our original sample. To do so, we rely on the following criteria: each person counts only once per hour. If a user is detected in k different zones within a certain 1-hour time period, each registered position will count as $(1/k)$ 'units of activity' for each of these k cells. The average number of users per hour and per day is $30,500$, which represents about 1% of the total population of the region of Dakar.

To assess the robustness of land use identification from mobile phone activity to sample selection, we apply the functional approach described above to 100 weeks drawn at random from our original sample (after removing outliers). Note that the days of the weeks are drawn separately, and they are therefore not necessarily chronologically ordered.

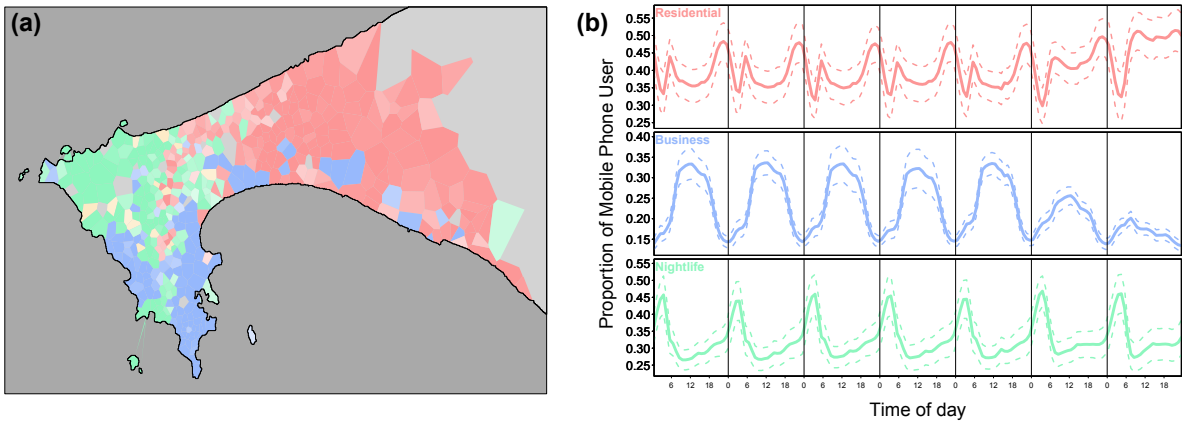


Figure 1: (a) Map of the region of Dakar displaying the three clusters. Colors vary from white to the most recurrent cluster identified in the 100 random sample. The color saturation depends on the number of times the zone was classified as the most recurrent cluster. The color code is red for *Residential*, blue for *Business*, green for *Nightlife* and orange for other types of land use. (b) Temporal patterns associated with the three clusters. The solid lines represent the average temporal profile computed over 100 random extractions, while the dashed lines represent the standard deviations.

First, we observe that three clusters emerge systematically, and represent on average 95% of the total surface. The remaining 5% correspond to other clusters with no clear patterns, probably associated with some local one-time events. We show on Figure 1b the average temporal profiles and the variability around this average for each of these three clusters. Each of the clusters can be roughly associated with a dominant land use:

- A Residential activity corresponds to a high probability of mobile phone use during early mornings, evenings and week end days.
- A Business cluster displays a significantly higher activity from 9am to 5-6pm during weekdays.
- A Nightlife activity profile is characterized by a high activity during night hours (1am-4am).

The Nightlife cluster (in green) covers the area of the international airport, and also the neighborhood of "la Pointe des Almadies", where live mainly wealthy people – most of the rich/sophisticated nightclubs are also located there. The Business cluster covers Dakar's CBD ("Le

Plateau"), where are located most of the companies headquarters, and also the port. Finally the residential cluster cover the rapidly growing parts of Dakar peninsula, who profit from the highway construction. It is worth noting that the results are consistent with the ones obtained with another mobile phone dataset in Spain (Lenormand et al., 2015), except that in the case of Dakar, the method is not able to distinguish between industrial (or logistic) and leisure nightlife activities (see Lenormand et al. (2015) for more details).

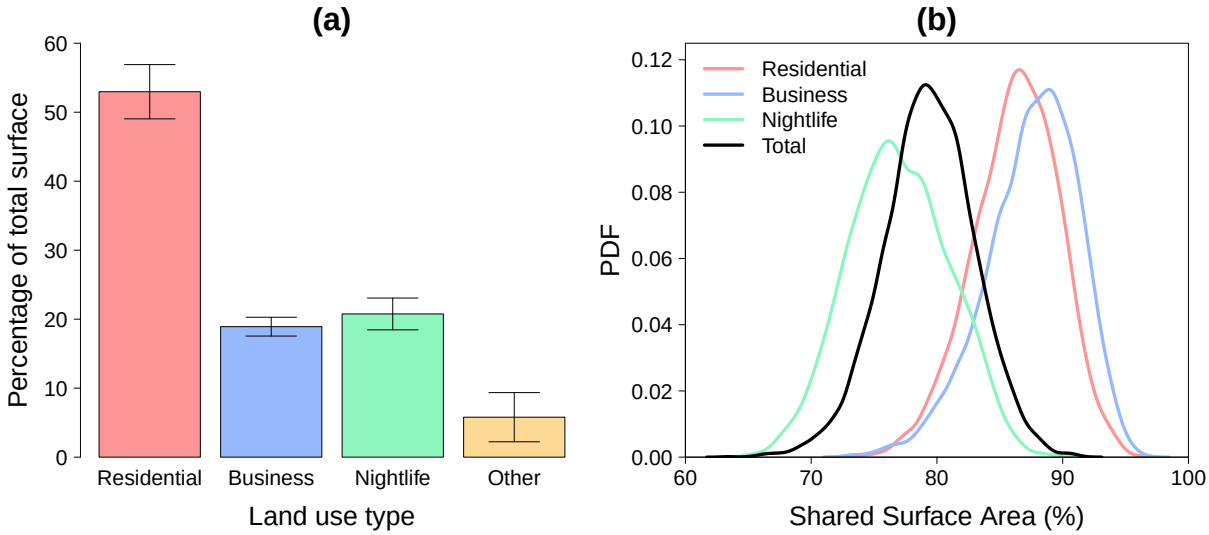


Figure 2: Uncertainty when inferring land use from mobile phone activity. (a) Area covered by the different land use types, expressed as a percentage of the total surface. The values have been averaged over 100 random samples, and the error bars represent the standard deviation. (b) Probability density function of the shared surface area between each pair of spatial distributions according to the type of land use.

As can be observed in Figure 2a, the area covered by the different types of land use is quite stable over the 100 samples, with the residential land use type representing on average about 55% of the total surface, while we observe about 20% for the Business and Nightlife clusters. Nevertheless, the stability of the proportion does not imply that they follow the same spatial distribution from one sample to another. To check for the stability, we compute the surface area shared by two spatial distributions S_l and S'_l of a given type l and obtained with two different samples. The expression for this quantity is

$$SSA(S_l, S'_l) = 2\frac{A_{S_l \cap S'_l}}{A_{S_l} + A_{S'_l}}, \tag{1}$$

where A_S denotes the surface area of spatial distribution S. Note that in our case $A_{S_l} \simeq A_{S'_l}$ (Figure 2a). Similarly, we can defined the total surface area shared by two spatial partitions P and P' (with the same number and types of land use) of the region of interest,

$$SSA(P, P') = \frac{\sum_l A_{S_l \cap S'_l}}{\sum_l A_{S_l}}. \tag{2}$$

The results are displayed in Figure 2b. The similarities between the 100 different spatial distributions is globally high, with on average 80% of shared surface area between the spatial partitions. The agreement is larger for *residential* and *business* clusters with an average shared surface area around 90%. For the *nightlife* land use type we find a result about 75%.

A map of the region of Dakar displaying the uncertainty associated with the land use identification is shown in Figure 1a. The colors represent the different land use types, where a zone has been assigned to the most recurrent cluster type over the 100 land use identifications. The color saturation is related to the uncertainty quantified with the number of times the zone was classified as given recurrent cluster: the color is darker if the uncertainty is low, paler otherwise. Most of the zones have been assigned the same clusters more than 80% of the time. This leads us to the conclusion that the identification of land use from mobile phone activity shows a high level of robustness to sample selection.

IV IDENTIFYING HOME-WORK LOCATIONS FROM MOBILE PHONE ACTIVITY

Geolocalized ICT data are also widely used to identify the most visited locations of an individual during his/her daily life trajectory, allowing to extract the origin-destination (OD) matrices of commuting flows, a fundamental object in mobility studies. A simple heuristic is that the most frequented place of a user in the late afternoon/evening and in the early morning can be identified as a proxy for his/her place of residence, while the most frequented area during working hours can be a proxy to his/her work (or activity) place. This simple assumption allows the accurate determination of mobility flows at intermediate geographical scales (see for example Tizzoni et al. (2014); Lenormand et al. (2014); Alexander et al. (2015); Toole et al. (2015)). However, the robustness of the results to sample selection has never been investigated.

For each of the 25 two-week periods and for each user, we apply the following home and work location extraction procedure:

- For each hour of the two weeks period (weekends excepted) during which an individual used his/her mobile phone, we identify the most frequently visited zone during this hour (based on his/her geolocalized mobile phone activity).
- Hours of activity are then divided into two groups, daytime hours (between 8am and 5pm) and nighttime hours (between 7pm and 7am).
- **Filter 1:** We keep only the users who have been 'active' at least ten 1-hour periods during daytime, and ten 1-hour periods during nighttime (spread over at least half of the days of the two-week period) .
- For both groups of hours (daytime and nightime), we identify the spatial unit in which the user has been localized the highest number of hours.
- **Filter 2:** We keep only users with a fraction of hours spent at "home" location and "work" location larger than one third of the total number of locations visited during nighttime and daytime, respectively.

The two filters allow us to discard users not showing enough regularity in order to estimate their main nighttime ('Home') and daytime activity ('Work') locations. The source code of this method is available online[1].

At the end of the process, after filtering out users living and/or working outside the region of Dakar, the remaining number of users is on average $\approx 65,000$. This number is quite stable over the 25 two-weeks periods, varying at most by 15% around this average. The resulting 25 two-weeks commuting networks can then be compared using several similarity metrics, such as the one described in Lenormand et al. (2016). We consider 2 commuting networks T and T', where T_{ij} is the number of users living in zone i and working in zone j, and we will use three different metrics, that encode different network properties. First, the common fraction of

[1] https://github.com/maximelenormand/Most-frequented-locations

commuters (CPC), varying from 0, when no agreement is found, to 1, when the two networks are identical, is estimated as

$$CPC(T,T') = \frac{2\sum_{i,j=1}^{n}\min(T_{ij},T'_{ij})}{\sum_{i,j=1}^{n}T_{ij}+\sum_{i,j=1}^{n}T'_{ij}}. \quad (3)$$

Second, we will consider the common proportion of links (CPL) that measures similarity in the networks' topological structure, and is calculated as

$$CPL(T,T') = \frac{2\sum_{i,j=1}^{n}\mathbb{1}_{T_{ij}>0}\cdot\mathbb{1}_{T'_{ij}>0}}{\sum_{i,j=1}^{n}\mathbb{1}_{T_{ij}>0}+\sum_{i,j=1}^{n}\mathbb{1}_{T'_{ij}>0}}. \quad (4)$$

Third, the common share of commuters according to the distance (CPC_d), assessing the similarity between commuting distance distributions is given by

$$CPC_d(T,T') = \frac{\sum_{k=1}^{\infty}\min(N_k,N'_k)}{N}, \quad (5)$$

where N_k stands for the number of users with a commuting distance ranging between $2k-2$ and $2k$ kms.

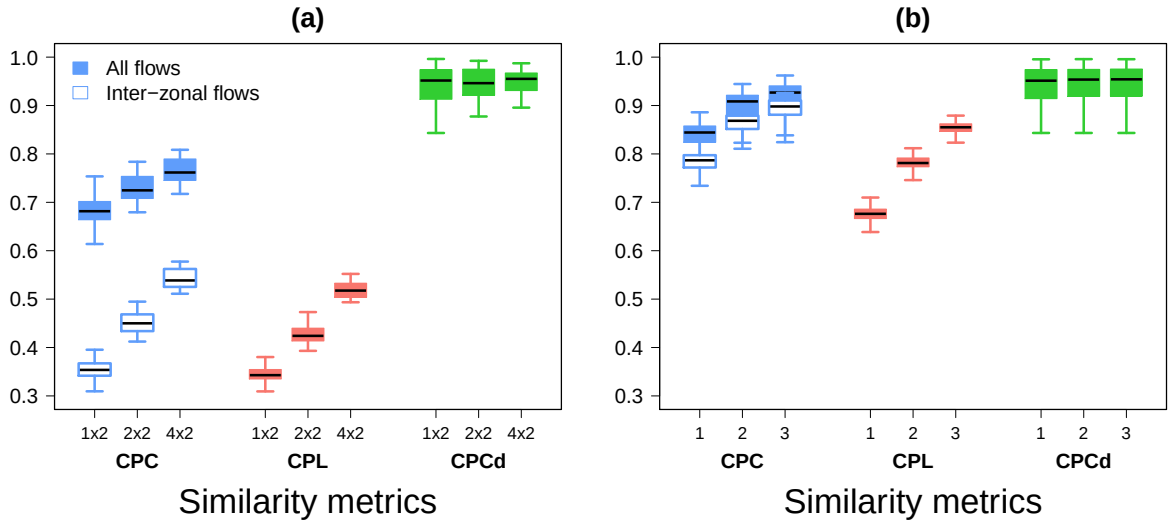


Figure 3: Uncertainty in the identification of the most frequently visited locations from mobile phone activity. (a) Boxplots of the CPC (blue), CPL (red) and CPC_d (green) between the 25 two-weeks commuting networks (1x2), 12 one-month commuting networks (2x2) and 6 two-months commuting networks (4x2). One-month and two-months commuting networks have been obtained by aggregation of consecutive two-weeks commuting networks. (b) Boxplots of the CPC (blue), CPL (red) and CPC_d (green) between the 25 two-weeks commuting networks, according to the grid cell side length (1 km, 2 km and 3 km). The CPC values obtained when considering only inter-zonal flows are also displayed.

The boxplots for the CPC, CPL and CPC_d values obtained by comparing the 25 two-weeks commuting networks are displayed in Figure 3a (1x2). The results are not completely conclusive, with different commuting networks showing a good agreement around 70% of commuters in common (considering both inter- and intra-zonal flows), but with a value falling down to 35% when only inter-zonal flows are considered. The CPL values are also quite low, around 35% of links are in common between the different networks. An encouraging result is that the common part of commuters according to the distance is very high, showing around 90% of similarity

between the 25 commuting distance distributions. However, it is important to keep in mind that these mixed results are obtained with a few thousand users for each two-weeks commuting network, drawn at a high spatial resolution with an average surface area equal to 0.5 km^2. Indeed, as it can be observed in Figure 3, both temporal and spatial aggregations of the networks greatly improve the results. The aggregation of consecutive two-weeks commuting networks allow us to perform pairwise comparisons of 12 one-month commuting networks (2×2) and 6 two-months commuting networks (4), and we clearly observe that the temporal aggregation improves the results. The two-weeks commuting networks can also be aggregated spatially, by projecting the data on a regular lattice (see Lenormand et al. (2014) for more details about the aggregation method), and results of the comparison of the two-weeks commuting networks according to the size of the grid cells (resp. 1 km, 2 km and 3 km) are represented in Figure 3b. Here again, we observe a significant improvement, with CPC values almost always larger than 0.75.

In order to go further, for each of the 25 two-week periods and for each user, we identify the home and work locations for each of the two weeks considered independently, by following the procedure described above. This allows us to assess the influence of the sampling of points along individual trajectories when identifying the home and work locations. We then compare the locations identified for each of the two weeks. Considering the high spatial resolution and the small time window, a good agreement is obtained, with an accuracy of $85.3 \pm 1.3\%$ for home (average $\pm$ standard deviation over the 25 two-week periods) and $79.1 \pm 2.7\%$ for work. Moreover, 60% of the inaccurate locations are less than 2 kms distant from each other. We can therefore conclude that the identification of users' home-work locations from mobile phone activity also shows a high level of robustness to sample selection.

V DISCUSSION

Data passively produced through information and communications technologies have been increasingly used by researchers since the middle of the 2000's to analyze a variety of human processes and activities. In particular, our understanding of human mobility has been deeply renewed thanks to these new sources. The longitudinal tracking of anonymized individuals opens the door to an enhanced understanding of human and social phenomena that could not be studied empirically with such a level of detail. However, these data may obviously suffer from a number of biases (Lewis, 2015), which include in particular sample selection. Systematic tests are then required for ensuring statistical validity, along with cross-checks between various data sources.

With this in mind, we performed two uncertainty analysis of results obtained with mobile phone data produced by millions of anonymized individuals and collected during an entire year. In the first part of the analysis, we assessed the uncertainty when inferring land use from human activity, estimated from the number of mobile phone users at different moments of the week. A good agreement was obtained between the land uses identified from 100 randomly selected samples of individuals, with on average 80% of shared surface area between land uses in the resulting maps. In the second part of the analysis, we investigated the influence of sample selection on the identification of users' home and work locations. We first examined the impact of the selection of users on the journey-to-work commuting networks extracted at the city scale. In our case-study of the city of Dakar, we showed that the level of uncertainty was highly dependent of the spatio-temporal resolution, and that good results were reachable with a reasonable level of aggregation. We then analyzed the effect of the sampling of locations along mobile phone

users' trajectories on the identification of their home and work locations. Most of the locations identified with different samples were the same, or very close to one another.

For these two spatial information retrieval tasks, our results suggest that the level of uncertainty associated with sample selection is low. Further work in this direction include the reproduction of such uncertainty analysis with other datasets coming from different countries and data sources. An important aspect of the rapidly growing 'new science of cities' (Batty, 2013), which heavily relies on new data sources, is to be able to reproduce results with different datasets, and to characterize and control to what extent the information provided by different sources are biased in a particular direction.

More studies in this spirit need to be done to strengthen the foundations of the field dedicated to the understanding of urban mobility and urban dynamics through ICT data. From a publication point of view, trying to reproduce previous results with different data sources, or to estimate the robustness of previously published results, might not be as appealing as proposing new measures and models, but is crucially important as well.

ACKNOWLEDGMENTS

Partial financial support has been received from the Spanish Ministry of Economy (MINECO) and FEDER (EU) under project ESOTECOS (FIS2015-63628-C2-2-R), and from the EU Commission through project INSIGHT. The work of TL has been funded under the PD/004/2015, from the Conselleria de Educación, Cultura y Universidades of the Government of the Balearic Islands and from the European Social Fund through the Balearic Islands ESF operational program for 2013-2017. JJR acknowledges funding from the Ramón y Cajal program of MINECO.

References

Ahas R., Silm, S.and J. O., Saluveer E., Tiru M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology 17*(1), 3–27.

Alexander L., Jiang S., Murga M., González M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies 58, Part B*, 240–250.

Batty M. (2013). *The New Science of Cities*. MIT Press.

Calabrese F., Ferrari L., Blondel V. D. (2015). Urban sensing using mobile phone network data: a survey of research. *ACM Computing Surveys (CSUR) 47*(2), 25.

de Montjoye Y.-A., Smoreda Z., Trinquart R., Ziemlicki C., Blondel V. D. (2014). D4D-Senegal: The Second Mobile Phone Data for Development Challenge. *arXiv preprint arXiv:1407.4885*.

Deville P., Linard C., Martin S., Gilbert M., Stevens F. R., Gaughan A. E., Blondel V. D., Tatem A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences 111*(45), 15888–15893.

Frías-Martínez V., Soto V., Hohwald H., Frías-Martínez E. (2012). Characterizing urban landscapes using geolocated tweets. In *SocialCom/PASSAT*, pp. 239–248. IEEE.

Isaacman S., Becker R., Cáceres R., Kobourov S., Martonosi M., Rowland J., Varshavsky A. (2011). Identifying Important Places in People's Lives from Cellular Network Data. In K. Lyons, J. Hightower, and E. M. Huang (Eds.), *Pervasive Computing*. Springer Berlin Heidelberg.

Kaisler S., Armour F., Espinosa J. A., Money W. (2013). Big Data: Issues and Challenges Moving Forward. In *2014 47th Hawaii International Conference on System Sciences*, Volume 0, pp. 995–1004.

Lenormand M., Bassolas A., Ramasco J. J. (2016). Systematic comparison of trip distribution laws and models. *Journal of Transport Geography 51*, 158–169.

Lenormand M., Picornell M., Cantú-Ros O. G., Tugores A., Louail T., Herranz R., Barthelemy M., Frías-Martínez E., Ramasco J. J. (2014). Cross-Checking Different Sources of Mobility Information. *PLoS ONE 9*(8), e105184.

Lenormand M., Picornell M., Garcia Cantú O., Tugores A., Louail T., Herranz R., Barthelemy M., Frías-Martínez E., Ramasco J. J. (2015). Comparing and modeling land use organization in cities. *Royal Society Open Science 2*, 150459.

Lewis K. (2015). Three fallacies of digital footprints. *Big Data & Society 2*(2).

Louail T., Lenormand M., Picornell M., Cantú O. G., Herranz R., Frias-Martinez E., Ramasco J. J., Barthelemy M. (2015). Uncovering the spatial structure of mobility networks. *Nature Communications 6*.

Louail T., Lenormand M., Ros O. G. C., Picornell M., Herranz R., Frias-Martinez E., Ramasco J. J., Barthelemy M. (2014). From mobile phone data to the spatial structure of cities. *Scientific reports 4*.

Pei T., Sobolevsky S., Ratti C., Shaw S. L., Zhou C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science 28*, 1988–2007.

Ratti C., Frenchman D., Pulselli R. M., Williams S. (2006). Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design 33*(5), 727–748.

Rosvall M., Bergstrom C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences 105*(4), 1118–1123.

Soto V., Frías-Martínez E. (2011). Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch*, HotPlanet '11, New York, NY, USA, pp. 17–22. ACM.

Tizzoni M., Bajardi P., Decuyper A., King G. K. K., Schneider C. M., Blondel V., Smoreda Z., González M. C., Colizza V. (2014). On the Use of Human Mobility Proxies for Modeling Epidemics. *PLOS Comput Biol 10*(7), e1003716.

Toole J. L., Colak S., Sturt B., Alexander L. P., Evsukoff A., González M. C. (2015). The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies 58, Part B*, 162–177.

Toole J. L., Ulm M., González M. C., Bauer D. (2012). Inferring Land Use from Mobile Phone Activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, New York, NY, USA, pp. 1–8. ACM.

Modelling, interpreting and visualizing uncertainties for the North Wyke Farm Platform baseline field surveys

Paul Harris*[1], Chris Brunsdon[2], Lex Comber[3], Hadewij Sint[1], Robert Orr[1], Michael Lee[1], Phil Murray[1]

[1] Rothamsted Research, North Wyke, Devon, EX20 2SB, UK
[2] National Centre for Geocomputation, Maynooth Unuveristy, Ireland
[3] University of Leeds, Leeds, West Yorkshire, LS2 9JT, UK

*Corresponding author: paul.harris@rothamsted.ac.uk

Abstract
This study demonstrates a new approach for the visualization of spatial uncertainty using data from three agricultural field surveys.

Keywords
caricRture; kriging; geographically weighted models; grasslands research

I INTRODUCTION

The North Wyke Farm Platform (NWFP) is a systems-based, farm-scale experiment with the aim of addressing grassland agricultural productivity and ecosystem responses to different management practices. The 63 ha site captures the data necessary to develop a better understanding of the dynamic processes and underlying mechanisms that can be used to model how agricultural grassland systems respond to different management inputs.

Via cattle beef and sheep production, the underlying principle is to manage each of three farmlets in three contrasting ways: (i) improvement of permanent pasture (i.e. 'business as usual' *green* farmlet); (ii) improvement through the use of legumes (*blue* farmlet); and (iii) improvement through innovation (*red* farmlet). The connectivity between the timing and intensity of the different management operations, together with the transport of nutrients and potential pollutants from the NWFP is evaluated using numerous inter-linked data collection exercises, operating at various spatial and/or temporal scales. Fig. 1a maps the NWFP experiment, where its 15 hydrologically-isolated sub-catchments are shown; some of which consist of multiple fields. In this study, we introduce some of the modelling and visualization opportunities that are possible with this rich data resource, with respect to baseline field survey data only.

II METHODS

We study three surveys that were conducted in the summers of 2012 and 2013. Baseline data entails that the *blue* and *red* treatments of NWFP farmlets are not as yet in place (thus the map in Fig. 1a is effectively all *green* during these surveys). Surveys sampled over a mixture of 25m and 50m grids for: (a) plant nutrients in 2013 (n = 544); (b) plant species in 2013 (n = 294); and (c) soil nutrients in 2012 (n = 250). These data sets (and many more) are freely available at www.rothamsted.ac.uk/farmplatform.

For plant nutrients, bulk stable isotopes of Carbon (δ^{13}C) and Nitrogen (δ^{15}N) along with Total Carbon (C), Total Nitrogen (N) and sward height were measured. For plant species, 18 species

were observed, where two species, Agrostis stolonifera and Loium perenne, clearly dominated. For soils nutrients, $\delta^{13}C$, $\delta^{15}N$, C and N, together with Bulk Density, Soil Organic Matter and pH were measured (Fig. 1b). Surveys can be analysed as a self-contained entity or related to each other.

Our study implements geostatistical and geographically weighted (GW) models to these surveys. Outputs are visualized to reflect different aspects of uncertainty, showcasing functions provided in the caricRture R package (Brunsdon 2016). Functions extend the 'sketchy' rendering techniques of Wood et al. (2012) to use with spatial data.

Using caricRture, we map a model output of interest (predicted/estimated to a finer 15m grid) via short lines that are coloured to demark the output range. To reflect uncertainty, the lines are drawn with varying degrees of 'sketchiness', where straight lines suggest relatively low levels of uncertainty, whilst highly sketchy ('hand-drawn') lines suggest relatively high levels of uncertainty. The technique provides a useful way of communicating uncertainty, where information that is commonly presented as two maps, are presented using only one. We present four case studies, each with a different visualization theme.

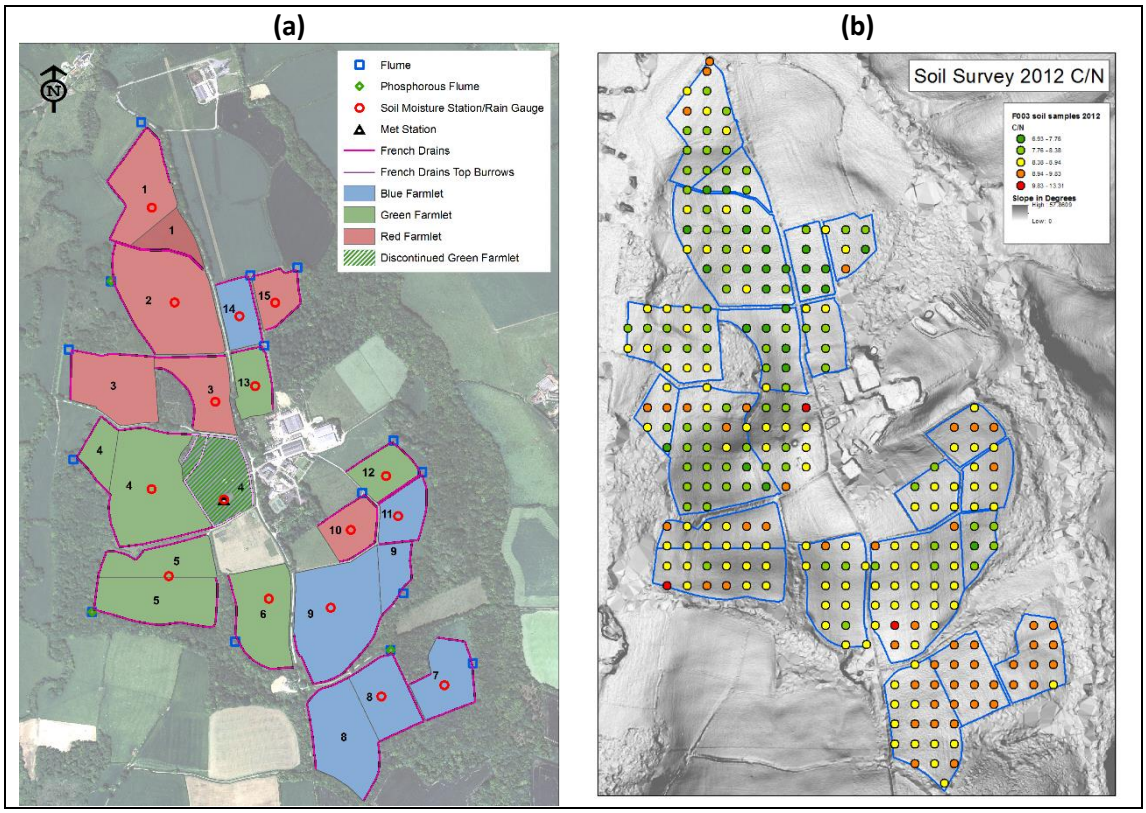


Figure 1:(a) The North Wyke Farm Platform and (b) Map depicting the C/N ratio for soils.

III CASE STUDIES

Visualization of spatial prediction uncertainty

In our first visualization, we predict C from the plant nutrients survey using universal kriging (UK) with a linear drift. Instead of using the UK variance as a measure of prediction uncertainty, we use the UK interpolation variance of Yamamoto (2000) which, unlike the UK variance, accounts for local changes in sample variance. The method is sensitive to the kriging neighbourhood size, and we choose N = 20. Fig.2a provides the UK prediction map, Fig. 2b provides the corresponding UK interpolation variance map, and Fig. 6a. provides the resultant 'sketchy' map - combining both UK predictions and their uncertainty. Thus the prediction levels

remain the same for maps in Fig 2a and Fig. 6a, whereas the high interpolation variances coloured dark green in Fig. 2b correspond to areas of high 'sketchiness' in Fig. 6a. Thus plant C predictions should be viewed cautiously in sub-catchments 2 and 7 (see also Fig. 1a).

Visualization of species metric uncertainty

Next, we suggest how to visualize species richness together with species diversity (for the plant species survey). Species diversity is a function the number of species present (i.e. species richness) and the evenness of how the individuals are spread (Hurlbert 1971). We choose to view species richness as our main output (Fig. 3a) and species diversity (Shannon's index) as its uncertainty or variance (Fig. 3b). To aid these visualizations we have smoothed both indices to the 15m grid. Fig. 6b provides the resultant 'sketchy' map - combining both species metrics. Species richness levels remain the same for maps in Fig 3a and Fig. 6b, whereas high levels of species diversity coloured dark green in Fig. 3b correspond to areas of high 'sketchiness' in Fig. 6b. Thus plant species are both rich and diverse, in for example, sub-catchments 14 and 15.

Visualization of local multivariate data structure uncertainty

Here, we apply a localised PCA (GWPCA, Harris et al. 2011) to all seven variables of the soils nutrients survey. We then map the local percentage of the total variance (PTV) accounted for by the first two components (PC1 and PC2) in Fig. 4a, where it appears that the seven soils variables tend to be relatively uncorrelated in the central sub-catchments of the NWFP, but relatively correlated elsewhere. From the usual PCA, the single global PTV value is 64.6% for the first two components, and in this instance, we map the absolute difference between this global value and each local PTV value in Fig. 4b. In Fig. 6c, we use that depicted in Fig. 4b, to reflect local structural uncertainty in our soils data with respect to how different the local PTV data is to that found globally. Thus data structure in sub-catchment 4 is most like that found, if we were to naively apply a non-spatial PCA to this data.

Visualization of local regression coefficient uncertainty

For our final visualisation, we apply a bootstrap methodology (Harris et al. 2015) to test for non-stationarity in the coefficients of a GW regression (GWR) as an alternative to a multiple linear regression (MLR), the null hypothesis. Our response is plant species diversity and our predictor is the C/N ratio for soils. GWR reduces AIC by 22 units from the MLR fit, suggesting clear value in a localised relationship. Fig. 5a maps the local regression coefficients for the C/N ratio for soils, which compares to a global coefficient of 0.015. Fig 5b maps *p*-values that have been formatted such that low *p*-values indicate coefficients which are significantly different to that found globally. Fig. 6d displays the corresponding 'sketchy' map, where high levels of 'sketchiness' indicate areas of significant non-stationarity in the coefficient estimates.

IV CONCLUDING COMMENTS

We have presented four case studies on the value of 'sketchy' maps to communicate uncertainty. Further 'sketchy' visualizations are possible when it comes to relaying uncertainty in the boundaries of spatial data, such as those of each NWFP sub-catchment, which we are currently working on.

References

Brunsdon (2016). Representing Uncertain Geographical Information with Algorithmic Map Caricatures. *Geophysical Research Abstracts Vol 18 EGU2016-2609,* EGU General Assembly 2016.

Harris P, Brunsdon C, Charlton M (2011) Geographically weighted principal components analysis. *International Journal of Geographical Information Science* 25(10), 1717-1736.

Harris P, Brunsdon C, Gollini I, Nakaya T, Charlton M (2015) Using Bootstrap Methods to Investigate Coefficient Non-stationarity in Regression Models: An Empirical Case Study. *Procedia Environmental Sciences* 27:112-115.

Hurlbert SH (1971). The Nonconcept of Species Diversity. *Ecology* 52(4), 577-586.

Wood J, Isenberg P, Isenberg T, Dykes J, Boukhelifa N, Slingsby A (2012). Sketchy rendering for information visualization. *IEEE Transactions on Visualization and Computer Graphics* 18(12), 2749-2758.

Yamamoto JK (2000). An alternative measure of the reliability of ordinary kriging estimates. *Mathematical Geology* 32, 489-509.

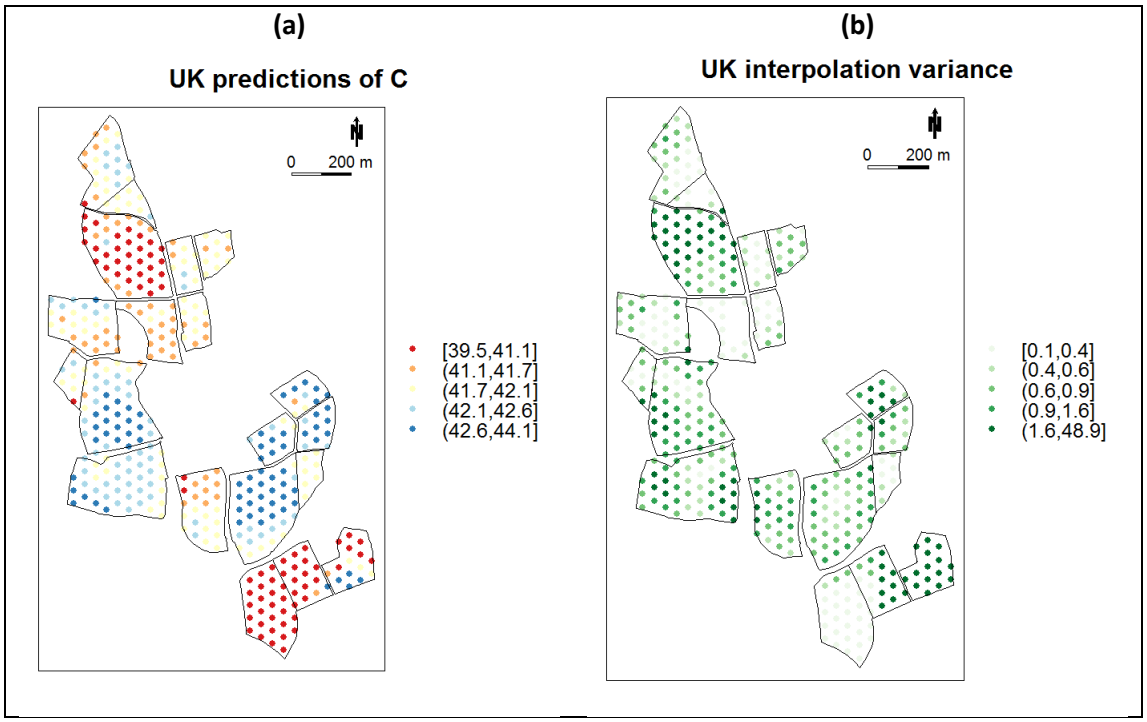


Figure 2: (a) UK predictions of C (plant nutrients) and (b) corresponding UK interpolation variances.

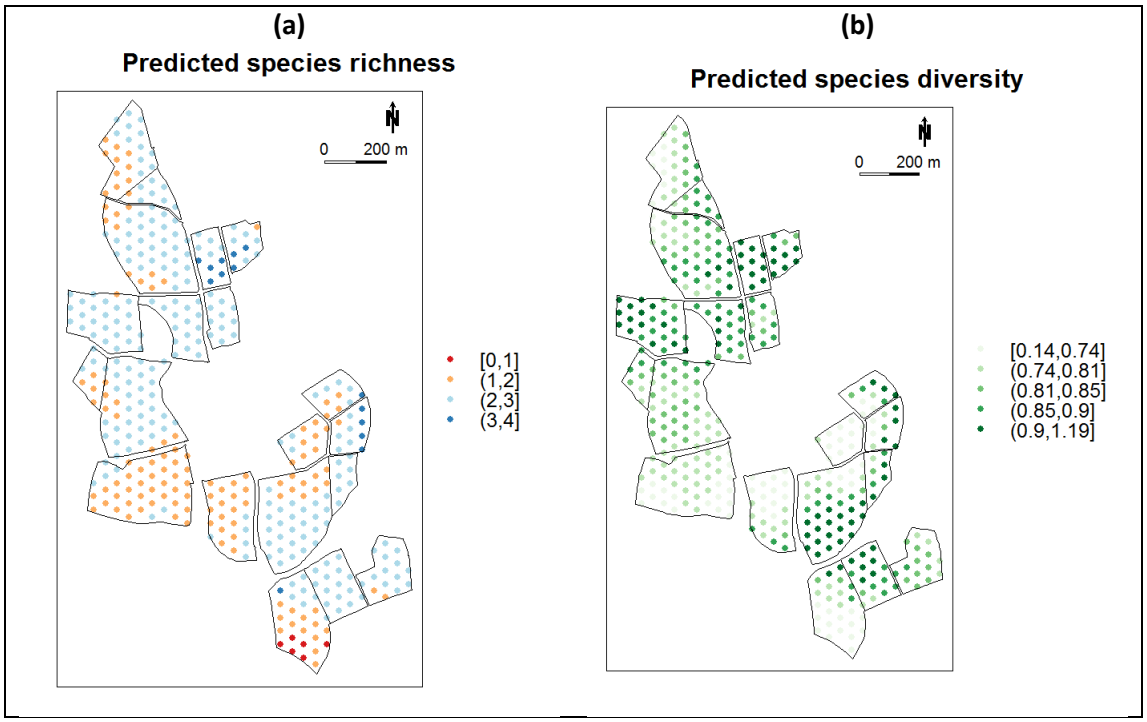


Figure 3: (a) Predicted species richness and (b) predicted species diversity.

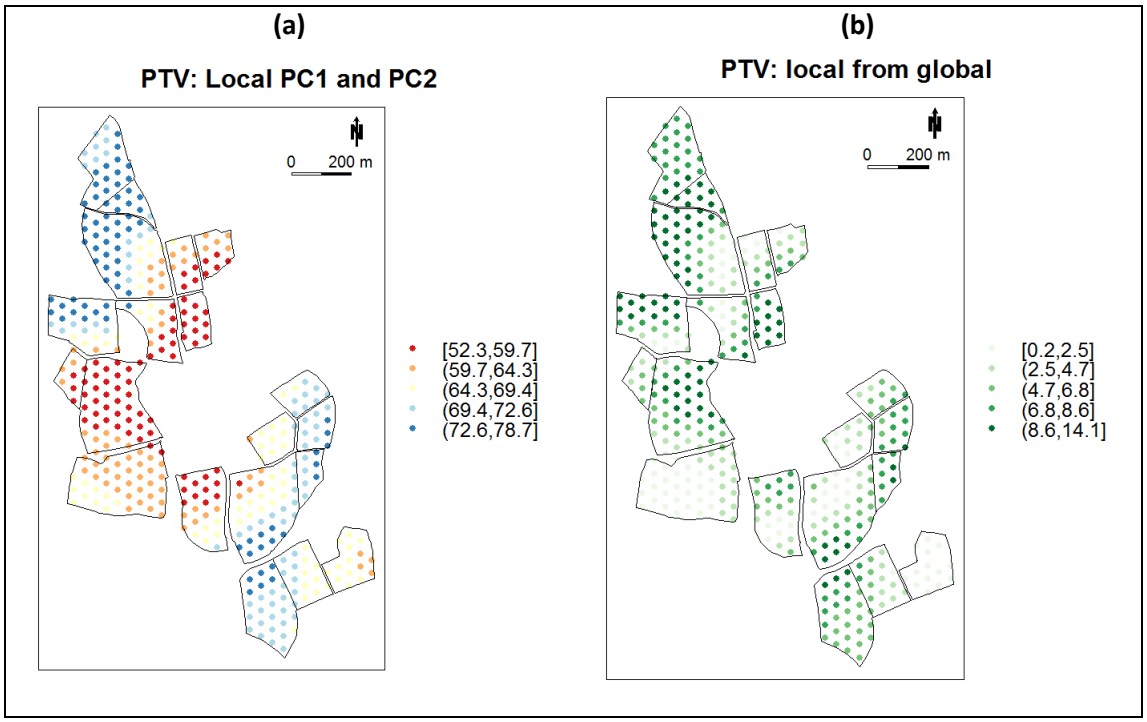


Figure 4: (a) Local PTV data from a GWPCA and (b) absolute difference in PTV from local to global.

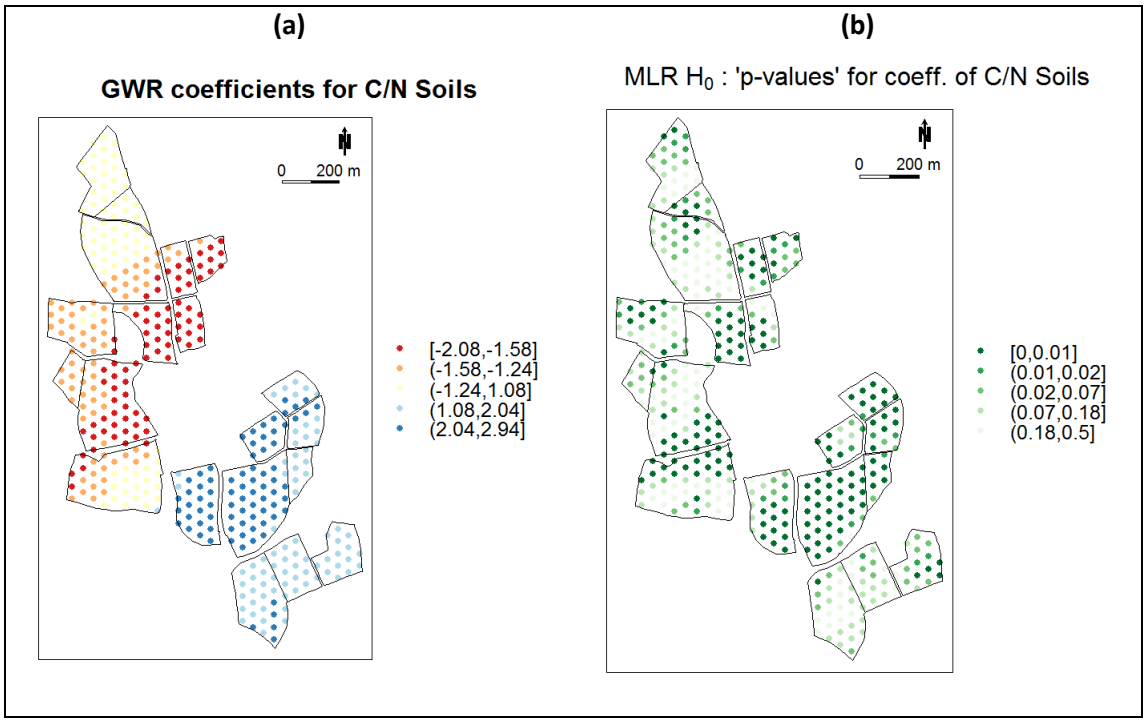


Figure 5: (a) GWR coefficients for C/N ratio for soils and (b) corresponding ‘p-values’.

 [ISBN: 978-2-9105-4510-5]

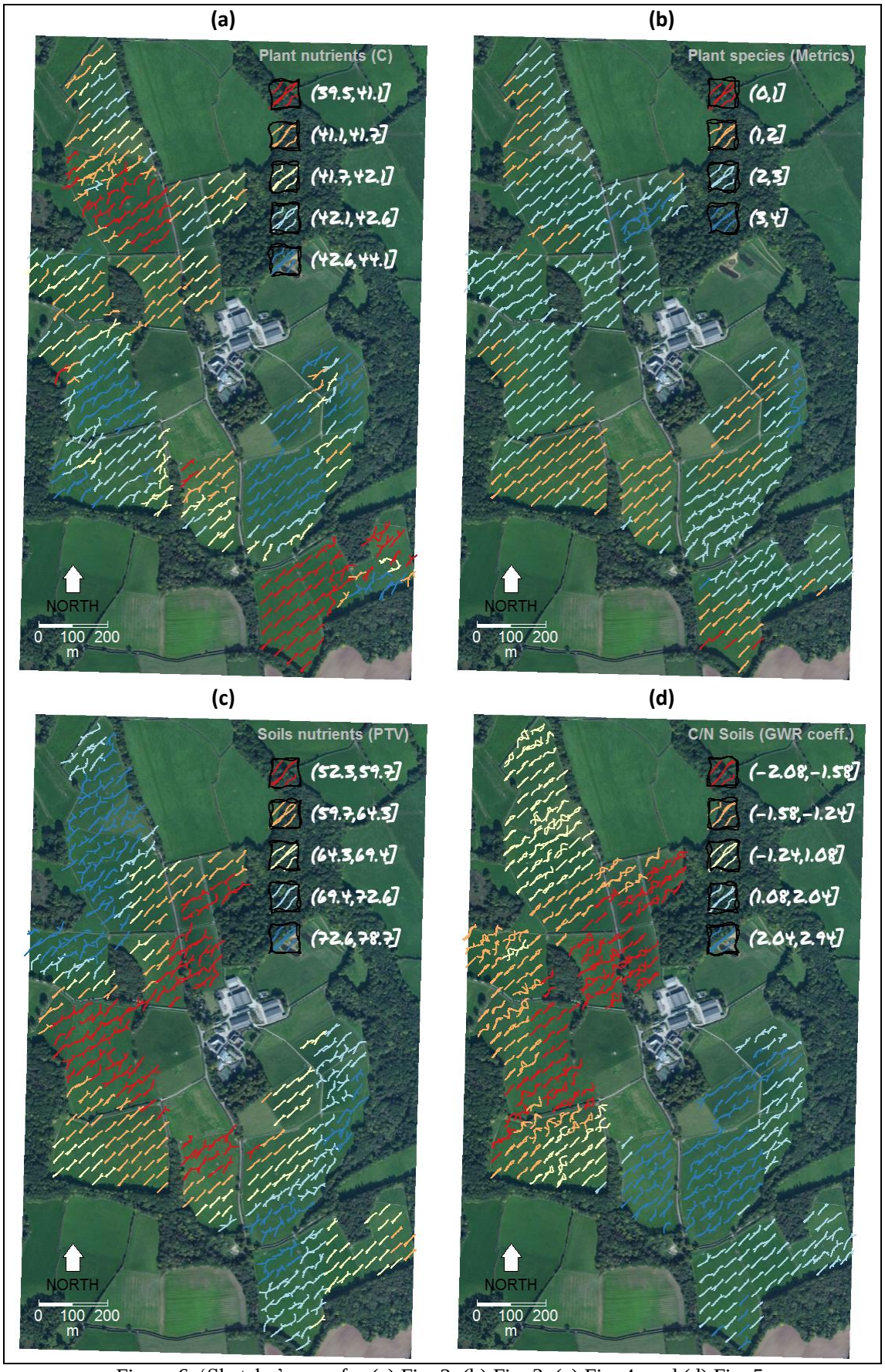


Figure 6: 'Sketchy' maps for (a) Fig. 2; (b) Fig. 3; (c) Fig. 4; and (d) Fig. 5.

Machine learning to deal with uncertainty in knowledge base for multivariate clustering applied to spatial analysis

Stephane Bourrelly[1]

Antonino Marvuglia[2]

Ian Vázquez-Rowe[3]

[1]University of Lyon 3 – UMR 7300 (ESPACE), France
[2]Luxembourg Institute of Science and Technology (LIST), Luxembourg
[3]Pontificia Universidad Católica del Perú, Peru

*Corresponding author: s.bourrelly@hotmail.fr

Abstract
We present a Tailor Made Machine Learning (TMML) methodology combining different clustering algorithms, spatial statistical methods and cartographic tools. The methodology is currently being programmed in an R package, especially designed to handle multivariate spatial datasets. We highlight the strengths of unsupervised clustering for the management of environmental health phenomena, also pointing out several uncertainty sources affecting the results of our analysis. In particular, we acknowledge that the traditional hierarchical clustering is usually applied without performing dynamic reallocations, integrating spatial key-concepts or discussing the quality of outputs. Therefore we describe the foundations of the TMML methodology, which is applied to deal with these uncertainties, as well as with the variety of possible outputs. The R functions are applied to the spatial dataset included in the package so to illustrate the procedure to apply for identifying the most accurate clustering output, in the context of a sustainable agriculture example in Luxembourg.

Keywords
Clustering, lattice, uncertainty, decision-making, agriculture.

I INTRODUCTION

When dealing with multidimensional datasets, clustering algorithms are often used to produce a reduced knowledge-base to simplify the representation of phenomena. The aim of multivariate clustering applied to spatial contexts is to form groups as heterogeneous as possible, composed by spatial units as similar possible. When clustering performs on lattice; spatial typologies are created in order to classify the administrative areas in meaningful clusters and summarise several dimensions of phenomena. In this way knowledge can be extracted from the multidimensional complexity with the purpose to conceive local policies, e.g. to ensure the economic growth and prevent socio-ecological conflicts (Delgado and Romero, 2016). The Hierarchical Clustering Analysis (HCA) is the most applied unsupervised method by public stakeholders, e.g. for the sustainable management of the agricultural activities (Boyacioglu and Boyacioglu 2008).

Unsupervised classification methods provide objective representations, unlike supervised methods which constrain the results by a known response variable. Unfortunately, even though

HCA presents a few advantages, the results are not optimal. Therefore some dynamic reallocation methods have been conceived to overcome this drawback (Hennig and Liao, 2010). However Levine (2000) notes that they are scarcely used in social sciences as well as the quality criteria which are rarely presented to discuss the statistical significance and the meaning of typologies. Moreover Levine points out that unsupervised clustering is usually used without integrating the key concepts of the scope of application; e.g. when algorithms are applied without taking into account the interconnectivity of the spatial units. Obviously, these shortfalls contribute to increase results' uncertainty, which might give rise to controversial decisions.

In this paper we briefly describe the methodology of a so-called Tailor-Made Machine Learning (TMML), which combines several clustering and data-mining algorithms with spatial statistic operators as well as cartographic tools. The TMML provides a methodology for taking into account the interconnectivity of spatial units. We present some of the functions of the TMML R package currently under development, to compare and map the outputs of the implemented clustering. Functions are applied to the case study shapefile of the TMML package, representing the environmental concentrations of chemical substances generated by the application of agricultural fertilizers in the administrative areas in Luxembourg.

II MATERIAL

The shapefile of the TMML package was derived from a cadastre spatial layer, representing the agricultural land use on the administrative areas (communes) in the Grand-duchy. This polygon layer describes the 108.328 georeferenced parcels contained in the cadastre records in 2009. These agricultural surfaces are differentiated from their main Agricultural Land Class (ALC) among the 23 official types, such as cereals, oilseed, legumes, etc. (Figure 1).

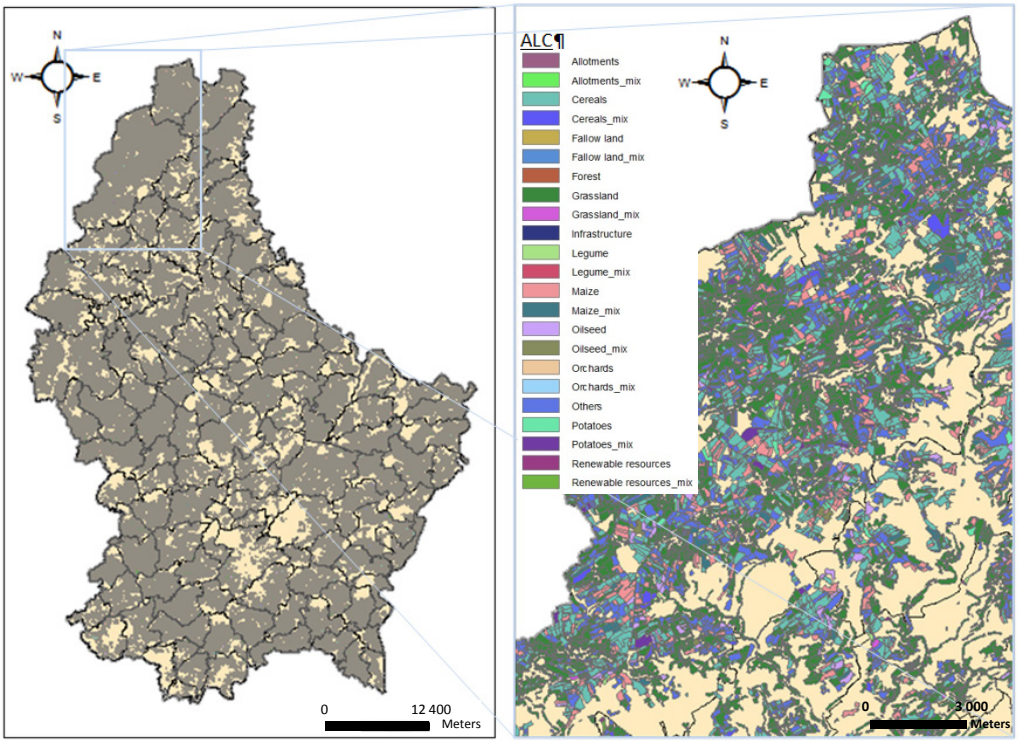


Figure 1: Luxembourg maps with 108.328 parcels named according to their main ALC.

In order to know the agricultural harvests in 2009 we used the agricultural statistics of the national database[1]. However, national statistics differentiate the agricultural surfaces from 27

[1] http://www.statistiques.public.lu/stat/ReportFolders/

different categories of crops; which have been matched according to the ALC description of cadastral data.

Fertilizers dosages (and in some cases yields data) have been derived from experts' communications, as well as from KTBL (2006). Chemical emissions from fertilizers application have been estimated according to (Nemecek and Kägi 2007). Finally the ArcGIS software has been used for aggregating, at the scale of communes, all the environmental emissions and computing the spatial indicators. We note x_k^j,the concentration level of the chemical substance 'j' in the administrative area of the commune 'k' for an environmental compartment, i.e. air, ground water or surface water (Table 1).

	AIR			GROUNDWATER		SURFACE WATER	
j	x_NH_3	x_NO_x	x_N_2O	x_P	x_PO4^{3-}	$x\_NO_3^-$	x_PO4^{3-}
substance	Ammoniac	Nitrogen oxide	Nitrous oxide	Phosphorus	Phosphate	Nitrate	Phosphate

Table 1: description of spatial indicators $x_\cdot^j$.

The command

- shapeLux = data(ml.chemichal)

-loads polygon shapefile of the TMML package. It describes the 116 communes of Luxembourg in 2009 (names, surfaces, index) and provides the values of the 7 x_k^j. The Figure 2 displays the spatial distributions of the concentrations of $\mathbf{NO_x}$ in air, $\mathbf{NO_3^-}$ in surface water and $\mathbf{PO4^{3-}}$ in groundwater, expressed in kg per km² of administrative area.

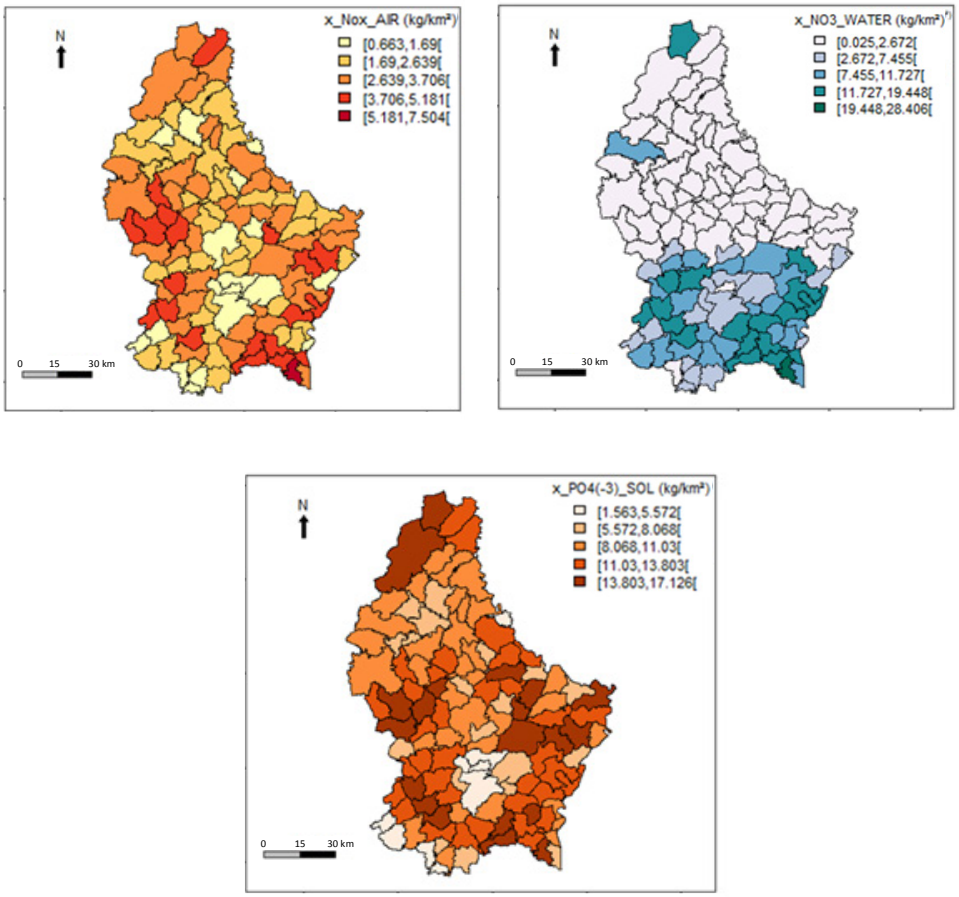


Figure 2: maps of $x_k^{NO_x}$ (upper left), $x_k^{NO_3^-}$ (upper right) and $x_k^{PO_4^{3-}}$ (lower).

III METHOD

The HCA algorithm performs an unsupervised classification for creating objective typologies. At the first step it groups the two most similar individuals (spatial units) so to form the first cluster (group). Then it iteratively maximises the inter-clusters inertia to group the other spatial units into nested cluster. Nested clusters are organized as a hierarchical tree called dendrogram, which allows specifying a number of clusters C.However with hierarchical clustering methods the composition of each of the generic clusters C_i is not optimal, due to the nested merging process. To overcome this drawback dynamic reallocation methods have been conceived in the literature for iteratively interchanging the cluster composition and minimising the intra-clusters inertia. The most popular algorithms are the Partitioning Around Medoïds (PAM) and K-means (Kaufmann and Rousseeuw, 1990).

Since the model hypotheses are different they provide different typologies. In addition they not allow objectively choosing a number of clusters; that is why they are usually initialised with the HCA outputs. The interpretation of typologies is a controversial issue, as only their statistical significance can be discussed from several quality criteria such as the R^2, the C-INDEX, the rate of inertia or the average silhouette (Hennig and Liao, 2010).

Moreover these clustering methods assume that individuals are independent, but each spatial unit influences the other ones located in its neighbourhood. Here we propose a methodology to integrate the spatial connectivity in the clustering analysis.

Firstly, in the TMML guidelines a conceptual neighbourhood pattern among those defined by Cliff and Ord (1981) should be chosen (Figure 3). In this way a neighbourhood matrix W is created. W contains weights w_{kl}, where $w_{kl} = 1$ if the spatial unit 'k' is directly adjacent to the spatial unit 'l' and $w_{kl} = 0$ otherwise. In order to test the interdependency between the adjacent values x_k^j, x_l^j the Moran's I statistic can be used (Gaetan and Guyon, 2010).

The tests can be replicated at several orders h of contiguity, i.e. for $h > 1$. In this case the neighbourhood matrix is termed $W(h)$ and the strength of spatial dependencies can be read on a correlogram. The correlogram chart displays the values of h as a function of the Moran's statistics $I(h)$ and their two-side confidence interval values, to denote the presence of a spatial autocorrelation (Gaetan and Guyon, 2010). In this way clustering algorithms can be applied to lagged values of x_k^j, noted xs_k^j to take into account the spatial dependency concept.

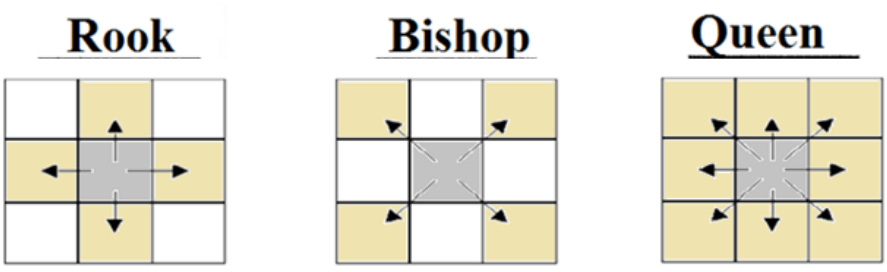


Figure 3: conceptual neighbourhood patterns at the order h=1.

IV RESULTS

The function

- ml.setC(x=X, update.C=4)

-performs a HCA and provides the chart of inertia differences between-clusters in decreasing orders, so to help in setting the number of groups C (Figure.4).

The function

- ml.clust.(shape=shapeLux, x=X, C=4, maps=T, bar.charts=T, ordinal=T, labCi = c("WEAK","MIDDLE","HIGH","MAJOR"))

-performs the HCA, K-means and PAM for merging the spatial units in C groups. It maps the distributions of class labels for the clustering outputs. With an ordinal typology the class labels 'labCi' are assigned according to the average centres of C_i When bar.charts=T bar charts are returned to summarise the characteristics of communes within the cluster C_i. The horizontal bars give differences in number of standard deviations, between the means of C_i and the overall average values for each chemical emission (Figure.4).

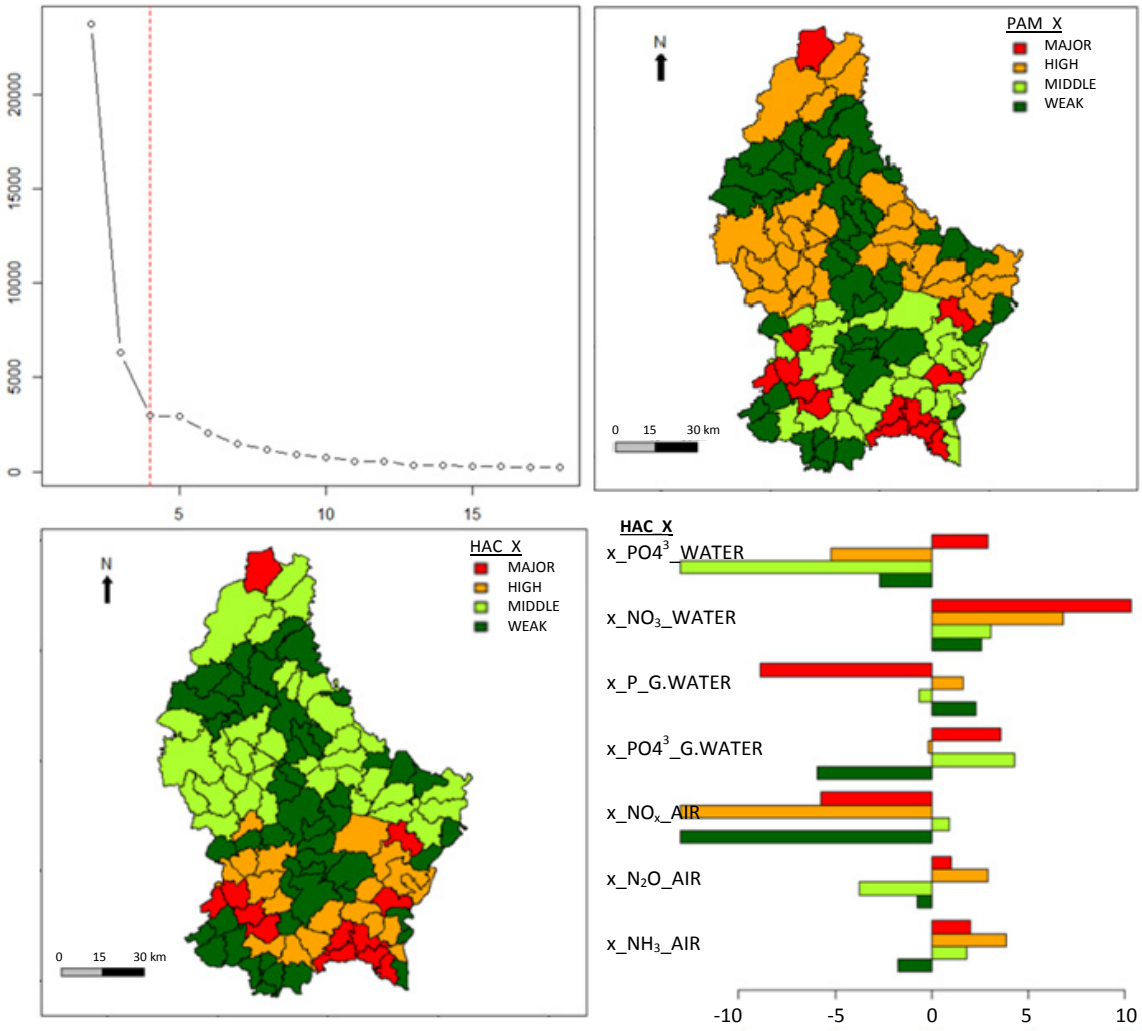


Figure 4: chart of inertia differences (upper left); maps of typologies for PAM (upper right) and HAC (lower left) with related bar chart (lower right).

The function

- ml.pattern(shape=shapeLux, x=X, j.lab="x_NOx", type="QUEEN", map=T Moran.test=T, correlogram="Moran", h.max=7)

-explores, for each spatial indicator j.lab, the interdependencies of values from a particular type of spatial pattern. The global autocorrelation of x_k^j is forecasted through the Moran test and the lag order of spatial dependencies, such as $h \leq$ h.max, can be assessed from the correlogram (Figure.5).

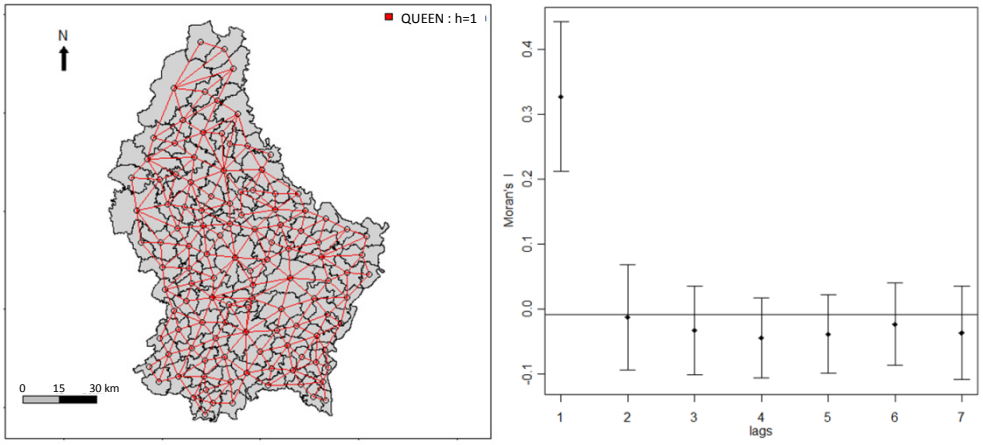


Figure 4: map of "Queen" connectivity pattern at the first order and Morans'I correlogram.

The function

- ml.clust.(…, …, spatial.patterns=rep("QUEEN", 7), h=c(1,1,1,2,2,3,5), mat.weight="W", quality=T, maps.xs=T)

-performs the HCA, K-means and PAM on the spatially lagged values of indicators $xs_{.}^{j}$, according to the spatial patterns, j-order dependencies h and standardised neighbourhood matrix, when mat.weight="W" (Figure.5). When quality=TRUE the quality criteria of clustering, highlighted in the method section, are returned.

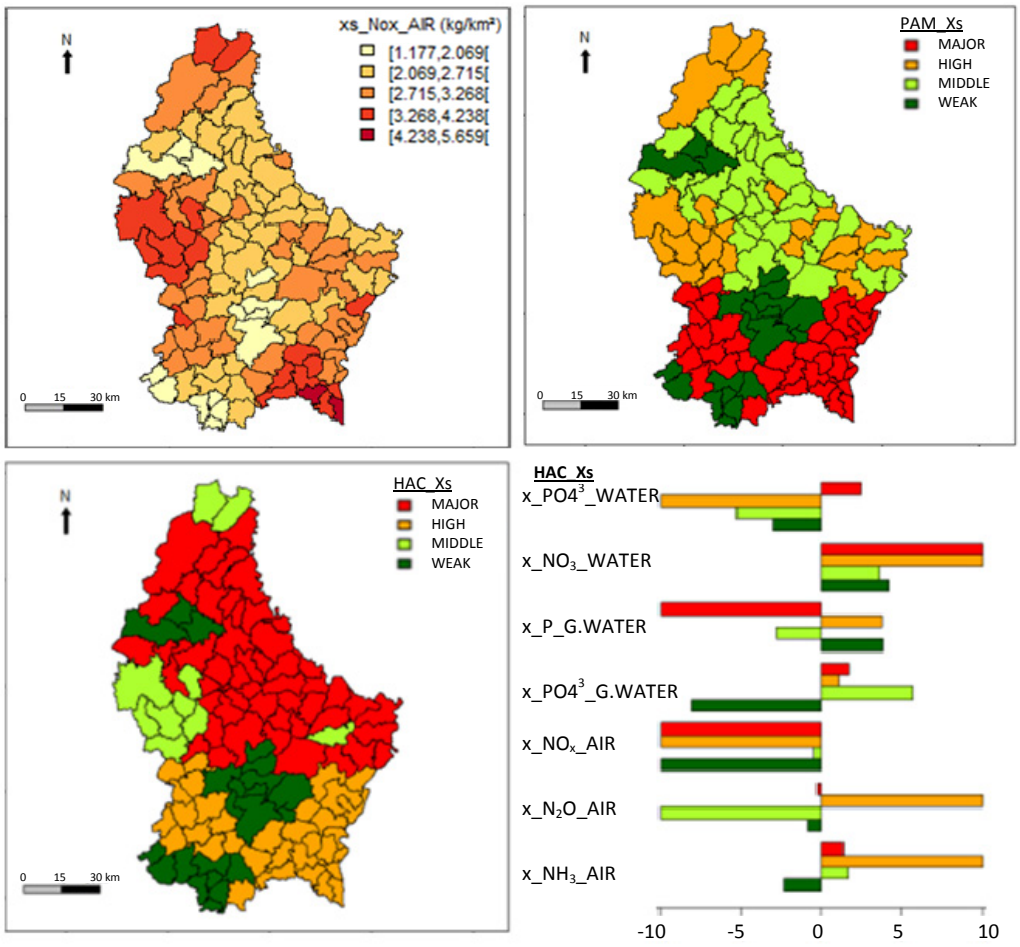


Figure 5: maps of xs_k^{Nox} and the typologies for PAM and HAC, with related bar chart.

V DISCUSSION

The variety of spatial typologies returned from the different clustering algorithms applied to the original or lagged spatial indicators, highlights the strong uncertainty induced in decision-making when the quality of outputs is not questioned.

Even though the uncertainties might be explored through some quality criteria, from the analyst's perspective it is important knowing to what extent the dynamic reallocations and the integration of spatial patterns really improve the statistical significance of the results.

In this respect an important uncertainty source in unsupervised clustering methods lays into the fact that the best clustering is usually chosen through statistical criteria. Indeed, the model selection issue is currently a recognised challenge in the unsupervised classification field (Hennig and Liao, 2010).

Ideally, the best clustering representation should be identified through a trade-off between the statistical significance of typologies and their spatial correlations with a decision criterion, identified by stakeholders. In particular, we point out the importance of defining the most suitable spatial typology for reducing the impacts of chemical emissions on eco-systems and therefore the cumulative socio-ecological inequalities. In future work, these latter could finally be estimated in line with the index of Morello-Frosch et al. (2011).

References

Boyacioglu H., Boyacioglu H. (2008). Water pollution sources assessment by multivariate statistical methods in the Tahtali Basin, Turkey. Environ Geol, 54, 275–282.

Cliff A., Ord J. (1981). The problem of spatial autocorrelation. In Scott A. London: Pion. 25-55.

Delgado A., Romero I. (2016). Environmental conflict analysis using an integrated grey clustering and entropy-weight method: A case study of a mining project in Peru. Environmental Modelling & Software, 77, 108-121.

Hennig C., Liao F. (2010). Comparing latent class and dissimilarity based. Department of Statistical Science, UCL, Department of Sociology. University of Illinois.

Gaetan C., Guyon X. (2010). Spatial Statistics and Modeling. Berlin: Springer.

Kaufman L., Rousseeuw P. (1990). Finding groups in data: an introduction to cluster analysis. New York: Wiley.

KTBL - Kuratorium für Technik und Bauwesen in der Landwirtschaft (Ed.) (2006). Faustzahlen für die Landwirtschaft, Darmstadt, Germany (in German).

Levine J. (2000). But What Have You Done For Us Lately. 29(1), 35-40

Morello-Frosch R, Zuk M, Jerrett M, Shamasunder B, Kyle AD. (2011). Understanding he cumulative impacts of inequalities in environmental health: implications for policy. *Health Aff (Millwood)* 30, 879–87.

Nemecek T., Kägi T. (2007). Life Cycle Inventories of Swiss and European Agricultural Production Systems. Final report ecoinvent V2.0 No. 15a. Agroscope Reckenholz-Taenikon Research Station ART, Swiss Centre for Life Cycle Inventories, Zurich and Dübendorf, CH, retrieved from: www.ecoinvent.ch.

Generation of Plateau-Approximated Fuzzy Zones

Hazaël Jones[1], **Serge Guillaume**[2], **Patrice Loisel**[3], **Brigitte Charnomordic**[3], **Bruno Tisseyre**[1]

[1]Montpellier SupAgro, UMR ITAP, Montpellier, France
[2]Irstea, UMR ITAP, Montpellier, France
[3]INRA, UMR MISTEA, Montpellier, France

*Corresponding author: hazael.jones@supagro.fr

Abstract

When modelling spatial real data in environmental context, the border between two zones is rarely a sharp edge. Often, the transition from one zone to another is a gradual process. Fuzzy zones intend to model this uncertainty in zones by allowing a spatial point to belong partially to different zones with different membership degrees. Generating fuzzy zones is not easy to perform and can be time-consuming and difficult to interpret. It is also difficult to ensure the convexity of the generated fuzzy zones.

Our idea is to generate plateau zones instead of fuzzy zones. Plateau zones are an approximation of fuzzy zones. The interest of using plateau zones lies in the following: easier spatial coherence and convexity of zones, preservation of uncertainties, level and number of plateaus can be chosen depending on the application, interpretation and easier operations between zones.

Our algorithm is based on quantiles of spatial data in order to produce some isocontours. According to the desired number of plateaus, it is possible to adjust the quantile values to find next plateaus for each resulting zones. The goal of this representation is to provide the user a simplification of the spatial representation and to preserve uncertainties in order to use them in the decision process.

Keywords

Fuzzy Zones, Plateau Zones, Spatial Data, Zoning Algorithm, Spatial Coherence

INTRODUCTION

When summarizing spatial data to consider zones (agronomy, geography, etc.), the border between two zones is rarely a sharp edge. In this context, there are no obstacles to detect but management zones to define. Often, the transition from one zone to another is a gradual process (e.g. between clay soil and sand soil, we can have a transition zone with more and more sand).

Fuzzy zones could be used to model this uncertainty/gradual change in zones by allowing a spatial point to belong to several zones with different membership degrees. Fuzzy zones keep the zone uncertainties until the final decision is taken, in order to be able to use this information to relax operational constraints.

Many methods exist to build zones based on segmentation (Pal and Pal (1993), Pedroso et al. (2010), Roudier et al. (2008)) but little work has been done on the generation of fuzzy zones (see Crane and Hall (1999), Philipp-Foliguet et al. (2009)). In Philipp-Foliguet et al., the mem-

bership degrees are assigned based on the distance to typical zones and attribute values, it is then difficult to ensure convexity of the generated fuzzy zones.

Plateau zones (see Kanjilal et al. (2010)) are an approximation of fuzzy zones as shown in Figures 1 and 2. Each plateau corresponds to a level with a membership degree. Kanjilal et al. (2010) have studied the interest of plateau zones for computation, but there are no computation methods to generate these plateau zones. This article is a first step to provide a method for such generation.

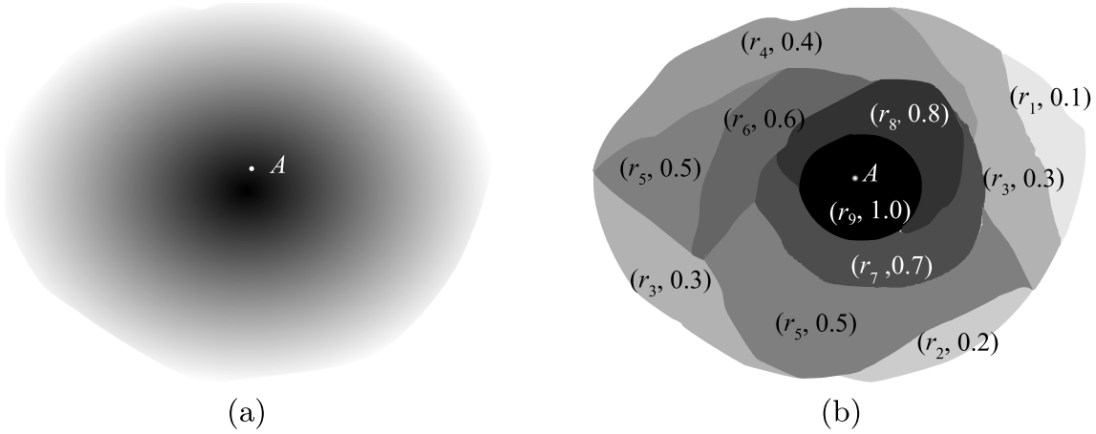


Figure 1: An example of a fuzzy zone modelling an air-polluted area (a) and its representation as a plateau zone (b) *source : Figures from article Kanjilal et al. (2010).*

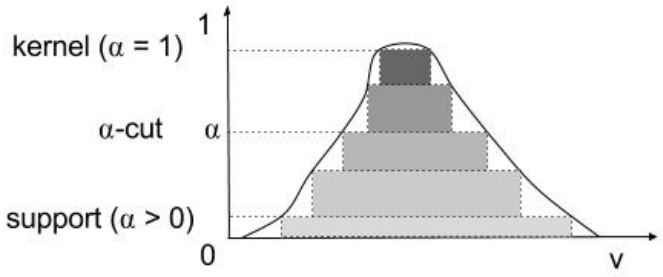


Figure 2: A 2-dimensional illustration of a fuzzy zone and its approximation by plateau zones.

Plateau zones present an interest for several reasons:

- Enhanced spatial coherence and convexity of membership degree for each zone,
- The level and the number of plateaus can be chosen depending on the application. Particularly in agriculture, it is not possible to apply a continuous treatment between one level and another.
- Interpretation: it is more efficient to use a map with plateau zones than continuous fuzzy zones in the decision making phase. Experts often reason using categories and not using a continuum of fuzzy values.
- Effective geometrical operations can be calculated between zones (difference, union, intersection, etc.) thanks to the simplification obtained by plateau (see Kanjilal et al. (2010)).

The goal of this representation is to provide the user a simplification of the spatial representation and to preserve uncertainties in order to use them within the decision process.

THE ALGORITHM FOR BUILDING PLATEAU ZONES.

The algorithm is based on quantiles of spatial data. The use of quantiles on spatial data intends to produce some isocontours. An isocontour (level set) is a line with the same value all along its length. According to the desired number of plateaus, it is possible to slightly adjust the quantile to find the next plateaus for each resulting zone. Quantiles used on spatial data are an efficient way to ensure the imbrication of plateaus with a spatial coherence. If expert knowledge is available on the data (thresholds to divide the data, knowledge about data distribution, etc.), it is possible to use it to set the quantile values. When no *a priori* information is available, quantiles can be chosen in order to divide data into equal proportions. The choice of the best quantile is not the focus of this paper.

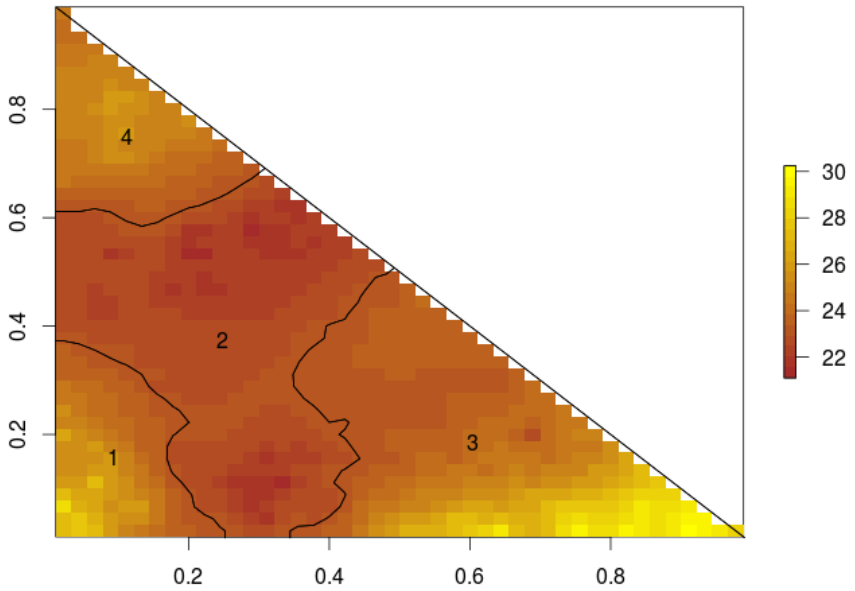


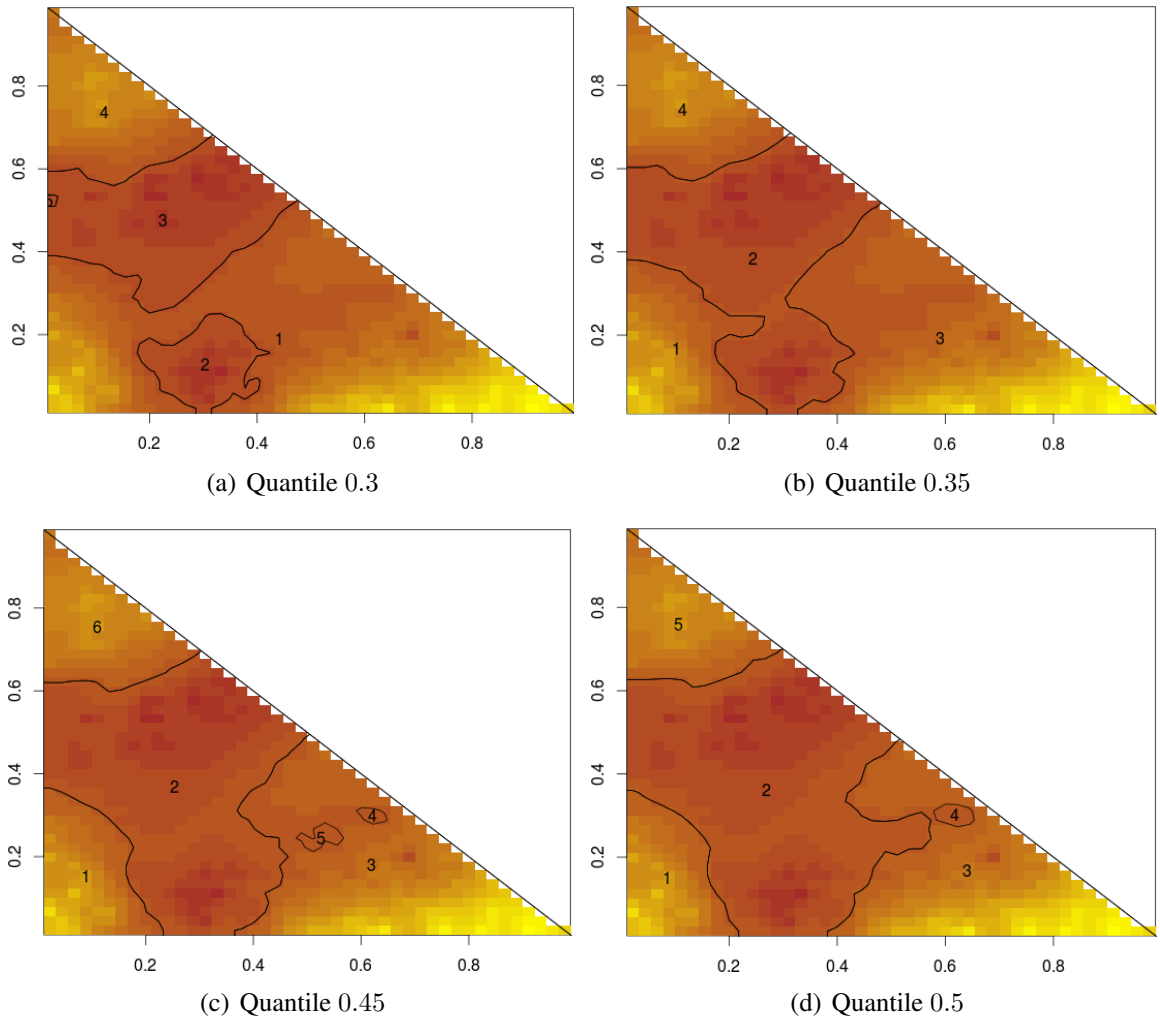
Figure 3: Initial zones based on a quantile.

Figure 3 shows the initial zoning given by one quantile (0.4) on simulated mono-dimensional geo-referenced data (generated using a Gaussian random field). Four distinct initial zones are obtained, and our goal is to first obtain plateaus from them, and to generate approximated fuzzy zones from these plateaus. We will focus on Zone 2 to illustrate the algorithm.

Figures 4(a), 4(b), 4(c) and 4(d) shows how the shape of Zone 2 evolves when the quantile is slightly adjusted. When applied to Zone 2, the algorithm can provide a computation of the kernel, α-cuts and support of this zone (see Figure 2 for an illustration of this concepts). Figure 4(a) shows the potential kernels of Zone 2.

Algorithm 1 is the proposition for building approximated fuzzy zones from plateau regions. For an initial zone and an initial quantile, our procedure BUILDPLATEAUZONES produces *nPlateau* zones associated to *nPlateau* quantile values. It is based on the following principle: as the initial zone (based on quantile 0.4) is selected to divide zones properly, it is granted a membership of 0.5. When we define zones thanks to quantiles inside this zone, the membership increases (see Algorithm 1, lines 6-7), and when we define them outside of this zone, it decreases (lines 8-9). The deeper (further) the zone is inside (outside) the initial zone, the more (less) it is characteristic of the zone. The algorithm allows a membership degree between 0 and 1 (line 7 between 0.5 and 1 - line 9 between 0 (0 excluded) and 0.5). A plateau with a membership degree of 0 is not relevant as it will not be part of the zone. Formula $0.5 - \frac{n}{nPlateau+1}$ (line 9) avoids this case.

 [ISBN: 978-2-9105-4510-5]



(a) Quantile 0.3 (b) Quantile 0.35

(c) Quantile 0.45 (d) Quantile 0.5

Figure 4: Zoning for different quantile values

Algorithm 1 Algorithm for building plateau zones

1: **procedure** BUILDPLATEAUZONES(*initialZone, quantile, step, nPlateau, data*)
▷ *listPlateaus* is a list of zones with associated memberships degrees
2: *listPlateaus* $\leftarrow \emptyset$
3: ADDZONEDEGREE(*listPlateaus*, *initialZone*, 0.5)
▷ compute new zones
4: $n \leftarrow 1$
5: **while** $n \leq nPlateau/2$ **do**
▷ compute new internal zones (memberships > 0.5)
6: *newZoneInt* $\leftarrow$ COMPUTEZONEINT($initialZone, quantile - n * step, data$)
7: ADDZONEDEGREE(*listPlateaus*, *newZoneInt*, $0.5 + \frac{n}{nPlateau}$)
▷ compute new external zones (memberships < 0.5)
8: *newZoneExt* $\leftarrow$ COMPUTEZONEEXT($initialZone, quantile + n * step, data$)
9: ADDZONEDEGREE(*listPlateaus*, *newZoneExt*, $0.5 - \frac{n}{nPlateau+1}$)
10: $n \leftarrow n + 1$
11:
12: **function** COMPUTEZONEINT(*initialZone, quantile, data*)
13: *relevantNewZones* $\leftarrow \emptyset$
▷ return a list of zones based on *quantile*
14: *newZones* = COMPUTEZONES(*quantile, data*)
15: **for all** $Z \in newZones$ **do**
16: **if** $Z \subseteq initialZone$ **then** ADDZONE(*relevantNewZones*, *Z*)
▷ return the largest zone that belongs to eligible zones
17: **return** (BIGGERZONE(relevantNewZones), *quantile*)
18: **function** COMPUTEZONEEXT(*initialZone, quantile*, data)
19: *relevantNewZones* $\leftarrow \emptyset$
▷ return a list of zones based on *quantile*
20: *newZones* = COMPUTEZONES(*quantile, data*)
21: **for all** $Z \in newZones$ **do**
22: **if** $Z \supseteq initialZone$ **then** ADDZONE(*relevantNewZones*, *Z*)
▷ return the smallest zone that belongs to eligible zones
23: **return** (SMALLESTZONE(relevantNewZones), *quantile*)
24: **function** COMPUTEZONES(*quantile, data*)
25: *zonesFromQuantile* $\leftarrow \emptyset$
26: *isoContours* $\leftarrow$ GENERATEISOCONTOURS($quantile, data$)
27: *zonesFromQuantile* $\leftarrow$ CLOSEZONES($isoContours, data$)
28: **return** (*zonesFromQuantile*)

When zones are generated thanks to a new quantile (functions COMPUTEZONEINT and COMPUTEZONEEXT), relevant new zones must be included in the initial zone for COMPUTEZONEINT (and must include the initial zone for COMPUTEZONEEXT). As a single quantile can generate multiple contour lines, if several zones are relevant, we choose the one with the largest area for COMPUTEZONEINT (smallest for COMPUTEZONEEXT). An illustration of this choice can be seen on Figure 4(a) where zone 3 will be chosen as it is bigger than zone 2. COMPUTEZONES returns the list of zones generated on a quantile isocontour basis, it builds isocontours based on data (line 26), and closes them respecting the field border (line 27). The initial zone is built following the same principle but initial zoning is edited by removing/expanding zones which

are too small.

Some functions are not specified in the algorithm to be more concise. ADDZONEDEGREE adds a zone and the associated membership degree to the final list of plateaus *listPlateaus*. ADDZONE adds a zone in the relevant zones. BIGGERZONE and SMALLESTZONE select the largest zone (respectively, the smallest) between a set of relevant zones. GENERATEISOCONTOURS produces isocontours from data values corresponding to a quantile and close them respecting the field border (CLOSEZONES).

The result of the computation on Zone 2 is given in Figure 5. The dark green line represents the border of the kernel (plateau of level 1) of Zone 2. 0.75 plateau is in green, 0.5 in black, 0.3 in blue, 0.17 in light blue. The same procedure is followed separately for each initial zone.

Thanks to the use of quantiles, the shape of Zone 2 plateaus is not just a dilatation-erosion of the initial zone because plateaus are directly based on data.

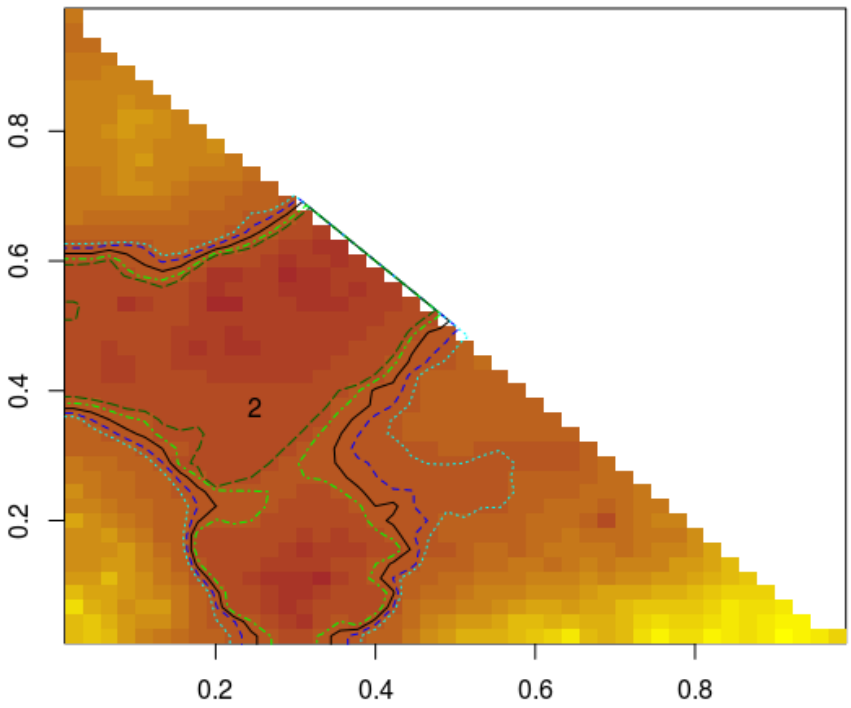


Figure 5: Final result with 5 plateaus for zone 2 (dark green (kernel), light blue (support)).

The algorithm allows the computation of fuzzy regions thanks to a plateau approximation. The computation is efficient (2.15 seconds for 5 plateaus on Intel core i7, 8 cores 2.7GHz, data size = 968) and provides plateaus on spatial data. Plateau zones are relevant because they provide both a simplification of the spatial representation and also the preservation of uncertainties.

This preliminary work can be extended in the following directions: First, it will be interesting to analyse the sensitivity of parameters. The initial quantile choice and the quantile step will influence the results. The bigger the value of the step, the wider the plateaus, the larger the uncertainty of each zone. Furthermore, a study considering the number of plateaus should be done in order to decide how many plateaus are needed to give the most relevant approximation. Finally, we want to show the interest of this representation using real agronomical data.

References

Crane S. E., Hall L. O. (1999). Learning to identify fuzzy regions in magnetic resonance images. In *Fuzzy Information Processing Society, 1999. NAFIPS. 18th International Conference of the North American*, pp. 352–356. IEEE.

Kanjilal V., Liu H., Schneider M. (2010). Plateau Regions : An Implementation Concept for Fuzzy Regions in Spatial Databases and GIS. In *Proceedings of the 13th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 624–633.

Pal N. R., Pal S. K. (1993). A review on image segmentation techniques. *Pattern recognition 26*(9), 1277–1294.

Pedroso M., Taylor J., Tisseyre B., Charnomordic B., Guillaume S. (2010). A segmentation algorithm for the delineation of agricultural management zones. *Computers and Electronics in Agriculture 70*(1), 199–208.

Philipp-Foliguet S., Gony J., Gosselin P. H. (2009). FReBIR: An image retrieval system based on fuzzy region matching. *Computer Vision and Image Understanding 113*, 693–707.

Roudier P., Tisseyre B., Poilvé H., Roger J.-M. (2008). Management zone delineation using a modified watershed algorithm. *Precision Agriculture 9*(5), 233–250.

Modeling process chain of SPOT images for resources uncertainty to monitor change in forest cover

Aimé Richard Hajalalaina[*1, 2], Dominique Hérvé[3], Eric Delaitre[4], Thérèse Libourel[4, 5]

[1]Centre Universitaire de Formation Professionnalisante, Université de Fianarantsoa, Madagascar
[2]Laboratoire de Recherche Appliquée Multidisciplinaire, Université de Fianarantsoa, Madagascar
[3]Institut de Recherche pour le Développement (IRD UMR 220), BP 64501, Montpellier, France
[4]Espace DEV, 500 rue JF Breton, Montpellier, France
[5]LIRMM, 161 rue Ada Université de Montpellier II, Montpellier, France

*Corresponding author: arhajalalaina@yahoo.fr

Abstract. In this paper, process chain and knowledge-based models of SPOT satellite images are proposed to help scientists, working in the field of the environment and in particular of the forest, to solve uncertainty of spatial resources (data, process) to monitor the change in forest cover which usually results of deforestation. Indeed, deforestation mobilizes all research, various methods of satellite image processing on forest dynamics are proposed. The SPOT images available, at present, are voluminous and heterogeneous in terms of spatial, radiometric and temporal resolutions (sensors HRV, HRVIR, HRG, NAOMI). The use of these SPOT images may solve uncertainty of data. In addition, phenomena such as deforestation require the analysis of time series of satellite images and the development of automated and reusable processing chains for monitoring change of forest cover. We propose to formalize these processing chains from modeling an abstract and concrete models based on existing standards in terms of interoperability (International Standard Organisation ISO and OGC Open Geospatial Consortium). The use of these standards solves uncertainty of process. These processing chains modelled will be capitalized, and diffusible in operational environments. Our modeling approach uses work-context concepts. These concepts need organization of human resources, data, and process in order to establish a knowledge-based connecting the two latter. This knowledge-based will be used to solve uncertainty of SPOT images resources for monitoring change in forest cover.

Keywords: forest cover, resources uncertainty, modeling, knowledge base, processing chain, satellite image, SPOT, work-context.

I INTRODUCTION

Our concerns are on the application of computer science in the field of environment remote sensing. Specialists in remote sensing produce hypotheses which they validate from experimental protocols. Now, remote sensing and computer experts work together to automatize these protocols in processing chains. This automatization poses many problems due, on the one hand, to the volumes of satellite images coming from different sensors and, in the other hand, to the proliferation of more or less complex image processing methods required by environment remote sensing.

In this context, it is necessary to put in place systems which help to store and manage important image-streams as well as their processing, by taking into account their various origins (different sensors, different spatial, spectral and temporal resolutions). The restitution and the exchange of these pieces of information are a real challenge in terms of

interoperability. This later has an advantage in reducing the uncertainty of image-processing methods through the capitalization and the mutualization of experiments between remote sensing specialists.

A lot of research-work, which represent high accuracy of classification, has been published on the study of forest cover using SPOT images, namely Hajalalaina et al. (2013), Souza et al. (2005)., Achard et al. (2002), Kimes et al. (1999), which has contributes to improving the knowledge on forest dynamics. But a data and processing formalization is necessary to ensure much larger spread of the research results to the scientific community. It helps to resolve the problem of uncertainty in the SPOT-image processing methods on the study of forest cover.

The formalization of data and processing is realized in a work-context acquired by the MDWeb platform put into practice by Desconnets et al. (2007). This platform proposes a view of human resource organization. It allows the plateform administrator to reference the future users from predefined roles and rights, and to reference the ressources (data/processings) within a metadata-base.

We propose to use this of this formalization to coordinate the collection of SPOT images and their processing, necessary for monitoring forest-cover. The objective is to capitalize, harmonize and spread the resources which allow a better understanding of forest dynamics at different spatial and temporal resolutions. It is part if several achievable actions within different period of time. The first one is the formalization of SPOT image processing to ensure their sharing, their re-use and their interoperability. The second one, which is a medium-term objective, is setting-up of a platform for the sharing a mutualization of experiments on SPOT image processing methods on the study of forest cover.

First, we have collected the current norms and formalization of spatial data and processing, then proposed work-context models for the formalization of SPOT image processing chains for monitoring of forest-cover.

II RESOURCE FORMALIZATION

The formalization corresponds with the description of the resources used by respecting the norms and standards which are in force. They are in the size of metadata. Various formalizations respecting the norms in force deal with syntactic and semantic aspects for the metadata. This definition and the application of these norms help to resolve the heterogeneity of the resource used. Dublin Core in BNF (2008) and ISO 19115 (2003) are the two norms which are valid and most used for the formalization of georeferenced data, like SPOT image, while ISO 19119 (2005) allows us to describe their processing. The formalization of processing chains use notion of work-context. The work-context is closely linked to the field of expertise. It can then be envisaged that it is built, through consensus, by experts in the area. Every scientist (the final user) will have at their disposal this context which they can develop or enrich in their turn. The construction of any work-context corresponds with the progressive on gradual organization of useful data and processing references, following three steps: human resource organization, data organization, and processing organization according to Libourel et al. (2010). Moreover even if this approach has not been standardized yet, it uses a relatively simple graphic symbolic language, called Simple Workflow Model (SWM) suggested by Lin et al. (2008), which allows the user scientists to handle easy appropriation concepts in a simple language.

The work-context of SPOT images in field of monitoring forest-cover is made up of three models bellow: human resource organization (figure 1) which manages the description of the platform users as well as that of their different roles and access rights, data organization (figure

2) which manages the description of SPOT image category according to sensors, processing organization (figure 2), which manages the description of SPOT images processing categories, work-context knowledge-base (figure 3) and the abstract model of processing chain of SPOT images to monitor change in forest cover (figure 4). This abstract model can be implemented by designing concrete model by using Orfeo Toolbox (2015) image processing.

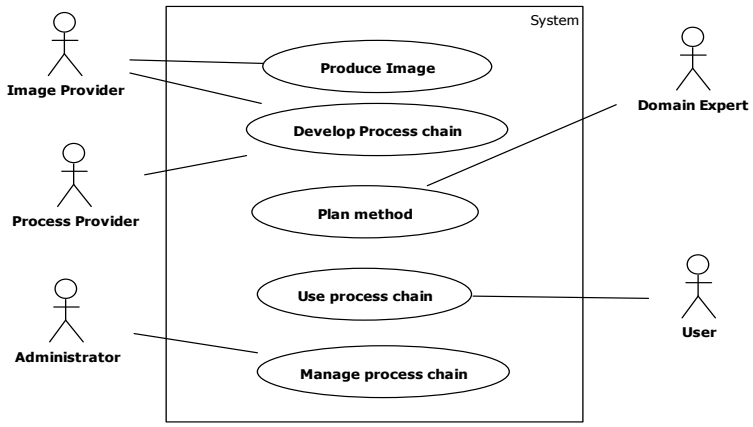


Figure 1 : Human resource organization model

Figure 2: SPOT images organization and their processing models

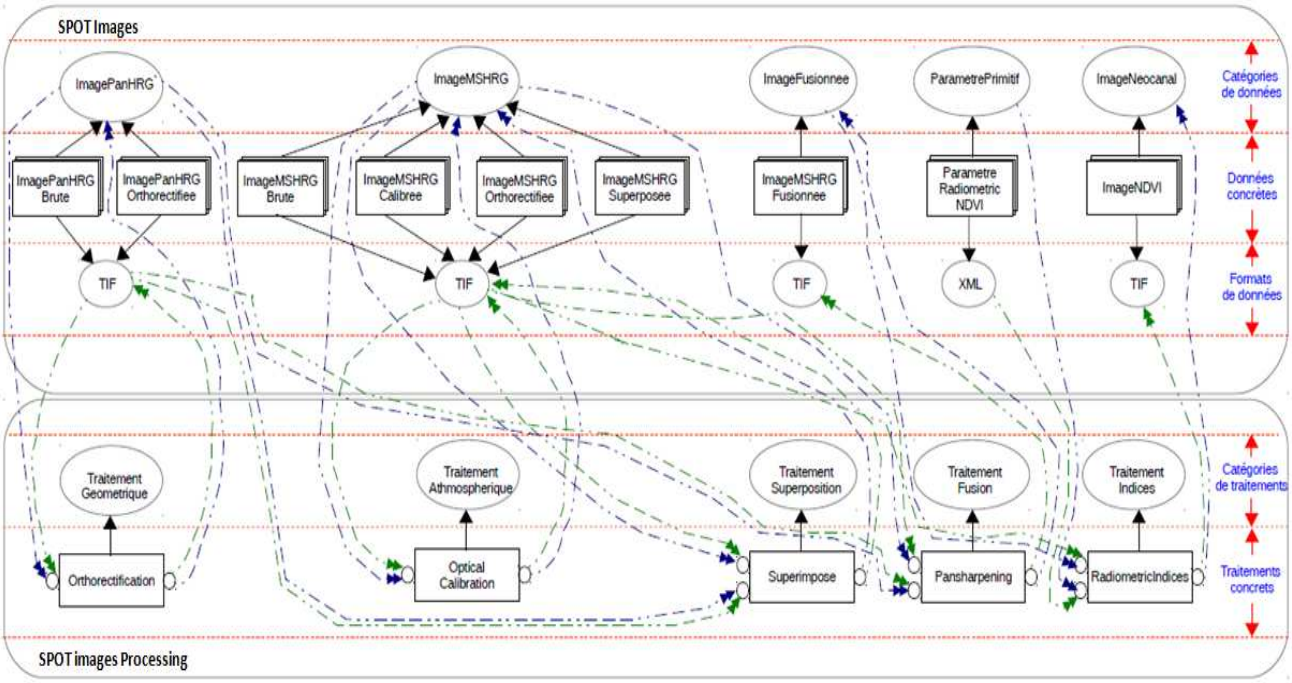


Figure 3: Extract of Work-context knowledge-base of SPOT images model

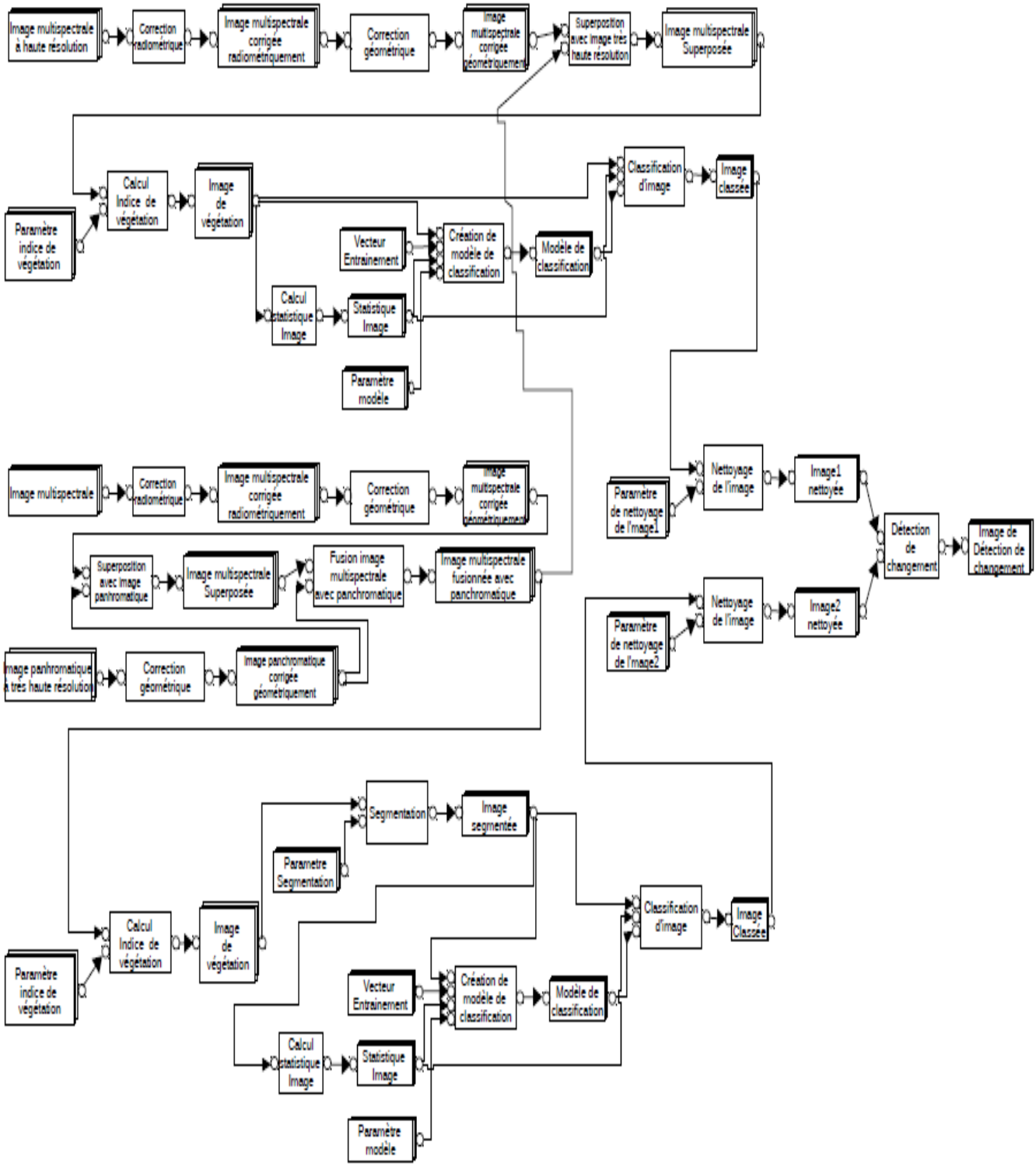


Figure 4: Abstract model of multi-resolution and multi-temporal SPOT images processing chain to monitoring change in forest cover

III ASSESSMENT OF PROCESSING CHAINS MODELS OF SPOT IMAGES

The manual processing chains of SPOT images will always remain a tough intellectual process. As a matter of fact, this type of chain requires the help of a remote sensing expert all along the realization process. In addition, new images entering the chain require the re-start of

the process from the beginning till the end. This makes their re-use difficult for users who are not specialists in the field and it does not make it possible to capitalize the experts' knowledge to assist the occasional users in the valorization of new images. We can say that the manual processing chains are not adapted to the SPOT images actualizing critical phenomena like the monitoring of forest cover.

Faced with this advantages of the manual processing chains, it interest of our proposals on formalizing processing chains lies in the automatization of different processing on SPOT satellite images.

Processing chain re-use: Once processing chain is formalized, its execution (implementation) is carried out automatically by taking into account new images. This is what makes the chain re-usable and also makes several repetitions of the processing possible without expert assistance for using the new images to carry out their analyses corresponding to their needs. Those properties are actually adapted to the monitoring of the forest cover, in view of changing and evolving character of SPOT images. It presents a major interest for capitalizing the experiments on SPOT images processing methods in order to overcome the problems of uncertainty in the processing of forest cover evolution.

Interoperability of processing chains: The formalized chain can therefore, a priori, be implemented in a distributed environment of Grid or Cloud type. Thus, it would then now be interesting to display these chains in a distributed manner in order to profit from this possible interoperability. As a matter of fact, in the contexts of forest cover monitoring, the distributed dimension will bring about an important advantage to this approach for it allow to divide up the experiments and spread on the whole set of the system to put in place, the contribution of separate elements to the knowledge-bases of work-context. In addition, normalized exchange protocols allow distributed requests which can be interesting for simultaneous research of SPOT image processing methods used in monitoring forest cover, which makes it possible to fills up the gaps on the uncertainty of SPOT image processing methods.

Cataloging processing chains: Making catalogs is the first step that will lead to an exchange or a spread of resources. Often the catalog is the origin of the initiatives for the rapprochement between the actors who wish to share their SPOT image processing chains for the monitoring of forest-cover. We would like to stress that it does not answer a move to share the processing chains but the one to share the metadata. We take profit from this possibility that the processing chains modeled and formalized in this way can be included into a catalog.

Processing chain exchange: We can take profit from the processing chain exchange by taking into account the different human resources which are proposed for monitoring change in forest cover. As a matter of fact, in contract to the spread which can be made through a downloading website, the exchange of SPOT image processing chains generally requires an interaction between the actors involved. This encounter is often initiated for legal reasons in order to draw up an exchange convention that respects the objectives of each party, or each of the parties. The notion of exchange is sometimes mixed with that of partnership, much broader. A partnership is a process of strategy cooperation between at least to actors and whose objective is to achieve a goal through the common use of material, intellectual, human, financial means. It often has a political dimension. The exchange of processing chains is interesting in setting up a process concerning the physical realization of one of the aspects of partnerships between the different actors. We stress that the notion of exchange implies a bilateral relation (the producer versus the receiver) which focuses on a flow while the notion of sharing also allows considering all the related aspects, in particular the appropriation and sharing of the knowledge which can result the exchange.

Spread of processing chains: More modest on the level of interactions, the objectives on the spread are not a challenge for the capacities of appropriation of processing chains by the actors with who the producer will have little or even no interaction. It can be pointed out here that there is a marked opposition with the co-production objective which, instead of relying on work groups, prefers to spread the produced processing chains without any discussion. It's the logic of the action of the researchers who wish to make their processing chains accessible to a great number of people.

Mutualization of processing chains: The interest in the formalization of chain-processing is to encourage the multualization of experience experiments between the researchers and is justified by various reasons. The reason is first of all to avoid useless duplication of efforts in the constitution of chain processing, avoiding at the same time redundancy of the result fragmentation which comes out. Considering the compartmentalization related to the multiplication of autonomous and non-coordinated chain-processing, the formalization of these chains represents an adequate solution to these defects.

Finally, our proposals on the formalization of processing chains constitute necessary conditions for the mutualization of experiments for monitoring forest cover. In fact, the mutualization of data implies formalizing the production and spread supervision modalities, according to an interoperable mode of a piece of information ready for use and whose quality is checked, for reason of common usage. By renforcing the cooperation between producer and users of processing chains, the mutualization also favours the exchange of experience and good practice through networking.

Processing-chain sharing: At last, it can be said that our proposals on the processing-chain formalization facilitate the sharing of experiments to monitoring change in forest cover. As a matter of fact, the data sharing aims at providing the community of experiment-users with coherent formats and structures, which will enable them to fulfill their missions better and produce their experimentations according to common and pre-defined requirements. Conversely, this interoperability and the improvement of processing-chain availability can contribute to reinforcing the relationships between the producer and user organizations and therefore can help to fill up some gaps on the use of SPOT satellite images for the monitoring of forest cover.

Experiment-sharing platform: Starting from the advantages offered by over proposals on the modelling and formalization of SPOT image processing chains for the monitoring of forest cover we are going to present other advantages which come out. We can state that the formalization of processing chains can be valued by the setting-up of a platform for sharing, mutualization re-use and the spread of experiments on the valuation of SPOT images for the monitoring of forest-cover.

The knowledge on the SPOT image processing methods for the monitoring of forest-cover are modelled then formalized in the form of processing-chains. The knowledge acquired on SPOT image-processing has to be made available to the actors. This capitalization is conceived so that everyone's experiment does not remain continued to individual level, but serves the collective in a move for knowledge sharing allowing to reduce the uncertainty in concerned field. The preservation and transmission of experiments and knowledge through formalized processing-chains facilities the implementation of new experiment protocols using SPOT images. More other, the capitalization and valuation of experiments using processing-chains is part of knowledge management. It means that the starting point is strong hypothesis that all experience or knowledge can by organized, referenced, enriched on supports which are adapted and exchanged as knowledge that other people can make their own. This puts the stress on the re-use, interoperability and sharing, the re-use, the spread of processing-chains to

valve the new SPOT images available as well as the new innovating methods for the monitoring of forest-cover.

The interests in formalizing processing-chains can be valued by the proposal on setting up a platform while allows the sharing, the mutualisation the spread, the naming of SPOT image processing-chains through catalogs for the monitoring of forest-cover. In addition, the existing platforms at the world level propose functionality for knowledge management following this view; we have proposed platforms for sharing SPOT images to deal with the uncertainty due to the lack of image of this type.

Knowledge management consists in grousing, sharing, and updating knowledge. It requires not only setting up mechanisms and procedures for uniting, organizing, presenting and spreading SPOT image processing-chains which are modelled and partners, but also proceeding to the assessments of these operation-results.

We have demonstrated how to produce SPOT image processing-chains which are reusable, sharable, spreadable, interoperable, and mutualizable. The step which can follow is implementation of a platform for sharing experiments. In view the interest the valorisation of our proposals, we can draw up a deduction that the solutions we found represent a help in the formalization of SPOT image processing-chains to capitalize the knowledge of the monitoring of the evolution of forest-cover.

IV CONCLUSION AND PERPECTIVES

This article shows the possibility for sharing and mutualizing the experiments or existing SPOT image-processing to deal with the uncertainly of the monitoring of forest-cover using the technique of remote sensing. The proposal on resource modelling and formalization for their sharing is one of the efficient techniques for resolve the problems of heterogeneity of SPOT image-processing. In fact the experiments or the processing methods of these images used for the monitoring of forest-cover will be re-usable for non-remote-sensing-expert-users. This decreases the uncertainly of the use of SPOT image processing in the forest area.

Our wok is limited to modelling, formalization and design of a resource-knowledge base related to the work-context. This latter has its advantage because if can be enriched by new recently-produced resources. It helps integrate new satellite images and recent processing methods for analysing forest cover.

Next, the perspective of this work is the integration of resources formalised in that way. This formalization makes the resources interoperable, which makes it possible to integrate the resources easily into the MDWeb tools for the sharing and mutualization and WPS (Web Processing Services), for the execution. For the execution of processing-chains via the web, we propose the web WPS Service server OGC standard (2012) since 2005.The specification of the WPS is in the form of a generic interface allowing describe and carry out or execute image processing chains of satellite images, according to Machet et al. (2008) and Eberle and Strobl (2012), which we produced in this work. The WPS is based on the http protocol and XML language. The processing-chains presented in this work will be converted into XML files, and will be integrated into WPS for execution.

References

Achard F., Eva H.D., Stibig H.-J., Mayaux P., Gallego J., Richards T., Malingreau, J.-P. (2002). Determination of deforestation rates of the world's humid tropical forests. Science 297:999-1002.

BnF, (2008). Guide d'utilisation du Dublin Core (DC) à la BnF : Dublin Core simple et Dublin Core qualifié, avec indications pour utiliser le profil d'application de TEL, Version 2.0. Bibliothèque nationale de France /Direction des Services et des Réseaux / Département de l'Information bibliographique et numérique, France.

Desconnets J-C., Libourel T., Clerc S., Granouillac B., (2007). Cataloguing for distribution of environnemental ressources. *10th AGILE, International Conference on Geographic Information Science*, Aalborg University, Danmark.

Eberle J. and Strobl C., (2012). WEB-Based Geoprocessing and Workflow Creation for Generating and Providing Remote Sensing Products. Geomatica, Vol.66(1), pp.13-26. Canadian Institute of Geomatics.

Hajalalaina A. R., Grizonnet M., Delaître E., Rakotondraompiana S., Hervé D., 2013. Discrimination des zones humides en forêt malgache, proposition d'une méthodologie multirésolution et multisource utilisant ORFEO ToolBox. *Revue Française de Photogrammétrie et de Télédétection*, n° 201, pp. 37-48.

ISO19115, (2003). Geographic Information Metadata, ISO 19115. International Organization for Standardization (ISO), Genève, Suisse.

ISO19119, (2005). Geographic Information Service, ISO 19119. International Organization for Standardization (ISO), Genève, Suisse.

Kimes D.S., Nelson R.F., Salas W.A. and Skole D.L., (1999). Mapping secondary tropical forest and forest age from SPOT HRV data. *in International Journal of Remote Sensing*, 20:3625–3640.

Libourel T., Lin Y., Mougenot I., Pierkot C., (2010). A platform dedicated to share and mutualize environmental applications. In J. Filipe, J. Cordeiro J. (eds.), *in Proceedings of the 12th international conference on enterprise systems. ICEIS, International Conference on Enterprise Systems, 12.*, Madere Funchal, 8-12 juin 2010. Setubal : SciTePress, p. 50-57.

Lin Y., Mougenot I., Libourel T. (2008). Un nouveau langage de workflow pour les sciences expérimentales. *In INFORSID'08 : Atelier ERTSI Evolution, Réutilisation et Traçabilité des Systèmes d'Information*, Fontainebleau, France.

Machet E., Kamhi M., Jacquin M., Le Page M., Dejoux J.-F. and Dedieu G., (2008). Web Processing Service pour le traitement des images satellites. *CESBIO, CNES, Toulouse, France.*

OGC. (2012). Web Processing Service 2.0 Standard Working Group. http://www.opengeospatial.org/projects/groups/wps2.0swg. Accessed: 2016-04-29

Orfeo ToolBox, (2015). The ORFEO Tool Box Software Guide Updated for OTB-5.2.1. https://www.orfeo-toolbox.org/packages/OTBSoftwareGuide.pdf. Accessed: 2016-04-29. Centre d'Etudes Spatiales (CNES), Toulouse, France

Souza C., Firestone L., Silva L.M.,, Roberts D., (2003). Mapping forest degradation in the Eastern Amazon from SPOT 4 through spectral mixture models. *in Remote Sensing of Environment*, 87:494–506.

Spatial accuracy quantification in mapping

Uncertainty quantification of interpolated maps derived from observations with different accuracy levels

Gerard B.M. Heuvelink*[12], Dick Brus[2], Tom Hengl[1], Bas Kempen[1], Johan G.B. Leenaars[1] and Maria Ruiperez-Gonzalez[1]

[1]ISRIC - World Soil Information, The Netherlands
[2]Wageningen University Research, The Netherlands

*Corresponding author: gerard.heuvelink@wur.nl

Abtract
Most practical applications of spatial interpolation ignore that some measurements may be more accurate than others. As a result all measurements are treated equally important, while it is intuitively clear that more accurate measurements should carry more weight than less accurate measurements. Geostatistics provides the tools to perform spatial interpolation using measurements with different accuracy levels. In this short paper we use these tools to explore the sensitivity of interpolated maps to differences in measurement accuracy for a case study on mapping topsoil clay content in Namibia using kriging with external drift (KED). We also compare the kriging variance maps and show how incorporation of different measurement accuracy levels influences estimation of the KED model parameters.

Keywords
Africa, geostatistics, interpolation, kriging, measurement error, regression, soil

Spatial interpolation errors are an important source of uncertainty in many spatial modelling applications and analyses. Geostatistics provides the tools to quantify interpolation error through the so-called kriging variance or by using spatial stochastic simulation, but in standard kriging the measurement error of individual observations is rarely addressed explicitly. It is usually represented as a component of the nugget variance of the semivariogram, but this implicitly assumes that all measurements are unbiased and have the same random measurement error variance. In reality, different measurement precisions and accuracies may occur because the data used may be a merge of field estimates and laboratory measurements, may be measured using different instruments and laboratory methods, or may be derived using the same methods but in different laboratories. Often data are also measured indirectly through proxies, such as when soil properties are estimated from soil spectroscopy signals that are converted to soil property values using statistical methods such as Partial Least Squares Regression (Brown et al. 2006, Leone et al. 2012). In recent years observations are also increasingly generated through crowd-sourcing and volunteered geographic information initiatives, which may suffer from large measurement errors (Goodchild and Li 2012). These initiatives can yield large volumes of data at cheap or zero cost, but their accuracy will usually be less than that of institutional data. In this work we extend kriging with external drift to the case in which each individual observation can have a different measurement error variance. As

a result observations with small measurement error variance carry more weight than observations with large measurement error variance, both in regression modelling and kriging.

The methodology builds on well-known approaches in geostatistics that go back as far as Delhomme (1978) and is also presented in text books such as Chilès and Delfiner (1999, Section 3.7.1). It boils down to modification of the kriging matrix by adding the measurement error variances to the diagonal elements of the covariance matrix. In case of correlated errors, the off-diagonal elements will also be affected. Solving the kriging system using the modified kriging matrix automatically decreases the kriging weights of observations with larger measurement error variances. Also, the influence of measurements on estimation of the trend coefficients is reduced when the measurement error variance is larger. While this is all well known, it is rarely applied in practice. This is unfortunate, because differences in measurement errors may have a large impact on resulting maps and hence the prediction accuracy can be markedly improved if these differences were taken into account. Differences in measurement error variances are typically also not included in estimation of the semivariogram and trend coefficients. In this presentation we show that measurement error variance can fairly easily be included in parameter estimation (both for estimation of the variogram parameters and regression coefficients) by taking a maximum likelihood estimation approach. Further to that, we account for systematic measurement errors by representing the (unknown) systematic error as a zero-mean random variable that is equal for all observations from the same source. Uncertainty about the variogram parameters can be incorporated by taking a Markov Chain Monte Carlo approach.

The statistical methodology is largely known and fairly straightforward but requires adaptations of existing software implementations. We implemented the methodology as R scripts, which in future will be extended to scalable R functions. We use a digital soil mapping application using topsoil texture data from the Africa Soil Profiles database (Leenaars 2013) and the LandPKS project (Herrick et al. 2013) to map topsoil clay content for Namibia. We compare prediction maps and prediction error variance maps with those obtained when measurement error is ignored. Results show marked differences and indicate that measurement errors should not be ignored, particularly when there are large differences in accuracy levels between observations within the conditioning dataset. We also explore the sensitivity of mapping results for different degrees of spatial autocorrelation of measurement errors.

One important reason that the methodology has not often been applied in practice is that it requires that the measurement error variances of all observations are known. In reality, this is seldomly the case because data come from many sources and their accuracies are rarely recorded. In our case study we used the texture triangle and expert judgement and expert elicitation (O' Hagan et al., 2006) to quantify the measurement error variances, but these are no substitute for real values and hence it is important that point data used for spatial interpolation are routinely accompanied by measures of their accuracy.

Extension from linear multiple regression and regression kriging to non-linear machine-learning regression methods, such as artificial neural networks, support vector machines and random forests, is less obvious. One approach might be to duplicate more accurate observations or assign weights to observations depending on their measurement error variance,

but this can only partly solve the problem and has a large ad hoc character. More satisfactory solutions should be found, but these should be sufficiently generic and work for the entire family of machine-learning methods, because in many practical machine-learning applications blends of multiple algorithms are used to optimise performance. This is important too, because in recent decades spatial interpolation makes use more and more of explanatory information contained in covariates and machine-learning algorithms are increasingly popular because they are more flexible than linear methods and usually produce more accurate predictions (e.g. Hengl et al, 2015).

References

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma 132(3–4)*, 273–290.

Chilès, J-.P., Delfiner, P. (1999). *Geostatistics. Modeling Spatial Uncertainty*. New York: Wiley.

Delhomme, J.P. (1978). Kriging in the hydrosciences. *Advances in Water Resources 1(5)*, 251–266.

Goodchild, M.F., Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics 1*, 110–120.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E. (2015). Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. *PLoS ONE* 10(6): e0125814.

Herrick, J.E., Urama, K.C., Karl, J.W., Boos, J., Johnson, M.V.V., Shepherd, K.D., Hempel, J., Bestelmeyer, B.T., Davies, J., Guerra, J.L., Kosnik, C., Kimiti, D.W., Ekai, A.L., Muller, K., Norfleet, L., Ozor, N., Reinsch, T., Sarukhan, J., West, L.T. (2013). The global Land-Potential Knowledge System (LandPKS): Supporting evidence-based, site-specific land use and management through cloud computing, mobile applications, and crowdsourcing. *Journal of Soil and Water Conservation 68(1)*, 5A-12A.

Leenaars, J.G.B. (2013). *Africa Soil Profiles Database, Version 1.1. A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset)*. ISRIC Report 2013/03. Africa Soil Information Service (AfSIS) project. ISRIC – World Soil Information, Wageningen, the Netherlands. 160 pp.

Leone, A.P., Viscarra-Rossel, R.A., Amenta, P, Buondonno, A. (2012). Prediction of soil properties with PLSR and vis-NIR spectroscopy: application to Mediterranean soils from Southern Italy. *Current Analytical Chemistry 8(2)*, 283–299.

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Chicester: Wiley.

Survey designs which maximize efficiency gains in ALS-based forestry plot imputation

Gavin Melville[1], Christine Stone[2], Jan Rombouts[3]

[1]Trangie Agricultural Research Centre, Mitchell Hwy, Trangie 2823, Australia
[2]NSW Department of Primary Industries Forest Research, Level 12, 10 Valentine Ave, Parramatta 2124, Australia
[3]One FortyOne Plantations, 152 Jubilee Hwy E., Mt Gambier 5290 Australia

*Corresponding author: Gavin.Melville@dpi.nsw.gov.au

ABSTRACT

The use of airborne laser scanner (ALS) data to estimate forest resource inventory variables is now becoming widespread in Australia (Rombouts et al., 2010; Stone et al., 2011). ALS data is combined with survey plot data to construct model-based estimates of timber volume. In particular, volumes of timber products, sourced from plantations of *Pinus radiata* have successfully been estimated using nearest neighbor methods. One of the challenges in this approach is to construct samples which capture the full efficiency gains which are achievable using the multiplicity of variables which can be derived from the ALS data. Some of the sampling design approaches that have been investigated include random sampling, grid sampling, stratification, systematic selection and balanced sampling. Estimates have been examined from both a design-based and model-based perspective (Melville et al., 2015). This talk will present results based on several approaches including a novel method specifically designed for imputation which optimizes the survey design by using the distance properties of the sample in the space defined by the auxiliary variables.

I INTRODUCTION

The use of remote sensing to measure key inventory metrics over large areas of forest has become widespread in recent years (e.g. Breidenbach et al., 2010; Latifi et al., 2010; McRoberts, 2012; Hudak et al., 2014; Rombouts et al., 2014; Dash et al., 2015). One of the remote sensing techniques currently being employed in Australian softwood plantations is LiDAR, also referred to as airborne laser scanning (ALS). It has been established that the use of LiDAR data in conjunction with ground measurement leads to efficient prediction of commercially important attributes including timber volume, basal area and stems per hectare (Rombouts et al., 2010).

LiDAR and/or other auxiliary data have been used as covariates in a range of prediction methods including imputation, linear regression and machine learning methods such as random forest. With imputation, for example, forest plots which have been measured on the ground are linked to non-measured plots according to their similarity in the covariate space defined by the auxiliary variables. Specific information relating to key attributes, such as timber volume, is assigned to non-measured plots from the most closely related measured plots as defined by the similarity structure.

Measurement of field plots is a time consuming and costly process requiring specialized inventory crews who often need to access remote and challenging terrain. In order to be effective the field plots must cover the full range of variability in the key attributes. Therefore, appropriate

methods of selecting the field plots are essential in constructing imputation estimates which are efficient, robust and economical.

There are a variety of sampling designs which can lead to a good spread in the attributes of interest. These include methods which are focused on structural attributes such as stratification, and spatially systematic designs such as grid sampling. Methods have also been developed which utilise the auxiliary data explicitly including stratification based on LiDAR variables (Hawbaker et al., 2009), multivariate methods such as sensor-directed response surfaces (SRDS - Lesch, 2005), and methods employing geographical coordinates such as generalized random tessellation stratified sampling (GRTS - Stevens and Olsen, 2004). More recently methods have become available which construct a sample which is simultaneously balanced with respect to multiple design variables and these have also been applied to forest inventory (Grafström et al., 2014). The present study investigates a new type of sample, termed nearest centroid (NC - Melville and Stone, 2016), which uses a multivariate clustering algorithm to select a sample of field plots.

In this paper the NC sampling method is presented, and data from a *Pinus radiata* plantation in South Australia are used to illustrate the technique. Comparisons are provided with other types of plot selection strategies.

II MATERIALS AND METHODS

2.1 Inventory approach

We define the "area of interest" (AOI) as the finite region over which predictions will be made together with (if separate) the finite region over which ground-based measurements will be made. The AOI is tessellated into a set of non-overlapping contiguous pixels called "virtual" plots which are used to define the population. Virtual plots are constructed to have similar dimensions to the ground-based plots and each virtual plot is associated with a set of metrics which is calculated from the LiDAR data and used as auxiliary information. Virtual plots are further characterised as "reference" plots - the set of plots which are selected for ground-based measurement, "candidate" plots - the set of plots from which the reference plots are chosen, and "target" plots - the set of plots for which predications are made. Therefore the approach which is used in an actual forest inventory can be separated into the following steps:-

1. Define the area of interest
2. Define the area containing the candidate plots
3. Construct a population of virtual plots in areas (1) and (2)
4. Select a sample of reference plots from area (2)

2.2 Imputation

The k-nearest neighbour approach involves calculating the distance or similarity in the auxiliary space between measured reference plots and target plots in order to determine which reference plots are most similar each target plot. The auxiliary variables are chosen because of their ability to predict the variable of interest which, in commercial forests, is typically timber volume. For every target plot, the k-nearest neighbour imputation estimate is the weighted mean of the variable of interest from the k most-similar reference plots, calculated as

$$\tilde{y}_i = \frac{\sum_{j=1}^{k} w_{ij}\, y_j^i}{\sum_{j=1}^{k} w_{ij}}$$

where $\tilde{y}_i$ is the prediction for target plot i , y_j^i is the jth nearest reference plot to plot i, and w_{ij} is the weight assigned to plot j (McRoberts et al., 2007). In this study $w_{ij} = 1$ was employed throughout.

