Within-field zoning using a region growing algorithm guided by geostatistical analysis

Lamia Zane, B. Tisseyre, S. Guillaume, Brigitte Charnomordic

Within-field zoning using a region growing algorithm guided by geostatistical analysis

L. Zane[1], B. Tisseyre[2], S. Guillaume[2], B. Charnomordic[1]
UMR MISTEA, INRA, building 26, 2 place Viala, 34060 Montpellier,
UMR ITAP, Montpellier SupAgro/Cemagref, building 21, 2 place Viala, 34060 Montpellier, France. Phone : +33-499-61-23-35. Fax : +33-499-61-24-36

Abstract

Region growing methods are of potential interest to define within-field zones and resulting site-specific management. These methods are unsupervised and based on regions which grow from the initial seeds according to homogeneity criteria. However, the determination of seed number and seed locations has strong repercussions on the zoning output. This paper proposes an approach to allow knowledge inherited from geostatistical analysis to guide the seed initialization (seed number and location) of a region growing-based segmentation method. In this study, the segmentation method is a general division/merging procedure, which in this study is used for merging and region growing. An original point is the possibility to use it either for irregularly located data or data arranged on a regular grid. The initialization of the segmentation method is guided by a prior analysis where a few parameters of the semi-variogram model are used to set i) the number of seeds required to initialize the region growing procedure, ii) to decide their relative locations (i.e. minimal distance between seeds) and iii) to identify potential outliers as seeds that may flaw the growing procedure.
Both methods were tested on two data sets : i) three hypothetical fields of known distribution and known spatial organization ii) a real field where yield monitor data were obtained. A qualitative analysis of the results is presented, as well as the evolution of the variance explained by the model.

Key words : zoning, segmentation algorithm, growing region, geostatistics

Introduction

The concept of precision agriculture relies on the differential management of within-field zones. Therefore zoning using high or medium resolution data is an essential step to implement differential treatments. In precision agriculture, zoning is mainly based on algorithms of unsupervised classification (Vrindts et al., 2005, Taylor et al., 2007). Therefore, classification methods are the most used tools for management zone delineation. The most common implementations are based on the k-means algorithm or its fuzzy extension (fuzzy c-means). These classification algorithms yield classes, which require a projection onto the geographical space to define zones. This transformation is likely to lead to a fragmentation of this space, which may be incompatible with the management constraints.
An alternative to classification is the introduction of segmentation algorithms that are routinely used in image processing. Segmentation methods differ from classification algorithms as they aim to produce a division of the geographical space which is a direct mapping onto homogeneous zones. Segmentation can be defined as the process of dividing an image into a set of connected and homogeneous regions. Each region is

different from its neighbors; the union of two adjacent regions must be homogeneous. Two main groups of segmentation methods can be distinguished: border oriented methods and region oriented methods. In the field of Precision Agriculture, there are no predefined objects to identify, so only the region-based methods apply, like the one proposed by Pedroso et al (2010). Among segmentation methods, region growing is of potential interest for precision agriculture. In these methods, regions grow from the initial seeds according to a homogeneity criterion. Indeed, they are unsupervised and they are dedicated to define homogeneous regions of the space which fits with requirements of within-field zones delineation for site-specific management. However, they present one major drawback : the map resulting from the zoning process highly depends on the initialization step; the seed number and their locations. This paper proposes to test an approach to allow knowledge inherited from geostatistical analysis to guide the seed initialization (seed number and location) of a region growing-based segmentation method. The segmentation method is a general division/merging procedure, which in this study is used for merging and region growing. An original point is the use of this method on irregularly located instead of data arranged on a regular grid.

Commentaire [J1]: Not in reflist

Material and Methods

Segmentation algorithm :

The seeded region growing approach aims to segment an image into regions with respect to a set of Z seed regions (Adams and Bischof, 1994; Mehnert and Jackway, 1997). Each seed region is a connected component comprising one or more points and is represented by a set Z_i, where i= 1,2,. . . ,Z. Let T be the set of all unallocated points that border at least one of the Z_i, i.e.

$$T=\{x \notin \bigcup_{i=1}^{Z} Z_i : N(x) \cap \bigcup_{i=1}^{Z} Z_i \neq \emptyset\}, \quad (1)$$

Where N(x) represents the set of neighbours of the point x. A single step of the algorithm involves examining the neighbours of each x $\in$ T in turn.

Finding zone neighbours on an irregular grid requires a specific process, which is a major difference when compared with segmentation methods in image analysis, see Pedroso et al. (2010). The Voronoi tessellation is used to convert each data point to a zone and to define the initial neighbourhood. Each point is associated with a unique polygon. Initially there are as many zones as data points, and the zone neighbourhood is equivalent to the point neighbourhood. At each step of the zone merging process, the zone neighbourhood is updated by considering all zones that share a vertex in the tessellation as neighbours.

Commentaire [J2]: Not in reflist

If N(x) intersects a region Z_j then a similarity measure, $\delta(x)$ of the difference between x and the intersected region is calculated. In the simplest case $\delta(x)$, is defined as indicated in relationship 2.

$$\delta(x)=| g(x)-mean_{y\in Z_j}\{g(y)\}|, \quad (2)$$

where g(x) is the value of the point x. If N(x) intersects more than one region then Z_j is taken to be that region for which $\delta(x)$ is a minimum. In this way a $\delta(x)$ value is determined for each x € T. Finally, The point s $\in$ T that satisfies :

$$\delta(\mathrm{s}) = min_{x \in T}\ \delta(x) \qquad (3)$$

is appended to the region corresponding to $\delta(s)$. The new state of the regions $\{Z_i\}$ then constitute the input to the next iteration. This process continues until all points have been assimilated.

Expert constraints derived from semi-variogram and data distribution

The segmentation method requires the determination of parameters which are difficult to identify without prior knowledge of the field:
- *Z*, the expected number of zones which determines the number of seeds,
- L_i the location of seed *i*, this location has i) to avoid potential outliers, ii) to take into account other seed locations and iii) to consider data values so that L_1, L_2,..,L_Z the locations of seed *1,2,...,Z* are representative of the magnitude of variation of the field.

The use of the semi-variogram was considered relevant to determine most of these parameters. Indeed, the semi-variogram model is interesting since it simultaneously summarizes erratic phenomenon and spatial organisation of the regionalised variable under study.

a. Number of seeds

The range (*r*) of the semi-variogram defines the distance beyond which there is no more spatial correlation. Beyond that distance, observed values on two different locations are considered independent and assumed to belong to two different zones. The average size of expected within-field zones was therefore determined from the range of the semi-variogram following equation 4. In a first approach, zones were considered round (circle) to simplify computation.

$$Z = \frac{A}{S_z} \text{ with } S_z = \pi\left(\frac{r}