2.3 NC sample

The NC sample operates by partitioning the target population into clusters of plots having minimum sums of squares of distances, from plots to the cluster centroids. The number of clusters is chosen to be equal to the required sample size, n, and the clusters themselves are not required to be spatially contiguous. The clustering procedure which was employed in this study is termed k-means clustering (Hartigan and Wong, 1979) and was performed using the standardized auxiliary variables and a Euclidean distance metric. K-means clustering is normally used for multivariate analysis of complex datasets. After calculating the cluster centroids one then finds the plots in the candidate set which are closest to the centroids in the auxiliary space and these become the reference plots.

An example is shown in Figure 1 which illustrates the selection of 8 reference plots by forming the target set into 8 clusters in the space defined by the auxiliary LiDAR variables p1m (proportion of heights greater than 1m) and mqh (mean quadratic height). The plots closest to the cluster centroids (centroids shown as black stars) are then selected as the reference plots. The term "nearest centroid" derives from the fact that the reference plots are the nearest plots to the target plot centroids. During imputation these plots become the nearest neighbours to target plots in the same cluster (provided they are sufficiently close to the plot centroids). Generally there would more than two auxiliary variables.

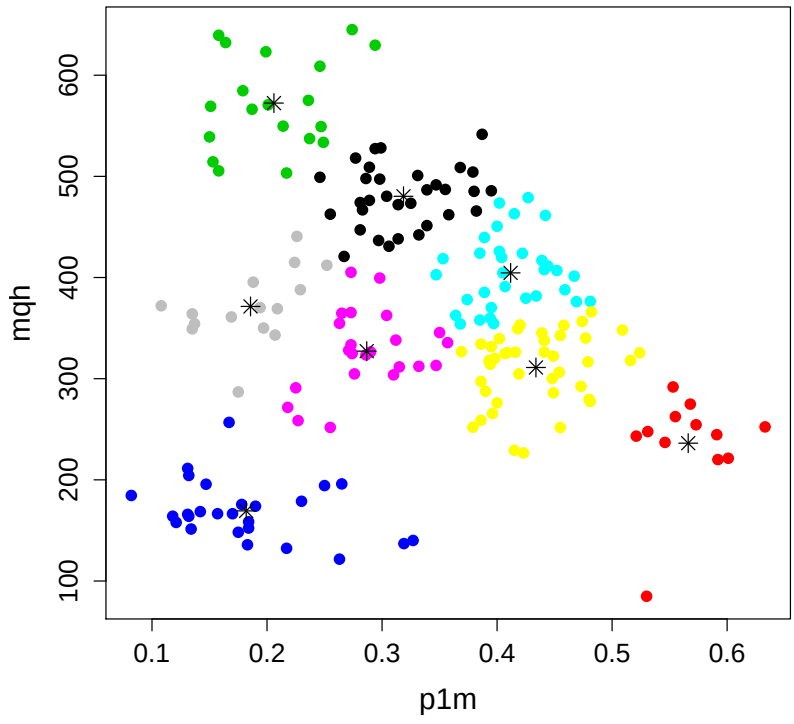


Figure 1: Illustration of NC sample using two auxiliary variables

2.4 Data

The *P. radiata* plantation used in this study is located in South Australia and occupies around 3300 hectares. The plantation is surveyed periodically to provide key product and management information. This study was based on data from 304 ground plots measured in 2012 and a list of plot attributes is given in Table 5.1 of the Final Project Report presented by Rombouts et al. (2014). The essential characteristics of the ground plots are that they have an area of 1000 m^2 and contain trees with an age range of 14-32 years and a mean age of 26.1.

LiDAR data were acquired in January 2012 by flying an aircraft equipped with an ALTM Orion device at an altitude of 800 m across the estate. The LiDAR first return data had a mean point density of 5.9 pulses m^{-2}. A detailed description of the LiDAR data specifications is provided in Table 5.2 of Rombouts et al. (2014).

Around 120 separate variables were extracted from the LiDAR point cloud and are summarized in Table 3. These variables were available for each virtual plot. Ten of these variables were eventually selected for imputation according to how well they were able to predicted the key attributes of interest using an approach described in Rombouts et al. (2014).

LiDAR	first return	last return	description
pground	Y	Y	proportion of ground returns
p>x	Y	Y	proportion of heights > x (x=1,2,5,10m)
sd	Y	Y	standard deviation of heights
skew	Y	Y	skewness of height distribution
kurtosis	Y	Y	kurtosis of height distribution
h	Y	Y	mean height
hmax	Y	Y	maximum height
hmax4	Y	Y	four highest in each plot quadrant
hx	Y		x% percentile height (x=10%∼100%)
d0.x	Y	Y	proportion of heights between 0% and x% hmax (x=10∼100%)
h>0	Y		mean height (heights > 0m i.e. vegetation)
mqh>0	Y		mean quadratic height (heights > 0m)
mqh>1	Y		mean quadratic height (heights > 1m)
scanangle	Y		mean scan angle in the plot
non LiDAR			
lop			thinning status (last operation)
nsq			site quality index
age			plantation age

Table 1: Auxiliary variables available for imputation models

2.5 Simulation approach

Three sampling methods were examined as part of this study. An approach often used in forest inventories is to place a grid over the AOI and select plots at the grid intersection points. LiDAR data are not required to construct either a grid sample or a completely random sample. The use of LIDAR data as *a priori* information in the sample design is aimed at obtaining efficiency gains from the plot selection process. In this study the sampling methods which were evaluated, in addition to random sampling, were locally balanced sampling (Grafström et al., 2014) and the proposed NC method.

In the sampling simulations below, the 304 study plots were divided at random into a target set (200 plots) and a candidate set (104 plots). Reference plots were selected from the candidate plots using each of three sampling strategies i.e. random, locally balanced, and NC. The prediction method used throughout was Euclidean imputation with $k = 1$. The number of reference plots was fixed at either 10, 25, 50 or 75 and the variable of interest was the timber volume in each plot.

The various sampling strategies were evaluated in terms of how well the known variable of interest was predicted. The simulations were repeated 10,000 times with each realisation comprising a new set of target plots, candidate plots and reference plots. Comparisons were done in terms of the relative bias (RB %) and the relative root mean squared error (RMSE %). The relative bias was calculated at the AOI level and is defined as

$$RB = \frac{\sum_j (\hat{Y}_j - Y_j)}{\sum_j Y_j},$$

where $\hat{Y}_j$ is the estimate of total timber volume over the AOI for the j'th realisation and Y_j is the actual total timber volume over the AOI for the j'th realisation. At the AOI level the relative root mean squared error is defined as

$$RRMSE = \frac{\sqrt{\frac{1}{B}\sum_j (\hat{Y}_j - Y_j)^2}}{\frac{1}{B}\sum_j Y_j},$$

where B is the number of realizations. Relative efficiency measures were also calculated for each sampling strategy, using the simple random sample as a benchmark. For any pair of sampling strategies the relative efficiency is calculated as the squared ratio of the AOI-level RMSE values.

III RESULTS

Table 2 presents the simulation results for the three sampling strategies and the four sample sizes which were investigated.

Sampling method	Sample size	Relative Bias (%)	Relative RMSE (%)	Relative efficiency
Random	10	-0.1	4.7	1.0
Balanced		0.2	3.8	1.5
NC		-0.1	3.1	2.2
Random	25	0.2	2.8	1.0
Balanced		0.1	2.3	1.4
NC		0.5	2.1	1.7
Random	50	0.3	1.8	1.0
Balanced		0.2	1.7	1.1
NC		0.7	1.6	1.2
Random	75	0.2	1.6	1.0
Balanced		0.3	1.6	1.1
NC		0.5	1.5	1.2

Table 2: Summary comparisons of bias, accuracy and relative efficiency of three sampling strategies for predicting mean timber volume using field plot measures and LiDAR data

The relative bias is very small with these data, irrespective of the sampling method. With small sized samples the relative RMSE varied from 4.7% using a random sample, to 3.1% using the NC sample. Hence the relative efficiency of the NC sample is 2.2 times that of the random sample. The efficiency gains are most pronounced when the sample size is small. For larger samples, e.g. n= 75, the relative efficiency of the NC sample, compared to the random sample, is 1.2. The relative RMSE results are also presented graphically in Figure 2.

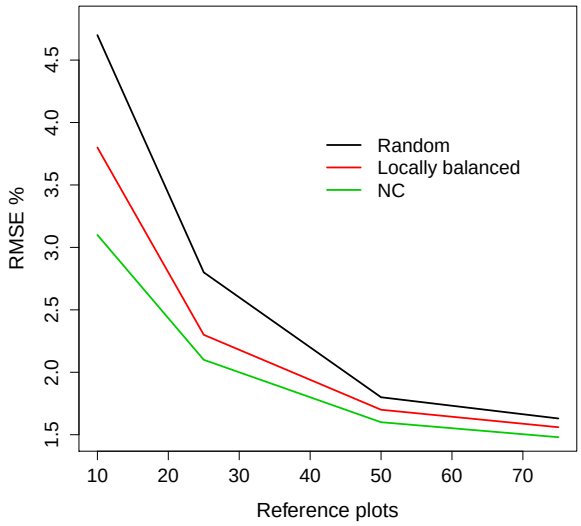


Figure 2: Relative RMSE % vs sample size for three sampling schemes

IV DISCUSSION

This paper describes an approach to inventory design which is specifically aimed at imputation and essentially different to other sampling methods. The results in Table 2 illustrate efficiency gains which are more than double that of random samples. The implications for inventory design are that surveys can be constructed with around half the number of plots that would be required using a conventional sampling approach. The method can be employed either within or across strata although the results in this paper were achieved without stratification.

Note that using remotely sensed data for survey design necessitates having the data available prior to plot selection. Therefore either the data need to be newly acquired or need to be available from a previous campaign. Where existing data are used, they need to be sufficiently recent. The key criteria in this respect is the extent of the correlation between the remotely sensed data and the variables of interest.

One of the primary advantages of the NC sampling method is its flexibility. Small area estimation (SAE) provides a good illustration of this. SAE uses reference plots which are mostly (or completely) outside the AOI to make predictions within the AOI. Most of the existing sampling strategies, including balanced sampling, are not suited to SAE because it is not possible to target the sample specifically to the AOI. There may or may not be reference plots in the sample which are similar enough to the target plots to provide acceptable imputation estimates. By partitioning the small area into defined clusters, the NC sample permits the selection of reference plots which are closely matched to the target plots.

It is proposed to make the clustering approach available as an R function (R Core Team, 2015) to enable sample selection in any application where auxiliary data are available for a population of discrete units such as forestry plots.

V CONCLUSIONS

The sampling method presented in this paper is an intuitive approach to forest inventory where imputation methods are to be used. The method has been trialled on *P. Radiata* datasets from four separate locations in eastern Australia. In every case it has proved to be superior to other sampling methods. With small samples in particular, relative efficiencies are substantially higher than random sampling methods and moderately higher than balanced sampling strategies.

References

Breidenbach J., Nothdurft A., Kändler G. (2010). Comparison of nearest neighbour approaches for small area estimation of tree species-specific forest using airborne laser scanner data. *European Journal of Forest Research 129*, 833–846.

Dash J. P., Marshall H. M., Rawley B. (2015). Methods for estimating multivariate stand yields and errors using k-nn and aerial laser scanning. *Forestry 88*, 237–247.

Grafström A., Saarela S., Ene L. T. (2014). Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Canadian Journal of Forest Research 44*, 1156–1164.

Hartigan J. A., Wong M. A. (1979). A k-means clustering algorithm. *Applied Statistics 28*, 104–108.

Hawbaker T. J., Keuler N. S., Lesak A. A., Gobakken T., Contrucci K., Radeloff V. C. (2009). Improved estimates of forest vegetation structure and biomass with a lidar-optimized sampling design. *Journal of Geophysical Research 114*, GE00E04.

Hudak A. T., Haren A. T., Crookston N. L., Liebermann R. J., Ohmann J. L. (2014). Imputing forest structure attributes from stand inventory and remotely sensed data in western Oregon, *Forest Science 60*, 253–269.

Latifi H., Nothdurft A., Koch B. (2010). Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: apredictors. *Forestry 83*, 395–407.

Lesch S. M. (2005). Sensor-directed response surface sampling designs for characterizing spatial variation in soil p*Computers and Electronics in Agriculture 46*, 153–179.

McRoberts R. E. (2012). Estimating forest attributes for small areas using nearest neighbors techniques. *Forest Ecology and Management 272*, 3–12.

McRoberts R. E., Tomppo E. O., Finley A. O., Heikkinen J. (2007). Estimating aerial means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. *Remote Sensing of Environment 111*, 466–480.

Melville G., Stone C. (2016). Optimizing nearest neighbour information - a simple, efficient sampling strategy for forestry plot imputation using remotely sensed data. *Australian Forestry*, to appear.

Melville G., Stone C., Turner R. (2015). Application of lidar data to maximize the efficiency of inventory plots in softwood plantations. *New Zealand Journal of Forestry Science 45:9*, 1–16.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rombouts J., Ferguson I., Leech J., Culvenor D. (2010, September 14–17). An evaluation of the field sampling design of the first operational lidar based site quality survey of radiata pine plantations in South Australia. In *Proceedings of the 2010 Silvilaser Conference*, Freiburg, Germany.

Rombouts J., Melville G., Kathuria A., Stone C. (2014). Operational deployment of lidar derived information into softwood resource systems. *Final Report for Project PNC305-1213, http://www.fwpa.com.au/rd-and-e/resources/611-operational-deployment-of-lidar-derived-information-into-softwood-resource-systems.html*.

Stevens D. L. J., Olsen A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association 99*, 262–278.

Stone C., Penman T., Turner R. (2011). Determining an optimal model for processing lidar data at the plot level: results for a *Pinus radiata* plantation in New South Wales, Australia. *New Zealand Journal of Forestry Science 41*, 191–205.

Validation of Copernicus High Resolution Layer on Imperviousness degree for 2006, 2009 and 2012

Christophe Sannier[1]*, Javier Gallego[2], Jochen Dahmer[3], Geoff Smith[4], Hans Dufourmont[5], Alexandre Pennec[1]

[1] Systèmes d'Information à Référence Spatiale (SIRS) SAS, France
[2] Joint Research Centre (JRC), Italy
[3]GAF AG, Germany
[4]SpectoNatura, UK
[5]European Environment Agency (EEA), Denmark

*Corresponding author: christophe.sannier@sirs-fr.com

Abtract
The validation of a dataset such as the Copernicus Pan-European imperviousness degree high resolution layer requires considerable effort. A stratified systematic sampling approach was developed based on the LUCAS sampling frame focusing on a 2 stage stratification approach. A two-stage stratified sample of 20,164 1ha square primary sampling units (PSU) was selected over EEA39 based on countries or groups of countries which area was greater than 90,000km² and a series of omission and commission strata. In each PSU, a grid of 5 x 5 Secondary Sample units (SSUs) with a 20 m step was applied. These points were photo-interpreted on orthophotos with a resolution better than 2.5m.
Initial results based on the binary conversion of the map by applying the 30% threshold indicate a level of omission and commission errors substantially greater than the required maximum level of 15% set in the product specifications. However, this assumes that complete information is available for each PSU which is not the case. An alternative procedure was applied to the quantitative continuous data considering the sampling error due to the SSUs selection which is expected to exhibit a more realistic assessment of the amount of omission and commission.

Keywords
Sratified systematic sampling approach, LUCAS, Binomial confidence interval, Omission error, Commission error

I INTRODUCTION

Pan-European High Resolution Layers (HRL) provide information on specific land cover characteristics, and are complementary to land cover / land use mapping such as in the CORINE Land Cover (CLC) datasets (Büttner et al. 2012) as part of the Land Monitoring Service (land.copernicus.eu) of the Copernicus programme, managed by the EC. The HRLs are produced from 20 m spatial resolution satellite imagery through a combination of automatic processing and interactive rule-based classification.

Five themes have been identified so far, corresponding with the main themes from CLC, i.e. imperviousness (the level of sealed soil), tree cover density and forest type, permanent

grasslands, wetlands and water bodies. Products with an initial pixel size of 20 by 20 m are aggregated into 100 by 100 m grid cells for final pan-European mosaic products. The imperviousness layer was the first to be produced during 2006-2008 from multi-sensor, bi-temporal and ortho-rectified satellite imagery, the same as used for the CORINE Land Cover 2006 update. The production of IMD2006 covered 38 European countries (32 EEA Member States and 6 West-Balkan countries). Since the 2006 production, a time series of imperviousness has been produced for reference years 2009 and recently 2012 over the whole area covered by the 39 member and cooperating countries of the European Environment Agency (EEA) representing a total of 6 million km². For each year it is available as a raster layer with 20 m resolution. At the time of undertaking this study, the area delivered for the year 2012 was around 90% of the total area and the study is based on this area (Spain, Greece, Cyprus and the French overseas regions are missing).

Built-up areas are characterized by the substitution of the original (semi-) natural land cover or water surface with an artificial, often impervious cover. These artificial surfaces are usually maintained over long periods of time. The imperviousness HRL captures the spatial distribution of artificially sealed areas, including the level of sealing of the soil per area unit. The level of sealed soil (imperviousness degree 1-100%) is produced using an automatic algorithm based on a calibrated normalised difference vegetation index (NDVI). A description of the Copernicus Imperviousness layer methodology was described by Gangkofner et al. (2010) for the 2009 update and by Lefebvre et al (2013) for the 2012 update. Similar methods were also applied in the USA for the development of the National Land Cover database (Xian et al. 2011).

A density threshold of 30% was used to derive the built-up layer from the imperviousness layer. This was not intended to be a separate product, but instead was calculated for the verification process only, because density products cannot be verified.

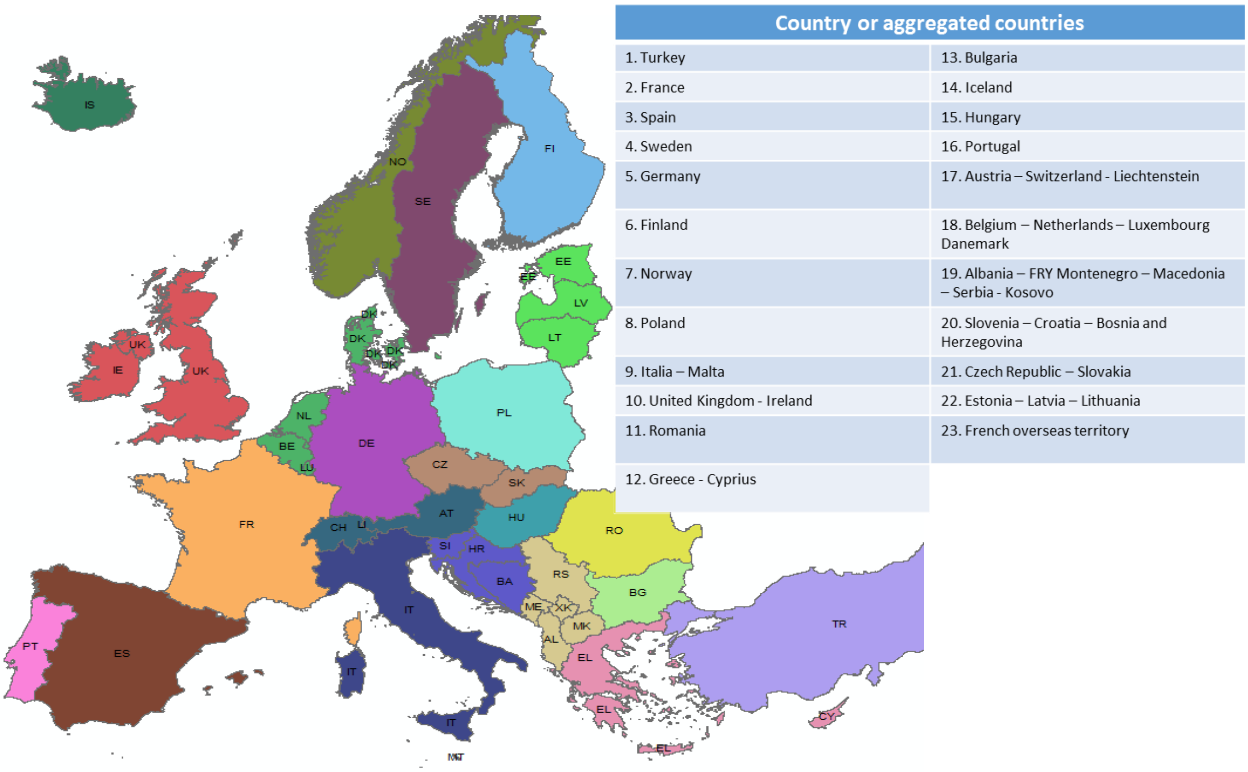
The objectives of this study were to develop and implement an accuracy assessment exercise (i) capable to confirm the results obtained by participating countries during the verification phase based on the built-up mask (Büttner, 2012), (ii) suitable to assess the accuracy of the imperviousness layer at EE39 (iii) whilst ensuring that the results can be analysed at bio-geographical region or large country level to ensure that there are no major regional differences.

II METHODS

The validation of this dataset over such a large area requires considerable and mostly unprecedented effort. Much of the literature on the assessment of the accuracy of impervious surface delineation from remotely sensed data tend to focus on relatively small areas (Ji & Jensen 1999, Yang et al. 2003, Chabaeva et al. 2007). The only comparable example is that of the National Land Cover Database of the Conterminous United States (Wickham et al. 2013) and a recent study from Hansen et al. (2014). However, this study focuses primarily on imperviousness changes and the accuracy assessment is therefore combined with that of the other land cover classes. In fact, relatively low accuracies are reported for the imperviousness change class, which is partially linked to the fact that imperviousness areas although constantly increasing, still occupy a very small portion of the total land area. In Europe, it is currently estimated that artificial surfaces represent less than 5% of the total EEA39 (Büttner et al. 2012) and it should

be expected that impervious surfaces would only represent a subset of this area, which makes it a very rare class and a particular challenge to assess its accuracy with a high degree of precision.

Wickham et al. (2013) applied a stratified random sampling approach using the land cover classes as strata. This ensured that sampling intensity is adapted to the occurrence of each class to reach a sufficient level of user's accuracy even for rare classes (Wickham et al. 2013). For the European Imperviousness layer, a similar approach was applied but based on a stratified systematic sampling approach using the EUROSTAT Land Use / Cover Area fram statistical Survey (LUCAS) sampling frame (Gallego and Delincé 2010) and focusing on a 2 stage stratification approach. The main advantage of using a LUCAS based approach is that a systematic approach ensures full traceability and it is also possible that sampling units will be shared for assessing several products, thus providing potential economies of scale. The first stage was to identify countries or groups of countries with an area greater than 90,000km² (see figure 1 below). This step allowed the analyse of the results for different countries and biogeographical regions to assess whether any heterogeneity in the quality of the data across different regions did emerge.



Country or aggregated countries	
1. Turkey	13. Bulgaria
2. France	14. Iceland
3. Spain	15. Hungary
4. Sweden	16. Portugal
5. Germany	17. Austria – Switzerland - Liechtenstein
6. Finland	18. Belgium – Netherlands – Luxembourg Danemark
7. Norway	19. Albania – FRY Montenegro – Macedonia – Serbia - Kosovo
8. Poland	20. Slovenia – Croatia – Bosnia and Herzegovina
9. Italia – Malta	21. Czech Republic – Slovakia
10. United Kingdom - Ireland	22. Estonia – Latvia – Lithuania
11. Romania	23. French overseas territory
12. Greece - Cyprius	

Figure 1. Level of reporting by country or aggregated countries as applied in the external validation.

Then as a second stage, for each country or groups of countries, omission, commission and commission change strata were determined as follows:

- Commission 2006-2009-2012-2015: Imperviousness Degree 30-100% in 2006-2009-2012

- Omission High Probability 2006-2009-2012: Imperviousness Degree 0-29% & CLC impervious classes 2006-2009-2012
- Omission Low Probability 2006-2009-2012-2015: Rest of the area 2006-2009-2012
- Commission Change 2006-2009: all changes [increased and decreased]
- Commission Change 2009-2012 : all changes [increased and decreased]

As indicated the commission strata were included to increase the precision of user's accuracy and the inclusion of a high probability omission error was to increase the precision of the producer's accuracy.

A minimum of 50-100 1ha square primary sampling units (PSU) were deemed sufficient to reach the required precision and were selected per stratum resulting in a total of 20,164 PSUs over the EEA39 area. Each 1ha PSU corresponds to a single 100m pixels in the aggregated imperviousness layer. A detailed procedure to map the imperviousness, involving field data or photo-interpretation, of all PSUs would be too time consuming and expensive. Therefore, in each PSU a grid of 5 x 5 secondary sample units (SSUs) with a 20 m step was defined (see figure 2). These SSUs were photo-interpreted against orthophotos with a spatial resolution better than 2.5m to determine if they were sealed. If a point falls on the boundary of an impervious element, a shifting rule is applied so that roughly half of the points in this situation are classified as impervious.

Figure 2: Example of SSUs organised in a 5x5 20m grid

III THEMATIC ACCURACY BASED ON 30% THRESHOLD

A 30% threshold was applied to the imperviousness degree product to convert the continuous density product to a binary mask (Büttner 2012).

Thematic accuracy is presented in the form of an error matrix made out of the results of the interpretation of the samples and their actual values in the impervious layer. As explained in (Selkowitz & Stehman, 2011), unequal sampling intensity resulting from the stratified systematic sampling approach should be accounted for by applying a weight factor (p) to each

$$\hat{p}_{ij} = \left(\frac{1}{N}\right) \sum_{x \in (i,j)} \frac{1}{\pi^{*}_{uh}}$$

sample unit based on the ratio between the number of samples and the size of the stratum considered:

(1)

where i and j are the columns and rows in the matrix, N is the total number of possible units (population) and π is the sampling intensity for a given stratum.

This is because the samples from smaller strata show a higher sampling intensity than those from the larger strata. Therefore, a correction for the sampling intensity will be applied to the error matrices produced, following the procedure described by (Selkowitz & Stehman, 2011) and applied by (Olofson et al., 2013). This leads to a weighting factor inversely proportional to the inclusion probability of samples from a given stratum. Not applying this correction could result in underestimating or overestimating map accuracies.

Initial results based on the binary conversion of the map by applying the 30% threshold indicate a level of omission and commission errors substantially greater than the required maximum level of 15% set in the product specifications. There is also considerable variability across different bio-geographical regions and group of countries and it should be noted that the 2012 layer exhibits higher accuracy to that of the 2009 and 2006 layer, which can be explained by the fact that there are always improvements made during the production of an updated dataset. However, omission errors appear to be heavily influenced by the presence of omissions in the low probability stratum which carry substantial weight. This may suggest that the stratification approach selected for omission probably needs to be revisited.

IV ANALYSIS OF IMPERVIOUSNESS DEGREE VALUES

In addition, the assessment based on the binary mask assumes that complete information is available for each PSU, which is not the case since the PSU level estimate of imperviousness is based on the SSUs. Therefore, an alternative procedure was applied to the quantitative continuous data considering the sampling error due to the SSUs selection, which is expected to exhibit a more realistic assessment of the amount of omission and commission errors.

If we had a complete information on the cell for our reference data, a reasonable measure of the commission φ and omission ψ errors would be:

$$\varphi = \frac{\sum_i pos(m_i - r_i)}{\sum_i m_i} \qquad \psi = \frac{\sum_i pos(r_i - m_i)}{\sum_i r_i} \tag{2}$$

where $pos(x)$ is the positive part, i.e. $pos(x) = x$ if x>0 and $pos(x) = 0$ if x≤0.

If the map reports a proportion m_i and the reference data give a proportion r_i ,

For each sampling unit of 100 m we have a quantitative value in the map (estimated % in the satellite image classification) and a reference value that is an estimation obtained from a sample of 25 points. The number of impervious points that we are using as reference value has a probability distribution due to the within-cell sampling. If the within-sampling is random, the number of points follows a binomial B(25, p). In our case the sampling scheme is systematic, but we use anyhow the binomial as an approximation.

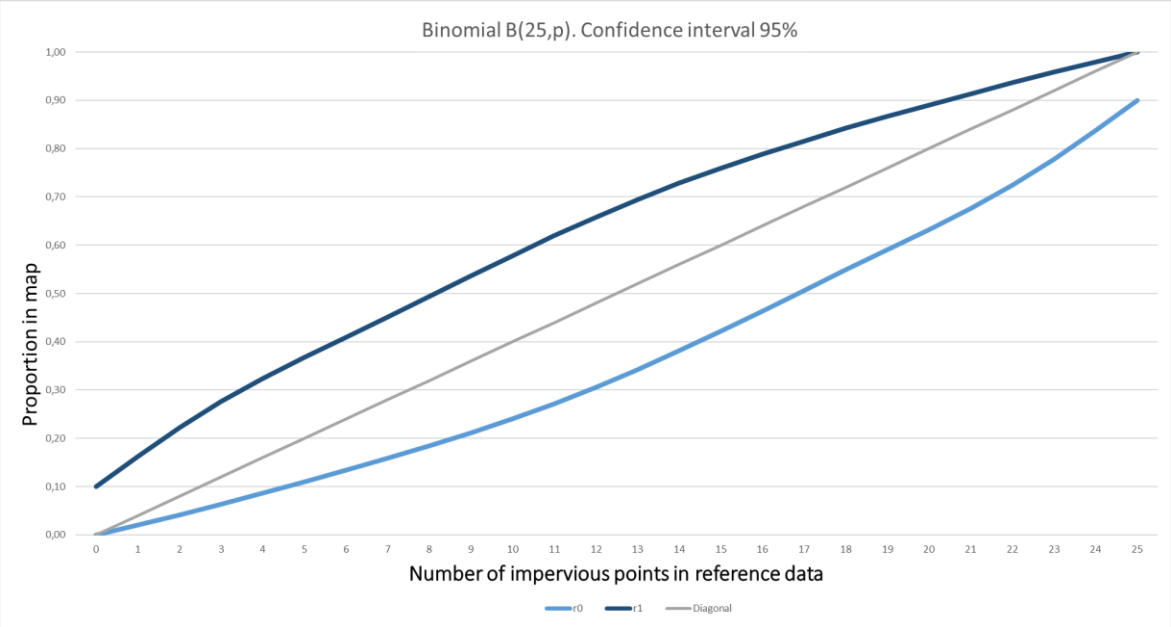


Figure 3: Representation of the behaviour of the 95% confidence interval for a 5x5 SSU grid over the whole range of imperviousness degree values

Therefore we cannot say that there is any significant disagreement if m_i lays within (r_{0i},r_{1i}), a confidence interval corresponding to B(25, r_i). Figure 3 represents the behaviour of the 95% confidence interval for B(25, r_i). Notice that only for proportions close to 0.5 we can apply the usual Gaussian approximation that leads to an interval approximately ($r_i \pm 2s_i$), while for proportions close to 0 or to 1 the intervals are strongly asymmetric.

A possible adaptation of the formulas above for the commission φ and omission errors ψ would be:

$$\varphi = \frac{\sum_i pos(m_i - r_{1i})}{\sum_i m_i} \qquad \psi = \frac{\sum_i pos(r_{0i} - m_i)}{\sum_i r_i} \tag{3}$$

The results obtained show that the amount of error is less than 10% for commission and just under 20% for omission, which exceeds the desired level of accuracy for this layer with regards to commission errors. But again this is heavily influenced by omissions included in the large area low omission strata. When removing just 39 points from the sample corresponding to these large strata (out of more than 20,000), the level of omission error is then reduced to 13%, which is within the acceptable range of accuracy expected for the imperviousness layer. This again suggests that the stratification approach for assessing omission errors needs to be revised to better target areas of potential omission.

V CONCLUSIONS

The results from the accuracy assessment exercise seem to suggest that the Copernicus Imperviousness High Resolution Layer appear to reach or even exceed the required level of

accuracy (less than 10-15% error for both omission and commission errors) particularly with respect to commission errors, but the stratification approach would need to be further improved to better target omission errors.

References

Büttner G., Kosztra B., Maucha G. and Pataki R. (2012) *Implementation and achievements of CLC2006*, ETC-LUSI, EEA, 65p.

Büttner G (2012) *Guidelines for verification and enhancement of high resolution layers produced under GMES initial operations (GIO) Land monitoring 2011 – 2013*, EEA, 47p.

Chabaeva, A., Hurd, J., & Civco, D. (2007). Quantitative assessment of the accuracy of spatial estimation of impervious cover. In *ASPRS Annual Conference Proceedigns 2007. Tampa, Florida*.

Gallego, J., & Delincé, J. (2010). The European Land Use and Cover Area-Frame Statistical Survey. *Agricultural survey methods*, 149-168.

Gangkofner, U., Weichselbaum, J., Kuntz, S., Brodsky, L., Larsson, K., & De Pasquale, V. (2010). Update of the European High-resolution Layer of Built-up Areas and Soil Sealing 2006 with Image2009 Data. 30th *EARSeL Symposium 2010: Remote Sensing for Science, Education and Culture*

Hansen, M. C., Egorov, A., Potapov, P. V., Stehman, S. V., Tyukavina, A., Turubanova, S. A., ... & Kommareddy, A. (2014). Monitoring conterminous United States (CONUS) land cover change with web-enabled landsat data (WELD). *Remote sensing of Environment*, 140, 466-484.

Ji, M., & Jensen, J. R. (1999). Effectiveness of subpixel analysis in detecting and quantifying urban imperviousness from Landsat Thematic Mapper imagery. *Geocarto International*, *14*(4), 33-41.

Lefebvre, A., Beaugendre, N., Pennec, A., Sannier, C., Corpetti, T. (2013) .Using data fusion to update built-up areas of the 2012 European High-Resolution Layer Imperviousness. In *Proceedings of the 33rd EARSeL Symposium Conference*, Matera, Italy, 3–6 June 2013; pp. 321–328.

Olofsson, P., Foody, G. M., Stehman, S. V., & Woodcock, C. E. (2013). Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, 129, 122–131.

Selkowitz, D. J., & Stehman, S. V. (2011). Thematic accuracy of the National Land Cover Database (NLCD) 2001 land cover for Alaska. *Remote Sensing of Environment*, June 2011, 115(6), 1401–1407.

Wickham, J. D., Stehman, S. V., Gass, L., Dewitz, J., Fry, J. A., & Wade, T. G. (2013). Accuracy assessment of NLCD 2006 land cover and impervious surface. *Remote Sensing of Environment*, *130*, 294-304.

Yang, L., Xian, G., Klaver, J. M., & Deal, B. (2003). Urban land-cover change detection through sub-pixel imperviousness mapping using remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, *69*(9), 1003-1010.

Xian, G., Homer, C., Dewitz, J., Fry, J., Hossain, N., & Wickham, J. (2011). Change of impervious surface area between 2001 and 2006 in the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, *77*(8), 758-762.

The polygon overlay problem in electoral geography

Romain Louvet[*1,2], **Jagannath Aryal**[2], **Didier Josselin**[1,3], **Christèle Marchand-Lagier**[4], **Cyrille Genre-Grandpierre**[1]

[1]UMR ESPACE 7300 CNRS, Université d'Avignon, France
[2]University of Tasmania, Discipline of Geography and Spatial Sciences, School of Land and Food
[3]LIA, Université d'Avignon, France
[4]CHERPA, Aix-en-Provence, LBNC, Université d'Avignon, France

*Corresponding author: romain.louvet@alumni.univ-avignon.fr

Abstract

We developed an algorithm for reducing geometric differences between a source and a target dataset. The algorithm tackles the polygon overlay problem in electoral geography before using areal interpolation methods. Our results show that improvement in matching between statistical areas and polling areas can reduce up to 40% of areal interpolation errors. This is applied to two case studies: the city of Avignon, France, and the city of Hobart, Australia.

Keywords

polygon overlay problem, areal interpolation, spatial disaggregation, spatial aggregation, COSP

I INTRODUCTION

When we explain elections based on socio-spatial context, one main methodological issue is the need to compare variables from different sources, *i.e.* electoral results with sociological and economical variables. The areal units used for mapping these variables, being designed for different purposes, usually come with different spatial resolutions and different boundaries.Therefore, it is an issue when studying the relationship between variables or trying to compare data over time (Fotheringham and Rogerson, 2013). This issue has been defined as one example of the Change Of Support Problem (COSP) called the polygon overlay problem when dealing with incompatible area to area spatial data (Gotway and Young, 2002). Being confronted with this problem, it is necessary to reallocate data from a source dataset to a target dataset, or in other words from the areal units with available data to the areal units of interest, by using areal interpolation methods.

Areal interpolation is a widely known and studied problem in spatial science and encountered with many type of data (Carson, 2013; Fotheringham and Rogerson, 2013). Many methods already exist (Tobler, 1979; Mugglin et al., 1999; Eicher and Brewer, 2001; Mennis and Hultgren, 2006; Reibel and Agrawal, 2007; Krivoruchko et al., 2011; Zhang and Qiu, 2011; Qiu et al., 2012; Lin et al., 2011), have been compared (Goodchild et al., 1993; Lam, 1983; Carson, 2013; Fotheringham and Rogerson, 2013; Do, 2015), and have been implemented. This paper does not propose a new method of areal interpolation but a process to reduce areal interpolation error by improving matching between source and target data.

Areal interpolation methods model spatial distribution based on strong assumptions such as homogeneity, isotropy, and stationarity. These assumptions, more fitted to natural phenomenon,

can fail to model human spatial distribution. Although intelligent and sophisticated methods can improve greatly their fitness to the actual spatial distribution, they are still subjected of generating interpolation errors. Therefore the polygon overlay problem is raising questions about the actual accuracy of analyses of electoral behaviour in their spatial context.

Another less common approach to this problem would be trying to reduce the mismatch between the source and target dataset. Such a method is based on aggregating algorithms which create new areal units optimizing an objective. This idea of an automated zoning procedure (AZP) was developed by Openshaw (1977) to solve the Modifiable Areal Unit Problem (MAUP) and has been applied as a solution to the polygon overlay problem by Martin (2003). Creating more fitted areal units could be indeed useful in the case of electoral geography because both geometric and attributes accuracy of polling areas are actually questionable (Bernard et al., 2015).

In this study, we combine areal interpolation with our algorithm, similar to AZP and sliver polygons eliminating tools, in order to reduce first the differences between the source and target dataset before using areal interpolation. Then it is expected that the results of areal interpolation will have fewer errors. This idea was applied to two case studies, one in France, and the other in Australia.

II MATERIAL AND METHODS

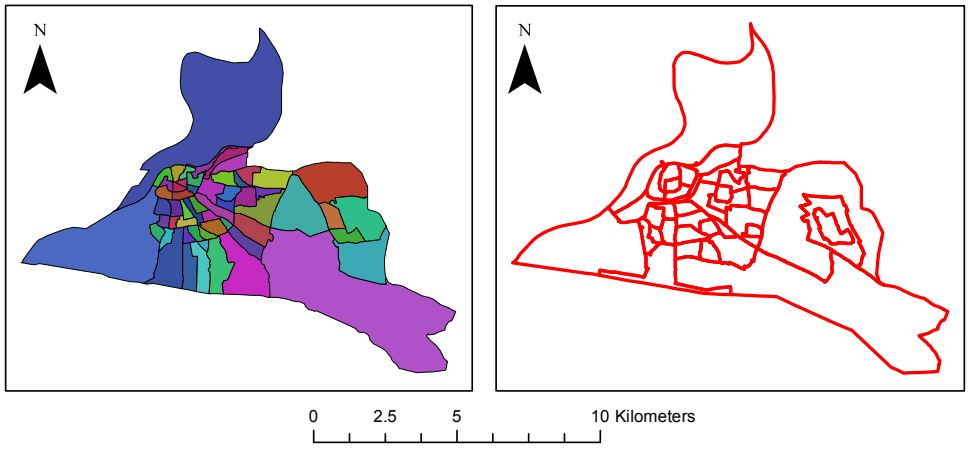


Figure 1: 2012 national election polling areas (left) and 2010 IRIS statistical areas (right), respectively target and source data of the city of Avignon, France

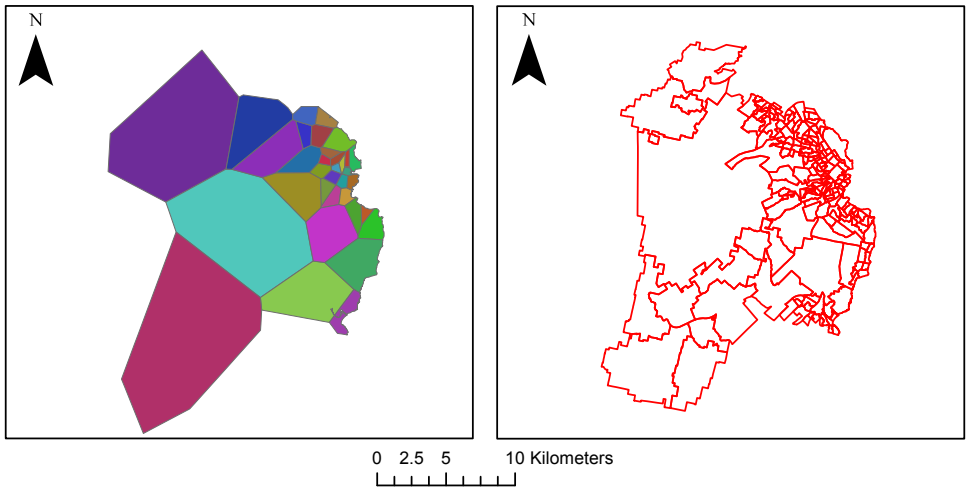


Figure 2: 2013 federal election polling areas (left) and 2011 SA1 statistical areas (right), respectively target and source data of the city of Hobart, Australia

We are comparing two study areas with similar populations but different densities and polling systems: Avignon (figure 1) and Hobart (figure 2).

We choose to use four areal interpolation methods, based on: area weighting, binary dasymetric, Kriging, and Geographically Weighted Regression. As ancillary data, we used building areas from land use for Avignon, and buildings polygons and points for Hobart. For both study areas, we used the number of dwellings as explanatory variable.

0. **given**
0.1. *source* **as** statistical area polygons
0.2. *target* **as** polling area polygons
1. *intersect* = geometric intersection of *source* **and** *target*
2. *selected layer* = selection **from** *intersect*
3. **loop for** each *feature* **in** *intersect*
3.1. **if** *feature* **in** *selected layer* **do**
3.1.1. *feature src* = *feature* identifier **in** *source*
3.1.2. *feature trgt* = *feature* identifier **in** *target*
3.1.3. *neighbours* = **get** *feature neighbours*
4. **loop for** each *feature neighbour* **in** *neighbours*
4.1. *neighbour src* = *neighbour* identifier **in** *source*
4.2. *neighbours trgt* = *neighbour* identifyer **in** *target*
4.3. **if** *neighbours trgt* == *feature trgt*
4.3.1. **pass**
4.4. **else if** *neighbour src* == *feature src*
4.4.1. *lengths* = **list** of shared line lengths between *feature* **and** *neighbour*
5. **set** feature new id **in** target **as** the identifier of its neighbour **with** maximum value in lengths

Figure 3: Our algorithm in pseudo code

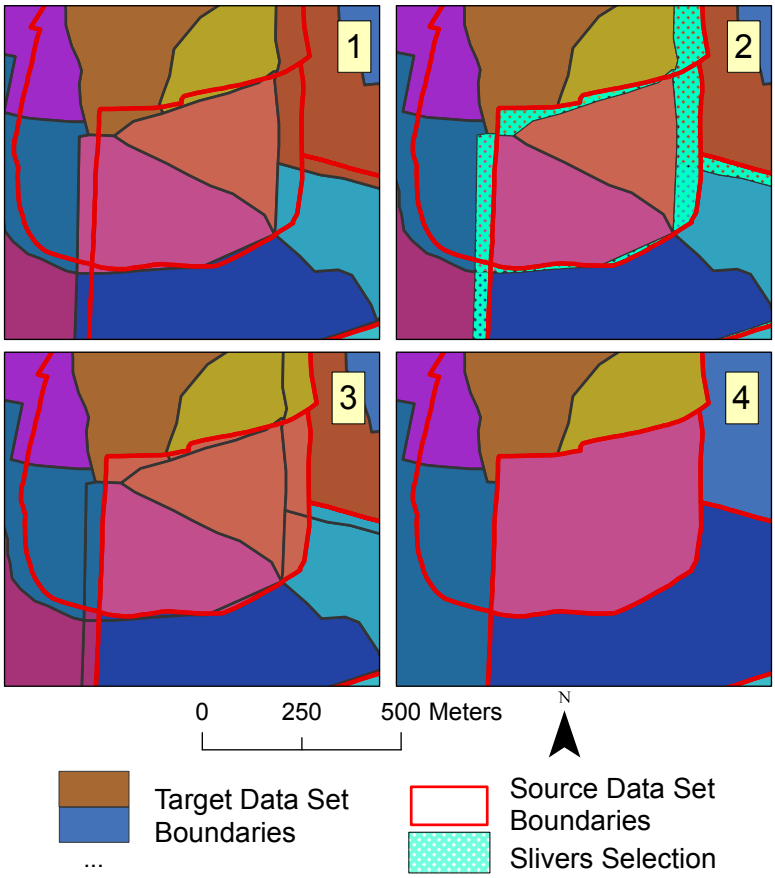


Figure 4: Steps of the implemented algorithm improving matching between source and target data by eliminating sliver polygons and aggregating polygons from the target data according to the source dataset boundaries.

In order to measure the accuracy of the areal interpolation methods we used population counts known at a disaggregated level. We used four error measure metrics. Three of them are central error values: the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), and the Median Absolute Error (MedAE). The fourth one we used is the Relative Absolute Value (RAV).

Our algorithm (see figure 3 & figure 4) works similarly to sliver polygons eliminating toolsby merging a selected intersection to the neighbour polygon sharing the longest arc. But this algorithm aggregates selected polygons only within defined boundaries, in our case the source data limits. In order to select the intersections we computed source and target area ratios and the thinness index of each intersecting feature.

III RESULTS & DISCUSSION

Our results comparing Avignon (figure 5, table 1), to Hobart (figure 6, table 2) show that our method can indeed reduce areal interpolation errors, regardless of the areal interpolation method, and that it is linked to how well source and target data originally match. Considering the RAE, our results even show that the total of areal interpolation error can be reduced by 10 percentage points, from 21.5% to 12.6% for Avignon in the case of Areal Weighting, which is approximately 40% half fewer errors. The comparison between Avignon and Hobart show that our algorithm's performance decreased slightly when source and target data already match well and the target dataset has a broader scale than the source dataset, (table 1, table 2).

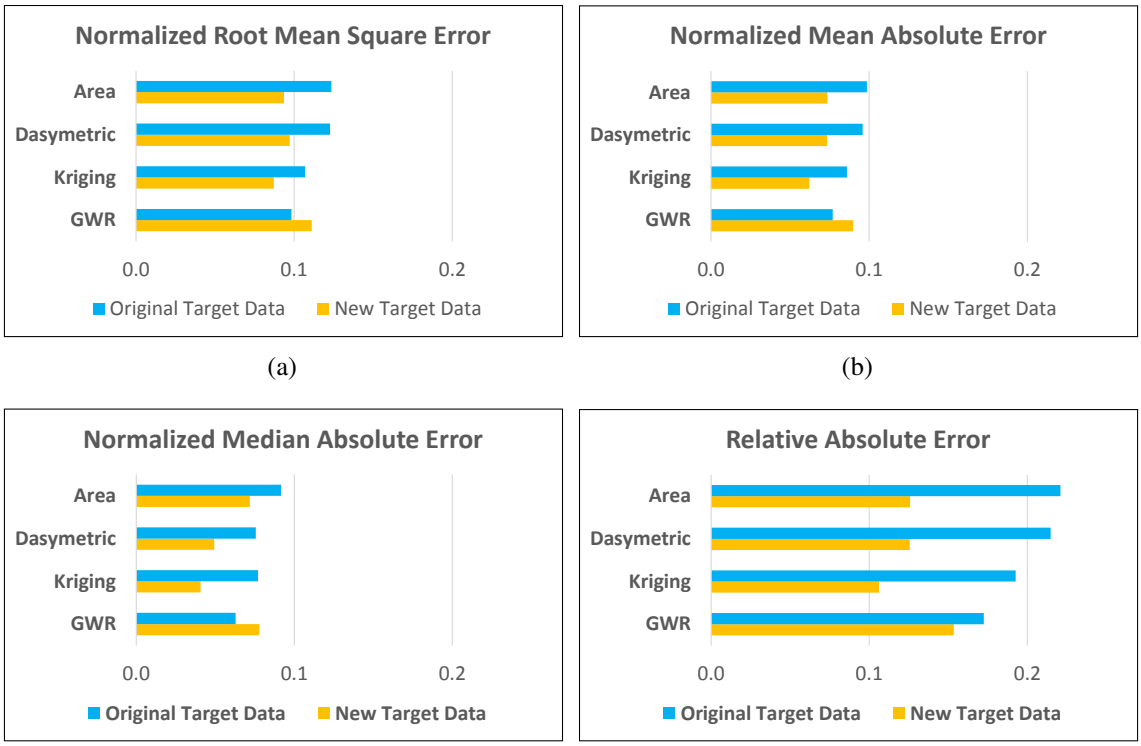


Figure 5: Avignon, Areal interpolation errors of population count, original polling areas and polling areas modified by our algorithm

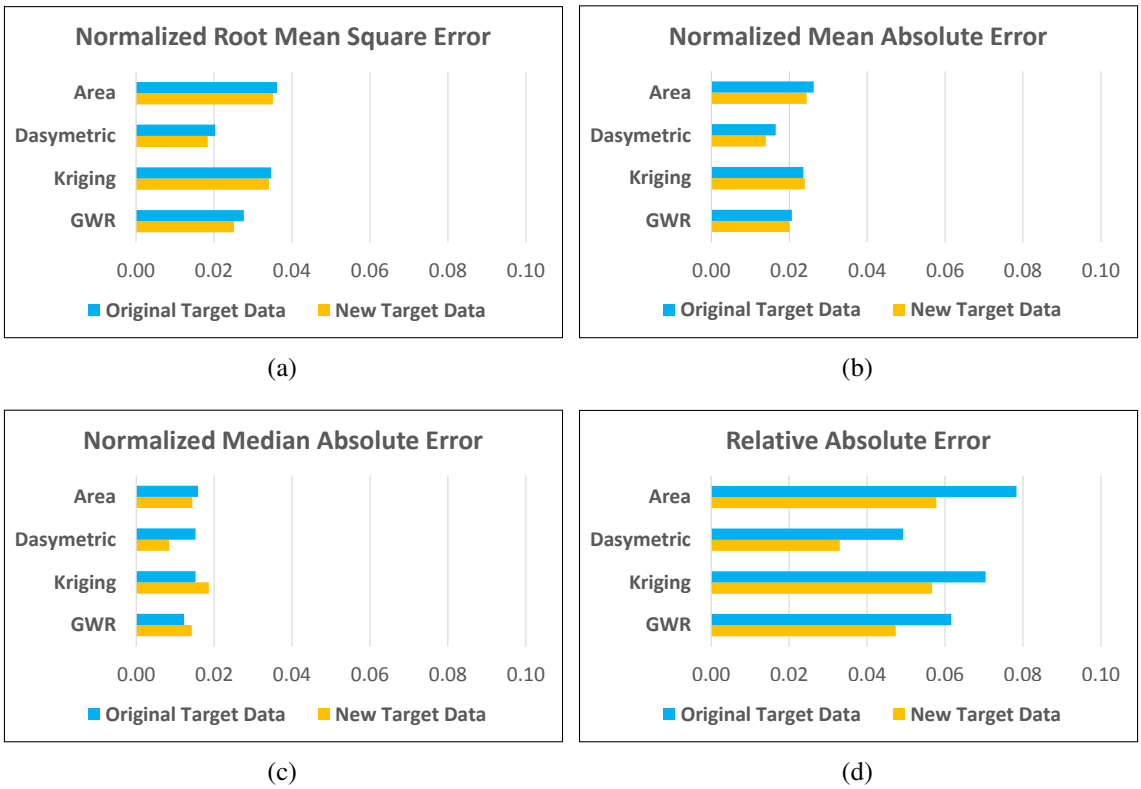


Figure 6: Hobart, Areal interpolation errors of population count, original polling areas and polling areas modified by our algorithm

Polling Area Geometric Precision			
	Original Target Data	New Target Data	Precision Variation (%)
Smallest Unit (ha)	7.9	9.3	-15
Number of Units	57	43	-25

Intersects with Statistical Area Geometric Precision			
	Original Target Data	New Target Data	Precision Variation (%)
Smallest Unit (ha)	0.001	0.9	-99.9
Number of Units	281	94	-66

Interpolation Errors Variation								
	RMSE		MAE		MedAE		RAE	
	%	people	%	people	%	people	%	people
Area Weighting	-23	-101	-24	-85	-21	-66	-43	-8489
Binary Dasymetric	-20	-86	-23	-75	-34	-90	-42	-7954
GWR	+14	+50	+18	+48	+25	+56	-11	-1697
Kriging	-17	-66	-27	-81	-47	-126	-45	-7714

Table 1: Avignon, geometric precision and interpolation errors, original target data and the new target

Polling Area Geometric Precision			
	Original Target Data	New Target Data	Precision Variation (%)
Smallest Unit (ha)	2.3	36.3	-94
Number of Units	35	34	-3

Intersects with Statistical Area Geometric Precision			
	Original Target Data	New Target Data	Precision Variation (%)
Smallest Unit (ha)	0.001	5	-99.998
Number of Units	415	238	-43

Interpolation Errors Variation								
	RMSE		MAE		MedAE		RAE	
	%	people	%	people	%	people	%	people
Area Weighting	-4	-9	-8	-13	-10	-10	-26	-1531
Binary Dasymetric	-10	-13	-16	-17	-45	-43	-33	-1204
GWR	-10	-17	-4	-5	+15	+12	-23	-1058
Kriging	-2	-5	+1	+1	+21	+21	-20	-1022

Table 2: Hobart, geometric precision and interpolation errors, original target data and new target data

References

Bernard L., Marchand-Lagier C., Josselin D., Louvet R. (2015, June). Some ways to estimate the effects of the bad voters registration on electoral participation. In *Congrès AFSP 2015*, Aix-en-Provence, France.

Carson B. D. (2013). *Testing Kriging-Based Areal Interpolation for Census-Based Socioeconomic Data*. Master's thesis, University of Redlands.

Do V. H. (2015). *Les méthodes d'interpolation pour données sur zones*. Ph. D. thesis, Toulouse 1.

Eicher C. L., Brewer C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science 28*(2), 125–138.

Fotheringham S., Rogerson P. (2013, April). *Spatial Analysis And GIS*. CRC Press.

Goodchild M. F., Anselin L., Deichmann U. (1993). A Framework for the Areal Interpolation of Socioeconomic Data. *Environment and Planning A 25*, 383–397.

Gotway C. A., Young L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association 97*(458), 632–648.

Krivoruchko K., Gribov A., Krause E. (2011). Multivariate Areal Interpolation for Continuous and Count Data. *Procedia Environmental Sciences 3*, 14–19.

Lam N. S.-N. (1983, January). Spatial Interpolation Methods: A Review. *The American Cartographer 10*(2), 129–150.

Lin J., Cromley R., Zhang C. (2011, March). Using geographically weighted regression to solve the areal interpolation problem. *Annals of GIS 17*(1), 1–14.

Martin D. (2003, March). Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science 17*(2), 181–196.

Mennis J., Hultgren T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science 33*(3), 179–194.

Mugglin A. S., Carlin B. P., Zhu L., Conlon E. (1999). Bayesian areal interpolation, estimation, and smoothing: an inferential approach for geographic information systems. *Environment and Planning A 31*(8), 1337–1352.

Openshaw S. (1977). A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling. *Transactions of the Institute of British Geographers 2*(4), 459–472.

Qiu F., Zhang C., Zhou Y. (2012, September). The Development of an Areal Interpolation ArcGIS Extension and a Comparative Study. *GIScience & Remote Sensing 49*(5), 644–663.

Reibel M., Agrawal A. A. (2007). Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review*, 10–1007.

Tobler W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association 74*(367), 519–530.

Zhang C., Qiu F. (2011, March). A Point-Based Intelligent Approach to Areal Interpolation. *The Professional Geographer 63*(2), 262–276.

Modelling uncertainty using geostatistics, a case study in Ecuador

Chicaiza Elena[*1] **and Buenaño Xavier**[1]

[1]Technical University of Madrid (UPM), Spain

*Corresponding author: gachisbar@gmail.com

Abstract
Spatial data includes positional and thematic components. In this study, the uncertainty analysis has been focused in thematic aspect. A total of 200 soil samples with metal determinations were analyzed. Graphical and numerical tools were employed to summarize and detect critical metals. Copper was identified as the cation of major concern in the study area. Kriging and Co-located Co-Kriging (with Zinc as secondary variable) were employed to carry out predictions. The leave-one-out cross validation method was employed to assess the variogram model used in estimation process, specifically mean square deviation ratio (MSDR) statistic was considered. Finally, unconditional simulations (500 realizations) were computed in order to get mean value and uncertainty of the random field of variable analyzed. The mean of variances obtained from simulations is 1.6 times higher than the mean of kriging variances.

Keywords
Uncertainty, simulation, validation.

I INTRODUCTION

According Griffith et al. (2015), spatial data comprise two components: attribute and location. In this work, we focused the analysis in attribute component.

Currently, the statement that descriptions of spatial phenomena are subject to uncertainty is now generally accepted as explained by Chiles and Delfiner (2009). In this context, Chiles and Delfiner (2009) argues Geostatistics can be defined as "the application of probabilistic methods to regionalized variables".

This study uses R Development Core Team (2013) open-source statistical software. In order to carry out the geostatistical analysis, two specific packages or libraries developed by Ribeiro Jr. and Diggle (2001) (geoR) and Pebesma (2004) (gstat) were employed.

The study area is located in the southeastern of Ecuador, in Zamora province. The area is inside a mining concession where informal mining activities are carried out. The accumulation of heavy metals in soils can bring human health problems, specially when these are exposed directly, for example in this kind of mining activities.

In this study, metals in 200 soil samples were analyzed. A multivariate data exploration analysis was carried out with some metals (Cu, Zn, Cd, and Pb). Kabacoff (2011) recommends the use of relatively new graphical statistical analysis tool named corrgram (see Figure 1) that summarize the relationships between variables very well.

The Figure 1 shows that these variables have a positive correlation (hatching in 45 degrees) with exception of Cd-Pb relation (hatching in -45 degrees); the color saturation in the figure repre-

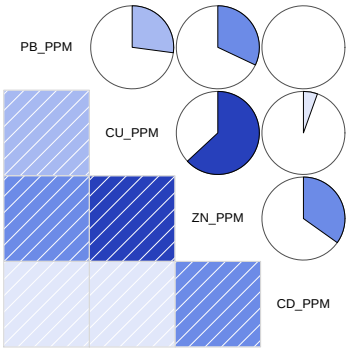


Figure 1: A corrgram of heavy metals soil samples.

sents magnitude of correlation between variables. Notice the significant correlation between Cu and Zn.

II CALCULATIONS

Copper has been identified as an element of major concern. For this reason, the Cu variable was chosen for the spatial estimation by two methods: a) Ordinary Kriging and b) Co-located Co-Kriging. It's interesting to estimate this variable with a covariate, in this case the Zn metal, in order to define the uncertainty threshold between the two methods employed. The Figure 2 shows variograms of Cu and Zn and the cross-variogram.

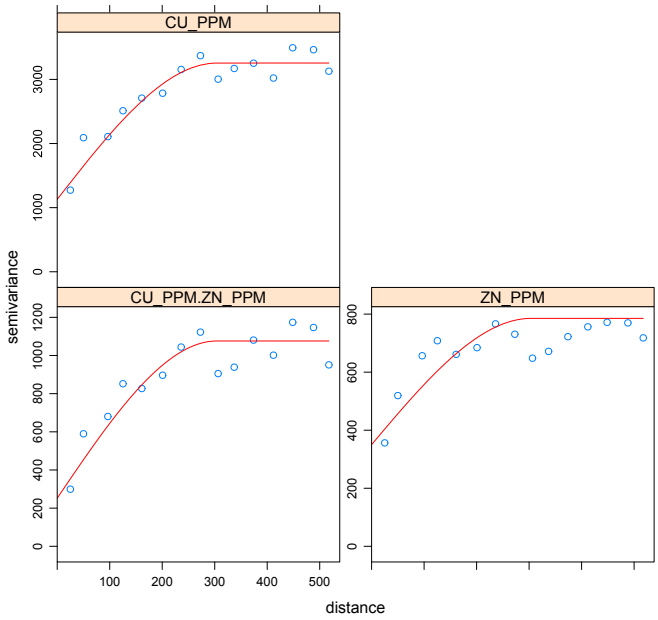


Figure 2: Direct variograms of Cu, Zn, cross-variogram and fitted functions.

A initial exploration data analysis (EDA) of Cu variable was carried out in order to identify outliers and a possible presence of trend that could limit the application of geostatistical theory.

Through directional clouds plotting, spatial trend was discarded. Directional variograms (0, 45, 90, 135 degrees) confirm this assumption.

An important aspect to notice is that exists a nugget effect in the variogram model of both elements analyzed (Cu and Zn). In this case, the nugget effect is originated in the presence of micro-structure.

The same approach was carried out to analyze Zn variable, and the results were similar.

Some variogram fitting techniques can be employed, but according Oliver and Webster (2014) residual maximum likelihood (REML) method is the current best practice. This approach was carried out in the present study.

According Goovaerts (2001), the choice of an approach for uncertainty modeling should be guided by the answer to these questions: What type of uncertainty model is being sought?, What is the support for the uncertainty assessment?, Why do we model uncertainty?.

In this context, it is important to carry out a validation procedure. There are some ways to validate the model decisions, one of them is the cross validation technique.

Bivand et al. (2008) denotes that cross validation splits the data set into two sets: a modeling set and a validation set. When the number of splits is equal to the number of observations, the procedure is called leave-one-out cross validation.

For the cross-validation, leave-one-out technique was used. There are some statistics to analyze the cross-validation. The most interesting is the mean squared deviation ratio (MSDR), which is the mean of the squared errors divided by the corresponding kriging variances according Oliver and Webster (2014). The formula is shown in Equation 1.

$$MSDR = \frac{1}{N}\sum_{i=1}^{N}\frac{\{z(x_i) - \hat{Z}(x_i)\}^2}{\hat{\sigma}_K^2(x_i)} \tag{1}$$

In these equation $z(x_i)$ is the ith datum at x_i, $\hat{Z}(x_i)$ is the kriged prediction there, and $\hat{\sigma}_K^2(x_i)$ is the kriging variance.

Also mean error (ME) and root mean squared error (RMSE) were computed. The results are shown in Table 1.

	ME	**RMSE**	**MSDR**
Ordinary Kriging	0.104	45.318	1.017
Co-Kriging	-0.085	40.688	1.099

Table 1: Uncertainty results with two geostatistical methods.

Another important tool to define uncertainty is simulation. Bivand et al. (2008) defines geostatistical simulation as the simulation of possible realisations of a random field, given the specifications for that random field (e.g. mean structure, residual variogram, intrinsic stationarity) and possibly observation data.

In this context, the drawing of envelopes is an interesting technique. Envelopes were computed assuming a (transformed) Gaussian random field model. According Ribeiro Jr. and Diggle (2001), simulated values are generated at the data locations, given the model parameters obtained by REML. Empirical variogram is computed for each simulation using the same lags as for the original variogram of the data. The envelopes are computed by taking, at each lag, the

maximum and minimum values of the variograms for simulated data. In this case, 99 simulations were carried out. Figure 3 shows the envelopes based in this criteria. The figure shows the range in that we would expect to model the variogram of Copper variable.

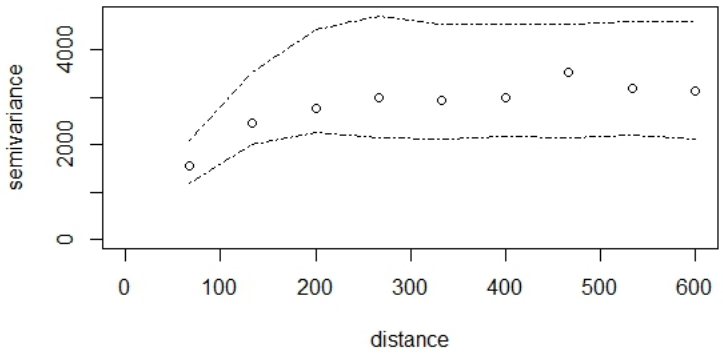


Figure 3: Envelopes of simulations.

Also, unconditional simulations that ignore observations and only reproduce means and prescribed variability, have been carried out. A mean value of 71, corresponding to the mean of prediction map was considering to carried out the simulations.

The results of four (4) first realizations are shown in Figure 4.

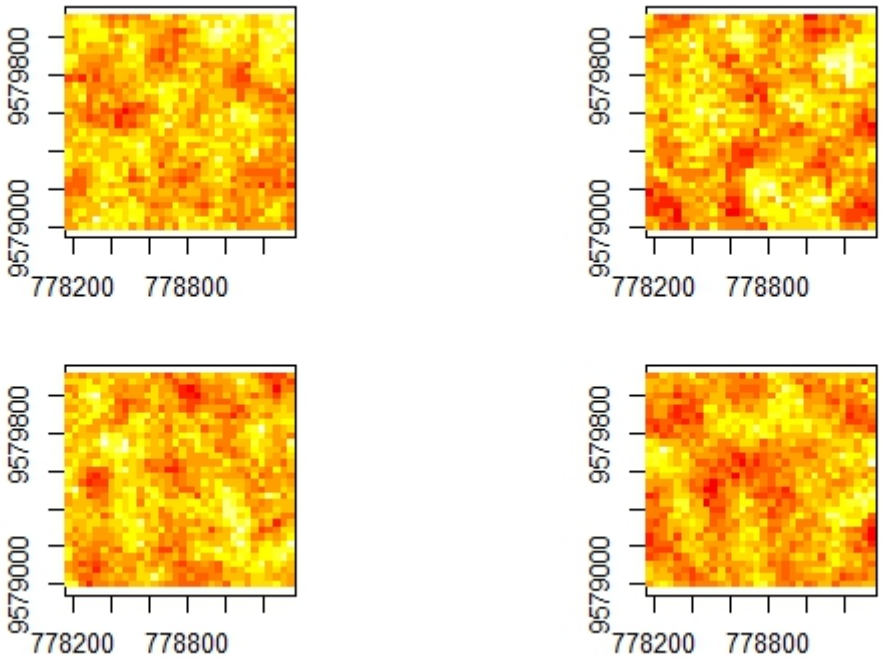


Figure 4: Unconditional simulation of 4 realizations with REML variogram parameters

A total of five hundred (500) realizations were carried out. The mean of variances obtained in these was computed; this mean is 1.6 times higher than the mean of kriging variances. This is a reasonable value, considering that observations were not included in simulations. Nevertheless

the high values of kriging variances and simulations's mean variances could be explained for nugget effect inclusion in the models. The mean of random field generated by simulations was 70.38 and the mean of kriging predictions was 70.96. The proximity of these values reflects that the random field was modeled correctly.

III RESULTS

ME of two methods is almost equal. This situation demonstrates that kriging (in general) is an unbiased estimator. RMSE of ordinary kriging is higher than co-kriging due to the auxiliary variable provides more information to the model prediction.

On the other hand, MSDR of Ordinary Kriging is slightly lower than MSDR of co-kriging, probably due to the characteristics of this statist. Lloyd (2010) in his study concerned with mapping precipitation amount, found that OK outperformed CK (with elevation as the secondary variable) and this was due, at least in part, to the weak global relationship between the two variables.

The map of kriging prediction is shown in Figure 5. The gradient color represents the Copper variable estimation and the contour lines represent variance prediction. The highest variances are located in the corner of southwestern sector.This zone exhibits a lower sampling density in relation with the rest of study area.

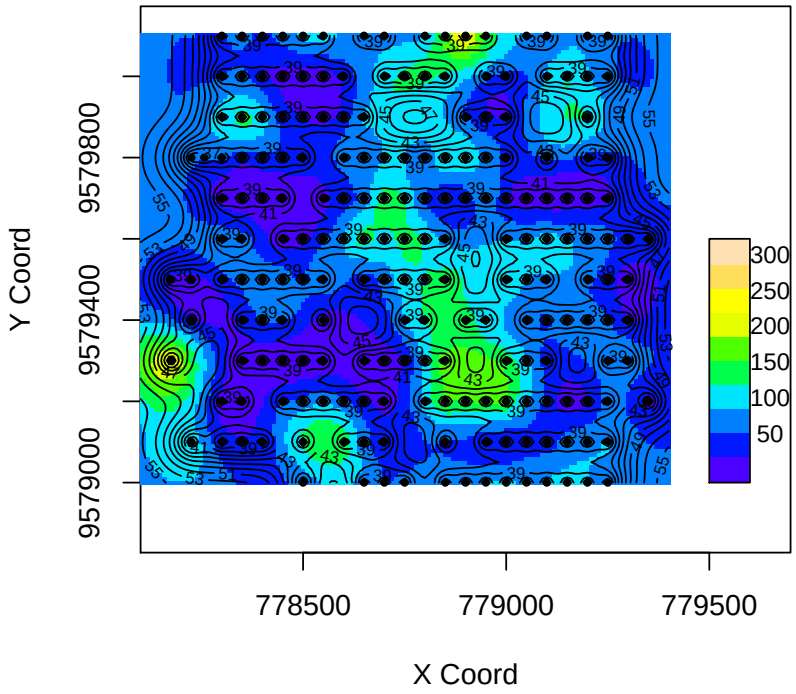


Figure 5: Map of kriging prediction and variance.

References

Bivand R. S., Pebesma E. J., Gomez-Rubio V. (2008). *Applied spatial data analysis with R. Springer*. Springer.

Chiles J.-P., Delfiner P. (2009). *Geostatistics: modeling spatial uncertainty*, Volume 497. John Wiley & Sons.

Goovaerts P. (2001). Geostatistical modelling of uncertainty in soil science. *Geoderma 103*(1), 3–26.

Griffith D., Wong D., Chun Y. (2015). Uncertainty-related research issues in spatial analysis. *Uncertainty Modelling and Quality Control for Spatial Data*, 3.

Kabacoff R. (2011). *R in Action*. Manning Publications Co.

Lloyd C. D. (2010). *geoENV VII – Geostatistics for Environmental Applications*, Chapter Multivariate Interpolation of Monthly Precipitation Amount in the United Kingdom, pp. 27–39. Dordrecht: Springer Netherlands.

Oliver M., Webster R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena 113*, 56–69.

Pebesma E. J. (2004). Multivariable geostatistics in s: the gstat package. *Computers & Geosciences 30*, 683–691.

R Development Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Ribeiro Jr. P., Diggle P. (2001). geoR: a package for geostatistical analysis. *R-NEWS 1*(2), 15–18.

A Hybrid approach for land cover mapping based on the combination of soft classifiers outputs and uncertainty information

Luisa M. S. Gonçalves[134*], Cidália C. Fonte[2,4]

[1] Polytechnic Institute of Leiria, School of Technology and Management, Portugal
[2] Department of Mathematics, University of Coimbra, Portugal
[3] NOVA IMS - NOVA Information Management School, Universidade Nova de Lisboa, Portugal
[4] Institute for Systems and Computers Engineering at Coimbra, Portugal

*Corresponding author: luisa.goncalves@ipleiria.pt

Abstract
In this article, the authors present a hybrid approach to produce more accurate land cover maps of diverse landscapes representing Mediterranean environments. The innovation of the proposed methodology is to use, in the combination of soft classifiers outputs, information about the classification uncertainty associated with each pixel and its neighbourhood.
The hybrid combined classification method developed includes the following steps: 1) definition of the training areas; 2) pixel-based classification using several soft classifiers; 3) computation of the classification uncertainty; 4) application of rules to combine the outputs of the pixel-based soft classifications and the uncertainty information obtained with uncertainty measures, on a pixel and neighbourhood based approach; 5) image segmentation; and 6) object classification based on decision rules that include the results of the combined soft pixel-based classification and its uncertainty. The proposed methodology was applied to an IKONOS and a SPOT-4 multispectral image with, respectively, 4m and 20m spatial resolution.
The overall accuracy of the hybrid classification obtained with the proposed methodology was higher than the one obtained for the individual pixel-based classifications, which shows that this approach may increase classification accuracy. The approach showed to be particularly powerful to obtain Land Cover Map for landscapes representing Mediterranean environments.

Keywords
Soft classifiers, uncertainty, land cover, IKONOS, SPOT

I INTRODUCTION

The use of a hybrid classification approach, which combines pixels and objects, has been shown to be suitable for the identification of Landscape Units that contain a variety of land cover objects using Very High Spatial Resolution images. With the combination of a set of classifiers outputs it is possible to obtain a classification that is often more accurate than the individual classifications (Gonçalves et al., 2010; Gonçalves, 2011). Although several approaches have been proposed for combining hard classifications, the development of methods to combine soft classifications and their integration in a hybrid pixel and object based approach is still a field of research. This study tests whether the combination of the outputs of a set of soft classifiers in a hybrid classification approach using uncertainty and contextual information can improve the accuracy of the results. In the proposed hybrid classification

method, the uncertainty information was used in two phases. First, for combining the outputs of three pixel-based soft classifications. This was done through the development of rules that incorporate the information provided by the pixel-based soft classifications and the results given by the application of an uncertainty measure. Second, in the classification of the obtained segmented objects, which represent the Land Units. These are classified through decision rules which include the results of the combined soft pixel-based classification and its uncertainty. The main objective of integrating uncertainty in the classification process is to avoid the use of misclassified pixels in the classification. The first step of the method was applied to different kinds of images, namely to two IKONOS multispectral images and a SPOT-4 multispectral image, with 4m and 20m of spatial resolution respectively, with different Mediterranean landscapes characteristics, to evaluate the outputs reliability. The second step was only applied to an IKONOS multispectral image to obtain a Land Unit Map (LUM).

II DATA

The study was conducted using three multispectral satellite images with High Spatial Resolution (HSR) from two regions of Portugal (Figure 1), two located in the Alentejo region and one in the Central region. An image of the Alentejo region near Alcácer was used, obtained by the IKONOS sensor, having 4 spectral bands namely (1) blue (0.45-0.52 µm), (2) green (0.52-0.60 µm), (3) red (0.63-0.69 µm) and (4) near infrared (0.76-0.90 µm), with a spatial resolution of 4m and a dimension of 11.8 km by 8.7 Km. The two other images were obtained by the SPOT 4 sensor, having 4 spectral bands, namely (1) green (0.50-0.59 µm), (2) red (0.61-0.68 µm), (3) near infrared (0.78-0.89 µm) and (4) short-wavelength infrared (1.58-1.75 µm), with a spatial resolution of 20m and a dimension of 16.4 km by 15.1 Km.

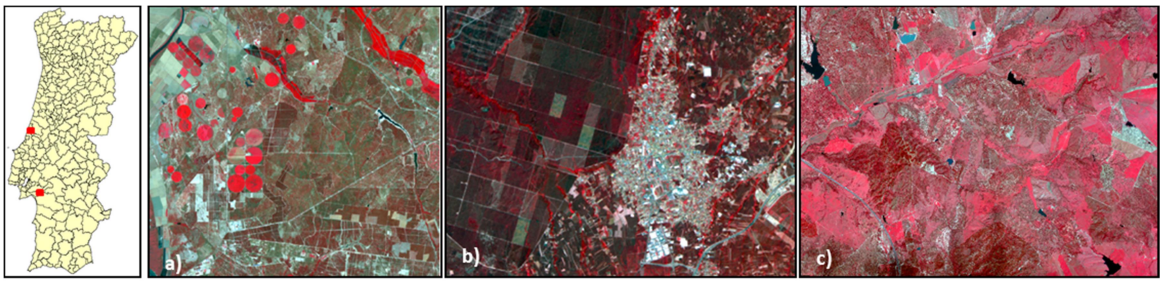


Figure 1: Multispectal images (false color) : a) SPOT-4 image of the Alentejo region, located in the municipal district of Setubal; b) SPOT-4 image of the Central region located in the municipal district of Marinha Grande; c) IKONOS image of the Alentejo region located in the municipal district of Alcácer.

The locations of the images chosen to carry on the study are representative of the main forest species and Mediterranean landscapes in Portugal. The SPOT-4 image is located in the municipal district of Marinha Grande, in the Central region, near the coast, and includes areas with different characteristics, such as: built-up areas; agricultural fields and forest. The dominant forestry species in the region are eucalyptus and coniferous trees.
The SPOT-4 image is located in the municipal district of Setubal, in the Alentejo region, is occupied mainly by agriculture, pastures, forest and agro-forestry areas. The dominant forest species are eucalyptus, coniferous and cork trees.
The IKONOS image is located in the municipal district of Alcácer, in the Alentejo region. It is also occupied mainly by agriculture, pastures, forest and agro-forestry areas, where the dominant forest species are eucalyptus, coniferous and cork trees.

III METHODOLOGY

The main goal of the classification is to obtain a LUM, where the objects that represent Landscape Units (LU) have a mean area of 0.5 ha and are classified using a hybrid

classification approach that integrates the combination of the outputs of pixel-based soft classifications and their uncertainty. The combination of the pixel-based soft classification is made through the application of rules that incorporate the information provided by the pixel-based classifications and the results given by the uncertainty measures on a pixel and neighbourhood based approach. This approach considers not only each pixel-based classification and its uncertainty information but also the classification and uncertainty of its neighbourhood. Three soft classifiers were used in this application: 1) the neural network Multi-Layer Perceptron Classifier (MLPC); 2) a pixel-based supervised fuzzy soft classifier based on the underlying logic of Minimum-Distance-to-Means (FC); and 3) Maximum Likelihood Classifier (MLC). The classifiers were trained using the same sampling protocol that included 100 pixels per-class. The classes used in this study to identify Surface Elements (SE), which are the basic elements of landscape, are: Eucalyptus Trees (ET), Cork Trees (CKT), Coniferous Trees (CFT), Shadows (S), Shallow Water (SW), Deep Water (DW), Herbaceous (H), Sparse Herbaceous (SH), Non Vegetation Area (NVA), Irrigated Herbaceous (IH) and Non Irrigated Herbaceous (NIH). These classification methods assign to each pixel and to the each class under consideration, in the case of MLPC, different degrees of assignment, in the case of FC, different degrees of possibility and in the case of MLC, different degrees of probabilities. This extra data provides additional information at the pixel level that allows the assessment of the classification uncertainty. To analyze if the use of an output combination of a set of soft classifiers in a hybrid approach classification can improve the accuracy of the results, a similar method, where the combination of the three classifiers and the classification uncertainty was not considered, is also presented. The hybrid classification method that was developed includes the following steps: 1) definition of the training areas; 2) pixel-based soft classification using several soft classifiers; 3) computation of the classification uncertainty; 4) development and application of rules to combine the soft classifications, incorporating the information provided by the pixel-based classifications the results given by the uncertainty measures, on a pixel and neighbourhood based approach; 5) image segmentation; and 6) object classification based on decision rules that include the results of the combined soft pixel-based classification and its uncertainty. The second method that does not take into consideration the output combination of the pixel-based classifications and their uncertainty includes three steps: 1) pixel-based classification of the image; 2) image segmentation and 3) object classification based on decision rules. This article focuses only on the first methodology since the second one was already presented in (Gonçalves et al., 2009a).

Combination of pixel-based soft classifications

The first phase of the algorithm developed to combine pixel-based soft classifications checks whether the same class is assigned to each pixel by all classifiers. If this condition is satisfied, the class is accepted. If the output classes for each individual pixel differed, the uncertainty information is compared and the class assigned with the lower value of uncertainty is chosen to be the one assigned to the pixel. In this approach the uncertainty measure E, developed by (Chow, 1970), was used to quantify the uncertainty at each spatial unit. This measure is given by

$$E = 1 - p(x_1) \quad (1)$$

where $p(x_1)$ is the largest degree of possibility or probability of the possibility distributions or probability distributions assigned to the pixel corresponding to the several classes. This measure is also called ambiguity measure.
If the classifiers have different results for a certain pixel but the ambiguity values are equal the results of the neighborhood classification is used to make a judgment.

To evaluate if the combined classification improves the results, the accuracy assessment was made with the same protocol used with the single classifiers and the results were compared.
One of the classifiers used was the MLP neural network, which is a non-parametric method and is the most commonly used in remote sensing. Details of the MLP can be found in Atkinson and Tatnall (1997) and in Brown et al. (2009). The MLP provides an activation level for every output class of each pixel, and for hard classifications each pixel is allocated to the class with the largest activation level. A soft classification may be derived from this classifier by considering the activation levels of the network output units for each pixel. These activation levels range from 0 to 1, and may be used as indicators of the uncertainty associated with the pixel allocation to the classes. The other classifier used was a pixel-based supervised fuzzy classifier based on the underlying logic of the Minimum-Distance-to-Means classifier. Details of the fuzzy classifier can be found in (Gonçalves *et al.*, 2009b). The third was a supervised Bayesian classifier similar to the maximum likelihood classifier. The maximum likelihood classifier is based on the estimation of the multivariate Gaussian probability density function of each class using the classes statistics (mean, variance and covariance) estimated from the sample pixels of the training set. Details can be found in (Gonçalves et al., 2009b).
To evaluate the classification accuracy of the individual soft classifications and the combined results, a stratified random sampling with about 100 pixels per class was selected considering the entire image scene, which also included mixed pixels. The number of pixels was chosen to obtain a standard error of 0.05 for the estimation of the accuracy indexes of each class. Each land cover class was sampled independently and the accuracy assessment was made with an error matrix.

Hybrid Classification to obtain a Landscape Unit Map

The LUM was built using the combined output of the pixel-based classification, its ambiguity information and the objects obtained with the segmentation algorithm. In the segmentation stage the whole image was partitioned into a series of closed objects, corresponding to the spatial patterns. The extraction of the objects was driven using the "Fractal Net Evolution Approach" (FNEA) segmentation method, implemented in eCognition software, which can be described as a region merging technique (Baatz and Schape, 2000).
In this study only one segmentation level was considered, chosen from a series of experiments done with different parameters, the results of which were visually analyzed. The criterium that led to their choice was the identification of meaningful image-objects i.e., groups of pixels that represented the LU existing in the study area, with a mean area of 0.5 ha. The next step was the development of rules that incorporate the information provided by the combined pixel-based classification within each object and the results given by the ambiguity measure *E*. The classification of the LU is similar to a decision tree which, for geographical objects, is a hierarchical structure consisting of several levels. At each level a test is applied to one or more attribute values. The application of a rule results either in a leaf, allocating an object to a class, or a new decision node, specifying a further decision rule. In this study eight LU classes were used: Water Bodies (WB), Agriculture and Pasture Areas (A), Non-Vegetated Areas (NVA), Broad-Leaved Forest (BF), Coniferous Forest (CFF), Cork Forest (CF), Agro-Forestry Areas (AFA) and Mixed Forest (MF). Table 1 shows the classification rules, based on the SE identified in the previous phase. The structure of the rules is based on the ones used on the study performed by (Gonçalves *et al.*, 2009a).
For the accuracy assessment, the sampling unit to assess the accuracy of the LUM was a fixed-area square plot sampling unit with an area of 0.5 ha. A stratified random sampling of about 50 samples per class was chosen, which guarantees a standard error of 0.07 for the Conditional

Probability of the Map (CPM) and Conditional Probability of Reference (CPR) estimates for each class, assuming that the classification accuracy is superior to 50%, which is acceptable because the construction of the LUM already involved a prior pixel-based classification and an analysis of the terrain. The accuracy assessment was made with an error matrix, where the p_{ij} entry is the proportion of area that is class *i* in the map and class *j* in the reference within the square areas with 0.5 ha. The CPM and CPR accuracy parameters were then derived from the error matrix.

Rules	Test	Class if true
Rule 1	Objects which have more than 10% of SE classified as tree crowns, regardless of species, with ambiguity of less than 0.5	Forest
	Objects which do not satisfy the previous test	Non-Forest
Rule 2	The mode of the SE, inside the object, with ambiguity of less than 0.5 is Deep Water or Shallow Water	Water Bodies
	The mode of the SE, inside the objects, with ambiguity of less than 0.5 is Herbaceous Vegetation or Sparse Herbaceous Vegetation	Agriculture
	The mode of the SE, inside the objects, with ambiguity of less than 0.5 is Non-Vegetated Area or Shadow	Non-Vegetated Areas
Rule 3	Eucalyptus Trees represent more than 75% of the trees or objects that have only Broad- Leaved Trees inside	Broad-Leaved Forest
	Coniferous Trees represent more than 75% of the trees or objects that have only Coniferous Trees inside	Coniferous Forest
	Cork Trees represent more than 75% of the trees or objects that have only Cork Trees inside and the percentage of Herbaceous or Sparse Herbaceous is inferior to Cork Trees	Cork Forest
	Objects that do not satisfy the previous test	Mixed or Non-
Rule 4	The percentage of trees is less than 50%; the percentage of Herbaceous or Sparse Herbaceous is superior to Cork Trees and 80% of trees are Cork Trees with ambiguity of less than 0.5	Agro-Forestry Areas
	Objects that do not satisfy the previous test	Mixed Forest

Table 1. Object's classification rules.

IV RESULTS

Combined Pixel Classifications

The accuracy assessment of the combined classification was made with the same testing datasets used to evaluate the individual classifications. Table 2 shows the Global Accuracy (GA) results.

Regions / Classifiers	Marinha Grande (SPOT-4)	Setubal (SPOT-4)	Alcacer (IKONOS)
FC	73.8%	70.4%	65.5%
MLPC	75.6%	72.4%	64.5%
MLC	86.9%	73.7%	66.9%
Combined	89.2%	82.7%	72%

Table 2. Global Accuracy of the classifications.

The GA of the combined classification of the Marinha Grande, Setubal and Alcácer regions was 89% , 82% and 72% respectively which represents a higher value than that of the most accurate individual classification. Figure 2, Figure 3 and Figure 4 show the classification results when each pixel is assigned to the class with a higher degree of probability for the MLC classifier, higher degree of possibility for the FC classifier, and with the largest activation level for the MLPC classifier for the three areas classified. The spatial distribution of the ambiguity measure E is also presented in the same figures, below the classification results. The regions with the larger ambiguity (darker zones) are the ones where the assignment degrees were lower.

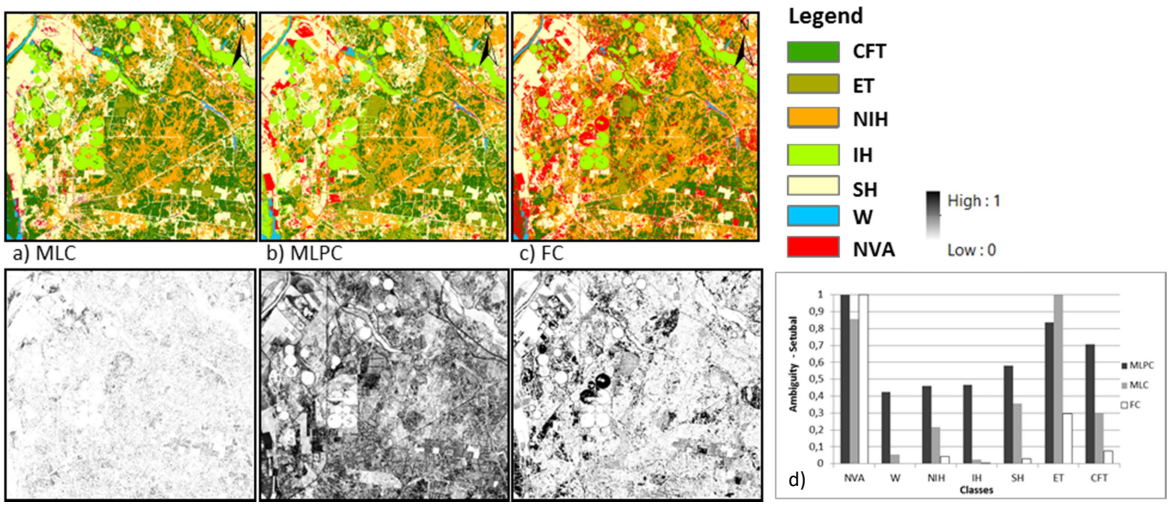


Figure 2: Hard version of the classification results of the Setubal area (above) and spatial distribution of ambiguity (below) with: a) MLC classifier b) MLPC classifier and c) FC classifier. Graph d) shows the mean ambiguity per class.

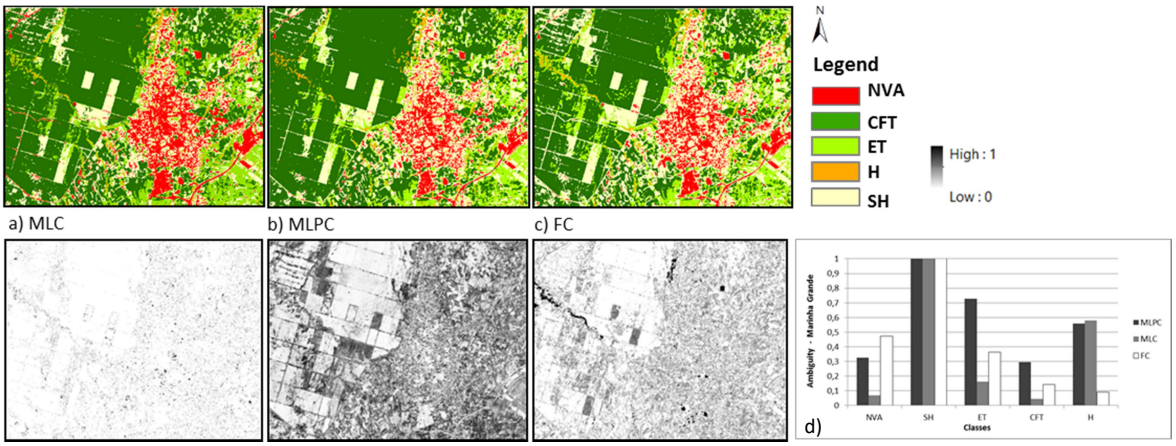


Figure 3: Hard version of the classification results of the Marinha Grande area (above) and spatial distribution of ambiguity (below) with: a) MLC classifier b) MLPC classifier and c) FC classifier. Graph d) shows the mean ambiguity per class.

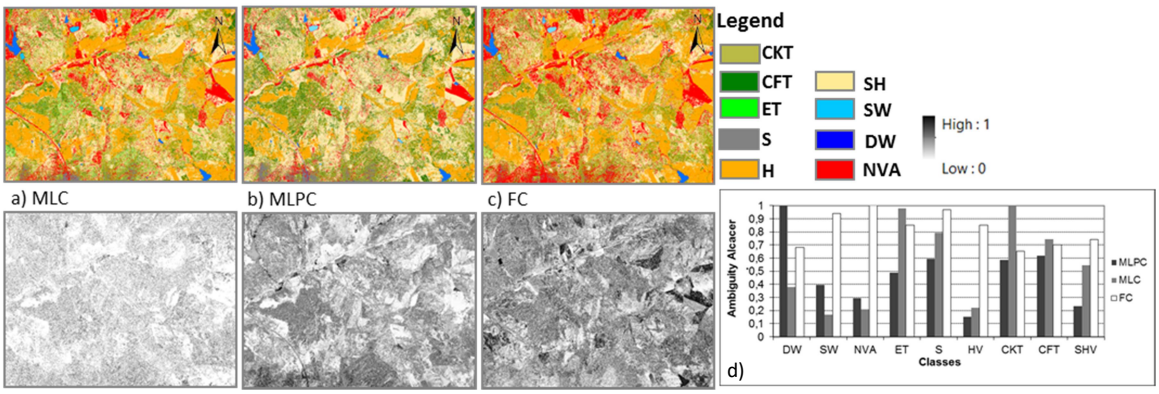


Figure 4: Hard version of the classification results of the Alcacer area (above) and spatial distribution of ambiguity (below) with: a) MLC classifier b) MLPC classifier and c) FC classifier. Graph d) shows the mean ambiguity per class.

The comparison of the mean ambiguity per class shows that for the Setubal area, almost all the classes were assigned with lower ambiguity with the FC classifier. The only exception was the Non-Vegetation Area which was assigned to the pixels with lower ambiguity by the MLC classifier. For the Marinha Grande area, SH was the class assigned to the pixels with more uncertainty by all classifiers, almost all the classes were assigned with lower ambiguity by the MLC classifier. The only exception was Herbaceous (H) that was assigned with lower mean ambiguity by the FC classifier.
For the Alcácer area, the comparison of the mean ambiguity per class shows that forest species, such as CKT and CFT, were assigned to the pixels with lower ambiguity with MLPC classifier. The class DW, SW and NVA classes were assigned to the pixels with lower ambiguity with MLC. The FC classifier presents the higher values of ambiguity. The mean ambiguity presented by the classifiers depends essentially on the characteristics of the image.

Land Unit Map

The Global Probability (GP) classification accuracy for the LUM, obtained with the hybrid approach which classifies the LU (segmented objects) using the combined classification and the ambiguity information, was 72%. The best result for the GP classification with the hybrid approach, using an individual classification (the ones that had the best GA results) and without the ambiguity information, was 59%. This shows that the accuracy increased significantly with the combined classification and the inclusion of the ambiguity information.
Figure 5 allows the comparison between the results of the CPR and CPM accuracies for the LUM obtained with both hybrid classification approaches. These show that the classification results obtained with the method using uncertainty are considerably better for almost all LU classes and this improvement is more evident for the forest classes. Figure 5 shows the final results of the classification with the proposed hybrid classification method.

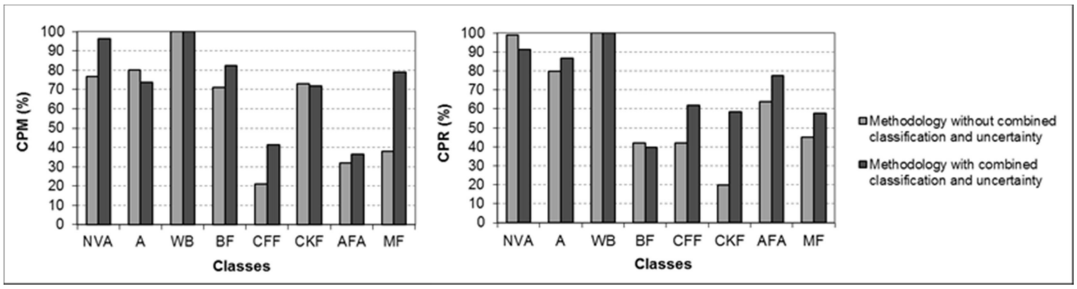


Figure 5: Conditional probability of reference (CPR) and Conditional probability of the map (CPM) obtained with the hybrid approach (LUM) with and without combined classification and uncertainty.

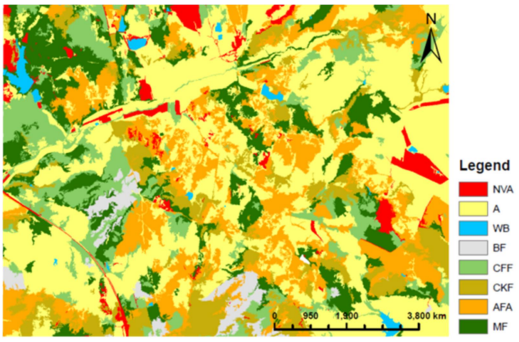


Figure 6: Land Unit Map.

V CONCLUSIONS

The goal of this study was to evaluate if the proposed hybrid combined classification method, using uncertainty information associated with the pixel-based soft classification, produces more accurate land cover maps of diverse landscapes representing Mediterranean environments. Another goal was to investigate if the proposed method that combines the outputs of different soft classifications, using the uncertainty information to choose the best class to assign to each pixel, when applied to different multispectral satellite images always produces more accurate Land Units. The method was applied to IKONOS and SOPT-4 multispectral satellite images with 4m and 20m spatial resolution and the results have shown that with the combined method the accuracy results in both cases were higher. With the proposed hybrid pixel-object classification, the global accuracy of the classification of Mediterranean Landscape Unit classes increased by 11% when compared to a similar classification method that does not use combined classification and ambiguity information. These results have confirmed that the information provided by the uncertainty measure was useful for the combined classification process and in the construction of the Landscape Unit Map because it allowed the determination of the best class to assign to the pixels.

Acknowledgements

This work has been supported by the Fundação para a Ciência e a Tecnologia (FCT) under project grant UID/MULTI/00308/2013.

References

Atkinson P. M., Tatnall A. R. L. (1997). Neural networks in remote sensing. *International Journal of Remote Sensing* 18, 699-709.

Baatz M., Schape A. (2000). Multiresolution segmentation – an optimization approach for high quality multi-scale image segmentation. In *Proceedings of Angewandte Geographische Informationsverarbeitung XII*. Beitrage Zum AGIT – Symposium Salzburg 2000, Karlsruhe: Herbert Wichmann Verlag, pp. 12-23

Brown K. M, Foody G. M., Atkinson P. M. (2009). Estimating per-pixel thematic uncertainty in remote sensing classifications. *International Journal of Remote Sensing* 30, 209-229.

Chow C. K. (1970). On optimum error and reject tradeoff. IEEE Transactions on Information Theory, 16, 41-46

Gonçalves L. M., Fonte C. C., Júlio E. N. B. S., Caetano M. (2009a). A method to incorporate uncertainty in the classification of remote sensing images. *International Journal of Remote Sensing* 30, 5489-5503.

Gonçalves L M, C C Fonte, E N B S Júlio & M Caetano (2009b). Spatial Data Quality From Process to Decisions, In *Proceedings of 6th International Symposium on Spatial Data Quality 2009*, Boca Raton: CRC Press, pp. 163-177

Gonçalves L. M., Fonte C. C., Caetano M. (2010). Using uncertainty information to combine soft classifications. In *Proceedings of the 13th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, IPMU 2010, Berlin Heidelberg: Springer-Verlag, LNAI 6178, pp. 455-463.

Gonçalves L. M. S. (2011). A hybrid approach classification of remote sensing images. In *Proceedings of the 31th EARSeL Symposium of the European Association of Remote Sensing Laboratories*, H. Halounová editors Czech Technical University, pp. 328-335.

Sampling to validate a global cropland map.

Javier Gallego*[1], Anne Schucknecht[1], François Waldner[2]

[1] Joint Research Centre. Ispra, Italy
[2] Université Catholique de Louvain, Belgium

*Corresponding author: javier.gallego@jrc.ec.europa.eu

Abtract
A global cropland map is being produced with a 300 m resolution in the SIGMA Project of the EU (Stimulating Innovation for Global Monitoring of Agriculture). The map, based on PROBA-V 300 m images of 2015, is not yet available but the collection of reference data is ongoing on a two-tier sample of plots photo-interpreted on publicly available Very High Resolution (VHR) images with the help of two versions of a Geo-wiki tool: a large sample (initially 40000 plots) photo-interpreted by volunteers (crowdsourcing) and a sub-sample of 4000 plots photo-interpreted by experts. A stratification based on a cropland probability map has been used to select a higher rate in the most difficult areas where the cropland probability is between 25% and 75%. Stratified systematic sampling with a variable number of replicates has been used. The sampling scheme has been assessed using as pseudo-truth the cropland probability map in the European Union. The conclusions suggest that the variance efficiency of the sampling scheme due to the improved spatial homogeneity is moderate. A more important gain appears in terms of traceability of the sample.

Keywords
Global cropland map; stratified systematic sample; crowdsourcing for validation.

I INTRODUCTION

The SIGMA project (Stimulating Innovation for Global Monitoring of Agriculture, http://www.geoglam-sigma.info/Pages/default.aspx) is elaborating a global cropland map based on PROBA-V 300 m images (Dierkx et al., 2014). The project is funded by the research framework program FP7 of the European Union. This paper presents the sampling scheme that has been elaborated for the validation of this map. The aim is ensuring at the same time 1) flexibility, 2) homogeneous spatial layout and 3) improved traceability. The solution we chose was a systematic sample with ranked replicates based on 1˚ latitude-longitude cells.

The concept of cropland for this map corresponds rather to “annual crops in the current year” rather than the standard concept of cropland in agricultural statistics (FAO, 1996). The chosen definition includes a minimum size threshold of 0.25 ha and a minimum width of 30 m. The SIGMA global cropland map is a binary map (cropland/non cropland) with the reference year 2015. The map will be delivered in latitude-longitude coordinates. Therefore the area of each pixel changes with the latitude. For the sample selection the same latitude-longitude coordinates have been used rather than a cartographic projection.

A two-tier sample is applied for the validation: a large sample of around 40,000 units is being photo-interpreted on very high resolution (VHR) images by volunteers that label as crop/non-crop a grid of 9 points inside each pixel. Experts photo-interpret the whole pixel after automatic segmentation for a smaller subsample of around 4000 units. Validators will interpret sub-pixels or polygons as cropland or non cropland based on a very high resolution imagery such as Google Earth and medium resolution time-series. Additionally, there will be the possibility to opt for "unknown/undetermined", if the validator does not know whether the sample is cropland or not.

II SAMPLING SCHEME

The collection of reference data for validation is being carried out before the cropland map to be validated is available. For this reason the map itself is not used for stratification as it is usually recommended (Olofsson et al., 2014). However a stratification has been built on the basis of the IIASA cropland probability map that has been produced by comparing different thematic maps (Fritz et al., 2015). Areas with very high or very low probability are considered easier to classify and need a lower sampling rate. Pixels with an attributed crop probability between 25% and 75% are sampled with a higher rate (table 1). The fact that the sampling plan does not use the SIGMA cropland map for the stratification might make the reference data set more easily usable to validate other cropland maps, as long as they have a compatible geometry.

A Global Agro-Environmental Stratification (GAES) provided by Alterra and FAO (Mücher et al., 2016), which defines strata based on agrosystem characteristics is applied. These zones will be used as post-strata for the computation, but were not used at the sampling stage.

Table 1. Strata definition from the IIASA cropland probability map

Stratum	Cropland probability
1	0
2	1-25%
3	25-75%
4	>75%

The sampling unit is the pixel of the map (300 m). The option of using 3 x 3 pixels windows to limit the impact of location inaccuracy has been discussed and discarded to reduce the photo-interpretation effort. Comparing the accuracy computed from the centre and the peripheral part of the 300 m unit should help quantifying the impact of location errors.

The usual approach to sample for the validation of a land cover map is independent random sampling in each of the classes of the map, considered as strata. When field work is necessary, sampling units are grouped into clusters to optimize logistics, but this has little interest when the reference data are obtained by photo-interpretation on publicly available images. An alternative approach is systematic sampling with a random starting point. Systematic sampling provides unbiased estimators of the parameter being assessed and is always more efficient than random sampling if the spatial correlation of the parameter of interest decreases with the distance (Bellhouse, 1988) and in particular for land cover data obtained from remote sensing (Dunn and Harrison, 1993). However systematic sampling has some drawbacks:

- It lacks flexibility if the sample size needs to be modified for practical reasons, adding more units in certain strata or reducing the initially foreseen sample (Stehman, 2009).
- Stratified systematic sampling may result in a loss of the spatial homogeneity of the sample distribution that is the basis of its good performance. This happens in particular when the strata are not defined by a limited number of large compact patches, but by scattered small areas.

- There is no unbiased estimator for the variance. The usual estimator for the variance may strongly overestimate it, although some alternatives have been proposed to reduce the overestimation (Wolter, 1984). We use a bi-dimensional adaptation of one of the formulas Wolter had assessed for unidimensional systematic sampling:

$$\hat{V} = \frac{\sum_{j \neq j'} w_{jj'} \delta_{jj'} \left(y_j - y_{j'}\right)^2}{2n \sum_{j \neq j'} w_{jj'} \delta_{jj'}} \quad (1)$$

where the weight $w_{jj'}$ is an average of the weights w_j and $w_{j'}$; $\delta_{jj'}$ is a decreasing function of the distance between *j* and *j'*, usually with zero value beyond a moderate distance.

An important advantage of systematic sampling appears when the estimates are politically sensitive and their acceptance is problematic or if there is a conflict of interest. For example if the validation is performed by the same team that has produced the map under validation, they can be tempted to eliminate difficult points, in particular if an economic interest is involved. It is more difficult to proof that a random sample has not been manipulated than it is with a systematic sample.

Figure 1 illustrates the irregular spatial layout of a random sample compared with a systematic sample and with a stratified systematic sample with the strata defined above based on the IIASA cropland probability map. In the stratified systematic sample the higher sampling rate for stratum 3 (higher uncertainty of cropland presence) has been applied by applying a grid step of 40 km instead of 70 km used for strata 2 and 4. A simple visual inspection suggests that the regular spatial distribution that gives some advantage to the systematic sampling is lost to a large extent with strongly mixed-up strata.

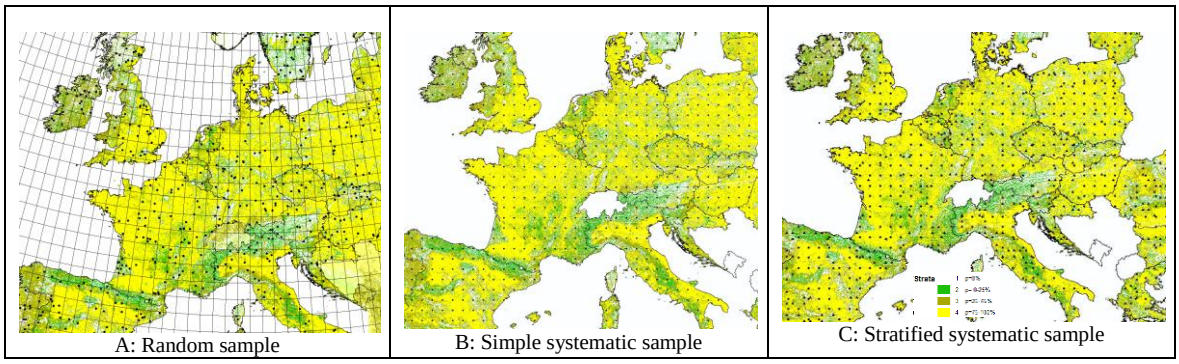


Figure 1: Random and systematic samples in central Europe in an equal-area projection.

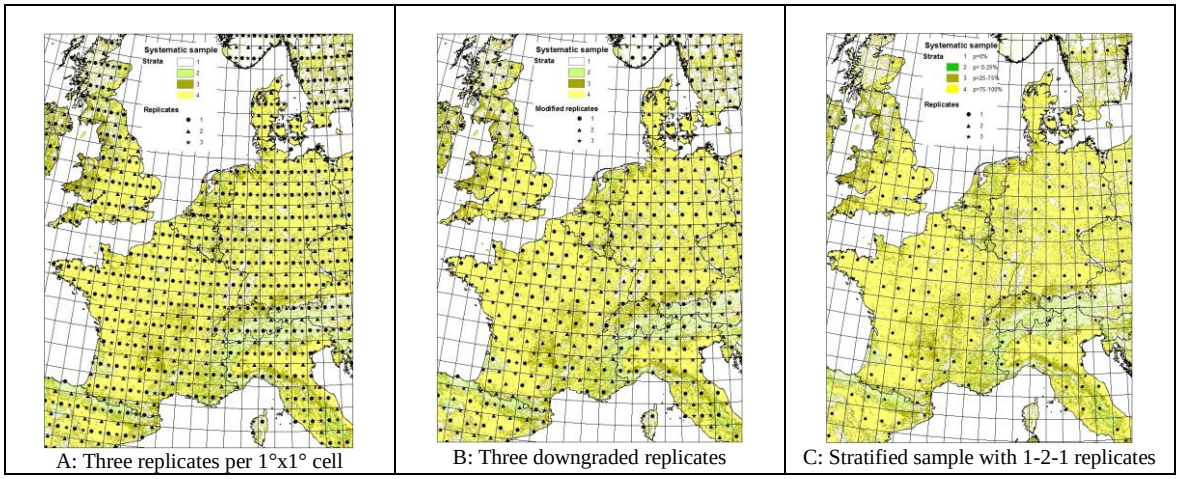
An additional issue with systematic sampling appears if we work with global products, such as the one we are considering, and we want to use geographic coordinates (latitude-longitude) in the sampling process. A straight forward systematic sampling in latitude-longitude is strongly unbalanced because the area of each cell changes with the latitude (111 km x 111 km x cos(lat) assuming a spherical shape).

One possible way to deal with this issue is keeping the systematic sample in latitude-longitude coordinates. This means we have an unequal sampling probability π and we have to apply 1/π as weight to the sample elements (Horvitz-Thompson estimator, Cochran, 1977). However this solution is not very efficient because the higher sampling rate far from the equator is not justified.

Figure 2 illustrates the approach we have used: Figure 2A represents a systematic sample built with three replicates in a 1º x 1º cell. Each replicate is the set of locations with the same relative position in each cell of the grid. Even if the area covered by the illustration is not very large, the higher density of the sample in northern latitudes is visible. This uneven density can be

rebalanced by modifying replicates in such a way that the number of points per replicate for a given latitude band (ring) is proportional to the land area in the corresponding ring (Figure 2B). Since the parallel at latitude α has approximately a length cos(α) compared to the equator, a proportion 1-cos(α) of points belonging to replicate 1 is downgraded to replicate 2, a proportion 2*(1-cos(α)) is downgraded from replicate 2 to 3 and so on. Very quickly we have a proportion rep*(1-cos(α))>1 and points may be downgraded for example from replicate 5 to replicate 8. In the current scheme the selection of points to be downgraded is purely random.

Figure 2B represents a stratified sample based on two modified replicates: stratum 1 is disregarded, 1 replicate is kept for strata 2 and 4 and 2 replicates are kept for stratum 3 (stratified sample 1-2-1). We can see that the geographical homogeneity is broken with some empty rows of cells. Some of the corresponding points have been dropped in the random selection of points to be downgraded from one to another replicate.



A: Three replicates per 1°x1° cell

B: Three downgraded replicates

C: Stratified sample with 1-2-1 replicates

Figure 2: Low density sampling schemes in latitude-longitude coordinates

Figure 3 gives an example of stratified systematic sample in latitude-longitude coordinates with 5 modified replicates for strata 2 and 4 and 10 for stratum 3 (stratified sample 5-10-5). It is similar to the stratified 1-2-1 above, but the irregularities are much smaller because the random subsampling for downgrading replicates happens only in the last replicate.

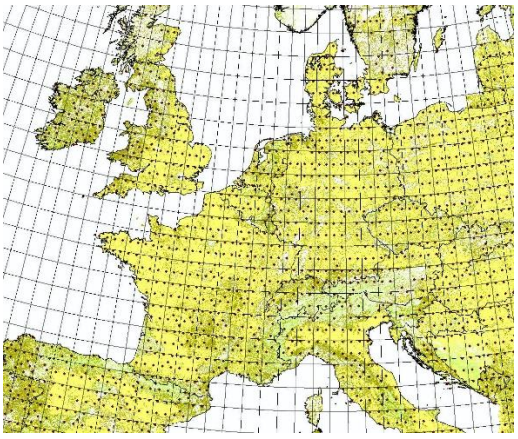


Figure 3: A higher density stratified sample with a systematic of replicates.

III SOME SIMULATIONS

At the moment of writing this paper, the SIGMA cropland map is not yet available and the reference data collection is still ongoing. For this reason we limit the scope of this paper to assessing the behaviour of the sampling schemes described above with two sets of data: the IIASA cropland probability map that we have used for stratification in SIGMA is used as pseudo-truth to be estimated. Considering a target on which we have an exhaustive knowledge allows us to make as many simulations as we wish and to distinguish the real variance of a sampling approach from the estimated variance computed with an empirical formula. We are aware however that the spatial structure of this map is not necessarily similar to the spatial structure of the inaccuracies of the SIGMA cropland map that are the target of our study. For the stratified versions we have chosen CORINE Land Cover (CLC) as data source (EEA 2007). For this reason, we have limited the simulation to the European Union (EU). The CLC land cover classes have been grouped into 3 strata: arable land, other agricultural classes and non-agricultural, that have been sampled with rates proportional to 3-2-1.

Table 2 summarises some of the results obtained. Simple random sampling (srs) is the benchmark with which the other standard errors are compared. Sampling in latitude-longitude coordinates, as if it were an equal area projection, introduces a bias of 7%. Therefore we do not even compute variances or relative efficiency, since such a bias is a sufficient reason to reject the method whatever the variance of the estimates. The same bias appears both with random and with systematic sampling: it has nothing to do with the sampling approach itself, but with the use of a reference system that is not equal-area. The other sampling schemes confirmed in the simulation to be unbiased. Simple systematic sampling has a relative efficiency of 1.22 (i.e. the variance is divided by 1.22 with the same sample size). This standard error could be computed in our example only because it was a simulation and we had an exhaustive knowledge of the sampled population. We can notice that using for systematic sampling the usual variance estimator for random sampling has a positive bias (overestimation) of 23%, leading to an "apparent efficiency" of 0.99. The estimator (Equation 1) based on local variance has a slightly positive bias and reports an estimated efficiency of 1.19. It should be noticed that the sample size in some approaches is not the same as the reference sample size of 1000 units we have chosen for the comparison. This highlights the rigidity of the usual systematic sampling towards the sample size.

Table 2. Comparison of mean and std. error of cropland probability for some sampling schemes.

Sampling method	Mean	n	Std error	relative efficiency
Simple random Sample (srs) equal area	53.1	1000	1.306	
Srs lat-long	49.45	****	****	****
Systematic latlong	49.45	****	****	****
Systematic equal-area	53.1	889	1.256	1.22
Syst. Eq-area usual variance estimator	53.1	889	1.391	0.99
Syst. Eq-area local variance	53.1	889	1.268	1.19
Systematic latlong balanced replicates	53.1	1000	1.213	1.16
Systematic latlong balanced replicates	53.1	100	4.07	1.03
Stratif random	53.1	1000	1.103	1.40
Stratified syst. equal area independent	53.1	959	1.056	1.59
Stratified systematic latlong balanced	53.1	1000	1.049	1.55

The systematic sampling schemes that we have labelled in the table as "balanced replicates" refers to the process described above to reduce the number of units in each replicate by a factor cos(latitude) (Figure 2B and 2C). If we consider a relatively large sample (n=1000 corresponding to a bit less than 3 replicates), we get a reasonably good efficiency of 1.16. In exchange a small sample of 100 units (a portion of one replicate) does not significantly improve the efficiency of a random sample of the same size. This happens because the subsample inside the replicate is random. At the same time the traceability of systematic sampling is lost to a large extent.

The stratification defined from CLC provides a more substantial improvement (efficiency=1.40). Combining stratification with systematic sampling roughly collects the advantages of both aspects in terms of efficiency. Independent systematic samples in different strata with different sampling steps in each stratum (figure 1C) gives a good efficiency in spite of the visually unpleasant pairs of points in different strata that are too close to each other. The type of scheme chosen for SIGMA (stratified systematic samplinglat-long balanced, illustrated in figure 3) behaves slightly worse, but still significantly better than stratified random.

IV SOME PROVISIONAL CONCLUSIONS.

The better performance of systematic sampling compared to random sampling is confirmed even if the relative efficiency is not particularly high. A relative efficiency above 1.5 for systematic sampling is usually not realistic. The rigidity of systematic sampling in terms of sample size can be removed by using a system based on a pattern of replicates. This type of system slightly reduces the efficiency compared to the pure systematic sampling. The usual variance formula for random sampling give a pessimistic bias if applied to systematic sampling. Better estimations are computed on the basis of local variance. The main advantage of systematic sampling is the traceability, i.e. the possibility to proof that no irregularities have been introduced in the sampling process.

Spatial sampling in a geographic reference that is not an equal-area projection, such as latitude-longitude can introduce a significant bias. This can be avoided with different adjustments, both in random and systematic sampling, but such adjustments in systematic sampling can strongly degrade its advantages, both in terms of variance and in terms of traceability.

The conclusions reported in this paper are based on an example of pseudo-truth, even if they are consistent with the conclusions of the wide literature on systematic sampling (see Bellhouse, 1988, for some references). Simulations with a pseudo-truth behaving in a more similar way to the errors of a land cover map would be useful, as would be, of course, the calculation of accuracy indicators on the SIGMA cropland map.

References

Bellhouse D.R., (1988). Systematic sampling, *Handbook of Statistics*, vol. 6, ed. P.R. Krisnaiah, C.R. Rao, pp. 125-146, North-Holland, Amsterdam

Cochran W., (1977) *Sampling Techniques*. New York: John Wiley and Sons

Dierckx, W., Sterckx, S., Benhadj, I., Livens, S., et al. (2014). PROBA-V mission for global vegetation monitoring: Standard products and image quality. *International Journal of Remote Sensing, 35*(7),

Dunn R., Harrison A.R. (1993). Two-dimensional systematic sampling of land use. *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 42 n. 4, pp. 585-601.

EEA (2007) *CLC2006 Technical guidelines*; EEA technical report 17/2007. http://www.eea.europa.eu/publications/technical_report_2007_17

FAO (1996). *Conducting Agricultural Censuses and Surveys*. FAO Statistical Development Series n. 6. FAO Publication, Rome.

Fritz, S., See, L., Mccallum, I., You, L., Bun, A., Moltchanova, E., Obersteiner, M. (2015). Mapping global cropland and field size. *Global Change Biology, 21*(5), 1980-1992.

Mücher, C.A., de Simone, L., Kramer, H., de Wit, A., Roupioz, L., Hazeu, G., Boogaard, H., Schuiling, R., Fritz, S., Latham, (2016). Global Agro-Environmental Stratification (GAES), SIGMA report D31.1.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment* 148, 42 - 57

Stehman, S. V. (2009). Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing, 30*(20), 5243-5272

Wagner, J. E., & Stehman, S. V. (2015). Optimizing sample size allocation to strata for estimating area and map accuracy. *Remote Sensing of Environment, 168*, 126-133.

Wolter K.M., (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, Vol. 79 No 388, pp. 781-790.

Exploring the Usefulness of Land Parcel Data for Evaluating Multi-Temporal Built-up Land Layers

Johannes H. Uhl*[1], Stefan Leyk[1], Aneta J. Florczyk[2], Martino Pesaresi[2]

[1] University of Colorado Boulder, Department of Geography, Boulder, Colorado, USA
[2] European Commission – Joint Research Centre (JRC), Institute for the Protection and Security of the Citizen (IPSC), Global Security and Crisis Management Unit, Ispra, Italy

*Corresponding author: johannes.uhl@colorado.edu

The increasing amount of free and open remote sensing data suggests that a number of multi-temporal land use and built-up land datasets derived from remotely sensed imagery will be made available soon. However, little research has been done regarding the approaches to evaluate spatiotemporal uncertainty of such products. Employing publicly available cadastral data with information on construction date may be useful for this purpose but requires developing a proper validation protocol. In this work we present preliminary results from a feasibility study that examines the potential use of land parcel data as reference data for spatiotemporal evaluation of such built-up land data, exemplified by the novel Global Human Settlement Layer (GHSL). The results indicate that alternative strategies shall be considered, as the use of parcel data tends to bias the evaluation results due to inherent mismatches between the two data sources, especially in rural areas.

I INTRODUCTION

The Global Human Settlement Layer (GHSL) is a methodology developed to identify and map built-up areas from Landsat satellite imagery (Pesaresi et al. 2013) and create a new global information baseline describing the spatial evolution of the human settlements in the past 40 years (Pesaresi et al. 2016). This global dataset is available at high spatial resolution (38m) and for various periods of time (around 1975, 1990, 2000, and 2014). GHSL data may provide new opportunities for population projections (Freire et al. 2016), disaster management and risk assessment (Freire et al. 2015), as well as for analysing and modelling urban dynamics and land use change.

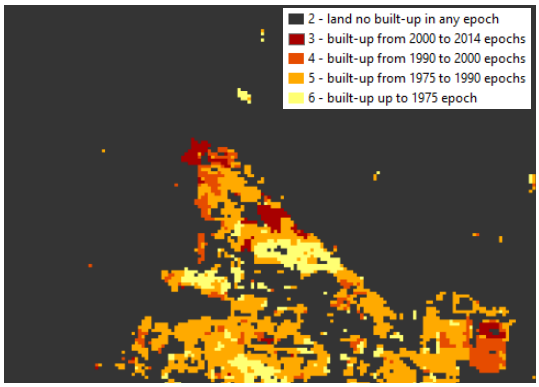
Before making such novel data products available to the research community, an extensive quality assessment is needed to demonstrate their usability. However, such assessments are difficult and rarely done due to the lack of reliable reference data, particularly for earlier time periods and in less developed regions. We carried out a first experiment to evaluate multi-temporal spatial data on built-up land such as GHSL or developed land cover classes in a typical land cover database using publicly available tax parcel (cadastral) data. How meaningful reference layers based on parcel data can be, represents an open question. This study is meant to shed light on the feasibility of such evaluations in order to establish a thorough protocol for validation studies in the near future.

II DATA AND METHOD

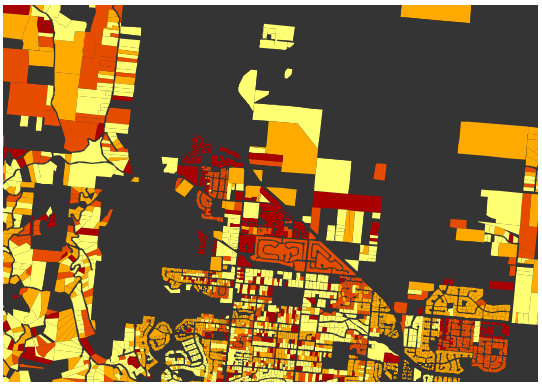
Open data policy makes cadastral and tax assessment data publicly available – often as GIS-compatible format – for many regions in the U.S. Often these parcel data contain rich attribute

information related to the type of land use, characteristics of the structure and the year when a structure in a parcel has been established (built year). To create spatiotemporal reference layers, the built year attribute is used to reconstruct snapshots of parcel-derived built-up areas that correspond to the GHSL time spans of built-up land (before 1975=class 6, 1975-1990=class 5, 1990-2000=class 4, 2000-2014=class 3) and its non-built-up land (class 2).

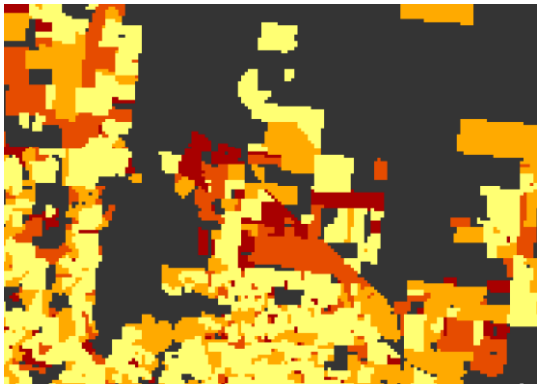
For each time span, GHSL built-up land is compared with parcel-derived built-up areas. This has been implemented in a raster-based approach where parcel polygons are rasterized to match the spatial resolution of GHSL (approx. 38m) using the time-span class as raster value (values 2-6). In order to comply with the GHSL-based built-up land definition, a pixel is classified as built-up when it overlaps with a built-up entity (Pesaresi et al. 2016). In this first experiment the parcel area is used as a proxy for built-up land despite well-known limitations (see below). This reference raster dataset is overlain with the original GHSL data and pixel-based confusion matrices are built to derive various accuracy metrics (Fielding and Bell 1997) for each time period. These metrics can be derived to quantify the classification accuracy of GHSL for each time span (cumulatively e.g., class 4 labels built-up land before 2000 including classes 5 and 6) and provide rich material for assessing its quality across space and time. Figure 1 shows GHSL built-up labels (a) and created reference data (b and c) for a subset of Boulder County, CO.



(a) GHSL raster data for Boulder, Colorado



(b) Parcels with built year (polygon objects)



(c) Parcels rasterized to GHSL resolution

Figure 1: GHSL built-up labels and corresponding reference data for North Boulder, Colorado.

Parcel data including built year information is publicly available for different administrative regions in the U.S. (states, counties, and cities) but not a standard product. In order to test parcel data for evaluation of GHSL in different settings with regards to development intensity, we included two rather rural (Figure 2) and two urban counties (Figure 3).

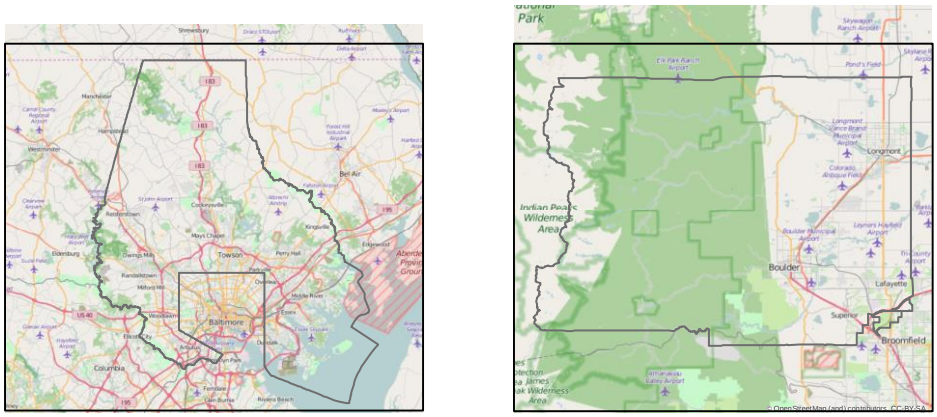


Figure 2: Rural study areas: Baltimore County (excludes the city of Baltimore), Maryland (left), and Boulder County, Colorado (right). Basemap: Open Street Map.

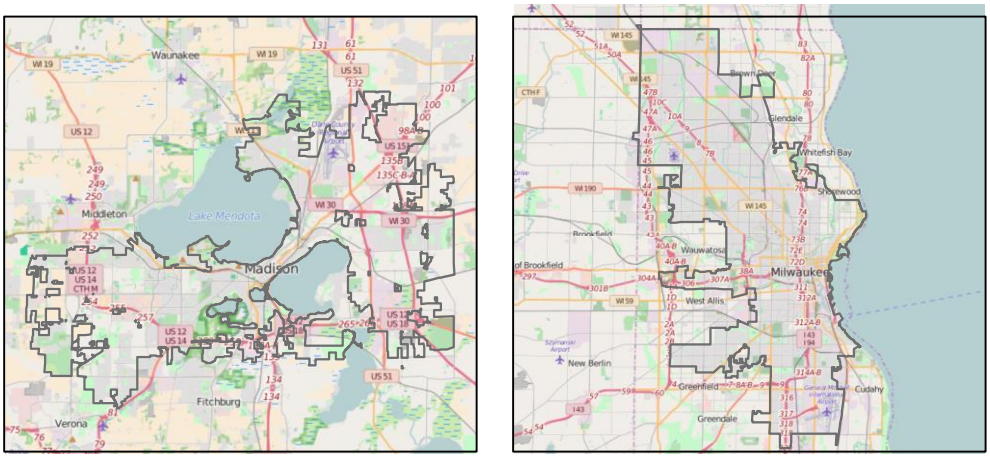


Figure 3: Urban study areas: City of Madison, Wisconsin (left), and Milwaukee County, Wisconsin (right). Basemap: Open Street Map.

III RESULTS

The results are based on the comparison of GHSL with the parcel-derived reference surface and intended to illustrate the feasibility of using parcel data to evaluate multi-temporal built-up land data. Confusion matrices are created for all time spans in both settings (Figures 4 and 5) and overall agreement metrics are shown in Table 1 for each study area. There are noticeable differences in agreement between rural and urban settings (e.g., Producer's Accuracy or Kappa).

	Average Producer's Accuracy	Average User's Accuracy	Kappa	Normalized Mutual Information	Overall Agreement	Average Omission Error	Average Commission Error
Madison (urb.)	0.497	0.428	0.284	0.153	0.568	0.503	0.572
Milwaukee (urb.)	0.369	0.304	0.220	0.095	0.538	0.631	0.696
Boulder (rur.)	0.275	0.636	0.155	0.062	0.673	0.725	0.364
Baltimore (rur.)	0.284	0.387	0.140	0.035	0.462	0.716	0.613

Table 1: Overall accuracy measures for the four study areas.

Confusion matrices for urban areas (Figure 4) show that while there is some variation among counties, in general the agreement for class 2 (non-built-up) and class 6 (built before 1975) is high. Yet, large portions of non-built-up areas (class 2) in parcel data have been classified as built-up before 1975 (class 6) in GHSL.

For rural areas very high agreement can be seen in non-built-up land. However, higher portions of built-up pixels in the parcel surface in all epochs (particularly class 6) have been classified as non-built-up pixels in GHSL (Figure 5) decreasing Producer's Accuracy. A possible explanation could be that portions of large rural residential parcels, which are mostly of agricultural use, are falsely assumed (and thus overestimate) built-up land in the reference data. These portions are often correctly classified as not built-up land (class 2) in GHSL resulting in this kind of disagreement.

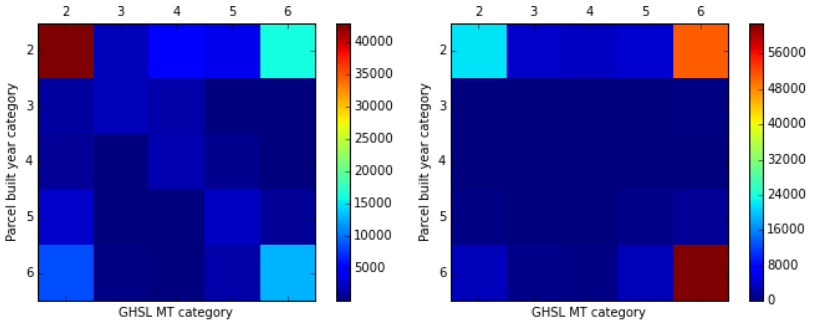


Figure 4: Overall confusion matrices for the urban study areas Madison, Wisconsin (left) and Milwaukee County, Wisconsin (right).

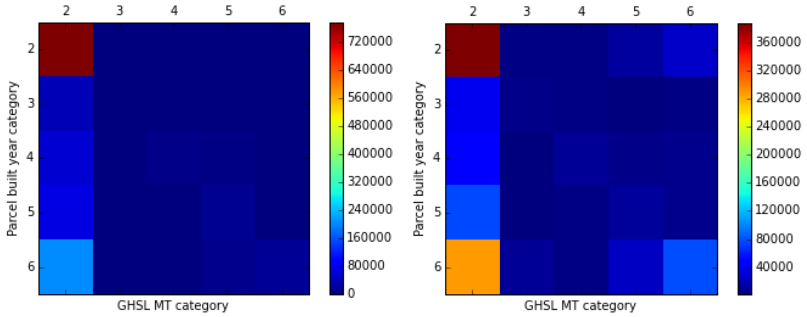


Figure 5: Overall confusion matrices for the rural study areas Boulder County, Colorado, and Baltimore County, Maryland.

Binary confusion matrices are used to derive agreement metrics comparing GHSL and parcel data for non-built-up labels (class 2) and built-up land in each epoch (classes 3-6). These metrics (Figures 6 and 7) also illustrate some interesting differences between rural and urban areas.

User's Accuracy (i.e., a pixel labelled class 4 in GHSL is also labelled class 4 in the parcel reference data) in rural areas increases over time and is higher for the built-up classes when compared with urban areas. Pixels of higher development intensity are more reliably detected as built-up in GHSL, and these locations are most likely within residential parcels in rural settings.

Producer's Accuracy (a pixel of class 4 in parcel data is classified as class 4 in GHSL) increases slightly in both settings over time but is much lower in rural areas due to the overestimated built-

up land in parcel data. Overall agreement (PCC) decreases towards more recent epochs due to an accumulation effect of disagreement in earlier epochs. Interestingly, Kappa, which accounts for chance agreement and is more conservative than PCC, increases over time in rural settings but decreases in urban areas. Possibly, this reflects higher detection rates in rural areas with improved technology, and at the same time a cumulative effect in urban areas similar to PCC.

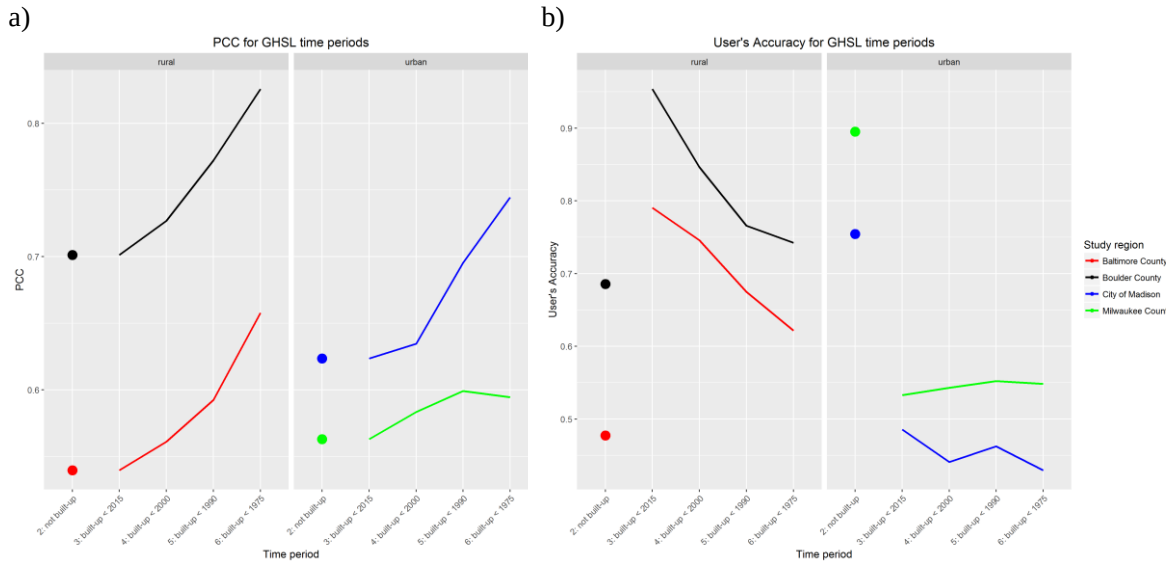


Figure 6: Trends in agreement metrics for urban and rural areas: a) PCC, and b) User's Accuracy.

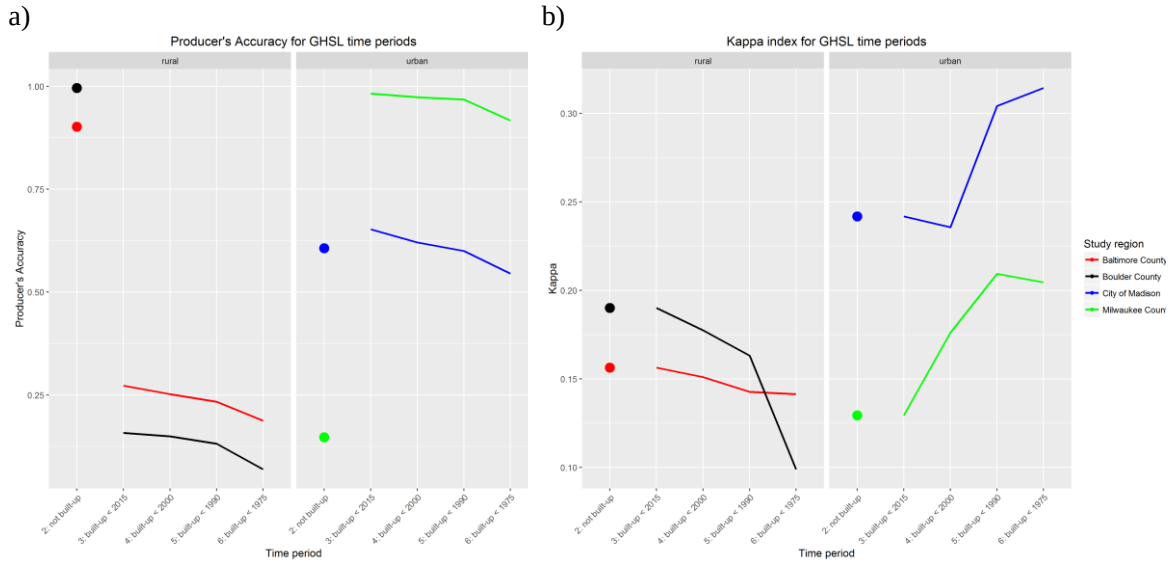


Figure 7: Trends in agreement metrics for urban and rural areas: a) Producer's Accuracy, b) Kappa.

IV DISCUSSION

Measuring the agreement in built-up land labels between publicly available parcel data and GHSL over time allows to investigate the usefulness of parcel data for evaluation of GHSL classes in rural and urban settings. While the built-year attribute provides a unique key feature for spatio-temporal evaluation purposes, the results show that the use of parcel data may bias the evaluation results due to inherent mismatches between the two data sources. Higher degrees of disagreement become evident in rural areas, where the residential parcels can be very large. Thus in order to establish robust protocols for thorough evaluation of GHSL the spatial integration of

building data, which are becoming increasingly available (e.g., LiDAR derived), will be tested to identify built-up areas within parcels more precisely thus creating more reliable reference datasets. Such an evaluation will also need to address uncertainty due to spatial offsets between datasets and the inherent uncertainty in the temporal information (e.g., (i) due to tear-downs and rebuilt structures, not reflected by the built-year attribute, (ii) the exact timestamp of an image and the assigned nominal class). Such an evaluation procedure will develop a broader understanding of the uncertainty across space and time in GHSL built-up area labels and inform the future user community about fundamental quality aspects if GHSL is applied to similar settings in other countries where no reference data are available. In future steps, the study area will be further extended, alternative criteria to differentiate urban vs. rural settings will be examined, and the appropriateness of the accuracy metrics employed will be critically reviewed.

References

Fielding, A. H., Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(01), 38-49.

Freire, S., Florczyk, A., Ehrlich, D. Pesaresi, M. (2015). Remote sensing derived continental high resolution built-up and population geoinformation for crisis management. In *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, 2677-2679.

Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E., Mills J., (2016). Development of new open and free multi-temporal global population grids at 250 m resolution. Proceedings of the 19th AGILE Conference on Geographic Information Science. Helsinki, Finland, June 14-17, 2016.

Pesaresi M., Ehrlich D., Ferri S., Florczyk A., Carneiro Freire S. M., Halkia S., Julea A. M., Kemper T., Soille P., Syrris V. (2016). Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. JRC Technical Report EUR 27741 EN; doi:10.2788/253582 (online)

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., Marin-Herrera, M. A., Ouzounis, G. K., Scavazzon, M., Soille, P., Syrris, V., Zanchetta, L. (2013). A global human settlement layer from optical HR/VHR RS data: concept and first results. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5), 2102-2131.

Urban objects recognition feasibilities by airborne hyperspectral and multispectral remote sensing

Sébastien Gadal[*1], Walid Ouerghemmi[*1]

[1]Aix-Marseille University, CNRS ESPACE UMR 7300

[*]Corresponding authors: sebastien.gadal@univ-amu.fr, walid.ouerg@gmail.com

Abstract

This paper explores the recognition uncertainty of urban objects by multiband imagery. The purpose is to recognize the urban objects by their spectral signature, using an external spectral library. Two Vis-NIR images were used for the study: a four bands Kompsat-2 multispectral image and a 16 bands Ricola's airborne hyperspectral image, two supervised classifiers were tested; a spectral based classifier, called the Spectral Angle Mapper (SAM), coupled to an external spectral library and a machine learning based classifier called the Support Vector Machine (SVM), in a second step the classification results obtained by the two classifiers were merged, the goal was to take advantage of both techniques, to optimize the classification result. The classifiers performance and the objects recognition feasibility were discussed for both images.

Keywords

Objects recognition uncertainty, spectral library, multiband imagery, urban environment.

I. INTRODUCTION

The cities space structure and territories result in a heterogeneous spatial and spectral mosaics of objects, which are combination of different manmade materials, urban vegetation species, and street networks. The recognition and characterization of urban objects becomes a difficult task in remote sensing image processing, requiring the uses of powerful data processing methods (e.g. Baillard et al, 1998). The use of multiband imagery is one of the approaches for identifying objects and urban space structures (e.g. Benediktsson et al, 2005), since it offer a precise characterization of the objects thanks to their spectral signatures. Two methods were selected for this study to deal with the complex context of urban objects recognition and classification. The first method called the Spectral Angle Mapper (SAM) (Kruse et al, 2003) is a simple and powerful classifier which is based on metric calculation, which is the angle formed between a spectral reference and a given pixel, each pixel is then classified depending on the value of the measured metric. Usually, the SAM method is applied using training samples acquired from the image, nevertheless, the sampling process could be complicated due to the presence of mixed pixels or to a lack of ground truth knowledge (e.g., Nidamanuri & Zbell, 2011). For this study, we used an external spectral library of urban materials in addition to the image-based spectral sampling. The use of such libraries in the exploitation of hyperspectral data is a growing and promising approach (Hueni et al, 2009; Zomer et al, 2009; Kotthaus et al, 2014). The second method called the Support Vector Machine (SVM) (e.g. Wu et al, 2004; Hsu et al, 2010) is a powerful classification method based on a machine learning algorithm, the method use training samples acquired from the image, to build optimal surfaces called the hyperplanes, and assign the

image pixels to these hyperplanes. Both classifiers were tested on the multispectral and hyperspectral images. Many study have tested and compared the efficiency of these method for urban land cover detection, burned areas detection, agricultural crops recognition, and geology mapping (e.g., J. Murphy et al, 2012; Hegde et al, 2014; Anggraeni & Lin, 2014). However, the fusion between these methods wasn't made. The second part of this work concerns the fusion of the SAM and SVM classification results, to enhance the overall classification accuracy and to take advantage of both methods.

II. STUDY ZONE AND MATERIALS

The study zone concerns the city of Kaunas (Lithuania) which is characterized by a large variety of urban materials including roofing materials, roofing's painting, roads, pavements, and urban vegetation. For the study we used a 16 bands airborne hyperspectral image acquired over the city centre of Kaunas, the camera includes a Finnish Vis-NIR sensor (RICOLA LTD), with a spectral range of 500-900nm, and a spatial resolution of $1m^2$. In addition to the hyperspectral image we used a 4 bands multispectral KOMPSAT-2 image, with a spectral range of 500-900nm, and a spatial resolution of $16m^2$. Both images were originally of radiance. They were converted to reflectance using a MODTRAN 4 radiative transfer model (Matthew et al, 2000).

To characterize the urban materials of the city, a spectral library was generated over Kaunas in July 2015. The spectral measurements were done in a black room using the superspectral Themis-Vision VNIR400H imaging camera (Fig 1.a). The sensor was developed by the NASA at the John Stennis Space Center. The camera is able to measure the electromagnetic radiation from 400 nm to 1000 nm using 1000 narrow bands; the spectral resolution is equal to 0.6 nm. The collected spectra were converted to relative reflectance values; using a standard white reference panel (99% of the Avian Technologies LLC white reference panel). The relative reflectance is equal to the brightness ratio of the sample grid to the white reference panel grid, multiplied by the standard spectral reflectance values (99%). The library includes thirty common urban materials, including roofing's material, roads, and vegetation types (Fig 1.b).

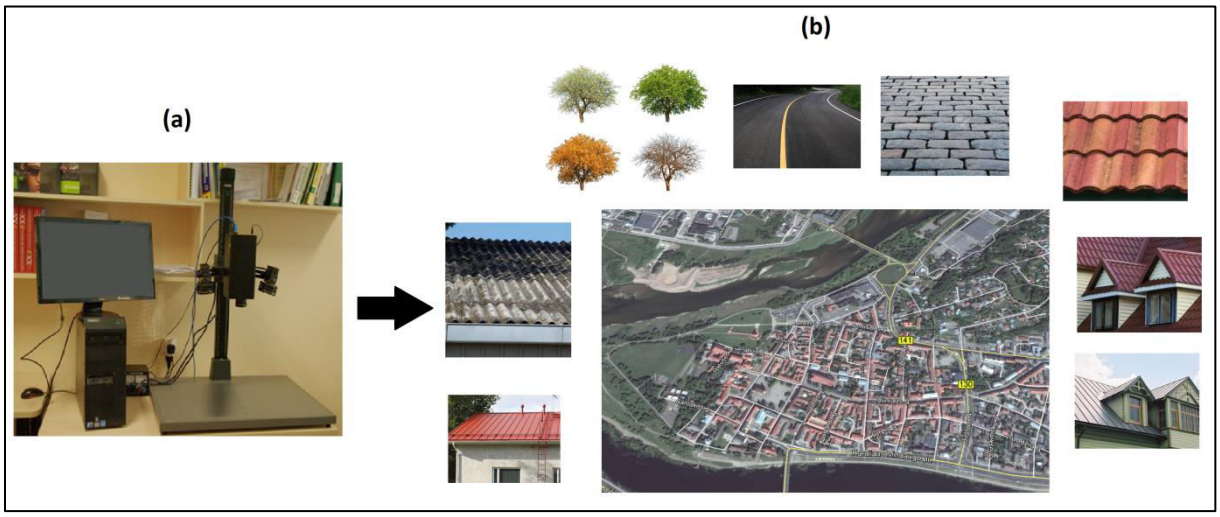


Figure 1: a) Themis Vision superspectral camera, b) Common urban objects in Kaunas city center (Lithuania) measured in laboratory.

In addition to the spectral library, we used a Geographic Information system (GIS) over the same study zone including many information's about the urban materials (i.e. photo, material type, color, date of acquisition, state) collected over Kaunas from December 2015 to March 2016. The GIS was used to validate the classification results and as a decision tool to extract training samples -if needed- for the classifications process.

III. IMAGE PROCESSING METHODS

a. PRE-TREATMENTS

Multiband images and spectral library were both converted to reflectance, concerning the images, an atmospheric correction using the Flaash module (Exelis Visual Information Solutions) were conducted, the correction is based on a modified version of the MODTRAN 4 radiative transfer code. Mid-latitude summer atmospheric model, 820 nm water feature, and no predefined aerosol parameters were used. The aerosol amount in air was estimated by visibility and was set to 40 km.

Concerning the spectral library, the relative reflectance was calculated with the Themis-Vision camera software, using a white reference panel that have a reflection coefficient of 99%, the relative reflectance value for a given pixel is then expressed as the brightness ratio of the sample grid to the white reference panel grid, multiplied by the standard spectral reflectance values (99%) (see Eq 1).

$$\rho_\lambda = \frac{NV_{sample}}{NV_{white\ reference}} \cdot 99\ \% \qquad (1)$$

b. CLASSIFICATION

Two classification methods were used for this study, which are the SAM and SVM, concerning the SAM method, the algorithm is based on a spectral classification scheme, and uses metric calculation. This is a powerful tool that permits to distinguish the classes by their spectral signatures. It decreases the shade influence to enhance the targets detection, and it didn't require any statistics about the distribution of the data. The metric used is the spectral angle θ between a reference spectrum *y* and an unknown spectrum *x* following this equation (see Eq 2):

$$\theta(x,y) = arccos\left(\frac{\langle x,y\rangle}{||x||||y||}\right) \qquad 0 \leq \theta \leq \frac{\pi}{2} \qquad (2)$$

Where $\langle .., .. \rangle$ is the dot product and $||..||$ is the 2-norm operator, smaller θ value indicate better match between *x* and *y*.

Concerning the SVM method, the algorithm is based on machine learning theory, below we introduce briefly the problem formulation, introduced by Vapnik (Vapnik, 1995) (see Eqs 3 and 4). Considering a two class problem in *n* dimensional space R^n, let's assume *l* training samples $x_i \in R^n$ with their corresponding labels $y_i \in \{-1, +1\}$, the method consists in finding the optimal hyperplane (i.e. surface) that separates the classes and maximizes the margin between them (i.e., distance to the closest training data points for both classes). The closest training data points to the hyperplane are called support vectors. We used the Radial Basis Function as a Kernel, the penalty parameter that controls the margin rigidity and misclassification was set to 100, no pyramidal

levels were fixed. The image was processed at full resolution without rescaling. The hyperplane was derived following Lagrangien problem below:

$$\text{maximize } f(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (3)$$

With the constraint: $0 \leq \alpha_i \leq C \ \forall\, i \in [1, l]$ and $\sum_{i=1}^{l} \alpha_i y_i = 0$, the normal vector of the hyperplane which give the maximal margin is given by:

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \quad (4)$$

Where C is the penalty parameter, K is the Kernel function, and α_i is non-zero vectors called the support vectors.

c. METHODS IMPLEMENTATION

In this study we present two classification strategies: a simple classification based on the separate use of SAM and SVM, and a classification based on the merging of SAM and SVM methods (Fig 2). The first step concerns the image pre-treatments. The radiance images are converted to reflectance using the MODTRAN 4 model; the multiband images generated by the laboratory camera Themis Vision were also converted to reflectance using the camera software (Hyper Visual Imaging Systems, Themis Vision Systems). The second step concerns the application of SAM and SVM methods. The SAM is coupled to an external spectral library in addition to an image-based library. The SVM is coupled to GIS database for a precise acquisition of training samples. The classification is followed by an artifact elimination to enhance the results interpretation and to delete some misclassified pixels. The last step concerns an original fusion strategy between the SAM and SVM which is guided by statistical accuracy indicators and ground truth knowledge.

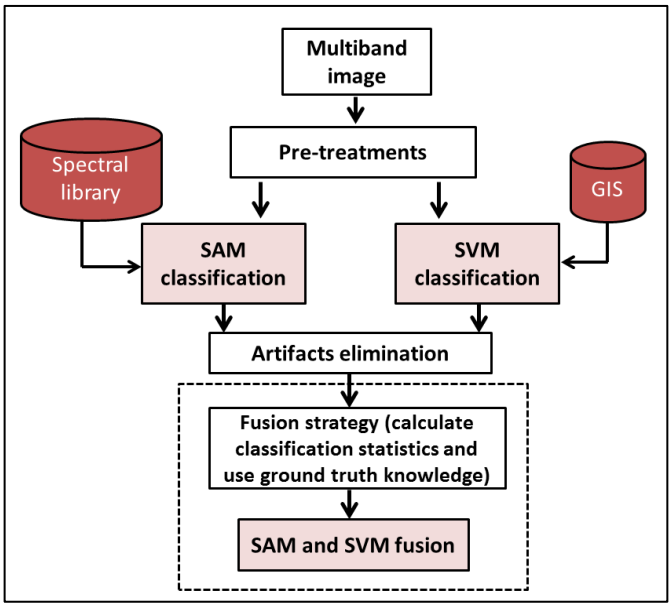


Figure 2: Method implementation

IV. URBAN MATERIALS CLASSIFICATION

a. CLASSIFICATION OF HYPERSPECTRAL RICOLA IMAGE

Concerning the SAM method, the results (Fig 3.a) reveal good identification of tile and red painted metal roofing's. The recognition is carried using the spectral library, nevertheless old tile with altered color was hardly recognizable by SAM, the incorporation of spectral samples of old tile to our library should enhance the detection of tile material. White painted metal and dark painted roofing's are recognized by spectral training samples extracted from the image; they were almost well recognized both. The trees are well recognized but sometimes confounded with grass. The pavements are weakly detected due to their high correlation with asphalt. Asbestos roofing's are not detectable due to their high correlation with roads. The results are evaluated by the Kappa coefficient and overall accuracy, which are respectively of 0.47 and 54%. The classification performance is moderately accurate due to the weak detection of some materials (i.e. vegetation and pavements mainly).

Concerning the SVM method, the results (Fig 3.b) reveal a good identification of both recent and old tile, nevertheless, an overestimation is observed, and the red steel roofing's are often misclassified with tile. The recognition is altered compared to the SAM results for this material. White steel roofing's, dark steel roofing's, and pavement are better recognized with SVM, the trees (coniferous) and grass are better separable, deciduous trees couldn't be detected due to the lack of dedicated validation points. Theoretically the SVM performs better than SAM, with a Kappa coefficient and an overall accuracy of respectively 0.79 and 82%. Nevertheless the algorithm tends to overestimate the pixels compared to the SAM. In the next step, the results of SAM and SVM will be merged to take advantage of both techniques.

Figure 3: Urban materials classification using a) SAM, b) SVM and c) SAM and SVM fusion.

The merging of the SAM and SVM (Fig 3.c) is done following the observations noticed previously. The SVM tends to overestimate the pixels classes for materials that are not

homogeneous in color or not highly reflective, which is the case of tile and red steel roofing's (Fig 3.b). These materials were detected with SAM method (Fig 3.a). The SVM estimates correctly the highly reflective materials (e.g. white roofing's, pavements) and the materials which are homogeneous in color (e.g. vegetation, some of the dark roofing's) (Fig 3.b). The white steel roofing's, the dark steel roofing's, the pavements and the vegetation were classified by SVM. The classification of coniferous trees was only possible with SAM, thanks to the spectral library (Fig. 3.a). The combination of SAM and SVM results (Fig 3.c) offered and interesting and precise pattern of urban materials over Kaunas city, with reasonable performances in terms of estimation indicators; the Kappa coefficient and overall accuracy were respectively of 0.71 and 75%.

b. CLASSIFICATION OF MULTISPECTRAL KOMPSAT-2 IMAGE

Concerning the SAM method, the results (Fig 4.a) reveal a correct identification of tile and red painted metal roofing's. The recognition is carried using the spectral library, an over estimation of the clay is noticeable. The detection of white steel roofing's is also possible. The detection of dark steel roofing's and asbestos roofing's is not possible using this 4 bands image. The pavements are partially detectable. Trees are detected partially and sometimes confounded with grass. The results are evaluated by the Kappa coefficient and overall accuracy, which are respectively of 0.48 and 57%, the classification performance is moderately accurate.

Concerning the SVM method, the results (Fig. 4.b) reveal a good identification of tile with moderate correlation with other materials. The red steel is weakly detected, in the other hand; the white steel roofing's and pavements are almost correctly detected, as for the SAM method, dark steel roofing's and asbestos roofing's couldn't be detected with this image. Concerning the vegetation; grass and trees are well distinguished, with an overestimation that could be noticed for the grass in comparison to the SAM method. Theoretically, the SVM method performs better than SAM, with a Kappa coefficient and an overall accuracy of respectively 0.8 and 84%. In the next step the SAM and SVM methods will be merged together.

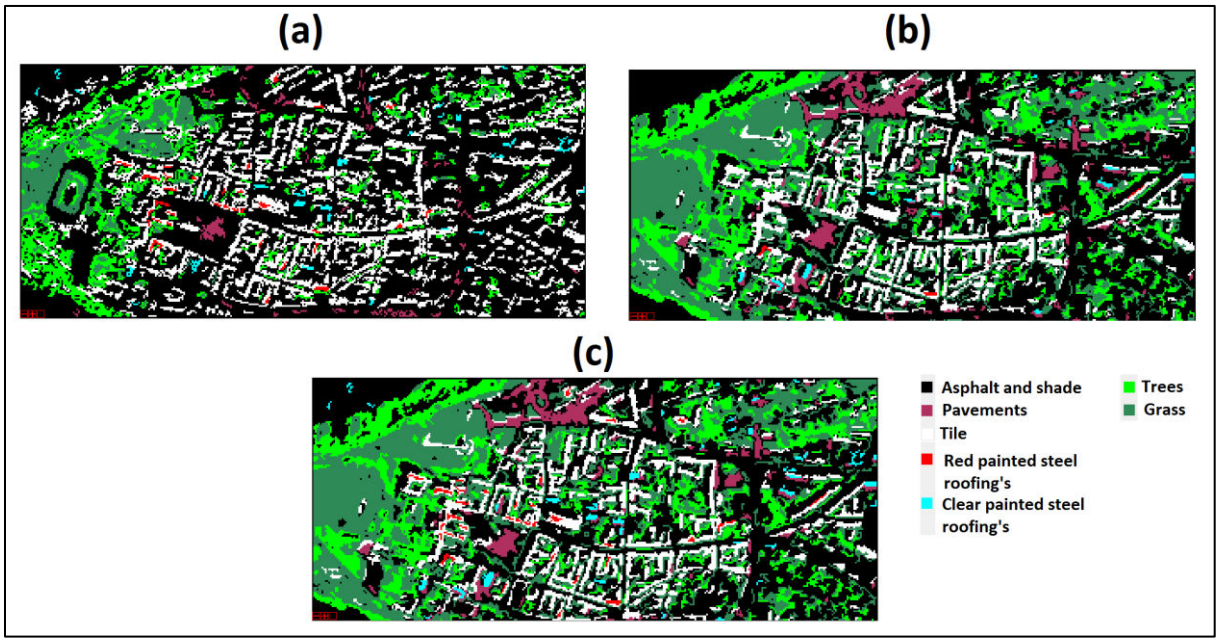


Figure 4: Urban materials classification using a) SAM, b) SVM and c) SAM and SVM fusion.

The SVM shows a better identification of tile, pavements and vegetation than SAM (Fig 4.b). The red steel roofing's is better detected with SAM (Fig 4.a), the white steel roofing's detection is correctly done with both methods. So both class of white steel roofing's are merged. The fusion of SAM and SVM (Figure 4.c) seems homogeneous in terms of class repartition with an accurate theoretical accuracy; the Kappa coefficient and overall accuracy were respectively of 0.81 and 85%.

V. CONCLUSION AND PERSPECTIVES

This study explored urban objects recognition using multiband imagery with two classification methods tested: the SAM and the SVM. The results obtained by both methods are merged to enhance the classification accuracy in a second step of processing. The SAM and SVM showed encouraging results in terms of detection with Ricola's 16 bands hyperpsectral image. Concerning the SAM, the materials were precisely recognized, nevertheless the external library couldn't cover all urban materials at the moment, therefore, image-based spectra have been used in addition to the spectral library to initialize the SAM. The overall accuracy of SAM was moderate due to non-sufficient detection of vegetation and pavements, the method accuracy is also sensible to the maximum angle affected to each class, which is tunable. Concerning the SVM, the method requires longer processing time than SAM, nevertheless it presented a better accuracy in terms of overall accuracy. The method inclines to overestimate certain classes, especially tile and vegetation. The results must be carefully examined and compared to ground truth knowledge before validation. In a second step we merged the SAM and SVM results. This fusion presents an encouraging opportunity for hyperspectral classification, and permits to overcome the limitations of both methods, while preserving an acceptable overall accuracy.
In addition to the hyperspectral image, we have tested our methods on a 4 bands multispectral image. The SVM seemed more suitable and powerful to process multispectral images than SAM. The moderate spectral resolution of the image is not suited for metric based classification. The fusion of SAM and SVM enhanced however -slightly- the overall accuracy and enhanced the identification of some materials (e.g. red steel roofing's). The enrichment of the spectral library will probably enhance the SAM classification performances.

Acknowledgments

This research was supported by the French National Research Agency (ANR) through the Hyep project (ANR-14-CE22-0016). We are thankful to Dr. Gintautas Mozgeris, Aleksandras Stulginskis University (Lithuania), for providing us with the hyperspectral image of 2015 and for allow us to perform the laboratory measurements with the Themis-Vision Camera.

References

Anggraeni, A., and Lin C. (2011). Application of SAM and SVM Techniques to Burned Area Detection for Landsat TM Images in Forests of South Sumatra. *International Conference on Environmental Science and Technology*, Singapore: pp. 160-164.

Baillard, C., Dissard, O., Jamet, O., and Maître, H. (1998). Extraction and textural characterization of above-ground areas from aerial stereo pairs: a quality assessment. *ISPRS Journal of Photogrammetry and Remote Sensing 53*, 130–141.

Benediktsson, J.A., Palmason, J.A., and Sveinsson, J.R. (2005). Classification of hyperspectral data from urban areas based on extended morphological profiles. In *IEEE Transactions on Geoscience and Remote Sensing,* vol. 43, no. 3, pp. 480-491

Hegde, G., Ahamed, J. M., Hebbar, R., and Raj, U. (2014). Urban land cover classification using hyperspectral data. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume XL-8, 751-754.

Hsu, C.-W., Chang, C.-C, and Lin, C.-J. (2010). *A practical guide to support vector classification*. National Taiwan University. http://ntu.csie.org/~cjlin/papers/guide/guide.pdf.

Hueni, A., Nieke, J., Schopfer, J., Kneubühler, M., Itten, K.I. (2009). The spectral database SPECCHIO for improved long-term usability and data sharing. *Computers & Geosciences 35*, 557–565.

Kotthaus, S., Smith, T.E.L., Wooster, M.J., Grimmond, C.S.B. (2014). Derivation of an urban materials spectral library through emittance and reflectance spectroscopy. *ISPRS Journal of Photogrammetry and Remote Sensing 94*, 194–212.

Kruse, F. A., Lefkoff, A. B., Boardman, J. B., Heidebrecht, K. B., Shapiro, A. T., Barloon, P. J., and Goetz, A. F. H. (1993). The Spectral Image Processing System (SIPS) - Interactive Visualization and Analysis of Imaging spectrometer Data. *Remote Sensing of Environment 44*, 145 - 163.

Matthew, M. W., Adler-Golden, S. M., Berk, A., Richtsmeier, S. C., Levine, R. Y., Bernstein, L. S., Acharya, P. K., Anderson, G. P., Felde, G. W., Hoke, M. P., Ratkowski, A., Burke, H.-H., Kaiser, R. D., and Miller, D. P. (2000). Status of Atmospheric Correction Using a MODTRAN4-based Algorithm. *SPIE Proceedings, Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI*. Vol. 4049, pp. 199-207.

Murphy, R. J., Monteiro, S. T., and Schneider, S. (2012). Evaluating Classification Techniques for Mapping Vertical Geology Using Field-Based Hyperspectral Sensors. In *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 8, pp. 3066-3080.

Nidamanuri, R. R., & Zbell, B. (2011). Transferring spectral libraries of canopy reflectance for crop classification using hyperspectral remote sensing. *Biosystem Engineering 110*(3). 231–246.

Plaza, A., Benediktsson, J.A., Boardman, J.W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., et al. (2009). Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment 113,* Supplement 1, S110–S122.

Samsudin, S. H., Shafri, H. Z. M., Hamedianfar, A., and Mansor, S. (2014). Spectral feature selection and classification of roofing materials using field spectroscopy data. *Journal of Applied Remote Sensing 9(1)*, 095079.

Villa, A., Chanussot, J., Benediktsson, J.A., and Jutten, C. (2011). Spectral Unmixing for the Classification of Hyperspectral Images at a Finer Spatial Resolution. In *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 521-533.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, New York: Springer-Verlag.

Wu, T.-F., C.-J. Lin, and R. C. Weng. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* 5, 975-1005.

Zomer, R.J., Trabucco, A., Ustin, S.L. (2009). Building spectral libraries for wetlands land cover classification and hyperspectral remote sensing. *Journal of Environmental Management 90*, 2170–2177.

International Conference on Spatial Accuracy (2016): Space-Time Kriging of Temperature over Van-Turkey

Pınar ASLANTAS[1], Okan YELER[2]

[1]Yuzuncu Yil University, Faculty of Agriculture, Landscape Architecture Department, TURKEY
[2]Yuzuncu Yil University, Muradiye Vocational School, TURKEY

*Corresponding author: pinaraslantas@yyu.edu.tr

Predictions of the variables at points with no measurements are obtained by interpolation techniques. Space-time interpolation techniques that consider variation both in space and time provided a new research area. Temperature is an important climatic parameter varying both in space and time. Like other meteorological, hydrologic and environmental variables, temperature is measured at specific locations. In order to obtain predictions for all grid locations, kriging methods have been applied for a long time. In here, space-time Ordinary kriging (ST-OK) and space-time Universal kriging (ST-UK) have been used in annual temperature estimation over Lake Van Basin using at 13 meteorological station observations for 2001-2011. Elevation, land cover, distance to Van Lake are used as secondary information in ST-UK. Elevation at 500 m resolution are obtained by Nearest Neighbour resampling of 3 arc second Shuttle Radar Topography Mission (SRTM). MOD12Q1 land cover data set has been downloaded from USGS organization. This dataset has 500 m spatial resolution and 17 subclasses. Distance to nearest coast variable is obtained by calculating the Euclidean distances of each SRTM pixel to the nearest boundary of the Van Lake coast vector. Annual temperature values are analysed and predicted at 500m*500 m resolution for 11 year-period. One-fold cross-validation is used to assess accuracy performance of both methods. R-square and Root Mean Square Error (RMSE) are calculated and evaluated for each technique. Comparison of kriging methods and inclusion of secondary information is assessed.

Key words: Space-time kriging, Van Lake Basin, Temperature

Introduction

Spatial kriging methods have been used for many years to predict variables at unmeasured locations in many disciplines. The first geostatistics and spatial kriging applications started in mining and geology. The variables used in these sciences can often be assumed constant in time. After understanding the usefulness and reliability of kriging in these disciplines, it was also introduced to many other disciplines within the earth and environmental sciences, such as meteorology, climatology, agronomy, soil science, hydrology, etc. Generally variables in these sciences vary both in time and space. Therefore the requirement of kriging methods for space-time interpolation is raised (Heuvelink and Griffith, 2010). If the data have been measured in different time and space locations, then more data may be used for prediction, and this allows obtaining more accurate predictions, helps to parameter estimation and helps to define spatial and/or temporal auto-correlation in measurements (Gething et al., 2007). In case of space-time kriging, to predict the value of the variable of interest at a specific location and time, past and

future measurements are used to predict on the specified time. This may add more complexity to the kriging procedure but may help to gain more accurate results.
In this study space-time Universal kriging (ST-UK) method is applied to average annual temperature values measured from 9 meteorological stations from 2001 to 2011 over the Lake Van basin of Turkey. The aim is to discuss applicability of space-time kriging methods on annual temperature values by using limited number of meteorological station.

Study Area and Data

The study area is Lake Van Basin that is located at the Far East part of Turkey (Figure 1). The area of basin is about 16.000 km^2. Lake Van basin has a high topography. The high mountains are located at the northern and southern parts of the basin. The mean elevation of basin is about 2200-2400 m., minimum elevation is about 1500 m. and maximum elevation is approximately 4000 m (Figure 2). Lake Van which is the biggest lake of country is located at the basin (Figure 2). The Lake is a depression state in the middle of high mountains. Lake has a surface of 3574 km^2, length of shoreline is 505 km, and a volume of 607 km^3. The lake stands at 1650 m. above sea level. The Lake is a closed lake without any significant outflow. With a maximum depth of 451 m and a volume of 607 km^3, it ranks fourth in water content among all the closed lakes of the world (Degens et al., 1984).

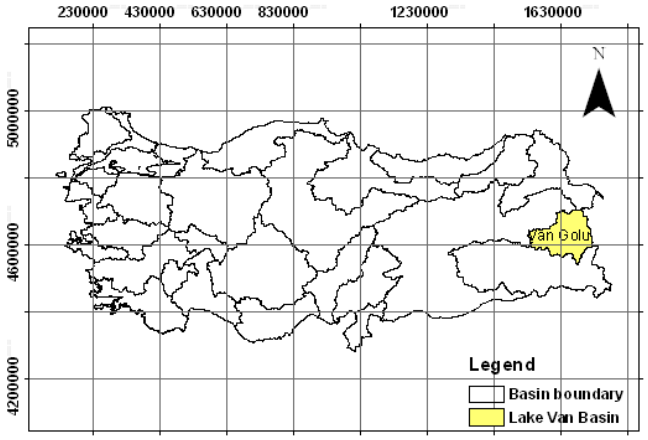


Figure 1: Location of Lake Van Basin on Turkey

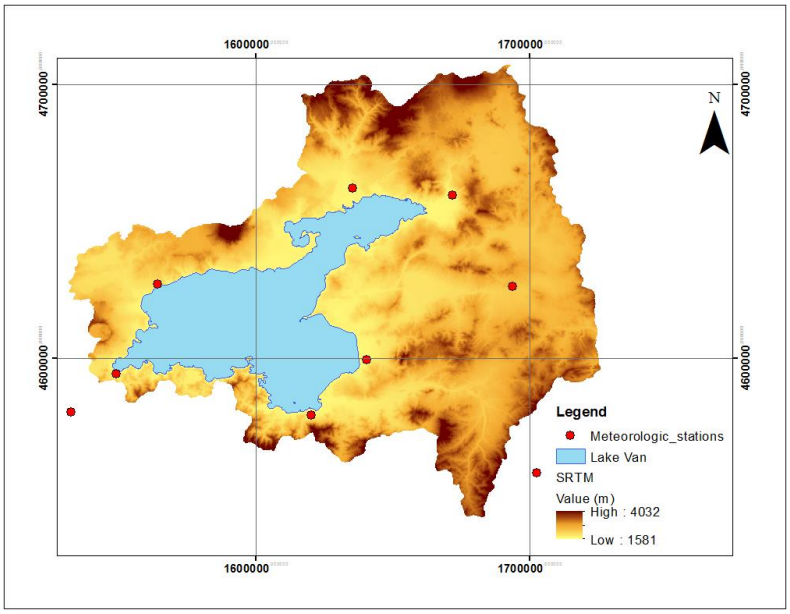


Figure 2: Location of Lake Van Basin on Turkey Lake Van, SRTM (90m) of basin, and distribution of meteorological stations over basin.

Data

The temperature data used in this study were obtained from the Turkish State Meteorological Service. The primary dependent data source was monthly temperature measured at 9 meteorological stations between 2001 and 2011. The spatial distribution of stations is not fairly uniform over the basin; when looking the overall distribution condensed placement can be seen near the Lake boundary (Figure 2). The highest average monthly temperature is between 24-25 ºC and is measured generally at the June. The lowest average monthly temperature is about -17 ºC and is measured at the January and Far East part of basin. It is observed in many studies that secondary information can often improve the spatial interpolation of environmental variables (Bostan et al., 2012; Lloyd, 2005; Hofierka, 2002; Boer, 2001). Therefore, three secondary variables correlated with temperature are selected to improve accuracy of predictions. As independent data source, an elevation map with 500 m spatial resolution was used (Figure 2). It was obtained by resampling the 3 arc second SRTM (the Shuttle Radar Topography Mission) (approximately 90 m spatial resolution) to 500 m spatial resolution. MOD12Q1 land cover data set has been downloaded from USGS organization. This dataset has 500 m spatial resolution and 17 subclasses. Distance to nearest coast variable is obtained by calculating the Euclidean distances of each SRTM pixel to the nearest boundary of the Van Lake coast vector.

Methodology

Annual temperature predictions are made on spatio-temporal framework. Space-time universal kriging (ST-UK) method is used to obtain predictions over the basin. Elevation, distance to nearest Lake Coast and land cover data sets are used as secondary variable. Seperable, productSum, Metric, sumMetric and simpleSumMetric variogram models are applied to sample variogram (Figure 3 and 4).

Space-time Kriging

Consider a variable z which varies in the spatial (s) and time (t) domain. Let z be observed at n space-time points (s_i, t_i), i=1, ..., n. These measurements constitute a space-time network of observations. However it is practically impossible to measure data point z at each spatial and temporal point. In order to obtain a complete space-time coverage, interpolation of z is required. The aim of space-time interpolation is to predict z(s_0, t_0) at an unmeasured point (s_0,

t_0), which is a node of a space-time grid. To predict z at these nodes, it is assumed to be a realization of a random function Z which has a known space-time dependence structure. Next $Z(s_0, t_0)$ is predicted from the observations and using the assumed space-time model (Heuvelink and Griffith, 2010).

The random function Z can be defined with a deterministic trend *m* and a zero-mean stochastic residual *V* as follows (1):

$$Z(s,t) = m(s,t) + V(s,t) \quad (1)$$

The deterministic trend *m* represents large-scale variations whereas the stochastic component *V* represents small-scale variations (Heuvelink and Griffith, 2010).

Results and Discussion

Space-time Universal kriging is performed to dependent variable: annual temperature. Figure 3 and 4 represent the plots of sample and modelled variograms to data. Optimum parameter value of the modelled variograms are used to select the best variogram model (Table 1). In addition, visual interpretation of modelled variograms are considered while selecting the appropriate model. According to these values productSum and sumMetric models resulted minimum and very similar optimum values. Therefore productSum model is selected for kriging operation (Figure 5).

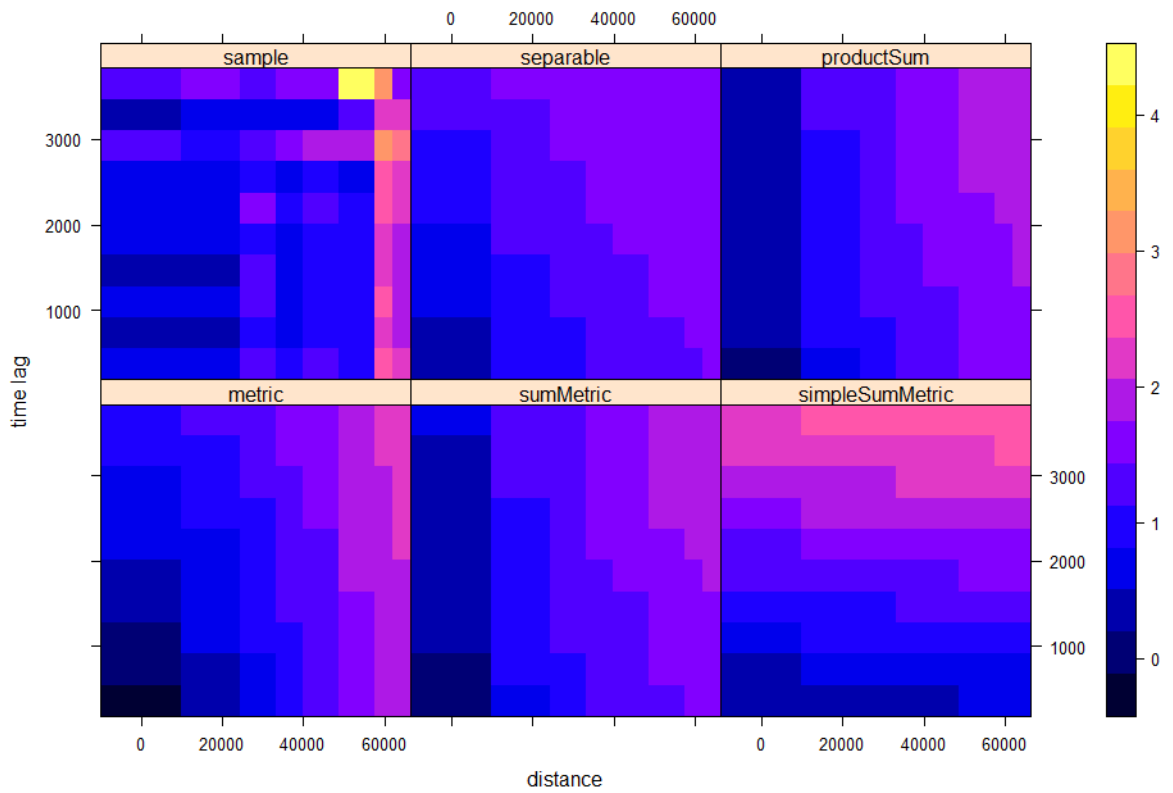


Figure 3: Plot of sample and modelled variogram models

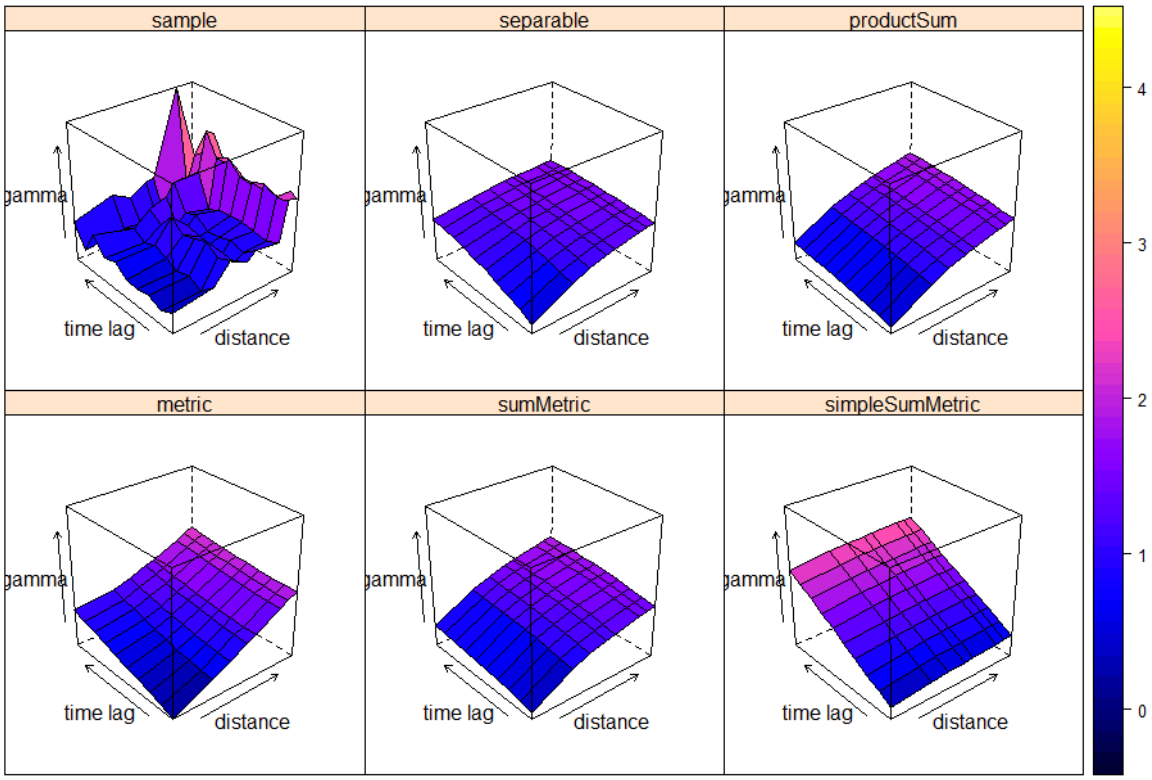


Figure 4: Wireframes of sample and modelled variogram models

Seperable	ProductSum	Metric	SumMetric	SimpleSumMetric
0,62	0,37	0,58	0,36	0,88

Table 1: Optimum Parameter Values of Variograms

Prediction maps of some selected years are represented at Figure 5. According to these maps, it is clearly seen that around the lake, temperature values are higher than the other regions. Around the Lake the area is more flat, after some distances elevation gets higher. So it can be said that distance to lake and elevation are highly correlated with temperature variable.

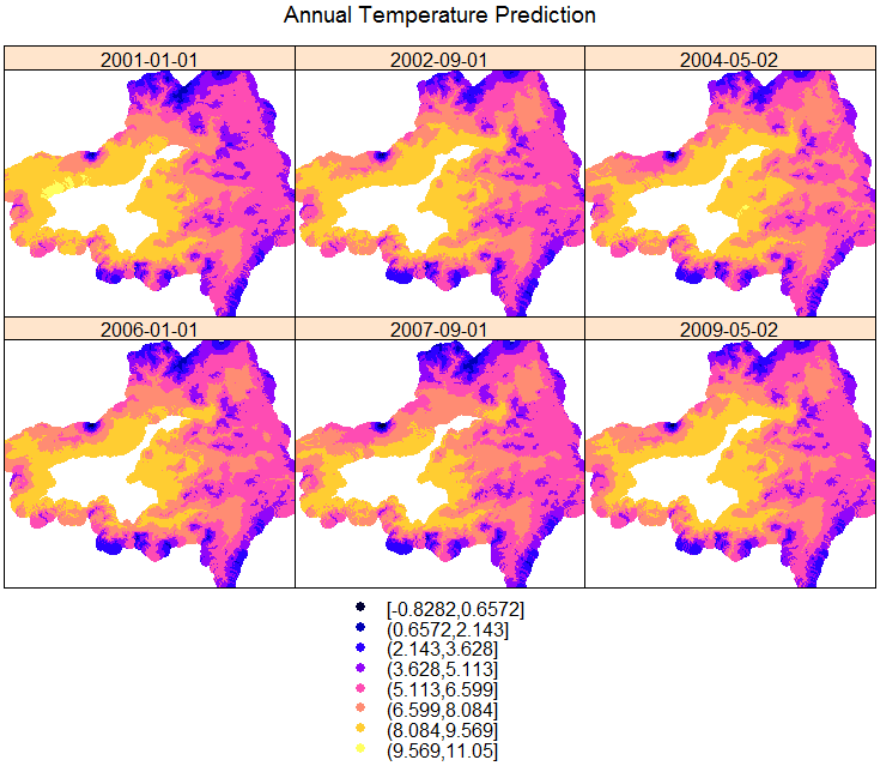


Figure 5: Prediction maps for some observation years

Conclusions

In this study space-time kriging method was applied to predict annual temperature of the Lake Van Basin, Turkey. Measurements obtained from nine meteorological stations were used for 2001-2011 time period. Secondary variables that vary in space but are static in time (elevation and distance to lake) and variable changes in space and time (land cover) were used by the space-time Universal kriging method. ST-UK method resulted with reasonable prediction values at space and time; however prediction values for each time scale are similar to each other. Limited number of observations were used at space-time kriging and gives satisfactory results with regard to spatial and temporal framework. However it is thought that using more dense observations would give more accurate results.

References

Heuvelink, G. B. M., Griffith, D. A., "Space–Time Geostatistics for Geography: A Case Study of Radiation Monitoring Across Parts of Germany", *Geographical Analysis* 42 (2010) 161–179.

Gething, P.W., Atkinson, P.M., Noor, A.M., Gikandi, P.W., Hay, S.I., Nixon, M.S., "A local space-time kriging approach applied to a national outpatient malaria data set", *Computers & Geoscienses* 33, (2007), pp 1337-1350.

Degens, E.T., Wong, H.K., Kempe, S., Kurtman, F., "A Geological study of Lake Van, Easten Turkey", *Geologische Rundschau* 73, 2, pp 701-734, (1984).

Bostan, P.A., Heuvelink, G.B.M., Akyurek, S.Z., "Comparison of Regression and Kriging Techniques for Mapping the Average Annual Precipitation of Turkey", *International Journal of Applied Earth Observation and Geoinformation* 19, pp 115-126, 2012.

Lloyd, C.D., 'Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain', *Journal of Hydrology* 308, pp 128-150, 2005.

Hofierka, J., Parajka, J., Mitasova, H., Mitas, L., 'Multivariate interpolation of precipitation using regularized spline with tension', *Transactions in GIS* 6 (2), pp 135–150, 2002.

Boer, E. P. J., Beurs, K. M., Hartkamp, A. D., 'Kriging and thin plate splines for mapping climate variables', *International Journal of Applied Earth Observation and Geoinformation* 3 (2), pp 146-154, 2001.

Spatial inference under uncertainty

Where (we think) the wild things are: comparing citizen generated and formal measures of wilderness

Alexis Comber[1*], Steve Carver[1]

[1] School of Geography, University of Leeds, UK

*Corresponding author: a.comber@leeds.ac.uk

Abstract
This paper compares mapped data describing the degree of wilderness form a formal data analysis – the Wilderness Quality Index (WQI) – against crowdsourced estimates of wilderness collected by the Geo-Wiki initiative. The analysis examines the quality of the crowdsourced wilderness reporting, and how that quality varies spatially, using a geographically weighted model. The results indicate globally a positive relationship between the 2 datasets, but with wilderness being under-estimated by the crowdsourced data. However, when local patterns are examined it is evident that in more wild areas (the Alps, eastern Norway, etc) the crowdsourced data have a stronger relationship with the formal data. This suggests that, while citizens may be good at identifying locales that at the extremes of a continuum of wilderness (*wild* and *not-wild*), they may not be so reliable in describing the wilderness in areas with a wilderness character between these extremes.

Keywords
Geographically weighted model; citizen science; wildness; Geo-Wiki

I BACKGROUND

This paper compares citizen measures of human impact from Geo-wiki with formal wilderness data to determine the degree to which citizens perceive wilderness. Geo-Wiki contributors record the degree of human impact [0-100%] and the land cover they perceive to be present. Different approaches have been used to map wilderness and activities at continental / global scales use a continuum / environmental modification spectrum (Kuiters et al. (2011). Other approaches such as the 2012 Google data that applied a 10km buffer to its road dataset under the assumption that roadless areas represent natural ecosystems that are undisturbed. The aim of this work was to compare two different datasets describing wilderness and wild land, one collected under the Geo-Wiki initiative (Perger et al., 2012; Fritz et al., 2012) and the other as part of a pan-European wildness project described in Kuiters et al. (2011).

Geo-Wiki uses Google Earth imagery and asks contributors to label features at randomly selected points. There have been a number of campaigns and interface developments. While much work has examined the utility and quality of the land cover information that can be derived from Geo-Wiki land cover data (See et al 2103; Foody et al 2014, Comber et al 2013), where users had to allocate scenes in Google Earth imagery to one of a predefined number of classes, as yet little work has considered the uncertainties, accuracies and errors associated with user contributed measures within a continuum, such as wildness. Specifically this work seeks to determine whether the effort involved in developing traditional maps of wilderness (as a suitability layer) results in significantly higher quality data products. Or whether, more

informal approaches, such as are facilitated by crowdsourced data, are able to support the development of data of quality to be scientifically useful.

II DATA & ANALYSIS

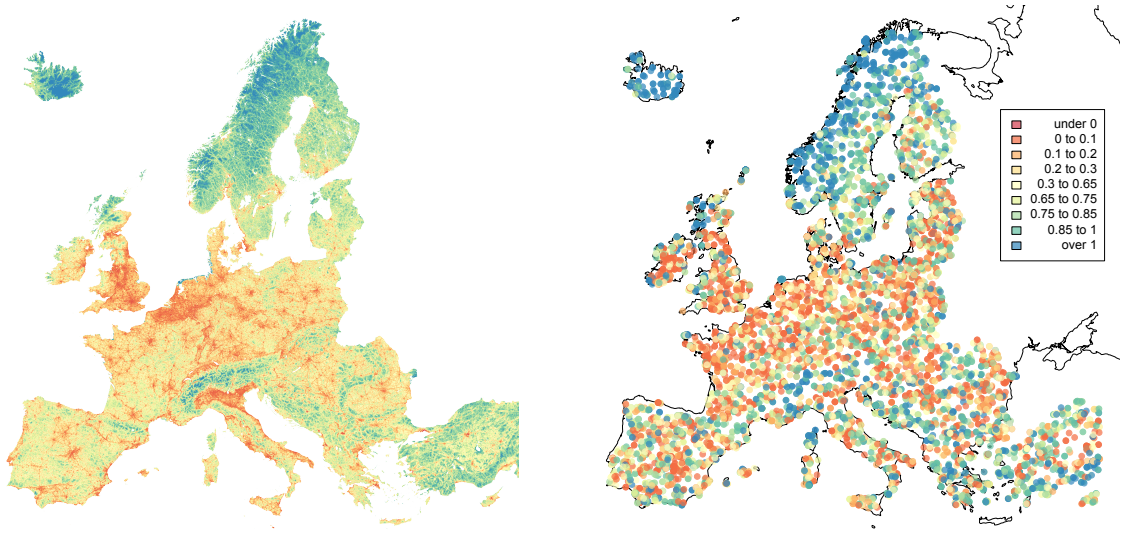
Figure 1 shows ~13,000 Geo-Wiki data points and formal data the wilderness quality index (WQI, Kuiters et al. (2011). It shows that, broadly, both datasets are describing similar distributions of wildness. Iceland, the Alps and North western Scotland has all have high wildness values in these areas.

The WQI data reports wildness measure in the range [0, 1] at 1km scale. The Geo-Wiki data has a number of attributes including land cover class (from a pre-defined set of 10 classes), a Human Impact score which was treated as the inverse of a wildness measure and subtracted from 1, and a self-reported measure of Geo-Wiki volunteer confidence for each point that was labelled.

The analysis used a standard OLS regression to model the relationship between Geo-Wiki volunteer measures of Human Impact, under the assumption that this describes the inverse of wildness, with the formal measure of wilderness quality index. For each Geo-Wiki data point, the coincident wilderness measure was extracted from the WQI data. The model describing the degree to which Geo-Wiki Human Impact predicts WQI data was constructed using the following:

$$y = \beta + \beta_1 x_1 \qquad \text{(Eqn 1)}$$

where y is the WQI at each data point and x_1 is the Geo-Wiki wildness value derived from the Human Impact value, and β_1 is the coefficient estimate for Geo-Wiki wildness. This identified constructed a global model of the relationship between these 2 variables.



a) b)

Figure 1: a) the Wildness Quality Index data, darker, bluer areas are more wild, and b) the wildness measures from the Geo-Wiki platform, with a small transparency term

Next a Geographically Weighted model was constructed to examine whether and how the coefficient estimates vary. The idea here is that global statistical models implicitly make an

assumption that the relationship between the dependent and independent variables is constant over the study area. In reality this is frequently not the case. In such situations, where spatial non-stationarity is thought to occur, local models can provide a more detailed and informative analysis. The Geographically Weighted (GW) framework (Brunsdon et al., 1996) supports a suite of local models in order to account for local variation in statistical relationships between variables (Gollini et al, 2015). A GW regression (GWR) was used to examine how the coefficient estimates of the degree to which Geo-Wiki measure predict WQI measures from of a standard OLS regression vary spatially. A GWR version of the OLS (Equation 1) is defined as follows

$$y_{(u_i,v_i)} = \beta_{0(u_i,v_i)} + \beta_1 x_{1(u_i,v_i)} \qquad \text{(Eqn 2)}$$

where (u_i, v_i) is a vector of two dimensional co-ordinates describing the location of i over which the coefficient estimates are assumed to vary. The outputs provide spatially varying estimates of the coefficients an the aim of generating these was to better understand how informal, citizen generated measures of landscape wildness relate to formal measures, how these vary spatially and how measures of wildness interact with the self-reported confidence value attached by volunteers to their labelling.

III RESULTS

Figure 2 shows the Geo-Wiki wildness scores (inverted from the Human Impact value) against the WQI data. The OLS model suggested a coefficient estimate β_1 of 0.356 for the Geo-Wiki wildness value, indicating that globally, each increase of 0.1 in the WQI data is associated with an increase of 0.036 in the Geo-Wiki data. The results of the all the regressions are summarised in Table 1. The distribution of the GWR coefficient estimates indicates that there is considerable local variation from the global model.

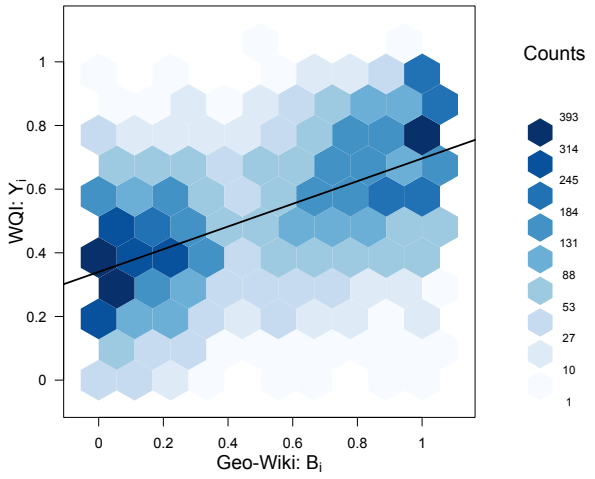


Figure 2: WQI against rescaled Geo-wiki wildness scores, with the slope of regression coefficient.

OLS	Estimate	Std. Error	t value	Pr(>|t|)	Sig
(Intercept)	0.3400	0.0031	108.5000	<2e-16	***
Geo-Wiki wildness	0.3563	0.0051	69.9700	<2e-16	***

GWR	Min.	1st Qu.	Median	3rd Qu.	Max
(Intercept)	0.1602	0.3190	0.3797	0.4718	0.7934
Geo-Wiki wildness	-0.0643	0.2046	0.2813	0.3385	0.5746

Table 1. The results of the OLS regression and the GW regression

Next the coefficient estimates were mapped to show how the coefficient estimates linking wildness score from Geo-Wiki volunteers relate to formal WQI wildness index scores. This is shown in Figure 3 and indicates distinct areas and gradients where volunteers found it difficult to accurately assign wilderness values. Clearly there is high agreement in the Alps and in eastern Norway but large differences between formal measures and volunteered ones in for example eastern France. Under the assumption that the formal WQI data are correct, this suggests that there are locales where people's perceptions of wilderness differ hugely from the way that it is formally recorded.

This paper presents the first analysis of cognitive aspects of how citizens perceive wilderness. Future work will seek to unpick the origins of these differences in relation to landscape context (for example the land cover and land use in these areas), user self-assigned confidence in their scores and their interaction. It will also consider the use of Geo-Wiki data to validate other datasets as has been done for many global land cover initiatives, and how to objectively quantify the accuracy of either dataset.

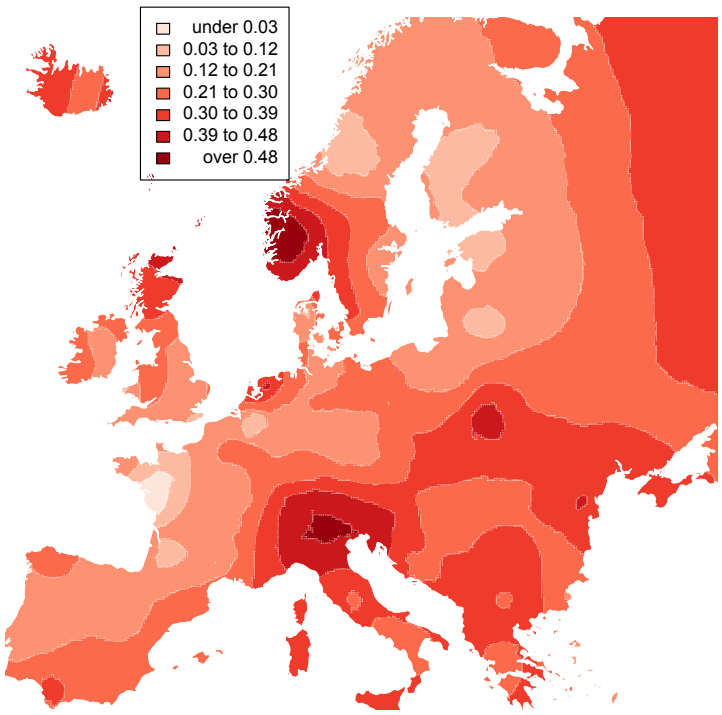


Figure 3: The spatial variation of degree to which informal Geo-Wiki measures of wildness predict formal measures from the WQI datset.

References

Brunsdon, C.F., Fotheringham, A.S. and Charlton M. (1996). Geographically Weighted Regression - A Method for Exploring Spatial Non-Stationarity, *Geogr Anal*, 28, 281-298.

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., Foody, G.M. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. *International Journal of Applied Earth Observation and Geoinformation*, 23: 37–48 http://dx.doi.org/10.1016/j.jag.2012.11.002

Foody, G.M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., Boyd, D.S. and Comber, A., (2014). Accurate attribute mapping from volunteered geographic information: issues of volunteer quantity and quality. *The Cartographic Journal* DOI: http://dx.doi.org/10.1179/1743277413Y.0000000070

Fritz S, McCallum I, Schill C, Perger C, See L, Schepaschenko D, van der Velde M, Kraxner F and Obersteiner M 2012 Geo-Wiki: An online platform for improving global land cover. *Environmental Modelling and Software* 31: 110- 123.

Kuiters, A. T., van Eupen, M., Carver, S., Fisher, M., Kun, Z., & Vancura, V. (2011). *Wilderness register and indicator for Europe*. Final Report - http://ec.europa.eu/environment/nature/natura2000/wilderness/pdf/Wilderness_register_indicator.pdf

Perger C, Fritz S, See L, Schill C, Van der Velde M, McCallum I and Obersteiner M 2012 A campaign to collect volunteered geographic Information on land cover and human impact. In: Jekel T, Car A, Strobl J and Griesebner G (Eds.) *GI_Forum 2012: Geovizualisation, Society and Learning*. Herbert Wichmann Verlag, VDE VERLAG GMBH, Berlin/Offenbach, pp.83-91.

See, L., Comber, A.J., Salk, C., Fritz, S., Van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F. and Obersteiner M. (2013). Comparing the Quality of Crowdsourced Data Contributed by Expert and Non-Experts. *PLoS ONE* 8(7): e69958. doi:10.1371/journal.pone. 0069958

Uncertain clustering of social specialization in metropolitan areas

Giovanni Fusco[1*], Cristina Cao[1]

[1] UMR ESPACE, CNRS / University of Nice Sophia-Antipolis, France

*Corresponding author: giovanni.fusco@unice.fr

Abstract: Instead of defining *a priori* target populations, sociodemographic variables and sector of residence are combined in order to identify geographically meaningful clusters of households on the French Riviera. A Bayesian classifier produces uncertainty-based clusters whose uncertain knowledge is represented through an interactive geo-dataviz solution. Cluster characteristics, sociodemographic content of geographic sectors, socio-spatial contrasts between neighboring sectors and proximity in variable space are explored taking into consideration the uncertain character of the available knowledge.

Keywords: Social Specialization, Uncertainty, Clustering, Bayesian Networks, Visualization, French Riviera

I INTRODUCTION

Social specialization of residential space within an urban area is the concentration of households according to some characteristics like social status, demography, ethnicity, etc. in different urban subspaces. The metropolization process is often associated to increased social specialization in large urban areas (Lacour and Gaschet 2008). Taken to the extreme, social specialization of residential space can produce residential segregation, creating socioeconomic dividing lines within the metropolitan space and undermining social and territorial cohesion (Massey 1985). In France, for example, important policies are carried out in order to assure that municipalities within metropolitan areas have a minimum social mix in terms of household income (Blanc 2010). Nevertheless, knowledge of social specialization of space in metropolitan areas is still incomplete and general assumptions have to be confronted to empirical data in case studies. Understanding the logics of social specialization is of course crucial in order to define policies like urban planning or housing. More specifically, which groups of households are most opposed in residential space? And which socio-demographic factors contribute most to these oppositions in space? How segregated are they in space? If we recognize that the answers to these questions are affected by high levels of uncertainty, which uncertainty-based methods could be used in order to describe our uncertain knowledge of social specialization? Finally, how could we represent and communicate most effectively this uncertain geographic knowledge?

In order to answer these questions, a new clustering approach is proposed integrating both sociodemographic descriptors of households and geographic distribution of their place of residence. The case study of the analysis is the metropolitan area of the French Riviera, a coastal conurbation of more than 1 million inhabitants stretching over 60 km west of the French-Italian border, and including the coastal cities of Nice (348 000 inhabitants in 2013),

Antibes (76 000 inhabitants), Cannes (74 000 inhabitants) and the Principality of Monaco (38 000 habitants), which is an independent city-state within French territory. These cities form nowadays an urban continuum and extend their influence over their alpine hinterland, which absorbs a growing part of the metropolitan population. Being a traditional residential destination for affluent retirees, a more recent hub of high-tech development and the necessary home for large populations of low-skilled workers of the tourist and residential economy, the French Riviera is particularly concerned by social specialization of residential space (Centi 1993, Billard and Madoré 2009, Fusco and Scarella 2011).

II METHODOLOGY: BAYESIAN CLUSTERING WITH SPATIAL CONSTRAINT AND INTERACTIVE GEO-DATAVIZ

Analyses of social specialization normally start by identifying target populations, whose segregation indicators are later calculated (Apparicio 2000). In our research, we opted for a bottom-up uncertainty-based approach: clusters of households were identified through data mining of selected sociodemographic variables within the 2008 Household Mobility Survey (which unfortunately does not cover the Principality of Monaco) combined with place of residence. Our starting hypothesis is that wealth differences are not the only factor beyond social specialization. A sample of 7539 households, representative of the population of the 94 sectors of the metropolitan area, has been analyzed through 16 variables describing social status, household structure, household demography and place in the workforce. Weights are attributed to variables in order to give approximately the same total weight to the four different dimensions of the analysis (Table 1), social status being slightly overweighted as its three variables cover more diverse issues. The *a priori* clustering of variables in four groups has been validated through Bayesian clustering algorithms based on mutual information among variables, given the empirical data. A 10-fold cross-validation procedure yields an average fit score of over 90% for this variable grouping.

Thematic Area	Indicator	Weight
Place in the Workforce	Occupation of the person of reference	1
	Number of working active people	1
	Number of unemployed people	1
Total Weight 4	Number of inactive people	1
Social Status	Maximum profession and socio-professional category among the spouses	2
	Maximum qualification among the spouses	2
Total Weight 6	Occupancy status of the dwelling	2
Household Composition	Presence of spouse in the household	1
	Number of household members	1
	Number of children	1
Total Weight 4	Number of other members	1
Household Demography	Number of minors (0-17 years)	0.8
	Number of young adults (18-29 years)	0.8
	Number of adults 30-59 years	0.8
Total Weight 4	Number of seniors (60-75 years)	0.8
	Number of elderly (more than 75 years)	0.8

Table 1 – The 16 indicators used to cluster households on the French Riviera.

Once the residence sector of the household is added as a further variable, different strategies of Bayesian clustering (Korb and Nicholson 2004) have been explored in order to produce an uncertainty-based socio-geographic clustering. More particularly, a naïve Bayesian classifier has been used with the variable weights of Table 1, with different constraints on maximal

number of clusters and minimal cluster content and with different weights for the new variable sector of residence. A minimal cluster content of 4% of the household population has finally been selected to avoid overfitting on the sample. As for the sector variable, weights beyond 3 forces the algorithm to consider it as the leading variable of the clustering: the likelihood of the resulting clustering is maximized by assigning the population of a given sector to one or two clusters only. This is clearly not the goal of our clustering, as we want to find dividing lines in the resident population that take into account place of residence together with and not instead of sociodemographic differences (the poor quality of clustering results based on place of residence only is also witnessed by the low contingency table fit on the 17 variables). The optimal compromise was found with a weight of 2 for the sector variable. In this case, the sector variable is only the fourth strongest variable in terms of mutual information with the clusters and the contingency table fit is 36,4% on all the variables.

This approach is different from previously developed research (Pallez et al. 2015): clustering is uncertain because household assignment to a given cluster is probabilistic and households can have several non-zero probabilities of being assigned to different clusters. Average probabilities in assigning a given household to a cluster range between a maximum of 0.97 (for cluster 11) to a minimum of 0.86 (for cluster 4). Individual households can have much lower probabilities, and sometimes even have similar probabilities of being assigned to two different clusters. Passing from the sample to the household population introduces additional uncertainties. Cluster labels are vague, too, in the sense that they are synthetic descriptions combining different variable values which are often (but not constantly) associated. The sociodemographic characteristics of clusters are thus described through Bayesian probabilities. Social specialization of metropolitan sectors with respect to these clusters is evaluated in terms of the classical dissimilarity index (Duncan and Duncan 1955), but different evaluations are proposed for different levels of uncertainty.

One of the main difficulties of the analyses was to convey the uncertainty associated with the results obtained. Several authors have already proposed approaches for graphical representation of uncertain information (MacEachren 1992, MacEachren and Howard 1993, Ehlschlaeger et al. 1997, Cedelnik and Rheingans 2000, Ward 2002). Within our research, we propose an interactive online geo-data visualization solution in order to explore the results of uncertainty-based analyses (Cao and Fusco 2015). Systems of dashboards for interactive visualization seem particularly useful in representing uncertain and complex phenomena like social specialization of space. First attempts of interactive representation of uncertain geographic data were proposed in the 1990s (Ehlschlaeger et al. 1997) even if they were not considered particularly effective in their applications (Evans 1997). Advances of the software interfaces have since been considerable and new applications seem to be both more user-friendly and scientifically sophisticated (Ban and Ahlqvist 2009, Kunz et al. 2011, Fusco et al. 2016). These applications link together interactive representations in the forms of maps, diagrams and text. In our case, different analyses are developed with respect to uncertain knowledge and represented in the geo-dataviz solution.

III RESULTS FROM THE FRENCH RIVIERA

Results of the Bayesian clustering of households in the French Riviera will be presented and commented through four visualizations, taken from the interactive geo-dataviz solution (Cao and Fusco 2015).

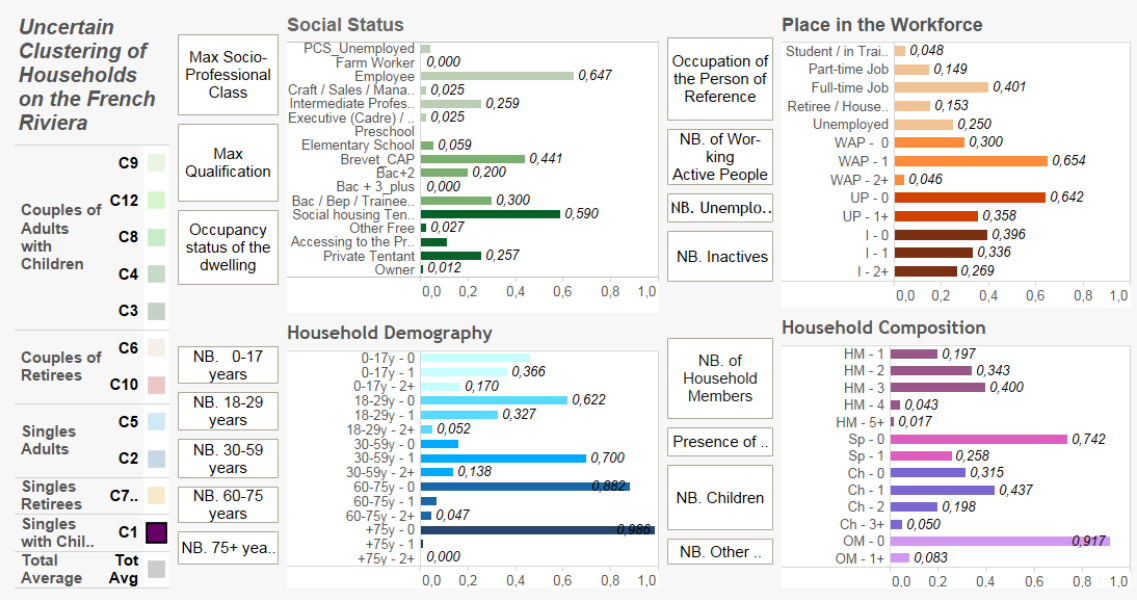


Figure 1: Probabilistic description of household clusters.

Figure 1 gives a probabilistic description of household clusters. The Bayesian classifier identified 12 clusters of households, which were later regrouped in 11 clusters after interpretation : five clusters concern families with children and differ in terms of social status, from highest (Cluster 9, families of skilled executives/professionals, very often owners of their dwelling) to lowest (Cluster 3, families of employees, with lower level of education and more frequently tenant than owner, sometimes social housing tenant); two clusters concern couples of retirees (of different social status); two clusters concern single adults (here social status is at least partially correlated with age); two clusters concern single retirees (they have been regrouped in one cluster only as their sole difference was the different age class 60-75 years or 75 years and more); a last cluster is specific to households of single parents with children, with difficult social situation (low education levels, employee not always with a full-time job and sometimes unemployed, tenant of social housing or in the private sector). The latter is, by no surprise, the most segregated cluster, with a dissimilarity index of 0.75. The exploration of the probabilistic values of the different variables for each cluster is of course richer than this simple summary of main features and lets the user better understand the vague content of the cluster labels and the subtle differences that sometimes exist between two relatively similar clusters.

Figure 2 represents through maps and diagrams the socio-demographic content of metropolitan sectors. Clusters of households are differently distributed in space. Every sector of the metropolitan area has a different probability profile in terms of cluster belonging of its resident population. Clusters can also be of particular importance for a given sector. In this visualization, like in the following ones, sectors can be linked to clusters according to different criteria: modal cluster, most overrepresented cluster (largest positive deviate from metropolitan

average), most underrepresented cluster (largest negative deviate from metropolitan average), most characteristic cluster (highest location quotient compared to the metropolitan average).

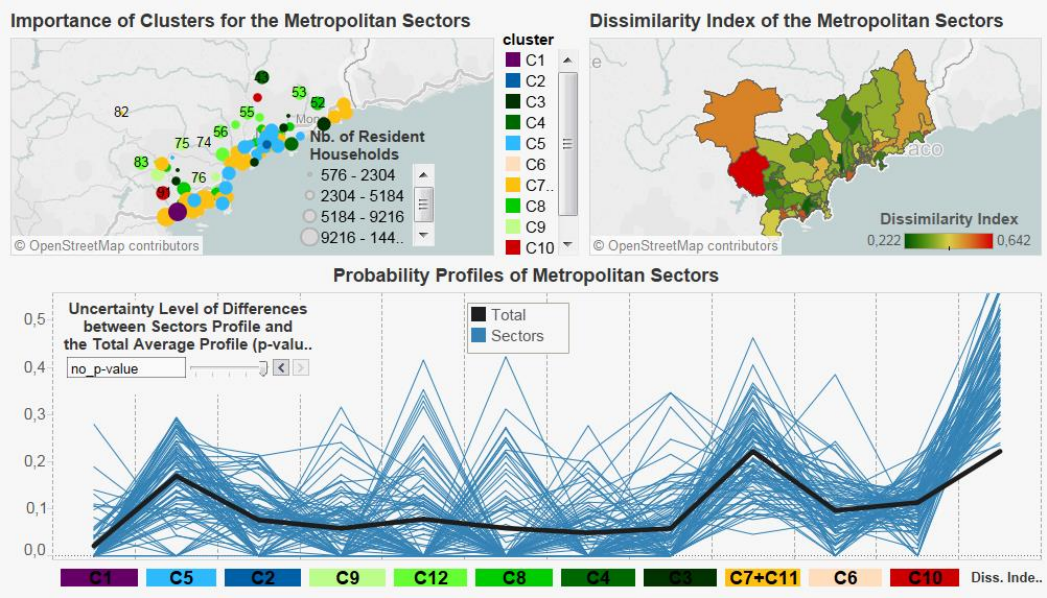


Figure 2: Sociodemographic content of metropolitan sectors.

Finally, the dissimilarity index measures which proportion of a sectors' population should move in order to attain the same probability distribution as the whole metropolitan area. Uncertain knowledge of population content of sectors can be explored in the visualization. Uncertainties derive both by the uncertain cluster assignment of households and by the sample process. Statistical significance of the observed deviates from metropolitan-wide values is evaluated against the null hypothesis that the percentage of a given cluster in the household population of the sectors is given by a binomial distribution whose expected value is the share of the cluster in the whole metropolitan household population. Given the sample size of every sector (always between 70 and 103), the binomial distribution can be approximated by a normal distribution. In the data-viz, the significance level can be set by the user: every time more certainty is required in knowledge, non-significant differences are omitted. The city-center of Cannes, for example, with strong and extremely significant over-representation of single retirees, is not necessarily the sector with the highest dissimilarity index, but becomes such once only the most significant deviations from the metropolitan values are retained. The user can thus interactively identify the deviates characterized by the lowest levels of uncertainty, which contribute significantly to the social specialization or residential space.

Socio-spatial contrasts between neighboring metropolitan sectors are represented in Figure 3. The diverse spatial repartition of clusters within the metropolitan area results in socio-demographic differences among contiguous sectors, which can have important impacts on the social functioning of the metropolitan area. The population content of sectors being known as probability distributions, their differences are measured as distribution divergence. More particularly, we use the Jensen-Shannon divergence (Lin 1991), which has the advantage of being symmetrical and defined even when some clusters are absent in a given sector. Socio-demographic distance between contiguous sectors is thus coherently measured with a probabilistic divergence. Once again, the uncertainties in socio-demographic content of sectors

can be used to produce different measures of socio-spatial contrasts, according to significance level of differences from the metropolitan-wide values. The user can thus identify those contrasts which are identified with the lowest uncertainty. The socio-spatial contrasts in the city of Nice (at the center of the metro area) are thus determined with higher levels of certainty than those west of Nice, where some differences are relatively uncertain.

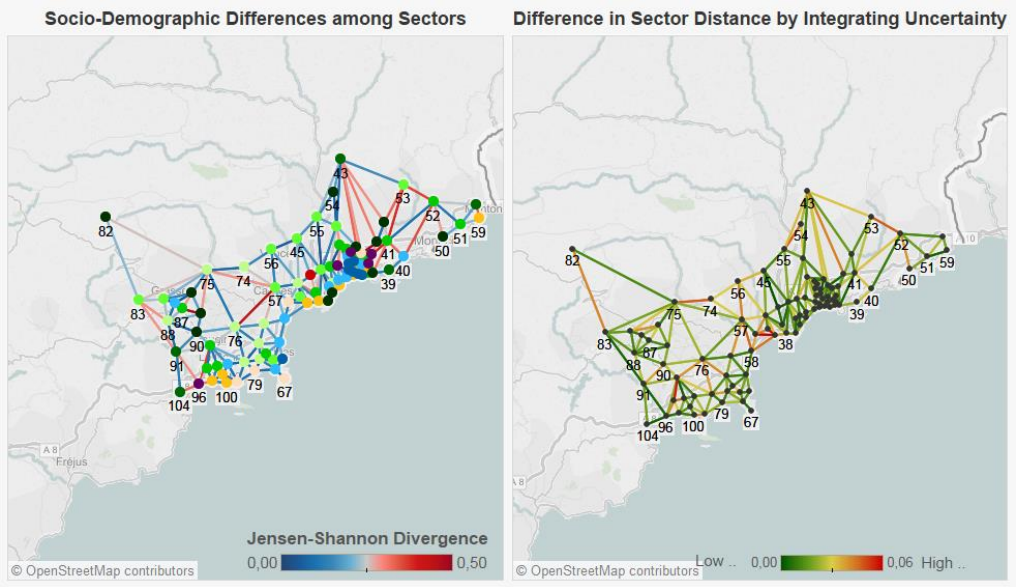


Figure 3: Socio-spatial contrasts between neighbouring metropolitan sectors.

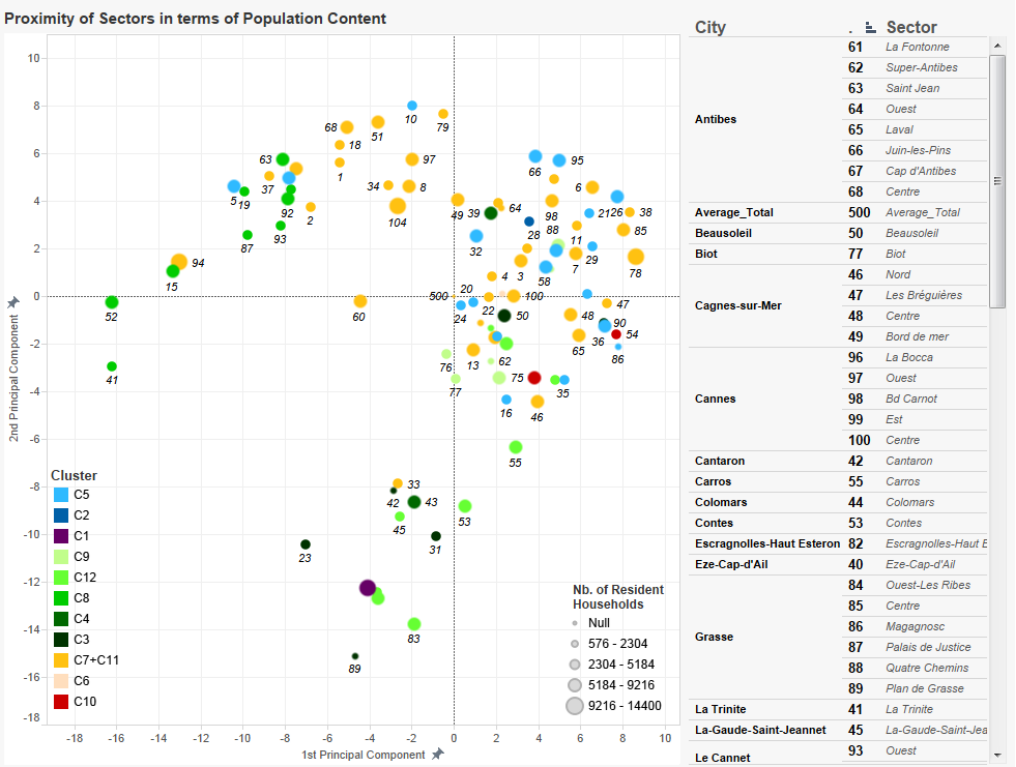


Figure 4: Proximity of metropolitan sectors in variable space.

Figure 4 represents proximity of metropolitan sectors in variable space. Jensen-Shannon divergence values make up a complete distance matrix among the metropolitan sectors. A principal component analysis can be performed in order to visualize the relative proximity of sectors in terms of socio-demographic content, on the main factorial plane. Geographical space proximity can thus be evaluated in conjunction with sociodemographic proximity in a multi-dimensional scaling approach. Once again, reducing the level of uncertainty through minimum significance levels modifies the results of the analysis. Through this analysis we can highlight the extreme diversity of social specialization within the city of Nice, a phenomenon which is not observable in other main cities like Cannes and Antibes. At the same time, the city of Cannes presents a marked opposition between its westernmost sector, strongly marked by the presence of single parents with children, in difficult social situation, and the rest of the city, where all sectors show very high presence of single retirees, and low presence of couples with children. All these results are particularly robust to uncertainty levels and can be considered among the most salient characteristics of our study area.

IV CONCLUSIONS AND PERSPECTIVES

In conclusion, with the example of the French Riviera metropolitan area, our research shows how knowledge of social specialization of residential space is uncertain. In this context, instead of defining *a priori* target populations, soft clustering techniques could be used to identify the most important sociodemographic divides in a given metropolitan area. For the French Riviera, position in the life-cycle as described by age and household composition seems even more important than social status in defining social specialization of space. Results of soft clustering benefit from appropriate geo-data-visualization solutions. In our example, a system of dashboards seems an appropriate way to describe complex phenomena like social specialization of space. Within this solution, knowledge uncertainty can be conveyed by interactive representations, whenever an appropriate calculus (in our case through the use of probabilities) can quantify the most relevant uncertainties.

The present work opens important perspectives in the search for uncertain geographically meaningful clusters. Concerning the analysis of social specialization of space, ethnicity is a crucial aspect which is missing in the French statistical information system and could only be integrated in the study through *ad hoc* surveys. As far as the methodology is concerned, the use of a naïve Bayesian classifier is a first solution and more sophisticated methods should produce better results. Multi-level clustering with hierarchical naïve Bayesian classifiers (Langseth and Nielsen 2006) could for example better exploit the structure of available information, where strong relations exist within groups of variables. The problem of identifying a correct weight for the geographic variable could be eliminated altogether by the use of algorithms of multi-objective optimization. We could thus optimize at the same time clustering likelihood based on the sociodemographic variables and dissimilarity index based on the geographic distribution of the clusters. The exploration of the Pareto front would then be instructive of role of the two criteria in the identification of socio-geographic clusters.

References

Apparicio P. (2000). Residential segregation indices : a tool integrated into a geographical information system. *Cybergeo: European Journal of Geography*, 134, http://cybergeo.revues.org/12063

Ban H., Ahlqvist O. (2009). Representing and negotiating uncertain geospatial concepts - Where are the exurban areas?, *Computers, Environment and Urban Systems*, 33(4), pp. 233-246.

Billard G., Madoré F. (2009). Les Hauts du Vaugrenier: un exemple atypique de fermeture résidentielle en France, *Mappemonde*, 1-2009, http://mappemonde.mgm.fr/num21/lieux/lieux09101.html

Blanc M. (2010). The Impact of Social Mix Policies in France, *Housing Studies*, 25(2), pp. 257-272.

Cao C., Fusco G. (2015). *Representing Uncertain Clustering. The case of Social Specialization on the French Riviera*, https://public.tableau.com/profile/fusco#!/vizhome/RepresentingUncertainClustering/Story

Cedilnik, A., Reinghans, P. (2000). Procedural annotation of uncertain information, In *Proceedings of Visualization '00*, IEEE Computer and Society Press, pp. 77-84.

Centi C. (1993). Les enjeux du modèle niçois. L'approche localiste du développement en question. *Revue Economique*, 44(4), pp. 687-712.

Duncan, O., B. Duncan (1955). A methodological analysis of segregation indexes. *American Sociological Review*, 20(2), pp. 210-217.

Ehlschlaeger C., Shortridge A., Goodchild M. (1997). Visualizing Spatial Data Uncertainty Using Animation, *Computers & Geosciences*, 23(4), pp. 387-395.

Fusco G., Cao C., Dubois D., Prade H., Scarella F., Tettamanzi A. (2016). Social Polarization in the Metropolitan Area of Marseille. Modelling Uncertain Knowledge with Probabilistic and Possibilistic Networks, ECTQG 2015 Proceedings, *Plurimondi*, 8 p.

Fusco, G. et F. Scarella (2011). Métropolisation et ségrégation sociospatiale. Les flux des migrations résidentielles en PACA. *L'Espace Géographique*, 40(4), pp. 319-336.

Korb K., Nicholson A. (2004). *Bayesian Artificial Intelligence*, Chapman & Hall / CRC.

Kunz, M., Regamey-Gret, A., Hurni, L. (2011). Visualization of uncertainty in natural hazards assessment using an interactive cartographic information system, *Natural Hazards*, 59(3), pp. 1735-1751.

Lacour, C. et F. Gaschet (2008). *Métropolisation et ségrégation*, Bordeaux: PUB.

Langseth H., Nielsen T. (2006). Classification using Hierarchical Naïve Bayes Models, *Mach Learn*, 63, pp. 135–159.

Lin J. (1991). Divergence Measures Based on the Shannon Entropy, *IEEE Transactions on Information Theory*, 37(1), pp. 145-151.

MacEachren, A. (1992). Visualizing uncertain information. *Cartographic Perspective*, 13, pp. 10-19.

MacEachren, A. M., D. Howard, et al. (1993). Visualizing the health of Chesapeake Bay: An uncertain endeavor, In *GIS/LIS Proceedings*, vol. 1, American Society of Photogrammetry and Remote Sensing/American Congress on Survey and Mapping, Bethesda MD, pp. 449–458.

Massey D. (1985). Ethnic residential segregation: A theoretical synthesis and empirical review. *Sociology and Social Research*, 69(3), pp. 315-350.

Pallez D., Serrurier M., Da Costa Pereira C., Fusco G., Cao C. (2015). Social Specialization of Space: Clustering Households on the French Riviera, *GECCO Companion 15 - Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*, ACM, New York, pp. 1447-1448.

Ward, M.O. (2002). A taxonomy of glyph placement strategies for multidimensional data Visualization, *Information Visualization*, 1(3/4), pp. 194-210.

Detection of outliers in crowdsourced GPS traces

Stefan S. Ivanović[*1], Ana-Maria Olteanu Raimond[1], Sébastien Mustière[1], Thomas Devogele[2]

[1] Université Paris-Est, IGN, COGIT Lab., France
[2] Université François Rabelais de Tours, Laboratoire d'informatique France

*Corresponding author: stefan.ivanovic@ign.fr

Abstract
Nowadays, crowdsourced GPS data are widely available in a huge amount. A number of people recording them has been increasing gradually, especially during sport and spare time activities. The traces are made openly available and popularized on social networks, blogs, sport and touristic associations' websites. However, their current use is limited to very basic metric analysis like total time of a trace, average speed, average elevation, etc. The main reasons for that are a high variation of spatial quality from a point to a point composing a trace and a need for referential data for evaluation of their quality. In this paper we present a novel approach for filtering and detection of outliers in crowdsourced GPS traces in order to assess their spatial quality intrinsically and make them more suitable for more advanced uses such as updating referential road network of French Mapping Agency – IGN. In addition, we propose a new definition of an outlier in GPS data, adapted to intrinsic assessment of spatial quality.

Keywords
Crowdsourced GPS trace, filtering, outlier detection, machine learning

I INTRODUCTION

Intensive sports activities of professionals and amateurs are very frequently recorded by using GPS devices. The traces obtained are then made openly available to the community through social networks, blogs, sport and touristic communities' web-sites. Currently, they are mostly used for visualization purposes and some basic metric data analysis (e.g. total time of the trace, distance, speed). On the other hand, they have a huge potential to be used for more advanced purposes. (Bergman and Oksanen, 2016). The aim of our research is to use crowdsourced GPS traces as a potential data source for highlighting updates in a referential road network of the French National Mapping Agency (IGN).

In our work, secondary road network in mountainous area such as hiking, bicycle and tractor paths are in our focus. Even if not always necessary, this network is very important for production of touristic maps and for other different applications such as defence and sport activities. These paths are very challenging for continuous update due to their intermittent nature (e.g. they appear and disappear very often) and various landscape (e.g. forest, high mountains, seashore) which make the update process time consuming and the traditionally used methods such as stereo-restitution, insufficient.

The crowdsourced GPS traces are collected without any protocol, with low and heterogeneous frequency sampling and mostly by low class GPS devices. In addition, they are made available with few or inexistent metadata. Moreover, various errors are introduced by different external factors such as topography, canopy, etc. These errors can cause a significant bias and may limit

the usability of the traces in different analysis. Thus, using GPS traces for updating authoritative data makes their quality issues very important. To assess the quality of GPS traces, a first step is to detect and filter outliers. This paper focuses on detection of outliers in crowdsourced GPS traces in order to improve their spatial quality.

II OUTLIERS DETECTION IN GPS DATA

Many studies have analysed factors that influence the quality of GPS data. Environmental factors have been found most influencing. First, topography that reflects on positional error and a number of fixed positions (Lewis et al. 2007; Cain et al. 2005; DeCesare et al. 2005). Second, canopy cover, especially the density of its coverage, type of species (Klimànek 2010; Tucek and Ligos 2002) and height (Janeau et al. 2004) affect positional error. Third, obstacles degrade the quality of GPS signal by causing multipath effect (Tucek and Ligos 2002). Since they all confirmed the influences of the environment on GPS data quality, it is necessary to take these influences into account when dealing with such quality anomalies like outliers are.

We define an outlier as a GPS point whose metrics and geometrics characteristics differ significantly from the characteristics of other points composing a GPS trace. Overall, most of works on outlier's detection in GPS data has been considered GPS measurements' errors as outliers and treated them by various filtering methods (Ordonez 2011; Duran 2012; Knight 2009). In addition, few works defined outliers from geometric point of view, as points that differ from the rest of the traces following the same path (Etienne 2013; Gil de la Vega et al. 2015). In our work, we only consider a single trace which does not require presence of other traces. That is important in situations where only one trace or few traces exist for the same path not allowing to define a pattern, like in some challenging mountainous areas.

Red arrows in Figure 1 point examples of outliers whose positions cause values of speed, distance and angle between them and their consecutive points significantly different compared to other points of the same trace.

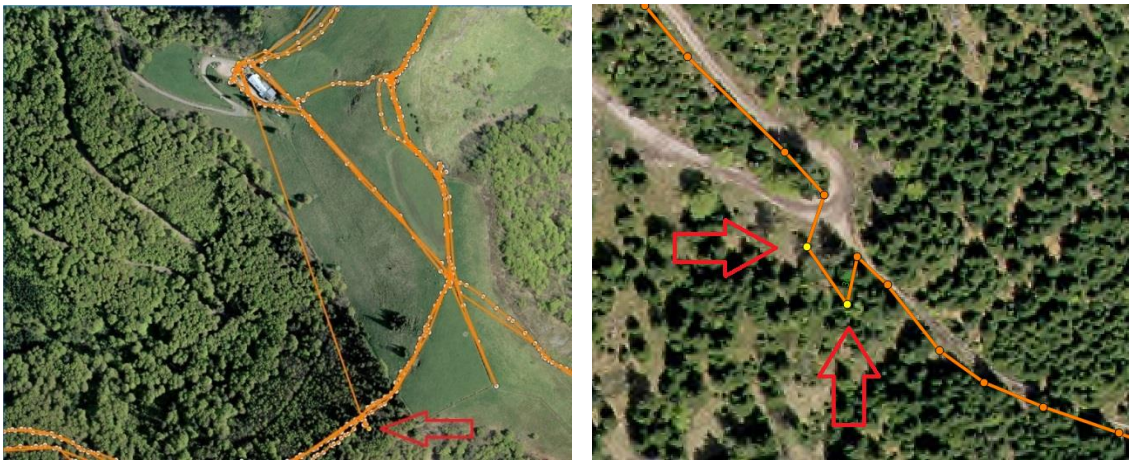


Figure 1: Examples of outlier points.

III TEST DATA

Crowdsourced GPS data recorded during sport activities such as running, cycling, walking and hiking in Vosges Mountains are collected from different websites and portals of French sport associations. We chose this area due to various landscapes with a plenty of secondary roads, as well as an important number of crowdsourced GPS traces.

In total, we selected 436 traces composed by 292337 points. GPS points are theoretically described by their position (WGS84), elevation and timestamps. Practically, some points have missing attributes, that is, missing elevation or timestamp or even both. Six indicators measuring the completeness of the attributes are computed. We found that 106206 points (36%) lacked timestamp, whereas the situation with elevation was far better – only 6580 points (2%) lacked elevation. Regarding the traces, we noticed that 157 (36%) had no timestamps at all, whereas 287 (66%) had at least one missing timestamp.

IV METHODOLOGY AND RESULTS

Due to the heterogeneous nature of crowdsourced GPS data and more random than systematic influences of environment and other known factors, the detection of outliers is a complex task. Many intrinsic and extrinsic indicators may be computed to detect outliers. However, determining relevant criteria and thresholds as well as how to combine them is also a challenging task. Thus, to manage these difficulties, we propose an approach based on supervised machine learning techniques in order to generate generic rules and thresholds.

Our approach is composed by four steps: i) the first one consists in filtering noise such as redundant points, negatives values, etc.; ii) the second is the computation of different intrinsic and extrinsic indicators; iii) the third one consists in using machine learning, which involves a supervised sampling followed by applying a classification algorithm; iv) finally, the generated rules and thresholds are applied on non-classified points of the test area.

First, within the step of noise filtering, redundant points (e.g. overlapping consecutive points) are filtered by means of zero distance values in a small time window. Negative speed values are used to detect structural errors in GPX files such as two traces irregularly merged into one or errors of GPS clock. This is an important step since it was spotted that in some cases negative speed values caused geometric anomalies of the traces. In total 4594 points (2%) were filtered thanks to this step.

To formalize outliers' detection, we propose intrinsic metrics and extrinsic indicators. The former are calculated from GPS measures only, such as distance, speed, direction and elevation between consecutive points. The later are calculated based on the analysis of the spatial context in which GPS points are recorded such as type and density of forest, slope and its curvature, proximity of obstacles (e.g. cliffs, buildings, forest) and other features (e.g. river, lake, building).

In total, 15 different indicators were computed. One representing direction change, three based on distance and speed respectively, taking into account their mean and variations between consecutive points, four derived from elevation (GPS and corresponding DTM) and four based on spatial context such as: proximity of obstacles (e.g. buffer of cliffs, buildings, forests), point in lake/river/building, point in the forest, type of forest. It is important to stress that due to missing attributes (e.g. lack of timestamps) for some points, some measures cannot be calculated (e.g. speed). The indicators are calculated for each point taking into account previous and next point in the trace.

In order to apply the machine learning techniques, first a sampling zone was chosen so that it represented faithfully the state and heterogeneities of the entire test area and the entire pattern of GPS points. Percentage of missing attributes (timestamps and elevation) of sampled points compared to the total percentage of missing attributes differed for only 2%. In total 2342 points were sampled and manually classified as outliers (3%) or regular points (97%). The learning process was conducted in WEKA software package using JRIP algorithm proposed by Cohen (1995). Evaluation of results by means of 30 fold cross validation is presented in a confusion matrix in Figure 2.

	Predicted		
Actual class		Outlier	Not an outlier
	Outlier	61	16
	Not an outlier	16	2249

Figure 2: Confusion matrix of 30 fold cross validation

The number of correctly classified outliers is almost four times bigger than the number of misclassified, while overestimation and underestimation are balanced. In addition, a global precision of the approach calculated from the matrix is 98%. Both results confirm high performance of the approach and generated rules, particularly taking into account presence and random distribution of missing attributes.

In total, five rules as well as thresholds were generated and presented in Figure 3.

1. AngleMean>=87°.54 and DistDiffMed>=1.05 => outlier
2. AngleMean>=71°.23 and SpeedRate>=1.5 => outlier
3. AngleMean>=74°.80 and DistDiffN<=0.21 => outlier
4. AngleMean>=83°.15 and SpeedRate <=0.85 => outlier
5. AngleMean>=56°.43 and DistMean>=8847.31m => outlier

Figure 3: Rules generated in machine learning

Where: i) AngleMean represents an average value of 3 direction changes; ii) DistDiffMed a relation between two consecutive distances (before and after a point that is evaluated) and a median distance of a trace; iii) DistDiffN a normalized value of two consecutive distances; iv) DistMean defines a mean distance of two consecutive distances and v) SpeedRate represents the velocity change rate being proposed by Winden et al (2016).

Finally, the generated rules were applied on the entire test area. As a result, 9309 points (3%) were recognized as outliers. In Figure 4, an example of a successfully detected outlier is illustrated.

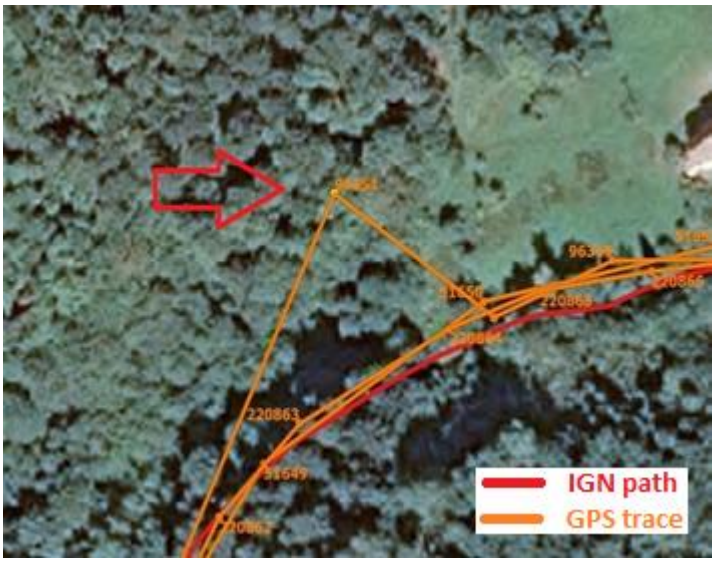


Figure 4: An outlier detected within non-classified points

Although the successfulness of the approach is high, there is a specific situation where it does not perform properly causing overestimations. This happens when treating traces with very high sinuosity and low spatial resolution such as illustrated in Figure 5.

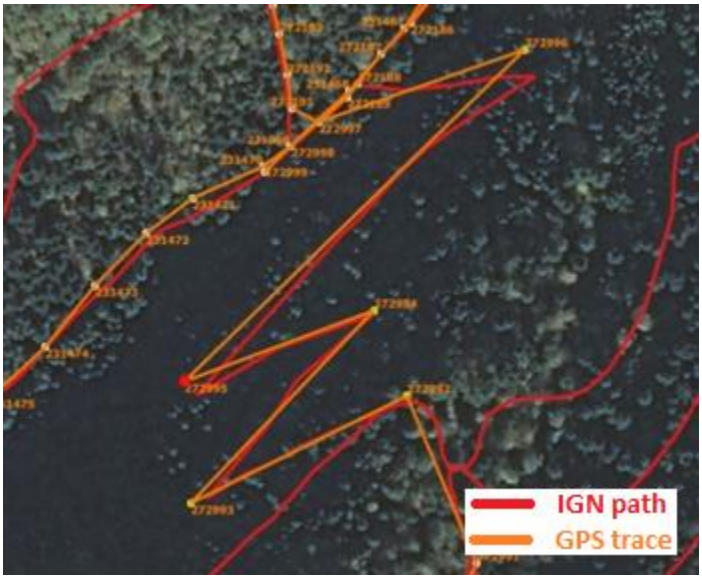


Figure 5: Misclassified outliers

All five points composing a part of the trace that fits referential road are identified as outliers. Such geometries of referential roads are not frequent (in our test area only 11), thus the overestimation produced are not numerous. However this should be treated by taking into account the sinuosity of referential roads.

Concerning the influence of external (environmental) factors, it is obvious that based on the generated rules, the impact and importance of the external factors on appearance of outliers were not discovered. It is likely that a link between them exists, however it is difficult to be detected and modelled in crowdsourced traces due to their extreme heterogeneity, and a lack of two very important information such as precision of GPS device and the quality of signal.

For example, the same obstacle produces a multipath and subsequently outlier while collecting data by low precision GPS, whereas it is not a case for high precision GPS under the same conditions. The same situation is with canopy cover. More precise device using better quality signal can produce a trace without outliers even in closed coniferous forest, while lower precision device supplied with low quality signal would cause outliers even in an open area.

V CONCLUSION

GPS traces are widely spread as crowdsourced data due to the significant number of sportsmen's and amateurs recording them during sport or leisure activities. Our research is aiming to use them for highlighting updates in referential secondary road network. To do that, it is necessary to assess the traces quality, particularly to deal with significant anomalies and of their geometry that we defined as outliers. In this paper, we presented a novel approach for formalization and detection of the outliers despite the high heterogeneity of crowdourced GPS traces and dominantly random influences of external factors on their quality. The outliers were modelled and detected using 15 intrinsic metric and spatial context indicators calculated for each point in a trace. The successfulness of the approach is generally high, except in the case where a sinuosity of a referential road is very high. By analysing generated rules and the nature of crowsourced data, we conclude that causes of the outliers are obviously highly heterogeneous. Thus, the most efficient way to model and detect them is by means of their intrinsic indicators, particularly taking into account that links between spatial context and crowdsouced traces' quality proved to be weak.

References

Bergman C., Oksanen J. (2016). Conflation of OpenStreetMap and Mobile Sports Tracking Data for Automatic Bicycle Routing. *Transactions in GIS* 00(00): 00-00

Cain III J.W., Krausman P.R., Jansen B.D., and Morgart J.R. (2005). Influence of topography and GPS fix interval on GPS collar performance. *Wildlife Society Bulletin 33*(3), 926-934

Cohen, W. W. (1995). Fast Effective Rule Induction. In *Twelfth International Conference on Machine Learning*, Tahoe City, California

DeCesare N.J., Squires J.R., and Kolbe Y.A. Effect of forest canopy on GPS-based movement data. *Wildlife Society Bulletin 33*(3), 935-941

Duran, A., Earleywine, M. (2012). GPS Data Filtration Method for Drive Cycle Analysis Application. In *SAE 2012 World Congress*, Detroit, Michigan

Etienne, L. (2011). Motifs spatio-temporels de trajectoires d'objets mobiles, de l'extraction àla détection de comportements inhabituels - Application au trafic maritime. PhD thesis. Institut de Recherche de l'Ecole Navale.

Gil de la Vega, P., Ariza-Lopez, F.J., Mozas-Calvache, A.,T. (2015). Detection of outliers in sets of GNSS tracks from volunteered geographic information. In *Agile Conference 2015*, Lisbon

Janeau G., Adrados C., Joachim J., Gendner J.P., Pépin D. 2005. Performance of differential GPS collars in temperate mountain forest. *C. R. Biologies 327*, 1143–1149

Klimànek M. 2010. Analysis of the accuracy of GPS Trimble JUNO ST measurement in the conditions of forest canopy. *Journal of Forest Science 56*(2), 84–91

Knight, N. L., Wang, J. (2009). A comparison of outlier detection procedures and robust estimation methods in GPS positioning. *J. Geodesy 62*(4), 699-709

Lewis J.S., Rachlow J.L., Garton E.O. and Vierling L.A. (2007). Effects of habitat on GPS collar performance: using data screening to reduce location error. *Journal of Applied Ecology 44*, 663–671

Ordoñez, C., Martínez, J., Rodríguez-Pérez, J., and Reyes, A. (2011). Detection of Outliers in GPS Measurements by Using Functional-Data Analysis. *Journal of Surveying Engineering137*(4), 150-155

Tuček, J. Ligoš J. 2002. Forest canopy influence on the precision of location with GPS receivers. *Journal of Forest Science 48*(9), 399–407

Van Winden, K., Biljecki, F. and van der Spek, S. (2016). Automatic Update of Road Attributes by Mining GPS Tracks. *Transactions in GIS* 00(00): 00-00

Locational Error Impacts on Local Spatial Autocorrelation Indices: A Syracuse Soil Sample Pb-level Data Case Study

Daniel A. Griffith*[1], Yongwan Chun[1], and Monghyeon Lee[1]

[1] University of Texas at Dallas, USA

*Corresponding author: dagriffith@utdallas.edu

Abstract

This paper focuses on propagation of location errors in spatial data analysis. Specifically, it investigates how location errors impacts local spatial autocorrelation indices that often are used to identify spatial clusters. Results of a simulation experiment using heavy metal soil sample points in Syracuse, NY, are summarized. In the simulation experiment, artificial location errors were introduced to perturb points, and then local Moran's *I* and Getis-Ord local statistics were calculated. The results show that location errors have an impact on the identification of spatial clusters. Some significant spatial clusters with no location error become insignificant ones with location errors, and some insignificant ones with no location error become significant ones with location errors. More severe deviations from the true results are observed with larger location error, as expected.

Keywords

Location error, local spatial autocorrelation, heavy metal soil sample, uncertainty

I INTRODUCTION

Location error occurs when the spatial information of a geotagged observation deviates from its true locational information. If spatial data have locational error, the error may increase any uncertainty associated with a spatial data analysis. Although uncertainties may render only slightly incorrect modelling results, they also can be completely fatal to an analysis of georeferenced data, and undermine the outcome of the spatial data analysis (Fisher, 1999). Any spatial model output may have incorrect results that have been corrupted by uncertainty propagated to the output. Local indicators of spatial association (LISA; Anselin, 1995) and Getis-Ord statistics (G_i^*; Ord and Getis, 1995) are well-known indices measuring local spatial autocorrelation and identifying spatial clusters. This paper summarizes simulation experiment results demonstrating how location error affects these spatial autocorrelation indices.

II DATA

Simulation experiments furnishing the basis of the analysis summarized in this paper utilized soil samples collected from across the City of Syracuse, NY over three years, mainly during the summers of 2003 and 2004. These samples were used to measure a suite of six heavy metals (Fe, Pb, Rb, Sr, Zn, and Zr), assayed in a chemistry laboratory using a NITON XL-700-series X-ray fluorescence (XRF) instrument (Griffith, 2008). These samples were assayed based on a 120-s testing time and NIST 2,711 standard reference materials (SRM); measurements are in milligrams of heavy metal per kilogram of soil, or ppm (Griffith, 2008). We focus on Pb because lead poisoning is one of the critical issues in public health today, and soil Pb levels tend to be strongly related to human Pb poisoning. The total number of collected

soil sample points is 3,574. We exclude five points that are outside of the city boundary, and averaged 284 duplicate sample assays by location. Consequently, this study utilized 3,290 distinct soil sample points (see Figure 1). The Pb levels were subjected to a logarithm transformation—$\mathrm{LN}(\mathrm{Pb}/s_{\mathrm{Pb}} - 12)$—so that their frequency distribution better conforms to a bell-shaped curve (Griffith, 2008).

The simulation experiments involved the following steps: (1) randomly adding location error to a subset of the soil sample location points, (2) aggregating the transformed Pb measures by census geography (i.e., census tracts and census block groups), and (3) calculating local Moran's *I* and G_i^* statistics. The City of Syracuse has 57 census tracts and 147 block groups in its 2000 census geography maps. One census block group (ID: 29002) has been merged with its neighborhood block group (ID: 29001), which is in the same census tract as it, because this census block group does not contain a soil sample point.

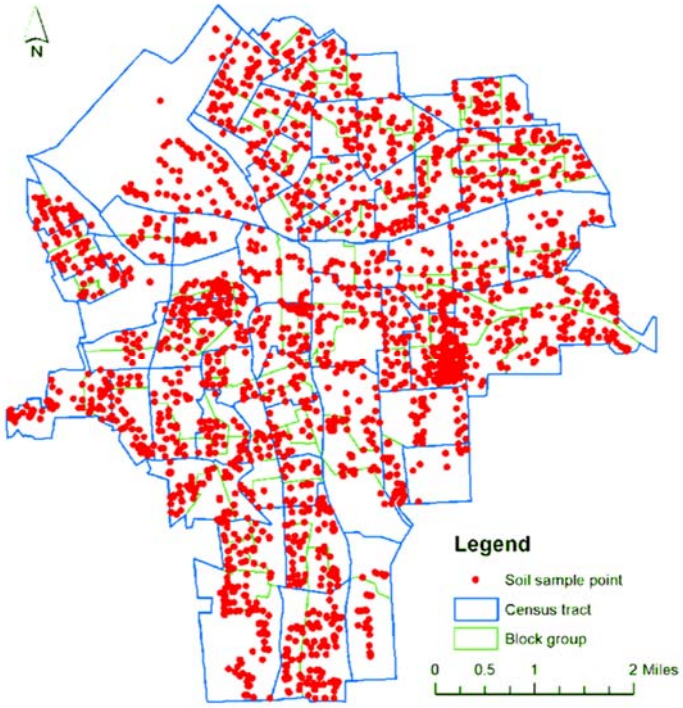


Figure 1: Soil sample points, census tracts, and block group in Syracuse, NY.

III METHOD

After extensive data cleaning, we artificially and randomly added locational deviations to the soil sample points in the simulation experiments. The simulation experimental design may be described as follows:

1) Randomly sample a selected percentage (beginning with 10%) of the 3,290 soil sample points;
2) Assign a specified amount of location error (beginning with 10m) with a random direction to each sampled point, constraining the perturbed locations to remain within the city limits;

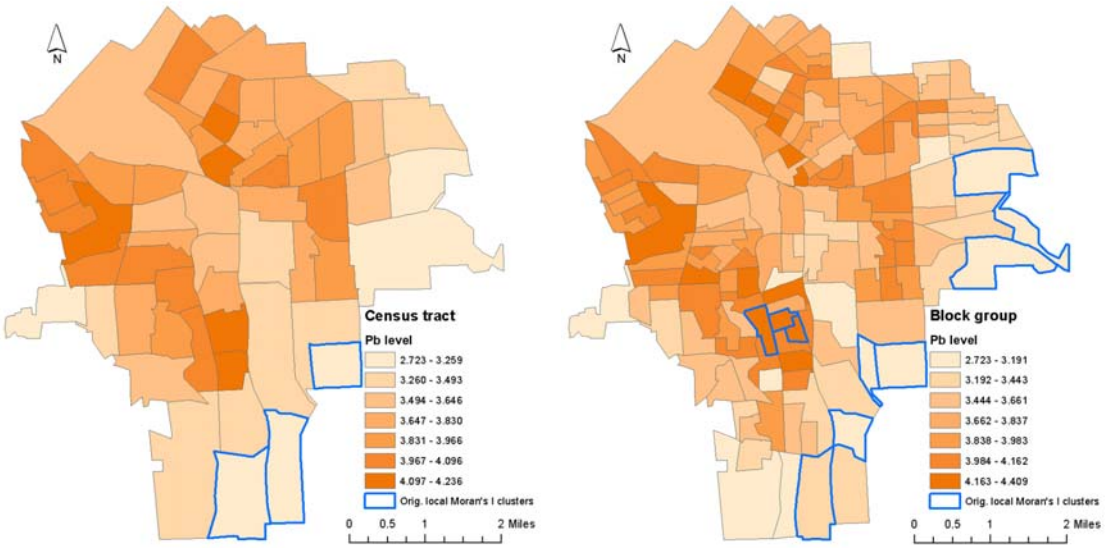
3) Aggregate transformed Pb measures in each census geography administrative unit (i.e., census tract or census block group), and then calculate local Moran's *I* and G_i^* statistics, recording clusters (e.g., hotspots and coldspots) for each simulation replicate;
4) Repeat Steps 1) to 3) 1,000 times (the number of replicates);
5) Repeat Steps 2) to Step 4) for 25m, then 50m, then 75m, and finally 100m of location error; and,
6) Repeat Steps 1) to 4) for 20%, then 30%, then 40%, and finally 50% of the soil sample points.

This simulation experimental design generates 25 different sample sizes (five percentages by five levels of location error), each with 1,000 replicates (to exploit the Law of Large Numbers).

IV RESRULTS

This section summarizes aspects of the two extreme simulation cases: results for the minimum (a 10% sample size, and 10m of added location error) and maximum (a 50% sample size, and 100m of added location error) error levels. Intermediate error level results are between these minimum and maximum results. Statistical significant levels for evaluating the local Moran's *I* have been adjusted using a Bonferroni correction based upon an effective sample size that adjusts for latent spatial autocorrelation (Chun and Griffith, 2013).

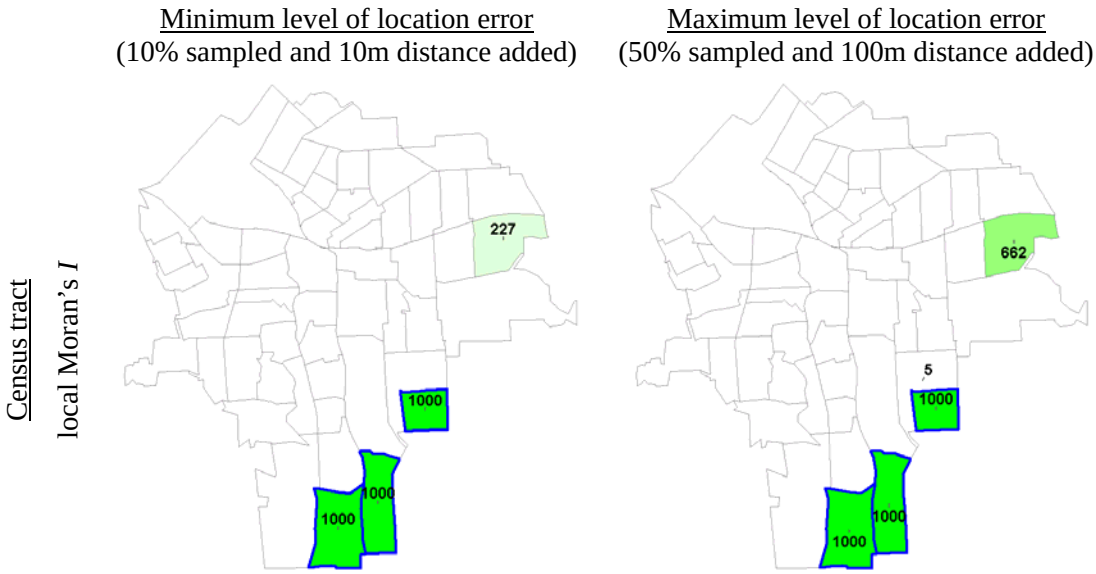
Figure 2 represents the original aggregated Pb level maps for 2000 census tracts and census block groups. The eastern and southern areas of Syracuse have relatively low Pb levels, whereas northwest and downtown areas of the City have relatively higher Pb levels.

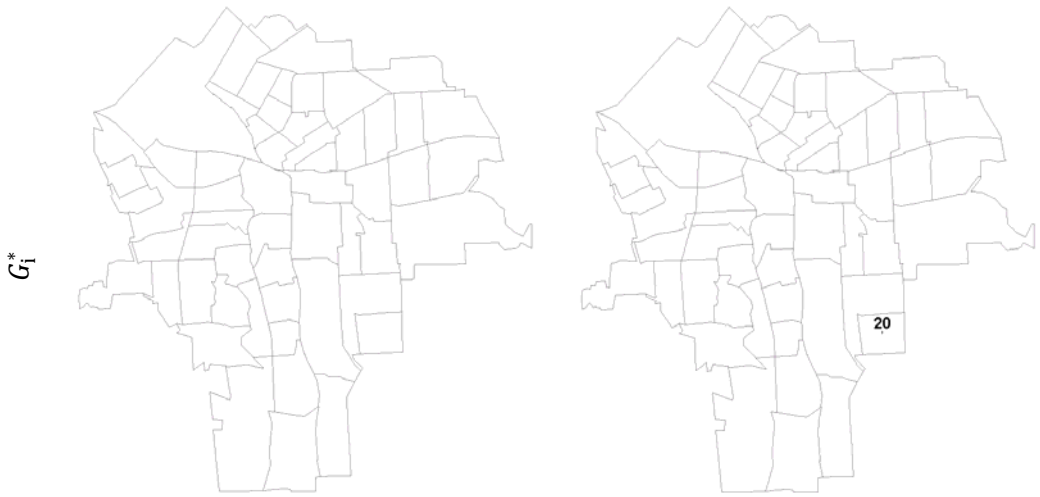


(a) (b)

Figure 2: Pb level and its local Moran's *I* cluster maps for 2000 census tracts (a), and census block groups (b) in Syracuse, NY (Note: no G_i clusters exist).

In Figures 3 and 4, red and blue boundary regions represent original (i.e., true) statistically significant positive High-High (HH; a high local Moran's *I* value region is surrounded by high local Moran's *I* value regions) and Low-Low (LL; a low local Moran's *I* value region is surrounded by low local Moran's *I* value regions) local Moran's *I* geographic clusters. Meanwhile, the G_i^* maps reveal no original geographic clusters. Polygons filled with green, rather than white, are geographic clusters; the numbers that overlay them represent their aggregate detection frequencies as significant geographic clusters across the simulation experiments. For example, if a region has a red boundary, is filled with dark green, and 1,000 overlays it, this region is an original HH cluster identified in every one of the 1,000 simulation replications; in other words, location error does not affect its detection. A blue boundary region filled with light green and overlaid with 200 is an original LL cluster identified in only 200 of the 1,000 simulation replications; location error obscures or hides this geographic cluster in eighty percent of the simulation replicates. Another possible scenario is that a region has a grey boundary, which means this region is not an original cluster, but filled with green and overlaid with a number; this is a false cluster that location error creates in a certain percentage of the simulation replicates. Finally, many regions have a grey boundary and are filled with white; these are neither original nor location error created clusters.





(c) (d)

Figure 3: Local Moran's *I* and G_i^* location error simulation results for census tracts.

Figure 3 portrays census tract simulation results. When the minimum level of location error is introduced (see Figure 3a), three original LL local Moran's *I* cluster regions are colored with dark green and overlaid with 1,000. However, one non-original cluster region has been detected 227 times out of 1,000 replicates. In this latter case, location error affects the final outcome. When we introduce the maximum level of location error (see Figure 3b), even though the three original cluster census tracts are not affected, one more non-original cluster region emerges five times as a cluster. Furthermore, the census tract that has been identified as a cluster 227 times with the minimum level of location error now has been detected 662 times with the maximum level of location error. One implication here is that as the amount of location error increases, uncertainty propagation also increases and impacts spatial analyses and output. Statistically significant original hotspots or coldspots in terms of G_i^* are not identified at the census tract geographic resolution. However, G_i^* identifies a census tract as a significant coldspot 20 times in the 1,000 simulation replicates (see Figure 3d).

Minimum level of location error	Maximum level of location error
(10% sampled and 10m distance added)	(50% sampled and 100m distance added)

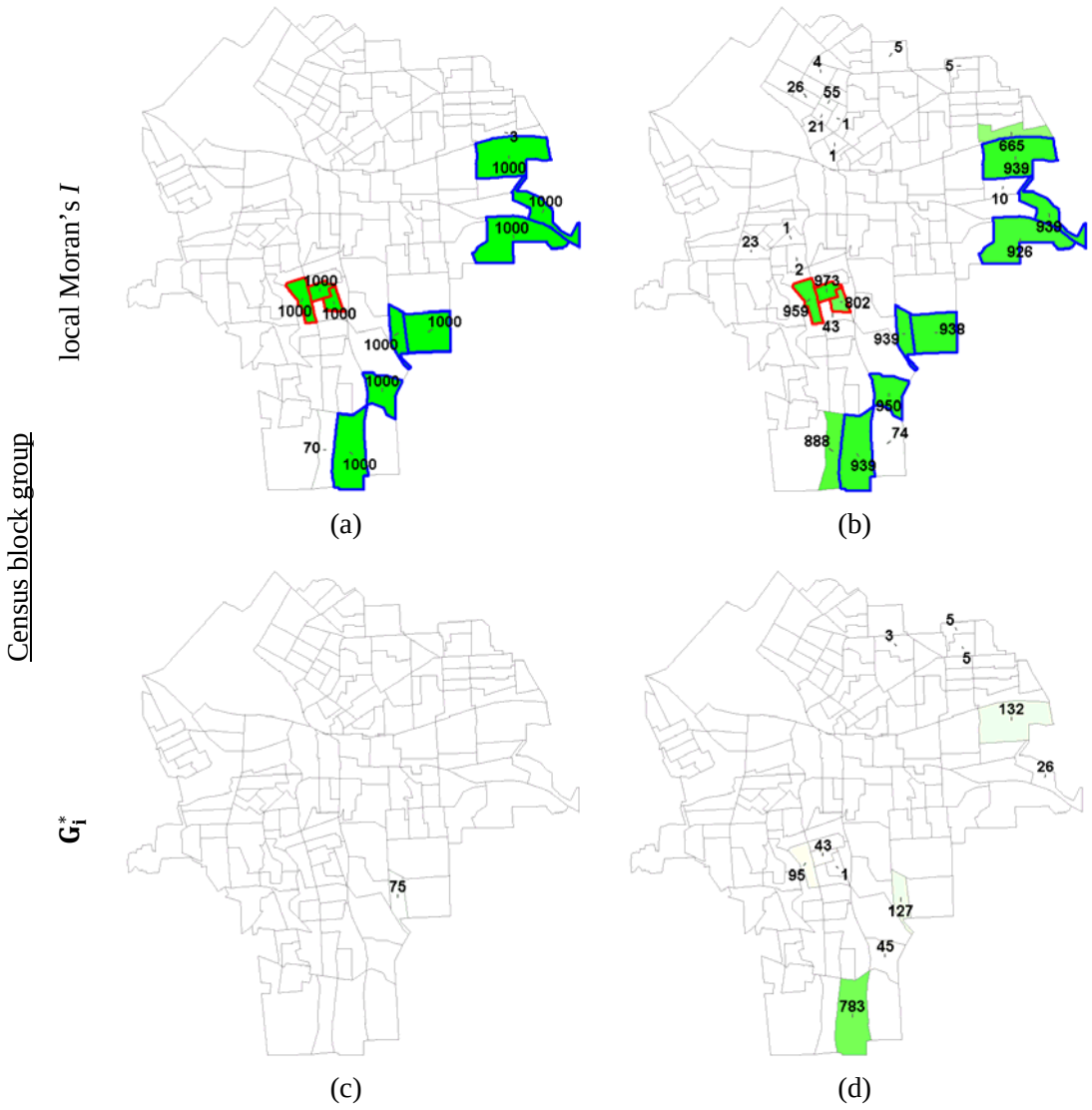


Figure 4: Local Moran's *I* and G_i^* location error simulation results for census block groups.

Figure 4 portrays the census block group simulation results. Significant G_i^* cluster regions do not exist at this geographic resolution, either. However, the local Moran's *I* simulation results clearly show location uncertainty propagation. In Figures 4a and 4b, the City has 10 significant original local Moran's *I* cluster regions: three of them are HH cluster regions, and seven of them are LL cluster regions. With the minimum level of location error, all of the original clusters are identified in every one of the 1,000 simulation replicates. In other words, the location error does not affect identification of the original clusters. However, two non-original clusters have been detected three and 70 times in the 1,000 replicates. They emerge because of location error. Uncertainty propagation tends to increase as the level of location error increases. With the largest level of location error, none of the original cluster regions is identified as a meaningful cluster in all 1,000 replicates. Nevertheless, detection of these clusters still occurs over 900 times; location error impacts cluster detection in this case. In terms of false positives, 16 census block groups are detected as significant geographic clusters even though they are not original clusters. In addition, the census block groups that have been

detected as significant cluster regions with the minimum level of location error are identified only 665 and 888 times out of 1,000 replicates. These findings imply that location error can generate severely biased outcomes. Similarly, G_i^* results also have been affected by location error. One non-original cluster census block group has been detected 75 times with the minimum level of location error, whereas 11 originally nonsignificant census block groups have become significant geographic clusters when introduction of the maximum level of location error.

Table 1 summarizes the number of detected significant geographic clusters in the City, both the original ones and the location error created ones. This table includes both false positives and false negatives: location error creating geographic clusters that are not true clusters; and, location error obscuring or hiding geographic clusters that are true cluster. All red numbers in Table 1 denote that location error affects local spatial autocorrelation cluster detection results. Because the census block group geographic resolution tends to have units smaller than census tracts, in terms of area, location error impacts tend to be more severe at this geographic resolution. Census block groups have more soil sample points that potentially can be perturbed across unit boundaries with a certain magnitude of location error.

		Original data		**Location error added to data (1,000 replicates)**			
				Minimum (10%; 10m)		Maximum (50%; 100m)	
				Cluster	Not cluster	Cluster	Not cluster
Census tract	local Moran's *I*	Cluster	3	3	0	3	0
		Not cluster	54	1	53	2	52
	$\mathbf{G_i^*}$	Cluster	0	0	0	0	0
		Not cluster	57	0	57	1	56
Block group	local Moran's *I*	Cluster	10	10	0	0	10
		Not cluster	136	2	134	16	120
	$\mathbf{G_i^*}$	Cluster	0	0	0	0	0
		Not cluster	146	1	145	11	135

Table 1. A cross-tabulation of the number of significant soil Pb cluster regions in Syracuse, NY.

V FINDINGS

A majority of areal units remain unchanged in terms of geographic cluster detection in the presence of location error. Nevertheless, location error does tend to change a considerable number of true geographic clusters, and it also artificially creates geographic clusters that actually do not exist. The principal implication is that location error can propagate to the final output of a spatial analysis, compromising its quality and precision. Furthermore, location error seems to affected local Moran's *I* results more than G_i^* results.

VI ACKNOWLEDGEMENTS

This research was supported by the National Institutes of Health, grant 1R01HD076020-01A1; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Institutes of Health.

References

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis* 27, 93–115.

Chun, Y., Griffith, D.A. (2013). *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. Thousand Oaks: SAGE Publications.

Fisher, P.F. (1999). Models of uncertainty in spatial data. *Geographical information systems* 1, 191–205.

Griffith, D.A. (2008). Geographic sampling of urban soils for contaminant mapping: How many samples and from where. *Environmental Geochemistry and Health* 30, 495–509.

Ord, J.K., Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27, 286–306.

Modeling spatial risk of the Foot-Mouth-Disease epidemic in South Korea

EunHye Yoo[1] **and JiYoung Lee**[2]

[1]Department of Geography, University of Buffalo (SUNY), USA
[2]Department of Geoinformatics, Seoul University, S. Korea

*Corresponding author: eunhye@buffalo.edu

Abstract: The 2010/2011 food and mouth (FMD) disease epidemic in South Korea spread nationwide and resulted in substantial economical damage. Its rapid transmission throughout the country despite the government's control policy argues for the need to improve our understanding of its spatial dimensions. The spatial point patterns of FMD incidences revealed that FMD incidences formed statistically significant spatial clusters during the early phase of epidemic but its significance and the spatial dimensions dynamically evolved over the course of epidemic. We developed a log-Gaussian Cox process (LGCP) model to quantify and to make inferences on the spatial distribution of FMD risk, while accounting for both known risk factors and unexplained variation. The FMD risk surface estimated from the LGCP model captured the spatial heterogeneity of FMD risk driven by environmental and landscape factors and unknown risk variations at both local and regional scales. This study demonstrated that point pattern analyses and statistical models enabled investigators to improve our understanding on the spatial dynamics of FMD, which may provide useful information for decision makers to make efficient and effective control strategies for future outbreaks.

Keywords: foot and mouth disease, spatial point pattern analysis, log-Gaussian Cox process

I INTRODUCTION

Foot and mouth disease (FMD) is considered by far the most serious infection between livestocks due to its rapid transmission (Keeling, 2005). In South Korea (S. Korea) the outbreaks of FMD during November 2010 and April 2011 spread nationwide and resulted in a total of 3.48 millions susceptible animals being culled (Park et al., 2013). The spatial pattern analysis of infectious diseases may reveal the shape of disease and guide how to control their future spread. However, most studies on the recent FMD incidences in S. Korea have overlooked the potential association between the observed FMD cases and spatial risk factors, and merely focused on descriptive statistics (Moutou and Durand, 1994; Alexandersen et al., 2003). In this study, we aim to fill the gap of FMD literature by examining the spatial patterns of infection risk and their dynamic changes. We will make an inference on the biological processes underlying the observed FMD cases using a log-Gaussian Cox process (LGCP) model.

II MATERIAL AND METHODS

2.1 Data

The official public health records of the FMD epidemic in S. Korea were used, which include the location of farms on which FMD infection was confirmed, the reported date of infection, the type of livestock infected. The types of livestocks infected include cattle, pig, deer, and goat, but we focused on cattle and pigs in this paper. Assuming that the spatial variation of FMD prevalence is affected by both the environmental factors and the spatial configurations of

livestock farms, we incorporated 2010 national census data on the number of domestic animals and farms. Figure 1(A) shows the elevation, (B) and (C) are the spatial distribution of livestock density and the average temperature of December of 2010 and January of 2011. The driving distance between each grid cell at the same resolution to the index farm in Andong is illustrated in Figure 1(D). The index farm was considered as an initial seed of infection because there was hardly any FMD cases reported in S. Korea prior to this epidemic in 2010 (Park et al., 2013).

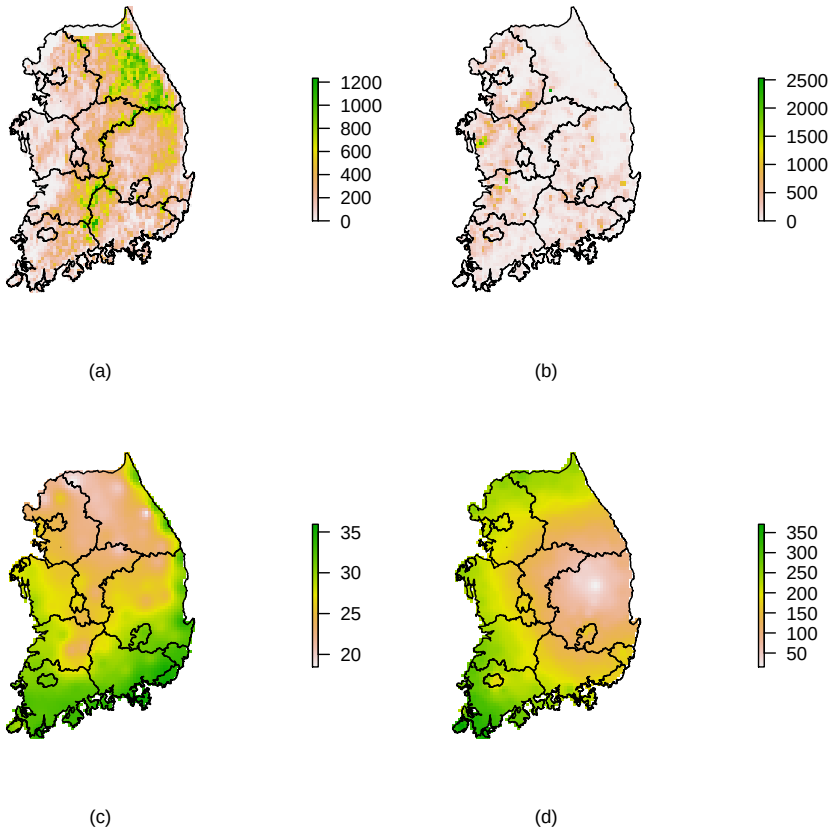


Figure 1: Spatial distributions of risk factors; (a) Elevation (in meter), (b) Density of livestock farm holdings per non-overlapping 5×5 km^2 areas, (c) Average temperature in winter months (in Fahrenheit), (d) Driving distance from the point source of infection (in kilometer).

2.2 Spatial point pattern analysis

The K function describes the extent to which there is spatial dependence in the arrangement of georeferenced point locations of events in the study area (Gatrell et al., 1996). This function can be estimated from an observed event distribution and its behavior in a particular situation is assessed. One of the standard reference distribution is homogeneous poisson model. Given that FMD is infectious disease which forms a spatial cluster, homogeneous model is least sensible to the observed FMD cases. As an alternative, we estimated the inhomogeneous K-function $\hat{K}^{iso}_{inhom}(r)$ (Baddeley et al., 2000) where the spatially varying intensity is estimated from the observed FMD cases. To make an inference about observed point patterns under the consideration of non-stationarity, we simulated envelops with the inhomogeneous K function. An observed process is assumed to be different than the null process model, if the estimated K function at a specific distance lie outside the envelop (Moller and Waagepetersen, 2003).

2.3 A log-Gaussian Cox Process model

A special case of Cox procesess where log-intensity is a Gaussian process is a log-Gaussian Cox Process (LGCP) model (Møller et al., 1998; Diggle, 2013). We modeled the spread of FMD in S. Korea as a poisson process that was partially driven by observed risk factors. This conceptualization enabled us to link the observed global heterogeneity in the dispersion of FMD to the spatial patterns of environmental factors. On the other hand, we viewed the unexplained spatial variability of the point process as outcomes of infectious cases transmitting the disease to nearby susceptible. Assuming that FMD incidence locations $\{\mathbf{s}_i, i = 1, \ldots, n\}$ are independently distributed conditional on a Gaussian spatial random field $U(\mathbf{s})$, the LGCP model for FMD is written as

$$\begin{aligned} \{\mathbf{s}_i; i = 1, \ldots, n\}|\mathbf{U}(\mathbf{s}) &\sim Poisson(\Lambda(\mathbf{s})) && (1)\\ \log \Lambda(\mathbf{s}) &= \mu + \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + U(\mathbf{s}) && (2)\\ Cov\{U(\mathbf{s}), U(\mathbf{s}')\} &= \sigma^2 \rho[|\mathbf{s} - \mathbf{s}'|/\theta] && (3) \end{aligned}$$

the vector of covariates $\mathbf{X}(\mathbf{s}) = [X_1(\mathbf{s}), \ldots, X_4(\mathbf{s})]$ consists the elevation X_1, the density of livestock holdings X_2, average temperature of winter months X_3 at the event location $\mathbf{s}$, and X_4 denotes a driving distance from the index farm. We included X_4 as a risk factor for FMD risk, because the initial FMD case on November 28th, 2010 served as a source of infection in subsequent cases and the risk of infection decreases as susceptible is further away. Based on related studies (Ferguson et al., 2001; Grubman and Baxt, 2004), which showed that cold temperatures aided the virus to persist outside the host for a longer time period, we incorporated the spatial variability of winter temperature as one of known risk factors in our LGCP model. The unexplained spatial variation $U(\mathbf{s})$ in FMD risk is a stochastic component (the latent Gaussian process) whose spatial structure was modeled as a function of the separation vector between any two incidence locations with a correlation function specified with two parameters — a range θ and a sill σ^2. In geostatistical jargon, a range refers to the maximum distance at which spatial dependence exists and a sill is a maximum variance of the process.

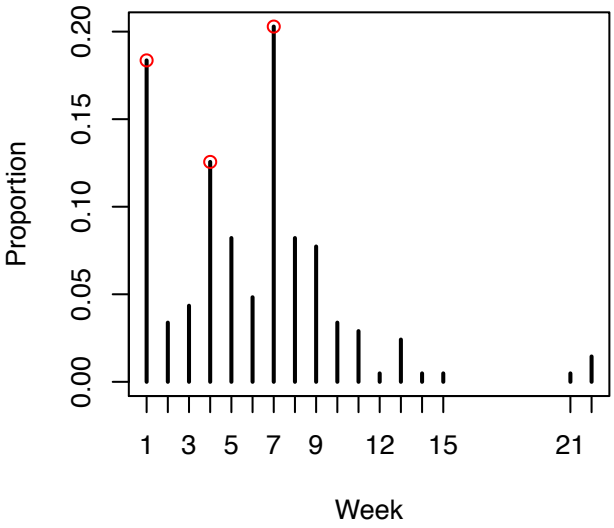
The target of inference of the LGCP model in Equation 2 and 3 includes $\{\beta, \theta, \sigma, U(\mathbf{s})\}$. Møller et al. (1998) originally used the method of minimum contrast estimation, which is often computationally formidable (Diggle, 2013). We addressed this computational challenge by using two approximations: the Gaussian Markov Random Field (GMRF) approximation to the Matern correlation (Banerjee et al., 2004) and integrated nested Laplace approximation (INLA) for the marginal posterior distributions. More specifically, we used the *lgcp* function implemented in *geostatsp* library (Brown, 2015) where a computational grid that consisted of cells 5×5 km^2 in dimensions is superimposed over the study region A. The FMD risk surface $\Lambda(\cdot)$ was defined over a collection of cells $g_l, k = 1, \ldots, L$ that form a disjoint partition of A. Let $Y(g_l)$ be the case count within the lth cell, where the count follows a Poisson distribution with a FMD risk Λ_l. The risk is defined as product of an offset δ_l and the intensity function (or spatial random function) Λ_l, which is a function of known risk factors and a latent Gaussian process as

$$\begin{aligned} Y_l &\sim Poiss(\delta_l \Lambda_l) && (4)\\ \log(\Lambda_l) &= \mu + X_{1l}\beta_1 + X_{2l}\beta_2 + X_{1l}\beta_3 + X_{2l}\beta_4 + U_l && (5) \end{aligned}$$

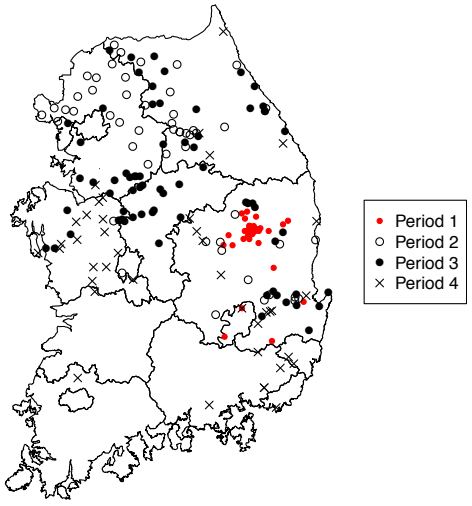
The spatial random effects $U(\mathbf{s})$ is approximated by U_l using GMRF where $\mathbf{s} \in g_l$ with Gaussian priors with a zero mean, $\theta \in [5, 20]$ km for range of the spatial random effects, and $\sigma \in [1, 4]$ for its standard deviation.

III RESULTS

The FMD epidemic in 2010/2011 has shown that more than 90% of FMD cases occurred at five provinces throughout the epidemic and over 60% of the total number of livestock farms in S. Korea reside within these five provinces. We further examined their relationships in the LGCP model by taking the total number of the farms in each town and treated them as population at risk. The observed FMD epidemic curve in Figure 2(a) showed that the major peaks reached on the week 4 and week 7.



(a) Time-FMD case distribution over the course of epidemics. Three peaks of the epidemic was denoted by the symbol of circle at week 1, 4, and 7.



(b) FMD incidences over four time periods.

Figure 2: Temporal and spatial profile of FMD incidences in S. Korea

We defined four time periods of the FMD epidemic to represent the natural phases of the epidemic: weeks 1-2 were the period of rapid spread followed by a sharp decrease (Period 1); in weeks 3-5 there was a medium- to long-distance spread combined with localized transmission from newly affected farms (Period 2); in weeks 6-8 the epidemic peaks (Period 3); in weeks 9-22 eradication was achieved (Period 4). The locations of FMD incidences are shown in the map of Figure 2(b) with different symbols per period. We characterized their spatial patterns using the inhomogeneous K-function, and the results are summarized in Figure 3. Spatial clustering is commonly found during the first three periods, although the spatial extents at which the clustering is observed are different with varying statistical significances. K-function for Period 1 in Figure 3(A), for example, shows statistically significant clustering across all distances, but the spatial cluster exists up to 20 km in Period 2. The K-function estimates for Period 3 (Figure 3(C)) shows statistically significant spatial clustering at medium- to long-distance. Lastly, the K-function in the last period in Figure 3(D) showed regularity without statistical significance.

Bayesian inference on the LGCP model parameters were conducted using the R-INLA package in the statistical computing language R (R Core Team, 2015). The posterior distribution of spatial intensity of FMD and spatial random effects are shown in Figures 4: The lower and upper bounds of the posterior mean of spatial intensity of FMD cases are shown in Figures 4 (A), (C), and (B), respectively. The corresponding posterior distribution of spatial random effects are shown in Figures 4(D), (F), and (E). The posterior distribution of parameters are also summarized in Table 1. The posterior mean of parameter estimates $\hat{\beta}$ are similar to $\hat{\beta}^*$

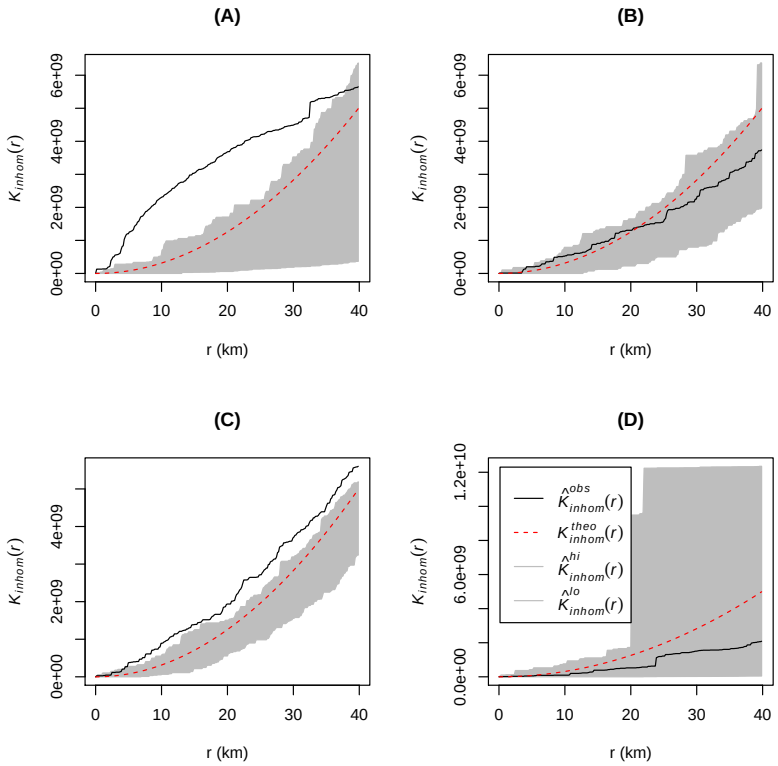


Figure 3: Inhomogeneous K function for four periods; (A) Period 1, (B) Period 2, (C) Period 3, and (D) Period 4. The solid and dotted lines, respectively, represent the inhomogeneous K function estimate and the K function under the null process, a theoretical expected value. The upper and lower simulation envelopes were shaded in a light grey colour.

except the coefficient of elevation $\hat{\beta}_1$ which was not statistically significant. It is clear that the influence of temperature appears to be relatively substantial compared to other risk factors in LGCP model, too. Both the FMD risk map in Figure 4 and the non-zero estimate of intercept $\hat{\mu}$ indicate that the risk of FMD was not constant over the study region but instead was spatially varying. Lastly, the posterior mean of range parameter ($\hat{\theta}$) is approximately 14.49 km with lower and upper bounds of 9.94 km and 20.12 km, respectively.

Table 1: Posterior distribution of LGCP model parameters

	Mean	Std. Dev.	0.025 quantile	0.975 quantile
(Intercept)	12.25	2.99	6.38	18.15
Elevation	-0.08	0.08	-0.23	0.07
Livestock holdings	0.30	0.08	0.14	0.47
Average temperature	-3.84	0.83	-5.48	-2.21
Driving distance	-1.10	0.23	-1.56	-0.66
Range	14,492.42	2,602.11	9,944.41	20,120.24
Std. dev.	1.12		0.94	1.32

IV DISCUSSION AND CONCLUSIONS

We described the FMD transmission process in S. Korea during 2010-2011 using spatial point pattern analyses. The application of inhomogeneous K function to the farm level FMD cases

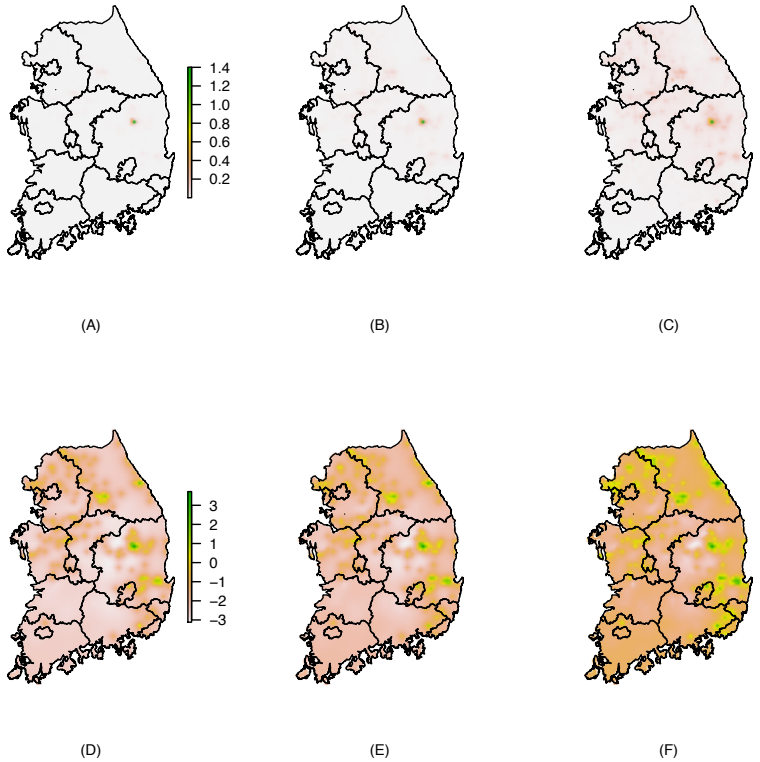


Figure 4: Posterior distribution of the spatial LGCP prediction; (A) 0.025 quantile, (B) Mean, and (C) 0.975 quantile of FMD risk (intensity) $E\{\Lambda|Y\}$; (D) 0.025 quantile, (E) Mean, and (F) 0.975 quantile of spatial random effects $E\{U|Y\}$

across four time periods suggested the presence of dynamically evolving the spatial patterns of FMD transmission over the course of the epidemic. We found that clusters at short scales during Period 2 and 3 were not statistically significant, which contrasts to the common assumption that contagious diseases form a spatial cluster. The absence of short scale clusters (less than 5 km) in both periods might be due to the enforcement of the government's control policy including culling and the movement restriction (Park et al., 2013).

Inhomogeneous K function analysis enabled us to account for the non-stationarity of the disease spread, but the spatial intensity was purely based on the observed FMD cases. For a further investigation of FMD transmission with respect to the behaviors and appearance of intensity function, we developed a LGCP model where spatially varying intensity of the disease was estimated via a log-linear function of environmental and socioeconomic factors. We also took a Bayesian approach to explicitly take into account parameter uncertainty. We found that LGCP models for FMD can be used as a tool for improving our understanding of the spread of FMD and optimizing disease control. For instance, our analyses demonstrated that temperature plays a critical role in determining the spread of FMD — probably because the virus tends to persist in lower temperatures. As shown in the map of FMD intensity in Figure 4(A)-(C), the area with high intensity is centered at the source of infection where the onset of FMD epidemic started. On the other hand, the maps of unexplained risk in Figure 4(D)-(F) illustrate the areas with high to low risk whose shapes and sizes vary over the study area.

The Bayesian estimates of the LGCP model parameters and prediction of random effects enabled us to assess the spatial variability of FMD incidences risk and to identify the spatial clustering under the explicit consideration of population at risk. The non-zero estimate of spa-

tial variability confirmed that FMD incidences formed a surprising tendency to cluster together even after the spatially inhomogeneous livestock farm density was taken into account. Perhaps LGCP may not be the only means of testing the hypothesis of spatial independence or interactions between events as variant K-functions can provide similar outcomes. However, LGCP models provide statistically rigorous inference and estimation and allow investigators to estimate the second order properties of the underlying process, that is, the spatial dependence of events. Furthermore, the LGCP inference using INLA approximation has relatively tractable moment properties, which allowed a simple method of parameter estimation and a flexible specification of the space-time covariance structure. We simplified the mean structure as a set of spatial covariates but this approach is naturally extendable to incorporate additional covariate information.

The proposed analysis/modeling framework is general and can be applied to other infectious diseases or FMD cases in other countries. The proposed LGCP model can fully account for the temporal aspect of FMD incidences by incorporating relevant temporal covariates and specifying spatio-temporal covariance structure (Diggle, 2013). In doing so, for example, we might have been able to take into account the changes in the susceptible population due to the culling policy. We expect that the longer time series of FMD incidence will improve our understanding (given situation) and allows us to evaluate a wide range of control strategies using the model.

References

Alexandersen S., Zhang Z., Donaldson A., Garland A. (2003). The pathogenesis and diagnosis of foot-and-mouth disease. *Journal of comparative pathology 129*(1), 1–36.

Baddeley A. J., Møller J., Waagepetersen R. (2000). Non-and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica 54*(3), 329–350.

Banerjee S., Carlin B. P., Gelfand A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL.: Chapman and Hall/CRC.

Brown P. E. (2015). Model-based geostatistics the easy way. *Journal of Statistical Software 63*(12), 1–24.

Diggle P. J. (2013). *Statistical Analysis of Spatial and Spatio-temporal Point Patterns*. Boca Raton, FL: CRC Press.

Ferguson N. M., Donnelly C. A., Anderson R. M. (2001). The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions. *Science 292*(5519), 1155–1160.

Gatrell A. C., Bailey T. C., Diggle P. J., Rowlingson B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 256–274.

Grubman M. J., Baxt B. (2004). Foot-and-mouth disease. *Clinical Microbiology Reviews 17*(2), 465–493.

Keeling M. J. (2005). Models of foot-and-mouth disease. *Proceedings of the Royal Society of London B: Biological Sciences 272*(1569), 1195–1202.

Møller J., Syversveen A. R., Waagepetersen R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics 25*(3), 451–482.

Moller J., Waagepetersen R. P. (2003). *Statistical inference and Simulation for Spatial Point Processes*. Boca Raton, FL: CRC Press.

Moutou F., Durand B. (1994). Modelling the spread of foot-and-mouth disease virus. *Veterinary Research 25*(2), 279–284.

Park J.-H., Lee K.-N., Ko Y.-J., Kim S.-M., Lee H.-S., Shin Y.-K., Sohn H.-J., Park J.-Y., Yeh J.-Y., Lee Y.-H., et al. (2013). Control of foot-and-mouth disease during 2010–2011 epidemic, South Korea. *Emerging Infectious Diseases 19*(4), 655.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sensitivity of DBSCAN in identifying activity zones using online footprints

David W. S. Wong[1*], Qunying Huang[2]

[1] George Mason University, USA
[2] University of Wisconsin-Madison, USA

*Corresponding author: dwong2@gmu.edu

Abstract
In mining spatial(-temporal) data for trajectory and activity analyses, a common task is to determine spatial clusters, which may represent activity zones. DBSCAN is a popular clustering algorithm. How the two parameters of DBSCAN that control the clustering algorithm may affect the results spatially has not been thoroughly investigated. This paper reported an incremental effort in conducting a sensitivity analysis by changing the parameter values. Preliminary results show that the two parameters work against each other to a certain degree. Increasing one parameter value (*minpts*) will break up larger clusters into smaller ones, while the other parameter (*eps*) controlled the spatial scale of clusters that can be detected.

Keywords
Spatiotemporal trajectories, spatial clustering, twitter, scale of cluster

I INTRODUCTION

Recent studies in GIScience and spatial analysis often exploited Twitter data, which may be considered as individual-level spatiotemporal data, to understand the mobility patterns of population. Many time-geography studies depict individual trajectories using space-time paths (sets of connected line segments) with data of similar nature to Twitter data. A fundamental issue is that this traditional representation depicts the trajectories of individuals with absolute certainty in space and time regardless if the data were gathered just for one day or several weeks, or derived from data like Twitter.

Twitter data may be regarded as data collected from semi-random spatiotemporal sampling of individual movements. Huang and Wong (2015) proposed a framework to use such data to depict the regular mobility patterns of individuals with certainty levels. The critical process is to identify zones that individuals visited regularly within the 24-hour period. Zones are formed by clustering visited locations within the same temporal window. These zones over different periods are connected with 3D cones to show the spatiotemporal (ST) trajectories. The variable sized ST cones depict the spatial variability of the trajectories and different colour hues on the cones indicate the levels of (un)certainty. A critical step in determining the zones regularly visited by the individuals employed a highly popular clustering method proposed by Ester *et al*. (1996), the density-based spatial clustering of applications with noise (DBSCAN). This clustering method has been adopted in many mining applications of spatial data (e.g., Birant and Kut, 2007; Huang and Wong, 2015; Zhou *et al.*, 2004). How DBSCAN is executed in determining the activity zones will affect the sizes of the ST cones and thus the depiction of the ST trajectories, introducing another uncertainty dimension through the method, in addition to the uncertainty in the data.

DBSCAN requires two parameters: the minimum number of points that can form a cluster (*minpts*) and the maximum distance within which two points belong to a cluster (*eps*). Using different parameter values may obtain different results and thus the spatiotemporal trajectory

identification results are likely dependent on these parameter values. Although some studies have suggested appropriate values for *minpts*, these recommendations were data dependent and are not generally applicable. Although conducting a full-scale sensitivity analysis of DBSCAN is warranted to derive some general rules or guidelines in using this clustering method, the objective of this paper is more limited: evaluating the impacts of varying DBSCAN parameters on zone identification and thus ST trajectory variability depiction.

II METHOD AND DATA

Following the recommendations provided by Ester *et al*. (1996), Huang and Wong (2015) set the *minpts* value to 4 and the *eps* value was determined using an iterative procedure. On the other hand, Zhou *et al.* (2004) used 20 meters as the *eps* value as this value approximated the positional error in GPS readings. As these recommended methods in determining the parameter values were based upon specific studies, they are not necessarily applicable to other research contexts, particularly in determining the regular activity zones of individuals based upon locations reported through social media data. Therefore, in this study, we will conduct a limited scope sensitivity analysis using a range of parameter values for DBSCAN to explore how the activity patterns and ST trajectories depiction will vary.

To assess the impacts of varying *minpts*, we use values range from 3 to 10, as Ester *et al.* (1996) suggested using 4. While Zhou *et al.* (2004) suggested an *eps* value of 20 meters, we test its impacts with a much larger range from 10 meters to 60 meters. The maximum *eps* value of 60 meters is probably sufficient to accommodate the locational uncertainty of human activity patterns at the intra-urban scale, but it may not be sufficient for activities conducted in the suburban or even rural areas. Therefore, in choosing our test data, we limited our choices to urban settings. For the sensitivity analysis, Twitter users in Washington, DC and Chicago, IL were used. From the tweets posted between January 1, 2014 and March 31, 2014 with users who used “Washington, DC” and “Chicago” in their profiles, we identified more than 9000 and 17,000 unique users in the two cities. After screening the data with other requirements (e.g., minimum number of geo-tagged tweets has to be 3 or more), 4,442 and 4,088 users from DC and Chicago, respectively, were used. Each of these users posted more than 3 geo-tagged tweets throughout the approximately two-year period (we retrieved the maximum number of tweets, 3,200, for each user).

In addition, we also selected one user likely with his/her residential locations in Washington, DC (we rarely can be definitely sure about the home locations of users) for detailed analysis. Specifically, we want to examine in detail how the clustering results are affected by different *minpts* and *eps* values. Using these data, we test different parameter values using DBSCAN to determine their activity clusters.

III RESULTS OF SENSITIVITY ANALYSIS

The two concerned parameters for DBSCAN control two properties of clusters to be identified. The *minpts* parameter controls how dense point locations will constitute a cluster. The *eps* parameter controls the spatial extent of a cluster. With smaller *minpts* values, more but smaller clusters are expected to be identified. Therefore, using smaller *minpts* values will likely increase the probability of committing type I error (false positive), including random locations that may not be regular activity locations. Using larger *minpts* values will likely produce fewer but larger clusters, likely increasing the probability of type II error (false negative) and collapsing distinctive clusters. On the other hand, how *eps* value may affect clustering results has not been clear, although larger *eps* values may possibly produce smaller numbers of clusters but larger in size.

Figure 1 summarizes the results by varying *minpts* from 3 to 10, and *eps* from 10 to 60 for both Washington, DC and Chicago, IL. When *minpts* increases from 3 to 10, the averaged number of clusters decreases. This result is intuitively expected. When the clustering process requires a larger minimum number of points in order to form a cluster, given that the total number of points to be clustered is fixed, fewer clusters can be formed. This general pattern was found in both Washington, DC and Chicago, IL and is valid regardless of *eps* value. While the impact of changing *minpts* value on the number of clusters seems obvious, how the changes in *minpts* value affect the process is not apparent.

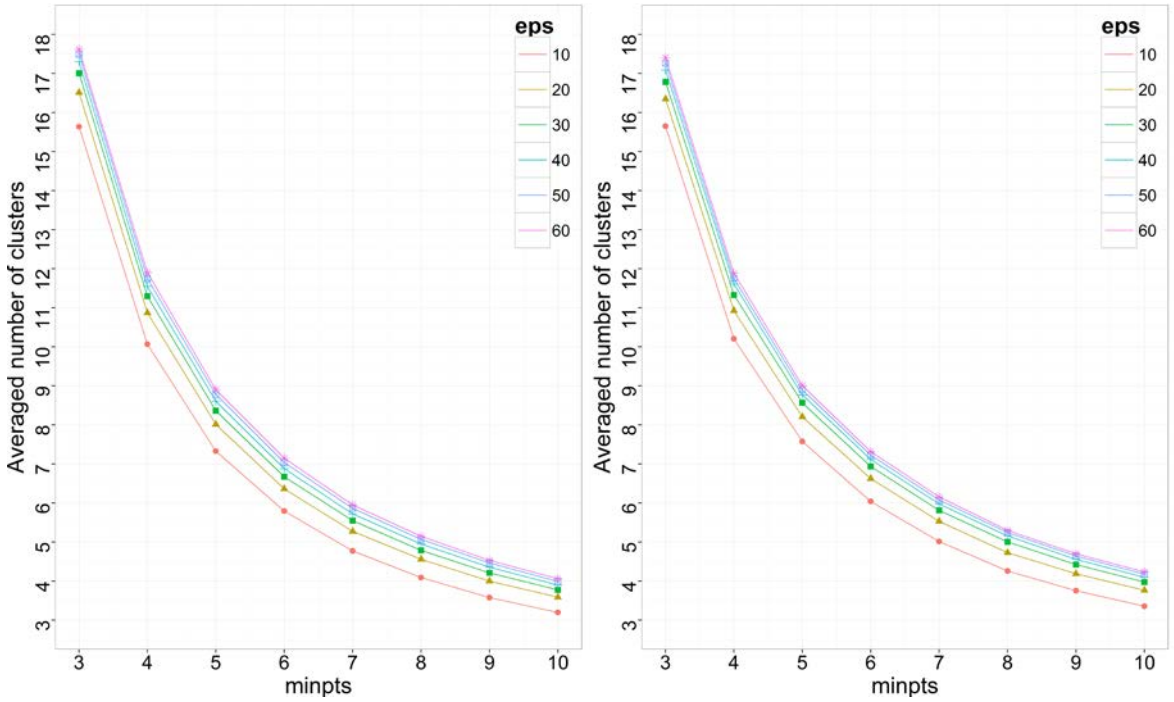


Figure 1: Averaged number of clusters with varying *minpts* and *eps* values (left: Washington, DC; right: Chicago, IL)

Figure 1 also shows that the average number of clusters increases slightly with increasing eps values when *minpts* is kept constant. When *eps* is raised, even points farther away can be included to meet the *minpts* requirement to form cluster. Then, clusters with larger *eps* values may likely be more spatially dispersed, or geographically larger in size. This effect will become more apparently when we examine the clustering results of locations from an individual Twitter user.

We selected a set of locations of tweets from a user to examine the changes in clustering detection when changing the DBSCAN parameter values. Figure 2 shows the clusters identified by DBSCAN using a subset of locations from geo-tagged tweets of a user in Washington, DC. The cluster boundaries were determined using the minimum bounding polygon. The *eps* value was kept at 20 meters, but *minpts* was altered from 5 to 9. In the aggregate analysis described above, we expected that when *minpts* increases, the number of cluster decreases because fewer clusters can be formed as more neighbouring points are required to be reachable within 20 meters for each point to be included in a cluster. Figure 2 shows that when *minpts* value increases, the process ignores smaller cluster and thus fewer

clusters are retained. Thus, fewer clusters were formed was not because clusters were competing for points to form larger but fewer clusters. However, the areal extents of clusters shrink when *minpts* increases, quite an unexpected outcome. In fact, larger clusters broke up into smaller clusters when *minpts* value increased (See the cluster(s) on the upper Figures d and e with *minpts* = 7 and *minpts* = 8).

Figure 2: Detected clusters using different *minpts* values and a fixed *eps* value of 20 meters on dispersed locations (The geo-tagged tweets are displayed as red dots and the boundary of a cluster is depicted in blue)

Results reported above show that when value of *minpts* increases, smaller clusters will be removed, but larger clusters will be broken up into more fragmented clusters. This unexpected outcome can be traced back to the design of the DBSCAN algorithm and how *minpts* is used in the clustering process. For each point to be added to a cluster, the point has to reach *minpts* number of points given the *eps* value. So it is possible that a point originally in a large cluster, likely at the edge, with *minpts* value, say 7, could find 7 points, but can no longer find 8 or 9 points within the *eps* value when *minpts* increases to 8 or larger values.

We also examined the varying clustering results when *eps* value changes while keeping the *minpts* value to 4. Figure 3 reports part of the results. With increasing *eps* value from 20 to 70 meters, smaller clusters are merged to form larger but fewer clusters. To a large degree, this general pattern is expected. However, results described in Figure 3 illustrate a fundamental issue in clustering analysis – spatial scale. Cluster detection is a scale-dependent process (e.g., Donnelly, 1978). The *eps* value essentially controls the spatial scale in which the clusters are to be determined. Thus, when *eps* values are small, clusters are small in their spatial extents, and vice versa.

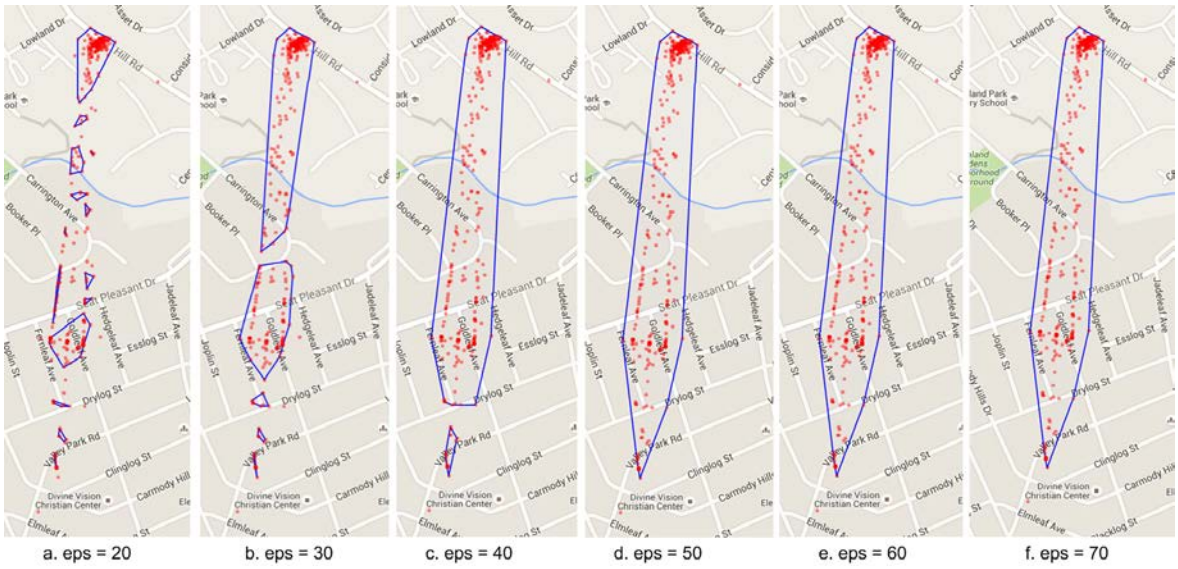


Figure 3: Detected clusters using different *eps* values and a fixed *minpts* value of 4 on dispersed locations

IV CONCLUSIONS

In this short paper, we report only partially results from the sensitivity study. Using larger *minpts* values not only remove smaller clusters, lowering the probability of identifying false positives, but break up larger clusters into smaller clusters. Breaking up larger clusters into smaller ones may not be regarded as "concerning" although the spatial structure of clusters of the entire study area is likely affected. While the results from changing the *eps* value were not surprising, the results highlight the role of the *eps* parameter in determining the spatial scale of cluster detection. Conceptually, determining an "optimal" scale for clustering detection is challenging, not to mention the spatial heterogeneity that may be presented in the data. Taking all these together, we may not be able to qualify the reliability of clustering results using the current tool.

The reported results just focus on the impacts of changing *minpts* or *eps* values, but not simultaneously, or their interaction. On the surface, the two parameters seem to be working against each other: increasing value of *minpts* not only reduces the number of clusters but also areas that fall within the clusters, while increasing *eps* value produces extensive clusters. How the two parameters interact needs to be scrutinized in the future. In addition, the spatial distribution of points to be clustered should also be considered since it may affect the performance of DBSCAN or other clustering detection algorithms.

References

Birant, D., Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering* 60, 208-221.

Donnelly, K. (1978) Simulations to determine the variance and edge-effect of total nearest neighbor distance. In I Holder (ed) *Simulation Methods in Archaeology*, pp. 91-95.

Ester, M., Kriegel, H-P, Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press. pp. 226–231.

Huang, Q., Wong, D. W. (2015). Modeling and visualizing regular human mobility patterns with uncertainty: an example using Twitter data. *Annals of the Association of American Geographers* 105(6), 1179-1197

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., Terveen, L. (2004). Discovering personal gazetteers: an interactive clustering approach. Paper presented at the *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, ACM, pp. 266-273.

On Reliability of Remote Sensing Data and Classification Methods for estimating transition rules of the land-use Cellular Automata

Yulia Grinblat*[1, 2], Michael Gilichinsky[3] and Itzhak Benenson*[1]

[1]Department of Geography and Human Environment, Tel Aviv University, Tel-Aviv,
[2]The Porter School of Environmental Studies, Tel Aviv University, Tel-Aviv
[3]Elbit Systems, Rehovot

*Corresponding author: juliagri@post.tau.ac.il, bennya@post.tau.ac.il

Abstract
Typically, the Cellular Automata (CA) models of Land-Use/Land-Cover (LULC) changes focus on estimating the rules of the LULC changes and analysis of the simulation results. However, the models put aside the uncertainty of the LULC maps that are used for establishing the transition rules. Our study questions the reliability of Remote Sensing (RS) data sources and classification methods applied for constructing these maps.
Based on four time intervals within a 36-year period, we construct LULC maps and estimate the transition probabilities between six LULC states. The LULC maps and transition probabilities matrices (TPM) are built based on the manual interpretation of high-resolution aerial photos and classification of multispectral Landsat images for the same years.
We consider the maps and TPM derived from the aerial photos as a reference data, and compare them to those constructed from the Landsat images classified with several methods: mean-shift segmentation algorithm followed by Random Forest classification method, and three pixel-based methods of classification: K-means, ISODATA, and maximum likelihood. Then, for each classification the TPM were compared to the referenced TPM.
The accuracy assessment of all maps obtained with the pixel-based methods is insufficient for estimating the LULC TPM. The LULC map obtained with the object-based classification method fit well to that based on the aerial photos, but the estimates of TMP are qualitatively different from those constructed from the aerial photos.
This article raises doubts regarding the adequacy of Landsat data and standard classification methods for establishing LULC CA model rules, and calls for the careful reexamination of the entire land-use CA framework.

Keywords
Cellular automata, Landsat images, land-use/land-cover changes, Markov transition probabilities matrices, validation of RS classification methods

I INTRODUCTION

Conceptual simplicity and the ability of explicit representation of landscapes and their changes make Cellular Automata (CA) a standard tool for simulating urban and regional land-use dynamics (Clarke et al. 1996; White and Engelen 1997), which potential for modeling Land-Use/Land-Cover (LULC) dynamics is widely recognized (Wu and Webster 1998; Pijanowski et al. 2002; Verburg et al. 2002). The major source of data for the CA modeling is Remote Sensing (RS) multispectral imagery classified for establishing land uses and covers.

It is often reported that the CA models are quite successful in predicting LULC changes, with the high overall fit (80-90%) between the real LULC data and model outputs. This is indeed true when the validation is based on comparing the *entire modeled area*. However, the period of time covered by the CA model is, usually, between one and few decades and the fraction of the modeled area that has been changed during such a period is, typically, few percent of the entire city area. As far as initial area is excluded from the comparison, the spatial fit between the predicted and real *changes* drops down (Hagen-Zanker et al. 2005; Pontius and Petrova 2010).

A hierarchy of reasons of limited capacity of the CA models for predicting LULC changes can be proposed: (1) CA framework as a whole is insufficient for predict LULC dynamics, due to, say, essential part of human bounded rational decisions in land planning and management; (2) The CA framework works, but wrong CA rules are chosen; (3) The CA framework works, the rules are properly established, but the data chosen for estimating parameters of the rules do not represented the real of the LULC changes. In this paper we deal with the latter and investigate the adequacy of the RS data for calibration and validation of the CA models.

II Testing the adequacy of classifications methods

Strangely enough, the adequacy of the RS classification for representing LULC *changes* remains on the margin of the CA modeling studies. Despite a series of publications that regard the erroneous consequences of misclassification (van de Voorde et al. 2009; van der Kwast et al. 2009) and sensitivity of the CA dynamics to the parameters of the CA rules (Liu and Andersson 2004; Jantz and Goetz 2005; Dietzel and Clarke 2006), the majority of modeling studies carelessly exploit the simplest methods of the RS images classification, take their outputs for granted, and focus on model calibration. This may evidently result in inadequate transition rules regardless of the calibration methods. The standard source of data for the CA model calibration and validation is 30m resolution LANDSAT multispectral imagery. We thus investigated the adequacy of different methods of LANDSAT images classification for establishing CA model rules.

The background of the CA model is Transition Probability Matrix (TPM) $\{p_{ij}\}$ - a set of probabilities, per time unit, of transition $S_i \rightarrow S_j$ between the states S_i and S_j of the LULC CA. Our study compares TPMs estimated based on the LANDSAT maps obtained by the different classification methods to the ground truth – the TPM that is estimated based on the manual interpretation of high-resolution aerial photos of the same area.

III Study area and Data preparation

The experimental area is the 15x6 km transect that starts in the center of the city of Netanya, Israel, and extends to surrounding agriculture areas. The period of comparison 1972 – 2008 (36 years) is divided into 4 intervals of 6 - 11 years, depending on availability of the LANDSAT images and aerial photos. Based on the manual interpretation of the high-resolution aerial photos, we have constructed the maps of Netanya LULC dynamics of LULC states. Six LULC states are considered: built-up areas (BU), roads (RD), agricultural (AG) and vegetation (VG) areas, open spaces (OS) and water surfaces (WA). In this short paper we present the results aggregated into three states only - – built-up (BU), agriculture (AG) and the rest states (RE) that aggregate the rest four LULC states.

Four popular in the CA studies pixel-based methods and one object-based method were applied for classifying the LANDSAT images. To remind, pixel-based classification methods consider pixels individually, while object-based methods recognize spatially continuous homogeneous domains of pixels, and then assign a land use to these segments (Lu and Weng 2007). All exploited pixel-based methods are traditional first choice of a CA modeler: K-means, ISODATA, Maximum Likelihood (ML) and hybrid classification. The object-based method we apply is two-staged: mean-shift clustering segmentation (Comaniciu and Meer 2002) is followed by a Random Forest classification (Breiman 2001).

IV The results

As can be seen in Figure 1 the fit between the LANDSAT-based maps and the map that is based on manual classification varies depending on the method. We do not present here the results of numerical analysis of this fit and focus on comparison of transition probabilities matrices. For this comparison, the TPMs all estimated for the time periods of different length, were normalized to the 10-year period.

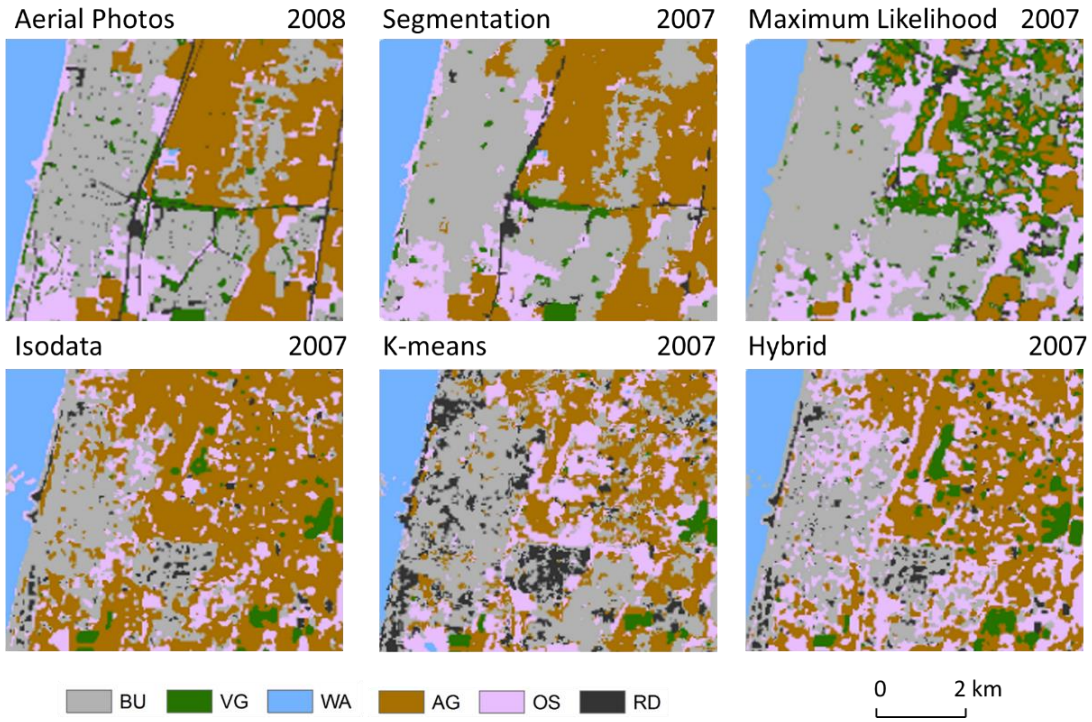


Figure 1. Land-use maps for the part of the study area obtained with the manual classification of the aerial photos and with several classification methods applied to the LANDSAT images

Table 1 presents the TPMs obtained with two of the applied methods, Maximum Likelihood and Segmentation, together with the TPM obtained for the manually classified map. Due to limited space, the TPMs are presented for the LULC uses aggregated into three classes and for the last time interval of the investigated period - 2000-2007, only. The aggregated land uses classes are as follows: BU - built-up areas and roads; AG - agricultural areas; RE –the rest of land uses - open spaces, vegetated areas and water surfaces.

		Maximal Likelihood				**Segmentation**				**Aerial photos**			
		BU	**AG**	**RE**	**Tot**	**BU**	**AG**	**RE**	**Tot**	**BU**	**AG**	**RE**	**Tot**
Transition probability	**BU**	**0.75**	0.05	0.20		**0.78**	0.13	0.09		**0.99**	0.00	0.01	
	AG	0.14	**0.64**	0.22		0.08	**0.86**	0.06		0.02	**0.95**	0.03	
	RE	0.32	0.09	**0.58**		0.13	0.13	**0.74**		0.07	0.10	**0.83**	
Area of transition, km^2	**BU**	**7.10**	0.46	1.90	9.50	**16.9**	2.80	1.8	21.5	**19.2**	0.02	0.11	19.3
	AG	5.50	**8.90**	25.9	40.3	3.7	**40.6**	3.0	47.3	0.87	**41.2**	1.30	43.4
	RE	11.9	3.30	**21.4**	36.6	2.3	2.20	**13.2**	17.7	1.70	2.30	**19.9**	23.9

Table 1. TPM for transitions between three land uses based of the 2000-2007(LANDSAT) / 1999-2008(Aerial photos) data. Probabilities are normalized to the 10 year period

As can be seen from Table 1, for the presented period, the TPMs obtained with the ML and Segmentation methods are qualitatively and quantitatively different from the TPM estimated based on the aerial photos. Most important, in reality, LULC states are changing in time

essentially less frequently than it is obtained based on the RS images classified with the ML method; for example, in reality, the probability of the AG→AG transition per 10 years is 0.95, while according to the ML map this probability is 0.64 only. The same is true for the rest of the pixel-based methods (not presented here). The TPM obtained with the Segmentation method is much closer to the reference TPM that the TPM of the pixel-based methods, but yet essentially biased towards more changes than in reality. Similar differences are characteristic of all other periods, as well as for the TPMs constructed for the basic, non-aggregated, set of the LULC states.

V Conclusions

We thus conclude that none of the maps obtained, based on the LANDSAT images, with the help of the popular pixel-based classification methods can be exploited for establishing CA transition rules. Object-based method provided better, but yet insufficiently precise estimates. We call for the revision of approach to the CA calibration and validation. An open depository of high-resolution, carefully validated, long-term series of the land-use/cover maps that reflect different types of LULC dynamics, and represent different types of land planning systems for different periods of population growth and economic development should be established. Instead of establishing a new database for every new CA model, one has to use these data series for calibration and validation of her/his new model. Only then, the model can be applied to the new dataset which, as we have demonstrated, must be constructed with the great care.

References

Breiman, L. 2001. Random Forests. *Machine Learning* 45 (1):5-32.

Clarke, K. C., S. Hoppen, and L. Gaydos. 1996. A self-modifying cellular automata model of historical urbanization in the San Franciso Bay area. *Environment and Planning B: Planning and Design* 24:247 - 261.

Comaniciu, D., and P. Meer. 2002. Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24 (5):603-619.

Dietzel, C., and K. Clarke. 2006. The effect of disaggregating land use categories in cellular automata during model calibration and forecasting. *Computers, Environment and Urban Systems* 30 (1):78-101.

Hagen-Zanker, A., J. van Loon, A. Maas, B. Straatman, T. de Nijs, and G. Engelen. 2005. Measuring performance of land use models: An evaluation framework for the calibration and validation of integrated land use models featuring Cellular Automata. Paper read at 14th European Colloquium on Quantitative Geography, September 9-13, at Tomar, Portugal.

Jantz, C. A., and S. J. Goetz. 2005. Analysis of scale dependencies in an urban land-use-change model. *International Journal of Geographical Information Science* 9 (2):217-241.

Liu, X., and C. Andersson. 2004. Assessing the impact of temporal dynamics on land-use change modeling. *Computers, Environment and Urban Systems* 28 (1–2):107-124.

Lu, D., and Q. Weng. 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing* 28 (5):823-870.

Pijanowski, B. C., D. G. Brown, B. A. Shellito, and G. A. Manik. 2002. Using neural networks and GIS to forecast land use changes: a Land Transformation Model. *Computers, Environment and Urban Systems* 26 (6):553-575.

Pontius Jr, R. G., and S. H. Petrova. 2010. Assessing a predictive model of land change using uncertain data. *Environmental Modelling & Software* 25 (3):299-309.

van de Voorde, T., F. Canters, J. van der Kwast, G. Engelen, M. Binard and Y. Cornet. 2009. Quantifying intra-urban morphology of the Greater Dublin area with spatial metrics derived from medium resolution remote sensing data. *2009 Joint Urban Remote Sensing Event*, Shanghai: 1-7.

van der Kwast, J., I. Uljee, G. Enegelen, T. Van de Voorde, F. Canters, and C. Lavalle. 2009. Using Remote Sensing Derived Spatial Metrics for the Calibration of Land-Use Change Models. Paper read at 7th International Urban Remote Sensing Conference.

Verburg, P. H., W. Soepboer, A. Veldkamp, R. Limpiada, and V. Espaldon. 2002. Modeling the Spatial Dynamics of Regional Land Use: The CLUE-S Model. *Environmental Management* 30 (3):391 - 405.

White, R., G. Engelen, and I. Uljee. 1997. The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics. *Environment and Planning B: Planning and Design* 24 (3):323-343.

Wu, F., and C. J. Webster. 1998. Simulation of land development through the integration of cellular automata and multicriteria evaluation. *Environment and Planning B: Planning and Design* 25 (1):103-126.

Evaluating performances of spectral indices for burned area mapping using object-based image analysis

Taskin Kavzoglu[1]*, Merve Yildiz Erdemir[1], Hasan Tonbul[1]

[1]Gebze Technical University, Department of Geomatics Engineering, 41400, Kocaeli, Turkey

kavzoglu, m.yildiz, htonbul@gtu.edu.tr

Abstract

Determining post-fire information is crucial for post-fire management activities and rehabilitation treatments. The use of robust and advanced approaches is needed to determine fire severity and thoroughly analyze post-fire rehabilitation period. Object-based image analysis (OBIA) is a powerful approach that has been successfully applied in many research problems in remote sensing arena. However, its use in forest fire and related studies including fire severity and burned area estimation is quite limited. This study was carried out in Antalya's Taşağıl district (Turkey) where according to Directorate of Forestry reports one of the largest wildfires in the Turkey occurred in 2008. The objectives of the present work are (i) to investigate the performance of object based analysis for burned area mapping; (ii) to compare the performances of widely-used burned area related spectral indices in identifying burned, slightly burned, water and non-burned areas from each other, and (iii) to delineate the boundaries of burned area. In this context, spectral indices of Normalized Burn Ratio (NBR), Normalized Vegetation Index (NDVI), Burned Area Index (BAI) derived from the satellite image were employed in analyses. Multiresolution segmentation and fuzzy membership function classifier were applied to the combinations of the selected indices (NDVI, BAI-NBR, NDVI-NBR) to discriminate burned, slightly-burned and non-burned areas from each other. Results showed that all combinations constructed in this study produced satisfactory results in terms of classification accuracy. However, the highest accuracy (98.37%) was achieved by NDVI-NBR index combination whilst the lowest accuracy (94.59%) was achieved when only the NDVI index was employed in OBIA process. It is hoped that with this work a contribution will be made for the government agencies to delineate fire perimeter and determine risk of wildfire for post-fire damage management.

Key words: Forest fire, Wildfire, NBR, NDVI, BAI, Object-Based Image Analysis

I Introduction

Wildfires are one of the most important natural disasters with respect to catastrophic consequences which cause serious social, economic and environmental problems. Particularly, the Mediterranean ecosystem of Turkey is suffered from increasing number of forest fires and severities due to human-induced activities or natural conditions. According to Turkish General of Directorate of Forestry report, 142,409 hectares of forest was burned between the years 2002 and 2014, most of which are located in the Aegean and Mediterranean regions of Turkey (OGM, 2014). In this context, burned area mapping is of crucial importance for fire management and post-fire damage estimation to determine fire behavior.

Several studies have investigated the use of remote sensing in burned area mapping on Mediterranean region (Mitri and Gitas, 2004; Kavzoglu et al., 2014; Pleniou and Koutsias, 2013). In addition; various spectral indices (e.g. NBR, NDVI, BAI) have been widely utilized

to monitor fire-induced vegetation changes, including burn severity and regeneration (Chen et al., 2011; Chuvieco et al., 2002; Veraverbeke et al., 2011).

OBIA considers spectral, textural and hierarchical information of objects, a group of neighboring pixels with similar characteristics. In contrast to pixel-based image analysis, it constructs segments from the objects and show high performances by producing more accurate thematic maps. OBIA, which deals with spectrally homogenous image objects instead of single pixels, may be more effective for burned area mapping (Mitri and Gitas, 2004). Mitri and Gitas (2002) states that "the combination of the object-oriented approach and the multispectral resolution data of Landsat TM showed very promising results in burned area mapping and in discriminating between burned and the other classes of confusion."

The aim of this study was to examine the efficiency of OBIA together with various spectral indices for the Antalya-Taşağıl wildfire in Turkey by using Landsat ETM+ images. Main objectives of this study are twofold; i) to distinguish burned and non-burned areas precisely, (ii) to assess the effectiveness of object-based image classification for burned area mapping.

II Study Area and Dataset

This study focuses on a large fire that occurred on 31th July 2008, in the Antalya-Taşağıl region (the central coordinates: 37° 03'N, 31° 10'E) of Turkey (Figure 1). The fire was human-induced and worsened by the prevailing winds. Pinus brutia is the dominant tree species at the region and other types of Mediterranean vegetation, (e.g. maquis) exist in the study area. The region is characterized by the Mediterranean climate, mean annual temperatures ranging between 10° and 18.5° C and with dry and hot summers. The mean annual precipitation ranges between ~450 and ~1020 mm.

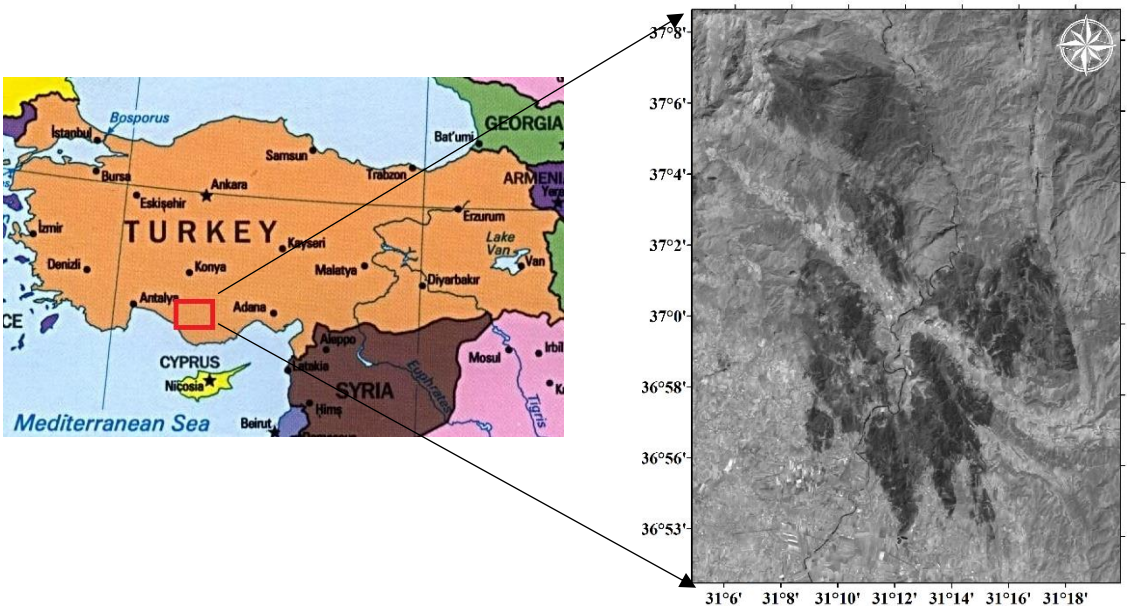


Figure 1: Location of the study area.

A cloud-free Landsat ETM+ image with an acquisition date of 12 September 2008 was obtained at no cost from the United States Geological Survey (USGS) archive (http://earthexplorer.usgs.gov/). The preprocessed (radiometrically corrected and geometrically registered) image was registered to UTM projection with WGS84 datum. A widely-used object based image processing software, eCognition Developer (v9.1), was used for segmentation and classification experiments in this study.

III Methodology

Spectral Indices

In this study, three spectral indices (NDVI, NBR and BAI) were calculated using the following equations (Table 1) to distinguish between burned and unburned surfaces for Landsat ETM+ image. The indices were calculated using the bands wavelengths: NIR (0,75 to 0.90 μm), R (0.63 to 0.69 μm), and SWIR (1.55 to 1.75 μm) for each spectral region.

Spectral Index	Abbreviation	Formula
Normalized Difference Vegetation Index	NDVI	$NDVI = \frac{NIR-R}{NIR+R}$
Normalized Burn Ratio	NBR	$NBR = \frac{NIR-SWIR}{NIR+SWIR}$
Burned Area Index	BAI	$BAI = \frac{1}{(0.1-R)^2+(0.06-NIR)^2}$

Table 1. Spectral indices used in this study (NIR: Near Infrared, R: Red, SWIR: Short Wave Infrared).

Veraverbeke et al. (2011) state that enhancing the NBR with Landsat's thermal band provides better seperability between burned and unburned lands. The latter index, the BAI, has a high discrimination ability for burned areas in the R-NIR spectral domain (Chuvieco et al., 2002). Furthermore, the NDVI has been widely used in monitoring fire affected areas (García-Haro et al., 2001; van Leeuwen et al., 2010).

Image Segmentation and Classification

Segmentation, the first step in the OBIA, splits an image into random object primitives to build up homogenous (segments or image objects) regions. Multiresolution segmentation algorithm (Baatz and Schäpe, 2000), which is a bottom-up region merging technique was utilized in this paper. The segmentation of Landsat ETM+ image was constructed by adjusting the scale, band weights and color-shape parameters.

The scale parameter is an unitless abstract which asses the maximum allowed heterogeneity of image objects. Optionally, user can define the scale parameter according to required level of detail in image. Higher scale parameter values construct larger objects; smaller scale parameter values generate smaller objects. Some studies emphasized the relation between scale parameter and number of objects (Addink et al., 2007). Moreover, several studies have underlined the variation in classification accuracy due to the user-defined scale setting (Kim et al., 2009; Kavzoglu and Yildiz, 2014).

In this study, bottom-up approach was performed to define land use/land cover objects. Based on a "trial and error" procedure, scale parameter was set to 50 and 15 respectively for level 1 and 2 in segmentation process. Membership function classifier was implemented in the commercial software eCognition (Developer v. 9.1). Membership functions allow users to define the relationship between feature values and the degree of membership to a class using fuzzy logic. Minimum and maximum value set the upper and lower limits of the membership function (Trimble, 2011). The fuzzy sets were defined by membership functions to identify feature attributes according to spectral and contextual information. Classifications were performed on two segmentation levels: "water" class at Level 1 (i.e., scale 50); 'burned area', 'slightly-burned area' and 'non-burned' class at Level 2 (i.e., scale 15).

IV Results and Discussion

Multiresolution segmentation and fuzzy membership function classifier were applied by using three different combinations of the selected indices (NDVI, BAI-NBR, NDVI-NBR) to discriminate water, burned, slightly-burned and non-burned areas from each other. In iterative steps, a two-level network of image objects was advanced. Membership functions were utilized by the thresholds of objects regarding to their spectral characteristics.

The Level 1 objects were initially categorized. SWIR band and NDVI values were utilized on membership functions of spectral information to distinguish water class, which was delineated similarly on the combinations. Level 2, main level of classification, was created smaller image objects. Three classes representing 'burned area', 'slightly-burned area' and 'non-burned area' were generated at this level.

In the first combination which is based on only NDVI, classification results were not sufficient to detect burn scars. It was observed that rocky land, vegetation and urban areas were mixed with burned area (Figure 2a). Mean brightness and NDVI values were used to define membership functions for burned area and slightly-burned area. NDVI based threshold values were selected as "0.089 / 0.20", "0.20 / 0.34" and "-0.52 / 0.85" for burned, slightly burned and non-burned class, respectively. At the end of the first step, the overall accuracy of the classification was estimated as 94.59%. Secondly, BAI-NBR combination was implemented to delineate the burned areas (Figure 2b). Image segments were classified as burned area class using NBR based threshold values "-0.8 / 0.1" and excluding the BAI based threshold values "-1.0 / 1.8". Similarly, slightly-burned area was determined using NBR based threshold values "-0.03 / 0.41" and excluding the BAI-based threshold values "-0.8 / 2.33". The related BAI-based threshold values excluded to eliminate confusion between burned and non-burned pixels. In non-burned area class determination, NBR based threshold values were selected as "-0.39 / 0.99". Although this approach produced good results for burned area discrimination, it was observed that several pixels were misclassified in coastal regions. In BAI-NBR combination, overall classification accuracy was estimated 97.23%. In third approach, NDVI-NBR combination was applied to distinguish burned areas from non-burned areas (Figure 2c). NBR index was utilized to differentiate burned areas. NBR based threshold value was selected as "-0.38 / -0.08" for burned area class. Slightly-burned area was determined by using NDVI and NBR indices together. Thresholds of NBR and NDVI for slightly burned area class selected as "-0.06 / 0.42" and "0.19 / 0.42", respectively. In non-burned area discrimination, NBR based threshold value was chosen as "-0.3 / 0.79". Finally, overall classification accuracy was calculated 98.37% for NDVI-NBR combination.

Since there is no information about post-fire perimeter map estimated by the Turkish Directorate of Forestry, the accuracy assessment was evaluated by using authors' previous research (Kaya et al., 2014). The related study was investigated to predict burn severity with using the differenced Normalized Burn Ratio (dNBR) algorithm in same region. Training and test datasets were collected according to mentioned above study's results. The overall classification accuracies of three combinations for burned area mapping were presented in Table 2.

Results presented in Table 2 showed that the NDVI-NBR approach increased the overall accuracy by about 4% compared to NDVI, while this increment was approximately 3% for the burned area class. It was also observed that producer's accuracy calculated for each class was higher for BAI-NBR combination than NDVI-NBR combination.

Combination		Producer's Class Accuracy (%)				Overall Accuracy(%)
		Burned	Slightly Burned	Non-Burned	Water	
1	NDVI	94.06	92.42	94.64	100	94.59
2	BAI-NBR	98.03	94.62	94.56	100	97.23
3	NDVI-NBR	96,66	97.48	98.31	100	98.37

Table 2. Accuracy assessment results for NDVI, BAI-NBR and NDVI-NBR analysis.

It is apparent from Figure 2 that the classified objects of burned area (especially in NDVI-NBR and BAI-NBR combinations) could strongly be distinguished from other classes. Additionally, users and fire agencies can extract burned area borders by using burned area mask for the purpose of fire perimeter definition.

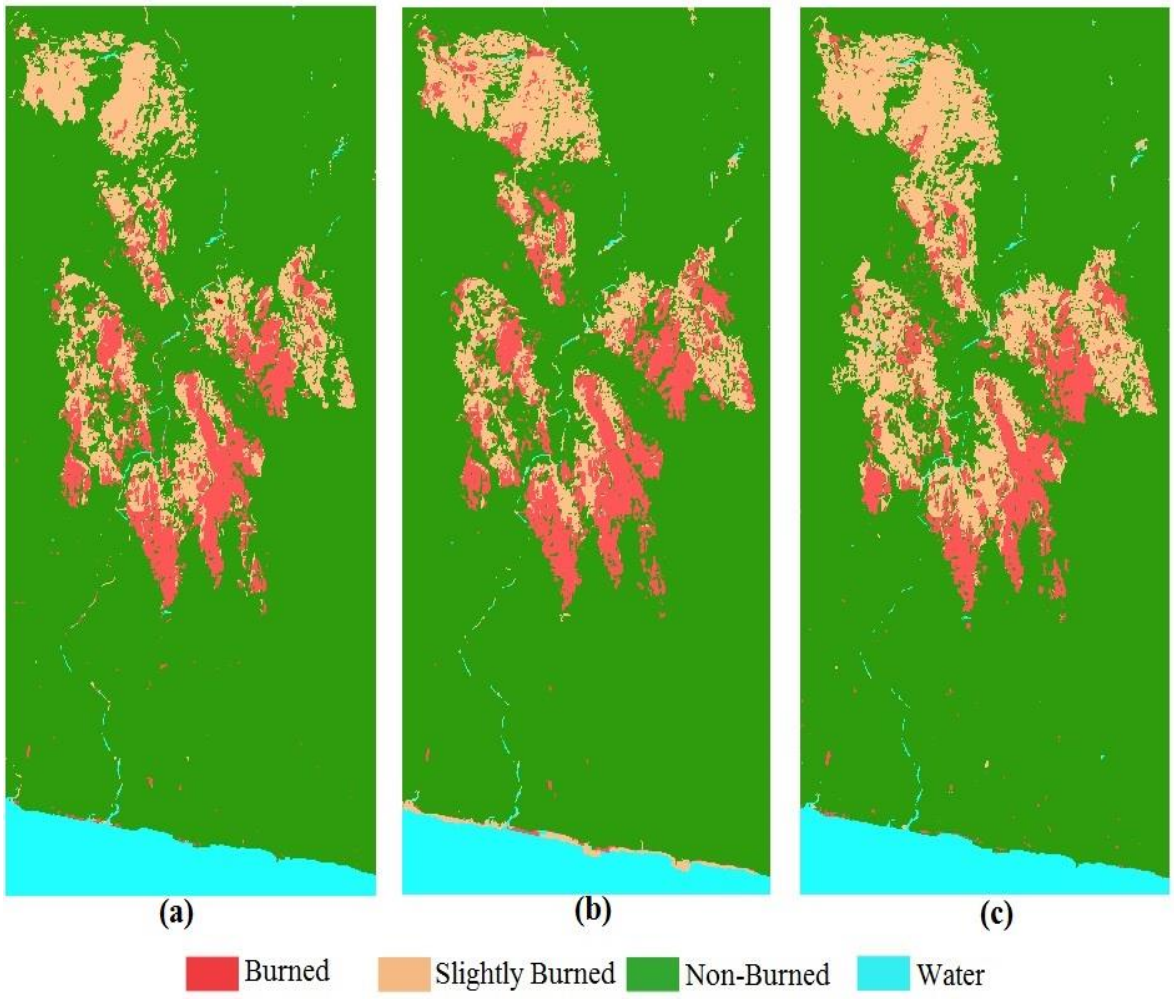


Figure 2: Classification results of three different combinations: (a) NDVI, (b) BAI-NBR, (c) NDVI-NBR.

V Conclusion

In this study, effectiveness of object-based classification was assessed in the determination of burned area mapping of Antalya-Taşağıl forest fire in Turkey. For this purpose, spectral indices of NDVI, NBR and BAI derived from Landsat ETM+ image were computed in analyses. Multiresolution segmentation and fuzzy membership function classifier were implemented to determine burned area.

The findings from this study make several contributions to the current literature. The combination of different spectral indices with OBIA approach showed very promising results. Furthermore, the use of membership functions contributed to reduction of misclassified pixels. The highest accuracy (98.37%) was obtained from NDVI-NBR combination in OBIA process. The following conclusions can be drawn from the present study, object-based classification can be used as a tool for rapid operational burned area mapping of Mediterranean forest fires. Unfortunately, no information about post-fire perimeter map was estimated by the Turkish Directorate of Forestry. The results obtained from these findings could provide valuable information for government agencies, planners and decision makers to construct management of forest fire strategies.

References

Addink E., Jong S. de, Pebesma E. (2007). The importance of scale in object-based mapping of vegetation parameters with hyperspectral imagery. *Photogrammetric Engineering & Remote Sensing*,. 73(8), 905-912.

Baatz M., Schäpe A. (2000). Multiresolution Segmentation: an optimization approach for high quality multi-scale image segmentation. In: Strobl, J., Blaschke, T., Griesebner, G. (Eds.), *Angewandte Geographische Informations*-Verarbeitung XII. Wichmann Verlag, Heidelberg, pp. 12-23.

Chen X., Vogelmann J. E., Rollins M., Ohlen D., Key C. H., Yang L., Huang C., Shi H. (2011). Detecting post-fire burn severity and vegetation recovery using multitemporal remote sensing spectral indices and field-collected composite burn index data in a ponderosa pine forest. *International Journal of Remote Sensing* 32(23), 7905-7927.

Chuvieco E., Martín M. P, Palacios A. (2002). Assessment of different spectral indices in the red-near-infrared spectral domain for burned land discrimination. *International Journal of Remote Sensing* 23(23), 5103-5110.

García-Haro F.J., Gilabert M.A., Meliá J. (2001). Monitoring fire-affected areas using Thematic Mapper data. *International Journal of Remote Sensing* 22 (4), 533-549.Kavzoglu T., Kaya S., Tonbul H. (2014, October 14-17). Detecting Burn Area and Burn Severity Using Modis Satellite Image Via Spatial Autocorrelation Techniques. In *5th Remote sensing and Geographic Information Systems Symposium (UZALCBS'2014)*, Istanbul/Turkey (in Turkish).

Kavzoglu T., Yildiz M. (2014, September 29-October 2). Parameter-based performance analysis of object-based image analysis using aerial and Quikbird-2 images. In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-7, pp.31-37.

Kaya S., Kavzoglu T., Tonbul H. (2014, December 15-19). Detection of Burn Area and Severity with MODIS Satellite Images and Spatial Autocorrelation Techniques. In *AGU Fall Meeting*, San Francisco/USA.

Kim M., Madden M., Warner T. (2009). Forest type mapping using object-specific texture measures from multispectral Ikonos imagery: segmentation quality and image classification issues. *Photogrammetric Engineering & Remote Sensing*, 75(7), 819-829.

Mitri G., Gitas I. (2002, November 18-23). The development of an object-oriented classification model for operational burned area mapping on the Mediterranean island of Thasos using Landsat-TM images. In *Proceedings of IV International Conference on Forest Fire Research / Wildland Fire Safety Summit*, Coimbra/Portugal.

Mitri G., Gitas I. (2004). A performance evaluation of a burned area object-based classification model when applied to topographically and non-topographically corrected TM imagery. *International Journal of Remote Sensing* 25, 2863-2870.

Pleniou M., Koutsias N., (2013). Sensitivity of spectral reflectance values to different burn and vegetation ratios: A multi-scale approach applied in a fire affected area. *ISPRS Journal of Photogrammetry and Remote Sensing* 79, 199-210.

OGM. (2014). Turkish General Directorate of Forestry Reports; Available from: http://www.ogm.gov.tr/ekutuphane/FaaliyetRaporu/Forms/AllItems.aspx. Accessed: 2016.04.05.

Trimble (2011). *eCognition developer 8.64.0* reference book. München: Germany Trimble Documentation.

van Leeuwen W., Casady G., Neary D., Bautista S., Alloza J., Carmel J., Wittenberg L., Malkinson D., Orr B., (2010). Monitoring post-wildfire vegetation response with remotely sensed time series data in Spain, USA and Israel. *International Journal of Wildland Fire* 19 (1), 75-93.

Veraverbeke S., Harris S., Hook S. (2011). Evaluating spectral indices for burned area discrimination using MODIS/ASTER (MASTER) airborne simulator data. *Remote Sensing of Environment.* 115(10), 2702-2709.

Irregularly Sampled Data in Space and Time: Using Poisson Kriging to Reduce the Influence of Uncertain Observations in Assessing the Risk of Aflatoxin Contamination of Corn in Southern Georgia, USA

Ruth Kerry[1], Brenda Ortiz[2], Ben Ingram[3], Brian Scully[4], EunHye Yoo[5]

[1]Brigham Young University, USA
[2]Auburn University, USA
[3]Universidad de Talca, Chile
[4]USDA-ARS, USA
[5]University at Buffalo, SUNY, USA

*Corresponding author: ruth_kerry@byu.edu

I. INTRODUCTION

Aflatoxin is a mycotoxin produced by fungi (*A. flavus* or *A. parasiticus*) which can contaminate several staple crops such as corn (Payne, 1992) and can cause liver cancer in humans and animals (Barrett, 2005). The Food and Drug administration office of the USA have set a limit of 20 ppb, total Aflatoxin, for interstate commerce of food and feed (FDA, 2015). Infection of corn with *A. flavus* or *A. parasiticus* is driven by high temperature and drought conditions (Guo et al. 2008) associated with particular climatic areas (Abbas et al. 2007), agro-ecological zones (Setamou et al. 1997) and soil types (Palumbo et al. 2010). In the Southeast of the U.S.A, corn planted as a summer crop is highly susceptible to Aflatoxin contamination. Rainfall variability and high temperatures in this region during summer, along with light textured soils can induce water stress and contamination. Also, lack of irrigation infrastructure in some areas further aggravates the situation (Brenneman et al. 1993). Salvacion et al. (2011) found that June maximum temperatures and precipitation were key predictors of Aflatoxin levels in southern Georgia (GA), USA.

There are several methods for accurate Aflatoxin measurement, but most are time-consuming and expensive (Papadoyanis, 1990) and conducted at harvest which does not allow implementation of in-season adaptation strategies to reduce the risk. Given the expense of Aflatoxin measurement, an important goal for agricultural extension services is to identify those years and counties most at risk of contamination to reduce unnecessary expense on testing in years and areas when there is little risk of contamination. Identifying such years and counties would also allow adaptation of management strategies in season to reduce contamination risk and identify areas where more resistant varieties (Chen et al. 2002) of corn should be planted or irrigation infrastructure improved.

This paper aims to use Poisson kriging to give a space-time summary of Aflatoxin contamination data collected irregularly, both spatially and temporally, over a 27 year period (1977-2004) in 53 counties in southern GA. This will then be used to evaluate a risk factors approach to identify counties with the greatest risk of Aflatoxin contamination based on weather, soil and crop variables. Due to the irregular sampling in space and time, and highly skewed data approaching a Poisson distribution typical geostatistical methods are unstable for analysing spatial and temporal patterns in this data.

II. METHODS

Data Collection

From 1977-2004, corn samples were collected at harvest using a grab sampling technique with an average of 3 replications in 53 counties in southern GA (Fig. 1a) for Aflatoxin assessment by the Plant Pathology Department, University of Georgia at Tifton, GA. Data was collected for 23-45 counties each year but exact locations within counties was not recorded. For all years combined there was a total of 705 observations and these data approached a Poisson distribution (Fig. 1b).

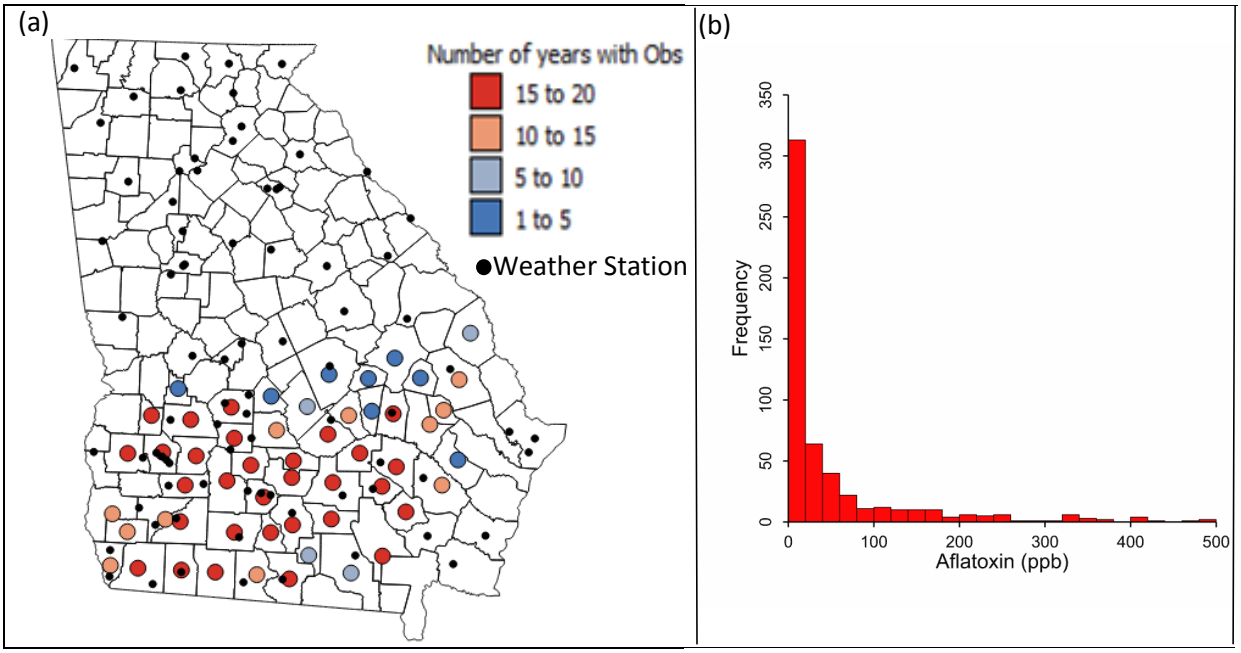


Figure 1. (a) Counties in Southern Georgia where Aflatoxin was measured, the proportion of years measurements were made and the location of weather stations (b) Frequency distribution of Aflatoxin values measured in southern Georgia 1977-2004.

Based on the findings of Salvacion et al. (2011), monthly maximum temperatures for June (June Tmax, °C) and June rainfall data (June RF, mm) were obtained for each year 1977-2004 from the Georgia Weather Network (http://georgiaweather.net). Fig. 1a shows the location of weather stations within GA. As all counties do not have a weather station, some have more than one, the weather stations are not located at the center of the county and have different installation dates, the weather variables were ordinary kriged (OK) to county centroids and a 1 km grid.

The area planted with corn per county was determined using The CropScape - Crop Land data layer produced by the National Agricultural Statistics Service (NASS, http://nassgeodata.gmu.edu/CropScape/). Unfortunately this information was only available from 2008-2009 onwards and so does not coincide with the period of Aflatoxin collection. So this and the soils data derived using it are quite uncertain and may not give an exact reflection of the proportion of each county under corn production because it assumes that the same rotations are used and that the proportion of agricultural land per county has not greatly changed over time.

A geo-corrected 1:250,000 map of soil associations (NRCS, 2006) was simplified and used to generate a map with 3 drainage classes: excessively, well and poorly drained soil. The percentage of

land areas with soil in each drainage class in the corn growing area (as identified using Cropscape above) was calculated for each county.

Geostatistical Analysis

Making sense of the Aflatoxin data with geostatistical methods was difficult. If a map of Aflatoxin contamination for each year for all counties in southern GA is to be produced then a variogram must be computed with an average of 37 data points and as few as 23. Variograms for individual years were unreliable and showed little spatial structure. This is typical of highly skewed (Kerry and Oliver, $2007_{a,b}$) and sparse data (Webster and Oliver, 1992). Poisson kriging (Monestiez et al. 2006) is ideal for data with a Poisson distribution and that have been irregularly observed in space or time. The proportion of years a county had Aflatoxin levels > 20ppb and > 100 ppb were Poisson kriged to county centroids and also to a 1 km grid. As proportions have a numerator and a denominator, the numerator was the number of years Aflatoxin levels were above one of the thresholds in a given county and the denominator was the number of years Aflatoxin was measured in that county. The influence of proportions for counties with fewer observations is down-weighted during variogram computation using the following weighted estimator

$$\hat{\gamma}_{Rv}(\mathbf{h}) = \frac{1}{2\sum_{\alpha,\beta}^{N(\mathbf{h})} \frac{d(v_\alpha)d(v_\beta)}{d(v_\alpha)+d(v_\beta)}} \sum_{\alpha,\beta}^{N(\mathbf{h})} \left\{ \frac{d(v_\alpha)d(v_\beta)}{d(v_\alpha)+d(v_\beta)} \left[r(v_\alpha) - r(v_\beta) \right]^2 - m* \right\} \quad (1)$$

where $N(\mathbf{h})$ is the number of pairs of counties ($v\alpha$,$v\beta$) whose denominator weighted centroids are separated by the vector **h,** and $m*$ is the denominator-weighted mean of the N area ratios. The usual squared differences, $[r(v_\alpha)- r(v_\beta)]^2$, are weighted by a function of their respective denominator sizes, $d(v_\alpha)d(v_\beta)/[d(v_\alpha)+d(v_\beta)]$, which gives more importance to more reliable data pairs based on larger denominators. Poisson kriging is a form of kriging with non-systematic errors and is parametric, modelling the noise attached to each observation with a Poisson distribution. Observations with small denominators receive less weight in kriging, by adding an error variance term to the diagonal of the kriging system (Monestiez *et al*. 2006, Goovaerts, 2005). Poisson kriging was done in SpaceStat (BioMedware, 2013).

Risk Factors Approach

By applying OK, risk factor data (June TMax, June RF, % Corn and Soil Type) were generated for each county and the nodes of a 1 km grid. The OK data was used to determine if risk factors exceeded a certain threshold converting it into indicator data (0/1). Table 1 shows the thresholds chosen for each variable. The thresholds for June TMax and June RF were chosen with respect to 30-year normals in southern GA and values receiving a (1) show hotter or drier than normal years. The indicator thresholds for other risk factors were determined based on natural marked breaks in the frequency distribution or were based on values associated with the tails of a normal distribution.

Once the number of risk factors above the specified threshold for each county in each year was determined, the relationship between these data and the Poisson kriged Aflatoxin data was assessed. This suggested broad groupings of years and counties with different levels of Aflatoxin contamination risk. These broad groupings were used to define grouping variables for Mann-Whitney U and Kruskal-Wallis H comparison tests to determine if there were significant differences in Aflatoxin levels based on the thresholds identified by the risk factors approach.

Table 1. Threshold values used for Risk Factor Indicators

Risk Factor	Threshold for Indicator (1/0)
June monthly maximum temperature (°C)*	>33°C
June monthly Rainfall (mm)*	<50 mm
Percent of county area growing corn (%)	>1.75%
Percent of county with well-drained soils (classes 1-4) (%)	>90%
Percent of county with excessively drained soils (classes 1-2.5)	>2.5%
Percent of years with 2 weather risk factors	>30 %

*Thresholds for June Tmax and June RF were chosen with respect to 30-year normal Tmax and RF in the area to show hotter and drier years than normal

III. RESULTS AND DISCUSSION

In trying to summarize the risk of Aflatoxin contamination in space and time and to verify if the risk factor approach proposed in this work is useful, an obvious starting point is to calculate summary statistics for each year and county. The means (not shown) were influenced by the maximum value and this was particularly pronounced in counties where a smaller proportion of observations had been made. This suggests that there can be great spatial and temporal variability in the Aflatoxin levels and that the small number problem is an issue with data for counties where few observations have been made giving very high or low Aflatoxin contamination levels.

Temporal patterns in Aflatoxin risk are mostly associated with weather variables, specifically June TMax and RF (Salvacion et al. 2011) which change from year to year and can help identify the specific years at greatest risk of contamination. The temporal results are not shown here but will be shown in the associated presentation to illustrate this point.

Analysis of Risk Data - Spatial Patterns

Analysing spatial patterns in risk of Aflatoxin contamination to identify the counties at greatest risk is more complicated than temporal analysis as it requires analysis of factors that are relatively stable in time (% corn and soil type) as well as weather variables. Fig. 2 a, c, e show the patterns in percent corn, well- and excessively-drained soil and Fig. 2 b, d, f show the indicators produced from these data using the thresholds in Table 1. Based on % corn, there is greatest risk (1) in south western GA (Fig. 2b) for well-drained soil risk is greatest in the west (Fig. 2d) and for excessively-drained soil the north is the highest risk area (Fig. 2f). Fig. 3a shows a map where the two weather risk factors (June TMax and June RF) have been combined and the proportion of years with two weather risk factors has been kriged to the 1 km grid. This gives a temporal summary of the areas most prone to drought in southern GA. When converted to an indicator based on the thresholds in Table 1, Fig. 3b shows the areas with a high (1, >30% of years) risk of drought. Fig. 3c shows the percentages of years with three or more risk factors for Aflatoxin contamination where June Tmax and RF are considered separate risk factors. The map has similarities to Figs. 2f and 3b suggesting that weather and excessively drained soils are the greatest risk factors for Aflatoxin contamination. Nevertheless, crucial importance of both weather factors is shown by the striking similarity in the patterns shown in Fig. 3a and b which show drought summary and Fig. 3e and f which show the Poisson kriged summary of Aflatoxin levels in all years.

 [ISBN: 978-2-9105-4510-5]

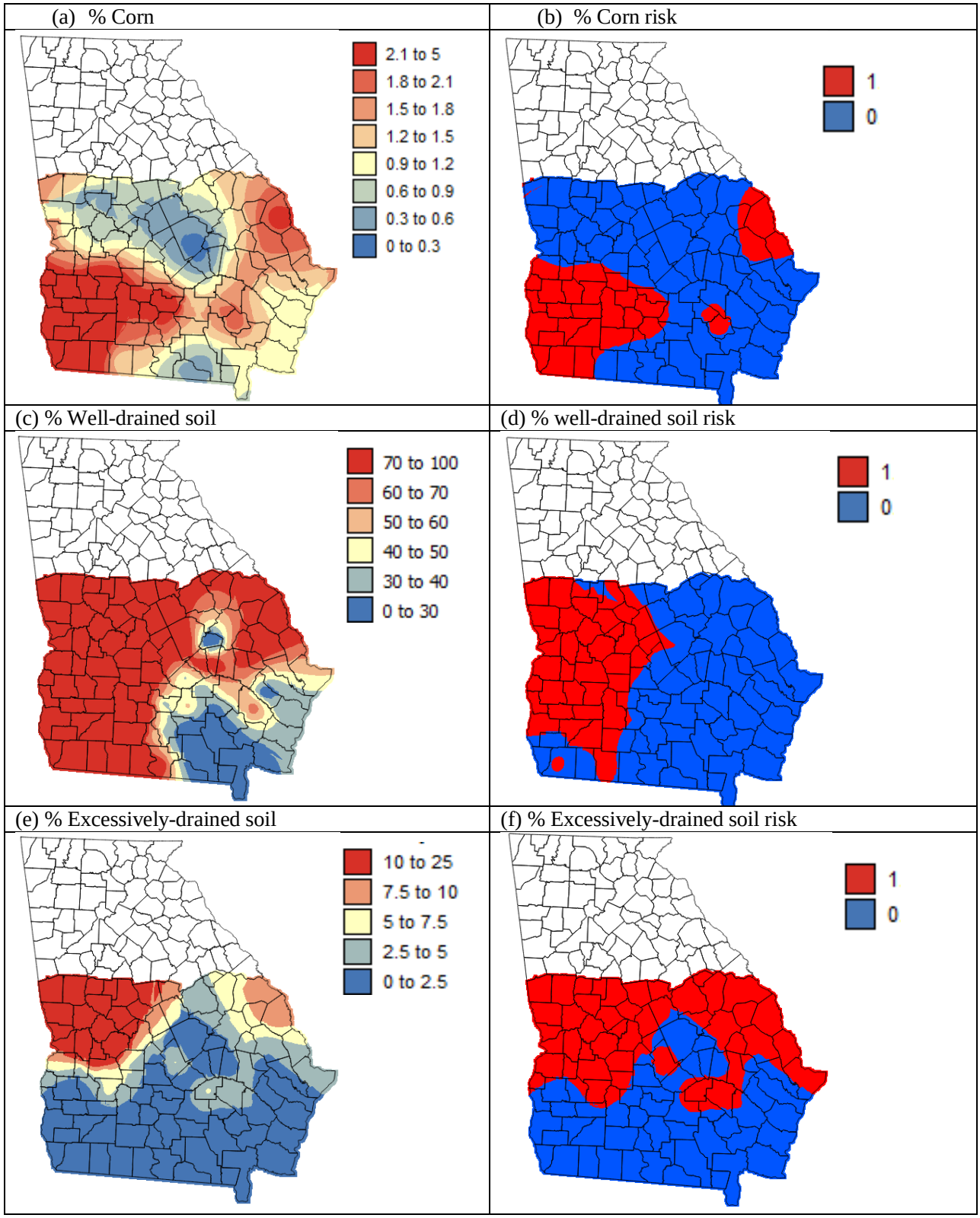


Figure 2. Risk factors and their associated indicators plotted kriged to a 1 km grid (a-b) percent of area growing corn, (c-d) percent of area with well-drained soils, (e-f) percent of area with excessively drained soils

 [ISBN: 978-2-9105-4510-5]

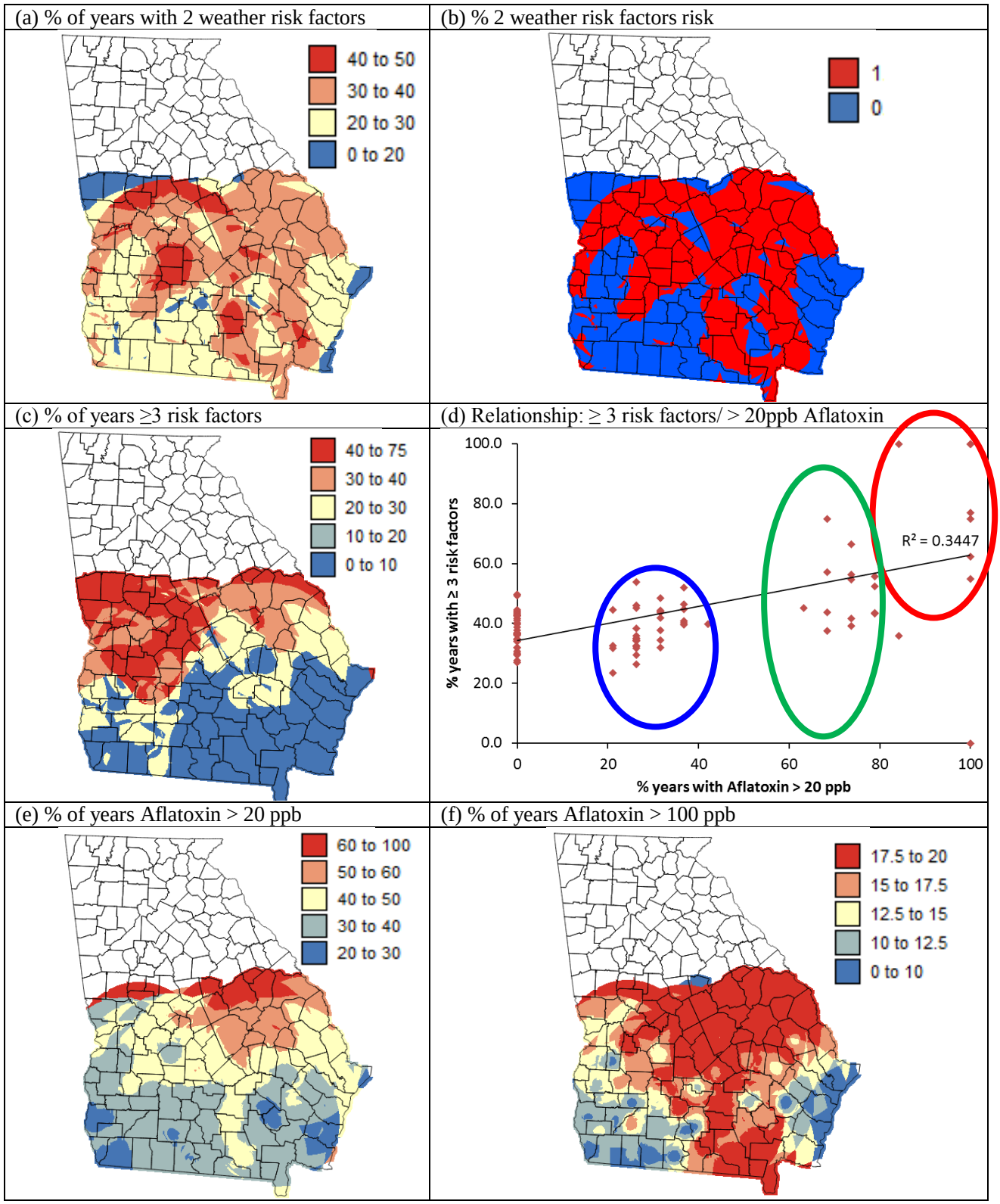


Figure 3. (a) Percent of years with 2 weather risk factors and (b) associated indicator, (c) percent of years with ≥ 3 risk factors, (d) Relationship between percent of years with ≥ 3 risk factors and percent of years and >20 ppb Aflatoxin, (e) percent of years with Aflatoxin > 20 ppb and (f) percent of years with Aflatoxin > 100 ppb.

Fig. 3d shows the relationship between the proportion of years with ≥3 risk factors and proportion of years with Aflatoxin levels >20ppb for each county. The correlation coefficient for this relationship was 0.59 which is significant at $p<0.001$. Distinct groupings of counties are visible in the plot and have been circled. Such grouping were not very well defined when just weather factors were considered and the correlation coefficient was lower, suggesting that the other risk factors (% corn and soil drainage types) help to distinguish the spatial differences between counties. The higher risk counties, circled in red and green in Fig. 3d, are the northern most counties in southern GA as well as those in the central area of the southern half of the state.

Comparison tests were performed to determine whether there were significant differences in Aflatoxin levels defined in particular ways based on risk levels identified from the scatter-graph produced (Fig. 3d) with the risk factor approach. Counties at 'high risk' of Aflatoxin contamination (with > 3 risk factors) were compared with those not at high risk (< 3 risk factors) by Mann-Whitney U tests. When average Aflatoxin levels were used for comparison there was no significant difference ($p=0.569$) but when Poisson kriged % years with > 20ppb and 100 ppb Aflatoxin data were used for comparison there were significantly higher, $p=0.002$ and $p=0.012$, proportions of years with Aflatoxin levels exceeding thresholds for 'high' risk years. Kruskal Wallis H tests comparing all risk classes identified in Figure 3d also showed more significant differences ($p<0.001$) and the expected order of classes when Poisson kriged data were used rather than average Aflatoxin data. This shows that average Aflatoxin levels do not give a good summary of the counties most and least at risk of Aflatoxin contamination while the Poisson kriged data which down-weight the influence of proportions based on low numbers of observations, do give a good summary.

IV. CONCLUSIONS

This study showed that when data have been irregularly sampled in space and time and have a Poisson distribution, Poisson kriging is a reliable way to generate a temporal summary of spatial patterns. Simple averages were shown to be unreliable where fewer observations were made and standard geostatistical methods do not work well when data have a Poisson distribution or have few data for individual years. Comparison tests showed that counties and years identified as having the greatest risk levels using the risk factors approach did have significantly higher Aflatoxin levels. This was not the case however, for average Aflatoxin due to unreliable averages because of irregular sampling. Identifying the weather conditions and counties associated with the highest contamination risk will allow for in-season adaptation strategies such as irrigation to avoid drought as temperatures and rainfall in June are carefully monitored with respect to 30 year normals. Also, testing can be focused in the highest risk counties and very little expensive Aflatoxin testing will be needed in low risk years and the highest risk counties could plant more resistant varieties of corn.

Future work should investigate including new variables in the risk factors approach, fine tuning of the thresholds of existing variables and quantifying the relative uncertainty of predictions associated with each risk factor. Using OK risk factor data means that uncertainty would obviously be greatest where distance from sampling points is greatest and for data like % corn where data from a different time period to sampling had to be used. The data kriged to a 1 km grid shows the potential for identifying high risk areas at the sub-county level. This should be more reliable when sub-county data on in-season corn NDVI values are used as an indicator of in-season drought stress in the risk factors approach. An online interactive Aflatoxin risk assessment tool that uses the risk factors approach outlined here is currently being developed and will include NDVI data. There is the potential that such an online tool could be adapted to other crops, states and even farms so that Aflatoxin levels may be better managed.

References

Abbas, H.K., Shier, W.T and Cartwright, R.D. (2007) Effect of temperature, rainfall and planting date on aflatoxin and fumonisin contamination, in commercial Bt and non-Bt-corn hybrids in Arkansas, *Phytoprotection*, 88, 41-50.

Barrett, J.R. (2005) Liver Cancer and Aflatoxin: New Information from the Kenyan Outbreak. *Environmental Health Perspectives.* 113, A837–A838.

BioMedware, Inc. (2013) *SpaceStat User Manual version 3.6*, BioMedware, Inc. Available at: http://www.biomedware.com/files/documentation/spacestat/default.htm

Brenneman, T.B., Wilson, D.M. and Beaver, R.W. (1993) Effects of diniconazole on aspergillus populations and aflatoxin formation in peanut under irrigated and non-irrigated conditions, *Plant Disease*, 77, 608-612.

Chen, Z.Y., Brown, R.L., Damann, K.E. and Cleveland, T.E. (2002) Identification of unique or elevated levels of kernel proteins in aflatoxin-resistant maize genotypes through proteome analysis, *Phytopathology*, 92, 1084-1094.

FDA (2015) http://www.fda.gov/ICECI/ComplianceManuals/CompliancePolicyGuidanceManual/ucm074703.htm accessed November 2015.

Goovaerts, P. (2005) Geostatistical Analysis of Disease Data: Estimation of Cancer Mortality Risk from Empirical Frequencies Using Poisson Kriging. *International Journal of Health Geographics*. 4, 31.

Guo, B., Chen, Z., Lee, R.D and Scully, B.T. (2008) Drought stress and preharvest aflatoxin contamination in agricultural commodity: Genetics, genomics and proteomics, Journal of *Integrative Plant Biology*, 50, 1281-1291.

Kerry, R. and Oliver, M.A. (2007) Determining the Effect of Skewed Data on the Variogram. I. Underlying Asymmetry. *Computers & Geosciences*, 33, 1212-1232.

Kerry, R. and Oliver, M.A. (2007) Determining the Effect of Skewed Data on the Variogram. II. Outliers. *Computers & Geosciences,* 33, 1233-1260.

Monestiez, P., L. Dubroca, E. Bonnin, J. P. Durbec, and C. Guinet. (2006). Geostatistical Modelling of Spatial Distribution of Balaenoptera Physalus in the Northwestern Mediterranean Sea from Sparse Count Data and Heterogeneous Observation Efforts. *Ecological Modelling,* 193, 615–28.

NRCS (2006) Digital General Soil Map of U.S., U.S. Department of Agriculture, Natural Resources Conservation Service, Fort Worth, Texas. Online_Linkage: URL:http://SoilDataMart.nrcs.usda.gov/

Palumbo, J.D., O'Keeffe, T.L., Kattan, A., Abbas, H. K. and Johnson, B.J. (2010) Inhibition of Aspergillus flavus in Soil by Antagonistic Pseudomonas Strains Reduces the Potential for Airborne Spore Dispersal, *Phytopathology*, 100, 532-538.

Payne, G.A. (1992) Aflatoxin in maize, *Critical Reviews in Plant Sciences*, 10, 423-440.

Salvacion, A. R., Ortiz, B. V., Scully, B. T., Wilson, D. M., Fraisse, C. W., Hoogenboom, G., Lee, R. D. (2011) Modeling Probability of Corn Aflatoxin Contamination Using Drought Index in South Georgia. *South East Climate Consortium Poster*

Setamou, M., Cardwell, K.F., Schulthes, F. and Hell, K. (1997) Aspergillus flavus infection and Aflatoxin contamination of preharvest maize in Benin, *Plant Disease*. 81, 1323-1327.

Webster, R., Oliver, M.A. (1992) Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*. 43, 177–192.

A comparison of optimal map classification methods incorporating uncertainty information

Yongwan Chun[*1], Hyeongmo Koo[1], Daniel A. Griffith[1]

[1] University of Texas at Dallas, USA

*Corresponding author: ywchun@utdallas.edu

Abstract
Uncertainty in spatial data attributes can produce unreliable spatial patterns in choropleth maps, but only a few studies have considered uncertainty in map classification processes. Unfortunately, a less desirable classification result often is generated by existing methods. For example, most observations are assigned to a single class while the remaining classes have a very small number of observations allocated to them. Also, selection of proper criteria for an optimal map classification is difficult. The purpose of this paper is to expand the discussion about incorporating data uncertainty for map classification by extending optimal map classification strategies with Bhattacharyya distance. The proposed method is illustrated with an application of soil lead contamination measurements in the City of Syracuse.

Keywords
Uncertainty, Map classification, Choropleth map

I INTRODUCTION

Most spatial data attributes inevitably contain uncertainty due to sampling and/or measurement error, among other sources of error (e.g., specification). The preferred form of visualization for these data is uncertainty as well as attribute values, which is limited because of existing technical and conceptual deficiencies. One common approach is to utilize a bivariate mapping technique that simultaneously represents estimates with their corresponding uncertainty information using additional visual variables. For example, using Bertin's (1983) graphic variables, such as color value and texture, with a general choropleth map (e.g., MacEachren et al., 1998, 2005; Xiao et al., 2007).

Sun et al. (2014) propose a class separability measure to incorporate uncertainty information in a map classification to produce more reliable spatial patterns. This approach produces a more elaborated classification result. But it is heavily affected by outliers, and, at worst, results in a number of classes with a single observation, allocating most observations to a single class. Sun et al. (2016) further propose a heuristic approach to overcome this imbalance issue of the class separability classification results; but untrained users still might find selecting values of criteria for achieving an optimal map classification to be difficult.

The purpose of this paper is to expand the discussion about incorporating data uncertainty for choropleth mapping. More specifically, this study incorporates optimal map classification strategies with Bhattacharyya distance.

II METHOD

Bhattacharyya distance is effective to quantify dissimilarities between two probability distributions (Bhalerao and Rajpoot, 2003), resulting in it commonly being used in feature selection and extraction (e.g., Choi and Lee, 2003; Reyes-Aldasoro and Bhalerao, 2006). When two observations conform to a normal distribution, Bhattacharyya distance between the two observations can be calculated with the following equation (Coleman and Andrews, 1979):

$$D_B(i,\ j) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{SE_i^2}{SE_j^2}+\frac{SE_j^2}{SE_i^2}+2\right)\right)+\frac{1}{4}\left(\frac{\left(\bar{x}_i-\bar{x}_j\right)^2}{SE_i^2+SE_j^2}\right) \quad (1)$$

where $\bar{x}_i$ and $\bar{x}_j$ are the estimates and SE_i and SE_j are corresponding standard errors for observations i and j. Bhattacharyya distance can index the dissimilarity between two observations having the same estimate with the first term of equation (1). In general, an optimal classification can be achieved by minimizing within class Bhattacharyya distance. A within class distance may be determined in two different ways. The first is the sum of pairwise distances within a class (SB), and the other is the maximum pairwise distance within a class (MB). Although the first method calculates a measure from all pairwise distances, the second method intends to control for the worst case to improve homogeneity within a class. Class breaks can be determined in order to minimize the sum of costs, which are within class distances here.

This study utilizes optimal classification methods incorporating uncertainty information by extending the optimal classification method developed by Cromley (1996). He shows that an optimal univariate map classification can be modelled as an analogy to a constrained shortest path problem, when an acyclic network is constructed by the rank ordered estimates of observations. In this network, a node represents an observation, and lines correspond to assigning observations to a class. Here, a cost (or impedance) of a line is defined by within class Bhattacharya distance. An optimal classification can be constructed by identifying a combination of lines minimizing a total cost.

Extending Cromley (1996), optimal map classification can be formulated as follows:

Minimize:	$\sum_{i,\ j} c_{i,j} d_{i,\ j}$	$\forall\ i,\ j\ \in N, i \le j$	(2)
Subjected to:	$\sum_i d_{i,\ k} = \sum_j d_{k,\ j}$	$\forall i\ \in In_k;\ j \in Out_k$	(3)
	$\sum_j d_{s,\ j} = 1$	$\forall j \in Out_s$	(4)
	$\sum_i d_{i,e} = 1$	$\forall i \in In_e$	(5)
	$\sum_{i,\ j} d_{i,j} = T$	$\forall\ i,\ j\ \in N, i \le j$	(6)
	$d_{i,\ j} \in \{0,1\}$	$\forall\ i,\ j\ \in N, i \le j$	(7)

where $d_{i,\ j}$ is a binary decision variable, $c_{i,j}$ is the cost of a line from i to j, In_k is a set of lines that terminate at node k, Out_k is a set of all lines that originate at node k, and nodes s and e respectively are the minimum and maximum values. Equation (2) is the objective function, and equation (3) ensures that all observations are assigned to a class. Equations (4) and (5) ensure that the first node is assigned to the first class, and the last node to the last class. Equation (6) constrains the number of classes, and equation (7) is a binary integer restriction.

III APPLICATION

The proposed method is illustrated with lead (Pb) contamination measurements collected from soil samples across the City of Syracuse (Griffith, 2008). These measurements are in milligrams per kilogram of soil (ppm), and are log-transformed so that their frequency distribution better conforms to a bell-shaped curve. A Pb surface was krigged with a Bessel function semivariogram model (26,402 pixel values were interpolated), and then the krigged surface was aggregated to 264 grid cell polygon values. Predicted measurements and prediction standard errors for these grid cells are utilized for map classification purposes. The proposed methods have been implemented as an extension of ArcGIS 10.1 using C# in the Microsoft .Net Framework 4. The optimization problems are solved by a branch-and-bound algorithm using Gurobi optimizer 6.5.0.

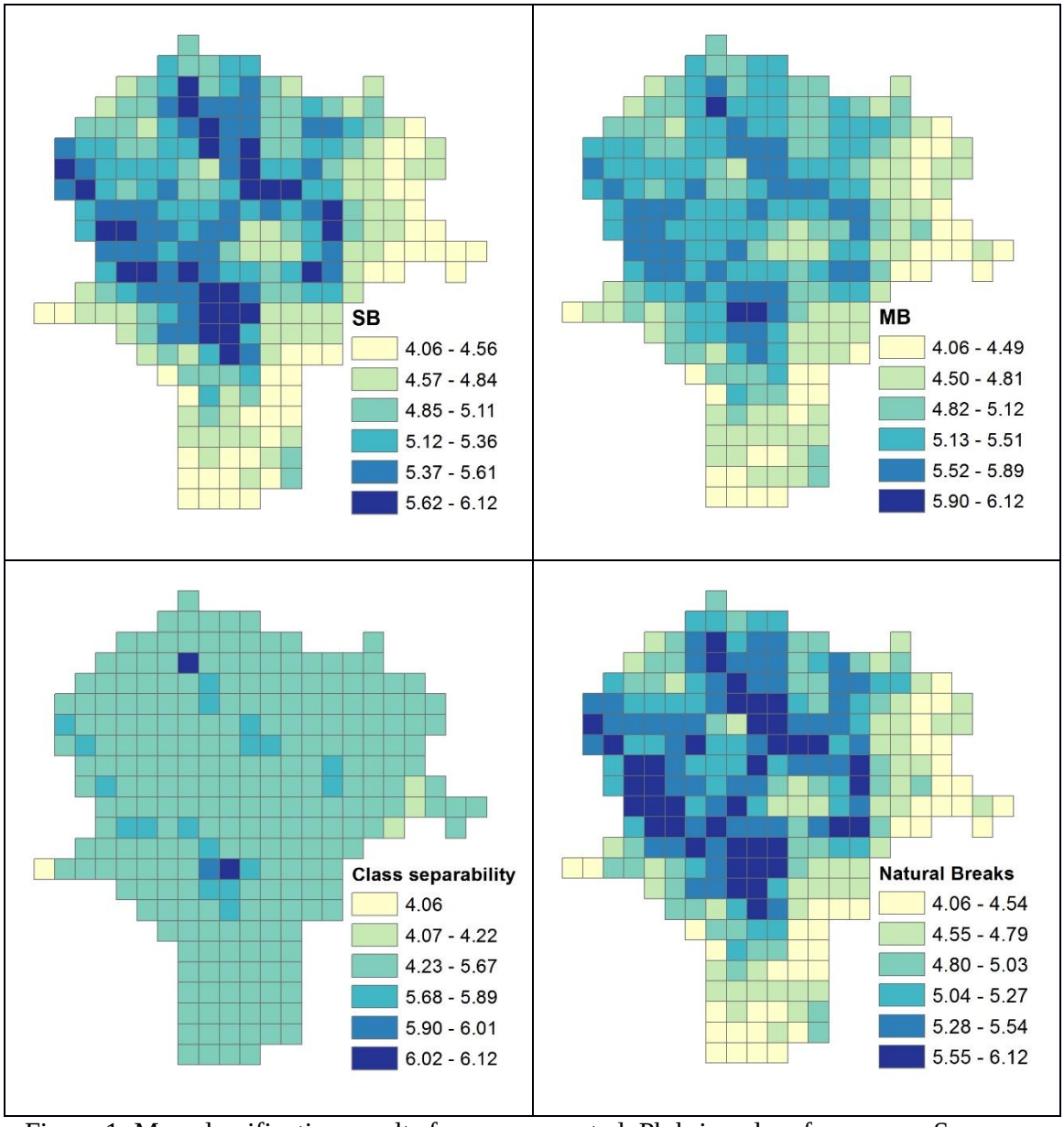


Figure 1: Map classification results for an aggregated Pb krigged surface across Syracuse

The classification results are compared with those from the class separability method by Sun et al. (2014), and from the standard natural breaks classification method. Figure 1 displays map

classification results for Pb levels. The class separability method produces a spatial pattern that is remarkably different from the other results. This method produces classes that contain only outliers on both sides of its frequency distribution. The other three map classifications show a relatively similar spatial pattern, although they have a noticeable difference for the last class. The last class by the natural breaks method has the greatest number of observations among the map classification results. In contrast, class 6 in MB has only three observations.

Table 1 presents the numbers of observations for output classes. The results of the class separability method display considerable variation in terms of the number of observation counts across classes. More specifically, most of the observations are assigned to class 3. However, classes 1 and 5 have only a single observation. In contrast, proposed classification methods have a more balanced number of observations across classes. The natural breaks classification also produces well balanced results, but this method considers only variances among estimates without a consideration of uncertainty (Jenks, 1977).

Table 1. The number of observations in classification result classes

	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Class 6**
SB	43	56	46	48	44	27
MB	32	59	55	76	39	3
Class separability	1	3	239	18	1	2
Natural breaks	38	50	42	43	50	41

IV DISCUSSION

This study extends map classification by integrating an optimal map classification (Cromley, 1996) and uncertainty information with various costs (e.g., distance here) and objective functions. Generally, the proposed methods produce homogenous classes by their respective criteria, and also achieve visually balanced classification results. Some limitations should be investigated in future studies. First, a performance evaluation for the proposed methods needs to be investigated. Because widely used evaluation methods examine only estimates (e.g. Jenks and Caspall, 1971; Xiao et al., 2007), these methods are not effective for evaluating the performance of the proposed map classification methods. Second, the proposed map classification methods are solved using a branch-and-bound algorithm. Thus, as the number of nodes increases, the amount of time needed to find an optimal solution tends to increase often at roughly an exponential rate.

V ACKNOWLEDGEMENTS

This research was supported by the National Institutes of Health, grant 1R01HD076020-01A1; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Institutes of Health.

References

Bertin, J. (1983). *Semiology Of Graphics: Diagrams, Networks, Maps*. Madison: University of Wisconsin press.

Bhalerao, A. H., Rajpoot, N. M. (2003). Discriminant feature selection for texture classification. *In Procedings Of The British Machine Vision Conference 2003*. United Kingdom.

Choi, E., Lee, C. (2003). Feature extraction based on the Bhattacharyya distance. *Pattern Recognition*, 36(8), 1703–1709.

Coleman, G. B., Andrews, H. C. (1979). Image Segmentation by clustering. *In Proceedings of IEEE*, 67, 773–788.

Cromley, R. G. (1996). A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International Journal of Geographical Information Systems*, 10(4), 405–424.

Griffith, D. A. (2008). Geographic sampling of urban soils for contaminant mapping: how many samples and from where. *Environmental Geochemistry and Health*, 30, 495-509.

Jenks, G. F. (1977). *Optimal Data Classification for Choropleth Maps*. Occasional Paper No. 2, Department of Geography, University of Kansas.

Jenks, G. F., Caspall, F. C. (1971). Error on choroplethic maps: definition, measurement, reduction. *Annals of the Association of American Geographers*, 61(2), 217–244.

MacEachren, A. M. A., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., Hetzler, E. (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information ScienceInformation Science*, 32(3), 139–160.

MacEachren, A. M., Brewer, C. A., Pickle, L. W. (1998). Visualizing georeferenced data: representing reliability of health statistics. *Environment and Planning A*, 30, 1547–1561.

Reyes-Aldasoro, C. C., Bhalerao, A. (2006). The Bhattacharyya space for feature selection and its application to texture segmentation. *Pattern Recognition*, 39(5), 812–826.

Sun, M., Wong, D. W., Kronenfeld, B. J. (2014). A classification method for choropleth maps incorporating data reliability information. *The Professional Geographer*, 67(1), 72–83.

Sun, M., Wong, D., Kronenfeld, B. J. (2016). A heuristic multi-criteria classification approach incorporating data quality information for choropleth mapping. *Cartography and Geographic Information Science*, 1–13.

Xiao, N., Calder, C. A., & Armstrong, M. P. (2007). Assessing the effect of attribute uncertainty on the robustness of choropleth map classification. *International Journal of Geographical Information Science*, 21(2), 121–144.

Scale effects in spatial accuracy

Getting the right spatial mix: optimising the size, type and location of renewable energy facilities

Alexis Comber*[1], Jen Dickiie[2]

[1] University of Leeds, Leeds, LS2 9JT, UK
[2] University of Stirling, Stirling, FK9 4LA, UK

*Corresponding author: a.comber@leeds.ac.uk

Abstract
Supply and demand modelling for facilities that require land related resources need to take into account the spatial distribution the resource. Using a Scottish case study, this short paper presents a method extension that to the p-median problem that identifies optimal locations and combinations of different types and sizes of land based biomass renewable energy facilities. Whilst there are many decision models and tools for siting other types of renewable energy (solar, wind, hydro), supply and demand tools decisions around land based biomass renewable energy do not exist. The p-median extension and optimisation allows the trade-offs between different decisions around land, for example related to agriculture, food security, biodiversity, flood risks, to be evaluated. In the context of Scotland, this methodology supports the many policy agendas around community initiatives, agri-renewables, circular economy, supply chains, local food agendas, carbon sequestration and green infrastructures.

Keywords
Supply and demand; renewable energy; biomass; land use

I INTRODUCTION

In much facility location analysis, the resources needed by a facility at potential locations are assumed to exist or that their supply is trivial. Much greater focus is given to identifying potential facility locations that meet the spatial distribution of demand, and where facility resources are considered these are typically done so from the perspective of transportation overheads, infrastructure and road networks.

Land based and biomass Renewable Energy (RE) facilities typically require consideration of the spatiality of supply as well as demand. The facilities require biomass feedstocks (inputs) that are usually derived from animal manures, crop residues, forestry / timber production. In the context of RE, as well as climate change mitigation, the objectives should also include net energy gains and minimal carbon impacts. Thus it is important to consider the resource catchments needed to supply RE facilities, rather than subsume financial transportation costs into site evaluation.

Recent work has extended the p-median problem to be able to handle the resource catchments needed to support facilities at potential locations in location allocation analyses (Comber et al., 2015) as well as demand. It has been further extended to locate multiple types of facility and to determine optima locations for pre-defined groups of different sized facilities which have

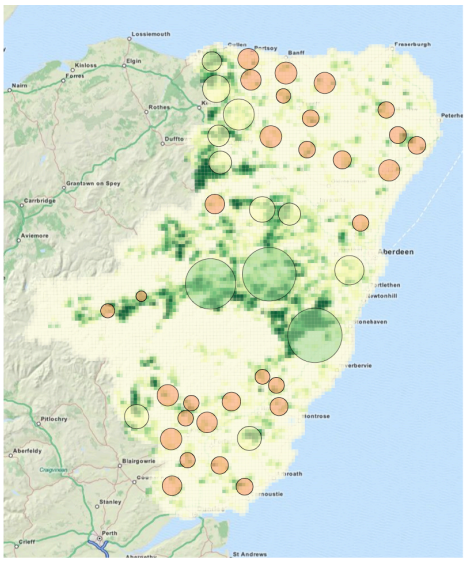
essentially imposed an optimal packing dimension onto the p-median problem (Comber et al, 2016).

This paper further extends these developments to the p-median problem to identify the optimal mix of facilities. The problem remains a packing problem but the extension seeks to identify the optimal mix as well as spatial arrangement of facilitates. The objective function remains the same (to minimise distances to demand and to forbid resource catchment overlap), but the search space is extended from sets of predefined facilities (e.g. 3 large, 10 medium and 20 small) to identifying the best combination of facilities for the study area being considered. The context for this extension is the need to accommodate competing land use demands, for example food security, flood defence or biodiversity.

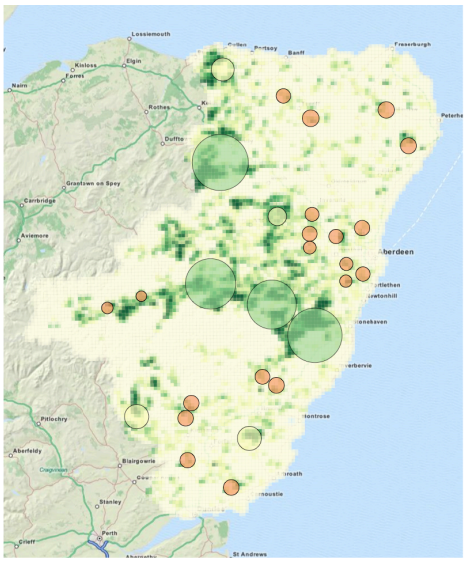
The spatial accuracy context of this work relates to the sensitivity of location–allocation algorithms and their associated evaluation functions to spatial consideration of resource catchments.

II INITIAL RESULTS

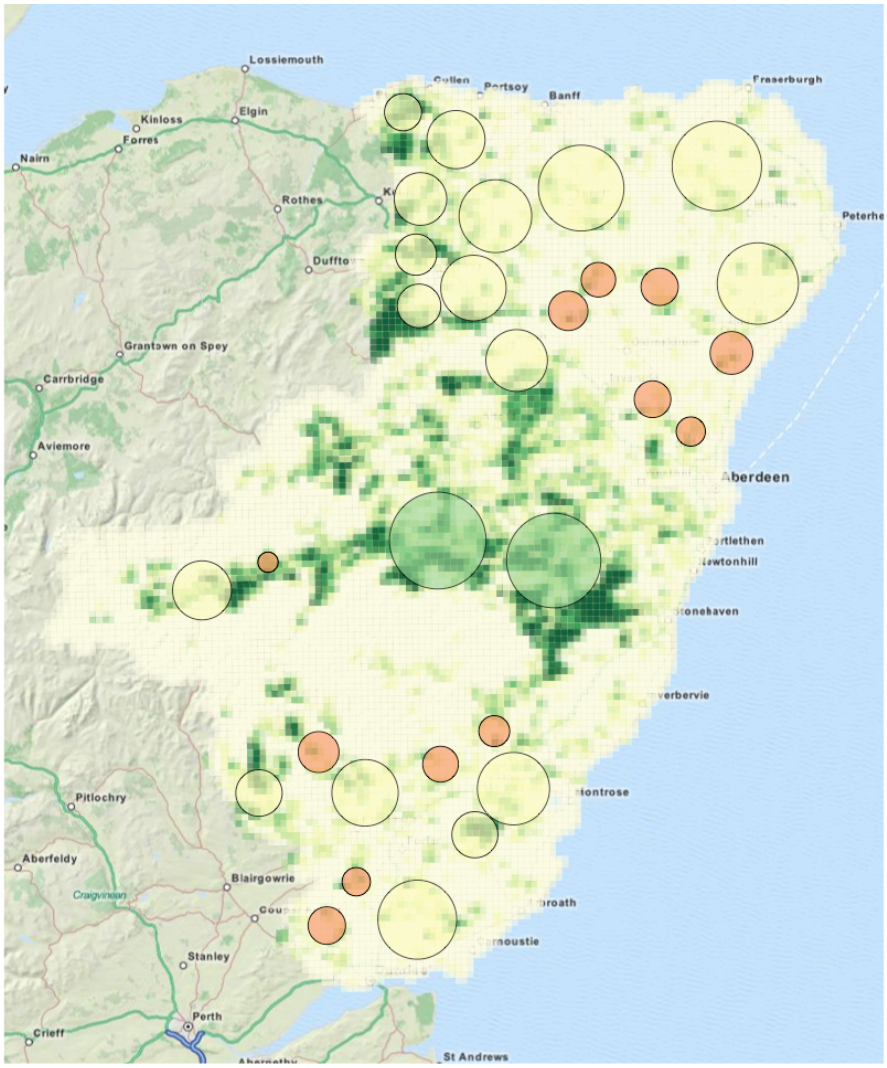
Consider the location, catchments and distribution of potential biomass RE facilities in Figure 1a along with the underlying spatial distribution of biomass from forestry. This shows an optimal arrangement for 3 Combined Heat and Power (CHP) units, with different capacities (energy outputs) and different input requirements for a case study in north east Scotland. The total potential amount of annual biomass from forestry is 1,895,677 t yr^{-1} and the mix of 3 20MW CHPs, 10 4MW CHPs and 30 1MW CHPs equates to a biomass feedstocks of 650,000 t yr^{-1} or 34% of the available supply and the production of 130MW.



a)



b)



c)

Figure 1: Optimal spatial arrangements for sets of biomass CHP facilities with the same overall generating capacity a) for a predefined mix of capacities, b) for a different mix of capacities, and c) the optimal, optimal mix

The question addressed by this paper is whether a different spatial arrangement of different mix of facilities sizes, such as suggested in Figure 1b, would achieve the same energy generating capacity output but use the biomass resource more efficiently, ie with less energy expended on transportation.

To achieve this, the algorithm has been further extended to search through *all possible combinations* of the 3 types of facility to achieve the same capacity. The optimal spatial configuration is shown in Figure 1c and suggests that the following combination of CHPs to achieve 120MW and minimising resource transportation costs is as follows: 2 x 20MW, 17 x 4MW and 12 x 1MW.

III DISCUSSION

There are a number of critical methodological issues and spatial accuracy considerations related to this work.

In terms of method development, this paper describes a further extension to the algorithm, first described in Comber et al., (2015), which accommodates the spatial distribution of the resources needed to supply the RE facility and satisfy the energy demand at any given location. It extended the p-median problem to prevent facility locations with overlapping resource

catchments from being selected. The first extension, described in Comber et al (2016), supported the allocation of a predefined number of different sizes of facility (3 small, 2 medium and 5 large, for example). However, this was based on a packing problem heuristic, which identified the location for n_1 largest facilities first, then the n_2 medium sized facilities and finally the n_2 largest facilities.

The extension reported here identifies the optimal mix of different sized facilities and considers all sizes of facility together. In each case the evaluation function was to minimise population distance to the resource supply. In this case this was done through a deterministic search. The resource catchment for each size of facility at each potential location was pre-computed and then the 'best' combination of n_1 20MW n_2 4MW and n_3 1MW facilities was determined. This was possible because of the relatively small number of potential sets of each $\{n_1, n_2, n_3\}$ combination. However for a larger study area, the number of potential combinations of facilities may preclude a deterministic search suggesting the need for search heuristics such genetic algorithms, or perhaps more pertinently to this type of study grouping genetic algorithms (Falkenauer, 1998) which have been shown to more effectively identify optimal groups sets of facilities than standard genetic algorithm (Comber et al., 2011).

In terms of the accuracy of the results generated by this method, they are subject to the usual considerations in any location-allocation application, namely the extent to which the evaluation function matches the problem specification and the reliability of the input data.

Here, optimality (the evaluation function) was determined by the degree to which potential facility locations minimised distance to population centres (ie they were not weighted by the population at those centres). There are two obvious areas to refine this function: first, to include population, for example, to generate demand weighted distances, reflecting differences in the spatial distribution of demand. However, this study used Output Area centroids as the demand locations. These are the smallest areal unit over which UK census data are reported (Martin, 1998) and were constructed such that they contain broadly an equal number of people (~350). Perhaps more interestingly data describing more nuanced measures of demand for RE could be used, such as the Call Credit Green and Ethical geo-demographic classification of the UK. Second, the object of any planning and policy in relation to renewable energy should be to minimise net carbon costs and maximise net carbon gains. This is rarely the case in the RE literature, which much effort seems to be spent on modelling transport costs in RE facility location decisions (eg Sultana and Kumar, 2012; Panichelli and Gnansounou, 2008; Sliz-Szkliniarza and Vogt, 2012). This is plainly daft as it renders the result of any analysis meaningless if there are radical changes in the price of oil. The optimal selection of suitable sites for RE facilities using these kinds of paradigms is critical if land based biomass resources are to be efficiently and maximally used to support a diverse range of objectives including food and energy security as well as environmental protection.

In summary the current method allows locations for multiple sizes of facilities to be evaluated. The algorithm could be applied to select optimal sites for multiple types as well as sizes of renewable energy facilities: CHPs, anaerobic digesters, gasification units etc. This would support truly holistic, strategic regional planning and well as community level energy initiatives. The latter are increasingly being supported in Scotland (eg http://www.localenergyscotland.org/cares).

The next steps in this work are multiple. First, to consider network distances to resources rather Euclidean distances, to explore how asymmetric, amorphous catchments may be incorporated, allowing them to fill the available space between already selected sites and to consider different combinations of feedstocks – domestic, forest and agricultural and household wastes – would could be used to refine the results of the analysis. Second, to develop methods for searching more efficiently, for example through heuristics, the use of *net*

energy gain as an evaluation function, the spatial distribution of demand that is receptive to RE and whether network distances to resources improve the outcomes.

This paper presents a framework for doing this: compute all possible combinations of RE supply, identify the resource catchments needed for each of individual facility at each potential location, and then develop a search through this very highly dimensional decision space. Future work will consider the development and application of evaluation functions that minimise the distances (and net energy and carbon costs) that resources have to travel to supply the facility as well as more nuanced measures of potential demand.

This research was funded by the EPSRC SECURE project, Reference EP/M008347/1 (http://www.gla.ac.uk/research/az/secure/).

References

Comber A, Dickie J, Elston D and Miller D (2016). What to put where? Extending the p-median problem to consider multiple facilities, multiple sizes and associated resource needs. Short paper AGILE 2016, Conference Helsinki, June 2016.

Comber A, Dickie J, Jarvis C, Phillips M and Tansey K. Locating bioenergy facilities using a modified GIS-based location-allocation-algorithm: considering the spatial distribution of resource supply. *Applied Energy*, 154: 309-316, 2015.

Comber, A.J., Sasaki, S., Suzuki, H. and Brunsdon, C., (2011). A modified grouping genetic algorithm to select ambulance site locations. *International Journal of Geographical Information Science*, 25(5): 807–823

Falkenauer, E., 1998. *Genetic algorithms and grouping problems*. London: John Wiley and Sons.

Martin, D., 1998. 2001 Census output areas: from concept to prototype. *Population Trends* 94, 19–24.

Panichelli L and Gnansounou E, (2008). GIS-based approach for defining bioenergy facilities location: A case study in Northern Spain based on marginal delivery costs and resources competition between facilities. *Biomass and Bioenergy*, 32: 289-300.

Sliz-Szkliniarz B and Vogt J, (2012). A GIS-based approach for evaluating the potential of biogas production from livestock manure and crops at a regional scale: A case study for the Kujawsko-Pomorskie Voivodeship. *Renewable and Sustainable Energy Reviews* 16: 752–763.

Sultana A and Kumar A, (2012). Optimal siting and size of bioenergy facilities using geographic information system, *Applied Energy*, 94: 192-201

Impact of compositional and configurational data loss on downscaling accuracy

Amy E. Frazier[1]* and Peter Kedron[2]

[1] Oklahoma State University, Stillwater, Oklahoma, USA
[2] Ryerson University, Toronto, Ontario, Canada

*Corresponding author: amy.e.frazier@okstate.edu

Abstract
Datasets collected at widely varying spatial scales are often merged to address questions related to global environmental change. Integrating data requires aggregation, which reduces data quality and introduces statistical biases collectively known as the Modifiable Areal Unit Problem (MAUP). These biases result from different forms of compositional and configurational data loss that occur during aggregation, but little is known about the relationship between data loss and MAUP biases for downscaling. This study uses the well-established process of landscape and surface metric scaling to examine how uncertainties related to the composition and configuration of land cover patterns propagate across scales when data are aggregated and ultimately impact downscaling results. Results suggest a link between compositional data loss and downscaling accuracy, particularly in the patch-based landscape paradigm. Further work is needed to determine if relationships exist between compositional and configurational data loss measures and downscaling error in the surface paradigm.

Keywords: scale and scaling; landscape ecology, spatial pattern metrics, heterogeneity, remote sensing

I. Introduction

In the age of data-driven science, diverse datasets collected at widely varying spatial scales are increasingly being merged to address questions related to global environmental change. However, integrating data collected at different spatial scales requires aggregation, which reduces data quality and introduces statistical bias. These biases, collectively known as the Modifiable Areal Unit Problem (MAUP), result from different forms of compositional and configurational data loss that occur during aggregation. MAUP biases can be minimized when aggregated data have a lower degree of heterogeneity (Holt et al. 1996; Steel and Holt 1996), but beyond this recognition, studies on the relationship between forms of data loss and MAUP biases remain limited.

Building on recent theoretical and methodological advancements in spatial science, remote sensing, and landscape ecology, we investigate how MAUP biases propagate across resolutions and whether the spread of bias, in the form of compositional and configurational data loss, can be used to forecast downscaling accuracy. Spatial pattern scaling provides a useful platform from which to examine these issues because research has established that several land cover pattern metrics exhibit consistent and robust scaling relationships across resolutions (Wu 2004). Yet, when these scaling relationships are extrapolated (i.e., downscaled) to predict metrics at a finer resolution, large errors typically result (Frazier 2014), likely from MAUP-driven aggregation biases (Frazier 2014, 2015a).

Until recently, examination of this hypothesis was limited to hard-classified landscape images, but the emergence of sub-pixel remote sensing classification techniques that preserve greater heterogeneity than their traditional, pixel-based counterparts have improved our ability to quantify data loss and preserve landscape heterogeneity, thereby opening the door for renewed investigations. Simultaneously, the proliferation of surface metrics in landscape ecology provides a means from which to compare this hypothesis across two different landscape paradigms: patch-based and surface. We use the well-established process of landscape and surface metric scaling (Turner et al. 1989; Wu 2004; Frazier 2015b) to examine how uncertainties related to the composition and configuration of land cover patterns

propagate across scales when data are aggregated and ultimately impact downscaling results. Predetermination of the impact of these biases on downscaling may eventually allow assessment of whether a landscape is a satisfactory candidate for downscaling.

II. Methods

Data were collected through a geographically stratified sampling of forest land cover in four forested ecoregions in the eastern United States (Omernik 1987). Within each ecoregion, we randomly sampled 125 20x20km plots from a continuous grid (Fig. 1a,b). We removed grid squares comprising urbanized areas greater than 500,000 people. We then clipped the national land cover map (NLCD), aggregated to parent classes including a 'forest' class, and the tree canopy cover (TCC) product, both 30m resolution, to sample boundaries. Each NLCD and TCC plot was then aggregated to 60, and the 60m raster aggregated to 120, 180, 240, 420, and 480m using majority rules (NLCD) or mean (TCC) aggregation (Fig. 1c).

We computed analogous patch-based and surface metrics to measure downscaling accuracy (Table 1) and a suite of metrics from each paradigm that measure landscape composition and configuration (McGarigal et al. 2009). Patch-based metrics were computed using Fragstats (McGarigal et al. 2012) and surface metrics using SPIP software (Image Metrology).

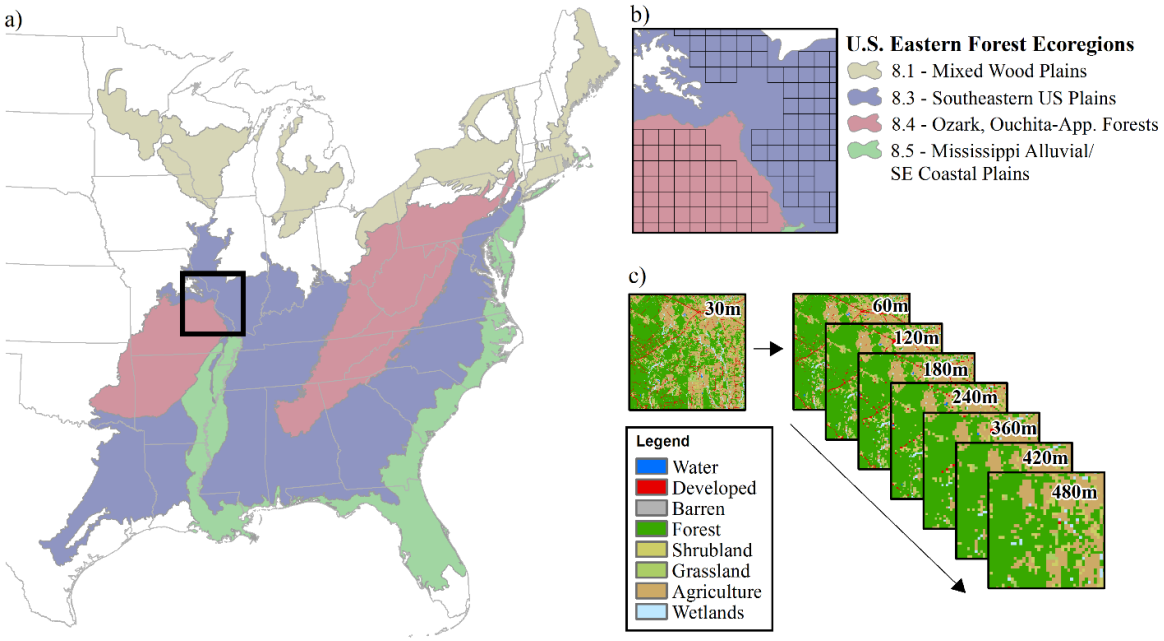


Figure 1. (a) Eastern U.S. forest ecoregions, (b) grid sampling scheme, and (c) NLCD aggregation.

Reserving the 30m raster, a scalogram was plotted for PD and Sds for the seven coarse resolutions for each plot, and various scaling functions fit to the scalogram (i.e., power law, first-, second-, and third-order polynomials). Wu (2004) and Frazier (2015b) demonstrate these curve types accurately model metric behavior across scale. To generate a measure of MAUP-induced error, those curves were downscaled to predict each metric value at 30m and those predicted values compared to true 30m PD and Sds values through a measure of relative error (Eq. 1), where M_p is the predicted metric from the scaling function, and M_t is the true metric value. Low E_{rel} values indicate the scaling function accurately predicts the metric at a finer resolution. The curve type producing the lowest average E_{rel} for each metric was selected as the candidate model for further examination.

Table 1. Analogous metrics for measuring downscaling accuracy and data loss in the two paradigms.

Patch-Based	Surface
Downscaling metrics	
Patch Density (PD)	Peak Density (Sds)
Compositional data loss metrics	
Total area of forest (AREA)	--
Largest patch index (LPI)	Maximum peak height (Sp)
Configurational data loss metrics	
Edge Density (ED)	Peak-Peak range height (Sy)
Percent of Like Adjacencies (PLADJ)	Moran's I index (Smi)
Mean Fractal Dimension (FRAC)	Surface fractal dimension (Sfd)

To capture compositional data loss within the patch-based paradigm, we computed the linear rate of change of the data loss metrics (Table 1) across the seven coarse resolutions. Finally, we used conventional, ordinary least squares regression to measure the influence of data loss on downscaling accuracy. Expressed as Equation 2, $\hat{y}_i$ is the estimated value of the dependent valiable E_{rel} for plot *i*, β_0 represents the intercept, and $\beta_{config(k)}$ and $\beta_{comp(j)}$ are coefficients for the independent variable x_{ik} and x_{ij}, our measures of either compositional or configurational data loss, and ε_i represents the error term. Independent regressions were completed for each ecoregion for each paradigm and a combined regression for all regions for each paradigm.

$$E_{rel}(\%) = \left|\left(M_p - M_t\right)/M_t\right| * 100 \quad (1)$$

$$\hat{y}_i = \beta_0 + \sum_k \beta_{config(k)}\, x_{ik} + \sum_k \beta_{comp(j)}\, x_{ij} + \varepsilon_i \quad (2)$$

III. Results and Discussion

Across the four ecoregions, a third-order polynomial model performed best for downscaling PD, and power law performed best for Sds, which is consistent with recent findings (Frazier 2015a). PD mean R^2 values were consistently 0.97. Model fit improved with the surface paradigm, and R^2 values were >0.99 for Sds (Table 2). Relative error statistics (Table 2) show average downscaling accuracies ranged from 28.3 to 39.8 for PD, and 27.7 to 39.3 for Sds. Thus, there is little correlation between model fit (R^2) and downscaling accuracy (E_{rel}), but results confirm prior findings that even when scaling relationships are strong and model fits are high, metric downscaling of aggregated data is not particularly accurate (Frazier 2014).

Table 2. Scaling model fit (R^2) and downscaling relative error (E_{rel}) across ecoregions and paradigms.

	PD				Sds			
Ecoregion	**8.1**	**8.3**	**8.4**	**8.5**	**8.1**	**8.3**	**8.4**	**8.5**
Mean R^2	0.97	0.97	0.97	0.97	0.998	0.996	0.997	0.998
Std. Dev. R^2	0.01	0.01	0.02	0.01	0.002	0.004	0.005	0.002
Mean E_{rel}	28.3	31.8	39.8	28.8	27.7	39.3	34.4	28.1
Std. Dev. E_{rel}	10.7	9.85	14.9	7.94	8.7	13.1	16.1	10.3
Sample (n)	124	125	125	125	124	125	125	125

Tables 3 and 4 summarize the results of OLS regression for the two paradigms. For the patch paradigm, the rate of AREA loss was consistently a significant predictor of relative downscaling error. ED was also significant for most models. Across all ecoregions, PLADJ was also significant. Model fit values ranged from 0.41 to 0.596. For the surface paradigm, models were poorly fit. No measures of compositional or configurational data loss were significant across all regions. Variance inflation factors indicated possible

collinearity of variables. These findings suggest further examination of alternative metrics is necessary. Of the two models with reasonable fit, the rate of loss of maximum peak height (Sp), a measure of compositional loss, was the best predictor of downscaling accuracy.

Table 3. Impacts of data loss on downscaling accuracy of PD (patch paradigm).

Ecoregion	8.1	8.3	8.4	8.5	Total
Intercept	18.174***	34.948***	40.101***	36.102***	40.802***
AREA	0.666*	1.250***	2.243**	1.914***	1.361***
LPI	-0.243	0.678**	-0.644	-0.602	0.029
ED	-2.566***	1.351***	0.984	0.934**	1.609***
PLADJ	0.625	1.173***	2.745	0.558	1.184***
Diagnostics					
R^2	0.478	0.596	0.467	0.410	0.454
Sample (n)	124	125	125	125	499

*p < 0.10, **p<0.05, ***p<0.01; FRAC was not significant in any models

Table 4. Impacts of data loss on downscaling accuracy of Sds (surface paradigm).

Ecoregion	8.1	8.3	8.4	8.5	Total
Intercept	34.375***	52.849***	41.33***	30.780***	39.288***
Sp	1.083	14.762***	10.790***	1.251	5.371***
Sy	-0.281	-6.769	-0.335	-0.114	-1.093
Sfd	-1.414***	-2.969***	0.361	-0.716	-1.140***
Diagnostics					
R^2	0.064	0.246	0.264	0.024	0.112
Sample (n)	124	125	125	125	499

*p < 0.10, **p<0.05, ***p<0.01; Smi was not significant in any models

IV. Conclusions

Results suggest a link between compositional data loss and downscaling accuracy, particularly in the patch-based paradigm—the greater the rate of AREA loss during aggregation, the more difficult it becomes to predict metric values at a finer resolution. Further work is needed to determine if any relationships exist between compositional and configurational data loss measures and downscaling error in the surface paradigm. An initial step would be to further establish correspondences between patch and surface metrics within these and other ecoregions. Future work should also address whether ecological characteristics impact the success these measures.

Acknowledgements: This work is supported by a grant to the authors from the U.S. National Science Foundation (#1561021) for "Data Complexity and Spatial Scaling: Prediction Accuracy and Implications for Emerging Landscape Paradigms".

References:

Frazier, A. E. (2014) A new data aggregation technique to improve landscape metric downscaling. *Landscape Ecology*, *29*(7), 1261-1276

Frazier, A.E. (2015a) Landscape heterogeneity and scale considerations for super-resolution mapping. *International Journal of Remote Sensing*,

Frazier, A. E. (2015b) Surface metrics: scaling relationships and downscaling behavior. *Landscape Ecology*, 1-13

Holt, D., Steel, D.G., Trammer, M., & Wrigley, N. (1996) Aggregation and Ecological Effects in Geographically Based Data. *Geographical Analysis* 28(3):244-261

McGarigal, K., Tagil, S., & Cushman, S. A. (2009). Surface metrics: an alternative to patch metrics for the quantification of landscape structure. Landscape Ecology, 24(3), 433-450

McGarigal, K., SA Cushman, and E Ene. 2012. FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. Computer software program produced by the authors at the University of Massachusetts, Amherst. Available at the following web site: http://www.umass.edu/landeco/research/fragstats/fragstats.html

Omernik, J.M. (1987) Ecoregions of the conterminous United States. *Annals of the Association of American Geographers* 77(1), 118-125.

Turner, M. G., O'Neill, R. V., Gardner, R. H., & Milne, B. T. (1989). Effects of changing spatial scale on the analysis of landscape pattern. Landscape ecology, 3(3-4), 153-162

Steel, D. G., & Holt, D. (1996) Analysing and adjusting aggregation effects: the ecological fallacy revisited. *International Statistical Review/Revue Internationale de Statistique*, 39-60

Wu, J. (2004) Effects of changing scale on landscape pattern analysis: scaling relationships. *Landscape Ecology*, 19:125-138.

A hierarchical scale setting strategy for improved segmentation performance using very high resolution images

Taskin Kavzoglu[1]*, Merve Yildiz Erdemir[1]

[1]Gebze Technical University, Department of Geomatics Engineering, 41400, Kocaeli, Turkey

(kavzoglu, m.yildiz)@gtu.edu.tr

Abstract

Land use/land cover (LULC) classification is a specific implementation to define terrain features to the closest real world object. Object-based image analysis (OBIA) has been proved to improve classification accuracy, particularly for very high resolution remotely sensed images. In this study, multiresolution segmentation algorithm was utilized in the image segmentation process using a pan-sharped Qickbird-2 image. Segmentation scales were determined by widely-used estimation of segmentation parameter (ESP-1) tool that produces rate of change graph (LV-RoC) in terms of local variance of the image. In this study, the LV-RoC graph of the image was evaluated to determine optimal scale values ranging from fine to coarse levels. An attempt was made to estimate optimal scale parameter for an image considering not only single-scales but also multi-scales for an image using a hierarchical scale setting strategy. Nearest neighbour classifier was used on single-scale segmented images and fuzzy classifier employing membership functions was used on multi-scale segmented image. Equal numbers of pixels for each class were randomly selected to estimate accuracy metrics (i.e. overall accuracy and kappa coefficient). The differences in classifier performances (~ 6%) were statistically significant according to McNemar's test. It was found that the proposed strategy has a great potential for LULC classification using very high resolution imagery.

Keywords: Object-based classification, Segmentation, Scale parameter, ESP, Quickbird-2

I Introduction

Parallel with the increasing use of very high spatial resolution images, object based classification has become more advanced (Kim et al., 2011). As stated by Blaschke et al. (2014), the object based classification technique that uses image objects instead of pixels has been widely used in last decade primarily applied to very high resolution images. Object-based image analysis (OBIA) offers several unique advantages in comparison to pixel-based classification, one of which is the removal of the so-called salt-and-pepper effect. In OBIA, shape, context and textural information of image objects are considered instead of individual pixels. At the same time, each object is considered both by its spectral, shape or texture features, and by its unique neighbours, its sub- and super-objects (Benz et al., 2004). The recent availability of OBIA provides various opportunities for sensitive and detailed LULC classification.

OBIA is generally applied in three stages: image segmentation, classification and accuracy assessment. Image segmentation process is the first and crucial step to define image objects. Although there exists several techniques, multiresolution segmentation introduced by Baatz and Schäpe (2000) has been one of the most commonly used image segmentation algorithm. Multiresolution segmentation produces highly homogeneous image objects in arbitrary resolution on different types of data and is applied to many problems in the field of remote

sensing (Baatz and Schäpe, 2000). It includes the setting of three major parameters, namely scale, shape and compactness. Setting of these parameters is of critical importance for the performance of a subsequent classification process (Witharana and Civco, 2014). Of these parameters, it is agreed that scale parameter is the most important one as it determines the objects size and thereby affects largely the resulting thematic map accuracy (Kim et al., 2011; Myint et al., 2011). The optimal value for scale parameter varies depending on the structure of the study area, land cover types and scale of the imagery (Myint et al., 2011).

Results of classification accuracy are affected by quality of segmentation process (Marpu et al., 2010). For this reason, many studies have already been carried out to determinate optimal scale parameter. In a previous research, Kavzoglu and Yildiz (2014) endeavoured to prepare guidelines for the selection of scale parameter. Besides selection of the scale parameter without relying on any method, evaluation methods of segmentation quality can be divided into two categories as supervised (Clinton et al., 2010) and unsupervised (Johnson et al., 2011; Gao et al. 2011).

In recent years, researches mainly focused on the use of statistical methods for determination of optimal scale parameter in segmentation (Witharana and Civco, 2014). However, only a few studies have led to automatically determining results that produce fast and applicable solutions. Woodcock and Strahler (1987) utilized local variance (LV) graphs to ascertain the spatial structure of each type of image. Then, the estimation of scale parameter (ESP) tool was developed by Drăguţ et al. (2010) for automated detection of scale parameter. The tool has been recently extended to multi-scale analysis based on single layers (Drăguţ et al., 2014). With this tool, LV was generated for each scale value from an image. Furthermore, a graph is generated using LV of image and rate of change (RoC) values of LV. Thus, LV-RoC graph is obtained on each scale step in an image considering a single layer.

The purpose of this study is to determine optimal scale parameter(s) of the image using the LV-RoC graphs produced by the ESP tool and perform usage of single- and multi-scale analysis on classification process. Nearest neighbour classifier was employed in single-scale classification process and membership functions were used in multi-scale classification process. The differences in classifier performances were analysed with McNemar's test, which is a statistical test for differences.

II Study area and Dataset

The study area chosen for this research covers approximately 80 ha area located in the Yomra district of Trabzon province, Turkey. A multi-spectral pan-sharpened Quickbird-2 satellite image having four spectral bands at 0.6 m spatial resolution acquired in May 2008 was used for the study site. It has high mountainous terrain covered by forest, sea and urban classes. For the purpose of evaluation, image was clipped to a subset of 1500x1500 pixels (Figure 1). eCognition Developer (v9.1), a widely-used object based image processing software, was used for all segmentation and classification experiments in this study.

The study area is mainly covered by eleven land use/cover features, namely bare ground, concrete surface, forest, asphalt road, gravel, pasture, shadow, water, red, blue and white roof.

III Image segmentation

Image segmentation is the first and crucially important process of object based classification, since the initial objects are created on this stage. The most widely used algorithm in image segmentation is multiresolution segmentation introduced by Baatz and Schäpe (2000). It is a bottom up region merging technique based on local homogeneity criteria. It merges neighbour and similar spectral pixels into image objects. Size of image objects is adjusted in conjunction

with segmentation parameters including scale, shape, compactness and band weights. Scale parameter sets the image size and considering most crucial parameter. In general, higher values for the scale parameter result in larger image objects, smaller values in smaller ones (Yildiz et al., 2012).

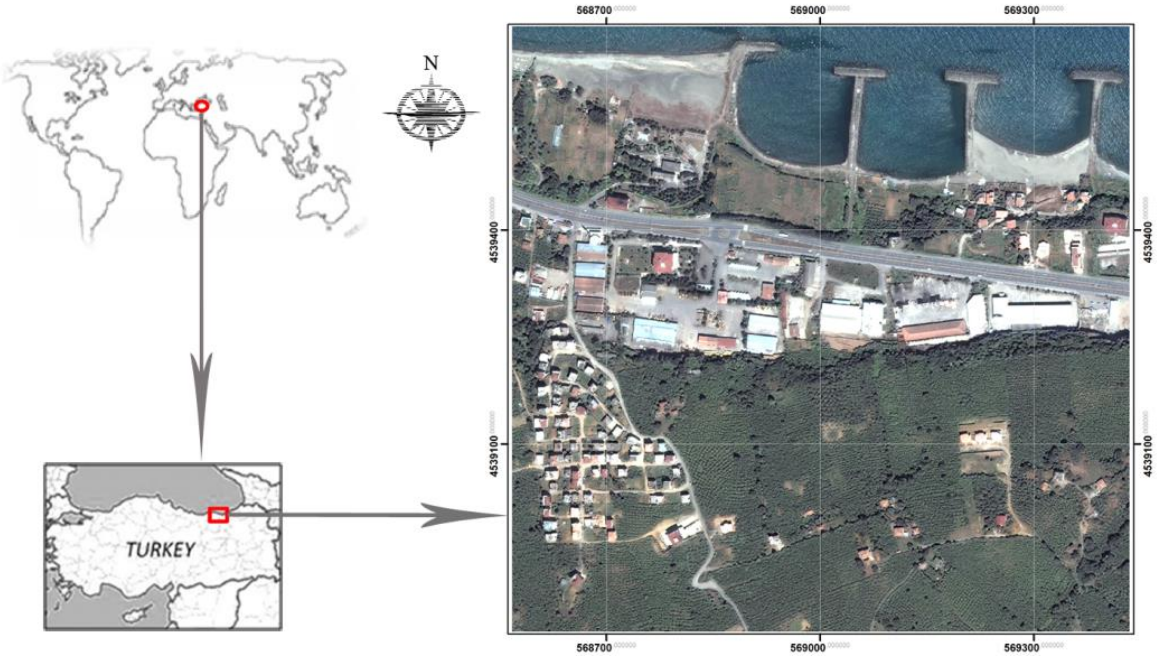


Figure 1. Location of the study area, Trabzon province of Turkey.

Estimation of scale parameter (ESP) tool was developed for automated detection of scale parameter by Drăguţ et al. (2010) who states that "the ESP tool iteratively generates image objects at multiple scale levels in a bottom-up approach and calculates the local variance (LV) for each scale image". Rate of change (RoC) values of LV is determined for each scale levels using Eq. 1. The RoC is calculated as:

$$RoC = \left\lfloor \frac{L-(L-1)}{L-1} \right\rfloor * 100 \quad (1)$$

where *L* is local variance at the target level and *L-1* is local variance at next lower level. The peaks in the LV-RoC graph designates the object levels at which the image can be segmented in the most appropriate way, according to data properties at the scene (Drăguţ et al., 2010).

IV Classification

Classification proccess which is the next step after the image segmentation uses segmented images. Two types of classifiers were applied in this study. Nearest neighbour classifier uses a set of samples of different classes in an attempt to assign class values to a segmented object. The procedure contains two major steps: teaching the system by giving it certain image objects as samples and classifying image objects in the image object domain based on their nearest sample neighbours. Membership functions allow describing the relationship between feature values and the degree of membership to a class using fuzzy logic. It can be defined by the degree of membership; for example, any value between one and zero (Definiens, 2008). In other words, Myint et al. (2011) states that "the membership function describes intervals of feature characteristics that determine whether the objects belong to a particular class or not".

V Results

This study aims to investigate optimal scale parameter(s) for single- and multi-scale segmentation. A pan-sharpened Quickbird image includes many land use/cover features such as urban, sea, forest and gravel was used in this study. For detecting optimal scale parameter(s), local variances for each of scale parameter were estimated using ESP tool in the image. As can be seen in Figure 2, the prominent peaks are apparent in the LV-RoC graph created using Eq. 1. Peaks at 26, 41, 53 and 77 on the graph appeared to be suitable values. The four scale levels were denoted s_ESP1, S_ESP2, S_ESP3 and S_ESP4, where S_ESP1 represented the finest object scale and S_ESP4 the coarsest scale. These scale parameter values specified by LV-RoC graph was individually applied to multiresolution segmentation for single-scale segmentation. Furthermore, all scale values were used in multi-scale segmentation using a hierarchical strategy (M_ESP). It should be noted that shape and compactness parameters were keep constant as 0.1 and 0.5, respectively.

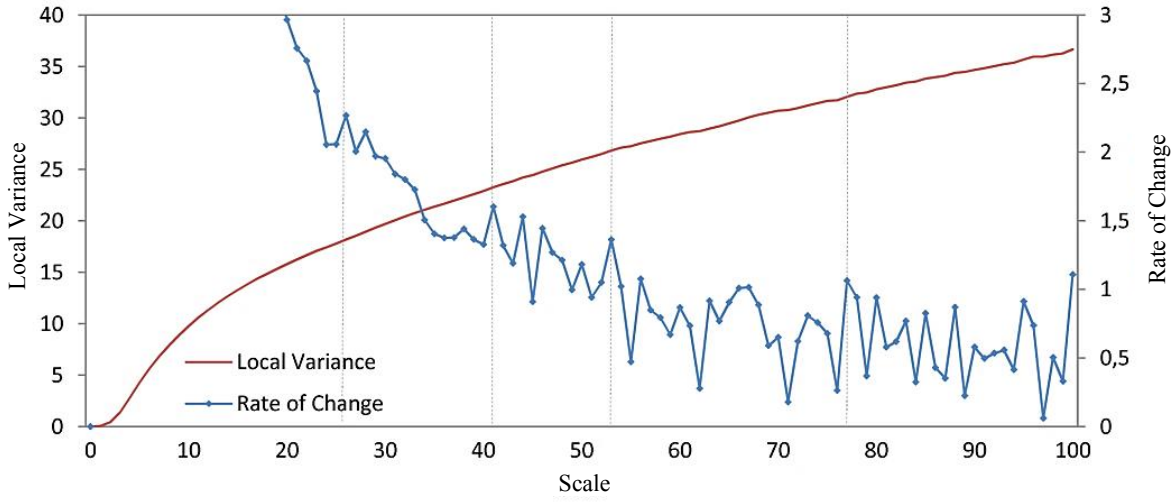


Figure 2. LV-RoC graph of the image. Scale values of 26, 41, 53 and 77 were selected as potential optimal scale values.

For single-scale segmentation, segmented images obtained from the estimated scale parameters were classifed using nearest neighbour classifier. Membership function classifier was applied on multi-scale segmentation and different segmentation scale levels were applied to detect feature classes. Each scale level was generated within a hierarchy, where scale parameters of 77, 53, 41 and 26 were set as scale levels of 1, 2, 3 and 4, respectively. Land use/cover classes were classified at mentioned levels. To be more specific, while water and gravel classes were extracted using scale level 1, blue roof and forest was extracted with scale level 2, asphalt road and white roof with scale level 3, and other classes with scale level 4.

Classifications were achieved using test datasets on the basis of contingency matrices. It should be noted that test datasets were prepared using random pixel selection. Equal numbers of samples for each class (700 pixels) were selected. For the assessment of classification results, overall classification accuracy (OA) and Kappa statistics were computed from the contingency matrices (Table 1).

	Scale Unit	Overall Accuracy (%)	Kappa
Single-scale	S_ESP1	86.10	0.847
	S_ESP2	84.65	0.831
	S_ESP3	81.18	0.793
	S_ESP4	80.39	0.784
Multi-scale	M_ESP	91.56	0.907

Table 1. Overall accuracies of single- and multi-scale segmentations.

The overall classification accuracies for single-scale parameters of 26, 41, 53 and 77 were estimated as 86.10%, 84.65%, 81.18% and 80.39%, respectively, suggesting superior performance when fine level scale values were used. For multi-scale segmentation applied through a particular strategy with fuzzy approach, a classification accuracy of 91.56% was achieved. As can be seen from Table 1, classification accuracies on single-scale segmentation range between 80% and 86% in terms of overall accuracy. All classified image results are shown in Figure 3.

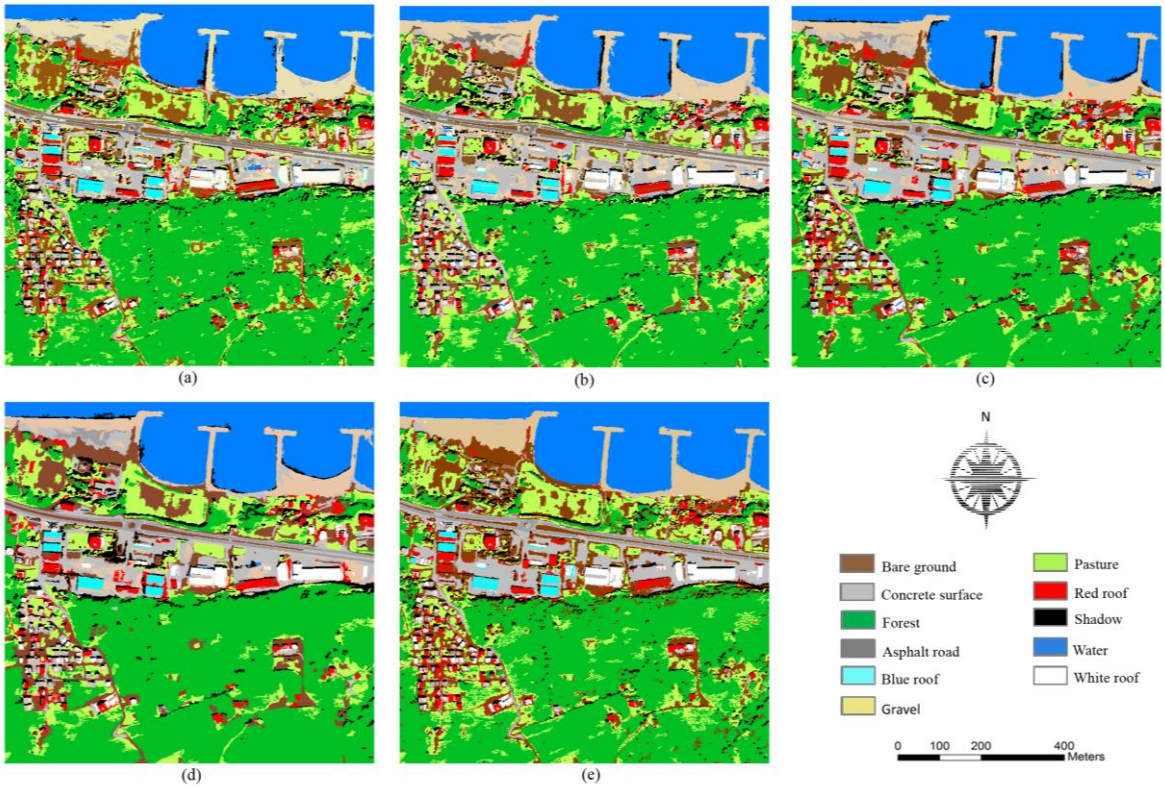


Figure 3. Single-scale classification results for (a) S_ESP1, (b) S_ESP2, (c) S_ESP3, (d) S_ESP4 and multi-scale classification result with hierarchical scale (e) M_ESP.

Although some classes (e.g., water and forest) were delineated easily, there were misclassification problems between some classes (e.g., gravel, bare soil, asphalt road and red building roofs). It should also be noted that low accuracies, compared with the accuracies produced for fine scale selection (i.e. 26 of scale value), were achieved in the classification of coarser scale selection (i.e. 53 and 77 of scale values). The reason could be related to training objects containing two or more classes in same object. Overall, the highest classification result

was estimated by multi-scale segmentation using membership functions. One reason for this can be that scale levels separated to identify objects of different classes and each land use feature was represented according to its spatial features. Another reason may be related to using membership functions in classification process. Thus, some classes were easily defined by object features such as spectral bands, NDVI, band ratio, brightness etc.

McNemar's test based on χ^2 distribution is a popular non-parametric test that is generally performed to compare the classification errors (Kavzoglu et al., 2015). In this study, McNemar's test was implemented to determine statistical significance of the differences between classifications according to scale selections (Table 2). It was found that there is statistical significance among all five classification results at 95% confidence interval ($\chi^2_{0.05}$ = 3.84).

	S_ESP1	**S_ESP2**	**S_ESP3**	**S_ESP4**	**M_ESP**
S_ESP1	-	210.72	463.59	539.13	930.44
S_ESP2		-	60.84	101.82	324.38
S_ESP3			-	17.19	132.64
S_ESP4				-	79.78
M_ESP					-

Table 2. McNemar's test result for single- and multi-scale segmentation.

VI Conclusion

Determination of the optimal scale parameter of an image is currently an important research for object based image analysis. The more representative segments are introduced to a segmentation process, the more accurate classification results can be produced. Land use/cover features of used imagery are usually composed of a complex combination of buildings, roads, water area, trees, and pasture. Scale parameter depends on certain structural features such as spatial resolution of image, land use/cover characteristics and study area.

In this study, LV-RoC graph derived from ESP tool was utilized for determine optimal scale parameter. Thus, each scale parameter was separately applied as single-scale segmentation and hierarchical usage of scale parameters was used in multi-scale segmentation. Nearest neighbour and membership function classifiers were applied on single- and multi-scale segmentation to perform the land cover classification, respectively.

The scale parameter of 26 was selected as an optimum scale parameter of single-scale segmentation after estimated segmentation levels, determined through LV-RoC graph analysis, were individually tested for multiresolution segmentation process. However, scale parameter of 77 produced the lowest accuracy of classification. The reason may be related to the fact that the large image objects produced for the coarse scale, thus several land cover properties were represented in a single object. Particularly, about 6% classification accuracy improvement was achieved when considering the optimal scale parameters in multi-scale segmentation process compared to single-scale segmentation. Consequently, overall accuracies of the classification based on multi-scale segmentation were higher than single-scale segmentation. As a result, it can be concluded that multi-scale segmentation shows better performance compared to single-scale segmentation. Main reason for poor performance of single-scale may be related to construction of objects including several land cover features, especially in coarse scale values. In summary, the analyses conducted in this study clearly showed that the multi-scale segmentation based on hierarchical scale setting strategy is an effective approach for classifying very high resolution imagery.

References

Baatz M., Schäpe A. (2000). Multiresolution Segmentation: an optimization approach for high quality multi-scale image segmentation. In: Strobl, J., Blaschke, T., Griesebner, G. (Eds.), *Angewandte Geographische Informations*-Verarbeitung XII. Wichmann Verlag, Heidelberg, pp. 12-23.

Benz U.C., Hofmann P., Willhauck G., Lingenfelder I., Heynen M. (2004). Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58(3-4), 239-258.

Blaschke T., Hay G.J., Kelly M., Lang S., Hofmann P., Addink E., Queiroz Feitosa R., van der Meer F., van der Werff H., van Coillie F., Tiede D. (2014). Geographic Object-Based Image Analysis – Towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*. 87, 180-191.

Clinton N., Holt A., Scarborough J., Yan L., Gong P. (2010). Accuracy assessment measures for object-based image segmentation goodness. *Photogrammetric Engineering and Remote Sensing*, 76(3), 289-299.

Definiens 2008. eCognition Developer 7.0 User Guide. Definiens AG, Munich, Germany.

Drăguţ L., Csillik O., Eisank C., Tiede D. (2014). Automated parameterisation for multi-scale image segmentation on multiple layers. *ISPRS Journal of Photogrammetry and Remote Sensing*. 88, 119-127.

Drăguţ L., Tiede D., Levick S. R. (2010). ESP: a tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. *International Journal of Geographical Information Science*, 24(6), 859-871.

Gao Y., Mas J.F., Kerle N., Navarrete Pacheco J.A. (2011). Optimal region growing segmentation and its effect on classification accuracy. *International Journal of Remote Sensing*, 32(13), 3747-3763.

Johnson B., Xie Z. (2011). Unsupervised image segmentation evaluation and refinement using a multi-scale approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(4), 473-483.

Kavzoglu T., Yildiz M. (2014, Sept.29- Oct. 2). Parameter-based performance analysis of object-based image analysis using aerial and Quikbird-2 images. In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-7, pp.31-37.

Kavzoglu T., Colkesen I., Yomralioglu T. (2015). Object based classification with rotation forest ensemble learning algorithm using very high resolution WorldView-2 image. *Remote Sensing Letters*, 6(11), 834-843.

Kim M., Warner T. A., Madden M., Atkinson D. S. (2011). Multi-scale GEOBIA with very high spatial resolution digital aerial imagery: scale, texture and image objects. *International Journal of Remote Sensing*, 32(10), 2825-2850.

Marpu P. R., Neubert M., Herold H., Niemeyer I. (2010). Enhanced evaluation of image segmentation results. *Journal of Spatial Science*, 55(1), 55-68.

Myint S.W., Gober P., Brazel A., Grossman-Clarke S., Weng Q. (2011). Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. *Remote Sensing Environment*, 115(5), 1145-1161.

Witharana C., Civco D. L. (2014). Optimizing multi-resolution segmentation scale using empirical methods: Exploring the sensitivity of the supervised discrepancy measure Euclidean distance 2 (ED2). *ISPRS Journal of Photogrammetry and Remote Sensing*. 87, 108-121.

Woodcock C.E., Strahler A.H. (1987). The factor of scale in remote sensing. *Remote Sensing Environment*. 21, 311-332.

Yildiz M., Kavzoglu T., Colkesen I., Sahin E.K. (2012, July 10-13). An assessment of the effectiveness of segmentation methods on classification performance, In: *10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Florianópolis, SC, Brazil, pp. 133-138.

Exploring Accurate Spatial Downscaling using Optimization

Michael Poss[*3] **and Didier Josselin**[1,2]

[1]UMR ESPACE 7300, CNRS, Université d'Avignon, France
[2]LIA, Université d'Avignon, France
[3]UMR CNRS 5506 LIRMM, Université de Montpellier, France

*Corresponding author: michael.poss@lirmm.fr

Abstract

This papers proposes to combine several downscaling processes based on expert knowledge and objective functions from Operations Research to fill a table of social data about illiteracy in the USA during the Thirties.

Keywords

Downscaling, social data, Operations Research, optimization, data about illiteracy, USA, 1930

I THE PROBLEM OF SPATIAL DOWNSCALING ON CENSUS DATA

This paper deals with the well-known downscaling problem (Bierkens et al. (2000)) related to the ecological inference fallacy (King (1997)), change of support problem (King et al. (2004)) or aggregation problem (Josselin et al. (2008)) in general. Many applications and research fields (environment, social science, geography) are concerned and this problem notably impairs census data analysis (Gehlke and Biehl (1934); Robinson (1950); Tranmer and Steel (1998)).

1.1 Example

Let us consider that we have information (attribute values) for a complete statistical population of a territory. We want to disaggregate this information into a set of smaller sub-areas that compose the whole territory, at a finer scale. To do so, we need to make assumptions on the way we distribute the values into the spatial partition elements. Fortunately, we also have a few relevant information about the complete population and about individuals aggregates, for all the categories of the variable, and also in each sub-area. But we do not know the exact distribution of the category-specific values in each area. We have to make assumptions to estimate it. This is illustrated in the Table 1 with an example about housing.

Region	**House owner**	**Tenant**	**Free accommodation**	***Inhabitants (millions)***
PACA	?	?	?	5
Rhône-Alpes	?	?	?	6.5
Centre	?	?	?	2.5
Quantity	8	5.6	0.4	**14**

Table 1: 14 millions of people from 3 French regions to distribute in 9 cells of the table according to different types of housing.

Depending on the number of sub-areas (e.g. regions) and classes of the attribute (e.g. housing types), there exist many ways to fill the partition of the Table 1. At this point, those are all

equivalent solutions to the problem since they respect the constraints of the (known) summed values in the row and column margins. For instance, the most simple table (that would correspond to a freedom degree of only one in a contingency table) with only 20 individuals and 4 cells, i.e. 2 categories for both variables, has already 5 possible solutions.

To find the solution that fits the reality the best way –in the unreachable case of knowing the complete table content, i.e. how many people really live in a given region with a given type of housing–, we sometimes need to use statistical tools or to fix complementary assumptions related to complementary knowledge.

However, we are not completely deprived of means. What is interesting here is that we can firstly set simple constraints based on aggregated information. Indeed, we know that:

- the sum of the values in every line or column should be equal to the corresponding margin values;
- the sum of the margin values should be equal to the total number of individuals.

To find what is the "good" solution, then we have to fix an objective to maximize or minimize, according to assumptions from experts. This is the main idea of this paper. On the one hand, getting reliable information using census is very costly and cannot be made very frequently. On the other hand, there are experts in social science who may know a lot about socio-spatial characteristics of people. It can strongly help in making assumptions to drive disaggregation process, although we know that we definitely cannot find the real and exact distribution of the data at a given time. However, it seems anyway better than reading in a crystal ball.

In this paper, we propose an approach mixing statistics and optimization to find an optimal data distribution in a table, according to aggregated constraints and expert knowledge. Our objectives are the following:

- getting a better accuracy in statistical data due to a downscaling process driven by a statistical criterion optimized using a mathematical solver;
- finding the solution the most related to a given statistical criterion or to expert knowledge translated to a new matrix of expected values;
- assessing the quality of the solution when taking into account uncertainty in downscaling estimations, providing min and max values in each cell of the contingency table.

This method is tested on data studied by Robinson (1950), for which we know the complete matrix. Thence it enables to compare the results of our method to the real observed data.

1.2 Our data set: USA census data about illiteracy in 1930 used by Robinson (1950)

We use the data about the illiteracy in the population of USA in 1930, which are reference data in the field of sociology [1]. They include the number and the percentage of illiterate people in each State and in regions, by population color and nativity. We study four groups of population:

- *native white with native parentage* people;
- *native white with foreign or mixed parentage* people;
- *foreign born white* people;
- *black* people.

In this paper, optimization methods are used for the whole matrix M and results are provided and only discussed about black people in USA in 1930. Indeed, we test a complementary hypothesis on the role of history (slavery in confederate states before civil war) in black population illiteracy (for instance, see map provided in the Figure 1).

[1] http://www.ru.nl/sociology/mt/rob/downloads/

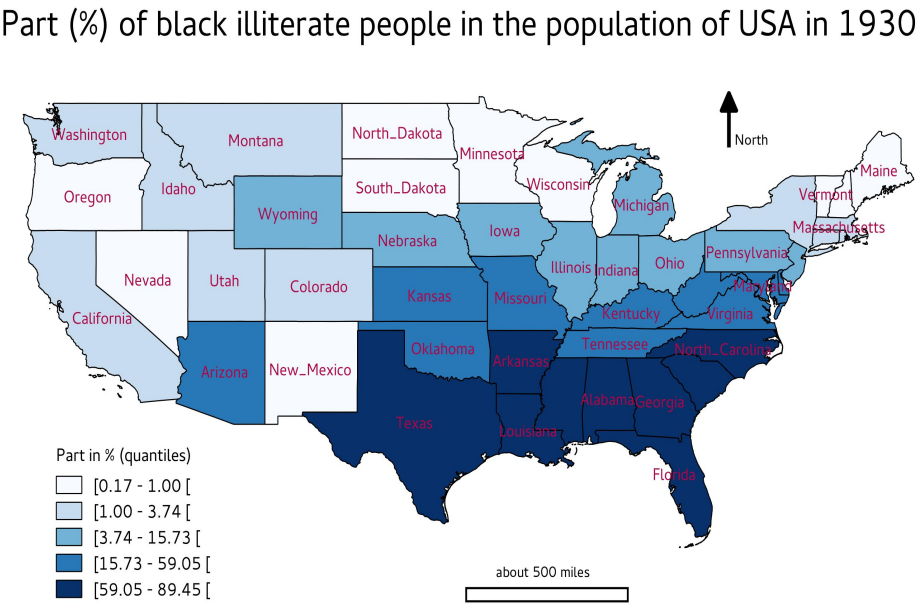


Figure 1: Illiterate black people in USA in 1930: a location related to history (unified *vs* confederate states in East-Southern).

Using these data, Robinson (1950) noticed that the correlation between illiteracy rate of black people and the different regions was close to 1. But if we consider the statistics at another more accurate scale (the States), it goes down to 0.2. This lead Robinson to define the *ecological fallacy* in social science and more generally on census data. On a practical point of view, it seems indeed very difficult to apply a knowledge learnt using the same statistics provided at a certain aggregated level, on a more disaggregated and finer level. That is the purpose of our proposition: trying to improve local estimation accuracy and downscaling process by modifying the data according to hypothesis and new constraints from aggregated statistics and from expert knowledge. Because we know the complete information about illiteracy in each type of population and in all the States in USA in 1930, we can compare and discuss the results of our method to the real data of the census. In this first stage, we fix the objective function criteria and the expert assumptions by ourselves to test the method.

II A METHOD TO DOWNSCALE CENSUS DATA

2.1 Downscaling in areal units

In geography, it is very common to disaggregate spatial data into more accurate layers. Usually, a hypothesis of proportionality is applied. The assumption is the following: space is somehow isotropic and densities are the same all over geographical space or inside a given type of characterized area. This corresponds to the $H0$ hypothesis in Chi^2 statistics. For instance, as in a contingency table, the expected number of individuals $\hat{x_{ij}}$ in a cell of the table is estimated by the product of the margin values of the corresponding column $x_{i.}$ and row $x_{.j}$, divided by the total number of individuals $x_{..}$. Here we consider that illiteracy and regions where individuals are located are independent statistical variables. If there are significant differences in the number of individuals between regions or between classes of the variable, this can lead to very various values in the cells of the table.

Another way to compute downscaling may be to find the solutions minimizing a certain statistical criterion. For instance, variance of the values in the matrix M of the table can be minimized. In this case, experts consider a kind of homogeneity in downscaling, i.e. that cell values should not differ too much from each others whatever the margin value weights. This approach is slightly different from the previous one, because a global statistical criterion is now considered. Moreover, it is possible to minimize many different functions $f(\theta)$ on such a table, according to several Lp-norms as we shall see in the next sections.

In an analog way, experts can use information about sub-areas peculiarities to provide probabilities or ranges (probability min and max) in each cell. Then the problem leads to find the solution that is the closest to the probability matrix M given by the expert. Practically, experts can provide bounds in which (s)he expects the values to be, in each cell of M. Whatever the function to minimize and levels of uncertainty (intervals), the optimal solution can be reached using an optimizer, as we propose in the following section.

2.2 Variability criteria to optimize

We denote by x_{ij} the value of row i and column j in the resulting matrix M and $\bar{x}$ the mean of all x_{ij}.

$$\theta_{ij} = |x_{ij} - \bar{x}| \tag{1}$$

We aim at minimizing 4 different objective criteria $f(\theta)$, each of them representing one approach to weight the variability among the different components of x. Then, given each specific objective criteria, we analyze the resulting optimal distributions of $\hat{x}_{ij}$ in M.

A first criterion is based on the L_∞-norm, which illustrates an equity process: outliers have an important weight in this solution.

$$f(\theta) = \max_{i,j}(\theta_{ij}) \tag{2}$$

A second is similar but takes into account squared residuals θ_{ij}. Outliers are even more considered.

$$f(\theta) = \max_{i,j}(\theta_{ij}^2) \tag{3}$$

The following case corresponds to the least absolute deviation case (L_1-norm) which depicts an efficiency purpose.

$$f(\theta) = \sum_{i,j} \theta_{ij} \tag{4}$$

This last criterion is the variance and corresponds to the L_2-norm (equality).

$$f(\theta) = \sum_{i,j} \theta_{ij}^2 \tag{5}$$

The drawback of the objective functions (2) and (4) is that many optimal solutions may exist, and in particular, optimal solutions that contain many components of x set to 0 which does not provide interesting insights. To avoid that situation, we smoothed the objective functions by replacing θ with θ' defined as

$$\theta'_{ij} = \theta_{ij} + 0.001.\theta_{ij}^2 \tag{6}$$

Smoothing ensures unicity of the optimal solutions and, more importantly, increase the number of positive coefficients.

These functions are tested with different data relating to different assumptions and in two cases:

- No information is provided about the range of the value in each cell (no information known or expressed about uncertainty of $\hat{x}_{ij}$);
- In each cell, the value is bounded and the solution must take into account this constraint to be included in the interval (uncertainty modeling for $\hat{x}_{ij}$).

2.3 Mathematical model: optimization under constraints

Let C_j be the given value for the sum of all elements in column $j \in \{1, \ldots, m\}$ and R_i be the sum of all elements in row $i \in \{1, \ldots, n\}$. We denote by x_{ij} the value of row i and column j in the resulting matrix M. The problem of finding the optimal values for x can be stated as the following optimization problem

$$(P) \equiv \begin{cases} \min & f(\theta) & \\ \text{s.t.} & \sum_{i=1}^{n} x_i = C_j & j \in \{1, \ldots, m\} \\ & \sum_{j=1}^{n} x_j = R_i & i \in \{1, \ldots, n\} \\ & x \geq 0 & \\ & \theta_{ij} \geq x_{ij} - \bar{x} & i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\} \\ & \theta_{ij} \geq -x_{ij} + \bar{x} & i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\} \end{cases}$$

where the function f models the preference used in the construction of the matrix M as explained in the previous section, and the last two groups of inequalities are a linearization for the definition of θ from (1). Notice that the objective functions involving $\max$ can be linearized by introducing the artificial optimization variable Θ linked to θ_{ij} through

$$\Theta \geq \theta_{ij} \quad i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\},$$

and optimizing the objective function $f(\Theta) = \Theta$. Finally, the model can be completed by adding bounds on variables x_{ij} as explained in the previous section.

Since the four cases of functions f presented in the previous section are convex quadratic functions, (P) turns to a convex quadratic and linearly constrained optimization problem, which can be solved very quickly with state-of-the-art optimization solvers (e.g. CPLEX, Gurobi).

2.4 Several cases and hypothesis to test

For all the studied cases, the process is similar: we set an hypothesis that will change the estimation of each value x_{ij} in M. Then we recompute the sums of individuals in rows and columns and we adjust the constraints accordingly (e.g. bounds). Several types of solutions are searched, based on margin values and according to different hypothesis:

- Models without optimization:
  - *[Variable independence $H0$].* As explained in the introduction, this model is based on the $H0$ table of contingency; so $\hat{x}_{ij} = x_{i.}x_{.j}/x_{ii}$;
  - *[Global Illiteracy Rate GIR].* In the second model, we apply the global rate (%) of illiterate people observed in the whole USA on each type of population in every State;

  - *[Population Illiteracy Rate PIR]*. In this model, we take into account the part of illiteracy in each type of population (%);
  - *[Effect of Old Confederacy EOF]*. Here we add a historical hypothesis based on whether or not a given State belong (or was close) to the Confederation area during the civil war in USA (1861-1865); indeed, this could partly explain the capacity of a State to potentially integrate black or foreign population including language learning facilities. The probability of illiteracy is multiplied by 2 for the Confederate states and by 1.5 for a few bordering States which used to allow slavery and to belong to the Union however.
- Models with optimization but without any local information about value range (no information about uncertainty, in all of them, we only use recalculated margin values in rows and columns):
  - *[Raw Data RW]*. We try to find out an optimal solution;
  - *[Global Illiteracy Rate GIR]*. We look for an optimal solution using GIR;
- Models with local bounded values (the solver finds a solution given new local constraints of uncertainty in each matrix cell); we use recalculated margin values in rows and columns and we fix min and max bounds in each cell of M using two different formula 7 and 8; are concerned by this procedure:
  - *[Global Illiteracy Rate GIR]* data (for the whole USA);
  - *[Population Illiteracy Rate PIR]* data (by type of population);
  - *[Population Illiteracy Rate PIR]* data (by type of population) updated by *[Effect of Old Confederacy EOF]* data.

For the models with local bounded values, we apply a function to find the minimum $inf(\hat{x}_{ij})$ and maximum $sup(\hat{x}_{ij})$ values of the interval centered on the estimated $\hat{x}_{ij}$:

$$sup(\hat{x}_{ij}) = round(\hat{x}_{ij}.k^{1/length(\hat{x}_{ij})}) \tag{7}$$

and

$$inf(\hat{x}_{ij}) = round(2\hat{x}_{ij} - sup(\hat{x}_{ij})) \tag{8}$$

These functions (equations 7 and 8) allow to have symmetric intervals, larger for low values of $\hat{x}_{ij}$ (indeed $length$ counts the digits of $\hat{x}_{ij}$) depending on k.

For models dealing with uncertainty, we consider two cases:

- $k = 2$; for instance $[109 - 185]$ and $[6023 - 8835]$
- $k = 4$; resp. $[0 - 294]$ and $[2363 - 12495]$

III RESULTS

We focus on the relation between illiteracy and black people in USA in 1930. To do so, we compute a LS regression model between observed data and estimations we obtained. We study the coefficient of determination r^2 and the slope of the regression. Closer to 1, better the quality solution, for both indicators.

In the Figure 2, we can notice that using optimization increases the estimation quality when expert assumption is weak (case with $H0$ and Global Illiteracy Rate GIR). For other hypothesis (Population Illiteracy Rate PIR and Effect of Old Confederacy EOF), optimization has no effect compared to estimations based on accurate hypothesis from experts.

Globally, Figures 3 and 4 show that:

- The results are better for bounded optimization methods;
- The difference between objective functions is not marked; indeed they all somehow compute a sort of variability minimization;
- Generally, solutions are better with $k = 2$ (more narrow intervals);
- Most of the time, functions including the max operator seem to get better estimates;
- The best optimization reaches a coefficient of determination r^2 of 0.86 and a $slop$e of 0.76, both quite close to 1.

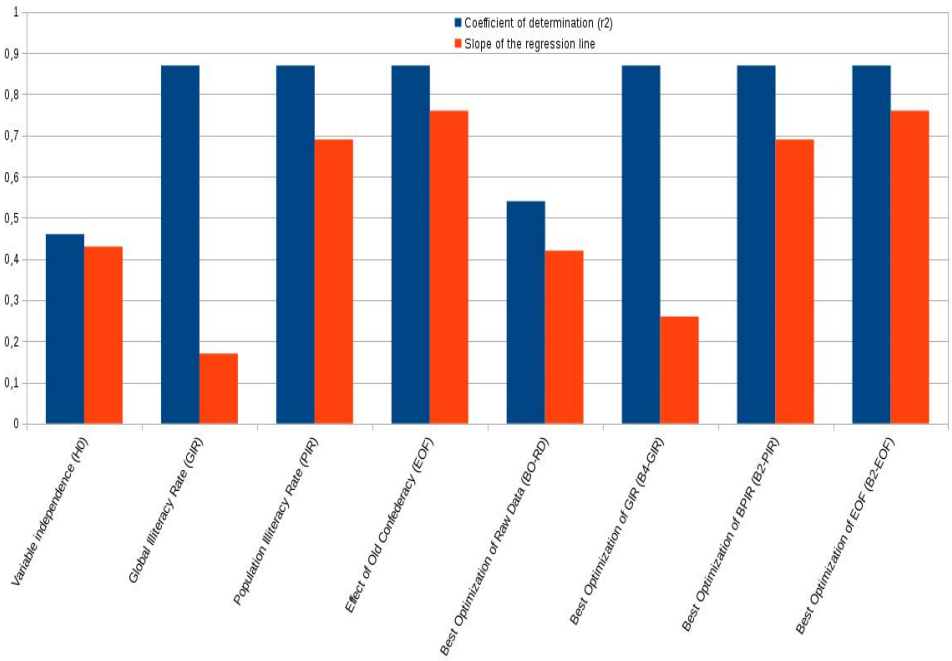


Figure 2: Comparison of different estimation methods according to the values of the coefficient of determination r^2 and of the slope of regression line (on left, 4 solutions without optimization; on right, 4 best solutions using optimization). Only the black population in USA in 1930 is considered.

IV CONCLUSION

In this paper, we explore how we can mix optimization methods and expert knowledge translated in statistical method to improve data estimation in contingency table downscaling. Although the results do not strongly demonstrate that the use of optimization is determinant on the result quality compared to other current methods from experts of the domain, it shows that, in certain conditions (weak assumptions or bounds knowledge from experts for instance), estimations can be improved, especially prediction quality (slope of the regression line between estimations and observed data).

Moreover, using optimization allows to find the most optimal solution among a large set of possible solutions, according to a given objective. These solutions can be then compared and participate in the data exploration.

Other experiments will be made, on the whole set of data whatever the type of population nativity, also on other types of data and with other compared methods (multiple regressions for example) to assess in which conditions, such a mixed approach may be useful for predicting data in contingency tables.

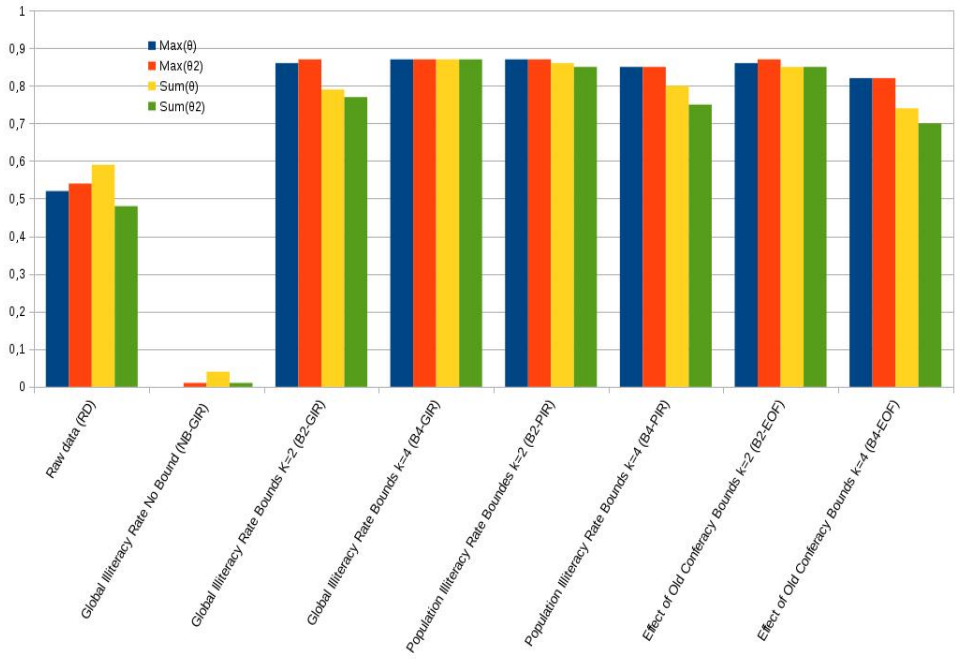


Figure 3: Comparison of different estimation methods according to the r^2 of the matrix M and to 4 different objective functions of θ. Only the black population in USA in 1930 is considered.

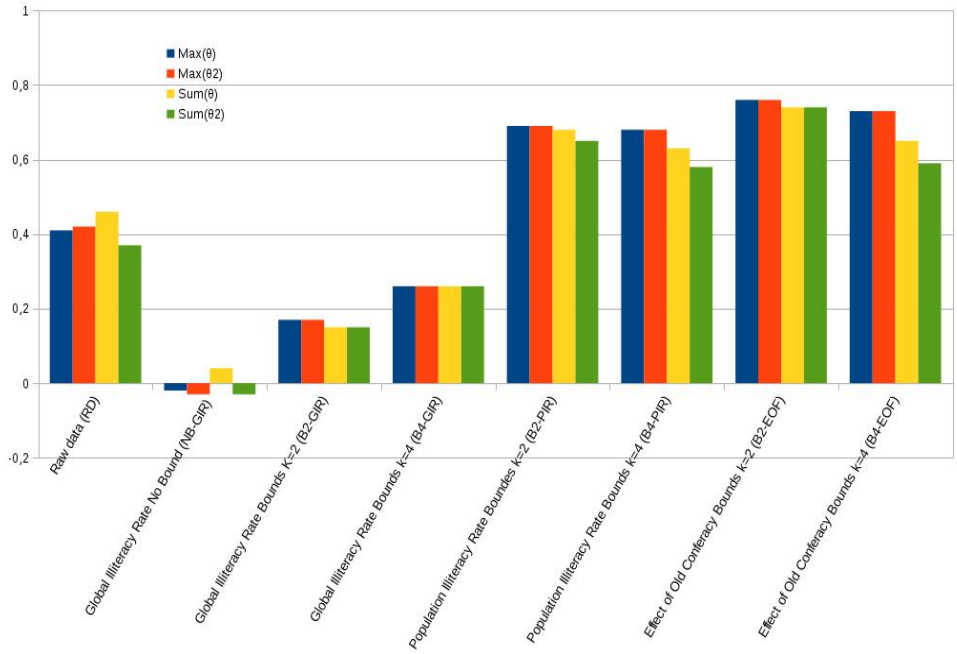


Figure 4: Comparison of different methods to estimate the slope of the regression line according to 4 different objective functions of θ. Only the black population in USA in 1930 is considered.

References

Bierkens M. F., A. F. P., de Willigen P. (2000). *Upscaling and Downscaling Methods for Environmental Research*. Springer, Series Developments in Plant and Soil Sciences.

Gehlke C., Biehl H. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association supplement 29*, 169–170.

Josselin D., Mahfoud I., Fady B. (2008). Impact of a change of support on the assessment of biodiversity with

shannon entropy. In *Spatial Data Handling, SDH'2008"*, Montpellier, June, 23-25, pp. 109–131.

King G. (1997). *A solution to the ecological inference problem. Reconstructing individual behaviour from aggregate data*. Princeton University Press.

King G., Rosen O., Tanner A. M. (Eds.) (2004). *Ecological Inference. New Methodological Strategies*. Cambridge University Press.

Robinson W. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review 15*, 351–357.

Tranmer M., Steel D. (1998). Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A 30*, 817–831.

Downscaling of soil moisture with area-to-point geographically weighted regression kriging and uncertainty analysis

Yan Jin[1,2], Yong Ge [*1], Jianghao Wang[1], Yuehong Chen[1,2], Xudong Guan[1,2]

[1] State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

[2] University of Chinese Academy of Sciences, Beijing 100049, China

*Corresponding author: gey@lreis.ac.cn

Abstract

Combine the visible/infrared remote sensing and passive microwave remote sensing has demonstrated its potential for capturing the high spatial resolution of the near-surface soil moisture. A geostatistical approach by integrating geographically weighted regression (GWR) and area-to-point kriging (ATPK) was proposed to downscale soil moisture. The proposed area-to-point geographically weighted regression kriging (GWATPRK) was verified on the AMSR-2 soil moisture product at 25 km resolution to obtain the downscaled soil moisture of 1 km spatial resolution. The results showed the GWATPRK can enhance the accuracy of the downscaled estimates. The best downscaled estimates owned the root mean square error value of $0.072m^3/m^3$ and the mean absolute error value of $0.056m^3/m^3$.The different auxiliary variables would affect the accuracy of downscaled estimates. Comparison of the downscaled estimates using different covariates with the *in situ* observations, the topography-corrected land surface temperature can improve the downscaled performance in the study area.

Keywords: downscaling, soil moisture, area-to-point kriging, geographically weighted regression, AMSR-2, MODIS

I. INTRODUCTION

Soil moisture (SM) is one of fundamental land surface parameters for modelling eco-hydrological processes and understanding impacts and conservation of ecosystems (Srivastava et al., 2013). Although the ground measurement methods (Bogena et al., 2010) can obtain the SM information with depth monitoring and high accuracy, the expanded applications at large scales included regional or global scales are impractical. The remote sensing of surface SM is considered to fill this gap, including visible/infrared remote sensing and microwave remote sensing. Microwave remote sensing technique has been widely applied to monitor SM owing to high SM sensitivity and ignorance of atmospheric conditions. Most of these satellites can provide the SM products at a fine temporal resolution (1 to 3 days) and a coarse spatial resolution (about tens of kilometers). The coarse spatial resolution couldn't satisfy the application requirements (Ge et al., 2015a). One the other hand, it is difficult to match the large-scale SM to the fine-scale ground observations in the remote sensing validation. Therefore, it is necessary to improve the spatial resolution of remote sensing products of SM.

A number of methods have been developed to downscale the passive microwave remote sensing products (Kim and Hogue, 2012; Yang et al., 2016). Combine the visible/infrared remote sensing and passive microwave remote sensing that has become more general, because it benefits the fine resolution of the former and takes advantage of the latter

which is high sensitive to the SM. Many SM downscaling algorithms focused on the relationship between the SM and the relative variables (Piles et al., 2011). Meanwhile, the geostatistical methods which consider the spatial correlation have received extensive concern in downscaling field (Atkinson et al., 2008) and growing applications in SM. The geostatistics perform well in spatial prediction and uncertainty analysis (Chiles and Delfiner, 1999). And the area-to-point regression kriging (ATPRK) (Kerry et al., 2012; Wang et al., 2015) is a flexible and well-performing downscaling method which incorporates regression kriging (Hengl, et al., 2007; Ge et al., 2015b) and ATPK (Kyriakidis, 2004). It can consider both the correlated variables and the change of support problem during the downscaling process. ATPRK has been shown the potential in downscaling the irregular geographical units (Zhang et al., 2014) but still less in remote sensing images and never be used in SM.

In this paper, a new geostatistical approach by integrating geographically weighted regression (GWR) (Fotheringham et al., 1997) and area-to-point kriging (ATPK) was proposed to downscale SM with high resolution visible/infrared variables (i.e., the Moderate Resolution Imaging Spectroradiometer (MODIS) products). The proposed area-to-point geographically weighted regression kriging (GWATPRK) as a downscaling strategy can improve SM estimates, because it retains the advantages of ATPRK and considers spatial nonstationarity relationship of variables. The approach was illustrated by an application for downscaling the SM remote sensing product of the Advanced Microwave Scanning Radiometer 2 (AMSR-2) (Imaoka et al., 2010) and a validation with *in situ* observations in the Heihe Water Allied Telemetry Experimental Research experiment (HiWATER) (Li et al., 2013).

II. STUDY AREA AND DATA DESCRIPTION

A. Study area

The study area is the upper reaches of the Heihe River Basin (HRB) in northwestern China, within latitudes 37.66 to 39.16 N, and longitudes 98.50 to 101.25 E. It includes the Babao River basin which is one of foci experimental areas of the comprehensive eco-hydrological program HiWATER. The district covers approximately 2452 km^2 with an elevation ranging from 2640 to 5000 m and is in a typical landscape of cold region with natural grassland as main vegetation. The topography of the study area with ground-based site locations and the AMSR-2 grid pixels (SM product of one day) was shown in Figure 1.

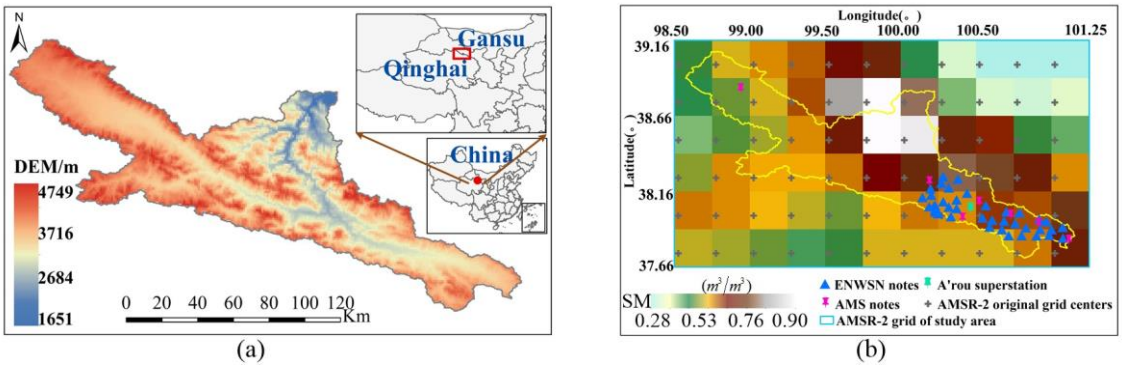


Figure 1: (a) The study area; (b) the ground-based site locations and AMSR-2 grid pixels.

B. In situ observations

In the upstream area of HRB, seven ordinary automatic meteorological stations (AMSs) and a superstation at A'rou have been installed to capture the atmospheric states. The Ecological and Hydrological Wireless Sensor Network (ENWSN) has been designed and deployed to monitor

the surface SM (Li et al., 2013). There are forty nodes of ENWSN to obtain the variation dynamically at depths of 4cm, 10cm and 20cm. In this study, one month (July 20-August 21, 2014) of *in situ* SM observations were collected. For validating the downscaled SM estimates from the AMSR-2 SM product, the average observations with 4 cm depth during satellite passes were considered, i.e., from 9:30 p.m. to 2:30 a.m. for the descending. Specifically, one period was adopted for validation because of the availability of the auxiliary MODIS images, i.e., 21 August, 2014 which was the day of Year DOY233.

C. Remote sensing data

The Level 3 descending surface SM product of AMSR-2 (version 001) at a 25 km spatial resolution and the Version 5 MODIS night products at a 1 km spatial resolution were employed. The daily SM product data were resampled into 25km×25km regular grids using the nearest neighbor resampling technique for completely overlapping the MODIS pixels. The daily LST was extracted from the MYD11A1 and MOD11A1 products. The NDVI was acquired from MYD13A2 product at a temporal resolution of 16 days. All of the MODIS data were projected and extracted consistently with large scale AMSR-2 data and the aggregated coarse MODIS images achieved full coverage. The 16-day NDVI product is cloud free. However, the single LST product owned a bad cloud free condition within the study area, it needed to combine the Terra and Aqua products. The data coverage fraction was 86% on DOY 233 and it was already high in available dates. Then the spline interpolation was used to interpolate the pixel values in uncovered region for obtaining the fully downscaled estimates.

III. DOWNSCALING STRATEGY

Assume that vector $Z(S_i)$ represents coarse SM at *n* spatial observation pixels $S_i=(U_i,V_i)$ ($i=1,2,\cdots,n$) and $Z(s_j)$ represents fine SM at nF^2 spatial observation pixels $s_j=(u_j,v_j)$ ($j=1,2,\cdots,nF^2$). *F* is the ratio between the coarse resolution and the fine resolution. In which, (U_i,V_i) and (u_j,v_j) state the geographic positions (latitudes, longitudes) of the centers of the corresponding pixels, respectively. The proposed downscaling strategy contains two phases. It first establishes GWR model among the SM and the visible/infrared variables to predict the spatial trend of the SM at fine resolution; then applies ATPK to downscale the regression residuals to obtain the fine resolution predictions. The general form of GWATPRK estimate is:

$$Z(s_j)=m(s_j)+R(s_j) \tag{1}$$

where $m(s_j)$ is the deterministic part which can be estimated by GWR and $R(s_j)$ is the ATPK parts for regression residuals. The detailed descriptions were displayed in the following sections.

The GWR model is a local regression that generates local coefficients to each pixel and has been successfully used in SM (Yu et al., 2008). Many Researches have indicated the feasibility of applying the universal relationship at different scale resolutions (Zhang et al., 2014). Let $x_k(S_i)$ ($k=1,\cdots,p$) be the value of the auxiliary variables at pixel S_i and $\beta_l(U_i,V_i)$ ($l=0,1,\cdots,p$) be regression coefficients with geographic position (U_i,V_i). The corresponding parameters of the fine resolution are $x_k(s_j)$ at pixel s_j and $\beta_l(u_j,v_j)$ with position (u_j,v_j). The changing regression coefficients $\beta_l(U_i,V_i)$ at different position can be estimated by using the least square method in basic GWR, in which one of the key steps is to choose a distance-decay function for representing the strength of the connectivity between pixels. The Gaussian function which is one of the most common distance-decay functions was selected in the following experiments. Furthermore, the GWR with a locally compensated

ridge term (Lu et al., 2014) was employed to address local collinearity problems in basic GWR. The relationship among the SM and the auxiliary variables at coarse resolution was given:

$$Z(S_i) = \beta_0(U_i, V_i) + \sum_{k=1}^{p} \beta_k(U_i, V_i) \cdot x_k(S_i) \quad (2)$$

Meanwhile, the regression residual was computed using the following equation:

$$R(S_i) = Z(S_i) - [\beta_0(U_i, V_i) + \sum_{k=1}^{p} \beta_k(U_i, V_i) \cdot x_k(S_i)] \quad (3)$$

Based on the hypothesis of the scale invariant, the relationship in Eq. 2 can be applied at fine spatial resolution:

$$m(s_j) = \beta_0(u_j, v_j) + \sum_{k=1}^{p} \beta_k(u_j, v_j) \cdot x_k(s_j) \quad (4)$$

where $\beta_l(u_j, v_j)$ equals $\beta_l(U_i, V_i)$ when the pixel s_j is covered by the large one S_i.

If the regression model was perfect, there would be no bias between the fitting result and the observations at coarse resolution, meaning that the regression part (Eq. 4) can attain the aim for downscaling. Because of the impracticality to search such ideal regression model, the regression residual (Eq. 3) should not be neglected. Consider that adding the residual into the regression prediction at fine resolution directly would ignore the change of supports, it was no downscaling step actually. The ATPK method was employed to downscale the regression residual to acquire the prediction at fine resolution by summing the weighted residuals at coarse resolutions. If a given pixel s_j was estimated as a combination of the $R(S_h)$ ($h = 1, 2, \cdots, m$) in m large neighboring pixels, the ATPK model can be written as followed:

$$R(s_j) = \sum_{h=1}^{m} \lambda_h \cdot R(S_h) \quad (5)$$

where λ_h are the weight coefficients for the prediction of the fine resolution and satisfy $\sum_{h=1}^{m} \lambda_h = 1$. The weight coefficients can be estimated by minimizing the prediction error variance, in which the main step for the implementation of ATPK is to obtain the point support semivariogram. A presented deconvolution procedure (Goovaerts, 2008) was utilized here.

Combine the regression predictions and downscaled fine residuals from the Eq. 4 and Eq. 5, the general form of GWATPRK estimate (Eq. 1) can be further written as:

$$Z(s_j) = \beta_0(u_j, v_j) + \sum_{k=1}^{p} \beta_k(u_j, v_j) \cdot x_k(s_j) + \sum_{h=1}^{m} \lambda_h \cdot \{Z(S_h) - [\beta_0(U_h, V_h) + \sum_{k=1}^{p} \beta_k(U_h, V_h) \cdot x_k(S_h)]\} \quad (6)$$

where the symbols of parameters stay the same as stated above. The coarse SM can be downscaled to the fine one through the Eq. 6.

IV. RESULTS AND DISCUSSION

A. Downscaled soil moisture maps

There were three groups of auxiliary variables as different independent variable(s) in downscaling. For comparison, a mentioned commonly used downscaling strategy known as ATPRK and a quadratic regression model (QRM) (Piles et al., 2011) which is one of the SM

downscaling models were adopt. The ATPRK was employed an ordinary linear regression.

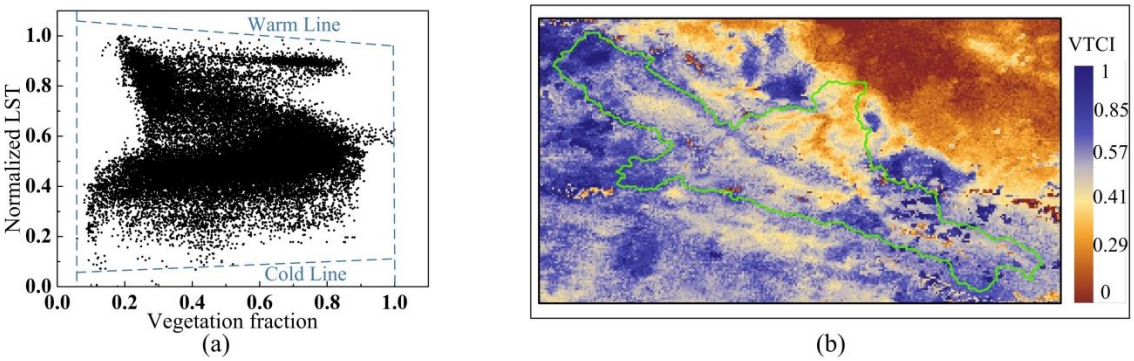


Figure 2: (a) The vegetation index-surface temperature trapezoidal feature space and (b) the calculated VTCI at 1km resolution on DOY 233

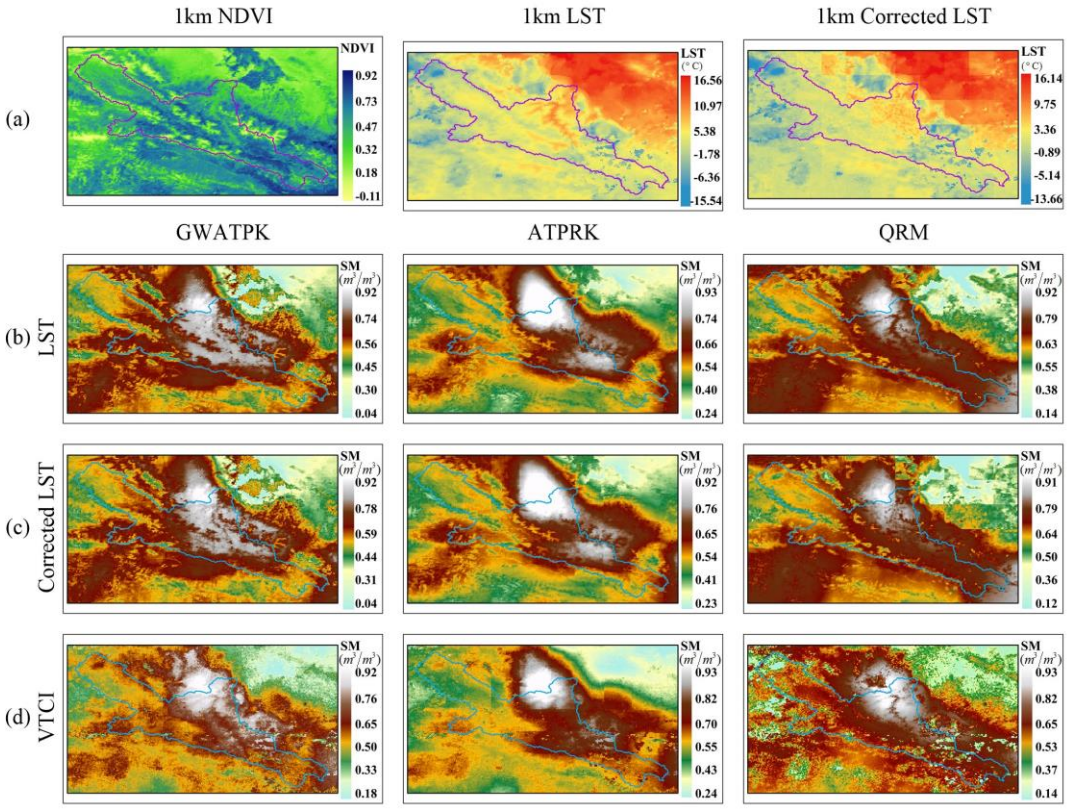


Figure 3: The images at 1km resolution on DOY 233. (a) The auxiliary variables; (b) the downscaled SM with LST; (c) the downscaled SM with corrected LST; (d) the downscaled SM with VTCI.

Firstly, the LST and NDVI were employed as the auxiliary variables which have already been closely linked to SM (Chauhan et al., 2003). The correlation analysis figured out that the LST and NDVI have correlation coefficients of 0.34 and 0.41, respectively. The variance inflation factors of LST and NDVI were both reported less than 10 which revealed the reasonability for employing these two independent variables without considering the multicollinearity. Secondly, considering the topographic effects in hilly surrounding area, a topography-corrected land surface temperature (Merlin et al., 2013) was used to replace the LST in the first group. Thirdly, we chose the vegetation temperature condition index (VTCI) which has a positive correlation with SM. It has been demonstrated to increase downscaling performance with VTCI as the only input (Peng et al., 2016). The normalized LST and the

vegetation fraction cover were employed to describe the feature space. Figure 2a and 2b showed the physically meaningful trapezoidal feature space and the calculated VTCI at 1km resolution, respectively.

The three downscaling approaches were realized on the three variables group. Figure 3a showed images corresponding to the auxiliary data at 1km resolution, meanwhile Figure 3b, 3c and 3d showed downscaled images for different approaches with varying variables. The downscaled results have a consistent tendency. The white parts in the images owned the highest SM values which were located in the low elevation district of the upstream area. All top right corners displayed the lowest SM values in accordance with the conditions of the highest temperature and less vegetation coverage. The GWATPRK and ATPRK could capture more similar spatial information of SM with the coarse image. And the downscaled images of GWATPRK represented stronger variability. The GWATPRK could obtain less minimum values of downscaled estimates which were closer to the minimum *in situ* observation. Furthermore, the values greater than around 0.7 m^3/m^3 of downscaled SM estimates may not accord with the natural cognition of the cold region. It caused by the large original SM values of the AMSR-2 data. And to focus on the downscaling methods, we skipped correcting for the AMSR-2 SM product here. In the following comparison, the GWATPRK with changed regression coefficients at each coarse pixel has obtained the good quality of which downscaled values were closer to the ground observations expectably.

B. Validation and comparison

To assess the accuracy of the downscaling strategy, the *in situ* observations were compared with the downscaled SM, including the mean absolute error (*MAE*) (m^3/m^3), the standard deviation of the error (*MSD*) (m^3/m^3), the root mean square error (*RMSE*) (m^3/m^3) and the correlation coefficient (*R*). It desires the downscaled estimates to be more approximate to the *in situ* measurements. The less *MAE*, less *RMSE*, and higher *R* would be expected. Considering one day of most growing season days was chosen and the corresponding AMSR-2 SM values were all greater than the *in situ* measurements. The mean of these differences between them would be removed from the downscaled estimates for a more suitable comparison with ground observations. The downscaled SM estimates were plotted against the *in situ* measurements and the accuracy of SM results of different downscaling approaches described in Figure 4.

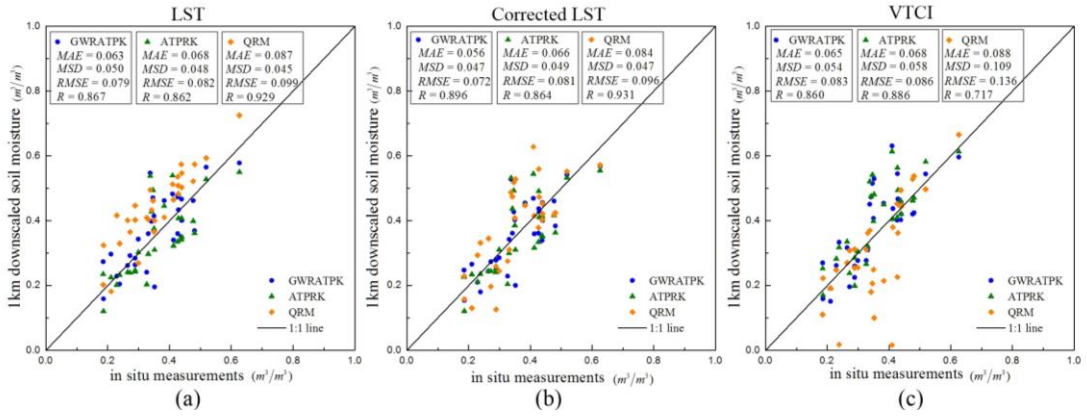


Figure 4: Downscaled soil moisture estimates versus *in situ* measurements and the summary of the comparison results. (a) The downscaled soil moisture with uncorrected LST; (b) the downscaled soil moisture with corrected LST; (c) the downscaled soil moisture with VTCI.

Although the downscaled estimates of the QRM had the higher *R*, the *MAE* and the *RMSE* were also higher. And its estimates were almost all over or below the 1:1 line. The results demonstrated that the GWATPRK downscaling strategy achieved better downscaled estimates

than other two methods in every variables group. The corrected LST can improve the estimates of SM in each method. The VTCI just reduced the *MAE* of QRM and increased the *R* of ATPRK, albeit the downscaled result of GWATPRK was better than other two. It was also clear that with the least *MAE* value of 0.056 m^3/m^3, a less *MSD* value of 0.047 m^3/m^3, the least *RMSE* value of 0.072 m^3/m^3 and a higher *R* value of 0.896, the GWATPRK downscaling method with corrected LST resulted in the best performance.

The downscaled error mainly comes from the input images, the models and the scale effect during the downscaling process (Zhao et al., 2015). The latter two are always complex and complicated, which are still affected by all the participating variables. Because that the study area is a hilly surrounding area, the topography-corrected LST is actually a less error variable for impacting SM than the original LST by eliminating the terrain influence. It worked well. In addition, the VTCI as a synthetic index didn't improve the accuracy expectably, in which the LST has not been corrected. The VTCI of ignoring the topography may not adequately capture the interactions with SM in this mountainous area. Meanwhile, it is best to apply the VTCI with wide range required vegetation coverage. It also reminded us again to use the temperature-vegetation indexes with more analyses of the area conditions. To select the appropriate variable will conduce to enhance the downscaled accuracy.

V. CONCLUSION

Downscaling passive microwave remote sensing SM product is an essential problem in SM monitor. To improve SM spatial resolution, GWATPRK was proposed to downscale SM by integrating GWR and ATPK. We evaluated different downscaling methods on the AMSR-2 SM product in the upper reaches of the HRB. The results showed the GWATPRK can produce downscaled images at a fine spatial resolution with greater qualities for heterogeneous region. And it would improve the accuracy of the downscaled estimates by using the topography-corrected LST in this study area. The input images as a significant part need considerations in downscaling process which would affect the accuracy of the downscaled results. Moreover, for an operational downscaled SM approach, further validation work will be carried out in the future studies.

ACKNOWLEDGMENTS

This work was supported by the Key Program of the National Natural Science Foundation of China (No. 41531174). The authors would like to thank all the data producers, and the anonymous reviews for their comments and suggestions to improve the quality of the paper.

References

Atkinson P M, Pardo-Iguzquiza E, Chica-Olmo M. (2008). Downscaling cokriging for super-resolution mapping of continua in remotely sensed images. *Geoscience and Remote Sensing, IEEE Transactions on* 46(2): 573–580.

Bogena H R, Herbst M, Huisman J A, et al. (2010). Potential of wireless sensor networks for measuring soil water content variability. *Vadose Zone Journal*, 9(4): 1002–1013.

Chauhan N S, Miller S, Ardanuy P. (2003). Spaceborne soil moisture estimation at high resolution: a microwave-optical/IR synergistic approach. *International Journal of Remote Sensing*, 24(22): 4599–4622.

Chiles, J. P., and Delfiner, A. (1999). Geostatistics: Modelling spatial uncertainty: Wiley Interscience. *New York*.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). Geographically weighted regression: the analysis of spatially varying relationships. *John Wiley & Sons*.

Ge, Y., Wang, J. H., Heuvelink, G. B. M., Jin, R., Li, X., and Wang, J. F. (2015a). Sampling design optimization of a wireless sensor network for monitoring ecohydrological processes in the Babao River basin, China.

International Journal of Geographical Information Science, 29(1), 92–110.

Ge, Y., Liang, Y., Wang, J., Zhao, Q., and Liu, S. (2015b). Upscaling sensible heat fluxes with area-to-area regression kriging. *Geoscience and Remote Sensing Letters, IEEE*, 12(3), 656–660.

Goovaerts P. (2008). Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Mathematical Geosciences*, 40(1): 101–128.

Hengl, T., Heuvelinkb, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33,1301–1315.

Imaoka K, Kachi M, Kasahara M, et al. (2010). Instrument performance and calibration of AMSR-E and AMSR2. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science*, 38(8): 13–18.

Kerry R, Goovaerts P, Rawlins B G, et al. (2012). Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. *Geoderma*, 170: 347–358.

Kim, J., Hogue, T. S. (2012). Improving spatial soil moisture representation through integration of AMSR-E and MODIS products.*Geoscience and Remote Sensing, IEEE Transactions on*, *50*(2), 446-460.

Kyriakidis P C. (2004). A Geostatistical Framework for Area-to-Point Spatial Interpolation. *Geographical Analysis*, 36(3):259–289.

Li X, et al. (2013). Heihe watershed allied telemetry experimental research (HiWATER): Scientific objectives and experimental design. *Bulletin of the American Meteorological Society*, 94(8): 1145–1160.

Lu, B., Harris, P., Charlton, M., & Brunsdon, C. (2014). The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. *Geo-spatial Information Science*, *17*(2), 85-101.

Merlin O, Escorihuela M J, Mayoral M A, et al. (2013). Self-calibrated evaporation-based disaggregation of SMOS soil moisture: An evaluation study at 3km and 100m resolution in Catalunya, Spain. *Remote sensing of environment*, 130: 25–38.

Peng J, Loew A, Zhang S, et al. (2016). Spatial downscaling of satellite soil moisture data using a vegetation temperature condition index. *Geoscience and Remote Sensing, IEEE Transactions on* 54(1): 558–566.

Piles M, Camps A, Vall-Llossera M, et al. (2011). Downscaling SMOS-derived soil moisture using MODIS visible/infrared data. *Geoscience and Remote Sensing, IEEE Transactions on* 49(9): 3156–3166.

Srivastava P K, Han D, Ramirez M R, et al. (2013). Machine learning techniques for downscaling SMOS satellite soil moisture using MODIS land surface temperature for hydrological application. *Water resources management*, *27*(8), 3127-3144.

Wang Q, Shi W, Atkinson P M, et al. (2015). Downscaling MODIS images with area-to-point regression kriging. *Remote Sensing of Environment*, 166: 191–204.

Yang K, Zhu L, Chen Y, et al. (2016). Land surface model calibration through microwave data assimilation for improving soil moisture simulations. *Journal of Hydrology*, 533: 266-276.

Yu G, Di L, Yang W. (2008). Downscaling of global soil moisture using auxiliary data//Geoscience and Remote Sensing Symposium. *IGARSS 2008. IEEE International. IEEE*, 2008, 3: III-230-III-233.

Zhang J, Atkinson P, Goodchild M F. (2014). Scale in spatial information and analysis. *CRC Press*.

Zhao W, Li A. (2015). A comparison study on empirical microwave soil moisture downscaling methods based on the integration of microwave-optical/IR data on the Tibetan Plateau. *International Journal of Remote Sensing*, 36(19-20): 4986–5002.

Uncertainty in 3D geodata

A proposal of a statistical test to control positional accuracy by means of 2 tolerances simultaneously

José Rodríguez-Avi[*1], Francisco Javier Ariza-López[1]

[1] University of Jaén, Jaén, Spain

*Corresponding author: jravi@ujaen.es

Abstract

We propose a new method for positional accuracy quality control of spatial data. This method is valid for 1D, 2D, 3D and nD dimensional data. Normality of errors is not required. The method is an exact statistical hypothesis testing based on multinomial distribution. The proportions of the multinomial distribution are defined by means of several metric tolerances. The proposed statistical test is exact, so the p-value can be derived by exploring a space of solutions and summing up the probabilities of each isolated case of this space. The performance of the test has been analyzed by means of a simulation procedure. The validity and the power of the contrast seem to be good enough. Some application examples are presented for the case of working with two tolerances.

Keywords

Multinomial distribution, Spatial Data, positional accuracy quality control

I INTRODUCTION

Spatial data are data referring to features that have a position in space, and for this reason positional quality is one of the most desirable characteristics of spatial data sets (SDS). Positional accuracy of SDS has traditionally been evaluated using control points. Following this idea there are a great many statistical Positional Accuracy Assessment Methodologies (PAAM) like: National Map Accuracy Standard (NMAS) (USBB, 1947), Engineering Map Accuracy Standard (EMAS) (ASCE, 1983), National Standard for Spatial Data Accuracy (NSSDA) (FGDC, 1998), STANAG 2215 (STANAG, 2002), ASPRS Accuracy Standards for Large-Scale Maps (ASLSM) (ASPRS, 1990), the new ASPRS Positional Accuracy Standards for Digital Geospatial Data (ASPRS, 2015), and many others.

The majority of PAAMs take as an underlying hypothesis the Gaussian distribution (Normal Distribution) of positional errors, and here are some examples taken from the existing PAAM. The EMAS shows a control (acceptance/rejection) perspective and proposes a set of statistical hypothesis control test based on the normality of data, the first one for testing the presence of bias (a t-Student test) and the second one for comparing variational behaviour of data (a Chi squared test). In another way, the NSSDA shows an estimation perspective, determining an accuracy value scaled to a confidence value (95%) and leaving to the user the final decision about the appropriateness of the data. In the NSSDA the planimetric case is computed using a Chi2, which can only be used under the assumption that both X and Y error distributions follow a Normal Distribution where homogeneity of variances is needed, and the altimetric case is computed using the value 1.96 as a scaling factor, which is only valid under the assumption that Z error distribution follows a Normal Distribution Function. Another PAAM that requires the strict normality of errors is the procedure stated in ISO 3951-1 (ISO 2013b), being applied in some European countries. The ISO 3951-1 method is a sequence of tests allowing the acceptance/rejection of lots that can be applied to spatial data based on their positional behaviour.

It is also important to notice here that prior to any statistical analysis of the data assuming normality, the outliers must be eliminated. The elimination of outliers in parametric distributions is a usual issue, and for the case of the normal distribution there are many methods, such as those presented in ISO 16269-4 (ISO 2010) (for more detail on outliers elimination methods see Barnett and Lewis 1994).

But several studies indicate that this hypothesis (normality of errors) is not true. There are some proposals, for instance, in the case of Global Navigation Satellite Systems error distribution, the Rayleigh distribution (Logsdon, 1995); and for the case of geocoding errors, a log-normal distribution (Cayo and Talbot, 2003). For the case of vertical errors in digital elevation models there are many references (e.g. Bonin and Rousseaux 2005, Oksanen and Sarjakoski 2006) indicating that error distribution is not Normal, and for this reason it is proposed to perform and express the results of quality control checks by means of percentiles (Maune 2007) of the observed distribution. The latter was introduced by the ASPRS into the Guidelines for reporting vertical accuracy for Lidar data (ASPRS, 2004) and the most recent standard from the ASPRS (ASPRS, 2015) follows this method by informing separately the vegetated vertical accuracy (supposed to be non-Gaussian distributed) and the non-vegetated vertical accuracy (supposed to be normally distributed). This situation is somewhat confusing for non-expert users.

The Normal Distribution Function and other distribution functions mentioned above are parametric distribution functions, which means that they can be determined by a set of parameters, for instance the mean (μ) and the variance (σ^2) in the case of the Normal Distribution Function. Parametric models summarize the behaviour of a population by means of a mathematical formula and parameters which control its shape (e.g. for the Normal Distribution Function the mean gives the position and the deviation the more or less flattened shape of the curve). Parametric distribution functions are very convenient when dealing with scarce data and computing capabilities are limited, as in the past. But now we live surrounded by very large data sets, and by very large computing capabilities, so we can use the observed distribution functions (Free-Distribution Functions or Parameters-Free-Distribution Functions) without major problems.

The PAAMs based on Free-Distribution Functions are scarce in positional accuracy assessment. For instance, the use of percentiles is proposed for dealing with height data of DEM when capture (e.g. by Lidar) is performed on a vegetal cover. However, the use of percentiles is merely descriptive, and no method for quality control is proposed. Recently, Ariza-López & Rodríguez-Avi (2014a) proposed a method that can be applied to data belonging to free-distribution functions. A step further would require a statistical method allowing control in the same way (mean value, variation, control of outliers) as performed when data come from a Normal Distribution Function.

Our aim is to propose a general positional accuracy control method for dealing with data following any kind of distribution function. It is a method based on the observed distribution function of the data and, in this way, can be applied to 1D, 2D or 3D data without the limitations of homogeneity of variances stated in traditional methods (e.g. EMAS, NSSDA, STANAG 2215, etc.). To achieve this aim we propose a method based on proportions of a multinomial distribution function in order to establish a strict control over data coming from any distribution function. The control is multiple and can test jointly proportions corresponding to tolerances related with, for instance, median values, extreme values (e.g. 95%), or the amount of outliers existing in the data set. The proposed control is based on an exact test, in the same way as the Fisher's exact test (Fisher 1922, Freeman and Halton 1951, Müller 2001).

II PROPOSED METHOD

In order to analyse a pass/fail model applied to positional accuracy control, an approximation is given in terms of the Binomial distribution (Ariza-López & Rodríguez-Avi, 2014). For each sampling point taken in a k-dimensional space, the value:

$$E_i = \sqrt{\sum_{i=1}^{k} (x_i - x_i^T)^2}$$

is calculated (where $(x_1^T, \dots, x_k^T)$ is the "exact value).

Given a tolerance T, we count the number of sampling points where E_i is greater than T. This can be seen as a realization of a Binomial distribution with parameters N and p. In this case we consider a hypothesis test where the null hypothesis is: the proportion of defectives points is less or equal than p_0, against H_1: the proportion of defectives is greater than p_0. In consequence, a right hand unilateral test is proposed and H_0 is rejected if $p = P[X > x | X \sim B(N, p_O)] < \alpha$ where x is the number of defectives (i.e. number of points whose error is greater than the previously specified tolerance) found in the sample.

This approximation is a pass/fail procedure and implies that we are not able to distinguish the "degree of bad quality" of a defective point (in position). In several circumstances is interesting to propose an ordination in the degree of fail. We propose a gradation in the degree of positional defectiveness, splitting the interval, the excess of error greater than the tolerance, into two parts. In this way we consider two metric tolerances defining three intervals for the positional error values measured in the sample. Each interval defines a quality level, and we can determine the minimum percentage of "correct" points in the first level (best level), and the maximum percentage of points in each of the other two levels (worse levels). For instance, and in order to control outlier points, we can stablish a percentage of at least a $p_1\%$ of points with an error less than a tolerance T_1 and at most a $p_2\%$ of points with an error between tolerances T_1 and T_2 $(T_2 > T_1)$ and, in consequence, at most a percentage of $p_3\% = 100 - p_1\% - p_2\%$ of points whit error greater than tolerance T_3. So, we classify the sample into three (instead of two, as before) categories, and consequently, the multinomial approximation instead of the binomial approximation is adequate.

Let us recall that the multinomial distribution is a multivariate extension of a binomial distribution when we classify the result of an experiment into more than two categories. So if we realize N independent experiments where we classify the results for exactly one of 3 categories, with probabilities π_1, π_2, π_3 , and $\pi_1 + \pi_2 + \pi_3 = 1$, then the mass probability function of a such multinomial $\mathcal{M}(N, \pi_1, \pi_2, \pi_3)$ is given by:

$$P[(X_1 = N_1, X_2 = N_2, X_3 = N_3) = \binom{N!}{N_1!\, N_2!\, N_3!} \pi_1^{N_1} \pi_2^{N_2} \pi_3^{N_3}$$

Where N_i is the number of points that belongs to the category i, which has a probability π_i.

We use the multinomial distribution to propose an exact test to decide whether a specified hypothesis is false or not. Firstly we decide two tolerance levels, T_1 and $T_2, T_2 > T_1$ to decide if a point is adequate $(E_i < T_1)$, slightly inadequate $(T_1 < E_i < T_2)$ or roughly inadequate $(E_i > T_2)$.

The null hypothesis implies than the proportion of points in each category is known. In consequence:

- $\mathbb{H}_0$: The sampling statistics has a multinomial distribution with parameters N, $\pi_1^0, \pi_2^0, \pi_3^0$ where π_i^0 is the hypothesized probability for category $i, i = 1, 2, 3$

We propose an unilateral test (we will only rejected the null hypothesis if the true error distribution is worse, and this situation takes place when the proportion of elements with tolerance less than T_1 is less than P_1, or when the other two proportions account for more than P_2 or P_3, because this implies a worsening in tails.

A sampling of size N is dropped, and we count the number of point that belongs in each category. The sampling statistics is $\boldsymbol{S} = (N_1, N_2, N_3)$, so that $N_1 + N_2 + N_3 = N$, where N_i is the number of points that, in a sample of size N belongs to category $i, i = 1, 2, 3$

Under the null hypothesis we can calculate the exact probability of obtaining a such point S and the probability of every point worse than S. As a result, to calculate the p–value we sum the probability, under the null hypothesis, of elements $(M_1,\ M_2,\ M_3)$ that verify:

- $M_1 < N_1$
- $M_1 = N_1$ y $M_2 < N_2$

And we reject the null hypothesis if the p-value obtained is lesser than α.

To prove the validity of our proposal we show two complementary simulations:

- The first one is in order to prove the validity of the contrast with respect to the Type I error.
- The second one is in order to prove the power of the contrast, that it is to say, the probability of rejecting the null hypothesis when this hypothesis is false.

Case $\boldsymbol{\pi_0} = (\mathbf{0.8, 0.15, 0.05})$	**Critical value**	**Sampling size (n)**						
		20	**50**	**75**	**100**	**150**	**200**	**300**
	0.5	0.5190	0.5205	0.5195	0.4980	0.4950	0.4885	0.4725
	0.3	0.3345	0.3015	0.3090	0.2880	0.2875	0.2905	0.2815
	0.2	0.2205	0.2000	0.2030	0.2025	0.1925	0.1960	0.1960
	0.1	0.1095	0.1020	0.0970	0.1050	0.0975	0.0890	0.1010
	0.05	0.0655	0.0565	0.0555	0.0470	0.0520	0.0450	0.0505
	0.01	0.0115	0.0130	0.0120	0.0095	0.0085	0.0090	0.0095

Table 1 - Validity of the exact contrast with respect to the Type I error

Table 1 shows the proportion of simulated points obtained when the null hypothesis is, in fact, the true population distribution (in this case, $\pi_0 = 0.80, 0.15, 0.05$) and Table 2 shows the power of the test when the true population distribution is worse than the null hypothesis, and in consequence, the test would be rejected.

Case $\boldsymbol{\pi_0} = (\mathbf{0.7; 0.25; 0.05})$	**Critical value**	**Sampling size (n)**						
		20	**50**	**75**	**100**	**150**	**200**	**300**
		True population probability: $\boldsymbol{\pi^1} = (\mathbf{0.69; 0.01; 0.30})$						
	0.5	0.6235	0.6235	0.6930	0.6255	0.6510	0.6380	0.6705
	0.3	0.4240	0.3865	0.3925	0.4455	0.4295	0.4660	0.4865
	0.2	0.2560	0.2810	0.2900	0.2870	0.3045	0.3435	0.3340
	0.1	0.1395	0.1890	0.1450	0.1755	0.1885	0.1955	0.2050
	0.05	0.1395	0.1160	0.0935	0.0820	0.1040	0.0960	0.1180
	0.01	0.0230	0.0180	0.0205	0.0220	0.0260	0.0285	0.0295

Table 2. - Proportion of times where the null hypothesis is rejected

Recall that in this last case, a binomial approach will not be rejected the null hypothesis because the probability of adequate point under it is similar to the true population distribution.

III APPLICATION EXAMPLES

We illustrate this procedure working with two different examples. In the first example $\mathbb{H}_0$ is true, and we are going to see what happens if a sample confirms this hypothesis or not and in the second example the true model is worse, in the sense of generating a higher number of errors than the desired model.

We work in a three-dimensional space, and we suppose that a spatial data product establishes that errors in X, Y and Z are distributed according to three Normal and independent distributions with μ=0 m and σ=1m. This is the null hypothesis for both cases. Under this hypothesis, the quadratic error in each element (e.g. point, line or whatever kind) is:

$$QE_i = x_i^2 + y_i^2 + z_i^2$$

which is distributed according to a chi-square distribution with 3 degrees of freedom. In this case, the probability that an element has a $E \leq 2$ m ($QE_i < 4\,m^2$) is approximately 0.74, and the probability that an element has a $E \leq 2.5$ m ($QE_i < 6.25\,m^2$) is 0.90. In consequence, if we take a sample of control elements of size N, calculate the errors and count the number of elements whose $QE \leq 4$ m2; the number of elements with $4\,m^2 \leq QE \leq 6.25\,m^2$ and the number of elements whose $QE > 6.25\,m^2$, these quantities will follow a multinomial distribution with parameters (N, 0.74, 0.16, 0.10).

Let suppose a sampling of size 30 obtained for such population.

Element	e_X [m]	e_Y [m]	e_Z [m]	QE [m^2]	Category
1	0.378	-1.044	-0.104	1.244	I
2	-0.607	-0.528	1.127	1.917	I
3	0.069	0.003	-0.711	0.511	I
4	-1.211	1.644	-1.073	5.322	II
5	-0.600	-0.744	-0.887	1.701	I
6	0.238	0.806	-0.278	0.784	I
7	-0.431	-0.187	-0.397	0.378	I
8	0.798	-0.059	2.211	5.529	II
9	1.518	0.004	0.266	2.376	I
10	1.302	1.442	0.217	3.822	I
11	-0.880	0.662	-0.063	1.218	I
12	-1.594	-0.549	-1.194	4.269	II
13	-1.274	-0.699	1.272	3.732	I
14	-1.633	0.855	-0.914	4.234	II
15	-0.884	-0.592	1.395	3.077	I
16	2.109	0.210	-0.667	4.938	II
17	0.497	-0.203	-0.498	0.536	I
18	0.296	-0.719	-0.441	0.800	I
19	1.429	1.264	0.098	3.649	I
20	-0.594	0.704	1.139	2.145	I
21	1.286	-1.149	-0.082	2.982	I
22	-0.717	-1.082	0.630	2.083	I
23	1.088	-0.793	-1.565	4.261	II
24	-0.546	-1.149	-0.561	1.935	I
25	-0.543	-1.332	0.220	2.117	I
26	0.533	-1.214	1.457	3.881	I
27	2.795	0.639	0.653	8.648	III
28	-0.196	1.201	1.514	3.775	I
29	0.970	-0.605	-0.190	1.343	I
30	-1.016	0.220	1.261	2.671	I

Table 3 Example with a hypothesized case (Null population =True population)

In this case, the sampling statistics is $S = (23, 6, 1)$, To obtain the p-value, we calculate the probability in a multinomial (30; 0.74; 0.16;0.10) of the point S and all points where the first value is less than 23 or if the first point is 23 the second point is lesser than 6. Finally, all these probabilities are summed and the p-value is 0.5362. In consequence, we do not reject the null hypothesis.

For the second example, we want to test if the same null hypothesis as before is true, but in a second population. In this population, errors are truly distributed according to three normal standard distributions, as before, but now, errors in X and Y are highly correlated (correlation coefficient equals to 0.6). We obtain a sample of size 30 for this new population, and we obtain the sampling that appears in Table 4.

For this sample we classify each point into its corresponding class according to its QE value, and the test estimator value is $S = (18, 9, 3)$. As before, to obtain the p-value, we calculate the probability in a multinomial (30; 0.74; 0.16;0.10) of the point S and all points where the first value is less than 18 or if the first point is 18 the second point is lesser than 9. Finally, all these probabilities are summed and the p-value is 0.02950. In consequence, we can reject the null hypothesis, to an alpha level of 5%.

Element	e_X [m]	e_Y [m]	e_Z [m]	QE [m^2]	Category
1	-0.3869	0.6570	1.7462	3.6306	I
2	-0.7862	-0.5534	0.2419	0.9828	I
3	1.4546	-0.3234	0.8716	2.9802	I
4	-0.0815	0.7061	-0.2738	0.5803	I
5	0.6979	0.6140	0.7460	1.4207	I
6	-0.0322	0.0424	0.2619	0.0714	I
7	1.0974	1.7072	1.0563	5.2346	II
8	-1.6764	0.3839	-1.2768	4.5880	II
9	-1.3978	-1.6323	-1.0792	5.7830	II
10	-1.0880	-0.6952	0.2615	1.7353	I
11	0.8909	-0.9406	1.4338	3.7342	I
12	0.6538	1.3700	-1.1415	3.6072	I
13	1.3760	1.1411	-1.2979	4.8800	II
14	-1.2167	-1.5128	0.0603	3.7727	I
15	1.5711	1.7410	-1.4285	7.5398	III
16	1.3679	1.5336	0.0297	4.2240	II
17	1.0336	-0.1400	0.5406	1.3802	I
18	1.5528	0.6278	-1.1301	4.0823	II
19	-1.9135	-2.0818	-0.4297	8.1803	III
20	1.4945	1.2784	-1.0525	4.9755	II
21	-1.6507	-1.2067	-0.2056	4.2232	II
22	-0.2347	-1.6093	1.2372	4.1755	II
23	1.1281	0.9067	1.0737	3.2474	I
24	1.5816	0.8454	-0.8331	3.9101	I
25	1.3497	0.6396	-0.9129	3.0641	I
26	1.3805	-0.0765	-0.2301	1.9645	I
27	1.6766	0.0924	2.7859	10.5810	III
28	0.9029	1.2889	1.0359	3.5496	I
29	-1.6151	-1.1481	0.1718	3.9560	I
30	-0.1513	0.4965	-0.5257	0.5458	I

Table 4 Example with a hypothesized case (Null population worse than True population)

IV CONCLUSSIONS

We propose a general positional accuracy control method, when we are interested in splitting points into three categories in respect with measurement errors. This method has the advantage that may be employed without any previous hypothesis about the undelaying distribution of errors, and it can be applied to 1D, 2D or 3D error data.

The method is an exact statistical hypothesis testing based on multinomial distribution with three parameters, N, π_1, π_2. In the case of error, we proceed to discretize data, through the classification into three categories given for two tolerances previously defined by the user. So, probabilities of the multinomial distribution are related to the probability of each category under the null hypothesis. The proposed statistical test is exact, so the p-value can be derived by exploring a space of solutions and summing up the probabilities of each isolated case of this space.

To probe the viability of this contrast a simulation procedure has been carried out and two examples are presented.

This contrast has some advantages, such as its easy realization and implementation, for instance in R program. Albeit we have applied it in a continuous underlying situation (measurement errors), it is also useful for discrete and categorical situations.

ACKNOWLEDGMENTS

Research in this paper has been partially funded by grant CTM2015-68276-R of the Spanish Ministry on Science and Innovation.

REFERENCES

Ariza-López, F.J., Rodríguez-Avi, J. (2014). A Statistical Model Inspired by the National Map Accuracy Standard. *Photogrammetric Engineering & Remote Sensing* 80(3), 271–281.

ASCE (1983). *Map Uses, Scales and Accuracies for Engineering and Associated Purposes*. American Society of Civil Engineers, Committee on Cartographic Surveying, Surveying and Mapping Division, New York.

ASPRS (1990). ASPRS accuracy standards for large scale maps. *Photogrammetric Engineering and Remote Sensing*, 56(7):1068-1070.

ASPRS (2004). *ASPRS Guidelines: Vertical Accuracy Reporting for Lidar Data*. American Society for Photogrammetry and Remote Sensing. Bethesda, Maryland.

ASPRS (2015). ASPRS Positional Accuracy standards for digital Geospatial Data, *Photogrammetric Engineering and Remote Sensing*, 81(3):53-75.

Barnett, V., Lewis, T. (1994). *Outliers in Statistical Data* (3rd edition). John Wiley & Sons, Ltd, Chichester.

Bonin O., Rousseaux, F. (2005). Digital terrain model computation from contour lines: How to derive quality information from artifact analysis. *GeoInformatica*, 9:253-268.

Cayo M.R., Talbot, T.o. (2003). Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics*, 2:10.

FGDC (1998). *FGDC-STD-007: Geospatial Positioning Accuracy Standards, Part 3. National Standard for Spatial Data Accuracy*. Federal Geographic Data Committee, Reston, USA.

Fisher, R.A. (1922). On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 85(1): 87–94.

Freeman G.H., Halton J.H. (1951). Note on an exact test treatment of contingency, goodness of fit and other problems of significance. *Biometrika* 38:141-149.

ISO (2010). *ISO 16269-4:2010. Statistical interpretation of data -- Part 4: Detection and treatment of outliers*. International Organization for Standardization, Geneve.

ISO (2013b). *ISO 3951-1:2013. Sampling procedures for inspection by variables -- Part 1: Specification for single sampling plans indexed by acceptance quality limit (AQL) for lot-by-lot inspection for a single quality characteristic and a single AQL*. International Organization for Standardization, Geneve.

Logsdon T. (1995). *Understanding the Navstar: GPS, GIS, IVHS*. Springer US.

Maune D.F. (ed) (2007). *Digital Elevation Model Technologies and Applications: The Dem Users Manual*. American Society for Photogrammetry and Remote Sensing. Bethesda, Maryland.

Müller M.J. (2001). Exact Tests for Small Sample 3x3 Contingency Tables With Embedded Fourfold Tables: Rationale and Application. *German Journal of Psychiatry*, 4(1):57-62.

Oksanen J., Sarjakoski, T. (2006). Uncovering the statistical and spatial characteristics of fine toposcale DEM error. *International Journal of Geographic Information Science*, 20:345-369.

STANAG (2002). *Standardization Agreement 2215: Evaluation of land maps, aeronautical charts and digital topographic data*. North Atlantic Treaty Organization, Bruxelles.

USBB (1947). *United States National Map Accuracy Standards*. Bureau of the Budget, Washington DC.

 [ISBN: 978-2-9105-4510-5]

Practical global elevation data error simulation

Ashton Shortridge, Joe Messina, Xue Li, and Nick Ronnei

Department of Geography, Michigan State University, USA

*Corresponding author: ashton@msu.edu

Digital elevation models (DEMs) contain error, and that this error results in uncertainty in scientific and applied applications that use DEMs. Although the fitness-for-use paradigm has been widely recognized in the geospatial data production community for over two decades, today there are no publicly-available datasets that actually meet this criterion. We report on two research threads to efficiently generate well-parameterized error models for medium-resolution global DEM products: the first thread concerns the development of error models to account for relationships between SRTM error and local environmental characteristics, and the second concerns the efficient generation of spatially autocorrelated error realizations.

I GLOBAL TERRAIN DATA AND ERROR

It is widely recognized that DEMs contain error. This is a particular problem for terrain data products of global extent, as it is difficult to conduct rigorous "ground-truth" exercises to assess data quality in remote portions of the Earth. The impact of spatial data error on applications is typically not reported, and the burden of dealing with the potential impact of data error falls squarely on users of spatial data (Fisher & Tate 2006). Although the fitness-for-use paradigm has been widely recognized in the geospatial data production community for more than 15 years, today there are no publicly-available datasets that actually meet this criterion. We think this problem has three main components: the difficulty in translating the many DEM accuracy assessments into error models, the computational complexity of implementing Big Data stochastic simulation, and the difficulty in using realizations in scholarly and applied research with terrain data.

This paper reports on our team's progress towards addressing the first two of these challenges. We describe a global SRTM error model that captures important aspects of error using secondary, globally available data, and we introduce an efficient simulation implementation based on process convolution, enabling the rapid generation of very large error realizations. The paper concludes with comments on the method's limitations and potential extensions.

II ELEVATION ERROR MODELING

In a continental-scale accuracy assessment, Shortridge & Messina (2011) linked SRTM error with globally available ecoregion, forest cover percentage, and topographic data. While a regression model with these variables (Table 6 in Shortridge & Messina 2011) accounted for substantial variance in SRTM error, the residuals were spatially autocorrelated, as Figure 2d demonstrates for a sample of locations in the United States. We extend this approach by proposing a basic regression kriging model (Hengl et al. 2007) to characterize SRTM or GDEM error as a linear function (parameters estimated via generalized least squares) of these variables along with a separate spatially autocorrelated error term, (calculated via simple kriging):

$$\hat{y}(\boldsymbol{s_0}) = \hat{m}(\boldsymbol{s_0}) + \hat{e}(\boldsymbol{s_0}) \quad (1)$$

Where y-hat is the DEM error estimate at location $\mathbf{s_0}$, m-hat is a linear model accounting for covariates measured at $\mathbf{s_0}$, and e-hat is an estimate of the error in the linear model at $\mathbf{s_0}$, with the estimate derived from simple kriging. This model can be developed in data-rich areas where reference elevations are available at locations $\mathbf{s_0}$ and then efficiently deployed in areas without reference elevation data to generate realizations. The e-hat component is estimated via unconditional simulation, and individual elevation estimates are found as:

$$\hat{z}(\mathbf{s_0}) = DEM(\mathbf{s_0}) - \hat{y}(\mathbf{s_0}), (\mathbf{s_0}) \in DEM \quad (2)$$

where z-hat is an estimate of the unknown actual elevation at $\mathbf{s_0}$ and DEM is the SRTM or GDEM elevation at $\mathbf{s_0}$.

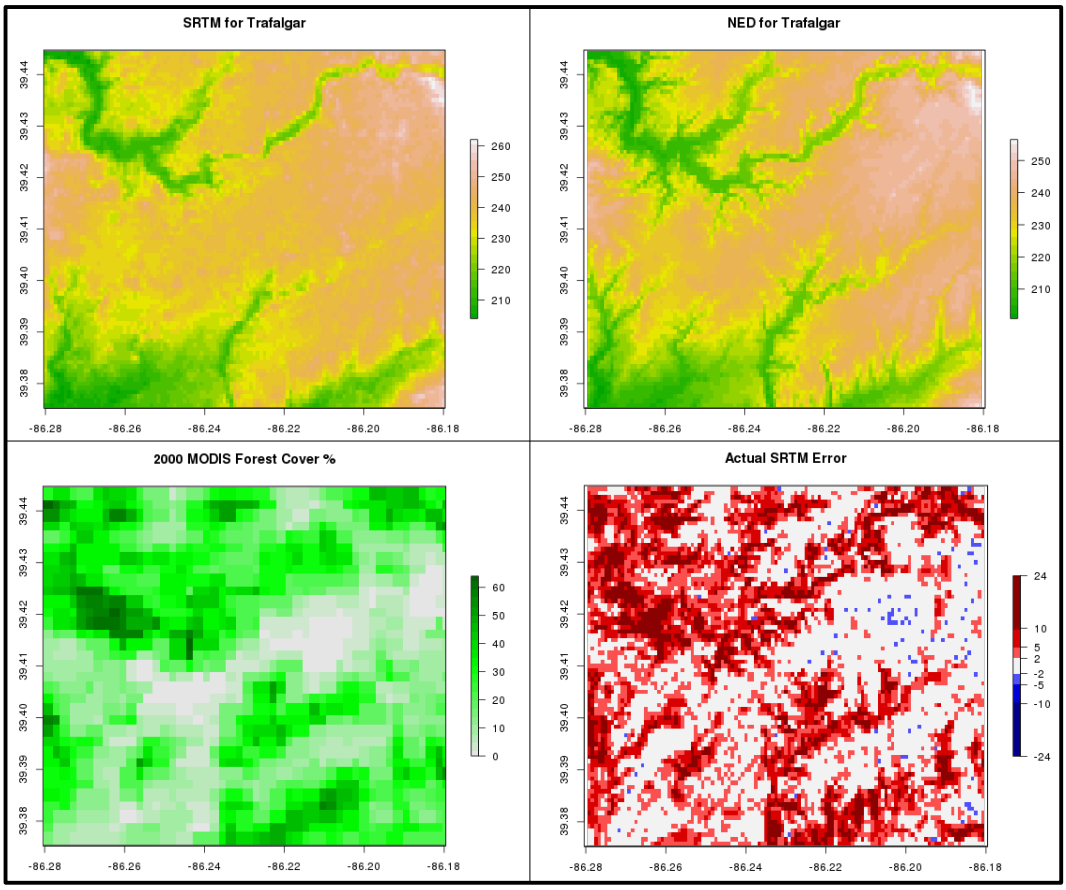


Figure 1: a) 3” SRTM DEM near Trafalgar, Indiana, USA; b) Resampled 3” NED DEM; c) 2000 MODIS forest cover %; d) SRTM errors.

We illustrate this model with a case study on rolling terrain in south-central Indiana, USA near the town of Trafalgar. The region extent is 8,670 m x 7,700 m, comprising 10,043 pixels of 3” SRTM data (Figure 1a). For reference data we use USGS National Elevation Dataset (NED) data, resampled to 3” to match SRTM (Figure 1b); mismatches are apparent between these elevation surfaces. MODIS forest cover percentage for the year 2000 (Figure 1c) is clearly linked to larger positive SRTM errors (Figure 1d). Implementing the continental error model described above, parameters are multiplied by local SRTM slope, aspect, and MODIS forest cover percentage values to produce a predicted error surface (Figure 2a), which is subtracted from the original SRTM to produce a modeled DEM (Figure 2b). This model is not perfect; residuals from (SRTM - Predicted Error) - GDEM are presented in Figure 2c. RMSE of the original SRTM is 5.66 m, while RMSE for the model is 4.43 m.

Further, errors are spatially autocorrelated. Using a USA-wide sample of prediction errors, an average variogram model was formulated; this nationally-derived model is plotted against the empirical variogram of errors for the study site on Figure 2d.

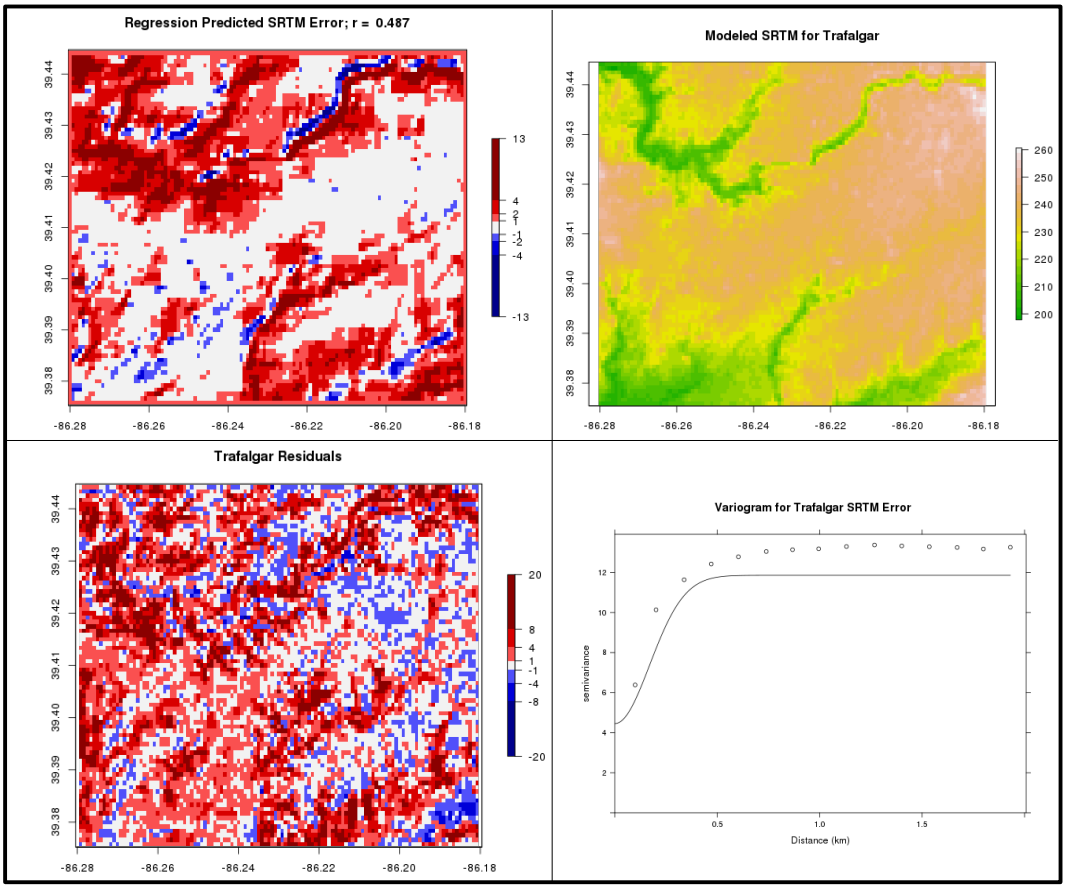


Figure 2: a) Predicted SRTM error using Shortridge & Messina (2011) regression model; b) Model DEM (SRTM – Predicted Error); c) Residuals (Model - GDEM; d) Empirical variogram (dots) for the study area; model (line) derived from USA-wide sample of locations.

III EFFICIENT ERROR REALIZATION PRODUCTION

A significant challenge in global error propagation modeling via stochastic simulation is the computational complexity of implemented geostatistical simulation models. Generating dozens of error realizations for large (greater than one million cell) DEMs presents practical challenges for many algorithms. We implement an approach of Ver Hoef & Barry (1998), further developed by Higdon (2002) to construct realizations of valid spatial covariance models by passing specially constructed moving averages over gridded white noise. Convolution filters are commonly used by geospatial analysts for feature detection and classification in image processing, but their utility in error modeling is less well-developed (Oksanen & Sarjakoski 2005 is the notable exception). As explained in Higdon, particular covariance models (e.g., Gaussian with a specified range) may be convolved into specific smoothing filter. The result is an unconditional realization with spatial stucture corresponding to a valid covariance model.

We have implemented this approach in R. Figure 3a demonstrates that realizations comprising millions of cells can be generated rapidly with this algorithm, making global deployment possible. The filter realization methodology can readily be adapted for specific areas. Figure 3b shows off a single error realization for the Trafalgar study site using a Gaussian covariance

model with parameters defined from a USA-wide sample: nugget 4.45, partial sill 7.41, and effective range 416 meters. One hundred similar realizations from this model for this study area were simulated in less than five seconds on the first author's older desktop PC.

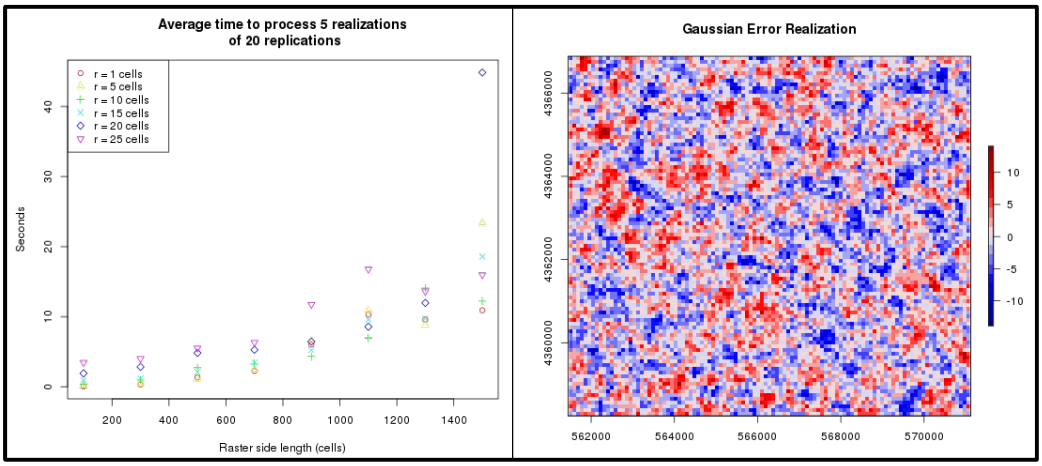


Figure 3: a) Simulation times for spatially-autocorrelated fields using process convolution; b) Example error realization for the Trafalgar study site (UTM zone 16 projection).

IV DISCUSSION AND CONCLUSIONS

This brief paper demonstrates how DEM accuracy research may be developed into DEM uncertainty model research. Regression kriging offers an appealing framework for capturing important error properties of global DEM products and efficiently reproducing observed pattern in new areas. Further, the use of convolution filters to produce the spatially-structured random component is very promising. There are limitations to both approaches: RK can not work on non-colocated data, and it uses a very simple measure of spatial dependence. Process convolution models do not appear capable of conditioning realizations to locally available reference data, and several papers have identified approaches that capture pattern with greater fidelity (e.g., Bolin & Lindgren 2013). A significant limitation of the current implementation is that a single covariance model is employed. However, these models are computationally cheap and implementable on a large scale, and clear potential exists for nonstationary variogram modeling. Ultimately, this is part of a larger effort to enable efficient deployment of DEM uncertainty propagation; we expect to have a pilot system running very soon.

References

Bolin, D. Lindgren, F. (2013) A comparison between Markov approximations and other methods for large spatial data sets. *Computational Statistics & Data Analysis*, 61: 7-21.

Fisher, P. F., Tate, N. J. (2006) Causes and consequences of error in digital elevation models. *Progress in Physical Geography*, 30(4): 467-489.

Hengl, T., Heuvelink, G. B., Rossiter, D. G. (2007) About regression-kriging: from equations to case studies. *Computers & Geosciences*, 33(10): 1301-1315.

Higdon, D. M. (2002) Space and space-time modeling using process convolutions. In Anderson, C. W, Barnett, V, Chatwin, P. C and El-Shaarawi, A. H., (Eds) *Quantitative Methods for Current Environmental Issues*. London: Springer, 37–54.

Oksanen, J. Sarjakowski, T. (2005) Error propagation of DEM-based surface derivatives. *Computers & Geosciences*, 31(8): 1015-1027.

Shortridge A., Messina J. (2011) Spatial structure and landscape associations of SRTM error. *Remote Sensing of Environment* 115(6): 1576-1587.

Ver Hoef, J. M. and Barry, R. P. (1998) Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning & Inference*, 69: 275:294.

How far is far enough? Towards an adaptive and "site-centric" modelling integrating co-visibility constraints for optimal land use

Valerio Signorelli*[1], Thomas Leduc*[1], Guillaume Chauvat[2]

[1] UMR AAU – CRENAU, Ecole Nationale Supérieure d'Architecture de Nantes, France
[2] Cittànova, Nantes, France

*Corresponding author: valerio.signorelli@crenau.archi.fr

In this article, we propose a renewed site-centric solution that allows us to characterize a specific region of interest by defining the extent of the surroundings that influence sunlight exposure. The proposed method is a mix of an adaptive refinement and a visual-based clipping technique. This method has been implemented in the SketchUp context and applied to three sites located next to the French historical thermal town of Aix-les-Bains.

I INTRODUCTION

Urban planning policies aim at defining the conditions of human settlements. Coherence and convergence of public action rely on the spatial continuity of its field of application for common issues. That is why French planning regulations tend to follow a concerted process led on several municipalities. They operate through planning regulations which are the expression of a political answer to issues emerging from territorial analyses. These analyses reveal the complexity of the territory by describing the local implications of the relations between cultural, physical and anthropological phenomena. Terrain features, settlement patterns, vegetation and infrastructures do not only influence environmental parameters, such as the amount of solar energy potential and daylight availability, in existing and planned urban fabric, but they also define the way in which inhabitants perceive their environment.

For this reason, urban planning practices should consider at the same time, and alongside urban regulations, perceptual, environmental and climate features in available and potential construction sites. Within this context solar exposure takes on great importance in terms of energy efficiency, quality of public and private spaces, and physical-perceptual enhancement of the local eco-system.

The availability of high resolution terrain and buildings models, the improvement in computation capability, and the development of 2.5D modelling simulation tools, based on image processing computation have provided, in the last decades, essential means for obtaining effective solar analysis from territorial to city scale (Prévot *et al.*, 2011; Morello *et al.*, 2010; Morello & Ratti, 2009; Floriani & Magillo, 2003; Tandy, 1967). However, simulations conducted on large and detailed raster grid models, the format commonly used by public agencies to deliver these basic data, are still costly operations, due to the amount of data needed, and tools based on image processing approaches do not permit yet an effective integration with vector-based solutions used in urban planning practices.

In large and complex topographic areas the influences of the various territorial features, even far from the actual position of the chosen sites, should be carefully considered. That is to say, by considering just the immediate surrounding of a site we can over- or under- estimate the solar contribution in terms of energy and sunlight. But which are the features to consider? Which is the level of detail they should provide? And how can we ensure an accurate simulation model able to provide reliable results in a reasonable amount of time? In other words, *how far is far enough*?

The aim of this explorative research is to propose a site-centric simplification method, based on 3D visibility analysis, in order to obtain a vector-based terrain model easier to handle, faster to compute. The entire process is integrated, as a series of extensions, for a well-known CAAD system, Trimble SketchUp. Three case studies, illustrating different levels of topographic constraints, will be used. Potentialities and limitations of the method will be highlighted in the discussion and further investigation will be proposed.

II METHOD

Pre-processing

The spatial datasets are provided by the IGN[1], a national French institute in charge of the management and updating of geodesic and leveling networks, aerial photographs, and geospatial data. More precisely the aforementioned datasets are part of the French Large Scale Reference system (RGE): Digital Elevation Model or DEM (RGE® ALTI 5 m) in raster format for the representation of the landform (supposedly free of vegetation, buildings, etc.), and 3D vector models of significant spatial features such as footprints of individual buildings, forest cover, etc. (BD TOPO® 3D).

The needed pre-processing operations has been made using the Geospatial Data Abstraction Library GDAL/OGR (GDAL, 2016). Specifically the merging of the geospatial data, obtained through "*gdal_merge.py*" command line tool in the context of raster-based tiles or "*ogr2ogr -update -append*" command line tool in the context of vector-based layers; the clipping of data sources to some specified bounding box by using "*gdalwarp*" in the context of raster-based tiles or "*ogr2ogr -clipsrc*" in the context of vector-based layers.

To generate 3D vector contour files from the input raster DEM, the "*gdal_contour.py*" command line has been used and the resulting polylines have been simplified using the "*ogr2ogr -simplify*" tool. The simplified contour polylines are then reused to build the various Terrain Models presented hereafter.

A two-step process: adaptive refinement and visual-based clipping techniques

The objective, after the conversion of the raster-grid model in a vector-based model, is to refine the virtual model of the terrain with an acceptable trade-off between the amount of data, and thus computation time, and data accuracy. Several terrain configurations have been produced and compared. However, only three of them will be detailed in this presentation.

[1] The *Institut national de l'information géographique et forestière* (National Institute of Geographic and Forestry Information, IGN), is a French public state establishment to produce and maintain geographical information for France.

The “reference” Terrain model (M1) has the same planimetric resolution (5 meters) for the whole region, as provided by the original IGN dataset. In this first model there is no difference between studied parcels and the surrounding landscape, and the entire complexity of the terrain is represented. A side effect of this model is that even the areas that do not influence the chosen sites are precisely modelled. The “mixed” Terrain model (M2), combines the high resolution model of the three selected sites, with the low resolution model of the rest of the whole region. The third model is the “local” Terrain model (M3), where just the selected site and their immediate surroundings are considered. The mixed Terrain model (M2) is therefore some sort of intermediary between two extreme solutions. On the one hand, the "local" Terrain model (M3) does not take into account far distances’ masks. On the other hand, the "reference" Terrain model (M1) is unnecessarily precise all over the wide region.

The three models have been developed using the contour lines obtained during the preprocessing phase, and then imported as SHP file (through a tool developed by one of the authors) in SketchUp. Through the existing tool “*Sandbox From Contours*”, the contour lines have been converted into a Terrain Model (Triangulated Irregular Network or TIN). Lastly, the building footprints have been imported as SHP file, drape on the terrain surface and extruded using the elevation values given as attributes.

The adaptive refinement of the Terrain models consists in the spatial union of two different sets of contours lines layers, with distinct spatial resolutions. Connections between this two datasets are automatically handled by the SketchUp “*Sandbox from Contours*” tool. The three images presented in Fig. 1 show a) the boundaries of the immediate surroundings of the studied site (the red polygon was obtained as a 300 meter radius buffer) b) the spatial union of the two contours lines datasets, with different resolutions, and c) a 3D sketch of the result in SketchUp.

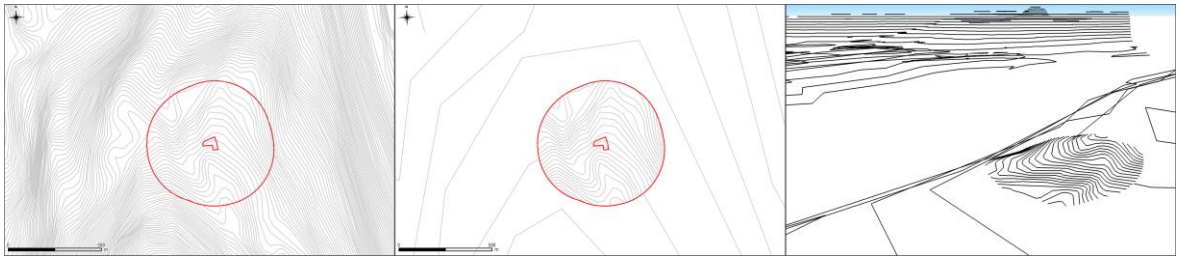


Figure 1: The refinement technique used to adapt the resolution of the Terrain model. The contour lines outside the red polygon of the studied site, have been replaced by lower resolution contour lines.

In order to further reduce the amount of data to be processed, a visibility-based clipping technique is used, considering that the hidden parts of landscape, from a given position, will not influence the studied area in terms of solar exposure.

A 3-step method has been implemented in SketchUp. First of all, we placed, over the three selected areas, a point grid with a fixed step equal to the resolution of the terrain obtained by IGN. For each of the sampling nodes, a 3D viewshed is computed using our extension based on the native ray casting engine of SketchUp. Finally, the spatial union of all these viewsheds is assessed and, to avoid any interpolation effect in all concavities during the TIN building phase, the convex hull of the resulting spatial union is delineated (see Fig. 2). The convex hull is then reused to clip the coarse-grained contour lines, and therefore (potentially) divide by two the area of the region to be taken into account.

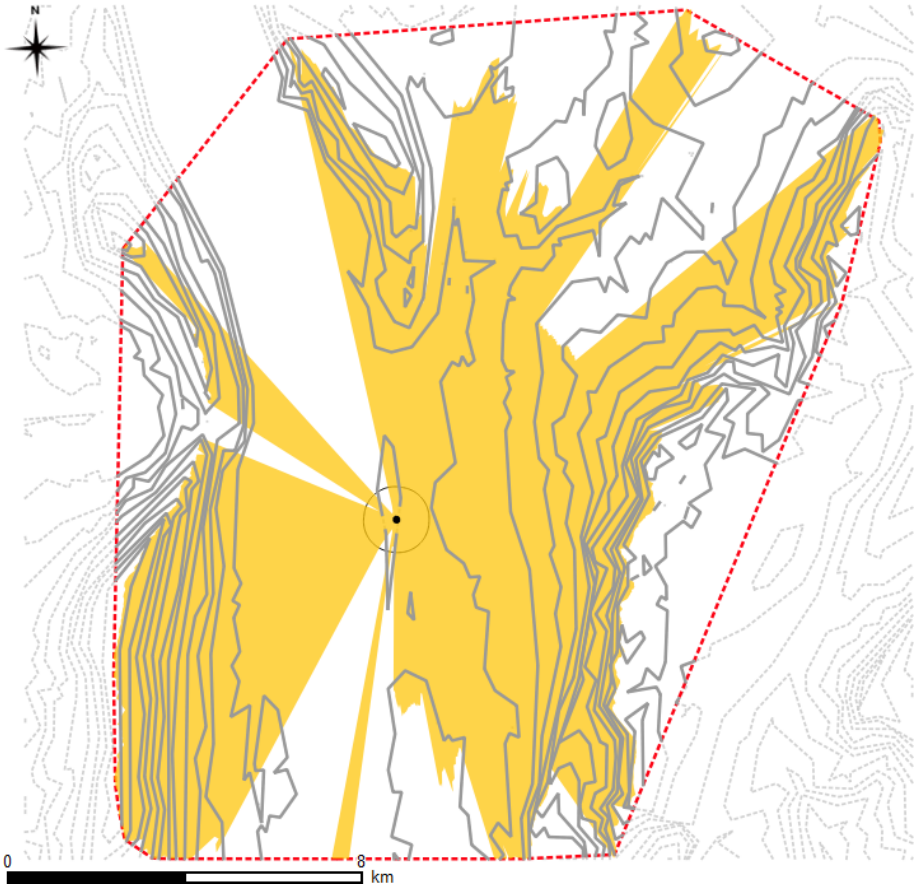


Figure 2: The visual-based clipping technique shows the amount of contour lines, and therefore of terrain, that will not be taken into account in the simulation phase.

Post-processing

The three terrain models developed (M1, M2, and M3) have been used in the solar simulation and the obtained outcomes compared. Two indicators have been considered in order to test the reliability of our simplification method: the beam (direct) solar irradiation values (Wh/m²) and the daylight duration (min.). Both indicators consider a standard clear sky model and do not take into account sky or model reflections. We decided to conduct the simulation on December, 21st as the day with the lowest sun angles of the year, thus lowest amount of irradiation and daylight.

The measure of irradiation per unit area, depends obviously on the Terrain model itself. More precisely, it depends on its own direction towards the various sun positions, insofar as irradiation understood as the sum of instantaneous density of solar radiation incident on the surface over the given time period is the scalar product of the normal to the face by the sunlight direction).

III USE CASE

The region of interest, of about 386.5km² (a 20 km-width square), is located in the French Alps, on the shores of the wide *Lac du Bourget*. In its south part, it embeds a 8-km wide valley oriented north-south which spreads between two mountain ranges (*Le Mont du Chat* and *Les Bauges*) peaking at around 1.5 km over the sea level. In the northern part, on the contrary, the valley broadens and the surrounding ridges sink down (see Fig. 3).

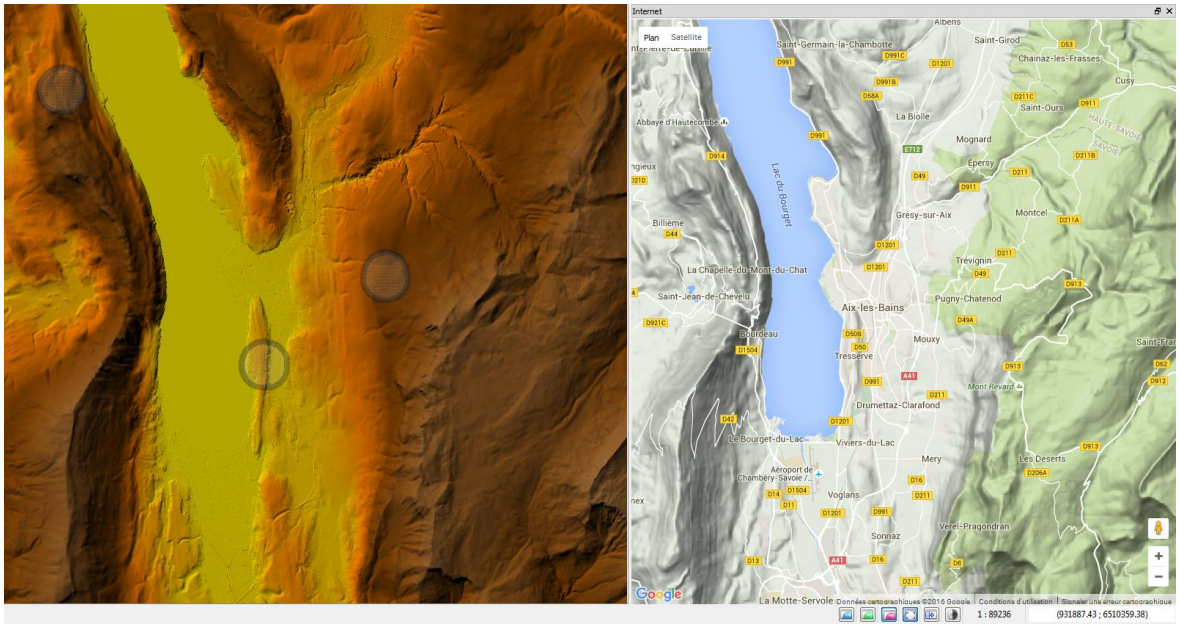


Figure 3: The region of interest with the 3 studied sites (*Ontex* on the NW side, *Pugny-Chatenod* on the East side, and *Tresserve* close to the *Lac du Bourget*).

This region covers a specific inter-communality (the *Communauté d'Agglomération du Lac du Bourget: Grand Lac*) consisting of 17 municipalities. Planning regulations mostly tend to concentrate urban development around the administrative center of municipalities. This aims to allow planning optimization, spaces preservation for natural and agricultural purposes. On this constrained territory, land pressure requires a global project which can be embodied by many strategic locations. We propose to evaluate the sun exposure on those strategic parcels, in the existing fabric or on its boundaries.

In order to select relevant parcels, we assessed the expected impact of terrain on municipalities' center by evaluating the highest aspect ratio (H/W) to the closest relevant ridges (see Table 1).

Municipality	**Aspect ratio**	**Municipality**	**Aspect ratio**
Tresserve	0.865	Bourdeau	0.274
Grésy-sur-Aix	0.579	Viviers-du-Lac	0.252
Méry	0.445	Aix-les-Bains	0.228
Drumettaz-Clarafond	0.433	Voglans	0.220
Le Montcel	0.430	Saint-Offenge	0.208
Mouxy	0.316	Brison-Saint-Innocent	0.184
Pugny-Chatenod	0.305	Ontex	0.147
Trévignin	0.295	La Chapelle-du-Mont-du-Chat	0.088
Le Bourget-du-Lac	0.284		

Table 1. Characterization of the impact of closest ridges for each municipality center.

Instead of *La Chapelle-du-Mont-du-Chat*, whose administrative center is constrained by the topography, we chose parcels in *Ontex*, on the slope of *Le Mont du Chat* (see Table 2, Fig. 4). We also chose locations on *Tresserve*'s hill and in hillside *Pugny-Chatenod*. Those different locations all embody development opportunities inside the existing urban fabric.

	***Ontex* site**	**Pugny-Chatenod site**	**Γresserve site**
Reference TIN	M11	M12	M13
Mixed TIN	M21	M22	M23
Local TIN	M31	M32	M33

Table 2. Various nomenclatures (models names) in use.

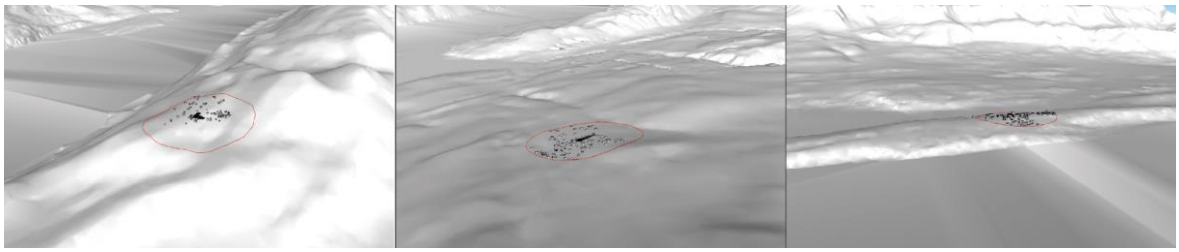


Figure 4: Zoom in the three sites (from left to right: *Ontex*, *Pugny-Chatenod*, and *Tresserve*). The close horizon limits are represented by red circles.

IV DISCUSSION

Predictably, values obtained from M1 are the lowest and indices' values provided by M3 are the greatest (see Fig. 5). Indeed, a fine modeling of the mountainous Terrain model add new masks to the mock-up and therefore decreases the solar potential of the Terrain patches.

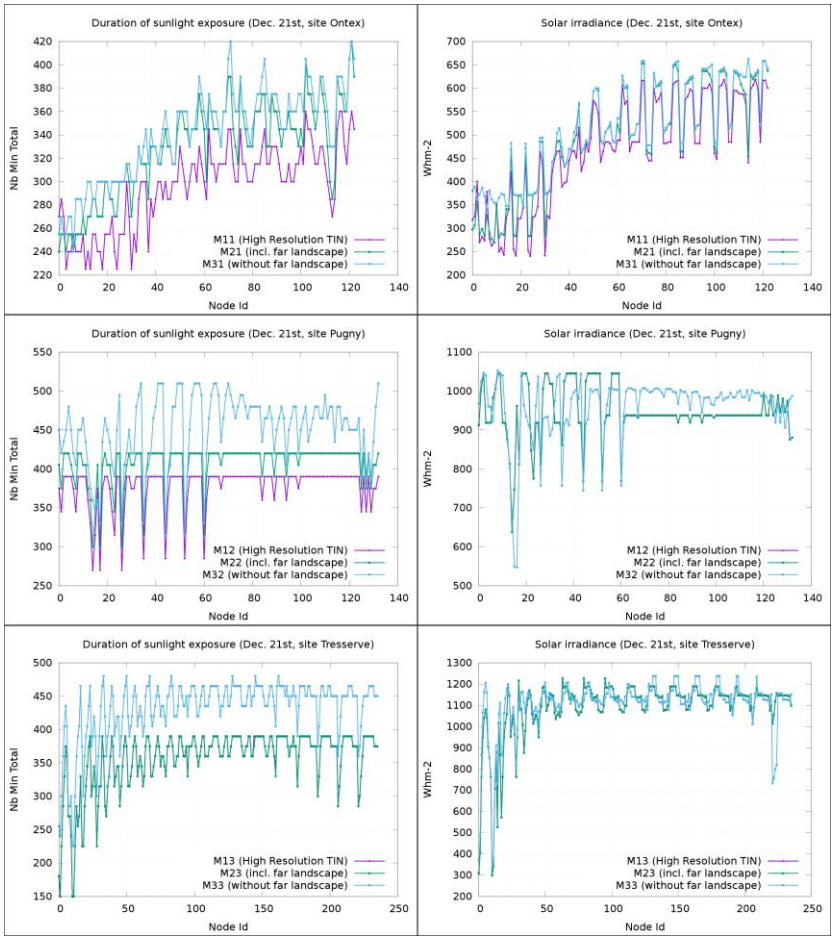


Figure 5: Comparison of the two indicators (duration of sunlight exposure and irradiance, Dec. 21st) for the three sites.

The respective average elevations of the three sites vary from 310m (*Tresserve*), to 604m (*Pugny-Chatenod*), and 711m (*Ontex*). The corresponding standard deviation in each site is included between 2 and 3.8m (with a maximum value in the site of *Tresserve*). Height range between lowest and highest points within a 300m buffer varies along the three sites. This range is 115m (from 645m to 760m) high at *Ontex*, 125m (from 555m to 680m) at *Pugny-Chatenod* and 75m (from 250m to 325m) at *Tresserve*. On that last site, the closest obstacles do not alter direct irradiance, since they are located in a northerly area. The *Pugny-Chatenod* site is located on a West-East slope, which alters early morning exposure (lower energy input). Eventually, the parcels chosen in *Ontex* are affected by southerly masks (see Fig. 6), that reduce mid-day exposure (higher energy input).

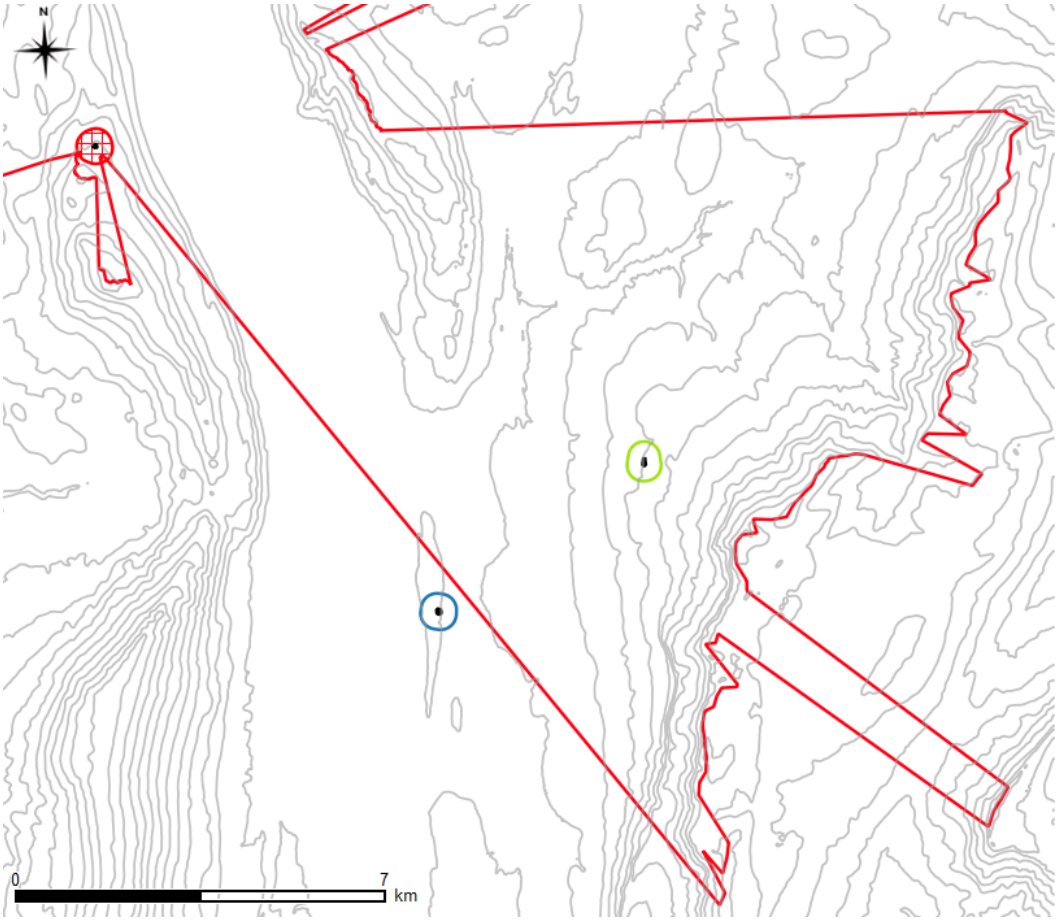


Figure 6: The southern edges of the "cumulative viewshed" of the *Ontex* site (NW) are close from measurement locations (from 180m to 670m with a peak of about 1150m high located just south at about 2.5 km).

As one can notice, for the site located on the top of a small hill close to the lake (*Tresserve*), the coarse-grained model M2 provides simulations results that perfectly match those obtained using the reference Terrain model M1. When the Terrain model gets hillier in the immediate surroundings, simulation outcomes are not as clear-cut. Thus, in the specific case of *Pugny-Chatenod*, where the site is located on a west-facing slope, the coarse-grained model M2 is accurate enough for the irradiation indicator (because it impacts mainly the early morning exposure with lower energy input). However, the mixed TIN M2 is obviously not precise enough to assess the precise daylight duration (there is indeed an average difference of about 30-minutes with the reference values). At last, in the particular case of *Ontex* site, where the Terrain model is particularly hilly nearby the measurement locations, the mixed TIN M2 is undoubtedly inaccurate.

V CONCLUSION

In our explorative analysis we prove the potentialities and limits of a vector-based approach for analyzing large and complex terrains. Further investigation will be also conducted in order to improve the usability of the SketchUp extension developed in this research. Thus, every territory covered by a DEM could be analyzed, by dynamically changing the terrain site.

Many parameters can also affect the modelling of a site. Closeness has to be defined according to the site location. We proposed a 300m buffer but this parameter could be subject of a sensibility analysis. This analysis suggests that there should be an optimal obstacle region radius for sunlight access assessment, depending on each site. Generally speaking, adaptive modelling should be the aim of all site analysis. Also, the accuracy of the skyline silhouette also alters the solar exposure. It relies on the distance between contour lines. A sensibility analysis should also be led on that parameter to determine how the level of detail of the terrain model affects the calculation of solar radiation indicators.

We proposed that the solar exposure could be defined efficiently thanks to a site-centric modelling. The relevant obstacle region around the observing point depends on the urban fabric. This approach could be extended to other case studies, for a wide set of density range: from the open field to the high medieval urban fabric.

References

Floriani, L. De, Magillo, P. (2003). Algorithms for visibility computation on terrains: a survey. *Environment and Planning B: Planning and Design*, 30(5), 709–728. http://doi.org/10.1068/b12979

GDAL (2016). GDAL - Geospatial Data Abstraction Library, Version 1.10.1. GDAL Development Team, Open Source Geospatial Foundation. Retrieved from http://www.gdal.org

Morello, E., Ratti, C. (2009). Sunscapes: "Solar envelopes" and the analysis of urban DEMs. *Computers, Environment and Urban Systems*, 33(1), 26–34. http://doi.org/10.1016/j.compenvurbsys.2008.09.005

Morello E., Carneiro C., Desthieux G., (2010). The use of digital 3-D information to assess urban environmental quality indicators. In Schmitt G. et al. (eds.), *Future Cities: Proceedings of the 28th International Conference on Education in Computer Aided Architectural Design in Europe*, Zurich, Switzerland, pp. 388-395.

Prévot, A., Rodriguez, D., Molines, N., Beckers, B. (2011). La modélisation 3D : une nouvelle voie pour les documents d'urbanisme ? Application à l'optimisation énergétique des bâtiments. *Revue Internationale de Géomatique*, 21(4), 557–583. http://doi.org/10.3166/rig.15.557-583

Tandy, C. R. V. (1967). The isovist method of landscape survey. *Methods of Landscape Analysis*.

Combination of error characterization and spatial error model to improve quality of digital elevation models

Tomaž Podobnikar[*1]

[1] University of Ljubljana, Faculty of Civil and Geodetic Engineering, Slovenia

*Corresponding author: tomaz.podobnikar@fgg.uni-lj.si

Abstract

A processing chain for spatial data cleaning has been developed, in order to produce more reliable digital elevation model (DEM). The first step of this heuristic process is error characterization in a combination of (geo)statistical, empirical and visual exploratory analyses, which were experimentally realized with clustering analysis. The second step is developing a spatial error model that uses homogeneous regions derived from terrain parameters, such as roughness and land cover, in relation to reference information. To support the methods of error characterization and spatial error model in spatial datasets, a novel 'locally systematic' error type has been proposed. This more complex error type characterizes (spatial) positional error field, which is supplementary to traditional random, systematic and gross error types. It is supposed to be the most typical with its geomorphological and topographic components. It comprehends geomorphological shapes, patterns, structures, i.e. semantics of the terrain surface that is usually oversimply considered as a random error. The results show that a comprehensive understanding of the error properties through error characterization supports effectively processing of the individual variables and from them a reliable error surface, whereas both are useful for effective removal different types of error, with an exception of the random one.

Keywords

DEM, digital elevation model, DTM, digital terrain model, spatial data quality, statistical/empirical/visual quality assessment, uncertainty, error characterization

I INTRODUCTION

The quantity and availability of digital spatial datasets rapidly increase, especially widely applicable earth observation (EO) data provided by a series of satellites. The digital elevation model (DEM), or often called as digital terrain model (DTM) was one of the first forms of digital geographical information which became available (Fisher and Tate 2006) and it is now a fundamental dataset to mapping our world. The main reason of recently improved DEM usability are new methods for data acquisition that have been introduced, especially from passive and active sensors on satellites, airborne, and unnamed aerial systems (UASs). These data are today considerably higher (spatial) resolution than before and therefore more comparable to the other EO. There is also a growing number of public accessible application platforms that additionally increase the usability of the DEM. However, the spatial characteristic of the DEMs quality is not always in accordance with their resolution and user's expectations. Additionally, there is still a gap between what the quality assessment experts can produce and the information that users can understand and use (Devillers et al. 2010).

The aim was to develop a solution that combines a number of different aspects of the DEM quality through error/uncertainty characterization and a spatial error model as an error surface or field, in order to get a better quality DEM. The proposed solution basically contributes in the DEM production chains with a spatial data cleaning procedure (detection, validation, correction; Kimball and Caserta 2004) as part of quality assurance. For both, the already produced DEM, or for implementation in the future DEM processing chain, a better quality DEM can be produced. An important background of the proposed solution is a conceptualization of the DEM taking into the account fitness for use – not only for multi-purpose demands where the main aim is to process a near-universal DEM – but also for specific demands to produce different DEMs for various applications. E.g., for the near-universal DEM we can define that the model excludes vegetation, snow cover, different buildings and bridges, but include glaciers. It is also good to know what kinds of errors (Leon et al. 2014) and their magnitude are people willing to accept.

II COMBINED APPORACH

The proposed solution is based on a combination of statistical, empirical and visual (with eyes inspection) quality control and assurance, i.e. on quantitative and qualitative approaches. The most significant elements of the spatial data cleaning process is a combined approach for *error characterization* and a *spatial error model* building (Figure 1) in order to generate a better quality DEM. Order of data cleaning methods is important; therefore, a decision-making process is necessary to adopt final procedure among several alternative possibilities resulting from different methods.

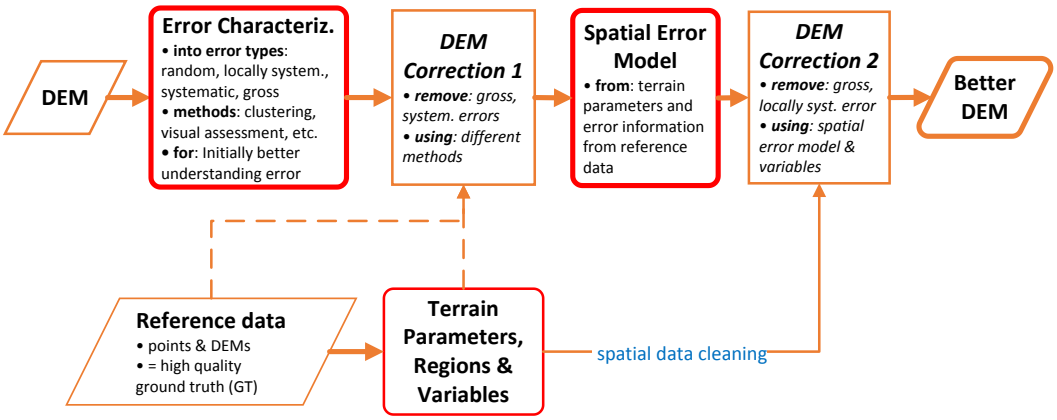


Figure 1: A processing chain of spatial data cleaning to produce an improved DEM from the already well processed DEM.

In order to effectively test the results of the proposed solution, we located the study areas in different parts of Slovenia, where DEM from contour lines, National DEM, lidar DEM, geodetic spot elevations database and other reference datasets, such as land cover and geological map were available; as well as the most typical landscapes were included.

Terrain parameters and regions

The expected DEM error/uncertainty/quality characteristics and magnitude are related to terrain parameters, i.e. terrain complexity/morphology, sampling density, which have been observed by numerous studies (e.g. Carlisle 2005, Oksanen and Sarjakoski 2006, Erdogan 2010). Podobnikar (2005) suggests correlating the error field with terrain skeleton (peaks, pits, saddles, ridges, talwegs), roughness, slopes, and rate of vegetation cover. Guth (2006) proposed geomorphometric variables that appear to be most important in terrain description and which

access SRTM quality: steepness of roughness, elevation-relief ratio or coefficient of dissection, organization, heterogeneity, profile curvature, plan curvature, average elevation, fabric direction, massiveness, modality, and limitation. The high-resolution lidar data and derived DEMs are expected to vary with land cover types, such as open terrain, high grass weeds and crops, brush lands and low trees, fully forested and urban areas (Zandbergen 2011). Many other papers classify the elevation error of lidar DEM similarly as open terrain, weeds and crops, scrub and bushes, forested, build-up areas (Hodgson and Bresnahan 2004). They classify the open terrain data as highest and fundamental accuracy, where the assessment of the other terrain types is supplementary. The ISPRS (by Sithole and Vosselman) provides fifteen benchmark lidar point clouds on the arbitrary defined representative environments, on which selected error types of the produced DEMs can be compared since 2003.

The DEM error is also determined by geology (rate of karst – karstness), hydrological network; anthropogenic impacts on terrain (e.g. stone quarries, buildings, transport network infrastructure); larger areas of plains, hills and mountains; precipitation and temperature (e.g. in karst areas with the same geology the terrain is considerably rougher in mountains than in lower areas). Nevertheless, the predicted error very much depends on the data sources used (e.g. photogrammetric, lidar), scale/resolution and related information entropy or fractal dimension (Wise 2012), as well as on the interpolation/filtering methods used and on particular producers' style of data processing.

Practically, the terrain parameters were derived from DEM, land cover data and geological maps. The terrain parameters were either continuous surfaces or discrete datasets. All these parameters were then separately classified to significant regions.

Error characterization (for the 1st correction)

Bearing in mind traditional error theory there are three types of errors: random, systematic and gross. Typical systematic (e.g. wrong vertical geodetic datum) and gross errors (e.g. wrong altitude of particular locations) can be detected and removed. In reality, most of the DEM errors are not random, not stationary (spatially variable) and not from and identical normal distribution (Zandbergen 2011). Statistical methods are appropriate for error assessment only if a substantial number of checkpoints is used, what is rarely feasible in real applications (Liu et al. 2012). Therefore, the validity of the most applicable root-mean-square error (RMSE) metrics is questionable, particularly because the main error in DEMs is so called systematic-like (Oksanen and Sarjakoski 2006) or locally systematic (Podobnikar 2008); which is not normally distributed (e.g. as tested with the Q-Q-plot). A robust spatial statistics (median, quantiles) or qualitative methods are needed in this case (Höhle and Potuckova 2011). Consequently, we propose four main groups that characterize a positional error in DEM:

- *random* error – (not removable – vertical and planimetric); differently autocorrelated and scale dependent (Fisher and Tate 2006)
- *locally systematic* error – (potentially partly removable and partly inherent – vertical only); it is: (1) very complex and most typical error in DEM, but usually oversimply considered as a random error or even uncertainty; (2) similar to the concept of low and medium frequency errors (Imhof 1982); (3) spatially autocorrelated (e.g. in relation to sampling density), in the context of the certain area (mostly constrained) geographic neighbourhood (e.g. in relation to underestimated altitude of peaks); (4) topographic and semantic with specific shapes and patterns (e.g. in relation to geomorphological forms and structures); in this context the DEM is considered as a surface and not as independent individual points

- *systematic* error – (mostly removable – vertical and planimetric)
- *gross* error – (mostly removable automatically or manually – vertical and planimetric); wrong attributes of altitudes are owing to not appropriate filtering of lidar point cloud data or interpolation; it is often a semantically recognizable error due to misinterpretation in manual digitalization of maps or in stereo-photogrammetry

There are numerous different measures and indices used to characterize DEM error into main proposed or other groups, such as to a group of (geo)statistical, empirical and visual; or to a group with an error assessment on a DEM only, or using reference datasets. Here are some basic examples of this kind of exploratory analysis: RMSE calculation, using shaded DEM or contour lines calculated from DEM, relating independent hydrological network and DEM, applying Monte Carlo simulations and sensitivity analyses, computing histogram or swath profile, etc. The proposed measures and indices are used for error identification and consequently for removal gross errors (e.g. with correction/transformation attributes, interpolation/filtering, filling gaps with other datasets) and systematic errors (e.g. with transformation or georeferencing).

To practically demonstrate the error characterization, we a developed statistical cluster analysis with visual interpretation, on the empirically processed terrain parameters: DEM 20 elevations, slope, aspect, curvatures, and an error field as difference between the DEM 20 and a lidar DEM, which was used as a reference ground truth (Figure 2). This result helps better understand patterns of different types of errors that occur in the target DEM, and removal gross and systematic errors (in our case local horizontal shifts).

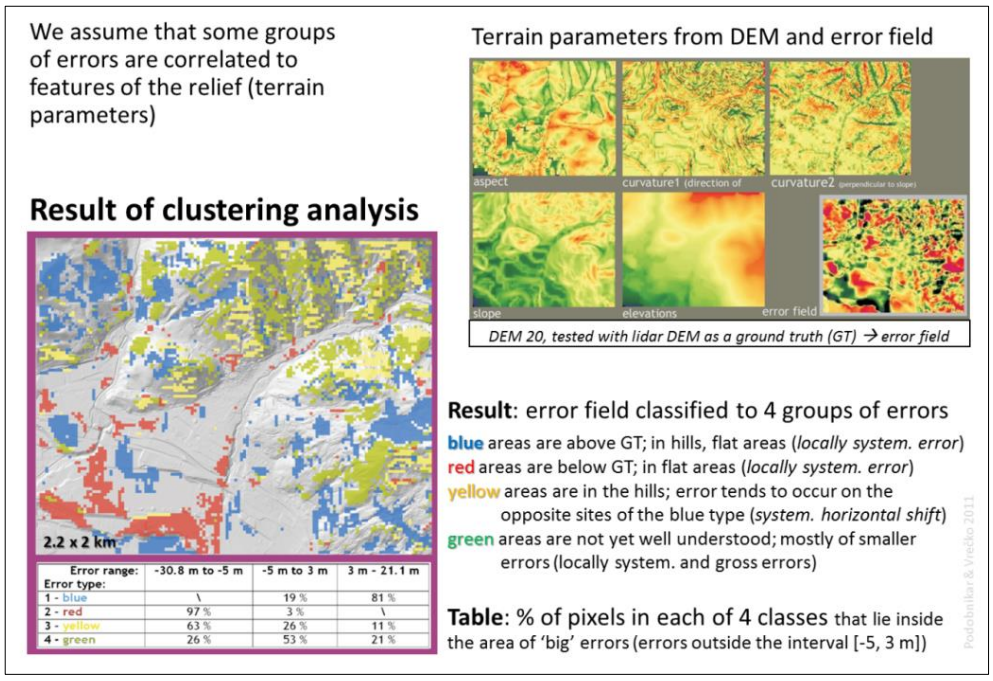


Figure 2: Error characterization with clustering analysis (area of Koroška Bela, NW Slovenia, 2.2 x 2 km).

Spatial error model (for the 2nd correction)

Our philosophy for spatial error model building is similar to reverse engineering, where we need to extract a knowledge/semantics (in our case through the error characterization process) and

then reproducing it (in our case with processing the spatial error model). An advantage of the spatial error model calculation is to better understand, control and remove the errors in DEM. This model has been also used for learning purposes, in order to understand the key issues of potentially unsuccessful DEM processing design and subsequently to improve the design.

Various kinds of DEM error model are known, such as spatial conditional stochastic models with Monte Carlo simulations (Podobnikar 2008), or interpolated from checkpoints with kriging (Leon et al. 2014). Comparable is an approach with (linear) regression models that define a relationship between DEM error and morphometric character (distribution and scale of elevation; Carlisle 2005). Due of the proved non-stationarity of parameters that vary across space, a local GWR technique is a valuable solution, too (Erdogan 2010). An empirical error model that assumes correlation of the DEM errors with terrain morphology, sampling density, and interpolation method, has been developed based on approximation theory (Liu et al. 2012). This model doesn't require a reference dataset, i.e. any checkpoints, but only the source data and the DEM.

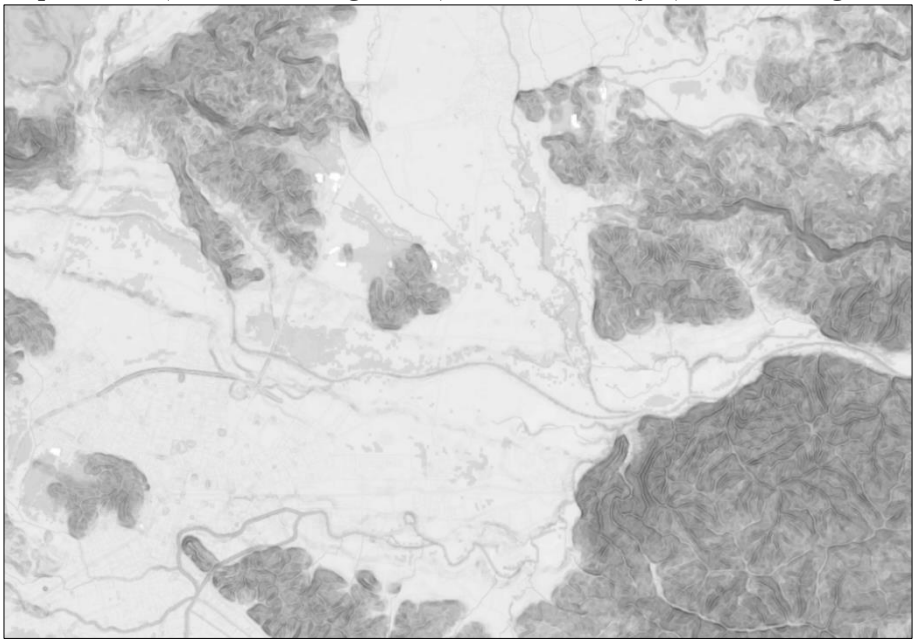
Our proposed spatial error model is empirical and predictive. It basically determines locally systematic error as a combination of different variables, where the procedure is following:

- *significant terrain parameters* are processed through defining indices that influence to a DEM error; this process requires knowledge acquisition about error sources and types
- *homogeneous regions* within each terrain parameters are calculated with classification (and employing segmentation methods, where the aim to establish these regions is similar to applying stratification used for better spatial data sampling): in each region a stationarity of the stochastic process is assumed, i.e. the error within regions needs to be homogeneous with a certain threshold; statistical tests using reference points and DEM, or operator's knowledge help in the region definition
- *variables* V_i are calculated for corresponding terrain parameters with an assignment of the unique error values to each region
- empirical *spatial error model (SEM)* is processed as: $\mathrm{SEM} = \sum_{i=1}^{n}(\boldsymbol{V_i} \cdot \boldsymbol{w_i})$, where V_i is variable (e.g. errors in regions of terrain parameter 'roughness'), and w_i is weight on interval [0, 1], where $\sum_{i=1}^{n} \boldsymbol{w_i} = \mathbf{1}$
- *spatial data cleaning* is applied using spatial error model and variables

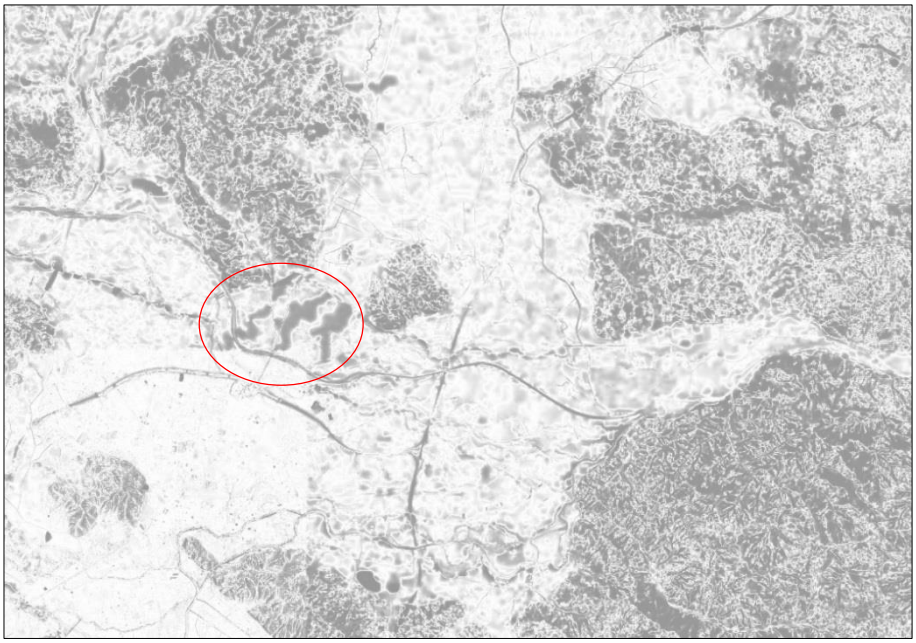
For the practical application, we developed a spatial error model (SEM). We propose following *significant terrain parameters*: RR for main terrain characteristics, RN for slope of terrain, an RG for land cover information and RO for terrain skeleton. *Homogeneous regions* for all four terrain parameters were calculated. The RR was classified into following four classes: flat surface (plains), low hills, hills, and mountains, according to the slope, profile curvature, and elevation. The standard region RN was classified into three classes: 0°–5°/5°–20°/> 20°. The RG was classified into categories of bushes, deciduous forest, mixed forest, coniferous forest and open areas. The last one, RO was classified into regions of peaks, pits, saddles, ridges and talwegs. RMSE values were assigned to terrain parameter's regions to get *variables* V_i. We applied exploratory statistical tests, with regards to our previous knowledge and experiences. Finally, the SEM was developed. Weights were assigned empirically: $w_1 = 0.5$ for RR, $w_2 = 0.27$ for RN, and $w_3 = 0.23$ for RG. A sum up of all three variables bearing in mind the proposed weights was calculated using map algebra operators. Finally, the RO was overlaid in order to calculate a spatial error model (Figure 3a).

For assessment of the spatial error model quality, we calculated an *absolute error* (Figure 3b). It was calculated as an absolute difference between a DEM from contour lines digitised from topographic maps in scale 1:25,000, and considerably better National DEM 12.5 used as a reference, which was produced mostly with stereo-photogrammetric methods (Podobnikar 2005).

The result (Figure 3) demonstrates very similar spatial error pattern for both, the predicted and the calculated error. It is obvious that the prediction model considers the right terrain parameters, variables, and weights. However, the most obvious differences between both are due to interpolation problems (encircled in Figure 3b) and other not (yet) eliminated gross errors.



a)



b)

Figure 3: A DEM error field (darker is area, larger is error – up to ~8 m): a) spatial error model (prediction); b) absolute error, calculated as difference between independent reference and target DEM 12.5 m (area of surroundings of Ljubljana, central Slovenia, 20 x 15 km).

The spatial error model (prediction) and the variables were separately used for removal particular locally systematic and gross errors. Figure 4 shows improvement of the DEM from contour lines with the variable of terrain skeleton (RO).

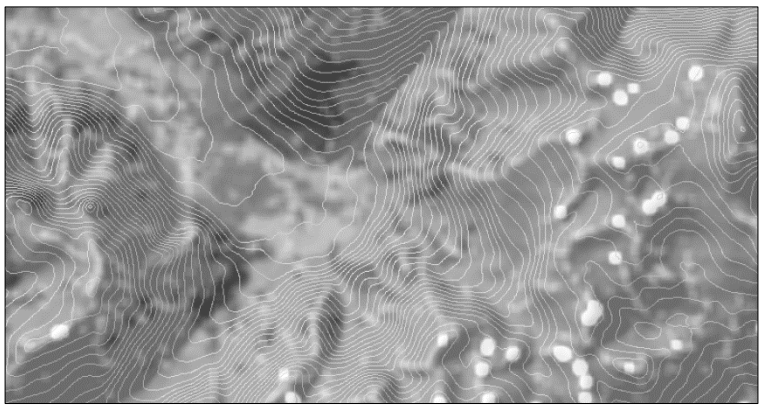


Figure 4: DEM from contour lines improvement with variable RO. The largest errors of underestimated surface (up to -3.5 m – white areas) were found around the sharpest mountain peaks (area of Julian Alps, NW Slovenia, 4.5 x 3 km).

III DISCUSSION AND CONCLUSIONS

The paper proposes a novel combined approach for spatial errors characterization and prediction with a spatial error model, in order to improve a quality of the digital elevation model (DEM). The method for error characterization with clustering analysis was developed, as well as a method for empirical spatial error model processing based on terrain parameters used to derive homogeneous regions and variables. Both methods use combined solutions with statistical, empirical and visual quality control, in order to describe spatial pattern of error and its magnitude. They are followed by DEM error removal and correction methods for attributes correction or filling gaps with other datasets. Furthermore, four main groups of error that characterize positional errors in DEM were proposed. Besides of the traditional random, systematic, gross, a locally systematic error has been suggested, with its very complex nature that includes geomorphological components. The latest is actually the most typical error in DEM, which is usually too simply considered as a random error.

The planimetric (horizontal) error of the DEM is actually more complicated to assess than of the vertical one. It is also difficult to separate between vertical and planimetric errors, especially on a flat surface. Considerably more challenging is to assess its geomorphological properties. Moreover, in the practice, it is needed a very good quality reference DEM for error characterization and removal, as demonstrated in this paper, which is usually not available. In the case when there is not available any reference data, it can considerably – but not absolutely helps to use a set of independently acquired DEMs, even if they are not perfect quality. Another problem is a complexity of the error field that cannot be easily distinguished from the roughness/rugosity of the terrain, what needs a special attention for the future research.

The resulted error model based on an absolute error is surprisingly high quality, and very nicely reflects the real situation. Besides of the support of error characterization methods, our expert knowledge was also implemented in this model. The solution with spatial error characterization and model considerably better explain DEM error field than the global measures, such as RMSE. An advantage of the proposed empirical spatial error model is that the error pattern has been better understood, controlled and the errors in DEM were removed/corrected.

The future design of the more automated solution needs implementation a mixture of complex experiences and methods. Our next step of research will be a reconstruction of full error field without only absolute values as in this study, applying more comprehensive sensitivity analysis and knowledge-based systems. We will step-by-step apply other terrain parameters, which may influence the quality of the error model. Since the error characters of InSAR or lidar DEMs are considerably different, the terrain parameters and corresponding classifications will consider these options with different scales, and not only the DEM from photogrammetric data sources as in this case. Moreover, our future error assessment methodology will better consider specific demands of particular users.

Acknowledgements

Part of the study was supported by the Water Science and Technology, and Geotechnics Program [P2-0180] of the Slovenian Research Agency (ARRS). The author is also grateful to A. Vrečko for her contribution in the clustering analysis.

References

Carlisle B.H. (2005). Modelling the Spatial Distribution of DEM Error. *Transactions in GIS 9*(4), 521–540.

Devillers R., Stein A., Bedard Y., Chrisman N., Fisher P., Wenzhong, S. (2010). Thirty years of research on spatial data quality: Achievements, failures, and opportunities. *Transactions in GIS 14*(4), 387–400.

Erdogan S. (2010). Modelling the spatial distribution of DEM error with geographically weighted regression: An experimental study. *Computers & Geosciences 36*(2), 34–43.

Fisher P.F., Tate N.J. (2006). Causes and consequences of error in digital elevation models. *Progress in Physical Geography 30*(4), 467–489.

Guth, P.L. (2006).Geomorphometry from SRTM: Comparison to NED. *Photogrammetric Engineering & Remote Sensing 72*(3), 269–277.

Hodgson E.M., Bresnahan P. (2004). Accuracy of Airborne Lidar-Derived Elevation: Empirical Assessment and Error Budget. *Photogrammetric Engineering & Remote Sensing 70*(3), 331–339.

Höhle J., Potuckova M. (2011). *Assessment of the Quality of Digital Terrain Models*. EuroSDR.

Imhof E. (1982). *Cartographic Relief Presentation*. Berlin and New York: Walter de Gruyter.

Kimball R, Caserta J. (2004). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. New York: John Wiley & Sons.

Leon J.X., Heuvelink G.B.M., Phinn S.R. (2014). Incorporating DEM Uncertainty in Coastal Inundation Mapping. *PLoS ONE 9*(9), 12 p.

Liu XH., Hu P., Hu H., Sherba J. (2012). Approximation Theory Applied to DEM Vertical Accuracy Assessment. *Transaction in GIS 16*(3), 397–410.

Oksanen J., Sarjakoski T. (2006). Uncovering the statistical and spatial characteristics of fine toposcale DEM error. *International Journal of Geographical Information Science 20*(4), 345–369.

Podobnikar T. (2005). Production of integrated digital terrain model from multiple datasets of different quality. *International Journal of Geographical Information Science 19*(1), 69–89.

Podobnikar T. (2008). Simulation and representation of the positional errors of boundary and interior regions in maps. In *Geospatial vision, new dimensions in cartography*, (Lecture notes in geoinformation and cartography), Springer, pp. 141–169.

Wise S. (2012). Information entropy as a measure of DEM quality. *Computers & Geosciences 48*, 102–110.

Zandbergen P.A. (2011). Characterizing the error distribution of lidar elevation data for North Carolina. *International Journal of Remote Sensing 32*(2), 409–430.

Spatial uncertainties simulation and propagation

Exploring the uncertainty of soil water holding capacity information

Linda Lilburne*, Stephen McNeill, Tom Cuthill, Pierre Roudier

Landcare Research, Lincoln, New Zealand

*Corresponding author: lilburnel@landcareresearch.co.nz

Abstract

Soil water holding capacity is an important soil property for understanding how much irrigation water is required and the quantity of nutrients that are likely to leach into groundwater. This soil profile level property is derived from horizon level data including soil water content, stone content, thickness and horizon type. Soil water content data is expensive to measure so is often estimated in a model that is based on more readily collected soil information. A model of the soil water content (or hydraulic response) has been developed and tested. Its inputs include sand and clay content, profile and horizon classifications. Thus, uncertainties in the derived estimate of soil water holding capacity are due to variability in the inputs to the soil hydraulic model, error in the model itself, and variability of the other key soil properties (stone content, thickness). The combined uncertainty is estimated in a new Soil Profile Simulator tool. It is based on a simulation approach that draws upon the statistical error model of the soil hydraulic model and expert information held in a national soil survey database in the form of probability distributions. This expert information characterises the variability of clay, sand, horizon thickness, stone content, and uncertainty in the classification information associated with the soil polygons. This paper describes the tool, reporting on the progress that has been made in deriving and visualising quantified estimates of uncertainty of soil water holding capacity. Recent advances in technology, including new R packages (aqp, VGAM) and Rserve have been behind this progress.

Introduction

Information on soil hydraulic properties is essential for the sustainable management of irrigated agricultural land. Providing too much water is wasteful and contributes to contamination of water through leaching and runoff of nutrients. Too little water will impact negatively on yield. One of the key soil properties for managers of irrigated land is the soil's water holding capacity or profile available water (PAW), which represents the amount of water held in the soil that is readily accessible to plants. In New Zealand, this is taken to be the water content held between soil moisture tensions of 10 and 1500 kPa. Although this can be measured directly, it is an expensive process, so PAW is commonly predicted using other soil properties that are easier to measure or observe. These include texture, carbon, bulk density and soil morphology. A relationship between soil properties is known as a pedotransfer function (PTF).

The soil hydraulic response has a number of features that make model predictions difficult. First, the response is bounded (0–100%), so either a bounded-value regression (e.g. logistic) is required or a transformation is needed to an unbounded domain. Second, the response is monotonic with respect to the tension: that is, the response strictly decreases as the tension increases. Finally, the response at a given tension is correlated with the response at other tensions. This complexity also adds to the challenge of estimating the error of the predicted response.

The soil properties used as inputs by the PTF are also uncertain. This paper presents progress with estimating and communicating uncertainty of PAW, whereby the error model of the soil hydraulic PTF is combined with probabilistic information on variability of the key soil inputs of the PTF held in a soil survey database. A new visualisation R package 'aqp' is used to visualise the results.

Methods

Soil hydraulic error model

In forming an empirical model for soil hydraulic response, we use soil sample data available from the New Zealand National Soils Database (NSD), which provides the response at tensions of 0 (total porosity), 5, 10, 20, 40, 100, and 1500 kPa. For each sample, texture (sand, silt, clay fractions) data are available, as well as the soil classification, and other factors describing the soil sample.

Our methodology for response prediction uses a vector generalised linear model (VGLM). This, according to Yee (2015), can be thought of as a generalisation of the generalised linear model (GLM) with a vector of responses, which is free from many of the restrictions that the GLM method imposes. We use a vector of responses formed from the logit-transformed 1500 kPa response (transformed to give an unbounded range), as well as the difference between the logit-transformed 100 and 1500 kPa responses, the difference between the logit-transformed 40 and 100 kPa responses, and so on, up to the difference between the logit-transformed 5 and 0 kPa response. The response at one of the specified tensions is formed from the VGLM prediction for 1500 kPa, plus a succession of differences for lower tensions. The uncertainty of each marginal response (at 1500 kPa or a difference in response between tensions) is formed from the estimated VGLM model.

An error model for the difference between the 10 and 1500 kPa responses (i.e. estimates of total available water within a soil horizon or layer), or indeed for any other convenient combination of responses, has been verified using independent validation data (McNeill et al. in prep.). Uncertainty limits, calculated in terms of containment intervals, are estimated by simulation of the aggregated response.

Soil Profile Generator tool

Lilburne et al. (2012) described how information on soil variability and uncertainty was being incorporated in a national-scale soil database for New Zealand called S-map (Landcare Research, 2015). The very limited amount of soil sample data meant that an expert knowledge approach was used to record information on the confidence of classification and base property attributes, the variation of soils in a polygon and their proportional reliability, and the range of values of key soil properties (Lilburne et al. 2008). Variability of quantitative soil properties is stored in the form of probability distribution functions (pdf). The lack of point soil sample data precluded a geostatistical approach to modelling and simulating PAW.

A new Profile Simulator tool has been developed that creates realisations of profiles based on the expert-derived information on soil variability and uncertainty. The key information used in this study on PAW is the PTF error model; expected variability of the stone, sand and clay content; type and thickness of functional horizons within a soil survey polygon; confidence in the Soil Order classsification, rock type of the fine material and drainage class, as well the reliability of the proportions of soil types within a polygon. Values for horizon stone, sand, and clay content,

and thickness were drawn from their respective pdfs. Each profile realisation is checked against a set of rules to ensure that it is still consistent with the soil definition. For example, the sum of the simulated horizon thicknesses must fit the pdf of the soil's depth. Realisations that do not fit the rules are discarded and regenerated. The PTF error for each soil profile realisation was simulated and the PAW calculated.

The tool is based on an architecture that links the S-map database in SQL Server with RServe – a server that responds to requests from clients by running a R script. Rserve enables R calculations to be performed on request without the need to start up an R session each time. A number of servlets have been developed that retrieve the stored uncertainty information from SQL Server, call a set of Monte Carlo functions to generate realisations of the key parameters required by the soil hydraulics error model, run this model on each set of parameter realisations, simulate model error, and finally, return the set of profile realisations with estimates of available water, in a range of useful formats (database tables, csv file, Rdata file). RServe was installed on a Linux-based server to allow for multi-threaded processing.

Visualisation

The 'aqp' package (Beaudette et al. 2013) is an R package designed for working with soil information. It contains a SoilProfileCollection class to simplify the process of working with the collection of data associated with soil profiles. It also includes tools for plotting soil properties by depth and their associated variability.

Results

Using a training dataset of 1007 points, a VLGM model was developed and tested on an independent dataset of 432 points (Figure 1). Histograms of the residuals for seven tension values, and the associated model diagnostics using training and validation data, showed similar patterns, indicating that the model was not overfitted. Simulated error for the 1500 kPa tension value and total available water for an example horizon is shown in Figure 2. Response distributions are formed by simulation, and can be processed to give containment intervals (e.g. 95%). While most response distributions appear symmetric, the response can in some cases be highly skewed where there is little training data (rare soil classes). The simulated water retention curves for the 500 realisations of a Barr_6a sibling (a moderately deep sandy loam) are shown in Figure 3.

The Soil Profile Simulator is fast, generating 10 000 realisations of a specified soil sibling as a Rdata file in 28 seconds. The tool can output the realisations as tabular data or as a SoilProfileCollection (the class used in aqp). Use of the R aqp package facilitates drawing graphs of the soil profile. Figure 4 compares the 500 realisations of estimated PAW with the PAW that is listed on the S-map fact sheets (123.17 mm). The fact sheet value is estimated using the mean soil hydraulic response with mean values for clay and sand content, along with mean horizon thickness, stone content, and modal horizon classification. The mean PAW of the 500 realisations is 122.96 mm with a standard error of 1.2356 and a standard deviation of 27.62. Figure 5 shows the 500 soil profile realisations in terms of stone, sand and clay content down the profile. The variability of PAW down the profile is shown in Figure 6, where the darker grey area is the 25% to 75 % quantile, and the dashed lines show the 95% confidence interval.

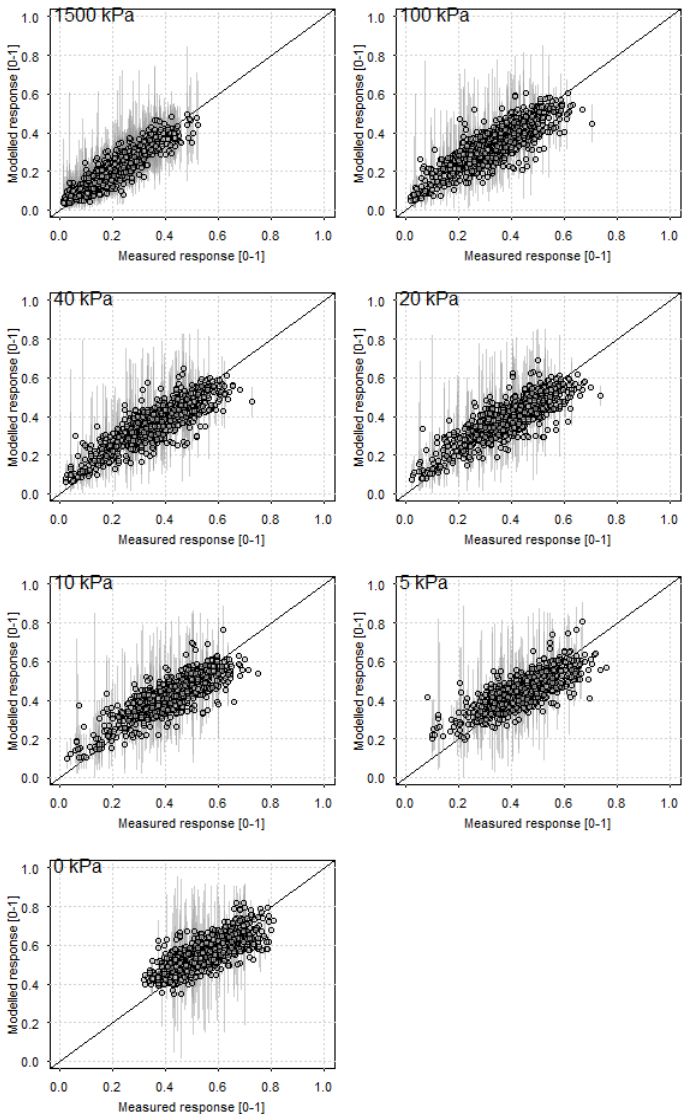


Figure 1: Measured-versus-fitted plots of the soil hydrological response for seven tension values. The vertical lines are plus and minus one standard error of the prediction uncertainty.

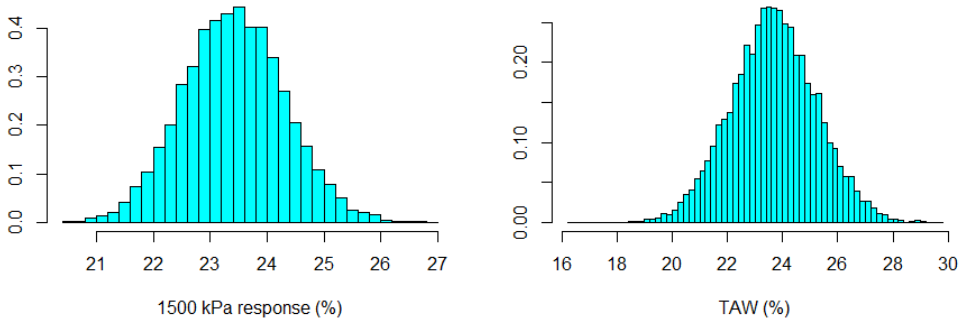


Figure 2: Histograms showing the simulated error (1500 kPa and total available water (TAW)) for an Allophanic soil with a loamy functional horizon (sand 15%, silt 57.5%, clay 27.5%).

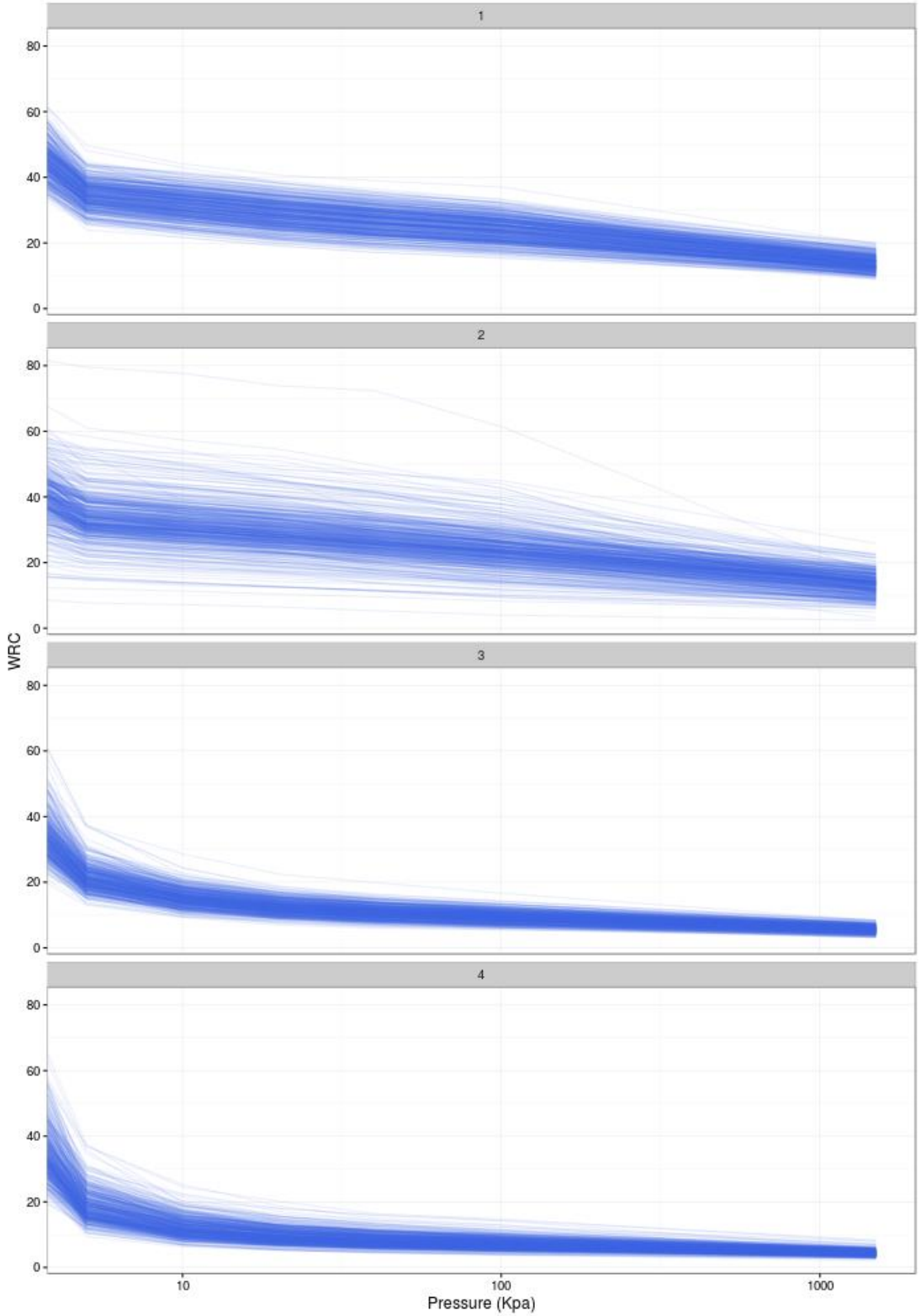


Figure 3: Simulations of the water retention curve for each of the four horizons of the Barr_6a sibling (a moderately deep sandy loam) as simulated by the soil hydraulic PTF error model.

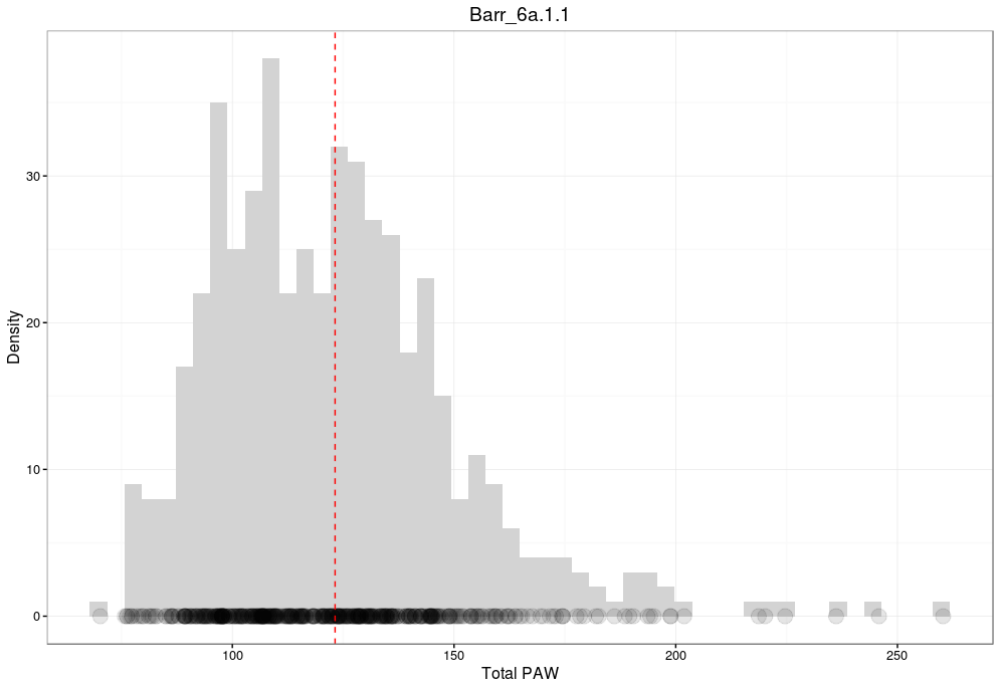


Figure 4: Histogram of the profile available water (PAW) of 500 realisations of the Barr_6a sibling. The red line shows estimated PAW where the mean values are used for the soil hydraulic response; clay, sand, stone content; and horizon thickness. The circles along the x axis indicate the distribution of the PAW estimates.

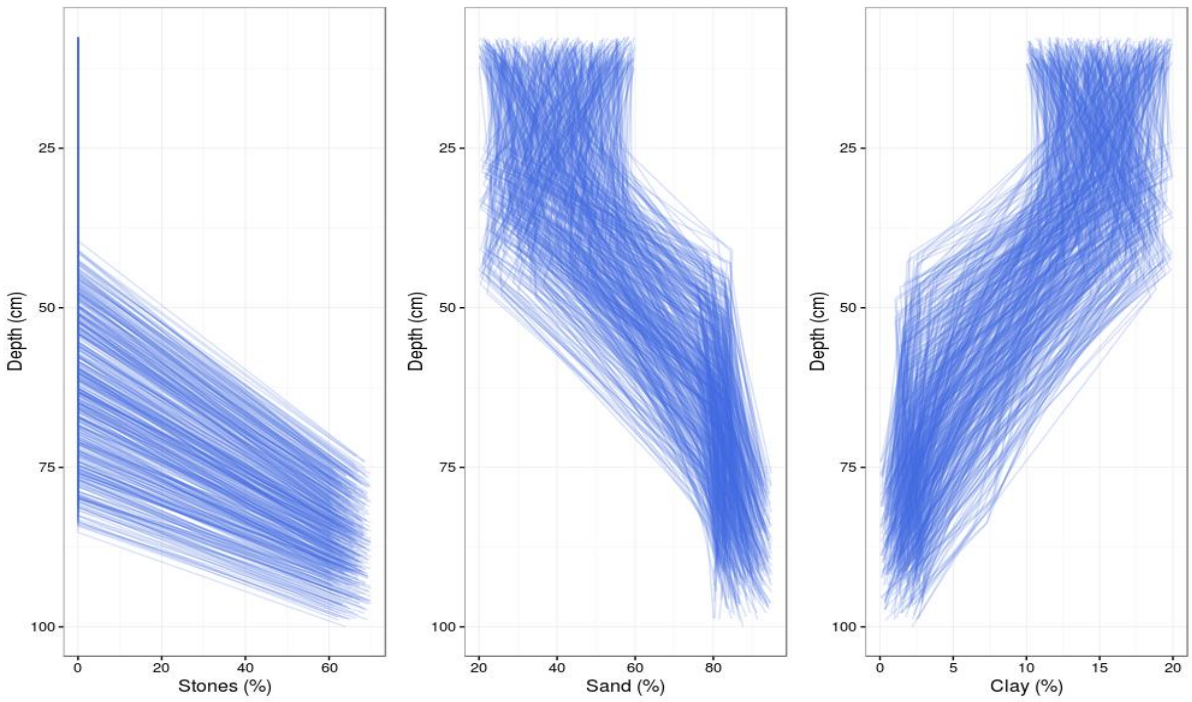


Figure 5: Soil profile realisations (n = 500) from the Soil Profile Simulator showing the variability of stone, sand and clay content down the soil profile of the Barr_6a sibling.

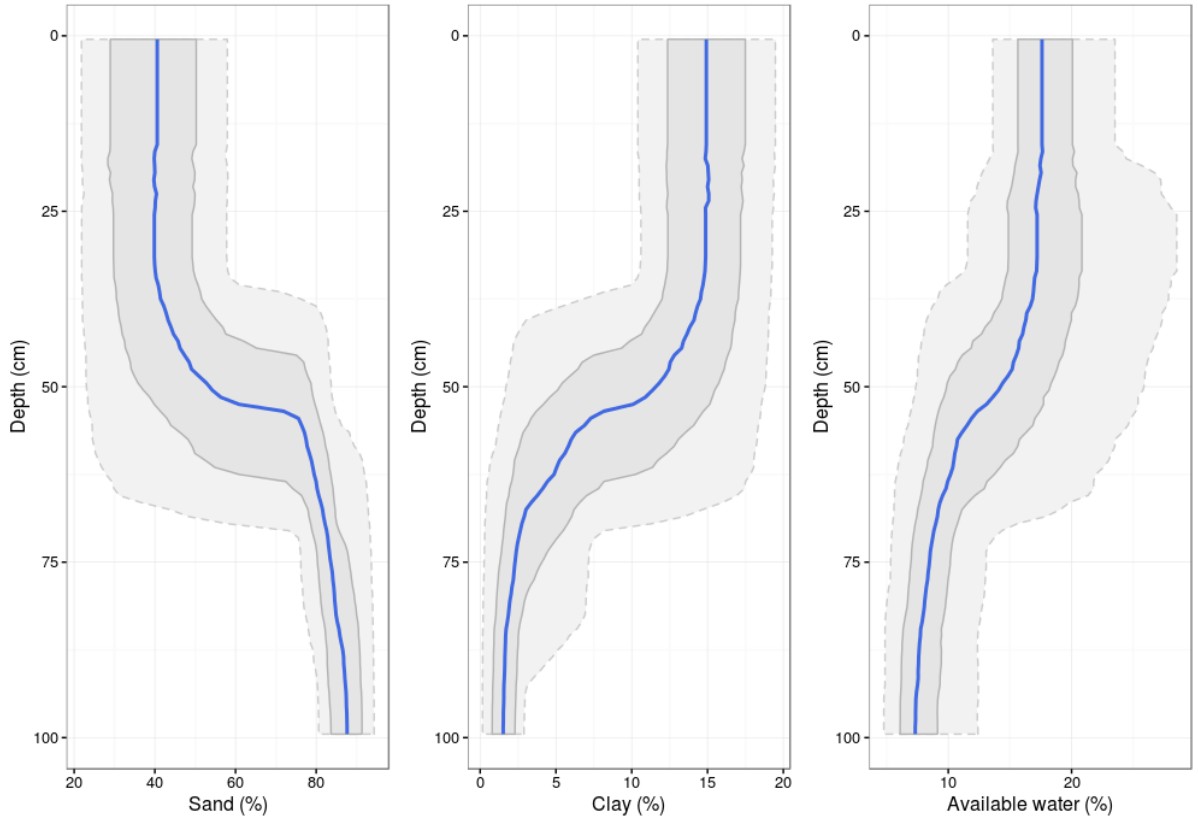


Figure 6: The variability of profile available water (PAW) down the Barr_6a soil sibling profile. The blue line is the median PAW from the 500 realisations, the solid and dashed grey lines indicate the 5% and 25% quantiles, respectively.

The speed of the Soil Profile Simulator tool also allows many siblings to be simulated according to the uncertainty of the proportions of each soil type within the soil survey polygons. Thus map realisations of PAW can be generated.

Discussion

Spatial uncertainty

The use of an expert approach to recording information about uncertainty within S-map is currently limited by the lack of any quantitative information about the spatial accuracy of the polygon line work. In the future, as we build in knowledge of soil-landform relationships, either explicitly or via digital soil mapping approaches, we will be able to represent this spatial uncertainty. For example, in a case where a soil type is linked to concave slopes and gullies, the weaker the membership of a DEM pixel in the concave-slope/gully class, the more uncertain the association with the soil class. This uncertainty can be added to the simulation tool. An alternative approach could be to use an expert approach to simplistically quantify the possible offset of each polygon line boundary.

Correlation

Another current limitation is the lack of information on both correlation between soil properties and spatial autocorrelation. The former was accounted for by using an approach whereby simulated profiles that did not meet known constraints were discarded. For example, S-map

contains pdfs of horizon thickness. Some horizons will be inversely correlated – if one horizon is thicker than the mean, then at least one other horizon must be thinner in order for the overall depth (up to 1 m) to be maintained. If the sum of the simulated thicknesses was not consistent with the range of the depth pdf, then that realisation was discarded and another regenerated. The tool stops after one thousand attempts to generate a valid realisation. However, lack of information on the spatial autocorrelation of soil properties means that this aspect cannot be addressed at this point.

Conclusion

Conventional soil survey databases do not lend themselves to exploring the impact of uncertainty in soil information. However, progress has been made in deriving and visualising quantified estimates of uncertainty in key soil properties related to the soil water holding capacity. This has been enabled by recent advances in technology, including new R packages (aqp, VGAM) and Rserve.

Acknowledgements

This work was funded by Landcare Research Core Funding. We thank Stella Beliss and Varvara Vetrova who kindly reviewed the manuscript, and Leah Kearns for editing it.

References

Beaudette, D.E., Roudier, P., O'Geen, A.T. (2013). Algorithms for quantitative pedology: A toolkit for soil scientists. *Computers & Geosciences 52*, 258–268.

S-map - New Zealand's national soil layer. DOI: http://dx.doi.org/10.7931/L1WC7. Accessed: 2016-05-16.

Lilburne, L., Hewitt, A., Ferriss, S. (2006). Progress with the design of a soil uncertainty database, and associated tools for simulating spatial realisations of soil properties. In: M. Caetano, M. Painho (Eds.), 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. Instituto Geografico Portugues, Lisbon, Portugal, pp. 510-519.

Lilburne, L.R., Hewitt A., Webb, T. (2012). Soil and informatics science combine to develop S-map: a new generation soil information system for New Zealand. *Geoderma 170***,** 232–238.

McNeill, S., Lilburne, L., Webb, T., Cuthill, T. (in prep.) Pedotransfer functions for hydrological properties of New Zealand soils using S-map information.

Yee, T.H. (2015). *Vector generalised linear and additive models*. Springer.

Simulation of realistic digitizing errors on geographical objects using constraints modeling the operator input process

Jean-François Girres[*1]
[1] Université Paul-Valéry Montpellier 3 – UMR GRED, France

*Corresponding author: jean-francois.girres@univ-montp3.fr

The geometry of vector objects in a geographical database is mainly obtained through manual input processes performed by an operator. It is well known that this capture process can cause errors - called digitizing errors - that affect the positioning of geographical objects or the computation of geometric measurements (i.e. length and area). To model these impacts, simulation methods, based on Monte-Carlo approaches, are generally proposed. However, several problems can occur by simulating digitizing errors with simple random processes (e.g. topological errors or non-respect of the object's shapes). Indeed, these problems would have been avoided by an operator performing a classical manual input process. Thus, in order to simulate digitizing error with a higher degree of realism, it seems important to understand and model some operator's behaviors. In this context, this paper proposes an approach to simulate digitizing errors using a set of constraints that try to reproduce the capture mechanisms of the operator. To contextualize these issues, principles and limitations of digitizing error simulation methods are presented in section 1, before detailing in section 2 the proposed constraints to model the operator input process. Finally, an experiment of the proposed methods is performed in section 3, before concluding.

I SIMULATION OF DIGITIZING ERROR : PRINCIPLES AND LIMITATIONS

The geometry of geographical vector objects is mainly produced through manual input process performed by an operator. This human capture process can generate accidental and systematic errors, called digitizing errors, that affect the positioning of the vertices of the geometry. These errors impact the positioning of objects, the computation of geometric measurements (length or area) or any analysis performed with the geometries of vector objects (e.g. buffers, overlaps). Thus, to model the positional uncertainty involved by digitizing error and assess its impact, several contributions (Keefer et al, 1988; Hunter and Goodchild, 1996) propose to use simulation methods, based on Monte-Carlo approaches. Using these approaches, simulated geometries are created by generating random errors in the positioning of each vertex of the geometry of the original objects. As proposed by Bolstad et al (1990), a normal law can be used to model the imprecision in the positioning of the vertices.

The simulation of errors on a geometry can be seen as the translation of the coordinates x and y of each vertex. As exposed on figure 1, two methods can be used to translate the vertices in a simulated geometry:

- in the first method (on the left), the coordinates of a vertex *P* are translated according to an angle *a* (generated randomly), at a distance *d* from the original vertex position (*d* is generated using a normal law)
- in the second method (on the right), the coordinates *x* and *y* of a vertex *P* are translated at distances *dx* and *dy* from the original vertex position

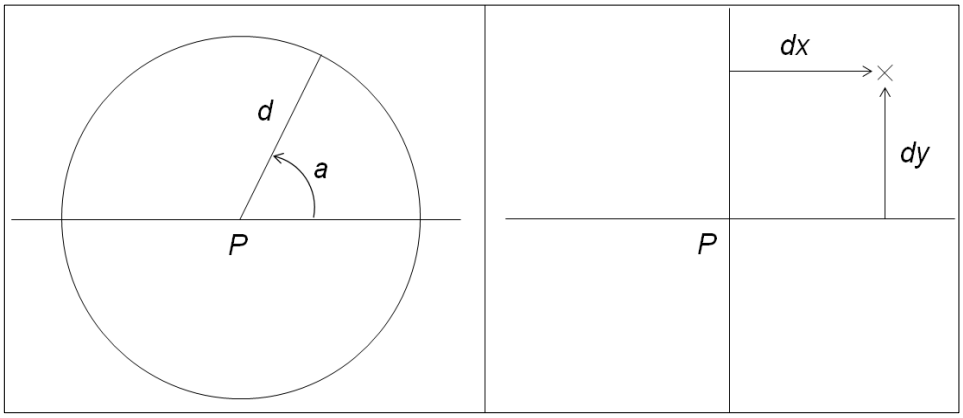


Figure 1: Two methods to translate the coordinates x and y of a vertex P

The first method, which is considered as more correct to represent accidental errors (Vauglin, 1997) is used in this study to simulate digitizing errors on a geometry. This operation is performed on each vertex of the original geometry in order to generate a simulated geometry. By repeating this operation on a large set of realizations (figure 2), it is then possible to study the sensibility of geometric measurements on simulated geometries (by comparing lengths or areas with the original geometry) and then assess the impact of digitizing error on measurements, as proposed by Goodchild et al (1999).

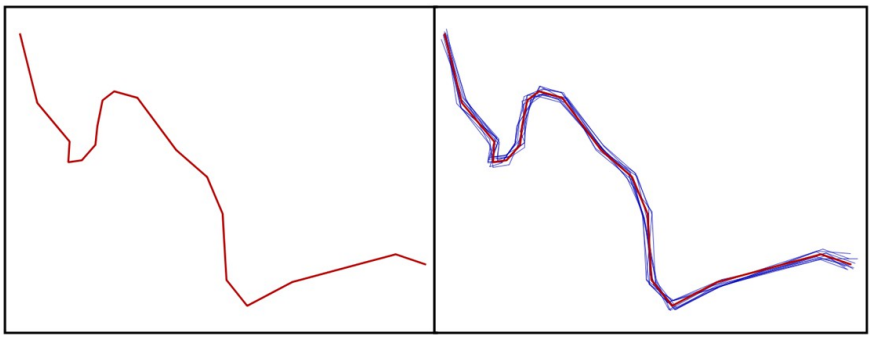


Figure 2: An original geometry (left) and 10 simulated geometries (right)

However, several problems, which would have been avoided by a "human" operator, can occur by simulating digitizing errors with simple random processes. The first type of problem is related to topological errors (e.g. self-intersection of edges) as exposed in figure 3. Several contributions have already been proposed to avoid such topological inconsistencies (Hunter and Goodchild, 1996; Bonin, 2002).

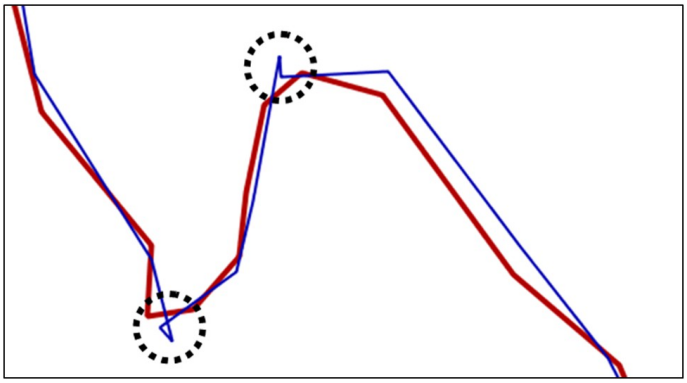


Figure 3: Topological inconsistencies on a simulated geometry (in blue)

The second type of problem deals with the non-respect of the original shape of the object after a simulation. For instance, as exposed in figure 4, the simulation of random digitizing errors on an original polyline with a straight shape (in red) can generate a distorted simulated geometry (in blue). As a consequence, the simulated polyline loses its original straight shape, and its length automatically increases (because a straight line minimizes the length). Thus, this simulation can be considered as unrealistic, because a human operator would have preserve the straight shape of the polyline.

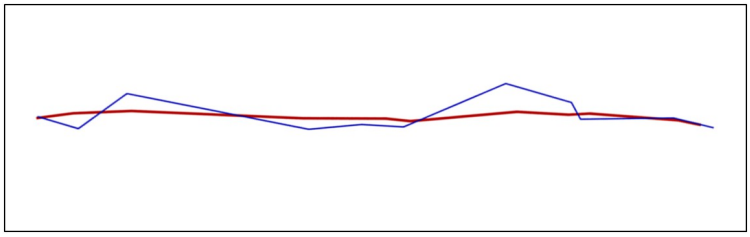


Figure 4: Non-respect of the original shape of a straight line using random simulations

These two examples illustrate a simple fact : the use of simple random models is not adapted to simulate human input processes. Indeed, the operator input process involves mobilizing cognitive mechanisms that are complex to model. So, in order to simulate digitizing errors with a higher degree of realism (and assess their impact on measurements), some operator's behaviors need to be understood and modeled in the simulation process. The next section of this paper will present propositions in order to model the operator input process.

II SIMULATION CONSTRAINTS TO MODEL THE OPERATOR INPUT PROCESS

In order to simulate digitizing errors with more realism, this section presents a set of constraints that try to reproduce the capture mechanisms of the operator. Based on basic assumptions about the data entry process, three types of constraints are integrated in the digitizing error simulation model in order to preserve the realism of the generated objects : the weighting of errors on vertices according to their angularity (1), the weighting of errors on nodes according to their degree (2) and the correlation of errors on vertices using progressive translation from the extremities (3).

2.1 Weighting of errors on vertices according to their angularity

The first type of constraint on the simulation process deals with the weighting of errors on vertices according to their angularity. This constraint is integrated in order to avoid the distortion of straight lines, as exposed in figure 4, that a human operator would have preserved.

The basic underlying assumption is that the angle between successive vertices of a geometry is a representation of the reality determined with the consciousness of the operator. So, if successive vertices of a geometry are aligned, a realistic error simulation should preserve this alignment. On the other hand, if brutal turns in the direction between successive vertices are observed, the risk of digitizing error might be more important. This basic assumption involves to take into account the angularity between successive vertices as a constraint on the simulation of digitizing error.

To model this constraint, a weighting q (between 0 and 1) of the distance d of the error on each vertex P of the geometry is determined according to the angle between successive vertices $P\text{-}1$, P and $P+1$. For instance, A, B and C are three successive vertices. To determine the weighting q_B of the error on the vertex B, its angularity α_B (in radian) is computed using equation 1.

$$\alpha_B = \frac{\left|\widehat{ABC} - \pi\right|}{\pi} \qquad (1)$$

Once the angularity of each vertex is determined, a function can be used in order to determine the weighting q according to the angularity α. Indeed, based on our assumption, the more the angle is close to π , the more the error should be small. But it does not mean that the error is null (which would be the case with $\alpha = 0$). Moreover, it is extremely rare to get angles between successive vertices where the angularity α_B is close to 0 or 2π.

As a consequence a function need to be defined in order to determine the weighting q of the error according to the angularity α. This function can take basic forms (e.g. $q=0.75\alpha+0.25$; $q=\alpha^{0.5}$) or more complex formulations, but the goal of this paper is not to define the best function to determine the appropriate weighting according to a given angularity. In the experiments (section 3), a basic form will be used.

Once the function is defined, the weighting of digitizing errors can be applied on each vertex of the geometry using a normal law $N(0, \sigma)$. As exposed on figure 4, which represents the simulation of errors using a weighting according to the angularity, the more the angularity is important, the more the the error on the vertex is important.

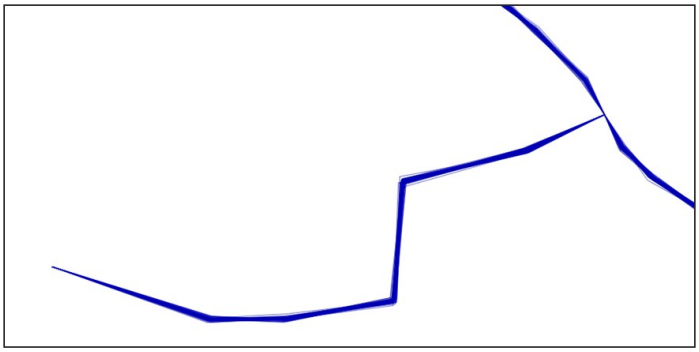


Figure 4: Simulation of errors weighted by the angularity between successive vertices

The example of simulation exposed on figure 4 shows that no errors is applied on the extremities of the polyline, where it is not possible to determine any angularity value. To determine the error applied on the extremities of a polyline, a second method, based on the degrees of the extremity nodes, is exposed in the following subsection.

2.2 Weighting of errors on extremity nodes according to their degree

Because it is not possible to apply a weighting based on the angularity between successive vertices on the extremities of a polyline, a method based on the degree of extremity nodes is proposed.

The underlying assumption formulated to determine this constraint is that the more a node is connected with other edges, the easier its localization should be for the operator, and the more precise its positioning should be. On the other hand, if an extremity is not connected with any other edge, we can expect that it is more difficult for the operator to locate the node precisely. As a consequence, the positioning of unconnected extremity nodes should be more imprecise, and the digitizing error should be more important.

Following this assumption and in order to simulate digitizing errors with more realism, we consider that the error applied on the extremities of a geometry have to be weighted according to the degree of the edges. Thus, for a node N with a degree d located at the extremity of a polyline, we propose to apply a weighting q_N as defined in equation 2, using a constant a.

$$q_N = \frac{a}{d} \tag{2}$$

As mentioned previously, the goal of this paper is not to define the most appropriate value of the constant *a* used to compute the weighting of errors on extremities. In this paper, we will use a value of *a=1*. Nevertheless, some other parameterizations of this constant could be experimented in further researches.

The weighting q_N is then applied on the simulated errors on the initial and final nodes of the geometry using a normal law *N(0, σ)*. Figure 5 illustrates the application of the weighting of errors on extremity nodes according to their degrees, associated with the weighting of errors on vertices according to their angularity. It shows that the error applied on unconnected nodes is more important than on connected nodes.

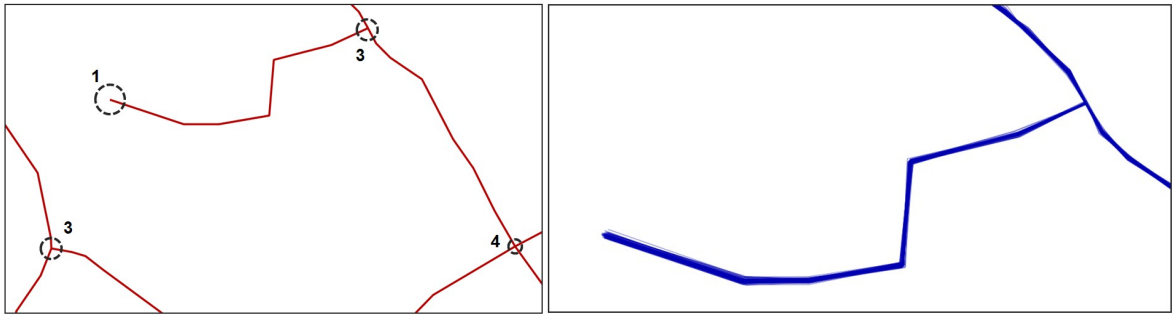


Figure 5: Determination of the degrees of nodes (left) for the weighting of errors on extremities (right)

This second constraint shows how to apply a weighting for the simulation of digitizing error on the extremities of a geometry according to its configuration. Finally, in order to take into account the dynamic dimension of the data capture process between the initial and the final node, a last constraint is proposed : the correlation of errors between successive geometries.

2.3 Correlation of errors between successive vertices

Several contributions (Heuvelink et al., 2007; De Bruin, 2008) consider that the simulation of errors on a geometry should integrate a correlation between successive vertices. It is admitted that an operator can generate a systematic error (i.e. a bias) during the input process of vector data. Nevertheless, it is difficult to simulate this error, because it depends of the operator and the way he works. But some other reasons (e.g. the dynamic dimension of the data entry process) can justify to integrate a correlation of errors between successive vertices.

To explain the integration of the correlation of errors between successive vertices, we can use the example of an operator capturing a road. If an error occurs at the extremity of the road (i.e. at the crossroad), we can assume that the operator will try to cushion progressively this error on the entire polyline, instead of correcting it brutally on a single vertex, what would generate a non-respect of the shape of the road. Indeed, the input process of vector data is a dynamic (and active) process. Thus, if the operator generates an error, he will try to correct the error without reducing his productivity. Following this example, we can consider that the error generated on the extremities (proposed in the previous subsection) is propagated progressively on the entire geometry, which supposes a correlation of errors between successive vertices.

To model the correlation of errors between successive vertices, we propose a method that generates a translation of each vertex, based on the errors on the extremities.

Using the method proposed in the precedent subsection, errors a generated at the extremities of a geometry according to the degree of the nodes, and following a normal law *N(0, σ)*. As a consequence a translation using vectors $\vec{v}$ and $\vec{u}$ is performed on the extremities *A* and *C*, and a translation is applied on an intermediate vertex *B* with a vector $\vec{w}$, computed with a weighted sum of vectors $\vec{v}$ and $\vec{u}$.

The weighting coefficients a and b are defined as the ratio between the length from the intermediary vertex to each extremity node with the total length of the polyline, as exposed on equation 3.

$$\vec{w} = a * \vec{v} + b * \vec{u} \quad (3)$$

where a is the ratio between the length of [AB] and the length of [AC] and b is the ratio between the length of [BC] and the length of [AC]

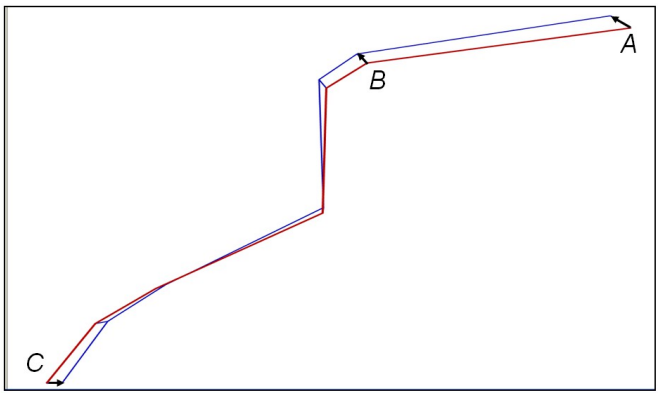


Figure 6: Translation of a vertex *B* according to the errors on the extremities *A* and *C*

It is important to notice that in figure 6, the error on intermediate vertices (using the angularity between successive vertices, as proposed on subsection 2.1) is not applied.

2.4 Combination of constraints for digitizing error simulation

The combination of the three constraints varies according to the type of geometries and their characteristics. For polylines, the three different constraints proposed to generate realistic digitizing error simulations are combined for each realization of the simulation as follows: (1) error simulation on the extremity nodes according to their degree, (2) translation of intermediary vertices according to extremity nodes errors (i.e. correlation of errors), (3) error simulation on translated vertices according to their angularity.

For polygons, it is necessary to take into account the characteristics of the represented objects. We propose to differentiate the objects representing a partition of the real world (e.g. administrative units) and objects that don't represent a partition of the real world (e.g. water surfaces). For polygon objects representing a partition of the real world, the same methodology as the one proposed for polylines is applied. Indeed, these objects suppose the existence of common borders. We can then suppose that a correlation between successive vertices exist, because the operator will start to capture a border from an initial node (the border between three units) to a final node. As a consequence, he will cushion the error from the extremities on the entire border. Thus, these polygon objects can be simplified as a graph.

For polygon objects which don't represent a partition of the real world, only the method of error simulation on vertices according to their angularity is applied. Indeed, because these geometries are closed and isolated, the notion of initial and final nodes doesn't exist. As a consequence, the method proposed to generate errors on extremities (according to the degree of the nodes) and translation between successive vertices (according to errors on extremities) don't need to be applied.

The different methods proposed in this section are experimented in the following section on real objects.

III EXPERIMENTS

To illustrate the functioning of the methods proposed to simulate digitizing errors, an experiment is proposed on a dataset of administrative units, extracted from the French BDCARTO database in the area of Handaye and Saint-Jean-de-Luz (south-west of France). The total area of these administrative units is about 133.43 sq. km.

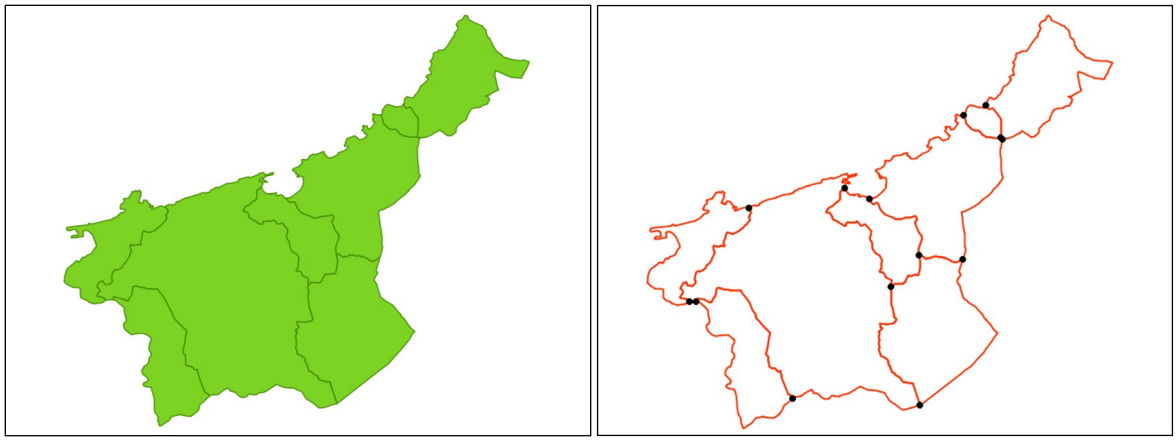


Figure 7: The administrative units used for the experiment (left) and their conversion in a topological graph (right, with edges in red and nodes in black)

To simulate digitizing errors, a normal law $N(0, \sigma)$ is used, with a value σ = 4.97 m. (determined using the automatic capture scale estimation method, see Girres, 2015 and Girres, 2012) and a function $q=0.75\alpha+0.25$ for the weighting of errors according to the angularity between successive vertices. Finally, 1000 realizations of the simulation are realized. To perform the simulation, administrative units, which are considered as a partition of the real world, are converted in a topological graph (figure 7). This method allows to extract nodes and to process each border between units as a polyline during the simulation. Finally, after the simulation, polygon units are re-built using the graph faces.

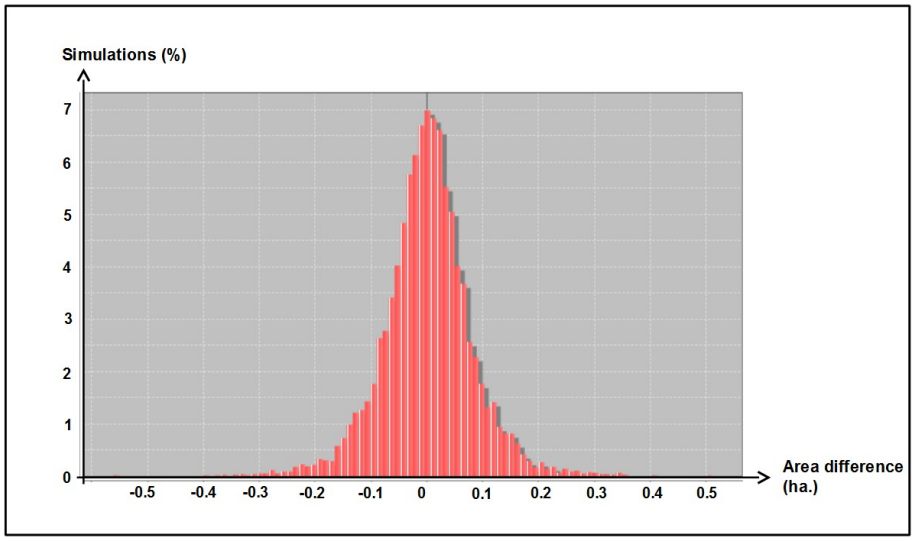


Figure 8: Distribution of area differences between original and simulated administrative units

Results of the experiments show that the distribution of area differences between the original dataset and the simulated datasets presents a value of σ_A = 0.008 sq. km. (= 0.8 ha). Assuming that this distribution follows a normal law, we can consider that the area imprecision is about +/- $3\sigma_A$ (with a confidence of 99%) which means an area imprecision of +/- 2.4 ha (0.018% of the original area).

As a comparison, the same experiment performed with no constraints on the simulation of digitizing errors would have generated an imprecision of about +/- 4.8 ha (or 0.036% of the original area), which means an area imprecision multiplied per two. These results show that the use of a simple random digitizing error simulation process can generate important over-estimations of the measurement imprecision, and that the integration of constraints in the simulation process allows to limit such exaggerations.

CONCLUSION

As a conclusion, this paper presents original methods in order to simulate digitizing errors with a higher degree of realism, by integrating constraints modeling the operator input process. Observations show that classical random processes used to simulate digitizing errors can generate topological inconsistencies and non-respect of the original shape of objects, and these problems would have been avoided by a human operator. Based on a set of three assumptions about the operator input process, three methods are then proposed to integrate constraints in the simulation process. Results of the experiments show that the use of these constraints reduce the measurement imprecision generated by a classical random process, in addition with the suppression of potential topological inconstancies. However, several improvements need to be realized on this method, especially the question of the parameterization of error weightings (according to the angularity of successive vertices or the nodes' degrees). Further researches should be investigated in order to provide the most appropriate ways to parametrize the different constraints. In spite of these implementation problems, this approach proposes new materials in order to model digitizing error with more realism, and assess its impact on the positioning of geographical objects or on geometric measurements.

References

Bolstad, P. V., Gessler, P. and Lillesand, T. M. (1990). Positional uncertainty in manually digitized map data. *International Journal of Geographical Information Systems 4*(4), 399-412.

Bonin, O. (2002). *Modèles d'erreurs dans une base de données géographiques et grandes déviations pour des sommes pondérées; application à l'estimation d'erreurs sur un temps de parcours*. PhD Thesis, Paris 6 University, France.

De Bruin, S. (2008). Modelling positional uncertainty of line features by accounting for stochastic deviations from straight line segments. *Transactions in GIS 12*(2), 165-177

Girres, J.-F. (2012). *Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques. Application aux mesures de longueur et de surface*, PhD Thesis, Paris-Est University, France.

Girres, J.-F. (2015). Estimation of geographical databases capture scale based on inter-vertices distances exploration, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences 2(3)*, 305-310.

Goodchild, M., Shortridge, A. and Fohl P. (1999). Encapsulating simulation models with geospatial data sets. In Lowell, K. and Jaton, A. (eds) : *Spatial Accuracy Assessment : Land Information Uncertainty in Natural Resources*, Ann Arbor Press, 123-129.

Heuvelink, G. B. M., Brown, J. D. et van Loon, E. E. (2007). A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Systems 21*, 497-513

Hunter, G. J. et Goodchild, M. F. (1996). A new model for handling vector data uncertainty in geographic information systems. *URISA Journal 8*(1), 51-57.

Keefer, B., Smith, J. and Gregoire, T. (1988). Simulating manual digitizing error with statistical models. In *Proceedings of GIS/LIS'88 Conference*, 475-483.

Vauglin, F. (1997). *Modèles statistiques des imprécisions géométriques des objets géographiques linéaires*. PhD Thesis, Marne-La-Vallée University, France.

How far spatial accuracy governs land-use changes monitoring frequency: the urban sprawl monitoring example

Jean-Pierre Chéry*[1], Jean-Stéphane Bailly[2], Valérie Laurent[3], Nathalie Saint-Geours[4]

[1]AgroParisTech, TETIS, France
[2]AgroParisTech, LISAH, France
[2]Irstea, TETIS, France
[4]ITK, France

*Corresponding author: jean-pierre.chery@agroparistech.fr

Abstract

In this paper, we illustrate how far spatial accuracy of a land-use map governs land-use changes monitoring frequency on a urban sprawl monitoring case study. From a specific Monte Carlo approach propagating uncertainties, conf4idence curves for minimal monitoring frequency to detect significant changes in urban sprawl indicators were built. Results showed that frequency decreased when uspscaling indicators but it also showed very low monitoring frequency for indicators at the lower level.

INTRODUCTION

When setting up land-use monitoring systems, spatial uncertainties and their impact on the detection of indicator changes are usually ignored. To capture a significant change in land-use indicators is thus strongly related to its spatial resolution, the velocities of the process it represents and the accuracy of the used indicators. As a consequence, the required monitoring land-use change frequency, corresponding to the mimimum time step to ensure a significant change in indicators, also depends on these three factors: indicator spatial resolution, change process velocity and spatial indicator accuracy.

In this paper, we illustrate how far spatial accuracy of a map governs land-use changes monitoring frequency focusing on urban sprawl monitoring system from satellite imagery in southern France. Urban sprawl is a major challenge for land use planners: it causes the sealing of lands closest to the urban centers into impervious areas, thereby transforming highly productive cultivated soils, increasing flood hazards, fragmenting natural habitats, and raising complex issues related to transportation and social diversity. To define new urban policies, integrated monitoring of urban sprawl is crucial. It is therefore important to choose an appropriate monitoring frequency (Allen and Lu, 2003). Such monitoring relies on urban sprawl indicators which suffer from uncertainties related to impervious area mapping process or from the population census methodology. In most urban sprawl monitoring studies, however, these uncertainties effects are not considered, and the monitoring frequency is chosen solely based on data availability. In practice, time lag between successive indicator calculations may range from 5 years to 25 years, or even be irregular. As for monitoring biological diversity (Yoccoz et al., 2001), having a carefully designed urban sprawl monitoring system in terms of choice of indicator, scale and temporal frequency is essential.

To support land-use planning, this paper proposes a framework to infer indicator monitoring frequencies, taking into account indicator uncertainties and scaling. This framework was ap-

plied to 3 urban sprawl indicators at municipality, inter-municipality, department, region scales on the former Languedoc-Roussillon region, France.

MATERIAL AND METHODS

Study area

Languedoc-Roussillon is a former region of mainland France that covers 27,376 km^2 along the Mediterranean coast, and hosts about than 2.7 million inhabitants. Since the 1960s, intense population pressure in the region combined with poor urban planning led to population concentration in the major cities and their peripheral extensions in the coastal zone. Languedoc-Roussillon region contains 1,545 municipalities aggregated in 130 inter-municipalities which facilitate cooperation on practical urban management issues, and 5 departments. Therefore, four increasing spatial supports of urban sprawl indicator were used in the study: municipalities, inter-municipalities, departments, and entire region.

Data

Dataset of impervious polygons covers the entire study area were obtained for 1997 and 2009 from the Dupuy et al. (2012) methodology, using satelitte image (RapidEye imagery for 2009) classification with numerous additive manual post-classification at 1:8,000 scale to reduce confusions.

Population data for along years were taken from censuses of the French Institute of Statistics and Economic Studies (INSEE), which was performed at municipality scale. INSEE also provides uncertainty measures for their data (INSEE, 2012), as a coefficient of variation (CV) along 5 municipality categories (Tab. 1). For small municipalities having less than 10,000 inhabitants, there was exhaustive sampling and therefore, no uncertainty.

	CV median	**Number of municipalities less than 10,000**
0 to 10, 000 inhabitants	0	1517
10, 000 to 19, 999 inhabitants	1.02	17
20, 000 to 49, 999 inhabitants	0.78	6
50, 000 to 99, 999 inhabitants	0.56	2
More than 100,000 inhabitants	0.39	3

Table 1: Median of the coefficient of variation of the number of inhabitants per municipality category (INSEE, 2012), and number of municipalities of the study area in each category

Urban sprawl spatial indicators

Urban areas usually consist in namely Morphological Urban Area (MUA) obtained after morphological closing of a set of delineated impervious polygons. Closing consists in successive dilation and erosion (with a given radius r), allowing to merge impervious polygons. Based on the MUA, Balestrat et al. (2010) proposed a collection of spatial indicators of urban sprawl. For this study, we selected three indicators from this collection: the MUA area A (area-based), the dispersion coefficient D (shape-based) corresponding to the ratio between the cumulated areas of MUA polygons whose areas are smaller, respectively bigger than 3 hectares, and the population density P (mixing spatial and population data, i.e. number of inhabitants). Maps of these three indicators may be produced for each of the four spatial supports (municipality, inter-municipality, department, and region) (Laurent et al., 2014b).

Urban sprawl indicators uncertainty simulation

Three uncertainty sources in the computation of urban sprawl indicators A, D and P were considered: i) the impervious polygons, ii) the inhabitant number per municipality, and iii) the radius r for dilatation-erosion used to build the MUA. Geometric uncertainties for both 1997 and 2009 impervious polygon layers were explicitely simulated using the approach developped Laurent et al. (2014b). Prior to the simulation, metrics on impervious polygons geometry errors were first computed based on the comparison of image-processed dataset with a reference dataset on 75 sampled municipalities. These metrics quantify both the shape of impervious polygons using the position error of polygon vertices (specific spatial covariance models) and the rates of omission and commission errors of the impervious polygons within municipality areas. Based on these metrics, random realizations of impervious polygons datasets were simulated for each municipality. First, the position of polygon vertices was randomly perturbed to account for geometric uncertainties. Then, small impervious polygons were randomly added or removed to account for polygon omission/commission. For more details, the reader is referred to Laurent et al. (2014a). Figure 1 presents a few examples of simulated impervious polygon simulations and resulting MUA.

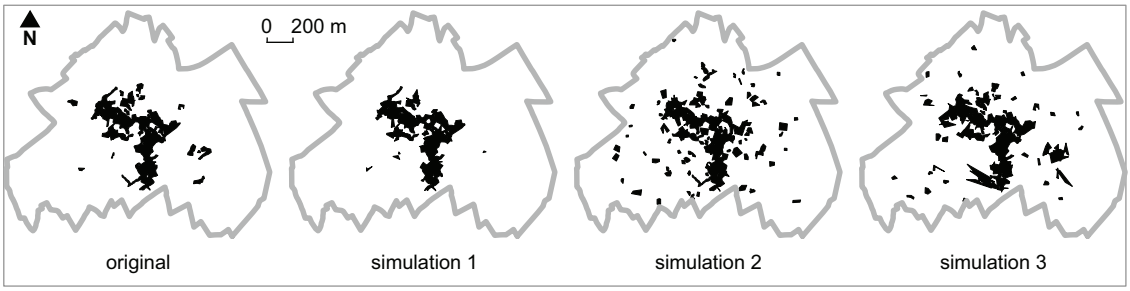


Figure 1: Original impervious polygons from the automatic dataset and three examples of impervious polygons simulations for the municipality of Jonquieres

For each municipality, the population was assumed to follow a Gaussian distribution $N(p, p * CV)$, where p is the municipalitys nominal population, and CV is the population coefficient of variation taken from Table1. Random realizations of population values were then drawn from $N(p, p * CV)$.

All sources of uncertainties were propagated through the urban sprawl indicator computation chain in a Monte Carlo framework. For each value of the radius r, a Monte Carlo experiment with N = 1,000 simulations was carried out while keeping total computational cost tractable.

Infering the right monitoring frequency under uncertainties

The "right" monitoring frequency is defined here as the minimal time lag Δ_t between two successive indicator calculations required to detect a significant change in the indicator. It depends both on indicator uncertainty and on the magnitude of the indicator change over time. Calculating Δ_t is thus considered as a problem of significant differences between two independent random variables. Let I_0 and I_1 be the indicator random variables at times t_0 and t_1 respectively, following independent Gaussian distributions $N(i_0, \sigma)$ and $N(i_1, \sigma)$. i_0 and i_1 are the indicator mean values (nominal values) at t_0 and t_1 and σ is the indicator standard deviation, assumed to be constant over time and infered from previous section 2.4. Assuming both unbiased indicator

means at a given time and a linear indicator trend with slope s, the random variable $\Delta_T = \frac{I_1 - I_0}{s}$, can be infered from the lowest significant difference $(I_1 - I_0)$ at confidence level $1 - \alpha$ as:

$$\Delta_t = \frac{q2\sigma}{s} \tag{1}$$

where q denotes the quantile of the standard normal distribution. We can then build confidence curves for Δ_t, as a function of s, for the usual confidence levels 90, 95, and 99% for instance. To choose the range of s values for the Languedoc-Roussillon region (s_{LR}), we calculated the average indicator slope over the study area between 1997 and 2009, assuming similar geometric and thematic uncertainties within the 1997 and 2009 impervious polygons datasets.

RESULTS

Urban sprawl uncertainties simulation

The nominal data (no uncertainty) and the simulation results at the municipality scale with 50 m radius r are presented in maps on figure 2. The nominal indicator maps (line 1 - Fig. 2) show the spatial patterns of the indicators. A has higher values along the coast and other plains than in the rugged inland areas. Conversely, D (dispersion coefficient) shows an opposite pattern, with higher values in the rugged inland areas, and lower values in the plains. The P (population density) map shows mostly low values with few municipalities located mostly in rugged areas having higher values Comparing the simulated indicator maps to the nominal maps, one notices that, for all three indicators, the simulated spatial patterns is similar to the nominal ones. This good match can also be seen on the scatterplots, where most points are located close to the $y = x$ line. For A, however, the simulated values were much higher than nominal values for municipalities having A smaller than 50,000 m2. This threshold of 50,000 m2 corresponds to a rectangle MUA of 500 * 100 m, which would be an unrealistically small village. We therefore infer that the nominal map underestimates A for those municipalities. This underestimation of A by the nominal maps may be due either to the presence of clouds in the satellite images, or to the omission of (parts of) polygons in the image classification. Both D and P present a general trend to overestimation. The CV (coefficient of variation) maps contain the uncertainty information: the higher the CV value, the higher the uncertainty. The three CV maps show similar spatial trends, with lower uncertainty along the coast and in other plains, and higher uncertainties in the rugged areas. The median CV value increases from A (25 %), to P (31 %), to D (58 %), meaning that uncertainty increases from A, to P, to D. The maximum CV values were very high (2.09 for A, 6.66 for D, and 14.57 for P), indicating locally very high uncertainties for some municipalities.

For convenience, results for other radius values are not detailly shown. Indeed, for large radius r values, more polygons are joined into the MUA, increasing A and, thereby decreasing D and P. For all three indicators, the median CV values slightly increase with increasing radius, indicating that increasing the radius slightly increases the indicator uncertainty.

The impact of indicators spatial upscaling is shown on the boxplots of figure 3. As expected, A increases with increasing entity size, because the MUA areas are being summed over larger entities. On the contrary, D decreases with increasing entity size. The entity size does not impact the P median. As expected (Heuvelink, 1998; Saint-Geours et al., 2014), for all three variables, uncertainty decreases (decreasing CV values) when upscaling. For A and D, the CV

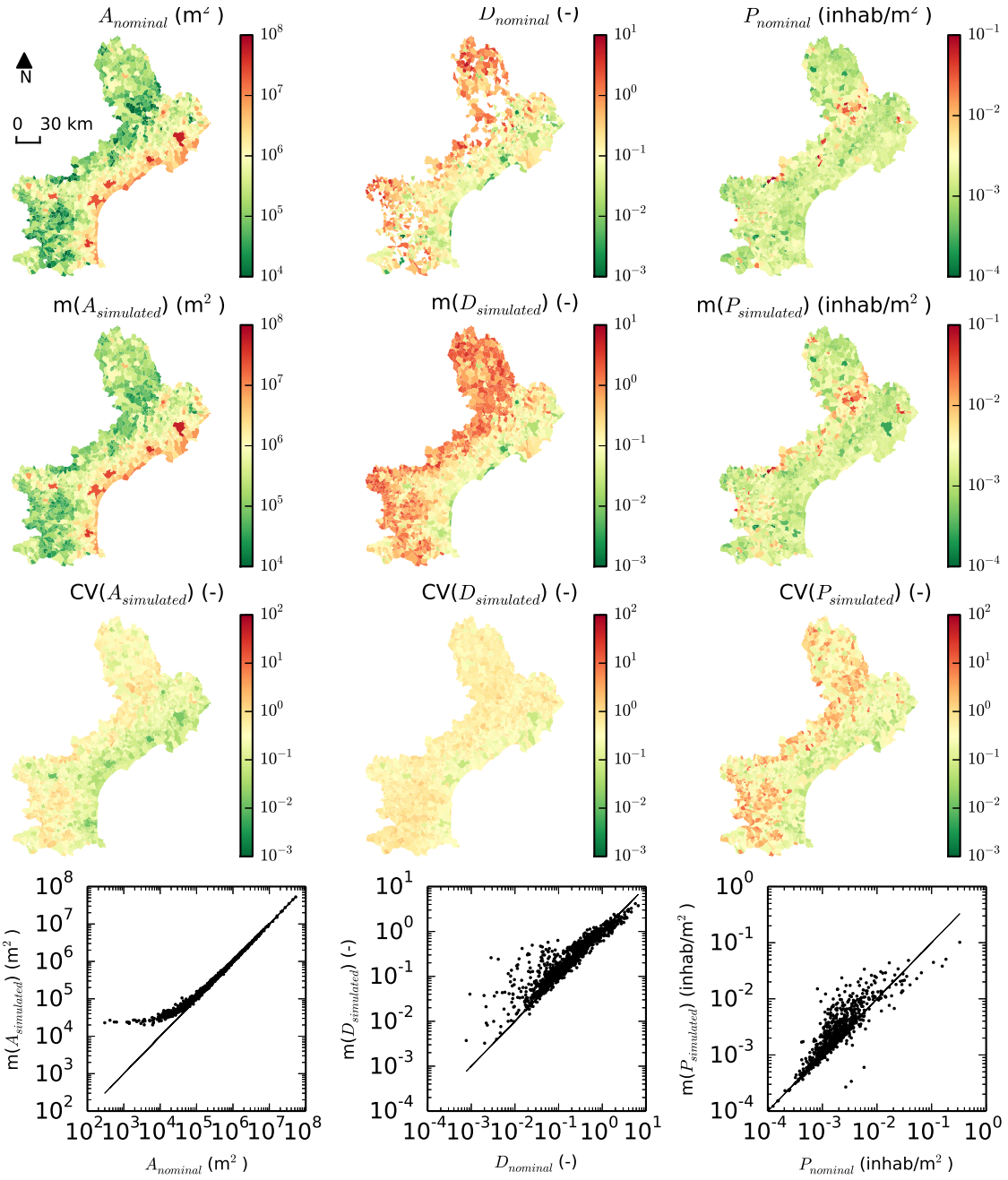


Figure 2: Results of Monte Carlo simulations, compared to the nominal results at municipality scale. The three columns present the results for the A, D, and P indicators. Line 1 shows the nominal maps, lines 2 and 3 the maps of the mean and CV respectively over the 1,000 simulations, and line 4 the scatterplots of the simulated versus nominal indicator values

values are divided by a factor 3 when upscaling from municipality to inter-municipality, by a factor 7 from inter-municipality to department, and by a factor 2 from department to region. For P, the factors are smaller: 3, 1.4, and 1.7, in the same upscaling order as for A and D.

Right monitoring frequencies

The confidence curves for the minimal monitoring time lag Δ_t, as a function of the slope s of the urban sprawl indicator trends, are presented in figure 4. Looking at the A plot at municipality scale, we can see that, for all three confidence levels, the curves present a hyperbolic decreasing

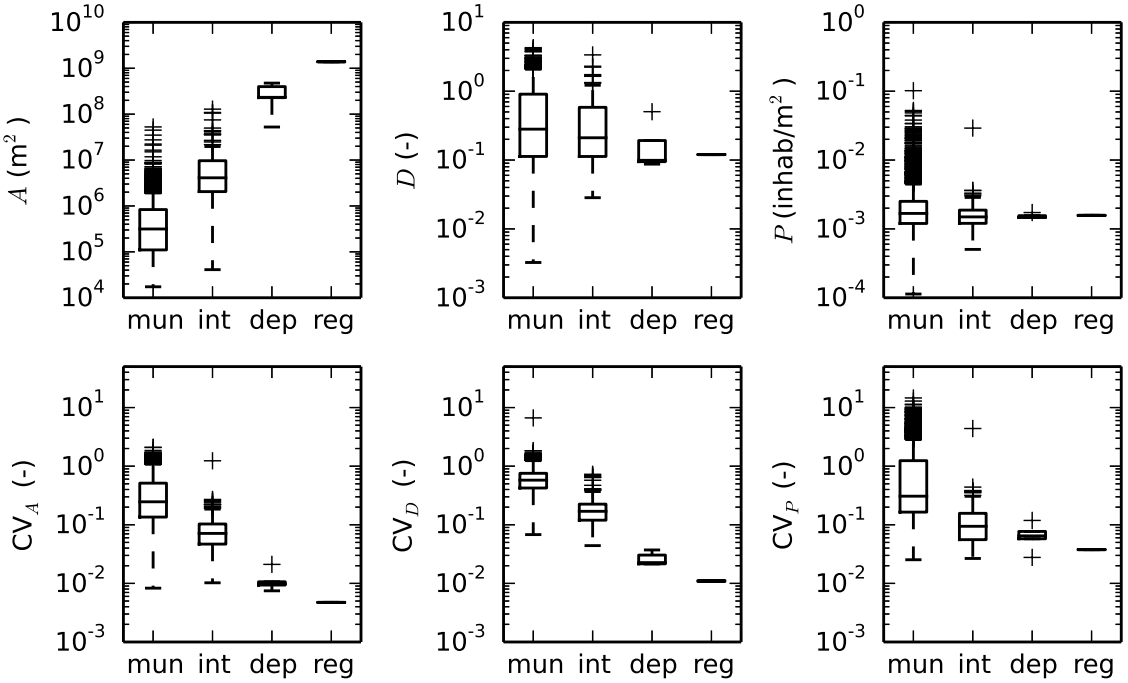


Figure 3: Influence of the aggregation scale on the indicator mean values and coefficients of variation (CV). The sample sizes are: 1,545 for the municipality scale (mun), 130 for the inter-municipality scale (int), 5 for the department scale (dep), and 1 for the region scale (reg).

shape, with Δ_t values starting at about 1,000 years for s close to zero, and decrease to about 10 years for s = 15% per year. This reflects the fact that smaller changes take longer than bigger changes to be detected. The curve for 99% confidence level shows higher Δ_t values than the 95%, followed by the 90% confidence level curves, reflecting the fact that to increase confidence that the detected change is a true change, and not due to uncertainties, more time is required. At municipality scale, the curves for the three indicators have similar magnitude, decreasing from 1,000 to 10 years. The magnitude of the curves decreases with upscaling, down to decreasing from 100 to 0.1 - 1. This is because the uncertainties decrease with upscaling, therefore translating into smaller Δ_t values. The bigger the aggregation scale, the more A, P and D drift apart in terms of Δ_t values at high absolute slope values: a significant change in A can be detected faster than a change in D, than a change in P. For all plots, even with a logarithmic y-axis, the confidence curves present a hyperbolic shape, indicating the highly non-linear relationship between s and Δ_t. Looking at the particular case of Languedoc-Roussillon for 1997-2009, the slope values are smallest for P, followed by A, and largest for D. The corresponding Δ_t are therefore ranked in the opposite order, with P having the highest Δ_t, followed by A, and D having the smallest Δ_t (value ranges indicated on Figure 4). One should note that, because of the logarithmic y-axis scale, the roughly constant vertical space between the confidence curves means that the difference in Δ_t in years is actually much higher for high Δ_t values than for lower Δ_t values. Because of averaging, the slope values decrease when increasing the aggregation scale. Because uncertainties decrease when upscaling, Δ_t also decreases when upscaling. The 12-year time lag between the 1997 and 2009 automatic impervious polygons datasets is sufficient to detect significant differences in A and D at region and department scales, and D at inter-municipality scale, but is not sufficient for P.

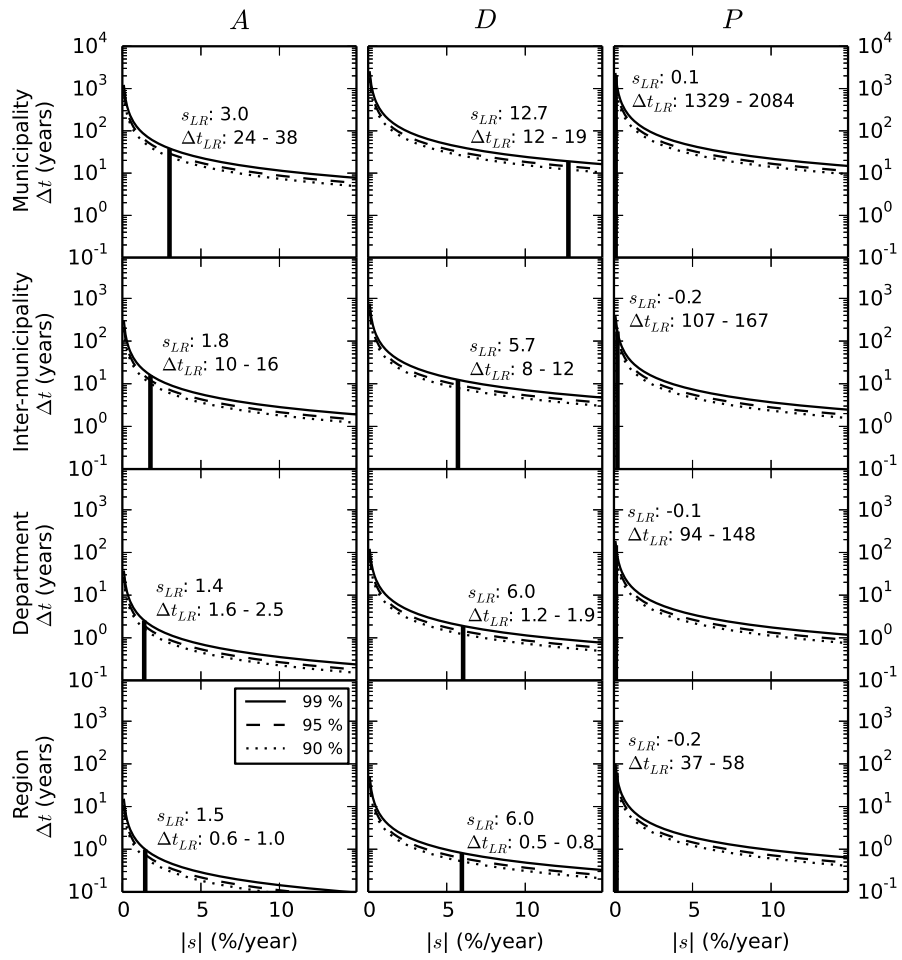


Figure 4: Confidence curves at 90, 95, an 99% confidence levels for the minimal monitoring time lag Δ_t as a function of the absolute value of the relative slope of the indicator changes s. The slope (s_{LR}) and time lag (Δt_{LR}) values at 90 and 99% confidence levels for the Languedoc-Roussillon case study are indicated on the plots in the axes units

DISCUSSION

As revealed by figure 4, minimum monitoring time lags vary strongly depending on the considered indicator and scale. These results are based on the average level of uncertainties for the entire Languedoc-Roussillon, whereas the uncertainties may strongly vary locally. Another caveat is that constant slope over time was assumed to build the confidence curves, and they should therefore be used with caution for inferring prospective time lag values. They are not adapted to non-linear cases, such as instantaneous abrupt growth of urban areas. Despite these limitations, we can attempt to compare the order of magnitude of the monitoring frequencies recommended by figure 4 to operational urban monitoring systems. In Europe, the monitoring time lag of the CORINE land cover datasets, decreased from 10 to 6 years. In the United States of America, the National Land Cover Database is updated every 5 years. These two systems are therefore appropriate for monitoring urban areas at scales as fine as the French department scale, but not inter-municipalities and municipalities. The theoretical monitoring time lag, as calculated from equation 1 is, of course, very important for establishing a new operational urban monitoring system, but other factors come into play. First, the theoretical time lag is different for each indicator and varies depending on the considered administrative scale. In this case study, the level of indicator uncertainty of the D indicator is simply too high to detect slight

changes. This is reinforced by the cyclic behavior of D (succession of dispersion and absorption of impervious polygons in the MUA), violating the linear trend assumption in the analysis. In this case, a maximum monitoring time lag has to be included to allow detecting intra-cycle changes. The time lag required for obtaining a new impervious polygon dataset is also important. Vector databases are typically updated less often than new imagery is acquired, although parts of the study area might be missing from the imagery. In addition, any governments available budget and number of qualified employees may also cause practical limitations. Finally, trade-offs are required.

CONCLUSION

A specific spatially explicit Monte Carlo simulation of urban sprawl indicator was developped according to the impervious areas geometric and thematic uncertainties, socio-cultural data uncertainties and closing operation uncertianty required to obtain urban areas. This simulation process allowed to propagate and scaling uncertainties up to three urban sprawl indicators calculation: MUA area (A), MUA dispersion coefficient (D), and population density (P), for the entire Languedoc-Roussillon region, France. As expected, upscaling the indicators to larger entities decreased their uncertainties. Confidence curves for minimal monitoring frequency to detect significant changes in indicator values showed that frequency decreased, for all four administrative levels. The highest frequency values were observed for the coarsest administrative scale (region), with values between 0.5 and 1 year for D and A, and 37 to 58 years for P.

ACKNOWLEDGEMENTS

This work was funded by the French Research Agency in the framework of the program Investissements d'Avenir through the support of the GEOSUD project (ANR-10-EQPX-20).

References

Allen J., Lu K. (2003). Modeling and prediction of future urban growth in the charleston region of south carolina: a gis-based integrated approach. *Ecology and Society 8*(2), 2.

Balestrat M., Chery J., Tonneau J. (2010). Construction of spatial indicators for decision making: interest in a participatory approach: the case of peri-urban languedoc. In *Proceedings of a symposium on Innovation and Sustainable Development in Agriculture and Food, Montpellier, France, 28 June to 1st July 2010.*, pp. hal–00539776. Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD).

Dupuy S., Barbe E., Balestrat M. (2012). An object-based image analysis method for monitoring land conversion by artificial sprawl use of rapideye and irs data. *Remote Sensing 4*(2), 404–423.

Heuvelink G. B. (1998). Uncertainty analysis in environmental modelling under a change of spatial scale. *Nutrient cycling in Agroecosystems 50*(1-3), 255–264.

INSEE (2012). Recensement de la population: La prcision du chiffre de population dans les grandes communes de mtropole. Technical report, Institut National de la Statistique et des Etudes Economiques.

Laurent V., Saint-Geours N., Bailly J.-S., Chery J.-P. (2014a, July 8-11). Local urban sprawl accuracy from image segmentation uncertainties simulation. In A. Shortridge, J. Messina, S. Kravchenko, and A. Finley (Eds.), *Accuracy 2014, 11th international Symposium on Spatial Accuracy Assessment in Natural Ressources and Environmental Sciences*, ISARA, East Lansing, Michigan, USA, pp. 146–150.

Laurent V. C. E., Saint-Geours N., Bailly J.-S., Chery J.-P. (2014b, 21-23 may 2014). Simulating geometric uncertainties of impervious areas based on image segmentation accuracy metrics. *South-Eastern European Journal of Earth Observation and Geomatics 3*(2), 37–40. Special Issue: 5th GEOBIA.

Saint-Geours N., Bailly J.-S., Grelot F. (2014). Multi-scale spatial sensitivity analysis: application to a flood damage assessment model. *Environmental Modelling & Software 60*, 153–166.

Yoccoz N. G., Nichols J. D., Boulinier T. (2001). Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution 16*(8), 446–453.

‘spup’ – a R package for uncertainty propagation in spatial environmental modelling

K. Sawicka*, G.B.M. Heuvelink

Soil Geography and Landscape Group, Wageningen University, PO Box 47, 6700 AA, Wageningen, The Netherlands

*Corresponding author: kasia.sawicka@wur.nl

Abstract

Computer models are crucial tools in engineering and environmental sciences for simulating the behaviour of complex systems. While many models are deterministic, the uncertainty in their predictions needs to be estimated before they are used for decision support. Advances in uncertainty analysis have been paralleled by a growing number of software tools, but none has gained recognition for universal applicability, including case studies with spatial models and spatial model inputs. We develop an R package that facilitates uncertainty propagation analysis in spatial environmental modelling. The ‘spup’ package includes functions for uncertainty model specification, propagation of uncertainty using Monte Carlo (MC) techniques, and uncertainty visualization functions. Uncertain variables are represented as objects which uncertainty is described by probability distributions. Spatial auto-correlation within a variable and cross-correlation between variables is also accommodated for. The package has implemented the MC approach with efficient sampling algorithms, i.e. stratified random sampling and Latin hypercube sampling. The MC realizations may be used as an input to the environmental models called from R, or externally. Selected static and interactive visualization methods that are understandable by non-statisticians can be used to visualize uncertainty about the measured input, model parameters and output of the uncertainty propagation.

Key words

R language, uncertainty analysis, uncertainty propagation, spatial models, Monte Carlo

I INTRODUCTION

Computer models have become a crucial tool in engineering and environmental sciences for simulating the behaviour of complex static and dynamic systems. However, while many models are deterministic, the uncertainty in their predictions needs to be estimated before they are used for decision support. Currently, advances in uncertainty propagation and assessment have been paralleled by a growing number of software tools for uncertainty analysis, but none has gained recognition for a universal applicability, including case studies with spatial models and spatial model inputs. Due to the growing popularity and applicability of the open source R programming language we undertook a project to develop an R package that facilitates uncertainty propagation analysis in spatial environmental modelling. The tool is intended for researchers and practitioners who understand the problems of uncertainty in data and models, and are looking for a simple, accessible implementation of the universal

methodology for uncertainty assessment. At the same time, it is designed to enable more experienced users to easily understand, customise, and possibly further develop the code.

A number of computational tools are readily available to tackle the uncertainty quantification problem to different degrees. These include both free software, like OpenTURNS (Andrianov et al., 2007), DACOTA (Adams et al., 2009) and DUE (Brown and Heuvelink, 2007), commercial, like COSSAN (Schuëller and Pradlwarter, 2006), or free, but written for a licenced software, e.g. SAFE (Pianosi et al., 2015) or UQLab (Marelli and Sudret, 2014) toolboxes for MATLAB. A broad review of existing software packages is available in Bastin et al. (2013). To the best of our knowledge, however, none of the existent software is specifically designed to be extended by the environmental science community. The use of powerful but complex languages like C++ (e.g. Dakota), Python (e.g. OpenTURNS) or Java (e.g. DUE) often discourages relevant portions of the non-highly-IT trained scientific community from the adoption of otherwise powerful tools.

The R programming language is an important tool for development in numerical and statistical analysis. R has advantages through its advanced statistical capabilities and high-quality graphical output (Ripley, 2001), and is gaining widespread use in science and education. Furthermore, through the use of R packages, the software can be used for a variety of geoscience analyses and visualisations. It has grown tremendously over the last 20 years, with over 8000 packages at the time of preparation of this paper. There is a number of R packages invoking uncertainty analysis through sensitivity analysis or use of a Bayesian framework for model calibration. We have found only one package named ‘propagate’ that deals with uncertainty propagation explicitly, using similar approaches as described in this paper. The package ‘propagate’, however, does not provide functionality for spatial models and variables.

II EMPLOYED (SPATIAL) UNCERTAINTY PROPAGATION ANALYSIS APPROACH

Uncertainty propagation aims to analyse how uncertainties in data (e.g. from measurement error, sampling, interpolation), combined with model uncertainties (e.g. in the model parameters and structure) propagate through the model (Heuvelink et al., 2007). Many environmental phenomena of interest are spatial, temporal or spatio-temporal in nature and often have strong correlations imposed by the physics and dynamics of the natural systems, making uncertainty evaluation difficult. The most frequently used approach represents uncertainty with probability distribution functions (pdfs). The pdf describes the relative likelihood for the random variable to take on a given value and typically it is viewed as a shape of the distribution, for example normal, uniform, lognormal or exponential. It is common for the pdf to be parametrized, i.e. to be characterized by distribution parameters. For example, the normal distribution is parametrized in terms of the mean and the variance, or uniform distribution is parametrized by minimum and maximum values. For situations in which pdfs can be estimated reliably, they have a number of advantages over non-probabilistic techniques. They include methods for describing cross- and auto- correlation between uncertainties, methods for propagating uncertainties through simple algebras or more complex environmental

models, and methods for tracing the sources of uncertainty in environmental data and models (Heuvelink 1998).

A frequently used method for the analysis of uncertainty propagation is the Monte Carlo (MC) method (Hammersley and Handscomb, 1979, Lewis and Orav, 1989). It is very flexible and can reach an arbitrary level of accuracy, and therefore it is generally preferred over analytical methods such as the Taylor series method (Heuvelink, 1998). The idea of the MC method is to compute the output of the model repeatedly, with input values that are randomly sampled from their marginal or joint pdf. The set of model outputs forms a random sample from the output pdf, so that the parameters of the distribution, such as the mean, variance and quantiles, can be estimated from the sample. The method thus consists of the following steps:

1. Characterise uncertain model inputs with pdfs.
2. Repeatedly sample from (spatial) pdfs of uncertain inputs.
3. Run model with sampled inputs and store model outputs.
4. Compute summary statistics of model outputs.

Note that the above ignores uncertainty in model parameters and model structure, but these can easily be included if available as pdfs. A random sample from the model inputs can be obtained using an appropriate pseudo-random number generator (Lewis and Orav, 1989). Note that a conditioning step will have to be included when the model inputs are correlated. Application of the MC method to uncertainty propagation with operations that involve spatial interactions requires the simultaneous generation of realisations from the spatially distributed inputs implying that spatial correlation will have to be accounted for (Heuvelink et al., 1989). For uncertain spatially distributed continuous variables, such as elevation, rainfall and soil organic carbon content, we assume the following geostatistical model:

$$Z(x) = \mu(x) + \sigma(x) \cdot \varepsilon(x) \quad (1)$$

where μ is the (deterministic) mean of Z, σ is a spatially variable standard deviation of the prediction of μ (spatial variability of σ reflects that in some parts of study area the uncertainty is greater than in other parts), and ε is a standardized, zero-mean, spatially auto-correlated residual modelled with a semivariogram or a correlogram (Diggle and Ribeiro, 2007, Webster and Oliver, 2007, Plant, 2012). The random sample is drawn from the pdf of ε to further calculate the realizations of Z.

The drawback of the MC method is that the accuracy of the uncertainty assessment is inversely related to the square root of the number of runs N. This means that to double the accuracy, four times as many runs are needed. In complex, multi-variable systems high accuracies are obtained only when the number of runs is very large (i.e. $N \geq 1{,}000$), which may cause the method to become extremely time consuming. The improvement on MC efficiency can be made by employing efficient sampling techniques (e.g. Latin hypercube sampling) and parallel computing.

III 'spup' (SPATIAL UNCERTAINTY PROPAGATION) PACKAGE DESIGN

The adopted approach for uncertainty propagation analysis dictates the general package design. The 'spup' package provides functions for examining the uncertainty propagation starting from input data and model parameters, via the environmental model onto model outputs (Fig. 1). The functions include uncertainty model specification, stochastic simulation and propagation of uncertainty using MC techniques, as well as several uncertainty visualization functions.

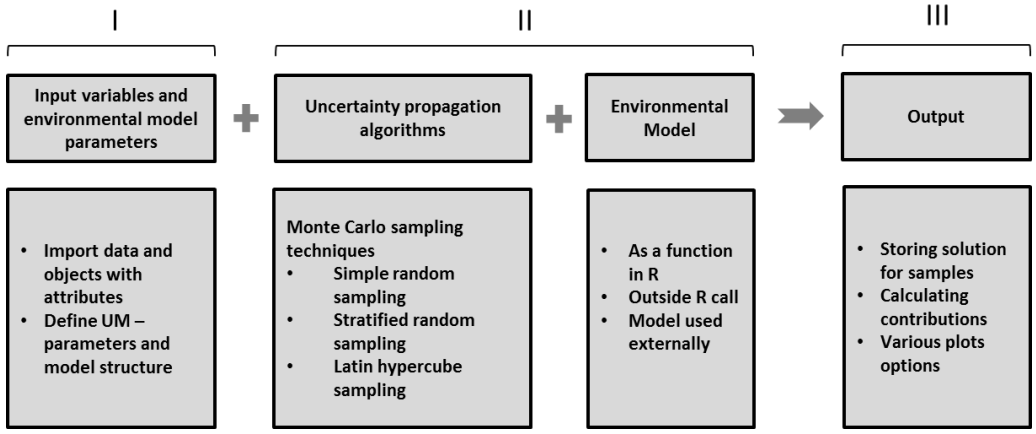


Figure 1 . The 'spup' package design. 'spup' comprises of functions for defining uncertainty model (I), quantifying uncertainty propagation (II) and storing output in a format of data or images.

Uncertain environmental variables are represented in the package as objects whose attribute values may be uncertain and described by probability distributions. Uncertainty assumption may also be ignored, in which case, during the model run the user works with μ (Eq. 1) as the model input that best represent the reality. Both numerical (e.g. air humidity) and categorical data (e.g. land cover) types are handled. Spatial auto-correlation within an attribute and cross-correlation between attributes is also accommodated for. The attributes may be independent in space, for which a marginal probability density function (mpdf) is defined at each point in space, or may co-vary in space, for which a joint probability density function (jpdf) is defined. Different shapes of marginal pdfs are supported, whereas joint pdfs may be defined for groups of attributes characterized with the normal distribution only. The specification of correlations between errors in space and cross-correlations between objects or attributes is made under the assumption that the correlations depend only on the distance between locations.

For spatially correlated variables the package relies on the unconditional Gaussian simulation implemented in the 'gstat' package (Pebesma, 2004). For drawing realizations of uncertain variables without assumed correlations the package has implemented the MC approach with efficient sampling algorithms, i.e. stratified random sampling and Latin hypercube sampling. The design includes facilitation of parallel computing to speed up MC computation. The MC realizations for uncertainty propagation quantification may be used as an input to the environmental models called from R, or externally.

Selected static (adjacent maps and glyphs) and interactive visualization methods that are understandable by non-experts with limited background in statistics can be used to summarize and visualize uncertainty about the measured input, model parameters and output of the uncertainty propagation.

IV APPLICATION EXAMPLE – MAPPING SOIL MOISTURE CONTENT FOR THE ALLIER CATCHMENT

As part of a research study in quantitative land evaluation, the World Food Studies (WOFOST) crop simulation model (van Diepen et al., 1989) was used to calculate potential crop yields for floodplain soils of the Allier river in the Limagne rift valley, central France. The moisture content at wilting point (Θ_{wp}) is an important input attribute for the WOFOST model. Because Θ_{wp} varies considerably over the area in a way that is not linked directly with soil type, it was necessary to map its variation separately to see how moisture limitations affect the calculated crop yield.

Unfortunately, because Θ_{wp} must be measured on samples in the laboratory, it is expensive and time-consuming to determine it for a sufficiently large number of data points for creating the prediction map by kriging. An alternative and cheaper way is to calculate Θ_{wp} from other indicators which are cheaper to measure. Because the moisture content at wilting point is often strongly correlated with the moisture content at field capacity (Θ_{fc}) and the soil porosity (Φ), both of which can be measured more easily, it was decided to investigate how errors in measuring and mapping these would work through to a map of calculated Θ_{wp}. Calculation of Θ_{wp} can be done using a pedo-transfer function, which in this case takes the form of multiple linear regression:

$$\Theta'_{wp} = \beta_0 + \beta_1 \cdot \Theta_{fc} + \beta_2 \cdot \Phi + \delta \qquad (2)$$

where Θ'_{wp} denotes measured moisture content at wilting point, β_0 , β_1 , and β_2 are the regression coefficients and δ denote residuals attributed to lack of model fit and measurement error. The regression coefficients were estimated using standard ordinary least squares regression, ignoring spatial correlation between the observations at the locations. The maps of Θ_{fc} and Φ were derived using co-kriging and accounted for spatial cross-correlation between Θ_{fc} and Φ. Each component on the right hand side of Eq. 2 is subject to uncertainty, which will propagate to uncertainty about Θ_{wp}. Following the adopted MC approach, for each variable and parameter the uncertainty model is defined and 1000 MC samples are drawn. For the spatial variables a linear model of co-regionalization (Wackernagel, 2003) is fitted with use of the 'gstat' package and possible realities are simulated. The joint pdf of the model parameters and structural error δ was estimated using Bayesian calibration (Van Oijen et al., 2005) (note, this is not included in the 'spup' package) and a random sample was drawn from their joint posterior distribution. 1000 realizations of Θ_{wp} was then calculated using Eq. 2 and summary statistics such as mean of prediction and standard deviation were derived.

If an uncertainty analysis with WOFOST would show that the errors in Θ_{wp} cause errors in the output of WOFOST that are unacceptably large, then the accuracy of the map of Θ_{wp}

would have to be improved. In order to decide how to proceed in such a situation, the contribution of each individual error source to the overall uncertainty in Θ_{wp} was determined as well. Figure 2 presents the results and these show that both Θ_{fc} and Φ, rather than model parameters and model structural error, form the main source of uncertainty. Thus, the main source of error in Θ_{wp} is the one associated with the kriging errors of Θ_{fc} and Φ.

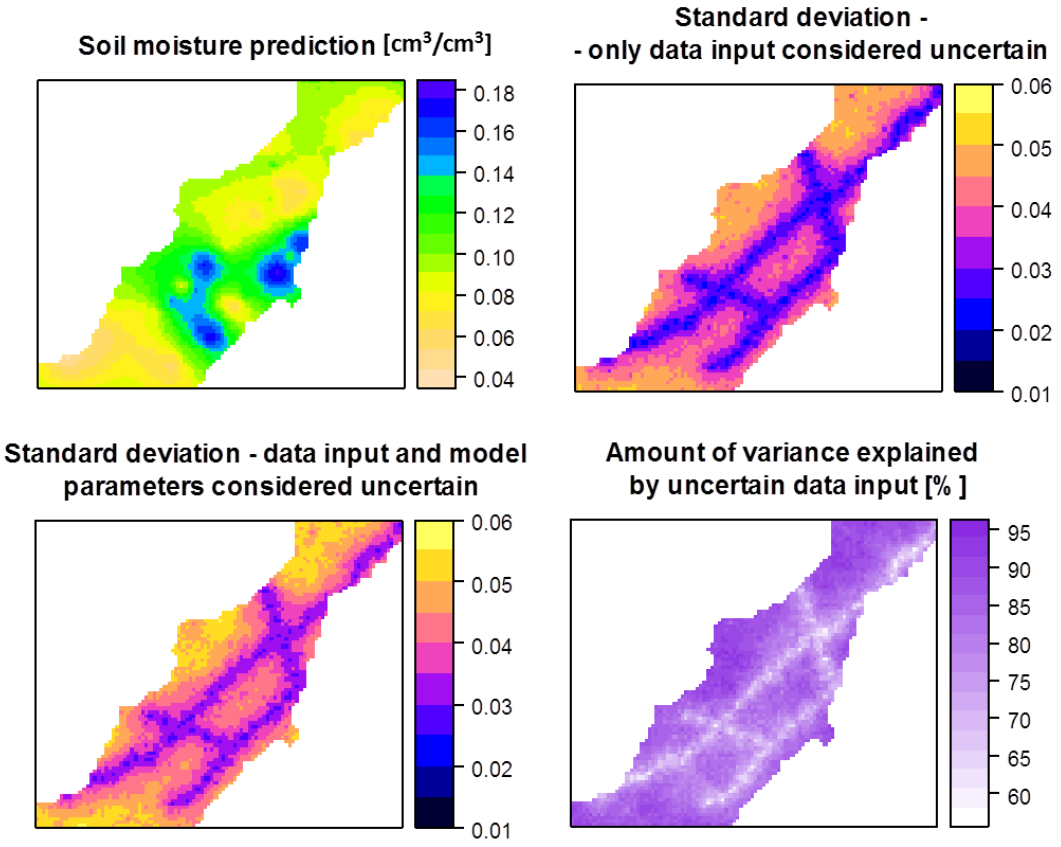


Figure 2 Results of uncertainty propagation for soil moisture prediction in the Allier catchment.

V CONCLUSIONS AND FURTHER WORK

We present a tool for uncertainty propagation assessment based on the uncertainty quantification framework described in e.g. Heuvelink et al. (2007). As the theoretical framework and implementation of the package progress, its application to real cases will be necessary, both to test the algorithms and usability of the tool, and to demonstrate the importance of assessing uncertainty in environmental data. The 'spup' package is being developed and used within the project "Quantifying Uncertainty in Integrated Catchment Studies (QUICS)". QUICS aims to carry out research in order to take the implementation of the Water Framework Directive (WFD) to the next level and improve water quality management by assessing the uncertainty of integrated catchment model water quality predictions. Currently, the potential case studies for the 'spup' application include uncertainty propagation analysis

with the LandscapeDNDC model (Haas et al., 2012) and German Schwingbach catchment data, and Metaldehyde Prediction model developed currently for the Severn Tent Water, water provider in the Midlands, UK. Finally, 'spup' will be introduced to the wider scientific community through CRAN (The Comprehensive R Archive Network), where many more challenges will be faced, including the time and resources required to implement an uncertainty assessment and the need to make uncertainty analyses understandable to non-statisticians.

References

ADAMS, B. M., BAUMAN, L. E., BOHNHOFF, W. J., DALBEY, K. R., EBEIDA, M. S., EDDY, J. P., ELDRED, M. S., HOUGH, P. D., HU, K. T., JAKEMAN, J. D., SWILER, L. P. & VIGIL, D. M. 2009. DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 5.4 User's Manual.

ANDRIANOV, G., BURRIEL, S., CAMBIER, S., DUTFOY, A., DUTKA-MALEN, I., DE ROCQUIGNY, E., SUDRET, B., BENJAMIN, P., LEBRUN, R., MANGEANT, F. & PENDOLA, M. OpenTURNS, an open source initiative to Treat Uncertainties, Risks'N Statistics in 520 a structured industrial approach. ESREL'2007 Safety and Reliability Conference, 2007 Stavenger, Norway.

BASTIN, L., CORNFORD, D., JONES, R., HEUVELINK, G. B. M., PEBESMA, E., STASCH, C., NATIVI, S., MAZZETTI, P. & WILLIAMS, M. 2013. Managing uncertainty in integrated environmental modelling: The UncertWeb framework. *Environmental Modelling & Software,* 39**,** 116-134.

BROWN, J. D. & HEUVELINK, G. B. M. 2007. The Data Uncertainty Engine (DUE): A software tool for assessing and simulating uncertain environmental variables. *Computers & Geosciences,* 33**,** 172-190.

DIGGLE, P. & RIBEIRO, P. J. 2007. *Model-based geostatistics*, Springer.

HAAS, E., KLATT, S., FRÖHLICH, A., KRAFT, P., WERNER, C., KIESE, R., GROTE, R., BREUER, L. & BUTTERBACH-BAHL, K. 2012. LandscapeDNDC: a process model for simulation of biosphere–atmosphere–hydrosphere exchange processes at site and regional scale. *Landscape Ecology,* 28**,** 615-636.

HAMMERSLEY, J. M. & HANDSCOMB, D. C. 1979. *Monte Carlo methods,* London, Chapman and Hall.

HEUVELINK, G. B. 1998. *Error Propagation in Environmental Modelling with GIS,* London, Taylor and Francis.

HEUVELINK, G. B. M., BROWN, J. D. & VAN LOON, E. E. 2007. A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science,* 21**,** 497-513.

HEUVELINK, G. B. M., BURROUGH, P. A. & STEIN, A. 1989. Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems,* 3**,** 303-322.

LEWIS, P. A. W. & ORAV, E. J. 1989. *Simulation methodology for statisticians, operations analysts, and engineers* Pacific Grove, Wadsworth & Brooks/Cole.

MARELLI, S. & SUDRET, B. 2014. UQLab: A Framework for Uncertainty Quantification in Matlab. *Vulnerability, Uncertainty, and Risk.* American Society of Civil Engineers.

PEBESMA, E. J. 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences,* 30**,** 683-691.

PIANOSI, F., SARRAZIN, F. & WAGENER, T. 2015. A Matlab toolbox for Global Sensitivity Analysis. *Environmental Modelling & Software,* 70**,** 80-85.

PLANT, R. E. 2012. *Spatial data analysis in ecology and agriculture using R*, CRC Press.

RIPLEY, B. D. 2001. The R project in statistical computing. *MSOR Connections. The newsletter of the LTSN Maths, Stats \textbackslash{\&} OR Network***,** 23--25.

SCHUËLLER, G. I. & PRADLWARTER, H. J. 2006. Computational stochastic structural analysis (COSSAN) – a software tool. *Structural Safety,* 28**,** 68-82.

VAN DIEPEN, C. A., WOLF, J., VAN KEULEN, H. & RAPPOLDT, C. 1989. WOFOST: a simulation model of crop production. *Soil Use and Management,* 5**,** 16-24.

VAN OIJEN, M., ROUGIER, J. & SMITH, R. 2005. Bayesian calibration of process-based forest models: bridging the gap between models and data. *Tree Physiol,* 25**,** 915-27.

WACKERNAGEL, H. 2003. *Multivariate Geostatistics: An Introduction with Applications*, Springer.

WEBSTER, R. & OLIVER, M. A. 2007. *Geostatistics for environmental scientists*, John Wiley & Sons.

Combining spatial and thematic uncertainty and sensitivity analysis for mountain natural hazard assessment

Jean-Marc Tacnet[*1], Guillaume Dupouy[1], Franck Bourrier[2], Frédéric Berger[2], Dominique Laigle[1], Nicolas Crimier[3], Kamel Mekhnacha[3], Michel Mémier[4]

[1] Université Grenoble Alpes, Irstea, UR ETGR, Snow avalanches Engineering and Torrent Control Research Unit – 2 rue de la papeterie, BP 76, F-38402 Saint-Martin-d'Hères Cedex, France.
[2] Université Grenoble Alpes, Irstea, UR EMGR, Mountain Ecosystems Research Unit – 2 rue de la papeterie, BP 76, F-38402 Saint-Martin-d'Hères Cedex, France.
[3] Probayes – 82 allée Galilée, F-38330 Montbonnot, France.
[4] Sintegra – 11 chemin des Prés, 38241 Meylan, France.

*Corresponding author: jean-marc.tacnet@irstea.fr

Abtract
This article presents an application of the Hybrid uncertainty propagation method to natural hazard assessment. It proposes a comparison with the usual probabilistic Monte-Carlo method, and propagates uncertainties related to both spatial and thematic variables.

Keywords
Uncertainty, sensitivity, Monte-Carlo, Hybrid.

I CONTEXT

Rockfalls and debris-flows are dangerous phenomena in mountains that cause severe damage to exposed assets and population (Figure 1). Risk assessment is based on both thematic and spatial information: physical phenomenon features such as height, speed, impact loads but also the extent of the phenomena is essential to be assessed (Figure 2).

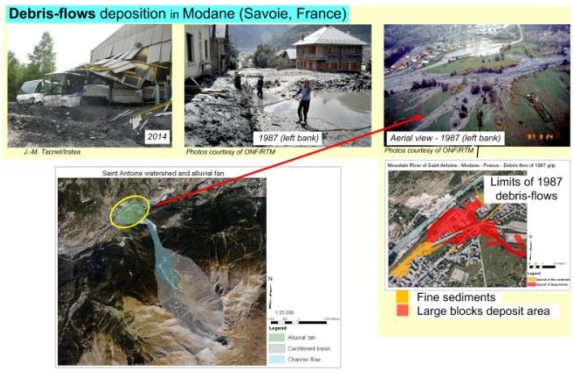


Figure 1: Examples of debris-flows effects and extents.

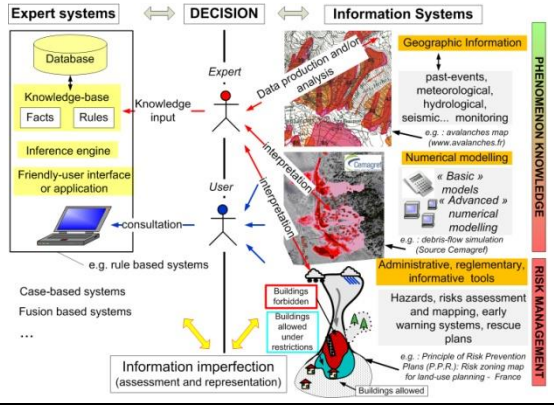


Figure 2: Spatial information is needed for natural risks assessment and management.

For both rockfalls and debris-flows hazard assessment, numerical models are used. In our case:

- For rockfalls, the code RockyFor3D described in Dorren et al. (2006) and Bourrier et al. (2009) simulates the 3D propagation of the rocks as a succession of free flights through the air and rebounds on the soil, modelled by a Digital Elevation Model (DEM). Using input thematic variables related to the falling rock characteristics, it provides spatialized outputs, namely kinetic energy, height, speed of the boulder.
- For debris-flows, Laigle et al. (2003) proposes the lave2D model, dedicated to the computation of the unconfined free-surface spreading of materials with complex rheology. It is based upon the steep-slope-shallow-water-equations which are solved using a finite volume technique which requires first to mesh the domain of interest. Equations are solved taking into account the material rheological behavior and the field topography represented by a DEM. Using input thematic variables related to the input hydrograph and rheological parameters, it provides spatialized outputs, namely flow height and speed.

Both thematic data and DEM can be affected by imperfection (imprecision, uncertainty) depending on the terrain morphology, data acquisition and processing methods. The main issues are therefore:

- To evaluate the quality of available DEM depending on its nature and acquisition process (Lidar, satellite, commercial maps etc.).
- To analyse and represent the effect of spatial information uncertainty on models' results.

This paper describes how innovative methods and tools are used to assess both thematic and spatial data uncertainties and to characterize the impact of the DEM uncertainties on the results of numerical modelling simulation.

II SPATIAL AND THEMATIC UNCERTAINTY PROPAGATION: SCIENTIFIC APPROACH AND MAIN RESULTS

The approach compares two uncertainty propagation approaches, to propagate both DEM and thematic uncertainty:

- The usual, probabilistic, Monte-Carlo method.
- A possibilistic, so-called Hybrid method proposed by Baudrit et al. (2006).

That last Hybrid approach considers the different aspects of information imperfection, especially its imprecision (lack of information, inaccuracy of measure…). It relies on the usual probability theory, the possibility theory as defined by Zadeh (1978) and Dubois et al. (2000), and the belief function theory detailed by Shafer (1976) and Smets et al. (1994). This method generalizes, under some restrictive conditions, the usual Monte Carlo method.

The DEM variability is modelled by a stochastic field using the new ModTer software developped by ProBayes under the ModTer project consortium, and detailed by Crimier et al. (2016) (Figure 3). It takes into account heterogeneity of DEM quality as an input of simulation model: it produces terrain simulations and confidence maps related to altimetric information.

Examples of results (quantile of rock passing heights) of a numerical simulation of rockfalls are proposed: they show the influence of data imperfection including those resulting from expert assessments on the simulation results (Figure 4). To demonstrate the effects of spatial data quality, a sensitivity analysis describes the contribution of the DEM uncertainty on the global uncertainty.

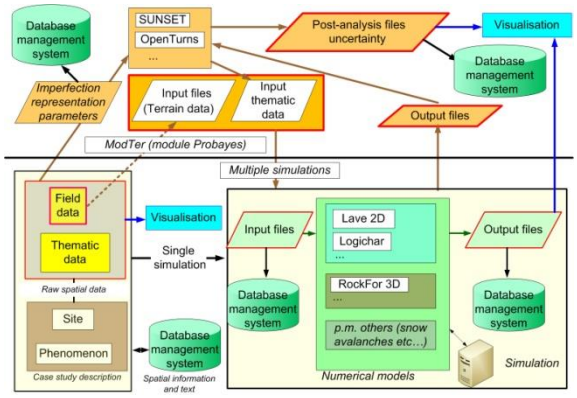


Figure 3: A global framework to evaluate and propagate both thematic and spatial uncertainties using new ModTer software

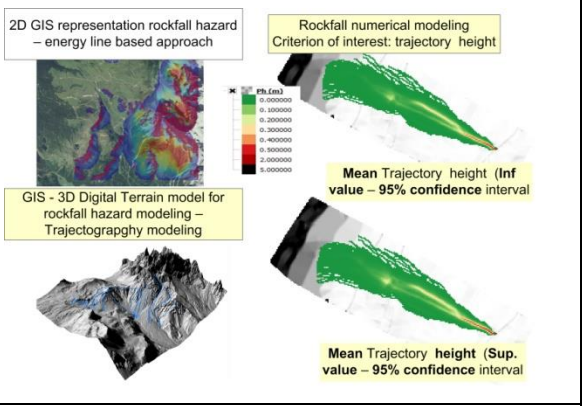


Figure 4: Example of rockfall simulation results

Acknowledgements: Those developments have been partially funded by the ModTer project: a RAPID project supported by the French Defence Procurement Agency (DGA) and the General Directorate for Competitiveness, Industry and Services (DGCIS).

References

Baudrit C., Guyonnet D., Dubois D. (2006). Joint propagation and exploitation of probabilistic and possibilistic information in risk assessment. *IEEE Transactions 14*, 593–608.

Bourrier F., Dorren L., Nicot F., Berger F., Darve F. (2009). Toward objective rockfall trajectory simulation using a stochastic impact model. *Geomorphology 110*(3-4), 68–79.

Crimier N., Mekhnacha K., Tacnet J.-M., Memier M. (2016). Terrain uncertainties modelling software : Modter. In *submitted to ESREL2016 conference*, Glasgow, Scotland.

Dorren L., Berger F., Putters U. (2006). Real-size experiments and 3-d simulation of rockfall on forested and non-forested slopes. *Natural Hazards and Earth System Sciences 6*(1), 145–153.

Dubois D., NGuyen H., Prade H. (2000). Possibility theory, probability and fuzzy sets : Misunderstandings, bridges and gaps. In *Fundamental of Fuzzy Sets*, pp. 343–438.

Laigle D. Hector A.-F., Hübl J., Rickenmann D. (2003). Comparison of numerical simulation of muddy debris flow spreading to records of real events. In *Proceedings of the 3rd International Conference 'Debris-Flow Hazards Mitigation'*, Davos, Switzerland.

Shafer G. (1976). *A mathematical theory of evidence.* Princeton University Press.

Smets P., Kennes P. Kennes R. (1994). The transferable belief model. *Artificial Intelligence 66*(2), 191–234.

Zadeh L.A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems 1*, 3–28.

Performing Multi-Temporal Spatial Data Analysis for Coastal Areas and Assessing Thematic Accuracy

Daniel Cohenca[1]and Carlos Antonio Oliveira Vieira[1*]

[1] Federal University of Santa Catarina, Geocience Department, Florianópolis, Trindade, SC 88040-900 Brazil

*Corresponding author: carlos.vieira@ufsc.br

Abstract

It was used TM Landsat images from 1985, 1994, 2004 and 2014 and Object Based Oriented Analysis (OBIA) image classification techniques in order to study changes in land use and land cover for two cities: Passo de Torres and Balneário Gaivota at the south region of Brazil. Based on these analyses, four informational classes were chosen: Natural Vegetation, Water, Urban and Agriculture Use. Agriculture use includes croplands, livestock activities and reforestation areas. It was used post-classification multi-temporal analysis. The aim of this paper is to perform multi-temporal spatial data analysis in order to identify trajectories of changes on land use and land cover, during the last three decades. This paper also discusses an alternative methodology to assess the thematic accuracy in multi-temporal analysis. The multi-temporal analysis results show that this study area has considerably changed in the last three decades. The most frequent transition was due to the transformation of natural areas into cropland, livestock and reforestation ones, from the year of 1985 to 2014. It was observed that most of transition for urban class occurred from natural vegetated areas, while few transitions occurred from agricultural areas, even considering these classes as being more representatives in the whole study area. This transition was especially common in coastal urban division into plots of land. Besides understanding those processes of change, this study points out the importance of better planning, licensing and monitoring of urban coast division into plots of land to prevent the continuous loss of natural ecosystems and impoverishment of environmental quality. The accuracy assessment discussion shows the importance in choose a better classification strategy in order to perform the multi-temporal analysis in remote sensed data. Furthermore, it is observed that a considerable amount of research needs to be undertaken before the spatial characterization of thematic accuracy associated with multi-temporal analysis can be adequately reported in standardized format and legends.

I INTRODUCTION

Brazilian coastal areas have experienced fast changes in recent years, on form and model of occupation driven by geographical, economic, social forces and public policies. The multi-temporal analysis of changes on land use and land cover is a method focused on understanding, measuring and analysing spatial and temporal changes, both qualitatively and quantitatively.

On the other hand, it is widely acknowledged that classification of remotely sensed imagery has variable and often poor quality. The cause and nature of these errors has been the subject of extensive researches (e.g., Lu and Weng, 2007) in order to improve the accuracy of remotely sensed products.

The aim of this paper is to perform multi-temporal spatial data analysis in order to identify trajectories of changes on land use and land cover, during the last three decades. This study also discusses an alternative methodology to assess the thematic accuracy in multi-temporal analysis.

II STUDY AREA AND MATERIAL

The study area covers the municipalities of Passo de Torresand Balneário Gaivota in coastal southern state of Santa Catarina and covers an area of 25,030 hectares (Figure 1). The landscape can be considered homogeneous with elevation ranging from sea level to 30 meters. This study area is entirely composed of Holocene coastal sedimentary plain (Dieh and Horn Filho, 1996), originally covered by herbaceous marshes, typical of wetlands or fixative dunes, shrub or tree (Falkenberg, 1999; CONAMA, 1999). In addition to the native vegetation, the area has diverse types of land cover including family farms, monoculture timber species (*Pinus* and *Eucalyptus* cultivation), rice cultivation, livestock and urban areas at different stages of occupation.

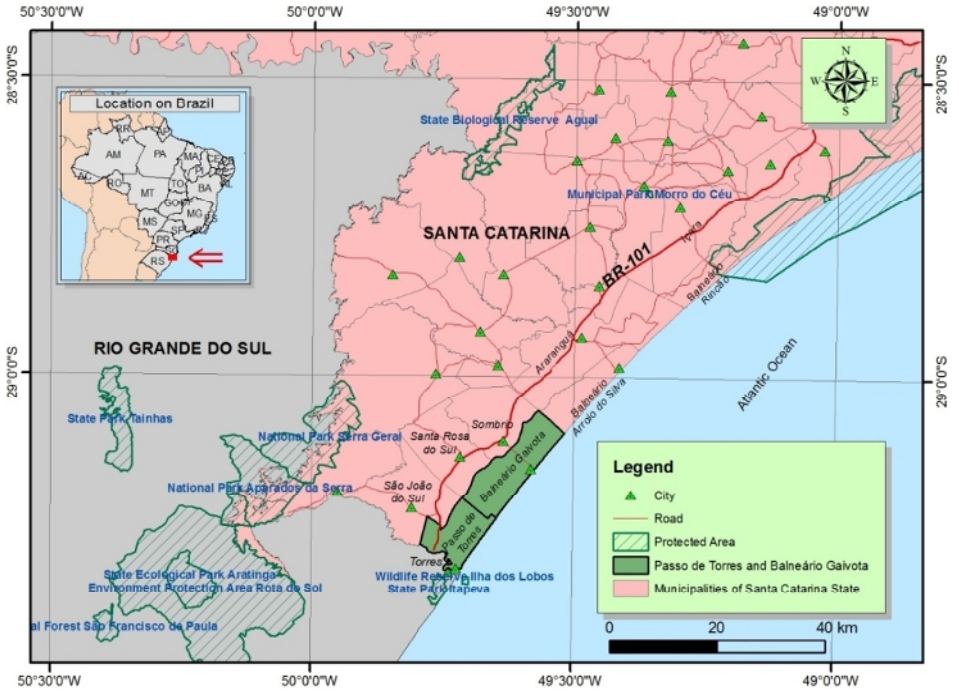


Figure 1: Study area location - municipalities of Passo de Torres and Balneário Gaivota.

III MATERIAL AND METHODS

Landsat TM scenes were selected at intervals of about 10 years, consisting in six spectral bands for each scene, for the images classification process (Table 1). It is important to note that the precise geometric correction process is crucial for the multi-temporal analysis (Morisette and Khorram, 2000; Mundia and Aniya, 2005; Singh, 1989). In order to perform this geometric correction procedure, it was used the image Landsat 8 / OLI 01/30/2014 as reference image (available with geometric correction by the United States Geological Survey - USGS). It was used a plane coordinates UTM system and the Datum WGS84. The resulting average quadratic error was less than 15 meters (half pixel) for all scenes. The resampling method used was the nearest neighbour, in order to maintaining the radiometric properties of the original images. After geometric correction, the four images were clipped using the feature IBGE municipal boundaries of the study area.

In addition, high-resolution images (e.g., SPOT and RapidEye Images) and photogrammetric data (Flights years 1978 and 1996) were used in the accuracy assessment procedures as reference data.

Satellite - Sensor	Date	Bands
LandSat 5 - TM	09/07/1985	1,2,3,4,5,7
LandSat 5 - TM	18/07/1994	1,2,3,4,5,7
LandSat 5 - TM	14/08/2004	1,2,3,4,5,7
LandSat 8 - OLI	30/01/2014	2,3,4,5,6,7

Table 1: Sensors, dates and bands used (orbit 220, point 080).

(i) Multi-temporal classification procedures

In the classification process, it was used an object oriented supervised classification method: OBIA (Object Based Image Analysis). According to Cohenca and Carvalho (2015) the OBIA method is based on image segmentation which consists of subdividing the image into homogeneous regions, called objects, which are the basic elements of the classification process (Benz*et al.*, 2004). Each object has spectral characteristics defined by its average value of pixel´s radiometry and also geometric characteristics such as: contextual information and texture. The segments are classified following a tree process, whose rules for the differentiation of the classes are defined by the interpreter (Francisco and Almeida, 2012).
For the whole set of bands were tested weights parameters, such as: scale, shape and compactness of the objects in each of the images, which define separability of objects and, consequently, the size and number of objects to be generated (Table 2).

Images	Scale	Shape	Compactness
TM -09/07/1985	10	0.1	0.8
TM-18/07/1994	10	0.1	0.9
TM-14/08/2004	10	0.1	0.9
OLI-30/01/2014	90	0.1	0.5

Table 2: Parameters used for segmentation of Landsat images.

The distinctive characteristics of atmospheric conditions on the acquisition of images and different sensors used, motivates each scene be classified separately. The evaluation of the quality of the segmentation was carried out through joint evaluation of the classification results (Darwish *et al*., 2003).

From field visits and analysis of high resolution images, were initially defined eleven thematic informational classes, considered representative of the diversity of types of coverage in the study area. In order to reduce source of imprecision and considering the objective of identifying the changes in coverage between native vegetation, areas of agriculture use and urban areas, regardless of the specific type of use in each segment, the initially defined eleven classes were grouped into four final classes: *Natural Vegetation, Water, Urban and Agriculture Use*. Agriculture use includes croplands, livestock activities and reforestation areas.

In order to assess the accuracy of the classification procedures, it was used the method proposed by Congalton and Green (1999). It was selected 200 pixels at random (50 samples

per class with a minimum distance of 100 meters between samples), for each of these scenes independently (Korting, 2007). Comparing classification results and reference samples were generated confusion (or error) matrices, from which were derived indexes, such as: overall accuracy, Kappa and Conditional Kappa coefficients for each scene.

(ii) Multi-temporal analysis

It was chosen an overlay vector method for comparing the thematic classifications performed for each analysed stage, considering the complexity and diversity of land uses and types of land cover in the study area for the multi-temporal analysis of changes.

Data analysis was quantitatively performed globally as well as regionalized, using the limits of the different settlement pattern identified, based on the main type of use, property size and dynamics of occupation that have been mapped in historical survey of the occupation process. This mapping was crossed with maps of land use change, allowing identify settlement patterns that have undergone major change and what were these changes in different time intervals.

Thus, it was analysed the trajectory of multi-temporal land cover transitions in order to understand the changes detected between the studied decades, in both: global and regionalized manner.

VI RESULTS

Overall accuracies, derived from the confusion matrix, ranged between 84% and 88%, and the Kappa coefficients ranged between 0.793 and 0.836 (Table 3). These accuracies values can be considered very satisfactory. According to the conditional Kappa coefficients, the class with the highest accuracy rating was the class water, followed by urban class. The classes that had greater confusion were Natural and Agriculture Uses. This kind of confusion seems to be due to a lower separability of these classes. Figure 2 shows thematic images for the years of 1985, 1994, 2004 and 2014.

Classes \ Images		1985	1994	2004	2014
Conditional Kappa	Water	0,973	0,947	1,000	0,895
	Natural	0,847	0,720	0,632	0,692
	Agriculture	0,783	0,636	0,803	0,787
	Urban	0,748	0,869	0,771	0,844
Kappa		0,836	0,793	0,800	0,807
Overall Accuracy		88%	84%	85%	85%

Table 3: Overall accuracies, Kappa and Conditional Kappa Coefficients.

The multi-temporal analysis results show that 19.8% of the study area has changed in the last three decades. Between 1985 and 2014, the most frequent transition (57.1% of changes) was due to the transformation of natural areas into agriculture use. The expansion of the urban class over natural areas responded for 24.3% of changes. The conversion of agriculture use into urban use responded for 5.1% of change, while conversion of agricultural areas to vegetation regenerated accounted only for 2.8% of change (Table 4).

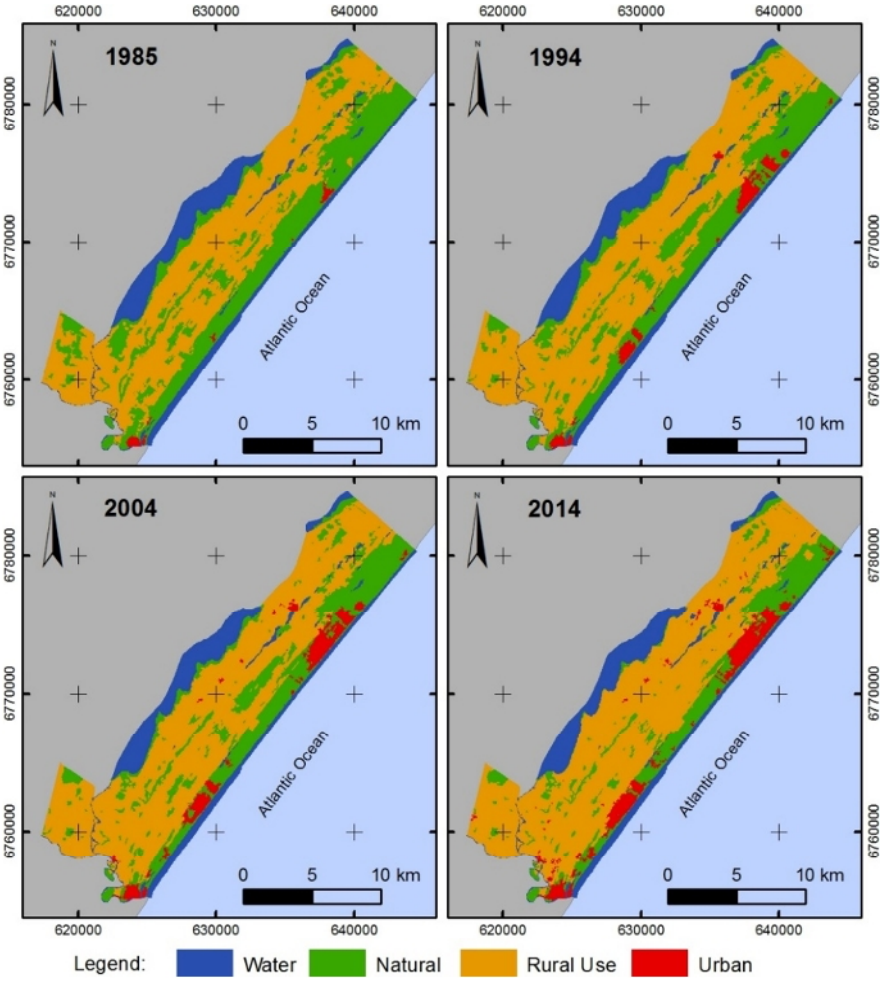


Figure 2: Image Classification results for the years of 1985, 1994, 2004 and 2014.

It was observed that 82% of the transition to urban class occurred from natural vegetated areas, while 18% occurred from agricultural areas, even considering these classes as being more representatives in the whole study area. This transition was especially common in coastal urban division into plot of land.

Remnants of natural class accounted for 17% of the study area in 2014, while in 1985 it was 32%, been particularly concentrated in these coastal urban divisions into plots of land.

Performing multi-temporal spatial data analysis operations on data of unknown accuracy will result in a product with low reliability and restricted use in the decision-making process, while errors deriving from one source can propagate through the database via derived products. Moreover, current accuracy assessment methods are based on non-spatial statistics derived from the confusion or error matrix, which compares the output of a classifier against known reference data. Although these measures are in widespread use, none of them considers the spatial distribution of erroneously classified pixels, either implicitly or explicitly.

Transitions \ Period	1985-1994		1994-2004		2004-2014		1985-2014	
	Area (ha)	%	Area (ha)	%	Area (ha)	%	Area (ha)	%
Natural → Agriculture Use	1090	49,2	815	63,5	1370	72,0	2822	57,1
Natural → Urban	522	23,5	322	25,1	346	18,2	1200	24,3
Agriculture Use → Urban	60	2,7	74	5,7	127	6,7	250	5,1
Agriculture Use→ Natural	411	18,6	58	4,5	26	1,4	138	2,8
Other Transitions	134	6,0	14	1,1	33	1,7	532	10,8
TOTAL	2217		1282		1902		4942	100

Table 4: Frequent transitions in land coverage.

Vieira and Mather (2001) presents one possible way to characterize the spatial distribution of the errors in a thematic classification is by generating a *distance image* showing the distance from individual pixels to the multivariate means of the classes to which they have been assigned (Figure 3c). Either the Euclidean distance or the Mahalanobis distance measure can be used. The former, however, implies spherical clusters in feature space, while the latter takes into account the covariance between the features on which the classification is based. The individual distances are scaled onto a 0-255 range, and displayed as a grey scale image. Darker pixels are spectrally "nearer" to their class centroid (in the sense of statistical distance), and are thus more likely to be classified correctly. On the other hand, pixels with higher distance values are spectrally further from the centroid of the class to which they were assigned, and are thus more likely to be misclassified. This process (presented in Figure 3b) allowed to get known the thematic uncertainties associated for each pixel of the classified image.

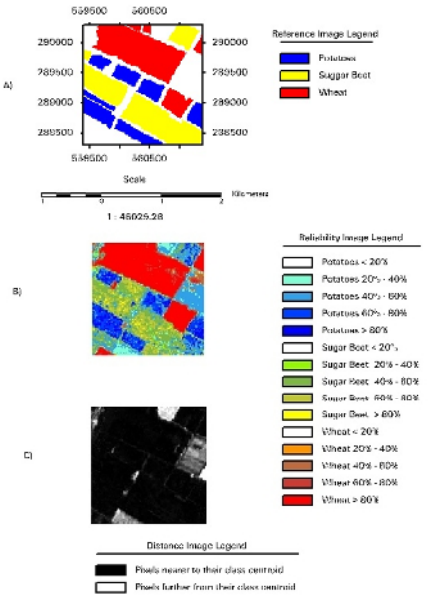


Figure 3: Spatial characterisation of classification errors using thematic image generated by Maximum Likelihood Classifier (100 by 100 pixels). A) Reference Image; B) Reliability Image; C) Distance Image.

Another possible way to characterise the spatial distribution of the errors in a thematic classification is by directly comparing thematic images with their respective ground truth maps. One of the products of this comparison should be an *error binary image* (Figure 3b), in which each point takes the value 0 (correctly labelled) or 1 (erroneously labelled).
On-going researches has been developed in order to propagate these uncertainties through the multi-temporal analyse model and a thematic uncertainty map could be generated, as suggested by Vieira and Mather (2001).

V CONCLUSIONS

The multi-temporal analysis results show that has considerably changed in the last three decades. Between 1985 and 2014, the most frequent transition was due to the transformation of natural areas into agricultural, livestock and reforestation ones. The expansion of the urban class over natural areas responded for 24% of changes. The conversion of agricultural use into urban one responded on exceeding of 5% of change, while conversion of agricultural areas to vegetation regenerated accounted only for less than 3% of change. It was observed that most of transition for urban class occurred from natural vegetated areas, while few transitions occurred from agricultural areas, even considering these classes as being more representatives in the whole study area. This transition was especially common in coastal urban division into plots of land. In 2014, remnants of natural class accounted for 17% of the study area, while in 1985 it was 32%, been particularly concentrated in these coastal urban divisions into plots of land.

Besides understanding those processes of change, this study points out the importance of better planning, licensing and monitoring of urban coast division into plots of land to prevent the continuous loss of natural ecosystems and impoverishment of environmental quality.

The accuracy assessment results point out the importance in choose a better classification strategy in order to perform the multi-temporal analysis in remote sensed data. Furthermore, it is also shown that a considerable amount of research needs to be undertaken before the spatial characterization of thematic accuracy associated with multi-temporal analysis can be adequately reported in standardized format and legends.

References

Benz, U.C.; Hofmann, P.; Willhauck, G.; Lingenfelder, I.; Heynen, M.(2004). Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 58, n. 3, p. 239-258.

Cohenca, D.; Carvalho, R. (2015). Comparação de métodos de classificação OBIA, Máxima Verossimilhança e Distância Mínima em imagem OLI/Landsat-8 em área de alta diversidade de uso do solo. In: *SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO*, XVII, 2015, João Pessoa. Anais... São José dos Campos: INPE, p. 1035-1042.

CONAMA (1999). Resolução Nº 261 de 30 de junho de 1999. Aprova parâmetro básico para análise dos estágios sucessivos de vegetação de restinga para o Estado de Santa Catarina. Conselho Nacional do Meio Ambiente-CONAMA, Brasília: DOU, n° 146, 02/08/1999. p. 29-31.

Congalton, R.; Green, K. (1999). *Assessing the Accuracy of Remotely Sensed Data - Principles and Practices*. Boca Raton: CRC Press, Taylor & Francis Group, 183p.

Darwish, A.; Leukert, K.; Reinhardt, W. (2003). Image segmentation for the purpose of object-based classification. *International Geoscience and Remote Sensing Symposium*, v. 3, p. III: 2039-2041.

Dieh, F.L.; Horn Filho, N.O. (1996). Compartimentação geológico-geomorfológica da zona litorânea e planície costeira do estado de Santa Catarina. *Notas Técnicas*, v. 9, p. 39-50.

Falkenberg, D.B. (1999). Aspectos da flora e da vegetação secundária da restinga de Santa Catarina, sul do Brasil. *Insula*, Florianópolis, n. 28, p. 1-30.

Francisco, C. N.; Almeida, C. M.(2012). Interpretação de imagens orbitais por meio de sistema especialista para mapeamento e cobertura da terra em região montanhosa. *Sociedade e Natureza*, v. 24, n. 2, p. 283-302.

Korting, T.S.; Fonseca, L.M.G.; Castejon, E.F.; Namikawa, L.M. (2014). Improvements in sample selection methods for image classification. *Remote Sensing*, v. 6, n. 8, p. 7580-7591.

Lu, D.; Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, v. 28, n. 5, p. 823-870.

Morisette, J.T.; Khorram, S. (2000). Accuracy assessment curves for satellite-based change detection. *Photogrammetric Engineering and Remote Sensing*, v. 66, n. 7, p. 875-880.

Mundia, C.N.; Aniya, M. (2005). Analysis of land use/cover changes and urban expansion of Nairobi city using remote sensing and GIS. *International Journal of Remote Sensing*, v. 26, n. 13, p. 2831-2849.

Singh, A. (1989). Review article digital change detection techniques using remotely-sensed data.*International journal of remote sensing*, v. 10, n. 6, p. 989-1003.

Vieira, C. A. O., Mather, P. M. (2001). On the Assessment of Spatial Reliability of Thematic Images. In: HALLS, Peter. (Org.). *Innovation in GIS: Spatial Information and the Environment*. UK: Editora London, Cap.01, p. 120-135.

Spatial model sensitivity analyses

Multiway sensitivity analysis of the fusion of earth observation, topography and social media data for rapid flood mapping

Didier G. Leibovici *, Julian F. Rosser

Nottingham Geospatial Institute
University of Nottingham, Nottingham, UK

*Corresponding author: didier.leibovici@nottingham.ac.uk

Abstract
Social media can be a valuable source of information for both detecting and mapping hazard events. Rapidly estimating flood inundation extent can aid decision-making during crises and help in damage assessment. A data fusion method, based on a Bayesian statistical model, which uses weights of evidence to calculate and combine the variables according to their influence in order to map the flooded areas, was previously developed for an experimental timeframe covering the flooding in Oxford, UK during January 2014. The method used three data sources: geotagged photographs, optical remote sensing and high resolution terrain mapping. In this work we aim at evaluating the sensitivity of this method. The multivariate sensitivity analysis proposed, uses an empirical non-parametric approach to generate the variations of the posterior probability of water presence from different levels of prior uncertainties of the input data sources. The sensitivity assessment is spatially depicted as a multiway correspondence analysis of the generated multi-entry table: uncertainty levels for each of the three data sources, the geographical space and the descriptive measures of the output variation, *i.e.* a 5-way array.

Keywords
crowdsourcing, social media, inundation,

I NTRODUCTION

Crowdsourcing has been used in various disaster management situations such as earthquakes (Barrington et al. 2011) and flood damage (Tomnod 2015). Typically, these crowdsourcing scenarios are task orientated. For example, the volunteers are assessing the aftermath of a disaster from satellite image interpretation. Increasingly, analysis may be undertaken on non-authoritative social media reports. For example, analysis of Tweets has been used to source data used in quasi-real time for flood extent estimation (Smith *et al.* 2015), where a hydrological flooding extent model with initialisations from the volunteered observed locations of flood. Rosser *et al.* (2016) proposed to use a data fusion method combining multiple evidence of water presence including social media data sources to estimate the inundation extent. This paper aims to perform a sensitivity analysis of that approach. The data fusion method utilises viewshed analysis (Panteras *et al.* 2015) and weights of evidence (Tehrany *et al*. 2014). The workflow adopted is shown in Figure 1 with the elements involved in the sensitivity analysis reported here highlighted in green (see section II). The three domains of data sources involved in the method are: Landsat-8 water mapping based on the Modified Normalised Difference Water Index (Xu, 2006), geotagged photographs sourced from Flickr to estimate the citizen's extent estimation based on viewshed analysis, and modelled topographic variables sourced from a high resolution Digital Elevation Model.

A classical approach to analyse the sensitivity of a model is to vary factors one-at-a-time and record the effects on the output, such as previously undertaken in multi-criteria mapping problems (Chen et al., 2013). Here, we suggest a multivariate sampling for each combination of the uncertainty levels for the three dimensions. After initial estimation of potential input uncertainties, we modulate these prior uncertainties at different levels to estimate the variation in output uncertainty from the Monte Carlo simulations. The output uncertainty of the posterior probability of water presence is summarised from the simulated distributions in each combination of the input uncertainties. The generated multi-entry table capturing the uncertainty associations (illustrating how sensitive the rapid inundation extend method is), is then analysed from multiway correspondence analysis (Leibovici 2010).

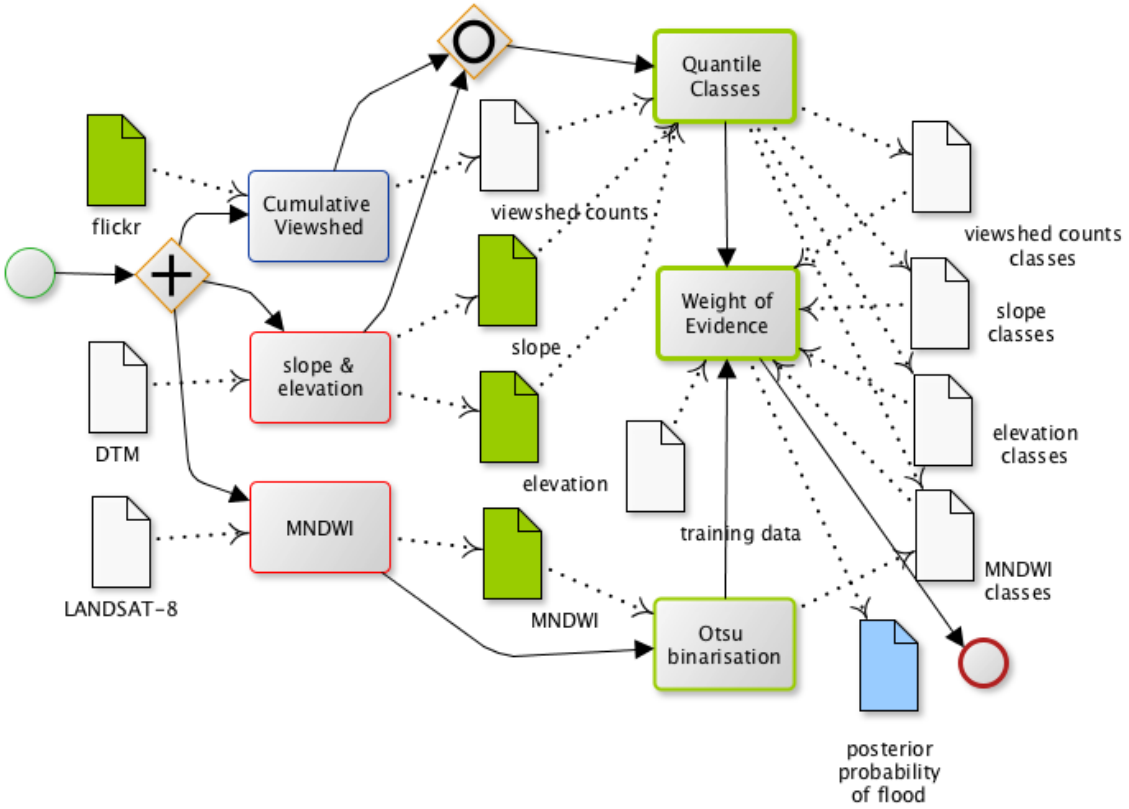


Figure 1: BPMN workflow of the rapid flood inundation data fusion method used in Rosser *et al.* (2016) – in green are the data input where uncertainty sampling takes place.

The method is quite computationally challenging as for each of the 6 174 000 pixels of the studied area, 64 000 evaluations of the workflow are performed: 4 levels of variations in each of the three dimensions with 1000 simulations. On a smaller area of 759 980 pixels (12% of the total area) 1 simulation took 108 seconds (1 core 2.7Ghz), so potentially taking 666 days for the whole study area without parallelisation. For the smaller area and parallelising the code using 20 cores and with only 100 simulations per cell, this is reduced to ~22 hours computation time and will be our first experiment.

II UNCERTAINTIES IN THE DATA FUSION WORKFLOW

Despite the use of quantile classes in the weight of evidence method (see details in Rosser *et al.* 2016) the uncertainty of the measurements or data capture could make the output posterior probability of water presence uncertain. The purpose of the sensitivity analysis is to evaluate

the effect of the uncertainties and to highlight how one data source alone, or in combination with another, can impact on the efficacy of the model.
The whole study area around the city of Oxford, UK, and its characteristics are described in Rosser *et al.* 2016; the data collection coverage for the January 2014 flood event are represented in Figure 2. Ground truth delineation of the flood extent was based on expert interpretation of colour infra-red aerial photography, provided by the Environment Agency Geomatics group (see Figure 6).

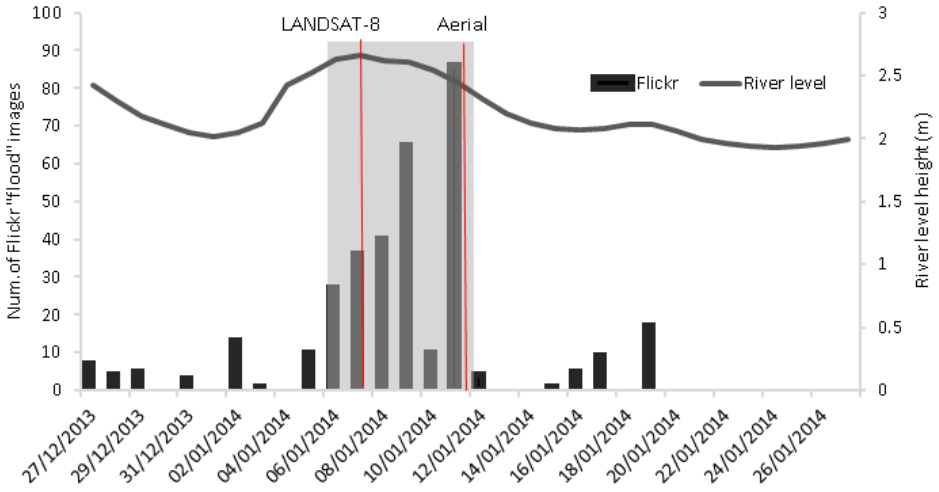


Figure 2: The number of Flickr flood images, Landsat-8 and aerial image (the latter used for ground-truth) acquisition times over the experimental time frame (shaded area). The gauged river level height is also shown.

Besides the representivity and adequation of the data sources for the study, which can be discussed as potential bias, the basis uncertainties taken into account per data source dimension are:

- Landsat-8 dimension

The satellite image matching the peak river level is used for the whole study period, so some variations could be expected in the MNDWI. In the first instance MNDWI uncertainty was estimated from the overall spatial variation of the MNDWI itself but a locally dependent (say 25 pixels' neighbourhood) variance map could be used on the index or directly on the wavelength bands. This gave a standard deviation of 0.22.

- topographic dimension

The Lidar DTM has a given horizontal and vertical uncertainty which changed due to resampling from 1m pixel size to 5m pixel size. To simplify the experiment, slope and elevation uncertainties were estimated from the observed variations globally, respectively 3 and 16 as 1 standard deviation.

- Flickr dimension

The geolocation of the social media data was resampled using as basis a 68% circular error, usually given by the mobile phone (corresponding to 1 standard deviation). As no such information was available in the Flickr data, we took as basis a value of 5m.

The weight of evidence method assumes conditional independence of the 'theme variables' (the predictors) given the outcome:

$$p(V_1 V_2 ... V_m \mid O) = p(V_1 \mid O) p(V_2 \mid O) ... p(V_m \mid O) \quad (1)$$

Assuming normal distribution of the multivariate input, we will sample the multivariate inputs as univariates with the additional assumption that the dependency to the outcome lies in the mean parameter of the distribution.

III MULTIVARIATE SENSITIVITY ANALYSIS

The multivariate sensitivity analysis applied here is based on error propagation through the data fusion workflow. Modulation of the basis uncertainties (section II) was undertaken with four levels of standard deviations: ½ *sd*, 1 *sd*, 2 *sd*, 3 *sd* where *sd* is the basis uncertainty as described in section II. Monte Carlo simulations were derived with normal distributions centred on the observed data values with standard deviations as the modulated uncertainties. A modified workflow taken from Figure 1 was used as the tasks in red ('MNDWI' and 'Slope & Elevation') did not need to run during the sensitivity analysis. Also, the other tasks ('Quantile classification', 'Otsu binarisation' and 'Weight of Evidence') were run using the initial estimated parameters or weights. Therefore the classes adopted the same cut-points and W+ and W- values for the weight of evidence associated to these classes for each simulation. Running the whole workflow with new class and weight estimations each time would encompass a sensitivity analysis, but also a reliability analysis of the workflow itself. Studying the variation of the W+ and W- values would nonetheless be an important reliability and model adequacy feature of the approach to investigate in future work.

The resulting uncertainty of the multivariate input variations has been summarised (using min, Q1, median, mean, Q3, max, var of the simulated distributions at each pixel) in the 5-way array shown in Figure 3.

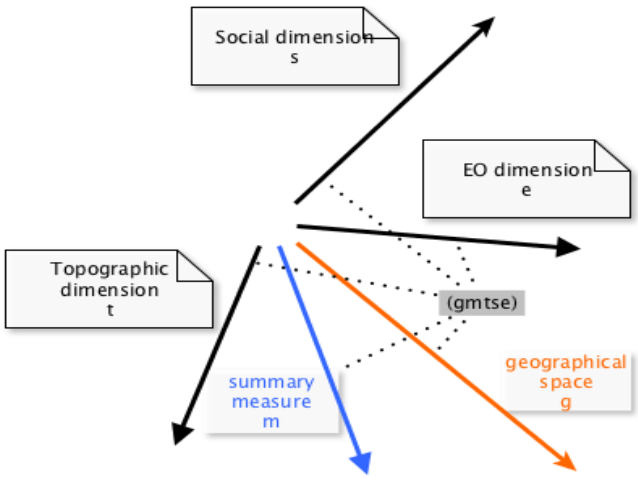


Figure 3: Multivariate sensitivity analysis output stored in a 5-way array, t, s and e indexing each the levels of input uncertainties, g the pixel location and m the distribution indicator of the output uncertainty.

To analyse this sensitivity captured in the 5-way array we can perform multiway correspondence analysis (see Leibovici 2010), allowing an association of the levels of input uncertainty to the intensity of the output uncertainty. The analysis can be performed either on the 3-way table when aggregating across the space and considering only the variance as a summary measure, or on the 4-way table keeping the space dimension. The 5-way analysis can be used to describe further potential associations within the output uncertainty distribution. Before doing so, it is worth looking at the overall variation of the output when aggregating the values over the spatial domain. Figure 4 depicts that overall the workflow is not sensitive to the earth observation

domain as the change in variation in these 4 levels is rather small. Note, also, that the maximum variance observed on the map over the 100 simulations was 4.5 E-4, and the variance over the pixels of the mean of the simulations was 1.12E-04! The patterns of variation between social media uncertainties and topographic uncertainties are very similar in each of the 4 levels of earth observation (MNDWI). Nonetheless, the earth observation uncertainty level 4 shows a larger spread over the topographic levels of uncertainty for social uncertainty level 3, therefore indicating some presence of interaction here. Globally as expected as the uncertainty increases (in social media and topographic data) so does the output variation: linear increasing trend at all earth observation levels. There is not much difference in topographic uncertainty levels except *t1* lower and with an apparent bigger increase in interaction with social dimension. The interaction topographic and social media is only convincing when looking at the difference s1-s4 for t1 versus the others.

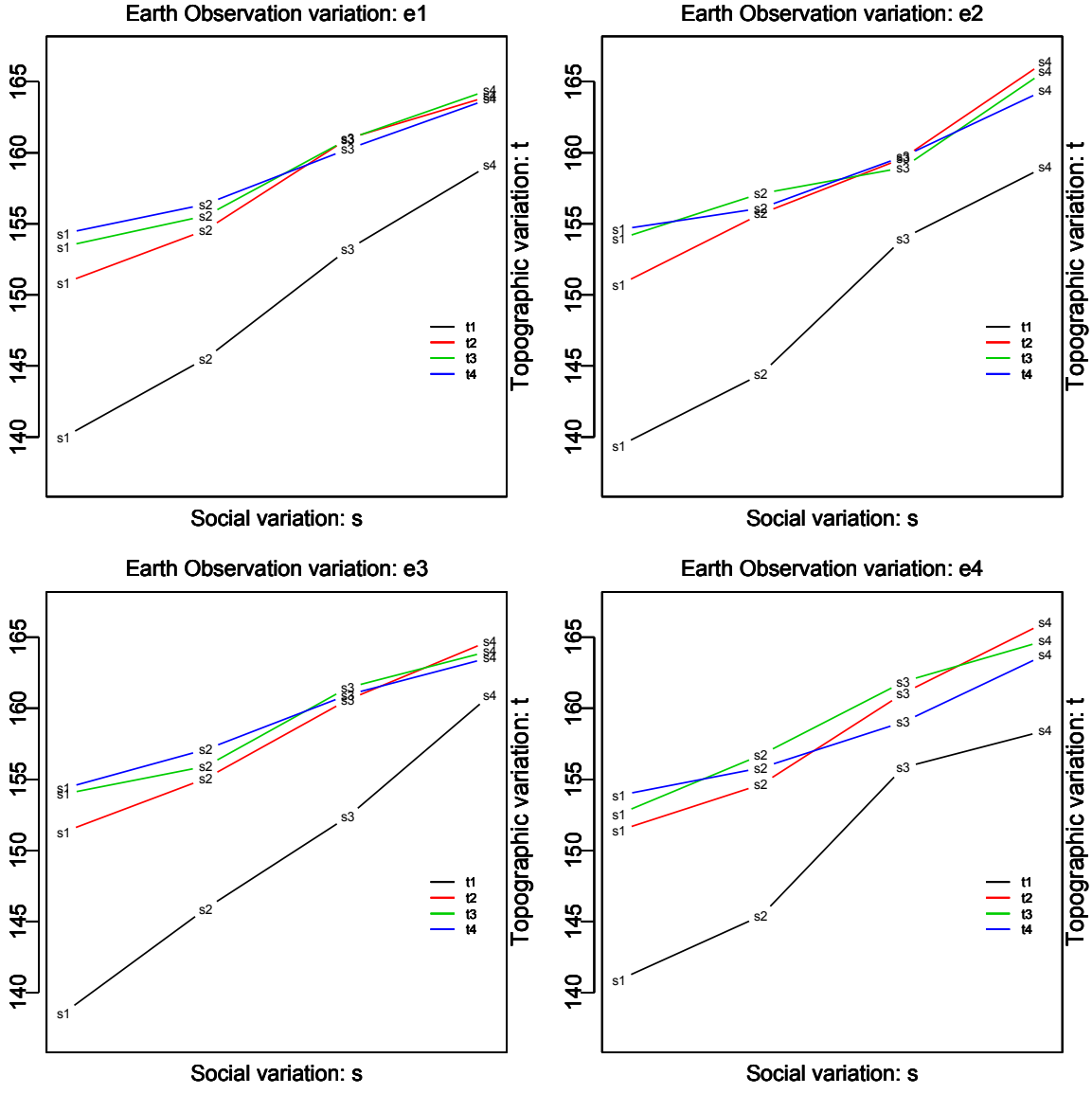


Figure 4: Spatially aggregated output variations (rescaled variance) for each uncertainty combination level (s, t, e).

When performing the multiway correspondence analysis (FCAk), the first principal tensor fitted by the algorithm is the 'independence' principal tensor (vectors of the margins which can

be also automatically set), then the following fitted tensors relate to lack of independence (Leibovici 2010). It is interesting to note that the FCAk of the spatially aggregated data (3-way table) gives 99.98% as representing the independence whilst the analysis on the 4-way table, which includes spatial differentiations, gives 91.95%. So, there is some spatial variation (in the output uncertainty), even though both are high values showing a very small lack of independence between the input uncertainty dimensions (both chi-square values are very small). For the 3-way analysis, the lack of independence has its two-way component breakdowns expressing 2% associated to the social variation margin, 2.7% associated to the marginal topological variation and 84.2% associated to earth observation variation margin, confirming no effect of the latter. Nonetheless, the pure interaction captured in the following (in the order of the decomposition) represents 7% of the lack of independence: *e3* with *t1* and *s4* opposed to *e4* with *t1* and *s1*. The 4% left (100 -(84.2+2+2.7+7)) are concentrated in 3 principal tensors representing nearly 1%, two of these as two-way decompositions associated to the 7% principal tensor. The 4-way analysis puts most of the lack of independence (51%) into a gradient in social media variation levels and spatial differentiation associated with marginal topographic and earth observation variations (Figure 5). The next principal tensor represents 7%, then a few at 1.5%, then a multitude accounting for about 30%. In Figure 5, the blue areas associated to higher social media uncertainty levels (*s3* and *s4*) coincide to the edges of the viewsheds, the brown areas correspond to areas with limited Flickr data.

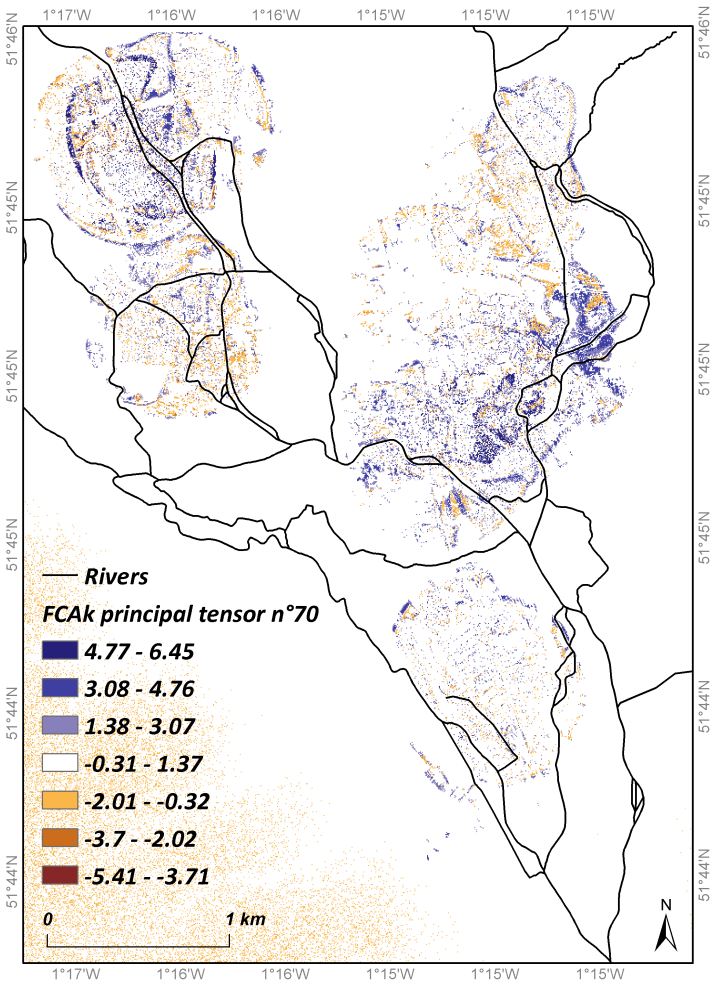


Figure 5: FCAk principal tensor n°70 associated margins to *t* and *e* representing 51% of lack of independence: gradient of *s* is *s1=-1.2*, *s2=-0.7*, *s3=0.4* and *s4=1.4 (*Values are mapped using an equal intervals).

IV DISCUSSION AND CONCLUSIONS

This paper illustrates a multivariate sensitivity analysis for a data fusion algorithm to estimate a flooding extent from a series of predictors. The approach defines each input variable as part of a sampling dimension, with a choice of prior levels of uncertainty per dimension. The multi-way correspondence analyses provide a characterisation of the potential multivariate associations sensitive to the outcome of the data fusion workflow. The example showed a small sensitivity globally and very small influence of the earth observation dimension (EO) without interaction with the other dimensions. If, overall, the increase in input uncertainty implied greater output uncertainty, the topographic dimension showed mostly a difference between its lowest uncertainty level (*t1*) against the other ones and in its interaction with the social media data dimension: bigger increase under t1 as the social dimension uncertainty than the other ones. Spatially, most of the sensitivity relates to the edges of the viewsheds when sampling with high uncertainty levels in the Flickr data, which can be expected.

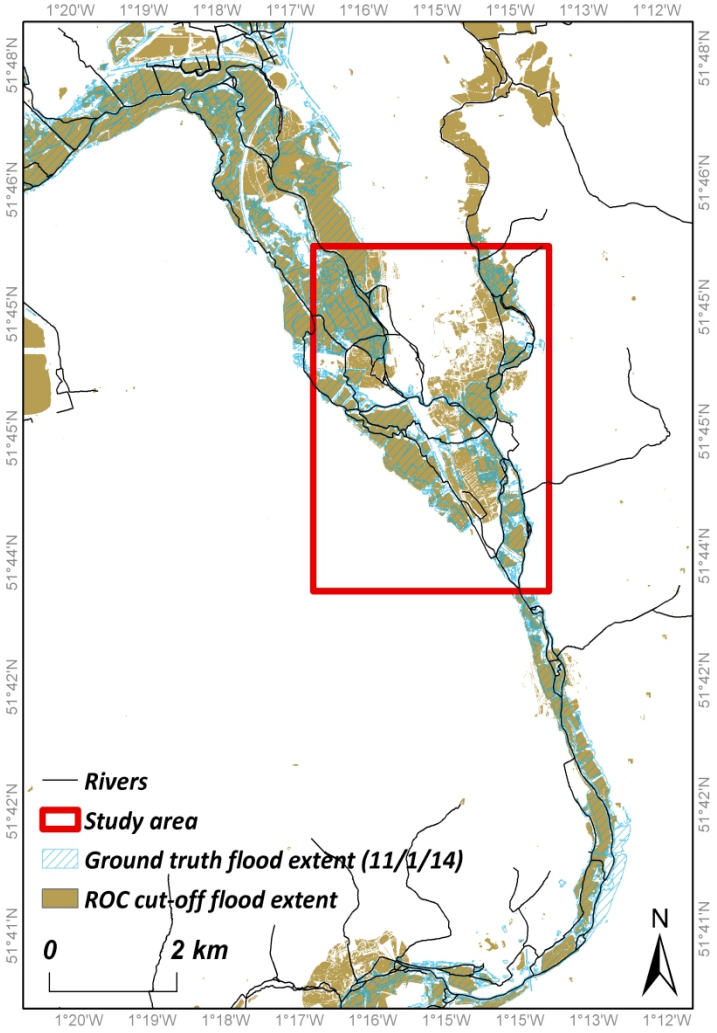


Figure 6: Sensitivity analysis study area, ground-truth and estimated flood extent (derived from Receiver Operating Characteristic).

In this sensitivity analysis, the spatial autocorrelation in the uncertainty sampling has been neglected or reduced simply to the difference in means of the distributions (normal distributions

centred on the observed values). A simple local uncertainty estimation discussed in section II could compensate some of this drawback as the variances could be locally correlated (even though this will still mean sampling at pixel level). Computationally, this would add a potentially expensive step in the initial uncertainty estimation of the inputs but would be done only once. The weight of evidence method producing the posterior probability of flood at a pixel given the predictors comes from the general form:

$$logit(p(O = 1 \mid V_1 V_2 ... V_m)) = logit(p(O = 1)) + \sum_j W^{\pm}(V_j) \quad (2)$$

using the prior and the weights of evidence (see Rosser *et al.* 2016 for more details). Our initial prior was uniform over the area and taken from Rosser *et al.* (2016) in order to allow direct comparison. Note that a Bayesian updating could be performed through deriving a *Beta* distribution from the Monte Carlo simulation (estimating the parameters from the mean and variance of the simulated sample) and using this to generate a new prior in the process. In the last 10 years Bayesian non-parametric methods (*e.g.* Gelfand et al. 2005) have developed which should allow flexibility in modelling spatial and multivariate dependency. However, for large datasets this could be still computationally prohibitive (Berrocal 2016).

References

Berrocal, V. J. (2016). Identifying Trends in the Spatial Errors of a Regional Climate Model via Clustering. *Environmetrics* 27 (2): 90–102.

Chen Y., Yu J. Khan S., (2013). The spatial framework for weight sensitivity analysis in AHP-based multi-criteria decision making, *Environmental Modeling Software*. *48* 129–140.

Gelfand, A.E., Kottas, A., MacEachern, S.N. (2005) Bayesian nonparametric spatial modeling with Dirichlet process mixing. J. Am. Stat. Assoc. 100, 1021–1035

Leibovici D.G., (2010). Spatio-temporal Multiway Decomposition using Principal Tensor Analysis on k-modes: the R package PTAk. *Journal of Statistical Software* 34(10), 1–34.

Panteras G., Wise S., Lu X., Croitoru A., Crooks A., and Stefanidis A., (2015). Triangulating Social Multimedia Content for Event Localization Using Flickr and Twitter. *Transactions in GIS* 19 (5): 694–715. doi:10.1111/tgis.12122.

Rosser J., Leibovici D., Jackson M., (2016). Rapid flood inundation mapping using social media, remote sensing and topographic data. Submitted to: *Natural Hazards.*

Smith, L., Liang, Q., James, P. and Lin, W., (2015). Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management, (Article first published online: 27 MAR 2015 DOI: 10.1111/jfr3.12154)*

Tehrany MS, Pradhan B, Jebur MN (2014) Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *J Hydrol 512:332–343. doi: 10.1016/j.jhydrol.2014.03.008*

Xu H., (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing. 27.* 3025-3033.

A GPU-based Solution for Accelerating Spatially-Explicit Uncertainty- and Sensitivity Analysis in Multi-Criteria Decision Making

Christoph Erlacher[*1], Seda Şalap-Ayça [2], Piotr Jankowski[2], Karl-Heinrich Anders[1], Gernot Paulus[1]

[1] Carinthia University of Applied Sciences – Department of Geoinformation and Environmental Technologies, Austria

[2] San Diego State University – Department of Geography, United States of America

*Corresponding author: c.erlacher@cuas.at

Abstract

This study presents a GPU-based approach to accelerate the time consuming parts of the variance-based spatially-explicit Uncertainty and Sensitivity Analysis (U-SA). Performance comparisons were conducted in respect to a CPU-based NumPy and a GPU-based CUDA implementation. Preliminary results reported herein suggest that the proposed approach will provide a quantitative decision quality measure in complex and comprehensive spatial multi-criteria decision making processes and will allow reasonable computational times making spatially explicit variance-based SA applicable and attractive for large-size problems. Furthermore, it will be beneficial for different application domains like natural hazard risk assessment, landscape assessment, infrastructure planning, environmental impact assessment or identification of land use strategies for sustainable regional development.

Keywords

Spatial Decision Support, GPU, CUDA, Parallelization, Spatially-Explicit Uncertainty- and Sensitivity Analysis, Multi-Criteria Decision Making, MCDM.

1. Introduction

Uncertainty and Sensitivity Analysis (U-SA) is a critical step in multi-criteria decision making. The major goal is to verify the robustness and stability of an implemented model with respect to the existing uncertainties. Saltelli et al. (2008) define sensitivity analysis as "*the study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources in the model input*". Uncertainties can refer to the criterion weights, the decision rules, the standardization process of the input criteria and the accuracy as well as the resolution of the input data (Hwang and Yoon, 1981; Ligmann-Zielinska and Jankowski, 2012; Ganji et al., 2016). Only a handful studies focused on spatially-explicit U-SA of model input factors (Ligmann-Zielinska and Jankowski, 2012; Ligmann-Zielinska and Jankowski, 2014; Şalap-Ayça and Jankowski 2016). As stated by Feizizadeh et al. (2014), Ligman-Zielinska and Jankowski (2014) and Şalap-Ayça and Jankowski (2016) spatially-explicit uncertainty and sensitivity analysis can be a time consuming process, which depends on the project area, the number of criteria and models runs associated with the sample size.

Therefore, this paper reports on the study of high-performance computing approach for spatially-explicit U-SA of the assessment of agricultural land units with a multiple-criteria decision making model. Two implementations, a CPU-based NumPy and a GPU-based CUDA (Compute Unified Device Architecture) solution, are compared regarding the computational time of performing the simulations. The Environmental Benefit Index (EBI) is used by the United States Department of Agriculture (USDA) to prioritize agricultural land units based on environmental benefits and select high-scoring units for USDA crop reserve program. This simple scoring model has been used by the Farm Service Agency (FSA) since 1990 to rank farmers' requests to enroll land into the Conservation Reserve Program during each general sign-up period (competitive bidding). The model includes five environmental factors at the top level ("Wildlife", "Water-Quality", "Soil-Erosion", "Enduring-Benefits" and "Air-Quality"). Detailed information concerning the EBI-Framework can be found in Şalap-Ayça and Jankowski (2016).

2. Theoretical Background

2.1. Multi-Criteria Decision Making

Multi-Criteria Decision Making (MCDM) techniques (Malczewski, 1999; Malczewski, 2006; Malczewski and Rinner, 2015) belong to the field of Spatial Decision Support Systems (SDSS) and consider a set of alternatives that are evaluated on the basis of conflicting criteria. MCDM problems include several objectives and each objective is expressed by a set of criteria where every criterion refers to a certain influence value that is indicated by the criterion weight. The combination of MCDM and Geographic Information Systems (GIS) facilitates the integration of the spatial aspects with MCDM. Geographical data are expressed as criterion maps, which can be differentiated into factor and constraint maps.

In short, the workflow of the decision making process includes the standardization of the criteria (transformation) regarding comparability and the application of decision rules like Weighted Linear Combination (WLC), Ordered Weighted Averaging (Yager, 1988; Malczewski, 2006) or Analytic Hierarchy Process (Saaty, 1980), which results in the monotonic rank-order of decision alternatives leading to final recommendation.

2.2. Spatially-Explicit Uncertainty and Sensitivity Analysis

As previously mentioned, sensitivity analysis is a critical part of MCDM in order to verify the robustness and stability of the implemented model. In spatial multi-criteria evaluation uncertainty can be potentially associated with the selection of decision criteria, criteria measurement (inaccuracy and errors) and expert's preferences that represent the criteria weights. Ligmann-Zielinska and Jankowski (2008) state, that a decision situation typically comprises aspatial as well as spatial aspects. However, the vast majority of reported sensitivity analysis studies and evaluation methods concern only the aspatial nature of decision situations while spatial characteristics are only included implicitly. Due to the fact that the spatial distribution of options and their criteria values may potentially influence the rank order of alternatives, the authors emphasize the consideration of spatial criteria (e.g. *proximity, compactness, contiguity, etc.*) and spatial weighting (varying the criterion importance over space and assigning different weights to different spatial units) in GIS-based multi criteria evaluation, which calls for new, spatially-explicit methods of U-SA. As described by Ligmann-Zielinska and Jankowski (2014) uncertainty analysis helps to quantify the variability of model outcomes, whereas sensitivity analysis focuses on the identification of decision criteria or criteria weights that are responsible for the variability. Ligmann-Zielinska and Jankowski (2014) examined the robustness of land suitability evaluation with help of Monte Carlo Simulation (MCS) and variance-based sensitivity analysis. Variance-based sensitivity

analysis is often recommended because it represents a model independent procedure and is applicable for spatially-explicit data. Furthermore, in contrast to one at a time SA (Si, first order sensitivity indices) variance-based SA also incorporates the interaction of input factors (ST, total order sensitivity indices). *The (S,ST) pair offers a succinct yet comprehensive measure of input influence* (Ligmann-Zielinska and Jankowski, 2014).

2.3. GPU-based Parallelization

In the course of this research project, a GPU-based prototype was developed focusing on an acceleration strategy for performing the MCS. MCS represents the most time consuming component of the variance-based spatially-explicit uncertainty and sensitivity analysis method, which depends on the number of criteria and model runs associated with the sample size. Tang and Jia (2014) present an approach that focuses on parallel Graphics Processing Units (GPUs) in order to enable and accelerate the sensitivity analysis of large agent- based modelling of spatial opinion exchange. A solution for accelerating time-consuming calculations is domain-specific and depends on the implemented mathematical functions. General-Purpose Computing on Graphics Processing Units (GPGPUs) are used for computer graphics and entertainment, or GPU co-processors that are constructed to accelerate massively parallel floating-point operations (Krömer et al, 2014). The GPU accelerated computing prototype utilizes the CUDA architecture for NVIDIA GPUs. CPUs (Central Processing Unit) consist of few cores and are optimized for sequential serial processing. GPUs have a parallel architecture that incorporates thousands of smaller but more efficient cores. The calculations are organized into threads, blocks and grids, and are distributed with respect to the streaming multiprocessors and cores. Detailed explanations on GPU programming can be found in Wilt (2013).

3. Methodology

The original spatially-explicit Sensitivity and Uncertainty Framework (Ligmann-Zielinska and Jankowski, 2014) refers to a Python implementation that incorporates two dimensional NumPy arrays for the weight samples (N) and the criterion maps (k). The weight samples are generated using the quasi-random Sobol's experimental design (Software Simlab) and the Ideal Point (IP) decision rule is used to compute the suitability surfaces (see Figure 1). The total number of model runs (R) is represented by the following formula:

$$R = (k + 2) * N \quad (1)$$

In order to compare the performance of the original Framework with the CUDA-GPU version two-dimensional NumPy arrays representing evaluation criteria are flattened in order to calculate the suitability surfaces (see Figure 2). Consequently, the dimension of the final suitability surface is represented by the number of model runs (rows) and the number locations (raster cells). Each row of the final suitability surface represents one suitability map that is associated with the corresponding weight sample and the used decision rule (see Figure 3).

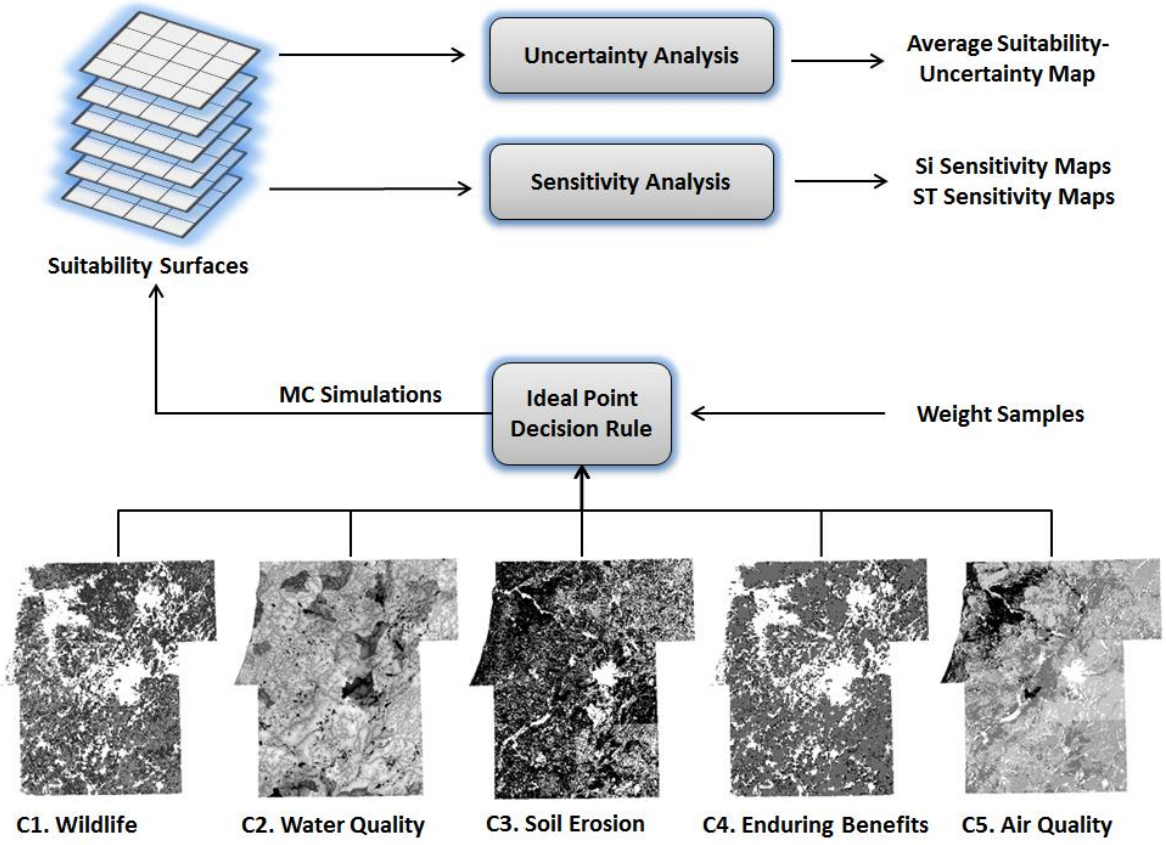


Figure 1: Displays the workflow to conduct a variance-based spatially-explicit U-SA.

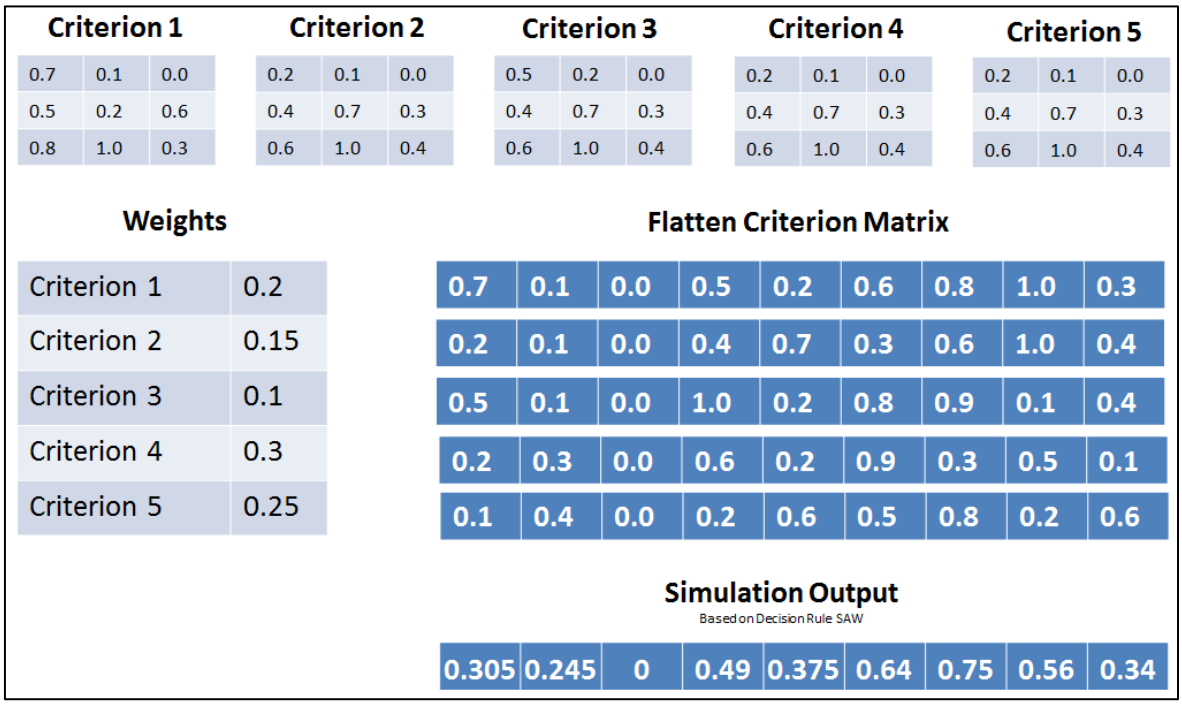


Figure 2: Represents the data structure to perform the GPU-based MCS.

Weight Sample					
#1	0.2	0.15	0.1	0.3	0.25
#2	0.25	0.1	0.15	0.2	0.3
#3	0.15	0.2	0.2	0.25	0.2
#4	0.3	0.15	0.15	0.2	0.2
#...	...	...	....		

Output of Simulations									
#1	0.45	0.1	0	0.72	0.35	0.61	0.79	0.55	0.38
#2	0.35	0.245	0	0.495	0.37	0.63	0.775	0.525	0.375
#3	0.315	0.23	0	0.545	0.38	0.635	0.775	0.535	0.35
#4	0.375	0.215	0	0.52	0.355	0.625	0.765	0.605	0.35
#..	...	...	...	...	...	...	...	...	...

Figure 3: Illustrates the structure of the suitability surfaces, where each row represents one suitability raster map based on the decision rule and the criterion weights.

The algorithm to perform the MCS for the CUDA-GPU implementation is divided into the host part (CPU) and the device part (GPU). Parallel portions of the application are executed on the device as kernels, whereas the host represents the program logic for memory allocation, data partitioning and recombination (see Figure 4). The data partitioning depends on the computing and memory capabilities of the device in order to avoid memory overflow. Therefore, the first task of the controller is to identify the available computing capacity of the device. For example, GPU memory could be allocated by another software or task. According to the determined GPU resources the number of partitions is calculated. For each partition the device memory for a subset of the suitability surfaces has to be allocated, which is necessary for calling kernels. A kernel is launched as a grid of thread blocks and each thread block consists of a specified number of threads (e.g. warp size). The number of blocks per grid and the number of threads per block depend on the number of rows and columns of the suitability surface.

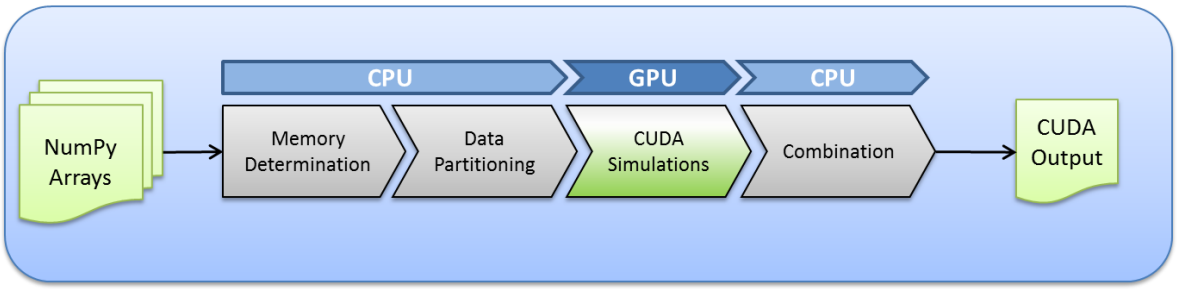


Figure 4: Overview about the CUDA-GPU based MSC computation.

Figure 5 illustrates a more detailed representation of the CUDA-GPU based Monte Carlo Simulations to generate the suitability surfaces.

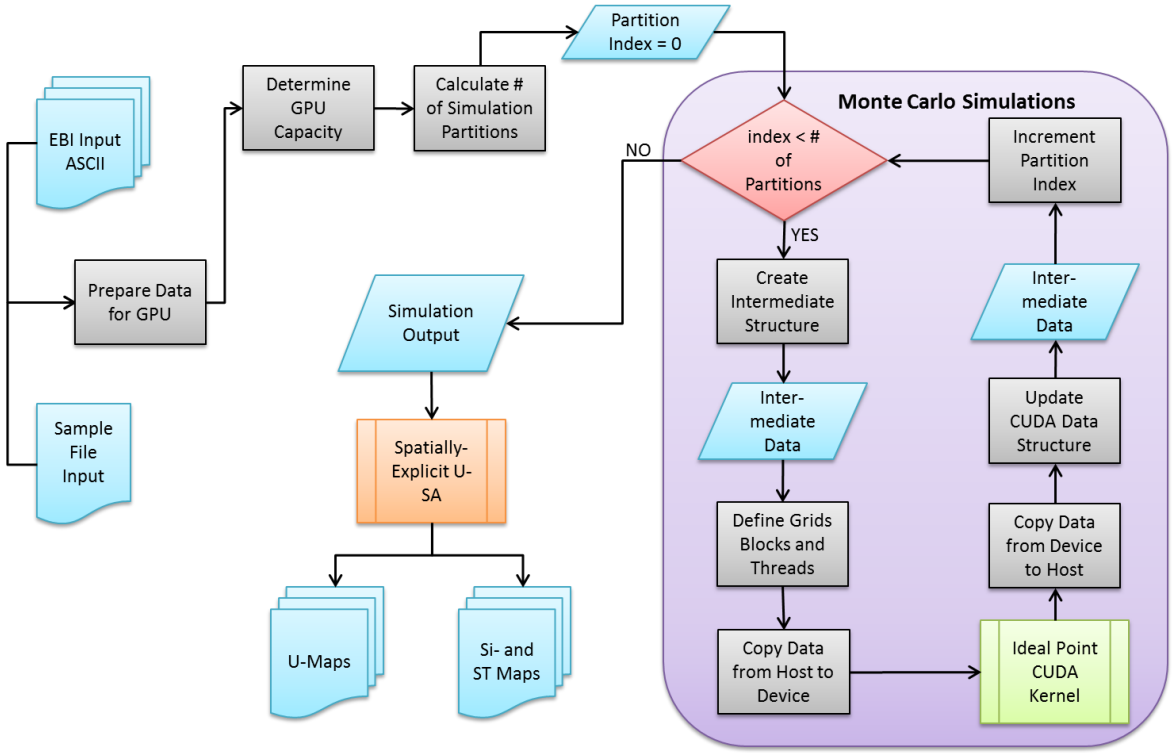


Figure 5: Detailed workflow of the CUDA-GPU based MSC computation.

4. First Speed Up Results

In order to facilitate the performance tests, different NVIDIA GPUs were investigated. The results confirm the efficacy of employing GPU-based solutions for computationally demanding, spatially-explicit U-SA. For example, for the Tesla k40c device, a computational acceleration up to a factor of 150 was achieved. The case study tested model inputs with 5 input factors, 1,475,464 locations, and 2,664 simulation runs.

Tables 1-3 illustrate CPU- and GPU specification overview and the performance comparison of different GPUs as well as a performance comparison regarding the CPU (NumPy – a package for scientific computing with Python) and CUDA (Anaconda NumbaPro) implementation.

Tesla – GPU Specifications	Tesla K40 Workstation	Tesla K20m Server
Peak double-precision floating point performance (Tflops)	1.43	1.17
Peak single-precision floating point performance (Tflops)	4.29	3.52
# of CUDA cores	2880	2496
Total Global Memory – GDDR5 (GB)	12	5
Memory bandwidth (GB/sec)	288	208

Table 1: Tesla GPU Specification Overview.

CPU Specifications	Intel Xeon E5-1620 v3 Workstation	Intel Xeon E5-2697 v2 Server
Intel Smart Cache (MB)	10	30
# of Cores	4	12
# of Threads	8	24
Processor Base Frequency (GHz)	3.5	2.7
Max memory bandwidth (GB/sec)	68	59.7

Table 2: Intel Xeon CPU Specification Overview.

Table 3 reports the relative speed ups for both Tesla GPU cards with respect to the computational demand of the CPUs. The Tesla K40c device (Workstation) is 1.809 times faster than the Tesla K20m device (Server).

Performance Comparison	Workstation	Server
Avg. Elapsed Time for CUDA (min)	0.1215	0.2198
Avg. Elapsed Time for NumPy (min)	18.6712	25.7937
CUDA Speed Up	153.6724	117.3508

Table 3: CUDA-NumPy Performance Comparison.

5. Discussion and Future Prospects

The implemented GPU-based prototype offers reasonable computation times to perform spatially-explicit U-SA, which makes this type of sensitivity analysis applicable and potentially attractive for distributed-output (spatially-explicit) models without using expensive servers or super computers. Furthermore, this solution allows the integration of further decision rules (WLC, OWA etc.) by changing the kernel function and can be applied for different application areas.

The proposed workflow incorporating the data preparation regarding the supported interface and the partitioning and recombination of the GPU-based computation is limited to local operations. For supporting the calculation of local variability (focal- and zonal operations) in respect to criteria values and weights, a conceptual definition for the data structure as well as partitioning- and recombination approach have to be adapted (Malczewski, 2011; Şalap-Ayça and Jankowski; 2016).

Further accelerations can be achieved by implementing additional CUDA kernels for the uncertainty and sensitivity computations. Additionally, the distribution of the CUDA calculations among many GPUs will be investigated too as the next step of reported here research.

6. Acknowledgement

Financial support for this work was provided by the National Science Foundation Geography and Spatial Sciences Program Grant No. BCS-1263071. Any opinion, findings, conclusions, and recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

References

Feizizadeh, B., Jankowski, P., Blaschke, T. (2014). A GIS based Spatially-explicit Sensitivity and Uncertainty Analysis Approach for Multi-Criteria Decision Analysis. *Computers and Geosciences*, Volume 64, pp. 81-95.

Ganji, A., Maier, H., Dandy, G. (2016). A modified Sobol' sensitivity analysis method for decision-making in environmental problems. *Environmental Modelling & Software*, Volume 75, pp. 15-27.

Hwang, C.L., Yoon, K., 1981. Multiple Attribute Decision Making Methods and Applications: A State of the Art Survey. Springer-Verlag, Berlin.

Krömer, P., Platoš, J., Snášel, V. (2014). Nature-Inspired Meta-Heuristics on Modern GPUs: State of the Art and Brief Survey of Selected Algorithms. *International Journal of Parallel Programming*, Volume 42, Issue 5, pp. 681–709.

Ligmann-Zielinska, A., Jankowski, P. (2014). Spatially-Explicit Integrated Uncertainty and Sensitivity Analysis of Criteria Weights in Multicriteria Land Suitability Evaluation. *Environmental Modelling & Software*, Volume 57, pp. 235-247.

Ligmann-Zielinska, A., Jankowski, P. (2012). Impact of proximity-adjusted preferences on rank-order stability in geographical multicriteria decision analysis. *J. Geogr. Syst*. Volume 14, Issue 2, pp. 167-187.

Ligmann-Zielinska, A., Jankowski, P. (2008). *A Framework for Sensitivity Analysis in Spatial Multiple Criteria Evaluation*. In: T.J. Civa et al., eds. GISience 2008, LNCS 5266, Springer-Verlag Berlin, pp. 217-333.

Malczewski, J. (1999). GIS and Multicriteria Decision Analysis. New York, Wiley.

Malczewski, J. (2006). Ordered weighted averaging with fuzzy quantifiers: GIS-based multicriteria evaluation for land-use suitability analysis. *International Journal of Applied Earth Observation and Geoinformation*. Volume 8, pp. 270–277.

Malczewski, J. (2011). Local Weighted Linear Combination. *Transactions in GIS*, Volume 15, Issue 4, pp. 439-455.

Malczewski, J., Rinner, C. (2015). Multicriteria Decision Analysis in Geographic Information Science, Advances in Geographic Information Science. New York, Springer.

Saaty, T.-L., 1980. The analytic hierarchy process. New York, McGraw-Hill.

Şalap-Ayça, S., Jankowski, P. (2016). Integrating Local Multi-Criteria Evaluation with Spatially Explicit Uncertainty-Sensitivity Analysis, Spatial Cognition & Computation, DOI: 10.1080/13875868.2015.1137578

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S. (2008). *Global Sensitivity Analysis*: The Primer. John While & Sons, Ltd.

Tang, W., Jia, M. (2014). Global Sensitivity Analysis of a Large Agent-Based Model of Spatial Opinion Exchange: A Heterogeneous Multi-GPU Acceleration Approach. *Annals of the Association of American Geographers*, Volume 104, Issue 3, pp. 485-509.

Wilt, N. (2013). The CUDA Handbook: A Comprehensive Guide to GPU programming. New Your, Addison-Wesley.

Yager, R.-R. (1988). On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Transactions on Systems, Man and Cybernetics, Volume* 18, Issue 1, pp. 183-190, IEEE Press Piscataway, NJ, USA.

Uncertainty propagation in urban hydrology water quality modelling

J.A. Torres-Matallana[*1,2], **U. Leopold**[1], **G.B.M. Heuvelink**[2]

[1]Luxembourg Institute of Science and Technology, Luxembourg
[2]Wageningen University, Netherlands

*Corresponding author: arturo.torres@list.lu

Abstract

Uncertainty is often ignored in urban hydrology modelling. Engineering practice typically ignores uncertainties and uncertainty propagation. This can have large impacts, such as the wrong dimensioning of urban drainage systems and the inaccurate estimation of pollution in the environment caused by combined sewer overflows. This paper presents an uncertainty propagation analysis in urban hydrology modelling. The case study was the Haute-Sûre catchment in Luxembourg for one yearly time series measured in 2010, and 10 individual rainfall events measured in 2011. The selection of model input variables for uncertainty quantification was based on their level of uncertainty and model sensitivity. Probability distribution functions were defined to represent the uncertainty of the input variables. We applied a Monte Carlo technique using a simplified model, EmiStatR, which simulates the volume and substance flows in urban drainage systems. We focus in loads and concentrations of chemical oxygen demand and ammonium, which are important variables for wastewater and surface water quality management.

Keywords: uncertainty propagation, urban hydrology modelling, EmiStatR, Monte Carlo

I INTRODUCTION

Uncertainty is often ignored in urban hydrology modelling (Mitchell et al., 2007; Bach et al., 2014). Engineering practice typically ignores uncertainties and uncertainty propagation, among others because of lack of user-friendly implementations (Schellart et al., 2010). This can have large impacts, such as the wrong dimensioning of urban drainage systems and the inaccurate estimation of pollution in the environment caused by combined sewer overflows (CSOs).

Six approaches to address uncertainty in the context of urban water systems are identified (Walker et al., 2003; Refsgaard et al., 2007; van Keur et al., 2008; Bach et al., 2014): 1) determinism; 2) statistical uncertainty; 3) scenario uncertainty; 4) qualitative uncertainty; 5) recognised ignorance; and 6) total (unrecognised) ignorance. Following van Keur et al. (2008), *determinism* applies when we have knowledge with absolute certainty about the system under analysis. The *statistical* approach is useful when it is possible to describe in statistical terms the uncertainty. This occurs when errors and uncertainties can be quantified by probability distribution functions (pdfs). The *scenario* approach, in contrast, applies when quantitative probabilities cannot be determined. It is used when possible outcomes can be listed without pretending that the list is exhaustive and without attaching probabilities to each possible outcome (Brown, 2004). The *qualitative* approach is used when *uncertainty cannot be characterised statistically, and not all outcomes are known* (Brown, 2004). Recognised ignorance occurs in a situation of awareness of lack of knowledge (van Keur et al., 2008). Finally, *total ignorance* is the state of *complete lack of awareness about possible outcomes* (van Keur et al., 2008). Amongst the

approaches described above, in this paper we will limit ourselves to the statical approach to characterise and propagate uncertainties.

Sources of uncertainty in the context of the evaluation and design of urban water infrastructure are identified by Neumann (2007): model structure, model parameters, errors in input data, and in numerical and computational procedures. Our focus is on model input as a primary source of uncertainty.

Uncertainty propagation analysis can be done analytically by means of the Taylor series method, or numerically by means of Monte Carlo simulation. This paper presents an uncertainty propagation analysis in urban hydrology modelling. For this we apply a Monte Carlo analysis using the EmiStatR model, which simulates the volume and substance flows in urban drainage systems. We use a case study from the Haute-Sûre catchment in Luxembourg for one yearly time series measured in 2010, and 10 individual rainfall events measured in 2011. We focus on substances as chemical oxygen demand (COD) and ammonium (NH_4), which are important variables for wastewater and surface water quality management.

II MATERIALS AND METHODS

2.1 The EmiStatR model

The EmiStat model (Klepiszewski and Seiffert, 2013) is a Microsoft Excel based model that was developed for the Luxembourgish water authority as a tool for the evaluation of planning scenarios of sewer systems. EmiStatR is the R implementation of EmiStat. It provides a fast estimation of combined wastewater emissions by imposing a strongly simplified representation of the real-world system. It can aid the planning and design of hydraulic structures and pollutant handling, without the requirement of extensive simulation tools. The EmiStatR model includes six main components to simulate combined sewage discharges of a catchment (Figure 1).

The sewer system under investigation includes tank structures to store first flush pollutant peaks. After filling of the storage volume a combined sewage overflow structure discharges subsequent volume and pollutant inflows exceeding the structures outflow from the wastewater treatment plant (WWTP) to the receiving water.

In EmiStatR a simple volume balance taking into account inflow volume, present storage capacity and outflow to the WWTP is implemented to simulate the volume in the tank structure. In case of an overflow the pollutant concentrations in the CSO are equivalent to the combined sewage inflow concentrations of the structure.

The pollutants typically taken into account are total COD and NH_4. The variable COD is the standard used in the framework of the dimensioning of CSO structures in Luxembourg. NH_4 represents a diluted substance which can have a significant impact on surface water quality due to possible transformation to ammonia (NH_3).

At the CSO tank structure a simple volume balancing takes place: 1) substance and volume flows are stored and discharged to the WWTP if the storage volume is not completely filled up; 2) if the storage volume is completely filled up the proportion of the volume inflow which is not discharged to the WWTP goes to the CSO.

2.2 Sewer systems of the Haute-Sûre sub-catchments

The study area is a sub-catchment of the Haute-Sûre catchment in the north-west of Luxembourg. The combined sewer system of the sub-catchment drains eight villages: Büderscheid

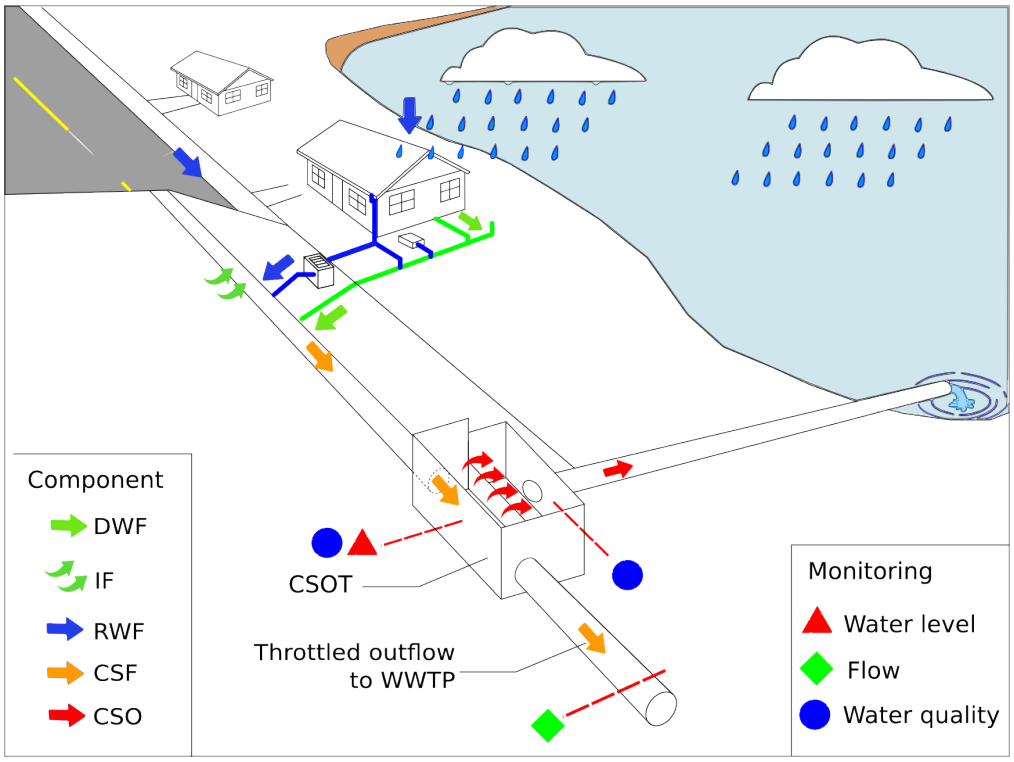


Figure 1: Main components of the EmiStatR model: 1) Dry Weather Flow (DWF) including Infiltration Flow (IF); 2) Pollution of DWF; 3) Rain Weather Flow (RWF); 4) Pollution of RWF; 5) Combined Sewage Flow (CSF) and pollution; and 6) Combined Sewer Overflow (CSO) and pollution. CSOT = CSO tank (Background adapted from: Sanitary-District (2015))

(BUD), Dahl (DAH), Eschdorf/Ost (ESO), Goesdorf (GOE), Heiderscheid (HEI), Kaundorf (KAU), Nocher (NOC) and Nocher-Route (NOR). The local sewer system downstream each village has a CSO tank to store pollutant peaks in the first flush of combined sewage flows. Table 2 shows the general characteristics of each CSO tank for each village. Figure 2 depicts their locations and the delineation of the catchment. The main land use types in the villages are residential, smaller industries and farms. Outside of the villages forest as well as agricultural arable and grassland are the dominating land uses. The receiving water bodies at CSO structures are tributaries of the river Sûre.

The general input data of the EmiStatR model for simulating the eight CSO structures is presented in Tables 1 and 2.

2.3 Model input uncertainty assessment

Following recommendations from Nol et al. (2010), not all model inputs were taken into account in the Monte Carlo uncertainty propagation analysis. Only those inputs that have a large uncertainty and to which the model is sensitive were included. The selection of model input for uncertainty quantification was based on a list of all model inputs, their level of uncertainty (low or high) and the level of model sensitivity (low or high). The level of uncertainty of the inputs was defined by expert judgement, literature research, measurements of different model inputs in the sewer system, and interviews with experts. The developers of the EmiStatR model, which have detailed knowledge about the processes in the system, were also consulted. The level of model sensitivity was derived by interpreting the model structure and components, interviews with experts, and model runs with EmiStatR. The main task to quantify input uncertainty is

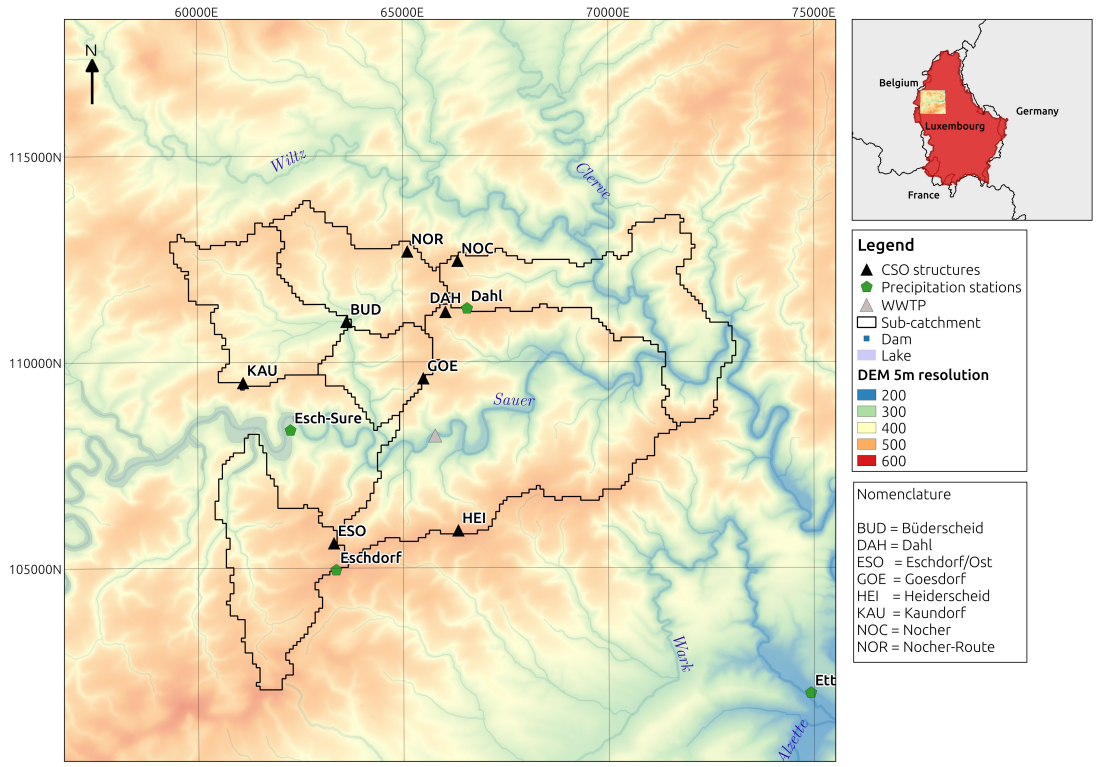


Figure 2: The Haute-Sûre catchment and villages of the sewer system. Location of CSO structures and precipitation measurement stations.

to define the probability distribution function (pdf) that represents the uncertainty of the variable chosen. The uncertainties of selected model inputs were characterized with pdfs following Heuvelink et al. (2007).

Measurement campaigns were done in Goesdorf from 28th April to 24th June 2011, in Kaundorf from 22nd June to 18th August in 2010 and from 20th July to 5th August in 2011, and in Nocher-Route from 18th November 2010 to 27th April 2011.

2.4 Uncertainty propagation

After the definition of model input uncertainties a Monte Carlo simulation was done to propagate input uncertainty to model output. The Monte Carlo method runs the EmiStatR model repeatedly, each time using different model input values, sampled from their probability distribution. The method thus consists of the following steps:

1. Repeat N times:
   (a) Generate a set of realisations of the uncertain model inputs
   (b) For this set of realisations, run the model and store the output
2. Compute and store sample statistics from the N model outputs.

Here, N is the number of Monte Carlo runs, i.e. the Monte Carlo sample size. Common sample statistics that measure the uncertainty are the standard deviation and the width of prediction intervals, which can be easily calculated from the N Monte Carlo outputs.

Variable	Value
Wastewater	
Water consumption, **qs**, [l/(PE · d]	150
Pollution COD, **CODs**, [g/(PE · d)]	104
Pollution NH_4, **NH4s**, [g/(PE · d)]	4.7
Infiltration water	
Inflow, **qf** ,[l/(s · ha)]	0.05
Pollution COD, **CODf**, [g/(PE · d)]	0
Pollution NH_4, **NH4f**, [g/(PE · d)]	0
Rainwater	
Precipitation time series, **P** [mm/min]	P1
Pollution COD, **CODr**, [mg/l]	80
Pollution NH4, **NH4r** [mg/l]	0
Storm water runoff	
Real flow time in the catchment, **tf** [min]	20

Table 1: General input data of the EmiStatR test model.

III RESULTS

3.1 Simulations of the CSO structures

Deterministic simulations of the EmiStatR model for the eight CSO structures were performed to compare these with the outcomes of the uncertainty analysis. The deterministic input variables of the simulations are given in Table 1. The individual deterministic input variables for each CSO structure are presented in Table 2. Nine output variables were evaluated. Three out of nine are related with water quantity: volume of filling of the CSO tank, $VTank$ [m^3]; overflow volume, V_{Ov} [m^3]; and overflow flow, Q_{Ov} [l/s]. Three other output variables are related with water quality in terms of COD: overflow COD load $B_{COD,Ov}$ [kg]; overflow COD average concentration, $C_{COD,Ov,Av}$ [mg/l]; overflow COD maximum concentration, $C_{COD,Ov,Max}$ [mg/l]. The remaining three output variables are also related with water quality but in terms of NH_4: overflow NH_4 load, $B_{NH4,Ov}$ [kg]; overflow NH_4 average concentration, $C_{NH4,Ov,Av}$ [mg/l]; and overflow NH_4 maximum concentration $C_{NH4,Ov,Max}$ [mg/l].

3.2 Model input uncertainty assessment

To assess the sensitivity of the model output to small changes in the model input, simulations with EmiStatR for each input variable with an increment and decrement of 10% with relation to the deterministic values (Tables 1 and 2) were performed. Basically, the variables $Aimp$, Qd, and V are the most sensitive variables of water quantity variables ($VTank$, V_{Ov} and Q_{Ov}). Regarding water quality in terms of COD the input variables $CODs$, $CODr$, $Aimp$, Qd, V, and P have greatest impact on output COD variables, whereas input variables qs, $NH4s$, qf, $NH4r$, $Aimp$, pe, Qd, V and P have the greatest impact on output NH_4 variables.

After evaluation of the model output sensitivity to input variables, and taking into account the degree of uncertainties of each input, we selected three input variables to be included in the uncertainty analysis: $CODs$, $NH4s$, and $CODr$. We leave the uncertainty propagation

Variable	BUD	DAH	ESO	GOE	HEI	KAU	NOC	NOR
Sub-catchment data								
Land use [-][1]	R/I	R/I	R/I	R/I	R/I	R/I	R/I	R/I
Total area [ha]	16.5	16.2	22.2	16.5	30.0	22.0	16.2	18.6
Impervious area [ha]	4.9	5.0	11.1	7.6	11.0	11.0	3.4	4.3
Population equivalents [PE]	289	460	705	611	676	358	260	326
Flow time structure [min]	10	10	7	10	6	10	10	10
Structure data								
Throttled outflow [l/s]	6	4	12	9	11	9	2	4
Volume [m^3]	90	270	330	190	220	180	166	157

[1]R = residential, I = industrial

Table 2: General input variables of the CSO structures of the EmiStatR test model.

analysis of $Aimp$ and P, which also score high on the sensitivity-uncertainty ranking, for future work.

The field measurements were the basis to characterise input uncertainty of $CODs$ and $NH4s$. Samples of COD and NH_4 in mg/l (91 in total for each variable) were analysed in the dry weather flow produced in the villages of Goesdorf, Kaundorf and Nocher-Route. An average wastewater amount was calculated for Goesdorf (153 l/(PE·d)), Kaundorf (112 l/(PE·d)) and Nocher-Route (94.3 l/(PE·d)). Table 3 presents the summary statistics of the dry weather flow measurements of COD and NH4 and the correspondent value of $CODs$ and $NH4s$. We use italic font for $CODs$ and $NH4s$ to emphasize that these are input variables of the model.

	COD [mg/l]	$CODs$ [g//(PE·d)]	**NH4** [mg/l]	$NH4s$ [g//(PE·d)]
Min	62	7	16.1	1.7
Mean	926	104	44.4	4.7
Max	3454	528	81.2	10.8
St. deviation	632	88	18.6	1.9

Table 3: Summary statistics of dry weather flow measurements for $CODs$ and $NH4s$ characterization.

Regarding $CODr$, no field measurements were available. Thus, expert judgement and values from the literature were used to characterise input uncertainty in $CODr$. A mean value of 80 mg/l and a standard deviation of 90 mg/l were used. For all three input variables, $CODs$, $NH4s$, $CODr$, we proposed a normal distribution to characterise input uncertainty with mean and standard deviation for each variable as defined above. In order to avoid negative values, we first transformed the variables by taking their natural logarithm.

3.3 Uncertainty propagation

We made a deterministic run of the model with values of the input variables as presented in Tables 1 and 2. Additionally, we performed 1,500 Monte Carlo simulations allowing $CODs$, $NH4s$,and $CODr$ as stochastic input variables with characteristics as defined in the previous section. In this way the total uncertainty of output variables due to input uncertainty was calculated. Figure 3 shows an example of model output and the total uncertainty band of 5 and 95 percentile for overflow concentration of COD (centre inset) and overflow concentration of NH_4

(bottom inset). In case of a rain event that produces a CSO, the uncertainty in the model output is quite large. Also, there is a systematic difference between the deterministic and the median run. The latter is always slightly above the deterministic run.

Next the contributions of each one of the input variables to the total uncertainty was calculated as well, by calculating the difference between the total uncertainty and the uncertainty obtained in the stochastic simulation of the other two variables. For instance, the uncertainty contribution of $CODs$ was calculated as the total uncertainty minus the uncertainty of the simulations running only $NH4s$ and $CODr$ in stochastic mode. Therefore, 4,500 additional Monte Carlo simulations were done to calculate the uncertainty contribution of the three input variables.

The results indicate that there is no uncertainty contribution of $CODs$, $NH4s$ and $CODr$ to output variables $VTank$, V_{Ov}, and Q_{Ov}. $CODs$ and $CODr$ contribute to uncertainty in $B_{COD,Ov}$ and $C_{COD,Ov,Av}$. $CODr$ has the most important uncertainty contribution to $B_{COD,Ov}$ (99.6%) and $C_{COD,Ov,Av}$ (90.4%). Finally, $NH4s$ contributes 100% of the uncertainty of $B_{NH4,Ov}$ and $C_{NH4,Ov,Av}$.

IV CONCLUSIONS AND FUTURE WORK

We applied an uncertainty analysis of the simplified urban drainage model EmiStatR. A characterisation of the input uncertainty of the three main input variables that control output uncertainty in water quality variables was done. We found that the uncertainty in the load of COD per capita per day in the sewage ($CODs$) and the concentration of COD in runoff ($CODr$) contribute to the uncertainty of the output variables: overflow COD load and overflow COD average concentration. $CODr$ has the most important uncertainty contribution in the load and concentration of COD. The load of NH_4 per capita per day in the sewage ($NH4s$) contributes totally in the uncertainty of overflow NH4 load and overflow COD average concentration.

Rainfall is one of the most important drivers in the definition of uncertainty of output variables as load and concentration of COD and NH_4. Therefore a detailed analysis should be done to quantify the contributions of rainfall input uncertainty in the total uncertainty. The rainfall uncertainty will be addressed in follow-up research as well as uncertainty due to impervious areas.

A further development of EmiStatR allows to analyse spatial model inputs. In this sense, a semi-distributed modelling framework can be implemented and several sub-catchments of the system modelled simultaneously, taking into account the inherent spatial variability of the inputs as rainfall and impervious areas and propagate the uncertainty of such inputs through the model. This also will be worked out in future research.

ACKNOWLEDGEMENTS

The work presented is part of the QUICS (Quantifying Uncertainty in Integrated Catchment Studies) project. This project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 607000 and the Luxembourg Institute of Science and Technology.

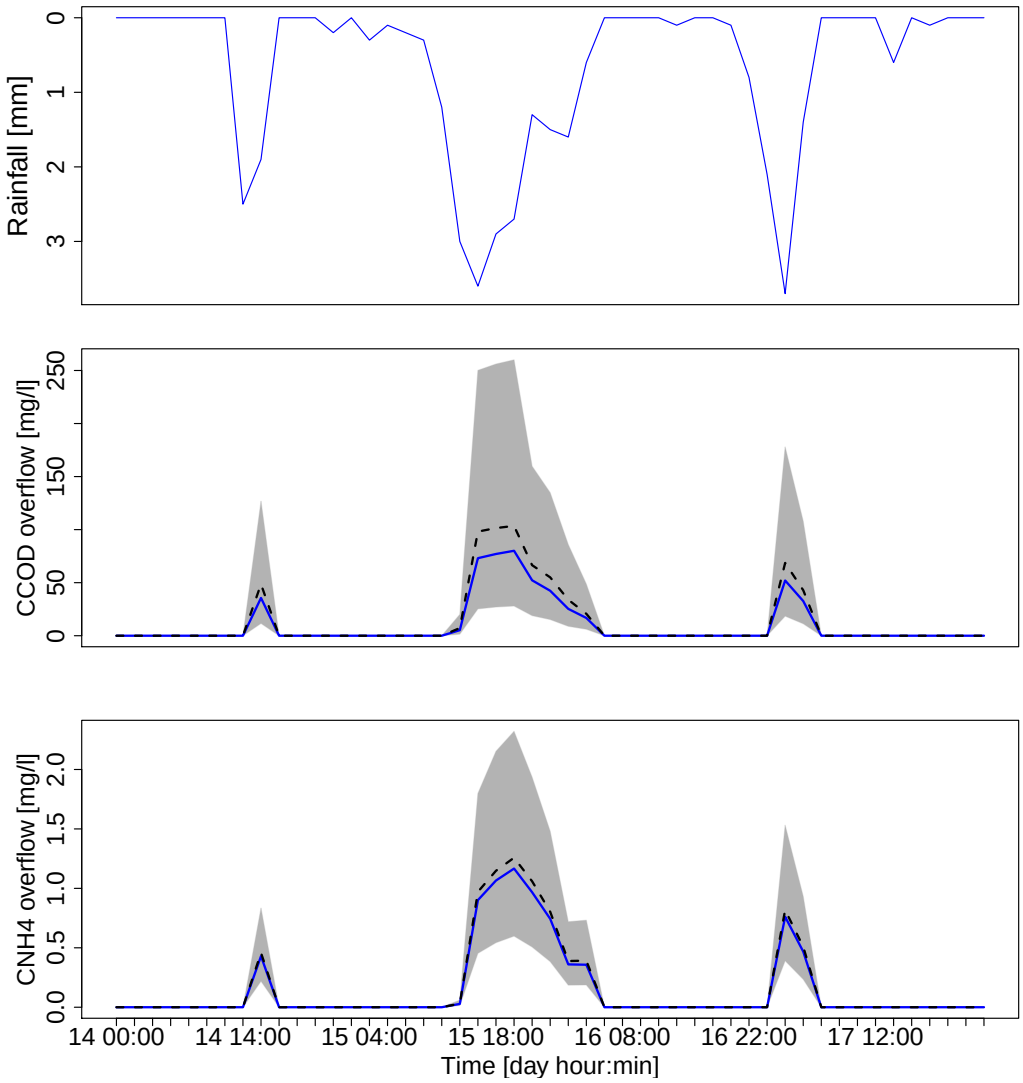


Figure 3: Precipitation, and overflow COD and NH_4 concentrations in deterministic and uncertainty outcomes. Location: Goesdorf. From 14 to 18 August 2010. Inset centre and bottoom: the blue line indicates the median, the black dotted line indicates the deterministic run, and the gray area indicates the 90 per cent uncertainty band.

References

Bach P. M., Rauch W., Mikkelsen P. S., McCarthy D. T., Deletic A. (2014). A critical review of integrated urban water modelling - Urban drainage and beyond. *Environmental Modelling & Software 54*, 88 – 107.

Brown J. D. (2004). Knowledge, uncertainty and physical geography: Towards the development of methodologies for questioning belief. *Transactions of the Institute of British Geographers 29*(3), 367–381.

Deletic A., Dotto C., McCarthy D., Kleidorfer M., Freni G., Mannina G., Uhl M., Henrichs M., Fletcher T., Rauch W., Bertrand-Krajewski J., Tait S. (2012). Assessing uncertainties in urban drainage models. *Physics and Chemistry of the Earth 42-44*, 3–10.

Grigoriu M. (2012). *Stochastic Systems: Uncertainty Quantication and Propagation*. London: Springer-Verlag.

Heuvelink G. B. M., Brown J. D., van Loon E. E. (2007). A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science 21*(5), 497–513.

Klepiszewski K., Seiffert S. (2013). Statistische Erfassung von Entlastungsbauwerken der Mischwasserbehandlung im Einzugsgebiet der Chiers. MIGR EMISTAT-MW. Technical report, TUDOR Centre de Ressources des Technologies pour l'Environnement, Luxembourg.

Mckay M. D., Beckman R. J., Conover W. J. (1979). A Comparison of Three Methods for Selecting Values of

Input Variables in the Analysis of Output from a Computer Code. *American Statistical Association and the American Society for Quality*.

Minasny B., McBratney A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences 32*(9), 1378–1388.

Mitchell V., Duncan H., Inman M., Rahilly M., Stewart J., Vieritz A., Holt P., Grant A., Fletcher T., Coleman J., Maheepala S., Sharma A., Deletic A., Breen P. (2007). State of the art review of integrated urban water models. In *Novatech 2007*, pp. 507–514.

Nash J. E., Sutcliffe J. E. (1970). River flow forecasting through conceptual models. Part 1 - a discussion of principles. *Journal of Hydrology (Amsterdam) 10*, 282–290.

Neumann M. B. (2007). *Uncertainty Analysis for Performance Evaluation and Design of Urban Water Infrastructure*. Ph. D. thesis, Swiss Federal Institute of Technology, ETH Zurich.

Nol L., Heuvelink G. B. M., Veldkamp a., de Vries W., Kros J. (2010). Uncertainty propagation analysis of an N2O emission model at the plot and landscape scale. *Geoderma 159*(1-2), 9–23.

Refsgaard J. C., van der Sluijs J. P., Hø jberg A. L., Vanrolleghem P. A. (2007). Uncertainty in the environmental modelling process - A framework and guidance. *Environmental Modelling and Software 22*(11), 1543–1556.

Sanitary-District (2015, May). *Combined Sewer Overflow*. Internet.

Schellart A. N. A., Tait S. J., Ashley R. M. (2010). Towards quantification of uncertainty in predicting water quality failures in integrated catchment model studies. *Water Research 44*(13), 3893–3904.

van Keur P., Henriksen H. J., Refsgaard J. C., Brugnach M., Pahl-Wostl C., Dewulf A., Buiteveld H. (2008). Identification of major sources of uncertainty in current IWRM practice. Illustrated for the Rhine Basin. *Water Resources Management 22*(11), 1677–1708.

Walker W., Harremoës P., Rotmans J., van der Sluijs J., van Asselt M., Janssen P., Krayer von Krauss M. (2003, March). Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integrated Assessment 4*(1), 5–17.

Zoppou C. (2001). Review of urban storm water models. *Environmental Modelling and Software 16*(3), 195–231.

How can big data be used to reduce uncertainty in stormwater modelling?

Nanée Chahinian*[1], Anne-Laure Piat-Marchand[2], Sandra Bringay[2], Maguelonne Teisseire[3], Elodie Boulogne[4], Laurent Deruelle[5], Mustapha Derras[5], Carole Delenne[4]

[1]IRD, HSM (CNRS, IRD, UM), France
[2]Univ. Montpellier 3, LIRMM (CNRS, UM), France
[3]TETIS, IRSTEA, France
[4]Univ. Montpellier, HSM (CNRS, IRD, UM), France
[5]Berger-Levrault, France

*Corresponding author: nanee.chahinian@ird.fr

Buried utility networks are often not fully documented and inaccurate positioning of network elements may impact hydraulic modelling applications. This study aims to put forward a methodology to extract network characteristics from data that is posted on the web or is available through sector specific databases, using heterogeneous text scrapping methods.

I INTRODUCTION

Urban growth is an ongoing trend and one of its direct consequences is the development of buried utility networks. With growing needs among consumers, new networks are being installed and more underground space is being occupied. Locating these networks is becoming a challenging task. Mispositioning of utility networks is an important problem for both industrialised and developing countries and will worsen as cities expand and their networks increase in size and complexity (Jamil et al., 2012; Metje et al., 2007). Over the past century, it was common practice for public service providers to install, operate and repair their networks separately (Rogers et al., 2012). Now local authorities are confronted with the task of combining data produced by different parties, having distinct formats, variable precision and granularity (Chen and Cohn, 2011). Although in certain countries contractors are now obliged by law to position all buried networks within set precision ranges, finding data related to older network branches is a cumbersome task. Once located these data are often unavailable at the desired precision or are prone to errors or omissions. Since the mid 90's, an increasing volume of data is posted on the web or is available through sector specific databases. An alternative and complementary approach to field surveys would be to track down archived data by using new methods of heterogeneous texts scrapping. Hence the objective of this work is to assess whether big data can be used to reduce uncertainty in stormwater modelling. The big data used in this study is original information scrapped from the web such as calls for tenders, newspaper articles, consumer complaints, etc. Information extracted with text mining techniques such as used in Kergosien et al. (2015) are particularly interesting to confirm or infirm the position, the depth, the material of buried network elements. Call for tenders often include additional technical descriptions (slope, local constraints, junctions, etc.) which may be used to build an attribute table, that contains the characteristics of the underlying pipes.

This study is a part of a global project which aims to recreate a stormwater and a sewage network in settings where no accurate information regarding the position or characteristics of buried utility networks is available. A previous study (Pasquet et al., 2016) put forward a methodology to detect manhole covers and inlet grates from aerial images. The results were encouraging but dealt only with the correct detection of the objects *i.e.* no information regarding the underlying

network was supplied. Positional errors of remotely sensed manhole covers and grates could lead to errors on rim elevation, which in turn impacts the pipe length and slope. Detection errors such as *i)* objects masked in the aerial photograph by trees or shade, *ii)* spots wrongly assumed to be manhole covers or grates, or even *iii)* visible but out of service elements, may strongly impact the shape of the network. The pipe characteristics (material, diameter and roughness) may also affect the output hydrographs. Consequently, it is important to assess the sensitivity of the hydraulic model's results to the network descriptors in order to determine the most relevant parameters, and the precision required when determining their values.

The paper is structured as follows. Section II describes the materials and methods including study site, the tools and methodologies used for the sensitivity analysis and the textual analysis. Section III and section IV present the main results of our study and the new perspectives that this preliminary analysis opens.

II MATERIALS AND METHODS

2.1 Study site and data

The study site is located in the "Hôpitaux Facultés" district of the city of Montpellier (Southern France), where heavy rains often cause over flooding of stormwater drains. It is an extension of the University of Montpellier's main campus which has already been used to test the manhole cover detection techniques proposed in a previous study (Pasquet et al., 2016).

The database consists of digital orthophotos at 16 cm resolution available through the city's open data platform; digital data on the wastewater network extracted from the GIS of Montpellier Métropole, the inter-municipality agency of the metropolitan area of Montpellier and data on the stormwater network supplied in digital format by the Municipality of Montpellier. A quick display of the three data layers showed that the manhole covers and grates located on the university campus were not reported in the city's GIS systems. A special request was made to the university's assets' management department which provided PDF files of the network maps that were georeferenced in QGis. Manhole covers and grated were digitised as point data, the water pipes as lines and all the descriptive data written in the PDF were entered manually in the attribute table.

2.2 Methodology

2.2.1 Hydraulic modelling and sensitivity analysis

The network built in the previous step was imported into the PCSWMM (Computational Hydraulics International) software for stormwater modelling. PCSWMM is built on the US EPA's SWMM5 engine and uses the Saint-Venant (1871) equations to model water flow.

The upstream catchment characteristics are determined based on the Municipality's dataset. The catchment is divided into three sub-catchments. Two synthetic storms are generated to run the model: a 2-hour duration Chicago design storm (Keifer and Chu, 1957) is used as a reference rainfall event and a Desbordes (Desbordes, 1974) double-triangle design storm with a 5 year return period is used to account for more intense events. No model calibration is undertaken. In order to determine the most important information to be searched for by data mining, sensitivity analyses are undertaken for the following elements:

1. **Slope.** In the framework of this project, rim elevation may be easily determined via remote sensing data (DEMs, airborne Lidar, etc). However, this does not mean that the inlet elevation will be known. If no further information is provided, pipe slopes should

deduced from rim elevation. Two tests are thus carried out to determine slope values: *i)* a constant slope is applied on the entire network; *ii)* the pipe slope is assumed to be equal to the terrain slope with a zero value in case of a counter slope. Figure 1 shows the profiles corresponding to the PCSWMM simulations (see section III).

2. **Network shape.** To test the influence of the missing elements, such as grates that are invisible on orthophotos or that are not reported in databases, the network shape is over-simplified (Figure 2).
3. **Roughness.** Pipes are assumed to be made of either PVC or concrete.
4. **Pipe diameter.** The diameter values are either increased or decreased by 30 and 50%.

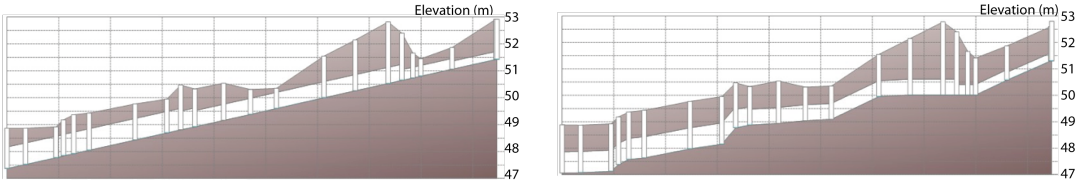


Figure 1: Recreating slope values.

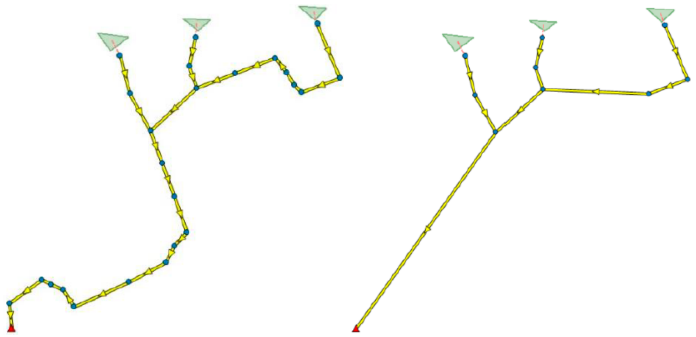


Figure 2: Network reconstruction. a) Real network b)Simplified network

2.2.2 Textual Analytics

Our process explore textual resources scrapped from the web to confirm or infirm information obtained from hydrologic modelling and sensitivity analysis. Our process is divided into 3 steps as shown in Figure 3. The first step is to gather a set of relevant documents. Once this corpus is built, textual documents are analyzed in two phases. The second phase aims to identify thematic, spatial and temporal features. The third step is dedicated to the linkage of these features.

For the first step, we manually gathered 30 documents identified as relevant. This set of documents is composed of one local planning map, one intervention report, three annual reports of public service concession holders, three technical notes, two prevention plans against natural floods, one scientific paper, one news article, two commercial brochures, four sewerage master plans, two municipal newsletters, one public inquiry file, one file of administrative acts, three study reports, a list of service delivery's contracts concluded, one convention document, three technical specifications. Six documents were identified as irrelevant.

For the second step, we used two different methods so as to identify three types of expressions: thematic, temporal and spatial, in our set of documents.

Our chain of treatments which aims to recognize spatial and thematic information is based on Unitex[1], which is a free collection of programs dedicated to corpus processing based on linguis-

[1]Unitex has been developed at LADL (Laboratoire d'Automatique Documentaire et Linguistique) http://www-igm.univ-mlv.fr/ unitex/

tic resources. Resources used by the software consist of: *i)* Dictionaries: description of simple and compound words with their lemmas and a set of grammatical and semantic codes; *ii)* Grammars: representations, as graphs, of linguistic phenomena on the basis of finite state automata; *iii)* Lexicon-grammar tables: matrices describing syntactical properties of some words in which rows correspond to verbs and columns to syntactical properties. We applied the transducers's cascade "CasEN_Quaero" which has been created in order to recognize named entities (Maurel et al., 2011). A transducer is a set of boxes which contains symbols (*e.g.* characters' sequences, a syntactical code) that we aim to recognize. Each graph in a cascade is applied, one after the other, to the corpus which has been modified by the previous graph (Friburger, 2002).

The recognition of temporal features was carried out using Heideltime[2]. Heideltime is a rule-based tagger that retrieves temporal informations (absolute and relative dates, durations) from texts. Expressions identified are normalized according to the TIMEX3 annotation standard[3] which describe temporal expressions, dates...).

Finally, the second phase of the process is then implemented as follows (Fig. 3):

- **Thematic identification:** we established a list of words related to the field of interest, composed of 77 words which describe objects such as manhole covers and stormwater or wastewater networks, and 62 terms that depict their characteristics (*e.g.* diameter, depth, flow) and measurement units (*e.g.* m 3/s , Eq-hab, m). We completed this list with terms extracted from two lexicons (UNESCO, 1992; Andréassian et al., 2002) and one reference work (Bourrier, 2008). This lexicon covers 1% of our corpus (percentage of annotated documents in the corpus). The automatic extraction of synonyms from 7 websites[4] provided a list of 1721 terms which had been projected on texts retrieved in step 1. Finally, 683 appeared in the corpus, thus expanding the initial list to 344%. The lexicon extended this way covers 4% of our set of documents.
- **Spatial identification:** we used resources embedded in the CasEN_Quaero cascade of the Unitex software in order to recognize spatial information (*e.g.* "au nord de la route R12 allant de Montpellier à Lunel", "le boulevard du Languedoc").
- **Temporal identification:** we chose to use Heideltime so as to detect time expressions (*e.g.* "l'appel d'offre signé du 12 mai", "du 03/06/14 au 21/07/14").

The chain of treatments was evaluated by one annotator on a sample of 3 documents which contains 21 904 tokens. Thus, 285 spatial features, 213 temporal expressions and 346 thematic features were identified.

The third phase, corresponding to the linkage operation, will link the spatial and temporal features obtained in the previous phase with the thematic marking (Fig. 3). Depending on their degree of proximity (in the document, in the section, in the paragraph or in the sentence), spatial and temporal markings will be retained or not.

III RESULTS

When no information is available on the inlet elevations or pipe slopes their values should be deduced from rim elevation data. The results obtained with the two approaches for this estimation are presented in Fig. 4, for the moderate and intense rainfall event. It can be seen that when assuming similar ground and underground slopes, the model underestimates both flood

[2]http://dbs.ifi.uni-heidelberg.de/index.php?id=129

[3]http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/timex3guidelines-072009.pdf

[4] http://www.cnrtl.fr, http://www.synonymo.fr, http://www.crisco.unicaen.fr, http://www.babla.fr, http://dico.isc.cnrs.fr, http://dictionary.reverso.net, http://dictionary.sensagent.com

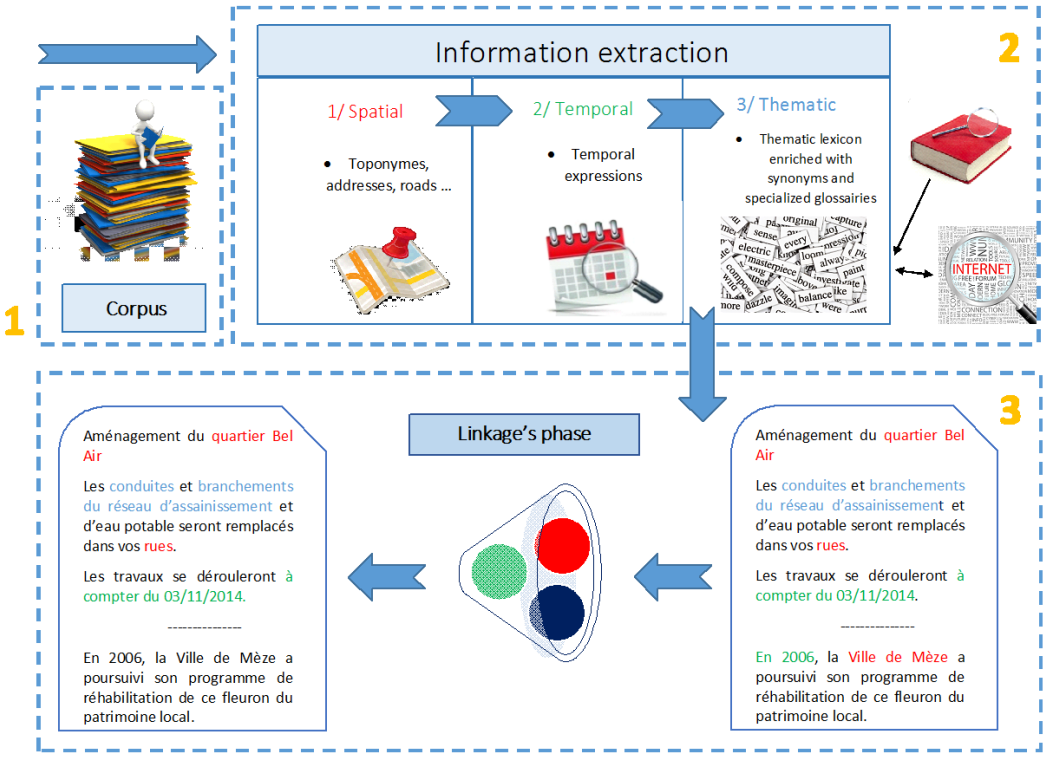


Figure 3: Global process of textual analysis.

volume and peakflow. Indeed, imposing a null gradient for counter slopes yields overflows and thus diminishes the routed volume. For these events using a mean, constant slope does not induce great differences on the simulated hydrographs: changing the slope by up to 43% yields comparable results. The two values tested in this case correspond to either the mean slopes given by the real database or to the difference between upstream and downstream rim elevations.

The results are further confirmed for the high intensity rainfall event. However, slight differences can be seen based on the selected slope value; one yields a better estimation of the flood volume while the other one's results are closer to the flood dynamics. This raises questions about the slope value that should be used for network reconstruction. Although French technical guides recommend using a 0.01 m/m slope when designing stormflow networks, it seems that common practice, especially when extending existing networks, is to set the slope according to the input and output gradients. For a given pipe burial depth, this would imply dividing the network into distinct sections to avoid counter slopes or unrealistic elevation values.

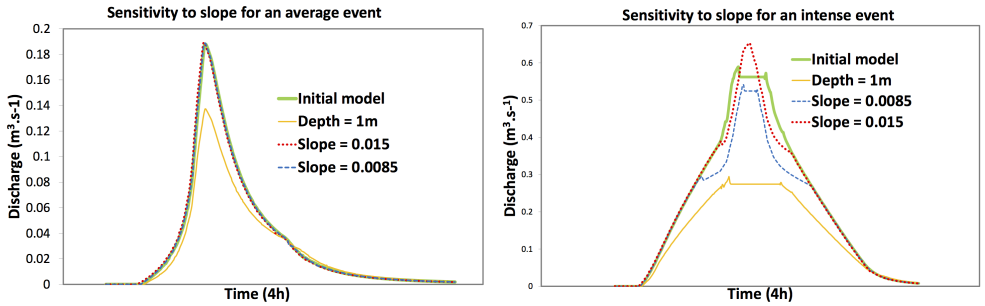


Figure 4: Sensitivity to slope for an average event and an intense event. The slope 0.015 correspond to the mean of all the real slope values, and 0.0085 is directly the slope between upstream and outlet.

Regarding the network's shape (Fig. 5) no real influence can be seen when modelling moderate intensity rainfall events. However, for heavier rainfall events, missing out inlet grates and consequent branching yields faster transfer times and upstream overflowing (Fig. 2).

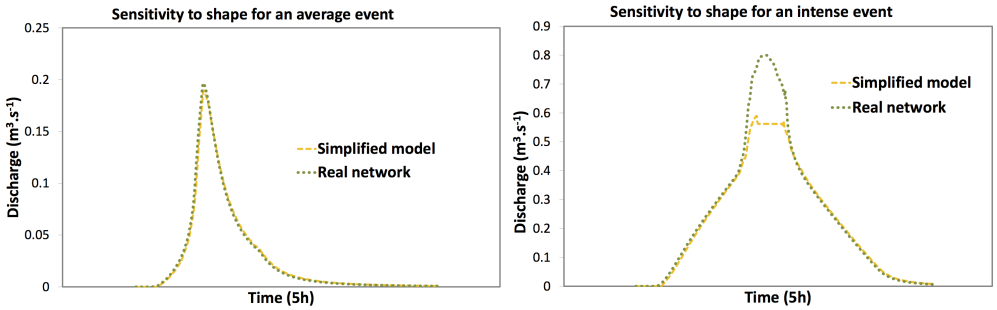


Figure 5: Sensitivity to missing inlet grates

For moderate rainfall, the roughness of the pipe material does not impact the output hydrograph. In case of high intensity events, high roughness may reduce flow velocity and yield greater over-flooding. However, roughness is often a calibrated parameter as its value cannot be easily determined by direct measurement. Consequently, lack of information on roughness can easily be overcome.

Finally, pipe diameter has nearly no influence on the output discharge as long as there is no transition from open-channel to closed-conduit flow. A significant reduction in pipe diameter will lead to over flooding.

Table 1 gives a ranking of the tested parameters in terms of impact on simulated flow hydrographs. Elevation and slope are the parameters which need to be estimated with the best accuracy. The aim of the textual analysis would be to identify any mention of these parameters and corresponding values in big data.

Rank	Parameter	Justification
1	Pipe slope	Faster transfer times if too steep
2	Pipe diameter	May cause overflowing if too small
3	Pipe roughness	Should be calibrated.

Table 1: Parameter ranking based on the sensitivity analysis

The first phase applied on the corpus allows the extraction of spatial, temporal and thematic features as illustrated in Figure 6. It can be seen that the program can accurately identify the name of the town ("ville de Vendome"), the timeframe of the document ("depuis 1994"; "l'automne 2014"). By using the thematic pattern, the type of wastewater network ("réseau séparatif"-separate wastewater network) and the capacity of the wastewater treatment plant ("35000 équivalents/habitants") are also identified. However, the text also indicates that 98% of the wastewater network is routed through a separate network but the information is not picked up by the program. To overcome this, a new pattern could be added but the selection rules cannot be extended endlessly. A rate of "acceptable loss" should be set with the specialists. The quantitative results of the feature extraction are summarised in Table 2.

Regarding the results of the sensitivity analysis we carried out, with the exception of the rim elevation, the parameters correspond to terms which are too generic, with a wide variety of uses. "Pipe diameter" ("diamètre de conduite") for instance, was mentioned in reference to karstic

	Spatial features	Temporal features	Thematic features
False positive	35	21	14
Missing	124	22	34
Correct	285	213	346
Recall	0.70	**0.91**	0.91
Precision	**0.89**	**0.91**	**0.96**
F-Measure	0.78	**0.91**	0.94

Table 2: Quantitative Evaluation of the feature extraction.

aquifers, water supply and flooded caves. "Pipe diameter" pointed also to supply catalogs. The term "slope" was the one that needed most disambiguation. "The general slope of the terrain", "Average catchment slope" and "sloping" ("en pente" in French) are very frequently used. Pipe slope values which are not positioned closely to the words "pipe" and "slope" are often missed. As a general rule, quantified information related to the network's slope can not be easily detected because often there are many intermediate words between them and many expressions are used when reporting this values. This also applies to "Rim elevation". In order to obtain quantitative information about these parameters, it is necessary to overcome this problem. It would be necessary to investigate other annotation techniques such as the one described in (Berrahou et al., 2013).

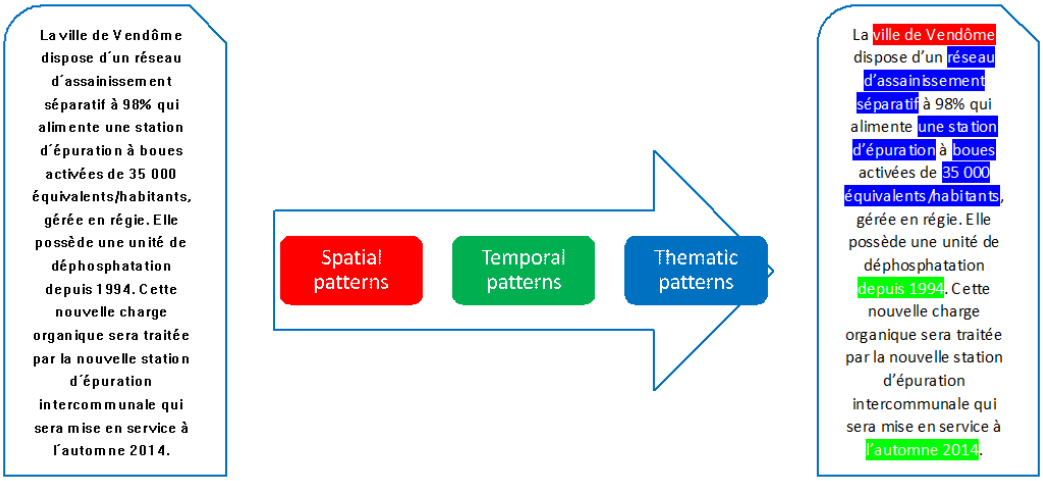


Figure 6: Example of an annotated text

IV CONCLUSION AND PERSPECTIVES

A methodological framework was put forward to use big data in order to improve the accuracy of stormwater modelling networks. A sensitivity analysis was carried out to determine the most sensitive parameters and text mining techniques were used to extract information about them in big data scrapped from the web. The primary results are encouraging as spatial, temporal and thematic patterns are correctly identified. However, quantitative information about the network's elements such as elevation and slope can be missed sometimes. Future work will look into overcoming this problem by either setting new patterns or creating an index which will give the expert an indication on the presence of possibly interesting numerical data in a document. The final step will consist in building a model based solely on the information gathered through the big data. The uncertainty associated with each characteristic will be assessed and hydraulic simulations using a classical modelling software will be carried out to assess the uncertainty

transmission in each step of the process until the simulation of the output hydrograph. Other types of documents could also be considered such as photos or scanned plans with captions. Automatic scrapping from the web based on research engines such as Google or Yahoo and contextual analysis of snippets will also be explored (Opitz et al. (2014)).

Acknowledgment
This study is part of the project "Cart'Eaux" funded by the European Regional Development Fund (ERDF).

References

Andréassian V., Sarkissian V., Chelmiki W., Stanescu V., Moussa R. (2002). Cemagref.

Berrahou S. L., Buche P., Dibie-Barthélemy J., Roche M. (2013). How to extract unit of measure in scientific documents? In K. Liu, A. L. N. Fred, and J. Filipe (Eds.), *KDIR/KMIS 2013 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing, Vilamoura, Algarve, Portugal, 19 - 22 September, 2013*, pp. 249–256. SciTePress.

Bourrier R. (2008). *Les réseaux d'assainissement Calculs - Applications - Perspectives*. Paris: Editions Tec et Doc/ Lavoisier.

Chen H., Cohn A. (2011). Buried utility pipeline mapping based on multiple spatial data sources: A bayesian data fusion approach. In *IJCAI-11, Barcelona, Spain*, pp. 2411–2417.

Desbordes M. (1974). *Réflexions sur les méhodes de calcul des réseaux urbains d'assainissement pluvial*. Montpellier: PhD Thesis USTL.

Friburger N. (2002). *Reconnaissance automatique des noms propres : application à la classification automatique des textes journalistiques*. Tours: PhD Thesis University of Tours.

Jamil H., Z. N., Mohid Yussof M. (2012). Underground utility mapping and its challenges in malaysia. In *FIG working week 2012. Knowing to manage the territory, protect the environment, evaluate the cultural heritage*, Rome, Italy, pp. 15.

Keifer J., Chu H. (1957). Synthetic storm patterns for drainage design. *Journal of Hydraulic's division* (83), 1–25.

Kergosien E., Alatrista-Salas H., Gaio M., Güttler F., Roche M., Teisseire M. (2015). When textual information becomes spatial information compatible with satellite images. In *KDIR*, pp. 301–306.

Maurel D., Friburger N., Antoine J., Eshkol-Taravella I., Nouvel D. (2011). Cascades autour de la reconnaissance des entités nommées. *TAL* (52), 69–96.

Metje N., Atkins P., Brennan M., Champan D., Lim H., Machell J., Muggleton J., Pennock S., Ratcliffe J., Redfern M., Rogers C., Saul A., Shan Q., Swingler S., Thomas A. (2007). Mapping the underworld: State of the art review. *Tunnelling and underground space technology* (22), 568–586.

Opitz T., Azé J., Bringay S., Joutard C., Lavergne C., Mollevi C. (2014). Breast cancer and quality of life: Medical information extraction from health forums. In C. Lovis, B. Séroussi, A. Hasman, L. Pape-Haugaard, O. Saka, and S. K. Andersen (Eds.), *e-Health - For Continuity of Care - Proceedings of MIE2014, the 25th European Medical Informatics Conference, Istanbul, Turkey, August 31 - September 3, 2014*, Volume 205 of *Studies in Health Technology and Informatics*, pp. 1070–1074. IOS Press.

Pasquet J., Desert T., Bartoli O., Chaumont M., Delenne C., Subsol G., Derras M., Chahinian N. (2016). Detection of manhole covers in high-resolution aerial images of urban areas by combining two methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9*(5), 1802–1807.

Rogers C., Hao T., Costello S., Burrow M., Metje N., Chapman D., ..., Saul A. (2012). Condition assessment of the buried utility service infrastructure: a proposal for integration. *Tunnelling and Underground Space Technology 28*, 202–211.

UNESCO (1992). *Glossaire international d'Hydrologie*.

Sensitivity analysis of spatio-temporal models describing nitrogen transfers, transformations and losses at the landscape scale.

Jordi Ferrer Savall[1*], Cyril Benhamou[1], Pierre Barbillon[2], Patrick Durand[3], Marie-Luce Taupin[4], Hervé Monod[2], Jean-Louis Drouet[1]

[1] UMR ECOSYS, INRA, AgroParisTech, Université Paris-Saclay, 78850, Thiverval-Grignon, France
[2] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France.
[3] UMR SAS, INRA, Agrocampus Ouest. 84215, Rennes, France
[4] UMR MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France

*Corresponding author: *jordi.ferrersavall@agroparistech.fr*, *jordi.ferrer.savall@gmail.com*

Abstract

Modelling complex systems such as agroecosystems often requires the quantification of a large number of input factors. Sensitivity analyses are useful to assess the impact of imput factors on model outcomes and to determine the appropriate spatial and temporal resolution of models. Comprehensive spatial and dynamic sensitivity analyses were applied to the model Nitroscape, a deterministic spatially distributed model describing nitrogen transfers and transformations in rural landscapes, simulating five years of farm management in an intensive rural area of 3 km2. 11 global imput factors were explored, characterizing spatial resolution of the simulations, the physical parameters of the landscape and the agricultural practices. Their impact on 29 spatially-distributed model outcomes was assessed through standard analysis of variance. Cluster analyses were applied to summarize the results of the sensitivity analysis on the ensemble of model outcomes. The purpose of this paper is to present the methods and preliminary results that we developed to carry out and to visualize comprehensive spatial and dynamic sensitivity analyses of multiple input factors on multiple outcomes.

Keywords
Please insert a few keywords here

Modeling nitrogen transfers, transformations and losses from processes integration at the landscape scale requires using complex models which are spatially and temporally explicit. Model predictions may vary considerably, depending on the initial conditions and input parameters. Therefore, it is important to characterize their robustness and variability through sensitivity analyses.

A variety of techniques have been developed to perform sensitivity analysis in spatially and dynamic models at different stages. Methods for exploring model inputs may range from the simplest one-factor-at-a-time screening techniques proposed by Morris (1991) to the exploration of representative combinations of factors in factorial designs, as described by Chen (2011). Recently, Moreau et al. (2013), proposed a method to take into account the spatial distribution and the granularity of input parameters in sensitivity analyses. Besides, Marrel et al. (2011) proposed a technique to extend variance-based methods well suited to analyse scalar outputs to dynamic and spatially explicit outcomes, rendering spatial maps of the sensitivity of each output to each model input. Finally, Ligmann-Zielinska (2013) showed that spatially explicit sensitivity analyses and analyses that aggregate outputs at different scales of description are complementary, but at the same time can lead to different decisions regarding input factor prioritization.

Here, we present an analysis of sensitivity of the NitroScape model developped by Duretz et al. (2011), with the specific aim to compare the impact of spatial granularity of the model to the impact of global parameters describing physical characteristics of the landscape and agricultural practices on spatially-distributed model outcomes. A central concern of the presented work is to provide tools for integrating the results of spatially

explicit sensitivity analyses applied on multiple model outcomes.

Nitroscape is a deterministic, spatially explicit and dynamic model describing the transfers and transformations of reactive forms of nitrogen (Nr) in rural landscapes of a few km^2 to a few tenth of km^2. It couples four modules characterizing processes of farm management, biotransformations and transfers of Nr by the atmospheric and hydrological pathways. It simulates Nr flows and losses within and between several landscape compartments: the atmosphere, the hydro-pedosphere (groundwater, water table and streams) and the terrestrial agroecosystems (livestock buildings, croplands, grasslands and semi-natural areas).

In order to evaluate the impact of global model inputs on the model outcomes, 11 parameters were selected. They characterized the spatial resolution of the model (size of the horizontal grid and depth of surface layers), the physical features of the landscape (soil characteristics) and the agricultural management (fertilization). A complete fractional factorial design with 11 categorical factors and 3 levels per factor was generated. The resulting plan comprised 243 units and the design resolution was 5, which allowed estimating main effects and two-factor interactions unconfounded.

The range of explored input values represented both the range of values observed in real settings and the degree of uncertainty in the measured or estimated values. Landscape structure (*e.g.* topography, land use), initial conditions (*e.g.* soil mineral content, agroecosystem characteristics) and forcing inputs (*e.g.* agricultural practices, meteorology) were common for all units in the factorial design. Their values were set from real measurements to represent an intensive rural area with mixed crops, pig farming and unmanaged ecosystems characterized by humid climatic conditions and little temperature contrasts.

Each run of the plan simulated nitrogen transfers, transformations and losses in a virtual landscape representing 300 ha of a catchment area with agricultural and farming activity, during 5 years, with daily outputs at the outlet and monthly outputs on the whole grid after an initialization period of two years. The impact of model inputs was evaluated on 29 model outcomes: 5 variables describing the catchment outflow, 9 spatially-distributed variables describing inter-compartment fluxes and 15 spatially-distributed variables describing the local state of the system.

NitroScape spatially-distributed output variables were aggregated to obtain either time-series describing spatially aggregated outcomes or maps of temporally aggregated outcomes. A principal component analysis (PCA) was applied to time series and spatial distributions to reduce their dimensionality. The influence of factors on each model outcome (expressed as aggregated data or in terms of its principal components) was explored through a standard analysis of variance (ANOVA) considering up to second-order interactions: the variability of each component of each outcome (day in time series, pixel in maps, or principal component), spread over the experimental design, was adscribed to every factor and two-factor interaction by partitioning the sum of squared deviations. The relative importance of factors and interactions was measured with sensitivity indices based on this partitioning. Finally, a cluster analysis and a PCA were applied to summarize the ensemble of results of the sensitivity analysis on the whole set of model outcomes, in order to identify outcomes with similar response to model inputs and to discriminate the effect of each factor and two-factor interaction on the ensemble of outcomes.

The methods here presented may be used to perform dynamic and spatial sensitivity analyses of models with multiple outcomes. The novel contributions of our proposal can be split into three aspects. Firstly, we simultaneously considered the spatial resolution of the model, the physical characteristics of the landscape and the agricultural practices in a single experimental design. This allows evaluating not only the main effect of the spatial resolution, but also its interactions with the other factors. Secondly, we aggregated either spatially or temporally each outcome and we carried out a global sensitivity analysis on the aggregated variables. This allows identifying patterns in the temporal sequences and in the spatial distributions of the sensitivity indexes that may arise from the model structure (seasonality and spatial structure of the landscape, respectively). And thirdly, we analysed the ensemble of results of the sensitivity analyses on multiple outcomes. This allows grouping together model outcomes that are mainly sensitive to the same inputs and identifying the factors that impact the ensemble of outcomes in a similar way.

References

Chen H., Cheng C. (2011). Fractional Factorial Designs. *Design and Analysis of Experiments, Special Designs and Applications 3*, 299.

Duretz S., Drouet J.L., Durand P., Hutchings N.J., Theobald M.R., Salmon-Monviola J., Dragosits U., Maury O., Sutton M.A., Cellier P. (2011). NitroScape: a model to integrate nitrogen transfers and transformations in rural landscapes. *Environmental Pollution* 159 (11), 3162–70.

Ligmann-Zielinska A. (2013). Spatially-explicit sensitivity analysis of an agent-based model of land use change. *Int. J. Geographical Information Science 27* (9), 1764-1781.

Marrel A., Iooss B., Jullien M., Laurent B., Volkova E. (2011). Global sensitivity analysis for models with spatially dependent outputs. *Environmetrics 22* (3), 383-397.

Moreau P., Viaud V., Parnaudeau V., Salmon-Monviola J., Durand P. (2013). An approach for global sensitivity analysis of a complex environmental model to spatial inputs and parameters: a case study of an agro-hydrological model. *Environmental Modelling & Software 47*, 74-87.

Morris M.D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics 33* (2), 161-174.

The transformed optimal transportation problem: sensitivity and segregation of the children-to-school constrained assignment in Lausanne

Théophile Emmanouilidis, Guillaume Guex, François Bavaud

University of Lausanne, Switzerland

*Corresponding author: Theophile.Emmanouilidis@unil.ch

Abstract: We present a fuzzy approach aiming at assigning origins (children home locations) to destinations (schools) while considering capacity and segregation constraints. Solving the classic instance of the optimal transportation problem generally leads to many equivalent solutions, and can be computationally demanding for large networks. Introducing an additional origin-destination mutual information term, measuring the assignment cohesiveness and penalizing hard assignments, allows faster computing of an unique solution. The algorithm is built on two freely adjustable parameters : the temperature T controlling the assignment cohesiveness, and the mixing parameter c, favoring social diversity at the destinations. The approach yields relevant results in school organisation and is currently used by the public administration of Lausanne to plan children-to-school assignment as well as to evaluate and simulate the development of the schooling infrastructures.

Key-words: optimal transport, regularization, fuzzy clustering, fuzzy logic, social mixing modelling, location-allocation uncertainty, spatial decision making, Lagrangian multipliers

I INTRODUCTION

May it be from an urbanistic, a social or from a governance point of view, the evolution of cities is a major challenge of our contemporary societies. By giving the opportunity to analyze spatial and social configurations or attempting to simulate future ones, GIS cannot be overlooked in urban planning and management. In five years, the population of the city of Lausanne has grown from $134'700$ to $140'570$ inhabitants while the numbers in public schools have increased from $12'200$ to $13'500$ schoolchildren. Demographic rise and dynamics constitute spatially heterogenous processes that have a direct impact on school organization and planning. They require the adaptation of the schooling infrastructures and may generate uncertainty in schoolchildren location-allocation. Nowadays most cities have divided their territory into multiple school districts to assign children to school. But sooner or later such districts become inaccurate as they define fixed boundaries in an evolving urban landscape. Beyond the demographic aspects, assigning children to public schools is a complex task as it has to satisfy political and public demands, as well as to meet legal requirements. To help to the administrative services in performing this task we have developed a fuzzy-logic approach for children to school assignment.

Models for schools and schoolchildren location-allocation have been investigated in operational research ever since the 60's (Clarke and Surkis (1968), Koenigsberg (1968), Heckman and Taylor (1969)): see the reviews of Caro et al. (2004), as well as Castillo-Lopez and Lopez-Ospina (2015) for recent years. All the models seek to to allocate geography units to facilities, while considering spatial (journey to school distance treshold, closest assignment), logistic (capacity of schools) (Delmelle et al. (2014)), segregation and or economic constraints (Antunes and Peeters (2000)). Minimizing the average distance for given children locations and school capacities amounts to the celebrated optimal transportation problem (see e.g. Villani (2009)), possessing in general many equivalent solutions, computationally demanding for large networks.

Minimizing the social segregation at schools also possesses a multitude of solutions whenever the number of children locations exceeds the number of social categories, as in real situations.

The intrinsic degeneracy of the minimal distance and the minimal segregation problems prevents the existence of a unique and stable spatial partition, made of the union of school assignation basins. By contrast, introducing an additional origin-destination mutual information term, measuring the assignment *cohesiveness* and penalizing hard assignments, permits to make the solution unique. It furthermore allows the quick iterative computing of the minimal distance and the minimal segregation problems, as well as convex combinations of the latter, provided the penalization term is large enough.

The formalism and issues presently addressed bear obvious connections with iterative fitting, doubly constrained gravity flows, regularized optimal transportation, and model-based clustering with fixed mixture proportions. It involves in addition a presumably original term, the *social segregation at destination*, and appears to be *analytically tractable* (iterative determination of the unique solution, duality theory and sensitivity analysis). All the ingredients can be combined into an unique objective functional, the *free energy*, depending upon two freely adjustable parameters, the *temperature* T controlling the penalization of hard assignments, and the *mixing parameter* c, favoring social diversity at the destinations (section II). Toy and real applications (section III) illustrate the formalism together with its arguably numerous virtues, and the conclusion (section IV) emphasizes its connections to sensitivity analysis and robustness.

II FORMALISM

Formalizing the above considerations necessitates the introduction of three categorical variables, namely O = "child domicile" with n modalities, G = "school locations " with m modalities, and A = "social characteristics at origin" with p modalities.

2.1 Notations and definitions

Consider the $n \times m$ matrix $N = (n_{ig})$ counting the number of children living at i and assigned at school g. Row margins $n_{i\bullet}$ (where "$\bullet$" denotes the sum over the replaced index) count the number of children at i. Column margins $n_{\bullet g}$ yield the absolute capacity of school g.

Relative $n \times m$ *couplings* $P = (p_{ig})$ with $p_{ig} = n_{ig}/n_{\bullet\bullet} = p(i,g)$ constitute the joint probability distribution of O and G. Its margins are denoted by $f_i = n_{i\bullet}/n_{\bullet\bullet} = p(i)$ (origin weights) and $\rho_g = n_{\bullet g}/n_{\bullet\bullet} = p(g)$ (destination weights). The assignment *cohesiveness* is measured by the origin-destination mutual entropy, that is

$$I[P] := I(O:G) = \sum_{ig} p_{ig} \ln \frac{p_{ig}}{f_i \rho_g} = H(O) + H(G) - H(O,G) = H(G) - H(G|O) \quad (1)$$

where $H()$ denotes the corresponding entropies. By construction, $I[P] \geq 0$, with equality iff $p_{ig} = f_i \rho_g$, that is iff O and G are independent.

Let $D = (d_{ig})$ denote the $n \times m$ matrix of domicile to school pedestrian distances. The transportation cost or *energy* is the average journey to school distance:

$$U[P] := \sum_{ig} p_{ig}\, d_{ig} \quad (2)$$

Finally, consider a "social variable" A (such as nationality, ethnicity, social class or sex) with p modalities $a = 1, \dots, p$, and let $Y = (y_{ia})$ denote the given $n \times p$ matrix specifying the proportion of social type a at origin i, i.e. $y_{ia} = p(a|i)$ with $y_{ia} \geq 0$ and $y_{i\bullet} = 1$. The joint distribution of social types and destinations is given by $Q = (q_{ag})$ with

$$q_{ag} = \sum_i y_{ia}\, p_{ig} = p(a, g) \qquad \text{with} \qquad q_{a\bullet} = \sum_i f_i y_{ia} =: r_a \qquad \text{and} \qquad q_{\bullet g} = \rho_g \ .$$

$r_a = p(a)$ is the overall proportion of type a. Let us measure the social segregation at destinations by the mutual information between A and G, that is

$$S[P] := I(A : G) = \sum_{ag} q_{ag} \ln \frac{q_{ag}}{r_a \rho_g} \tag{3}$$

By construction, $S[P] \geq 0$, with equality iff $q_{ag} = r_a \rho_g$, that is iff A and G are independent (absence of segregation at destinations).

2.2 Minimizing the free energy

The simultaneous control of the average distance, segregation and cohesiveness can be achieved by minimizing the *free energy*

$$F[P] := U[P] + c\, S[P] + T\, I[P] \tag{4}$$

among all couplings P with given margins f and ρ. Here $T \geq 0$ is a freely adjustable *temperature* parameter, and $c \geq 0$ a freely adjustable *mixing* parameter. Optimal transportation problem corresponds to $T = c = 0$. Setting $T > 0$ favors soft assignment, and setting $c > 0$ diminishes social segregation at destinations. $c < 0$ can also be adopted, *favoring social segregation*, provided $|c|$ is not too large: further analysis demonstrates that $F[P]$ is convex for $c \geq 0$ (and hence possesses an unique minimizer), but ceases to be convex for $c < -T$.

The minimization of $F[P]$ is a standard exercise in calculus. Setting to zero its first derivative (including the Lagrange multipliers term insuring that marginal constraints are satisfied) yields the identities

$$p_{ig} = f_i \rho_g \exp(-\beta\, d_{ig}) \exp(-\beta\, c\, \ell_{ig}) \exp(\beta(\lambda_i + \mu_g)) \tag{5a}$$

$$\exp(-\beta\lambda_i) = \textstyle\sum_g \rho_g \exp(\beta\mu_g) \exp(-\beta\, d_{ig}) \exp(-\beta\, c\, \ell_{ig}) \tag{5b}$$

$$\exp(-\beta\mu_g) = \textstyle\sum_i f_i \exp(\beta\lambda_i) \exp(-\beta\, d_{ig}) \exp(-\beta\, c\, \ell_{ig}) \tag{5c}$$

$$\text{where} \quad \ell_{ig}[P] := \sum_a y_{ia} \ln \frac{q_{ag}[P]}{r_a \rho_g} = \sum_a \frac{p(ia)}{p(i)} \ln \frac{p(ag)}{p(a)p(g)} \ . \tag{5d}$$

Here $\beta = 1/T$ is the inverse temperature or *coldness*, λ_i and μ_g the Lagrange multipliers insuring the origin, respectively destination constraints. The quantity ℓ_{ig} (possibly negative) behaves as a "social distance" in the expression $d_{ig} + c\, \ell_{ig}$, penalizing the assignment $i \to g$ whenever the dominant social characteristics a at i tend to be overrepresented at g.

Direct substitution shows the minimum free energy to be $F = \sum_i f_i \lambda_i + \sum_g \rho_g \mu_g$, an expression generalizing Kantorovitch duality theory of optimal transportation (e.g. Villani (2009)), and justifying the interpretation of the multipliers λ_i and μ_g as the unit embarkment cost at origin i, respectively disembarkment cost at destination g.

2.3 Iterative determination of the optimal coupling

Starting from some initial coupling such as $P^{(0)} = f\rho'$, and some destination multiplier value such as $\mu^{(0)} = 0$, a new coupling $P^{(1)}$ is obtained by successively applying (5d), (5b), (5c) and (5a). The process is then iterated. Convergence occurs provided the cohesiveness contribution is large enough compared to the contributions of both energy and social segregation. That is, the temperature T must be large enough to allow efficient regularization. Numerical experiments with R (2.15.2) for the toy network of section 3.2 shows that convergence occurs (say after ca. 1'000 iterations) provided both the following conditions hold:

i) $T \geq 0.003 \cdot d_c$, where d_c is the *Hausdorff distance* between the set (support) of sources $S = \{i|f_i > 0\}$ and the set of targets $T = \{g|\rho_g > 0\}$, that is $d_c = \max(d'_c, d''_c)$, where $d'_c = \max_i \min_g d_{ig}$ and $d''_c = \max_g \min_i d_{ig}$. Below that threshold, overflow occurs in (5b) or (5c).

ii) $T \geq |c|$, that is $c \in [-T, T]$. Outside that range, the iterative process may exhibit periodic or non-converging behavior (especially for small values of T), betraying instability of the fixed point (yet existing and unique for $T > 0$ and $c \geq 0$ since $F[P]$ is convex).

2.4 Boundaries and placement error

High temperature favors low cohesiveness, that is soft memberships (of children into schools) $z_{ig} = p_{ig}/f_i = p(g|i)$. The conditional entropy $H(G|i) = -\sum_g z_{ig} \ln z_{ig} \geq 0$ measures the uncertainty in the assignment of i. High values of $H(G|i)$ characterize origins i lying at the *boundary* of two or more assignment basins, and help selecting domicile candidates for school reassignments, if needed (e.g. changes in f, ρ or Y from year to year).

This being said, even origins far from the assignment boundaries obey $H(G|i) > 0$ for $T > 0$ in view of the softness of Z. Attributing a given origin to different schools is not always possible nor desirable in real situations, so Z might have to be *hardened* into $Z^\star = (z^\star_{ig})$, where $z^\star_{ig} = 1$ if $g = \arg\max_h z_{ih}$, and $z^\star_{ig} = 0$ otherwise (ties are resolved at random). In general, $Z^\star$ does not satisfy the capacity constraint, that is the quantity $\rho^\star_g = \sum_i f_i z^\star_{ig}$ generally differs from ρ_g: for f and ρ given, equation $\rho = fZ'$ possesses generally no hard solution Z. The relative placement error $\delta = \sum_g |\rho_g - \rho^\star_g|$ is expected to increase with T (figure 1). The absolute placement error $N\delta$ (where N is the total number of children) counts the number of misattributed children (excess or deficit) by the hardened assignment.

III ILLUSTRATIONS AND APPLICATIONS

3.1 Example A

$n = 600$ children of the same kind (no segregation considerations here) are uniformly placed on a 25×24 grid, and have to be assigned to $m = 6$ schools of identical capacity of 100 each, so that the average city-block distance (L_1 metric) is minimal (1). As expected, the fuzziness of the boundaries as well as the placement error diminishes with the coldness $\beta = 1/T$.

3.2 Example B

Consider the random placement, without overlapping, of $n = 40$ origins and $m = 5$ destinations, on a 13×13 grid endowed with the city-block distance, with uniform origin weights $f_i = 1/n$ and uniform destination weights $\rho_g = 1/m$. In addition, we focus on gender segregation ($p = 2$), and generate a gender gradient across the origins, such that the proportion of girls is maximum in the south-west corner, and minimum in the north-east corner. Figure (2) depicts the hardened memberships $Z^\star$ for each origin i, and also the proportion of girls versus boys at domiciles and at schools.

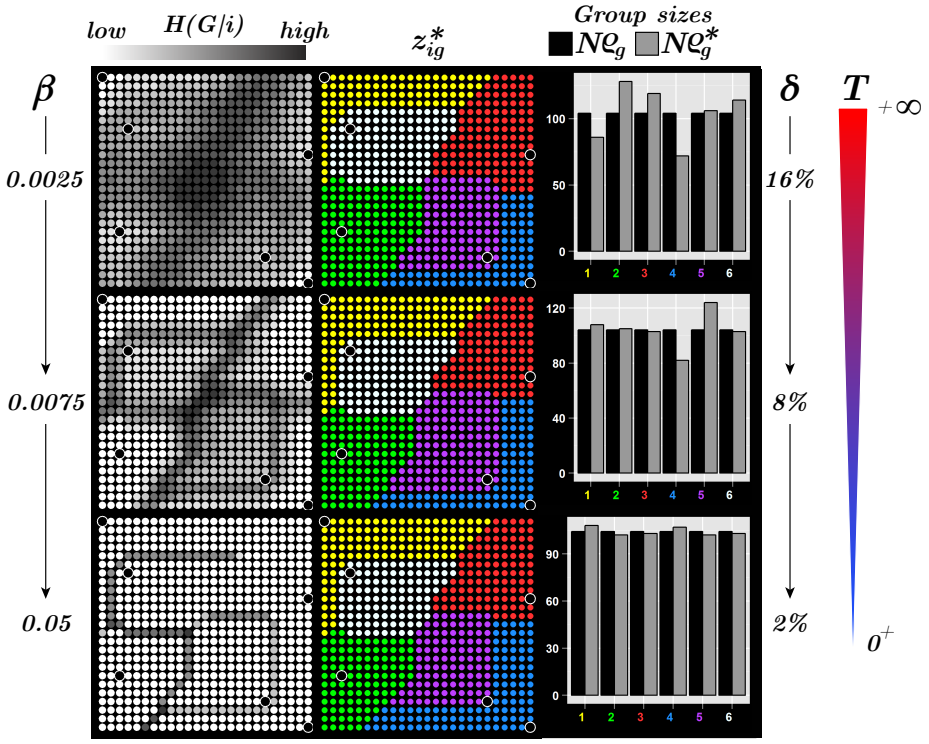


Figure 1: Example A: entropic boundaries $H(G|i)$ between school assignment basins, home-to-school assignment based upon the hardened membership $z^{\star}_{ig}$ (section 2.4), group sizes and resulting placement error δ, for decreasing values of the temperatures $T = 1/\beta$.

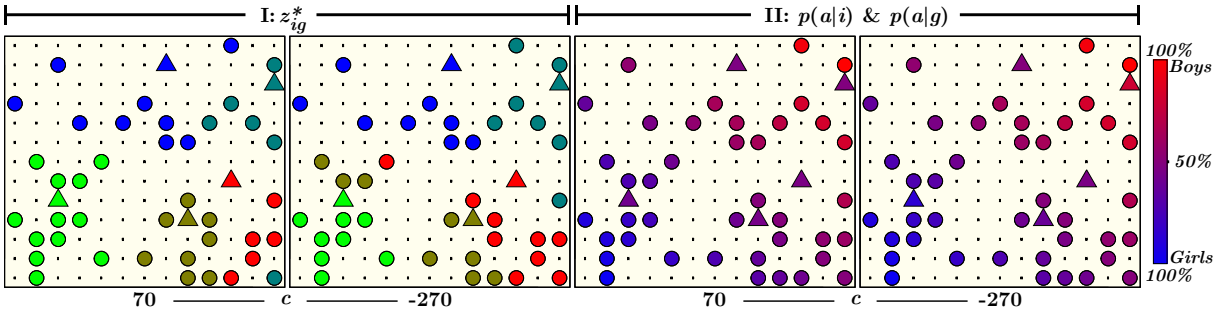


Figure 2: Example B. I: Hardened membership $z^{\star}_{ig}$ (I), II: the proportion of girls versus boys at origins $p(a|i)$ (circles) and at destinations $p(a|g)$ (triangles). With $d_c = 7$, $T = 16$ and while penalizing gender segregation at the destinations ($c = 70 > 0$) or, on the opposite, favoring it ($c = -270 < 0$).

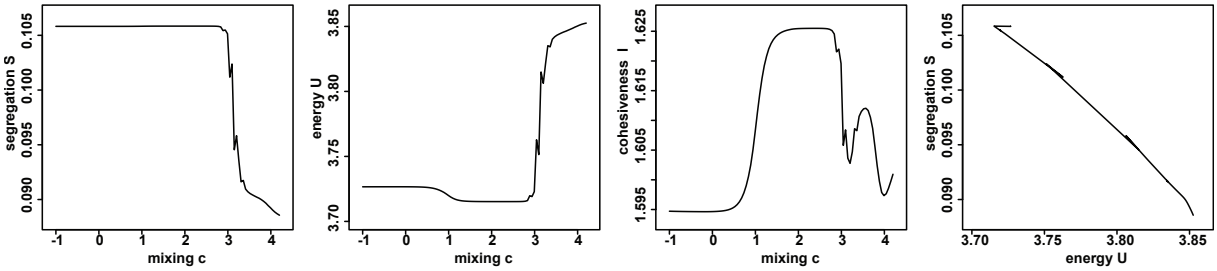


Figure 3: Example B: at low temperature $T = 0.02$, the iterative solution converges for $c \in [-1, 4.2]$. Mixings $c \in [0, 3]$ increase the cohesiveness I but leave U and S unchanged. Values $c > 3$ produce a sharp decrease in segregation S, and a sharp increase in energy U, depicting an effective anti-segregation regime together with an additional average distance cost.

3.3 Children-to-school constrained assignment in Lausanne

In the optimal transportation limit, that is in absence of social segregation considerations (i.e. $c = 0$), and for low temperatures $T \to 0^+$, the procedure has been shown (Guex et al. (2016), Emmanouilidis (2016)) to permit mapping of school attendance areas together with their fuzzy boundaries. As shown in Figure 4 (left), large values of the (spatially smoothed[1]) assignment entropy $H(G|i)$ (section 2.4) highlight fuzzy boundaries between basins, where the origin-to-destination attribution is most uncertain, and where the potential for school organisation and planning is most promising (e.g. balancing class sizes, or splitting up neighborhoods between two or more schools in order to promote social mixing). Furthermore, groups of attendance areas can be aggregated to design a new school district pattern. Mapping Lagrangian multipliers provides an overview of embarkment costs λ_i and disembarkment costs μ_g (Figure 4, right). Such a visualization can be especially useful for the comparison and evaluation of various potential scenarios involving changes of school capacities or new schooling infrastructures.

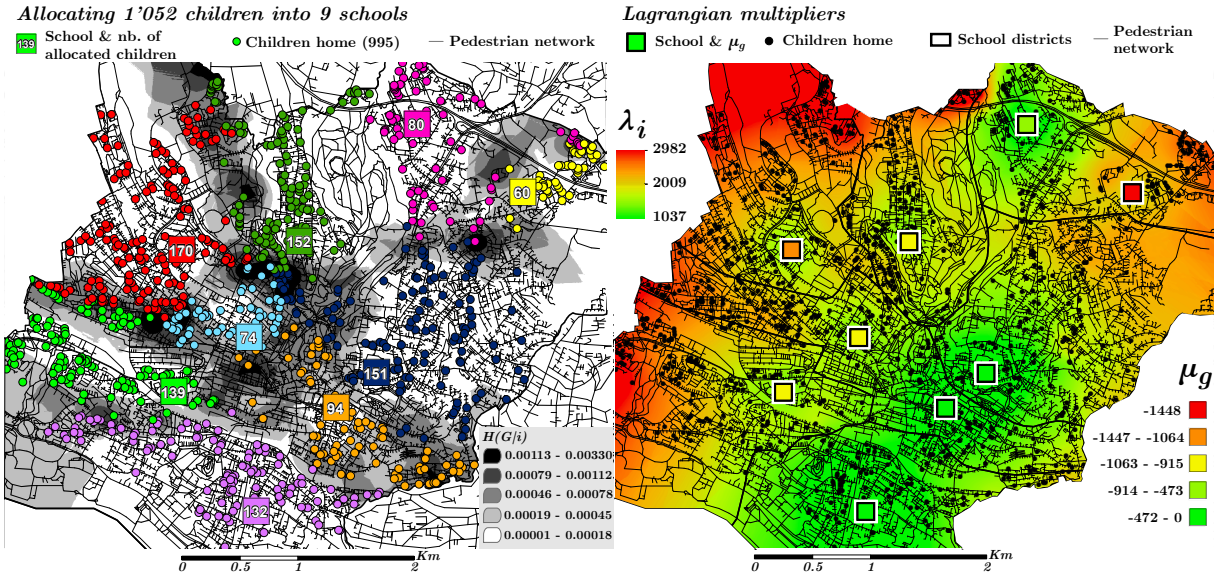


Figure 4: Resulting allocation plan for children entering secondary school in summer 2016 ($\beta = 0.2$). Left: hard assignment $Z^\star$ generates fairly compact attendance areas around schools. Children located in dark areas (high entropy $H(G|i)$) are more likely to be assigned to two or more schools, than the ones living in white areas. Right: the lack of schools in the western and northern outskirts of the city generates high (red orange) values of λ_i (spatially smoothed large embarkment costs). By contrast, children living in the center and in the south benefit from shorter journeys to school (green, yellow). The top-right school is too small to enroll all the surrounding children, which results in a large disembarkment cost μ_g.

3.3.1 Assignment with gender segregation constraint

Running the analysis on a similar dataset with adding the mixing parameter c leads to the results exposed in Figure 5. We here consider the assignment of 973 children into 9 schools. Cases I and II relates respectively the results with high ($c = -700$) and low ($c = 300$) gender segregation, while case III is free of segregation constraint ($c = 0$). Despite the almost perfect balance between the number of boys (486) and girls (487), their repartition within the city is locally heterogeneous (background maps of line B). Allocating children with high segregation ($c = -700$) generates almost five unisex schools (maps I-B & C). Four of them are located in

[1] in this paper, spatial interpolation is systematically performed through krigging (ArcGIS settings: type spherical, variable search radius, number of points:10).

the city-center while the fifth, on the top, is surrounded by a majority of girls. Highest values of c promotes, on the opposite, gender mix within schools (II-B & C). Results obtained with $c = 0$ are almost the same ((III-B & C), showing a good gender balance and suggesting that proportion of boys and girls around schools are almost equal. Accordingly, gender segregation here involves spatial mixing of overlapping hard assignments (I-A), while gender mixing creates a spatial partition of homogeneous connected groups (II & III-A).
Although the range of the assignment uncertainty ($H(G|i)$, line D) tends to narrow as c decreases, highest values characterize the same city parts in the three maps. The Lagrangian multipliers (line E) exhibit similar trends in cases I and II, where children of the west, north and far east suburbs have the longest journey to school (highest λ_i). With $c = 300$ (II-E) embarkment costs in these areas are more smoothed. The disembarkment costs μ_g vary little between the three cases, suggesting that gender segregation has, for this dataset, little effect on the former.

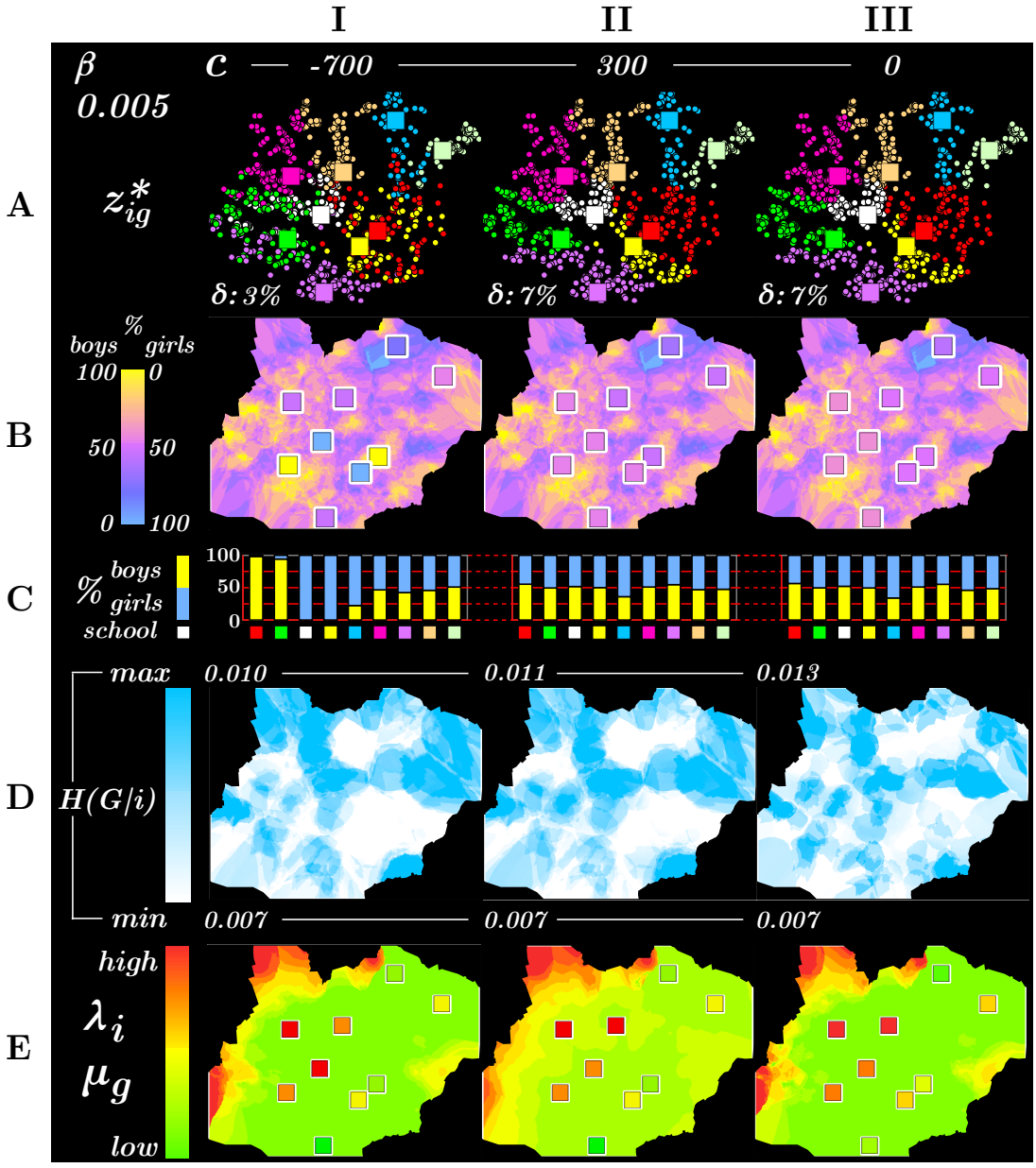


Figure 5: Children to school assignment with gender segregation constraint. A large negative value of c in (I) produces segregative allocation plan tending towards unisex schools. On the opposite, a large positive value of c promotes gender mixing within the schools.

IV CONCLUSION

Introducing some amount of randomness and uncertainty in an otherwise deterministic problem (the optimal transportation) results in a unique solution, easily and quickly computable, and *robust* with respect to small variations of child locations and populations, or school locations and capacities: once more, the regularizing virtues of the origin-destination mutual entropy are confirmed. Regularization permits in addition the effective identification and visualization of borders, of costs, and of alternative near-optimal solutions: those benefits are essential (and much appreciated) for adapting assignments to changes (turn-over of children at the origins, modifications of the pedestrian network) and for future planning.

As demonstrated in the present paper, this approach can be extended to handle in addition the issue of *social segregation*, by the introduction of a new *mixing parameter* controlling for the segregation. The resulting thermodynamic-like formalism permits to perform tractable *sensitivity analyses* of various kinds, assessing the amount of change in the quantities of interest induced by a change in the network geometry, origin or destination distributions, or by a change in the temperature or mixing parameters. In particular, the distance versus segregation trade-off can be made fully explicit (figure 3) - a crucial prerequisite for informed political decisions.

Instead of using pedestrian distances, the algorithm can be also be fed by alternative transportation costs, such as the total displacement time on uni- or multimodal transportation unoriented networks. Also, beside gender mixing, the segregation issues related to child age or ethnic origin can be tackled by the same approach, to be addressed in a future work. Among further possible developments, the question of dealing with flexible school penalized capacities (instead of with fixed capacities) seems worth investigating, as is the issue of imposing a minimal social mixing at each school, rather than on average only.

References

Antunes A., Peeters D. (2000). A dynamic optimization model for school network planning. *Socio-Economic Planning Sciences 34*(2), 101–120.

Caro F., Shirabe T., Guignard M., Weintraub A. (2004). School redistricting: Embedding gis tools with integer programming. *Journal of the Operational Research Society 55*(8), 836–849.

Castillo-Lopez I., Lopez-Ospina H. A. (2015). School location and capacity modification considering the existence of externalities in students school choice. *Computers & Industrial Engineering 80*, 284–294.

Clarke S., Surkis J. (1968). An operations research approach to racial desegregation of school systems. *Socio-Economic Planning Sciences 1*(3), 259–272.

Delmelle E. M., Thill J.-C., Peeters D., Thomas I. (2014). A multi-period capacitated school location problem with modular equipment and closest assignment considerations. *Journal of Geographical Systems 16*(3), 263–286.

Emmanouilidis T. (2016, 02). *Analyse de réseau piéton et gestion scolaire: sinuosité, centricité et transport optimal régularisé*. Ph. D. thesis, Université de Lausanne, faculté des Géosciences et de l'Environnement.

Guex G., Emmanouilidis T., Bavaud F. (2016). Transportation clustering: a regularized version of the optimal transportation problem. *Submitted for publication*.

Heckman L. B., Taylor H. M. (1969). School rezoning to achieve racial balance: a linear programming approach. *Socio-Economic Planning Sciences 3*(2), 127–133.

Koenigsberg E. (1968). Mathematical analysis applied to school attendance areas. *Socio-Economic Planning Sciences 1*(4), 465–475.

Villani C. (2009). *Optimal Transport, Old and New*. Springer.

Poster Session

A method for testing the similarity of spatial samples

M. Virtudes alba-Fernández[1], Francisco J. Ariza-López[2], José Rodríguez-Avi[1]

[1]Dpt. Ststistics and O.R. University of Jaén, Sapin
[2]Dpt. Cartographic engineering, geodesics and Photogrammetry. University of Jaén, Spain

*Corresponding author: mvalba@ujaen.es

Abstract

The purpose of this work is to propose a method to analyse the similarity between two spatial point patterns. Such similarity should be understood in the sense that both point patterns come from the same spatial point distribution. For this aim, first, we make use of a space-filling curve as a tool to linealize the space, with independence of its dimension, second, we model the count of points in the resultant grid by means of the multinomial law, and after that we test the homogeneity of both multinomial distributions by using the negative of the Matusita's affinity. Finally, we evaluate the performance of the procedure by means of a simulation study.

I INTRODUCTION

The understanding of spatial point patterns is one of the major challenges of geographical analysis and has interest in many sciences (e.g. biogeography, crop sciences, ecology, geology, etc.). Spatial patterns and spatial statistical sampling is also a major issue in spatial data quality assessment because sampling is a very common procedure in order to derive estimates or perform tests (Ariza-Lpez, 2002). For these reasons the evaluation of the spatial similarity of two observed samples can be of interest in many cases.

It is usual that an estimate of an attribute or property (e.g. concentration of a mineral, positional accuracy, etc.), is derived from a sample of points under some spatial distributional hypothesis for such sample (e.g. following a theoretic or an observed spatial pattern). We think that, previous to any error or uncertainty consideration about the attribute or property estimation, we must confirm the underlying hypothesis about the position of samples, in our case: the spatial distribution of the taken sample in relation to other observed spatial pattern.

Two common methods in use to investigate discrete point events are the distance- and area-based tests. In the case of distance-based tests, we use whatever distance (e.g. Euclidean) between two events in order to determine a random, clustered, or uniform spatial pattern to the points (Bailey and Gatrell, 1995). On the other hand, in the case of area-based tests, we count the number of point events within a predefined spatial area (e.g. a quadrant, a census unit, etc.) (Andresen, 2009).

Our proposal is very different to other techniques such as the Ripley's K function (Ripley, 1976). The K function is a distance-based test that characterizes point processes at many distance scales and allows the detection of different behaviors (e.g. random. clustering, inhibition) (Freeman and Ford, 2002); and it can be used to test different specific patterns (e.g. homogeneous Poisson process -complete spatial randomness-, Matern hard-core process, Strauss process, etc.) (Dixon, 2012). For the correct application of these functions it is needed to accept several assumptions that not necessarily are fully met in reality, and if conditions are not met, the output may be incorrect (Bolibok, 2008).

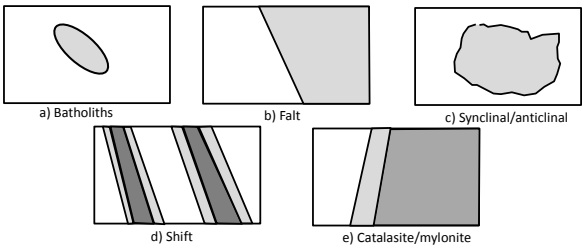


Figure 1: Examples of some geological structures that can influence spatial patterns: a) batholiths, b) falt, c) Synclinal/anticlinal, d) shift, e) cataclasite/mylonite

In this setting, we develop an area-based test centered on the counting of positional events (not attributes), which does not assume a theoretical model for the spatial distribution of the events (spatial distribution free). So, it can be applied for the comparison of two spatial samples (e.g. two control test samples, two field works, presences at two different times, etc.), with independence of sample sizes. For instance, the Figure 1 shows different geological structures that can determine the presence of point events in the space or the distribution of control samples in the space, and this method allows the comparison in such cases, and any other.

This method follows two steps; the first one is to make use of the space-filling curves (Sagan, 1994) as a tool to order the space, with independence of its dimension. The space-filling curves provide a partition of the space for a fixed level of neighborhood. The second one is to model the count of points in the resultant grid by means of the multinomial law. This way, the problem to test the similarity between sampled spatial distributions is equivalent to the problem of testing the homogeneity of two multinomial distributions. For doing this, we considered as a test statistic the negative of Matusita's affinity (Matusita, 1967).

On the other hand, the spatial distribution of samples may have an effect on the representativeness of the sample. This is especially true in complex situations, although, at the end of the day, for many spatial data properties, the spatial pattern can be considered uniform. To evaluate the usefulness of the proposal, we carried out a simulation study in order to analyze the effect of the space-filling curves and the test statistics used, as well as, the effect of the sample sizes and the level of iterations in the curves. In particular, we took the Hilbert's curve and the Peano's curve, and we analyze the uniform pattern and for other spatial patterns, specially, those patterns associated with some well-known geological structures. Our findings suggest the methodology is able to detect the similarity of two spatial samples of points for all the tried cases.

This paper is organized as follows, after this introduction follows the description of the approach, where we present the statistical basis and the procedure for its application; next a simulation experiment is developed using some spatial patterns and levels of the space-filling curves. Finally conclusions are presented.

II DESCRIPTION OF THE APPROACH

As said in the Introduction, the aim of this work is to propose a formal procedure to test whether two spatial distributions of points can be considered as similar or not. By Similar one should understand that coming from the same spatial pattern. To show how this approach works, let us consider the bivarate case, although for a general dimension, the procedure follows the same steps. So that, let $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be two independent samples of points from two spatial patterns with size n_i, $i = 1, 2$; and without loss of generality, let us suppose that both samples take values in the unit-square $S = [0, 1] \times [0, 1]$. The application

of a particular space-filling curve induces a partition on S with $M = 2^\nu \times 2^\nu$ squares, where ν represents the number of iterations in the space-filling curve construction. This way, for a given space-filling curve and a fixed level ν, the sampled points can be grouped into M classes, $C_1, C_2, \ldots, C_M$, or equivalently, taking values in $\Upsilon_M = (1, 2, \ldots, M)$. The order in M is induced by the space-filling curve considered. Note that the degree of neighborhood is different for each space-filling curve, specially, when the number of iterations increases.

Let $\pi_i^t = (\pi_{i1}, \pi_{2m}, \ldots, \pi_{iM})^t$ be the cell probabilities associated with each multinomial distribution, that is, $\pi_{im} = P[X_i = m]$, for $i = 1, 2$, and $1 \leq m \leq M$, and where the superscript^t denotes transpose. The application of a particular space-filling curve to both samples of points, provide us the number of points falling into each class in both samples. That is to say, we obtain the observed frequencies on each cell and hence, the maximum likelihood estimator of $\hat{\pi}_i$, say, $\hat{\pi}_i = \frac{n_{im}}{n_i}$, $i = 1, 2$, $m = 1, \ldots, M$. As a result, the problem of testing whether two spacial distributions are equal is equivalent to test whether two multinomial populations are equal. So that, our objective is to test the following null hypothesis

$$H_0 : \pi_1 = \pi_2. \tag{1}$$

For this purpose, several measures able to discriminate between multinomial populations can be used. Here, we consider a f-dissimilarity measure between multinomial populations because the smaller these measure is, the harder it is to discriminate between them. Specifically, it is considered the negative of Matusita's affinity (Matusita, 1967) defined by

$$T = -\sum_{m=1}^{M} \sqrt{\hat{\pi}_{1m}\hat{\pi}_{2m}},$$

which is a member of a class of test statistics based on f-dissimilarity (see Zoografos, 1998; Alba et al., 2009) and references therein for further theoretical results).

To decide when to reject H_0, we need to know the null distribution of T, or at least an approximation to it. Following Zoografos (1998), under H_0, the asymptotic null distribution of the test statisticn is given by

$$8\frac{n_1 n_2}{n_1 + n_2}(1 + T) \xrightarrow{\mathcal{L}} \chi^2_{M-1}. \tag{2}$$

So, we reject H_0 if

$$8\frac{n_1 n_2}{n_1 + n_2}(1 + T_{obs}) \geq \chi^2_{\alpha, M-1},$$

where T_{obs} represents the observed value of the test statistic T, and $\chi^2_{\alpha,M-1}$ denotes the $1 - \alpha$ percentil of the chi-square distribution with $M - 1$ degrees of freedom, $0 \leq \alpha \leq 1$; or equivalently, we reject H_0 if the corresponding p-value is less than or equal to α, that is, if

$$p = P_{H_0}\left[\chi^2_{M-1} \geq 8\frac{n_1 n_2}{n_1 + n_2}(1 + T_{obs})\right] \leq \alpha,$$

where P_{H_0} stands for the probability law under the null hypothesis.

As observed among others in Kim, 2009, Alba et al., 2009 or Jiménez-Gamero et al., 2014, the χ^2 approximation is rather poor for small and moderate sample sizes, and the approximation

of the null distribution by means of a parametric bootstrap estimator behaves better than the asymptotic one. From these results, we approximate the p-value by bootstrapping (see Alba-Fernández et al. 2009 for the theoretical properties of the bootstrap estimator).

The application of the procedure to study of similarity of two spatial samples of points follows the steps:

(1) Given the two spatial distributions of points $\{X_{1j}\}_{1\leq j\leq n_1}$ and $\{X_{2j}\}_{1\leq j\leq n_2}$, choose a space-filling curve and the number of the iterations in its construction, ν. These decisions will determine the value of M.
(2) Apply the space-filling curve and obtain $\hat{\pi}_i$, $i = 1, 2$.
(3) Calculate T_{obs}, the observed values of T.
(4) Approximate the corresponding p-value by bootstrapping.
(5) Conclude if the sample spatial distributions can be considered equal or not.

In addition, the bootstrap algorithm to approximate the p-value for testing (1) can be assessed as follows:

(i) Calculate T_{obs}, the observed values of T.
(ii) For $b = 1, \ldots, B$, generate $2B$ independent bootstrap samples, $\left\{X^{*b}_{1,j}\right\}_{1\leq j\leq n_1}$, $\left\{X^{*b}_{2,j}\right\}_{1\leq j\leq n_2}$ from the pooled multinomial distribution $M(n_1 + n_2; \hat{\pi}_{01}, \ldots, \hat{\pi}_{0M})$ where

$$\hat{\pi}_{0m} = \frac{n_{1m} + n_{2m}}{n_1 + n_2},\ m = 1, \ldots, M.$$

(iii) Calculate the values of T for each couple of bootstrap samples, say T^{*b}, $b = 1, \ldots, B$.
(iv) Approximate the p-value by means of $\hat{p} = card\left\{b : T^{*b} \geq T_{obs}\right\}/B$, respectively.

III SIMULATION EXPERIMENT

To evaluate the performance of this method, we have carried out a simulation study. The goal of this experiment is twofold, the first objective is to analyze the behaviour of the proposed methodology for small sample sizes with respect to the type I error for some space-filling curves and levels of sweep. In other words, if the procedure is able to conclude if two sample of points come from the same spatial pattern when they really do.

Together with this, the second task is to evaluate the power of the procedure, that is, if this method is able to detect two samples of point which have been generated by different spatial patterns. Next we briefly describe the simulation experiment and display the results obtained. All computations in this paper have been performed using scripts written in the R language (http://www.cran.r-project.org).

The methodology we propose is independent of the spatial pattern of sample of points, however, we are going to evaluate its behaviour by considering some spatial patterns related to the geological structures mentioned before (see figure 1). In particular, we will identify the uniform pattern with Pattern 1, the bivariate normal $N(\mu, \Sigma)$ with $\mu = (0.25, 0.25)$, $\Sigma = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}$ simulates a synclinal/anticlinar(Pattern 2) and the bivariate normal $N(\mu, \Sigma)$ with $\mu = (0.5, 0.5)$, $\Sigma = \begin{pmatrix} 0.25 & 0.8 \\ 0.8 & 0.25 \end{pmatrix}$ reproduces a Batholiths (Pattern 3) (see Figure 2) for a scatter plot of a sample of size 150).

So, we have generated two uniform samples on the unit-square of sizes $n_1 = n_2 = 25$, we have applied the method following the steps (1)-(5) described above with $B = 1000$ bootstrap

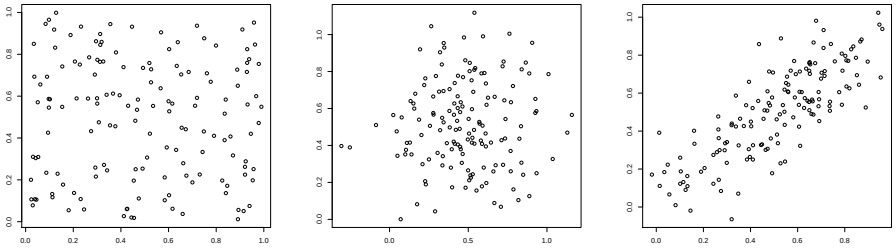


Figure 2: Example of Pattern 1 (left), Pattern 2 (center) and Pattern 3 (right).

replications. We repeated this 1000 times and we calculated the fraction of p-values less than or equal to 0.05 and 0.10 (denoted as $f05$ and $f10$ in tables), which are the estimated type I error probabilities for $\alpha = 0.05, 0.10$, respectively. We have repeated the whole experiment for the sample sizes $n_1 = n_2 = 50, 150$. Among the available space-filling curves, we have considered the Peano's and the Hilbert's curve (identified as n and h in tables). Note the value of M is related to the sample size and ν, and we try to find an agreement between them. Here, we try the values $\nu = 1, 2$. We have repeated this experiment for Patterns 2 and 3. The estimated type I error probabilities are shown in Table 1.

Looking at this table, we can conclude that the estimated type I error probabilities are quite close to the nominal ones in all the tried cases and the simulations results do not show differences between the space-filling curves. On the other hand, we have also studied if the methodology is able to distinguish between spatial patterns, and for this task we generated a sample of points following the Patter 1 of size $n_1 = 25$ and other sample of points from Pattern 2 of size $n_2 = 25$. We applied the methodology in the same conditions as before. We repeated the whole experiment 1000 times and we calculated the fraction of $p's$ values less than or equal to $0.5, 0.10$, which are the estimated power for $\alpha = 0.5, 0.10$ (now, $f05$ and $f10$ in tables). We repeated the experiment for samples of points following the Patterns 3. The estimated powers are shown in Table 2. From these results, we can say the procedure is able to distinguish clearly between two different Patterns.

$n_1 = n_2$	ν	curve	Pattern 1		Pattern 2		Pattern 3	
25	1		f05	f10	f05	f10	f05	f10
		n	0.042	0.099	0.052	0.100	0.048	0.096
		h	0.047	0.102	0.052	0.096	0.050	0.096
50	1		f05	f10	f05	f10	f05	f10
		n	0.055	0.103	0.053	0.105	0.054	0.106
		h	0.053	0.101	0.053	0.101	0.053	0.107
150	1		f05	f10	f05	f10	f05	f10
		n	0.054	0.104	0.058	0.107	0.046	0.093
		h	0.057	0.099	0.059	0.104	0.045	0.098
150	2		f05	f10	f05	f10	f05	f10
		n	0.056	0.104	0.048	0.094	0.060	0.108
		h	0.057	0.103	0.048	0.093	0.057	0.107

Table 1: Estimated type I error probabilities.

$n_1 = n_2$	ν	curve	Pattern 1 vs. Pattern 2		Pattern 1 vs. Pattern 3		Pattern 2 vs. Pattern 3	
25	1		f05	f10	f05	f10	f05	f10
		n	0.429	0.548	0.409	0.534	0.570	0.665
		h	0.427	0.557	0.366	0.501	0.537	0.635
50	1		f05	f10	f05	f10	f05	f10
		n	0.747	0.837	0.719	0.802	0.886	0.938
		h	0.748	0.837	0.715	0.802	0.881	0.928
150	1		f05	f10	f05	f10	f05	f10
		n	0.994	0.995	0.998	0.999	0.999	1.000
		h	0.995	0.995	0.998	0.999	0.999	1.000
150	2		f05	f10	f05	f10	f05	f10
		n	0.987	0.994	1.000	1.000	1.000	1.000
		h	0.989	0.995	1.000	1.000	1.000	1.000

Table 2: Estimated power.

IV CONCLUSION

To sum up, the procedure introduced in Section 2 takes advantage of the use of space filling curves as a way to linearize spatial distributions, and following the order induced by the space filling curve for a fixed level of its construction, the count of the number of points falling into each grid may be modeled by a multinomial distribution. For testing the homogeneity of two multinomial laws, the negative of Matusita's affinity is considered. Finally, the results of the simulation experiment reveals that the proposed method provide us a statistical tool in order to decide about the similarity of spatial samples.

ACKNOWLEDGMENTS

Research in this paper has been partially funded by grant CTM2015-68276-R of the Spanish Ministry on Science and Innovation.

REFERENCES

Alba-Fernández, V., Jiménez-Gamero, M.D. (2009). Bootstrapping divergence statistics for testing homogeneity in multinomial populations. Mathematics and Computers in Simulations, 79, 3375–3384.

Andresen M.A. (2009). Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach. Applied Geography 29, 333345.

Ariza-Lpez F.J. (2002). Calidad en la produccin cartogrfica. Madrid, ES, Ra-Ma.

Bailey T.C., Gatrell, A.C. (1995). Interactive spatial data analysis. Harlow, UK: Prentice Hall.

Bolibok L. (2008). Limitations of Ripleys k(t) function use in the analysis of spatial patterns of tree stands with heterogeneous structure. Acta Sci. Pol. Silv. Colendar. Rat. Ind. Lignar 7(1), 5-18.

Dixon P.M. (2012). Ripley's K function. Encyclopedia of Environmetrics, 2nd ed. John Wiley and Sons, Inc.

Freeman E.A., Ford E.D. (2002). Effects of data quality on analysis of ecological pattern using the k(t) statistical function. Ecology 83(1), 3546.

Jiménez-Gamero, M.D., Alba-Fernández, V., Barranco-Chamorro, I., Muñoz-García, J. (2014). Two classes of divergence statistics for testing uniform association. Statistics, 48 (2), 367–387.

Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. Annals of the Institute of Statistical Mathematics, 19,181–192.

Sagan H. (1994).Space-Filling Curves. Springer-Verlag.

Ripley, B.D. (1976). The second-order analysis of stationary point processes. Journal of Applied Probability 13, 255266.

Zografos, K. (1998). f-dissimilarity of several distributions in testing statistical hypotheses. Annals of the Institute of Statistical Mathematics, 50, 295-310.

Combining punctual and ordinal contour data for accurate floodplain topography mapping

Carole Delenne[*1,3], **Jean-Stéphane Bailly**[2], **Mathieu Dartevelle**[3], **Nelly Marcy**[1], **Antoine Rousseau**[3]

[1]Univ. Montpellier, HSM, France
[2]AgroParisTech, LISAH, France
[3]INRIA, Lemon, France

*Corresponding author: carole.delenne@umontpellier.fr

In the framework of hydrodynamic modelling, topography is classically obtained by an interpolation of punctual field elevation surveys. A methodology, based on block conditional simulations, is presented to enhance mapping accuracy, using contour lines extracted from flooded areas in remote sensing data.

I INTRODUCTION

Hydrodynamic models in two dimensions require a precise knowledge of the domain topography. However, as far as the modelling of great rivers or lakes is concerned, few accurate topographic records are generally available to accurately model the floodplain topography. The usual way to acquire accurate topographic information for hydraulics over floodplain remains the ground surveys, that provide punctual values at a very high cost. Based on these points, usual interpolation schemes often yield a too coarse and inaccurate topographic map for a realistic hydrodynamic modelling. Remote sensing data such as Lidar or stereo-photogrammetry data on non vegetated areas, constitute a good alternative to obtain large scale information, but yield other issues of acquisition and data processing. In addition, topographic Lidar operating with near-infrared lasers are still not suitable for an exhaustive survey of floodplains where submerged areas remain.

Progress in remote sensing data repeatability and spatial resolution now allows the automatic monitoring of water surfaces delineation from areal or satellite images either in optical or radar domains (see *e.g.* an application to the Inner Niger Delta in Ogilvie et al. (2015)). As for the recently launched Sentinel sensors, these data are becoming widely available with increasing spatial and temporal resolutions, allowing in turn the spatio-temporal monitoring of flooded areas. Flood dynamics from remote sensing data are known to be informative on floodplain topography for long (see *e.g.* Schumann et al. (2007) or Hostache et al. (2010)). Indeed, the extracted flooded areas may be considered as iso-elevation contour lines, as the hypothesis can be made that all the points located on the water/soil limit have the same elevation. If contour line elevation remains unknown, rank between the detected contour lines as well as between every contour line and surrounding topographic data points are known. Assuming this information, the next challenge is thus to develop a spatial data fusion method of ordered contour lines data and usual elevation data points to better model the floodplain topography.

Mixing punctual ranked and continuous data in spatial estimation was already proposed in literature, mainly to deal with highly skewed field distribution (Yamamoto, 2000) or for very large spatial datasets (Cressie and Johannesson, 2008). These methods used the uniform score transform and back-transform based on the standardized rank estimation (Saito and Goovaerts, 2000) to prevent the right estimate distribution and limit the smoothing effect of kriging. Uncertainty

assessment of resulting punctual kriging estimates was proposed thanks to the transformation process (Yamamoto, 2007). Other methods also exist to quantify the uncertainty of contour lines in random fields (Lindgren and Rychlik, 1995; Wameling and Saborowski, 2001). But in these studies, the objective was to quantify the uncertainty in contour lines location and not to quantify the uncertainty of contour lines level for a known location. Consequently, these different studies only partly touched the problem we face.

The objective of this paper is to propose a method to estimate the elevation of the ranked contour lines from data points. This method starts from block conditional simulations estimates filtered on rank statistics. After contour line elevation estimate, a usual kriging is performed to reconstruct a topographic map over the floodplain. Starting from a synoptic and theoretical example, the principles, advantages, properties and limitations of the proposed method are exposed. The spatial accuracies obtained in contour lines estimates and topographic map are thus compared to the usual block kriging estimates.

II MATERIAL AND METHODS

2.1 Case study generation

A reference exhaustive elevation field $Z(s)$ on a floodplain D was generated from a Gaussian spatial covariance model with range 20, sill 1 and no nugget effect. The resulting field is centered on $\mu = 10$ and respects distribution $N(\mu, \sigma)$ (see Figure 1-left): the blue colours represent lower floodplain elevations meanwhile the yellow and green stem for higher elevation areas. D is a 100 m$\times$100 m area and the dataset was generated on gridded D with resolution $r = 1$ m, r denoting both the remote sensing image spatial resolution and the desired end-user topographic map resolution.

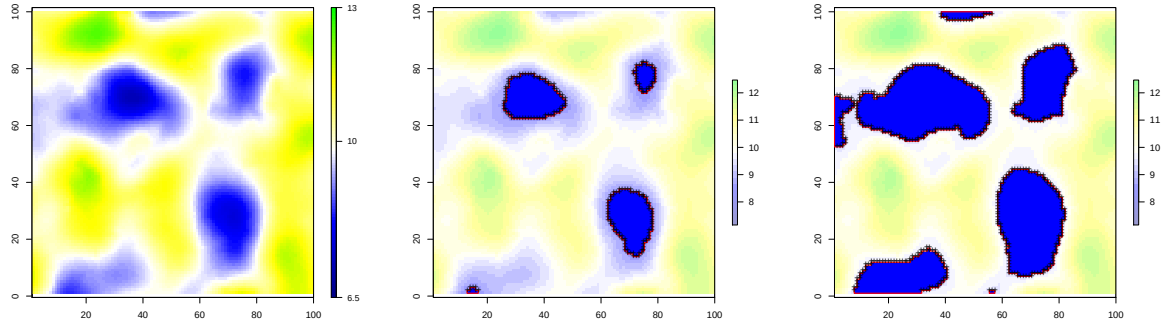


Figure 1: **Generated case study.** Left: Simulated elevation field over D; Middle: flooded area at time 1 (blue) and resulting rasterized contour (black crosses); Right: flooded area at time 2 (blue) and resulting rasterized contour (black crosses).

To reproduce the remote sensing data support (Ogilvie et al., 2015), flooded areas were generated and delineated at resolution 0.5 m for two different times t_1 and t_2 during rising water: when areas lower than 9.5 m (Fig. 1-middle) and 8.5 m (Fig. 1-right) respectively are flooded. Contour polylines were generated from these rasterized flooded areas with 0.5 m spatial resolution. The sets of two generated contour polylines C_1 and C_2 were thus reduced to a set of polylines vertices s_i^j ($j \in (1, 2)$ denoting the polyline) regularly located along the lines (black cross on figures 1). C_1 and C_2 contained $n_1 = 201$ and $n_2 = 520$ points respectively. Four

contour lines correspond to C_1 and five to C_2 (Fig. 1). When contour lines were connected to the D boundaries, only vertices within D were considered.

In addition, a sample of $n = 250$ topographic points $s_1 \dots s_n$ was randomly selected on D, excluding points selection at location closer than $b = 2$ m from polylines (Fig. 2-left). This ensures that no data points are located on a polyline, which would make too obvious the estimation of its elevation. We further assume that contour polylines elevation is unknown but that ranks between the polylines (i.e. the polyline vertices s_i^j) and between polylines and surrounding topographic data points are known and may be computed (Fig. 2-right).

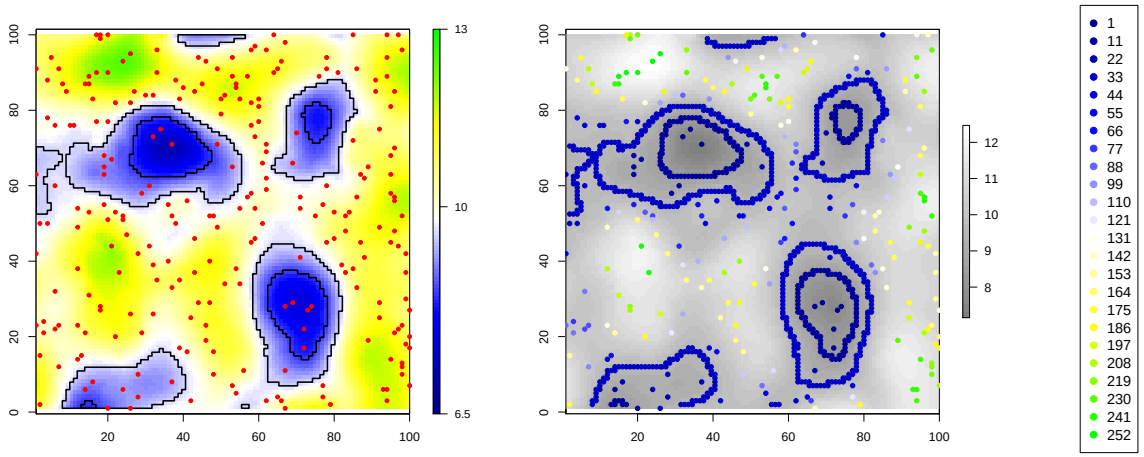


Figure 2: **Generated dataset.** Left: Simulated elevation field over D, contour polylines C_1 and C_2 (black lines) and data points (red points). Right: ranked data (points and polyline vertices) and according legend.

2.2 Method

The method we developed may be seen as a block conditional simulation process filtered by rank statistics, where a polyline C_j is seen as a block, *i.e.* a region corresponding to an unconnected set of vertices. Such simulation is known to be superior to kriging whenever interest lies in global statements for a region rather than inference on individual points. In the process described in the algorithm 1 hereafter, assuming a stationary random function, a spatial model (variogram) $\gamma(h)$ is first estimated and modelled from the 250 data points $s_1 \dots s_n$.

Once the variogram $\gamma(h)$ is modelled, an ordinary kriging is performed on each polyline vertex. From the kriging, a first estimate of polyline elevation results from vertices averaging. For polyline C_j, we denote further $\hat{z}_{ko}^j$ this estimate (corresponding to the usual block kriging estimate).

Next, a set of $N = 100$ conditional Multi-Gaussian simulations using the fitted variogram model and pathing through the 250 $s_1 \dots s_n$ is performed on each on the s_i^j contour vertices (Eq. 1):

$$z_{cs}(s_i^j) = \hat{z}_{ko}(s_i^j) + (z_{us}(s_i^j) - \hat{z}_{ko}^*(s_i^j)) \tag{1}$$

At the end of the conditional simulation, an estimate at polyline scale is computed by averaging simulated values on vertices. This remains consistent with the usual block kriging estimate

Algorithm 1 Contour polyline estimation process and field reconstruction

for j = 1 to j= 2 **do**
    Estimate and model the variogram $\gamma(h)$ from n data points
    Estimate ordinary kriging $\hat{z}_{ko}(s_i^j)$ on polyline vertices
    for n = 1 to N = 100 **do**
        Draw unconditional multigaussian simulation $z_{us}(s_k)$ and $z_{us}(s_i^j)$ on data location (k) and polyline vertices (i).
        Estimate ordinary kriging at vertices from simulated values at data locations $z_{ko}^*(s_i^j)$
        Compute conditional simulation on vertices (Eq. 1)
    Compute N polyline estimate by averaging vertices conditional simulations
    Filter keeping only polyline estimate conform to the initial ranks
    Compute polyline estimate $\hat{z}_j$ averaging the kept polyline estimates
    Compute standard deviation estimate from the set of kept polyline estimates
Affect polyline estimate $\hat{z}_j$ to each polyline vertice s_i^j
Perform ordinary kriging on all gridded D points from data $z(s_1) \cdots z(s_n)$ and polyline vertices estimates $\hat{z}_j(s_i^j)$

since vertices are regularly located along the polyline.

$$\hat{z}_k^j = \frac{1}{n_j} \sum_{i=1}^{n_j} z_{cs}(s_i^j) \tag{2}$$

We thus obtain N elevation estimates $\hat{z}_k^j$, $k \in (1 \dots N)$ for a given contour polyline C_j. Only the estimates respecting the rank are kept in the following, yielding a set of M estimates ($M \leq N$) for a given polyline. From this set, the final estimate of polyline C_j becomes:

$$\hat{z}_j = \frac{1}{M} \sum_{k=1}^{M} \hat{z}_k^j \tag{3}$$

with uncertainty characterized by the standard deviation σ_{C_j} of the $\hat{z}_k^j$ estimates.

In a final step, polyline estimate $\hat{z}_j$ is affected to each polyline vertex s_i^j in order to perform, in addition to the 250 data points a gridded map using ordinary kriging.

III RESULTS

Results obtained for this theoretical test case are given in Figures 3 and 4. Figure 3-left shows the estimated empirical variogram (dots) and fitted variogram model (lines). Figure 3-right shows for C_1 and C_2 the resulting distribution of the estimates using conditional simulation, the final estimation equal to the average of conditional simulation (black vertical line) compared to the true value (blue vertical line) or the ordinary kriging (interpolation) estimate (red vertical line). In this first application, rank was always respected in the conditional simulation and for the ordinary kriging but, as shown in the following, this will not always be the case. The estimated elevation obtained for C_1 is 8.53 m (8.45 using ordinary kriging) for a true value equals to 8.5 m, with a standard deviation of the estimate being only 5.37e-03 m. For C_2, the difference between the true elevation (9.5 m) and the estimated value is lower than one centimetre (3 cm using kriging) with standard deviation of the estimate equals to 3.99e-03 m.

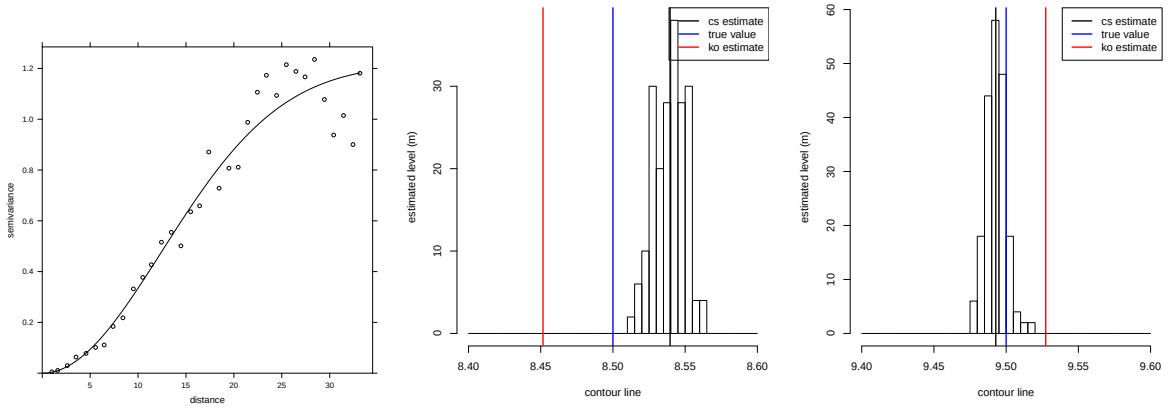


Figure 3: Left: Re-estimated empirical variogram from the $n = 250$ data points; Right: estimated elevation for C_1 and C_2.

Figure 4 shows the reconstructed fields from the method we proposed compared to *i)* the true field and *ii)* the field reconstructed only from the 250 data points without considering the contour polylines data. A visual comparison suggests better results for the proposed approach. However, artefacts can bee seen around the pool located at the bottom-right corner of the domain. In this area, the elevation increases with the distance to the pool minimum but decreases again at three places. This behaviour can be explained by the fact that very few topographic points are available around this pool.

On the contrary, the field reconstructed with the ordinary kriging method shows greater difference to the true field especially on the top part of the domain, despite a high concentration of ground-truth surveys.

To better assess the precision given by the two approaches, the root mean square difference e between results and true field is computed. The error value obtained without considering the contour polylines is equal to 0.232 m, and to 0.199 m for the proposed method.

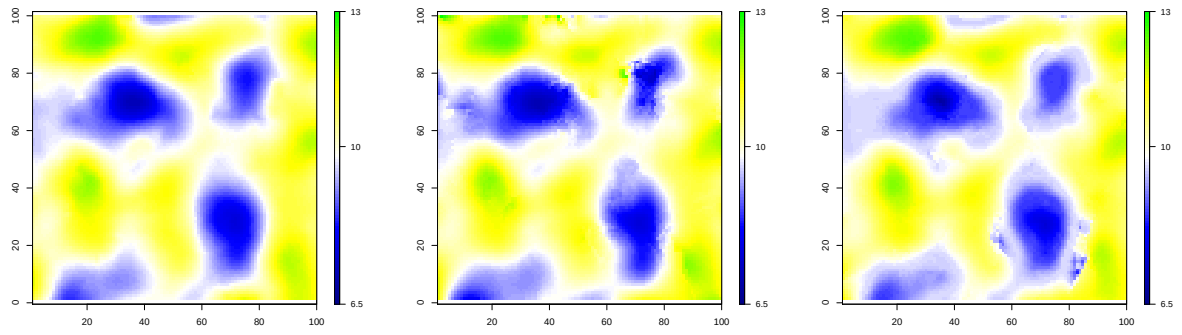


Figure 4: **Reconstructed elevation fields:** Left: true field D; Middle: ordinary kriging reconstructed field; Right: method (polyline conditional simulation) reconstructed field

This first test suggests that the sampling procedure may influence the results, at least through the location of acquisition points. To check the robustness of the proposed approach, the process described above was repeated 100 times, changing the 250 points data sampling.

Figure 5-left shows the contour polyline C_1 and C_2 error distribution obtained. Figure 5-right shows the mean distance to the true field distribution. Clearly, the proposed method using conditional simulation is more robust with few error variance. The standard deviation of the results obtained with this approach is 0.056 m (resp. 0.037 m) for the contour line 8.5 (resp. 9.5) compared to 0.183 m (resp. 0.1 m) using the ordinary kriging method.

Rank filtering occurred 10% of the times for the C_1 estimation and never for the C_2 estimation. This suggests that rank filtering is useful when contours are in extreme values.

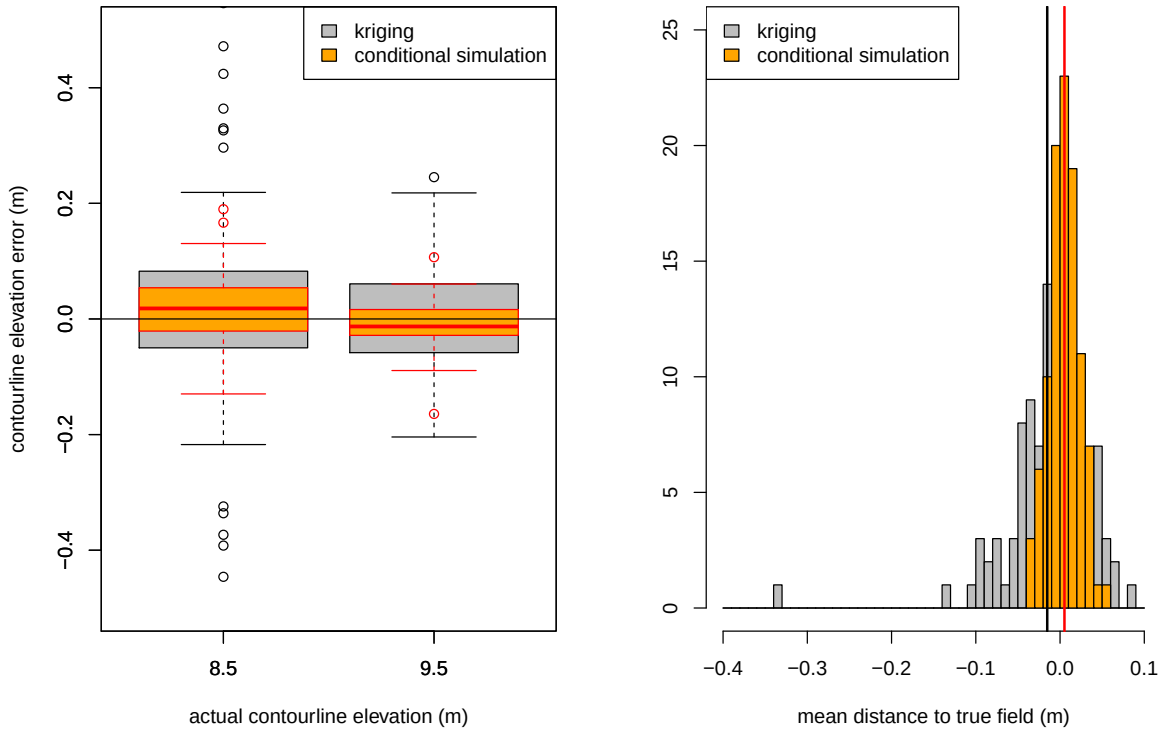


Figure 5: **Errors distribution on 100 random sample sets of 250 data points.** Left: C_j estimation error; Right : mean distance to the true field after reconstruction. The vertical lines represent the mean of the distribution.

Field survey being very costly, the aim is to reduce to the minimum the required elevation data points. The test is thus performed again with different numbers of points. An example of result obtained with a random sample of $N = 50$ points is given in Figure 6. It can be seen that the use of contour line enables to retrieve the pools shapes with better accuracy than an ordinary kriging based only on data points.

The comparison between true field and computed elevation fields with the two approaches is done using the root mean square error. These results are summarized in Table 1 for random samples obtained with $N = 10$ to $N = 250$. As expected, the error decreases if the number of available data points is high and the proposed approach always gives better results than the ordinary kriging.

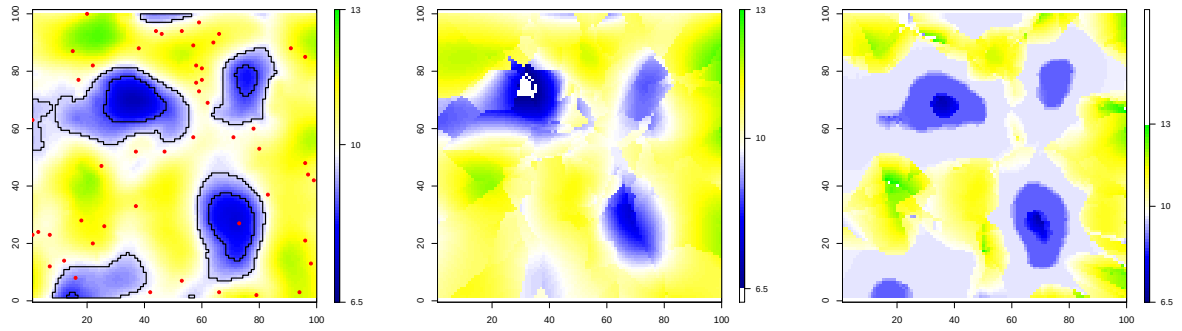


Figure 6: **Reconstructed elevation fields using only 50 data points:** Left: true field D; Middle: ordinary kriging reconstructed field; Right: method (polyline conditional simulation) reconstructed field.

	$N = 10$	$N = 50$	$N = 100$	$N = 150$	$N = 200$	$N = 250$
e_{ko}	1.064	0.635	0.361	0.403	0.435	0.232
e_{cs}	0.639	0.469	0.329	0.227	0.189	0.199

Table 1: Root mean square errors between true field and reconstructed elevation fields for different N.

IV DISCUSSION AND CONCLUSION

The objective of this work was to put forward a methodology to enhance the accuracy of the topographic description required by numerical hydrodynamic modelling, by using information of level lines available from remote-sensing data.

The first results obtained on a totally theoretical example, show that the topographic estimation benefits from such additional data; and the repetition of the process indicates that this result is robust to sampling. If the gain in accuracy may seem limited at this stage (about 3 cm in mean for 250 points), the influence of the sampling rate should be assessed more precisely on the two approaches, since the benefit taken from additional sources of information is expected to increase as less ground-truth data are available.

Moreover, a classical random sampling is obviously not the best approach to conduct a field survey. In the first presented test case, random sampling excluding the direct neighbourhood of the contour lines yielded a lack of data near the pool located at bottom-right part of the domain. This produced irrelevant estimation of the topography in some areas. Further tests are thus planed to infer guidelines in the sampling definition in order to decrease the number of data points required while maintaining a high accuracy in the results, thus enhancing the cost/benefice ratio. In the framework of hydrodynamic modelling, this work will integrate results of uncertainty and sensitivity analyses such as given in Guinot and Cappelaere (2009) or Delenne et al. (2012).

The methodology developed on theoretical test cases, will be assessed in a real-case study of the hydrodynamic modelling of the Vaccares lagoon for which amount of data are available for validation.

References

Cressie N., Johannesson G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 209–226.

Delenne C., Cappelaere B., Guinot V. (2012). Uncertainty analysis of river flooding and dam failure risks using local sensitivity computations. *Reliability Engineering and System Safety 107*, 171–183.

Guinot V., Cappelaere B. (2009). Sensitivity equations for the one-dimensional shallow water equations: Practical application to model calibration. *Journal of Hydrologic Engineering 14*, 858–861.

Hostache R., Lai X., Monnier J., Puech C. (2010). Assimilation of spatially distributed water levels into a shallow-water flood model. part ii: Use of a remote sensing image of mosel river. *Journal of Hydrology 390*(3-4), 257 – 268.

Lindgren G., Rychlik I. (1995). How reliable are contour curves? confidence sets for level contours. *Bernoulli*, 301–319.

Ogilvie A., Belaud G., Delenne C., Bader J.-C., Oleksiak A., Bailly J. S., Ferry L., Martin D. (2015). Decadal monitoring of the Niger Inner Delta flood dynamics using MODIS optical data. *Journal of Hydrology 523*, 358–383.

Saito H., Goovaerts P. (2000). Geostatistical interpolation of positively skewed and censored data in dioxin-contaminated site. *Environmental Science Technology 34*, 4228–4235.

Schumann G., Matgen P., Hoffmann L., Hostache R., Pappenberger F., Pfister L. (2007). Deriving distributed roughness values from satellite radar data for flood inundation modelling. *Journal of Hydrology 344*(1-2), 96 – 111.

Wameling A., Saborowski J. (2001). Construction of local confidence intervals for contour lines. In *Proceedings of the IUFR4.11 Confererence in forest biometry*.

Yamamoto J. K. (2000). An alternative measure of the reliability of ordinary kriging estimates. *Mathematical Geology 32*, 489–509.

Yamamoto J. K. (2007). On unbiased back-transform of lognormal kringing estimates. *Computacional Geosciences 11*, 219–234.

Monitoring spatial accuracy of oil palm cultivation mapping in southern Cameroon from Landsat series images

Prune Christobelle Komba Mayossa[1*], **Sébastien Gadal**[1]

[1]Aix-Marseille Université, CNRS ESPACE UMR 7300, France

*Corresponding author: prune.komba-mayossa@etu.univ-amu.fr

Abtract

Studying and mapping palm grove evolution allow understanding the impact related to its cultivation. Our study aims to map industrial palm grove using Landsat series images and measures the accuracy of the produced maps. It was carried out in SOCAPALM industrial plantation, located in southern of Cameroon. For the mapping and assessment of accuracy, per-pixel classification and confusion matrix method were used, respectively. We obtained high correlated maps (Kappa =0.92 in 2001 vs 0.86 in 2015). However, some confusions were observed between vegetation and oil palm classes for the two periods, affecting the maps accuracy. These confusions are caused by the presence of mixed pixels resulting from the spatial and spectral characteristics of palm groves, the method used to map and validate the map, and uncertainty related to dada. To increase the accuracy, we suggest (1) to use another mapping method such as super-resolution mapping, (2) develop a classification system of cartographic products.

Keywords

Elaeis Guineensis.Jaq, oil palm, remote sensing, spatial accuracy, monitoring.

I INTRODUCTION

Oil palm (Elaeis guineensis.jaq) is a perennial oleaginous plant from Central Africa. In Cameroon oil palm has a high economic importance with an industrialisation dating from colonial period (Elong, 2003). The incomes generated by oil palm cultivation have developed agro- industries such as SOCAPALM (Cameroonian Society of Palm groves). This activity in high-yield and low cost of returns (Riva, 2013), is behind of socio- environmental damages whom deforestation, loss of biodiversity, pollution, etc. Studying and mapping palm grove evolution allow understanding the impacts related to this culture. Our work is placed in the context of the long-term management of Oil palm resources in Congo basin. This paper is focused on the first part of this project: studying, mapping and monitoring SOCAPALM industrial palm groves using remote sensing. Several studies have shown the capability of remote sensing to map plant resources (Domaç and al, 2004; Li and al., 2015; Koh and al, 2011; etc.). The accuracy of the produced maps is main question for the study of evolution, mapping and characterization of palm grove, and for the reliability of our method.

This first part aimed to map palm grove and assess the spatial accuracy of produced map.

II STUDY AREA AND DATA

II.1 study area

Located in the Congo Basin, the Republic of Cameroon is a central African country with a surface area of 475,440 Km² and 20,000,000 inhabitants in 2015. Two main climatic zones are defined. In south the climate is equatorial and sub-equatorial, temperatures range between 25°C and 35°C, annual abundant precipitations up to 1000mm. In north the tropical climate is Sudanese (high temperatures and scarce rainfall) and Sahelian (irregular precipitations). SOCAPALM industrial plantations are located in south west of Cameroon, in the Ocean County, not far from the port city of Kribi (see figure 1).

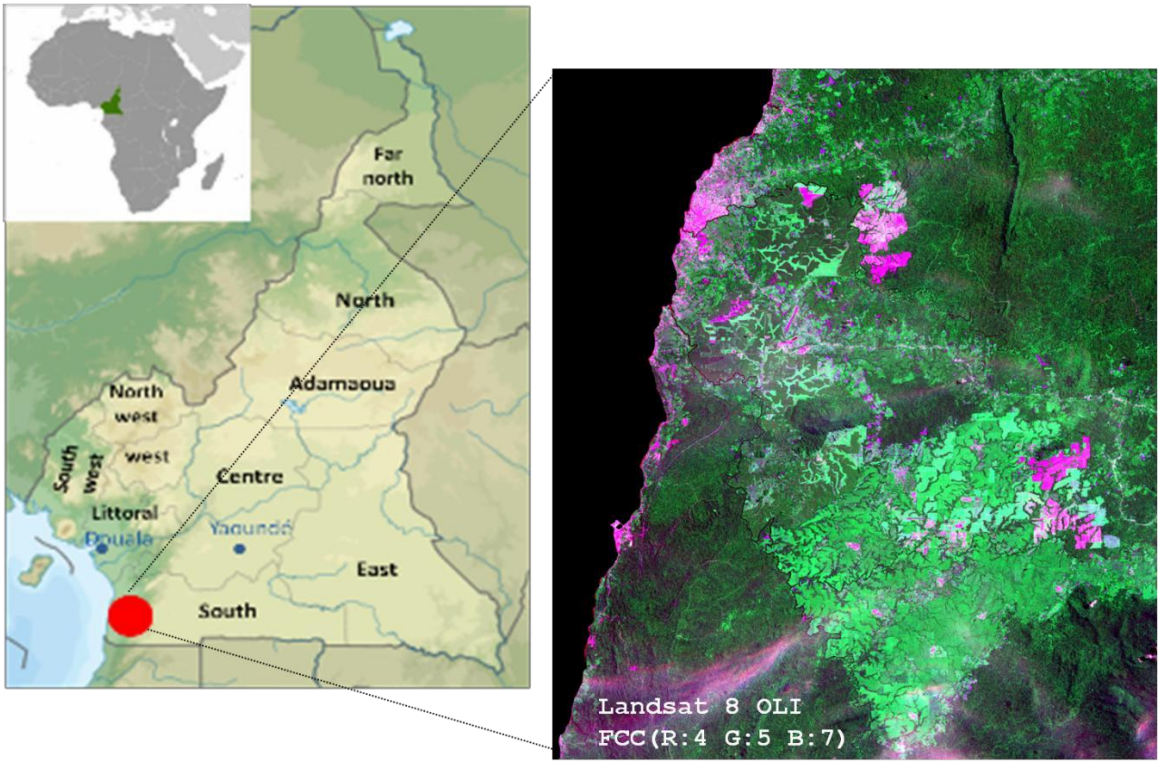


Figure 2: Localisation of study area, from CNRS ESPACE UMR 7300.

II.2 Data

Landsat 7 ETM+ and Landsat 8 OLI images of same season, but with different acquisition date were used (see Table 1). These images were acquired on the 26th April 2001 (ETM+) and 25th April 2015 (OLI), respectively.

Parameters	ETM+	OLI
Spectral range	0.45 -12.5µm	0.45-235µm
Spatial resolution	MS:30m/PAN:15m	MS:30m/PAN:15m
Swath width	183km	185km
Spatial coverage	Non-continuous	Non-continuous
Total number of bands	8	11
Mode	MS/PAN	MS/PAN
Date of acquisition of image	April 2001	April 2015

Table 1. Sensor specification of Landsat ETM^{+} and OLI.

III METHODOLOGY

Methodology adopted for this study is depicted in Figure 2. It involves mapping palm grove using spectral classification, evaluating the precision of produced maps for both Landsat images 2001 and 2015.

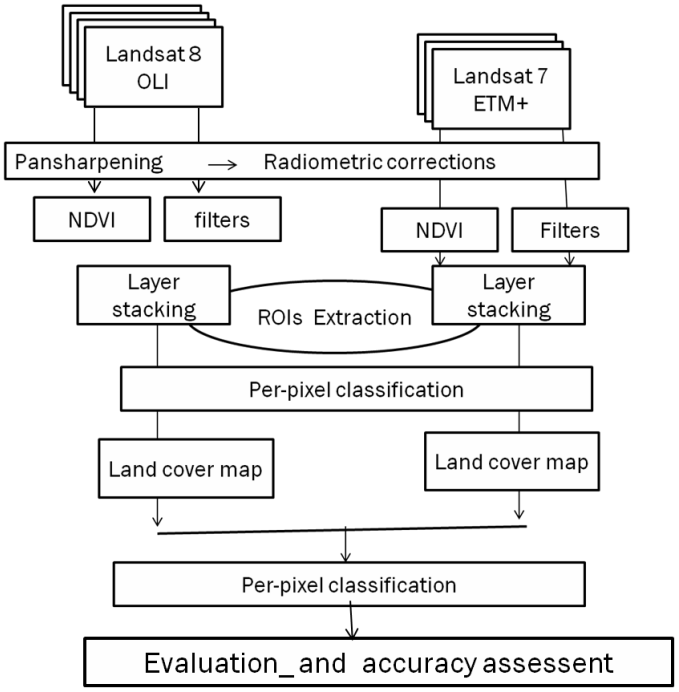


Figure 2: Flowchart of methodology adopted in this study.

III.1 Pre-processing

The pre-processing phase is composed of a chain of 6 calculs:

(i) Pansharpening method to improve the spatial resolution of the multispectral image (30m), merging it with the panchromatic image (15m), by the Brovet transform (Lacombe, 2008). The function resamples automatically the seven (Landsat 7) or eleven (Landsat 8) channels in the maximal resolution by using several methods; in our case, the cubic convolution method was used, to obtain a multispectral image of 15mx15m of resolution.

(ii) Radiometric corrections, images were calibrated in radiance by applying the equation:

L=a*CN+b, where *CN* is the digital count, *a* is the gain and *b* is the bias. The coefficients *a* and *b* of sensor calibration are given in the metadata files. The luminance

($W.m^{-2}.sr^{-1}$) was calculated for each band.

(iii) As the presence of cloud (30% of cloud cover) and water area, may disturb the analysis, they were masked.

(iv) The vegetation index was calculated according to the equation:

NDVI=NIR-RED/NIR+RED. It allows characterizing the vegetation cover in terms of level of Chlorophyll (Pouchin and al., 2002).

(v) To improve the detectability of objects, (Gadal , 2003) convolution filters were used: morphological (maximum filter) and directional filter with a kernel size of 3x3 pixels.

(vi) Eight regions of interest (ROIs) were defined according to the spatial organisation of landscape: growing, young, and mature oil palm, low vegetation, forest, bare ground/built-up areas and waterway.

III.2 PROCESSING

III.2.1 Palm grove mapping

First, we created an image with B, G, R, NIR, SWIR and NDVI channels, masked water area, extracted spectral signatures from all channels and computed radiance statistics. Second, to estimate ROIs separability, Jeffries-Matusita distance (JM) was used. This average distance between two classes (Wacker and Landgrab, 1972) takes values in the range [0-2]. Value over 1.8 indicates a very good separability, a value under 1.8 indicates poor separability.

Third, supervised maximum likelihood classification was applied on filtered channel and on other channels of the image.

III.2.2 Validation of classification

To validate produced maps, control areas were digitized with Arcgis GIS software in eleven as estimated for each image. The resulting manual classifications could then be crossed with the maximum likelihood classification result, to produce a confusion matrix and Kappa index.

III.2.3 Analysis of maps accuracy

The conventional methods of accuracy assessment of thematic maps were employed: confusion matrix (Congalton, 1991). The confusion matrix gives an overall evaluation of map accuracy and for classification results of each thematic class.

Kappa index assesses the correlation between obtained results (maps produced) and the ground truth. Kappa takes values in the range [0-1] and it's divided into five categories: very low agreement from 0 to 0.20; weak correlation from 0.21 to 0.40; moderate correlation from 0.41 to 0.60; substantial correlation from 0.61 to 0.80; high correlation from 0.81 to 1. This index (equation 1) is expressed in terms of overall accuracy observed (equation 2) and expected (equation 3).

$$K = \frac{a - b}{1 - b} \tag{1}$$

$$a = \frac{1}{N}\sum_{i=1}^{NC} x_{ii} \tag{2}$$

$$b = \frac{1}{N^2}\sum_{i=1}^{NC} (x_{i+}.x_{+i}) \tag{3}$$

Were <NC> is the number of classes; <N> the total number of observations; <x_{ii}> the number of observations in the row i; < x_{+i}> and <x_{i+}> total observations in the line i.

IV RESULTS AND DISCUSSION

IV.1 Classification by maximum likelihood and classes'separability

In general, the computed JM index to assess ROIs separability shows that, ROIs defined have good separability. Thus, JM distances are 1.99 for growing and young oil palms; but poor separability between forest class and mature oil palm. Confusions are expected between these last two classes (see Table 2).

IV.2 Accuracy and thematic confusions analysis

For both Landsat 8 and ETM+ images, maps produced have very good accuracy (see Table 2) with 90% in 2001 and 80% in 2015, respectively. Different classes of palm grove are recognized: 99.67% for growing oil palm, 80% for mature oil palm and 93.42 %, for mature oil palm in 2001. The same trend was obtained in 2015. Kappa values shows high correlation: Kappa = 0.92 (2001) and Kappa= 0.86 (2015).
However, despite of the high values of Kappa, some thematic confusions are observed between oil palm classes, or oil palm and vegetation classes. For example, we observed confusions between forest and mature oil palm (17%); or between young and growing oil palm (5.4%).
During the classification process, some pixels belonging to mature palm, where classified in to forest classes, for example.

The quality of classification is directly related to class separability, which itself depends on the variation among pixel from different classes as compared to within-class pixel variation. On one side, uncertainty to the data including several factors such as variable incidence, shading effects contribute to within class heterogeneity and alter the specificity of the signal.

On the other side, as some component may be common to different classes, (for example low vegetation pixels in the growing or young oil palm classes), distinct classes may share mixed pixels (Komba Mayossa, 2014).
On the validation map side, ground truth plays an important role in map accuracy. The knowledge of ground is an important key; the results were obtained from validation of digitized map and classification result. Landsat image, have 30mx30m of spatial resolution. The photo-interpretation is difficult especially in a heterogeneous landscape, as our study area where errors during the sampling step, cannot take to account within-plot heterogeneity. Thematic confusion caused by the presence of mixed pixels affects map accuracy (Chitroub, 2007). These mixed pixels result from spectral and spatial characteristics of the landscape studied objects, and from the method used to map.

IV.3 Prospects

The maps obtained have a good accuracy, but some thematic confusions were observed.
One limitation of spatial accuracy of oil palm grove mapping is bound to class heterogeneity, attendant mixed pixels and the method used. Several ways for improvement are possible. At first, per-pixel classification assumes that each pixel represents a single class only. Maximum likely hood algorithm, used to map palm grove, ignores the mixed pixel problem. One of the solutions is to use a technic which allows mapping at the sub-pixel scale, such as super-resolution mapping (Priyaa and sanjeevi, 2013; Muad and foody, 2010). Second, as the map validation method improves or decreases the accuracy of produced map, the development of a classification system of cartographic products can be most interesting. Indeed, further studies should be focused on the evaluation of reliability of the produced map by this classification system (Chalifoux and al, 2006).

Map 2001

GROUND TRUTH	**Growing Oil palm**	**Young Oil palm**	**Mature Oil palm**	**Low Vegetation**	**Shady Forest**	**Forest**	**Ground Built1**	**Ground Built2**	**Ground Built3**	**Waterway**	TOTAL
Growing Oil palm	**99.67**	2.87		0.01							102.25
Young Oil palm	0.33	**93.42**		5.4							99.15
Mature Oil palm			**80**		10	2.22					92.32
Low Vegetation		3		**94.59**		1.70					100
Shady Forest			2.88		**84.15**	7.07				6	100.12
Forest			1.22			**88.99**					112.45
Ground Built1							**100**				106.45
Ground Built2								**100**			100
Ground Built3									**93.55**		93.55
Waterway										**94**	94
TOTAL	100	100	100	100	100	100	100	100	100	100	**1000**

Overall accuracy 90% / Kappa=0.92

Map 2015

GROUND TRUTH	**Growing Oil palm**	**Young Oil palm**	**Mature Oil palm**	**Low Vegetation**	**Shady Forest**	**Forest**	**Ground Built1**	**Ground Built2**	**Ground Built3**	**Waterway**	TOTAL
Growing Oil palm	100										100
Young Oil palm		**100**		0.04							100
Mature Oil palm			**99.51**		1.58						101.13
Low Vegetation				**100**							100
Shady Forest					**97.55**	12					99.15
Forest			0.49		2.01	**97.22**					99.63
Ground Built1							**100**				100
Ground Built2								**100**			100
Ground Built3									**100**		100
Waterway										**100**	94
TOTAL	100	100	100	100	100	100	100	100	100	100	1000

Overall accuracy 80% / Kappa=0.89

Table 2. Confusion matrix and Kappa index (Land cover map 2001 and 2015)

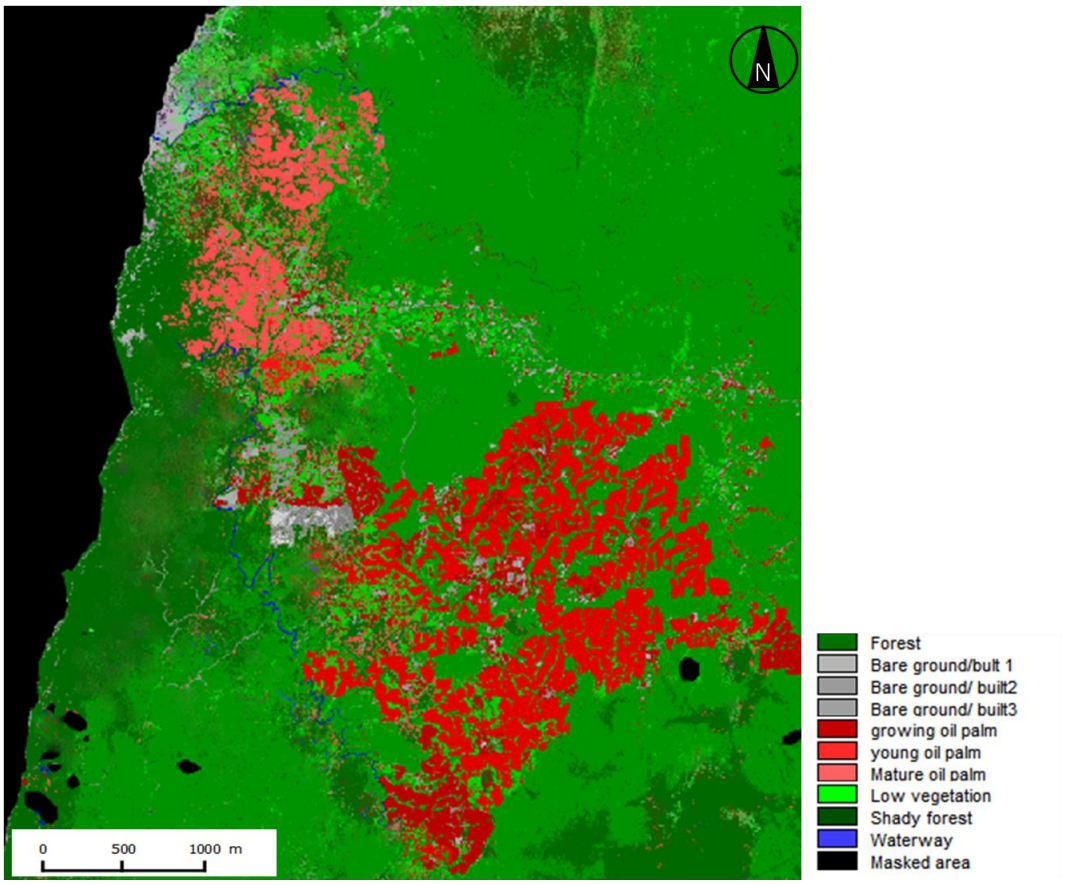


Figure 3: Landsat map in 2001(Landsat 7 ETM+), from CNRS ESPACE UMR 7300.

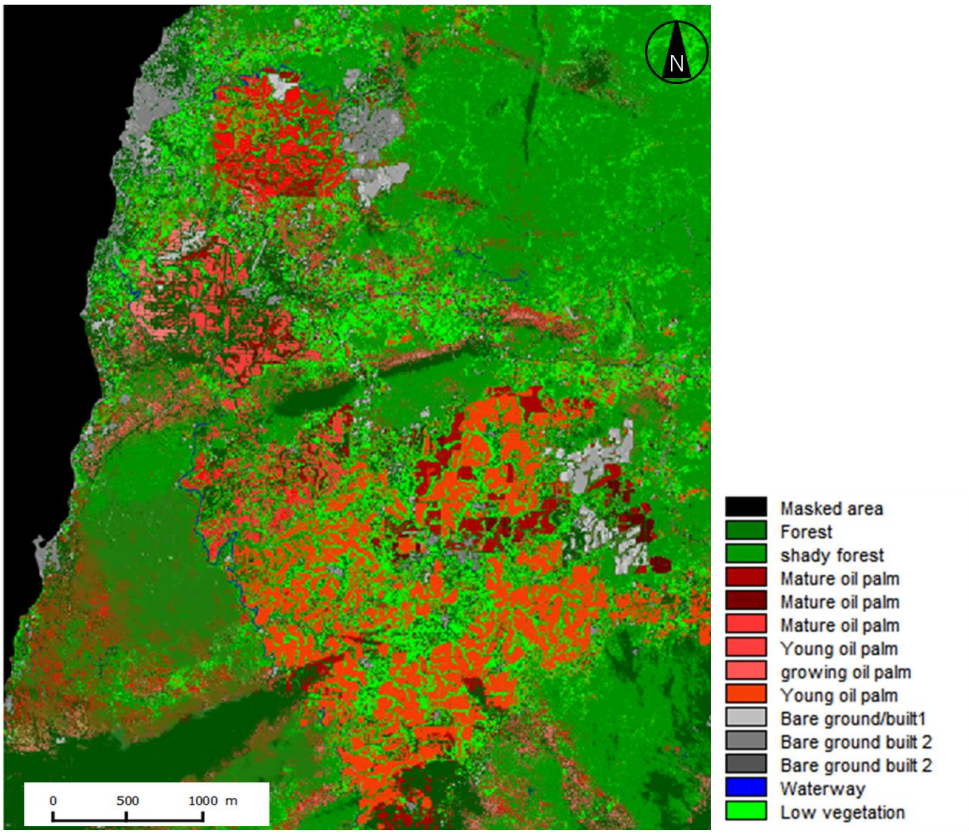


Figure 4: Landsat map in 2015 (Landsat 8 OLI-TIRS), from CNRS ESPACE UMR 7300.

References

Elong J.G. (2003). Les plantations villageoises de palmier à huile de la Socapalm dans le bas-Moungo (Cameroun): un projet mal intégré aux préoccupations des paysans. *Les Cahiers d'Outre-Mer. Revue de géographie de Bordeaux* 56(224), 401-418.

Chitroub S. (2007). Annalyse des composantes indépendantes d'images multibandes: Faisabilité et perspectives. *Revue de télédétection 7(1-4)*, 3-4.

Gadal S. (2003). *Reconnaissances multi-niveaux d'unités paysagères par segmentation automatique d'images satellites.* Télédétection des informations géographiques. Editions Anne-Elisabeth LAQUES.

Rival A., Levang P. (2013). *La palme de controverse : Palmier à huile et enjeux de développement,* édition Quae.

Priyaa A., Sanjeevi S. (2013). Super resolution mapping of multispectral and hyperspectral images of peechi reservoir, south india. *Image, 2010*.

Wacker A. G., Langreb D.A., 1972: Minimum distance classification in remote sensing. *LARS Technical Reports, 25 pp.*

Lacombe J.P. (2008). *Initiation au traitement d'images satellites: travaux dirigés, cahier2.*Ecole nationale Supérieure Agronomique de Toulouse.

Pouchin T., Debriej., Bourcier A. (2002). L'observation de la végétation de l'Afrique de l'Ouest par télédétection spatiale: l'apport de l'indice de végétation normalisé. *Science et changements planétaires/Sécheresse* 13(3) ,187- 94.

Koh L. P., Miettinen J., Liew S. C., Ghazoul, J. (2011). Remotely sensed evidence of tropical peatland conversion to oil palm. *Proceedings of the National Academy of Sciences*, *108*(12) , 5127-5132.

Li L., Dong J., Njeudeng Tenku, S., Xiao X. (2015). Mapping Oil Palm Plantations in Cameroon Using PALSAR 50-m Orthorectified Mosaïc Images. *Remote Sensing 7(*2), 1206-1224.

Domaç A., Zeydanli U., Yesilnacar E., Suzen M.L., 2004: Integration and usage of indices, feature components and topography in vegetation classification for regional *,20th ISPRS Congress, Istanbul,pp. 204-208.*

Editors: Jean-Stéphane Bailly, Daniel Griffith & Didier Josselin
Publisher: UMR 7300 ESPACE – Avignon, Case 41, 74 rue L. Pasteur, 84029 Avignon Cedex, France
Series: Actes Avignon - ISBN: 978-2-9105-4510-5 - Juillet 2016
Design: N. Brachet (UMR ESPACE – CNRS)
https://colloque.inra.fr/spatial-accuracy2016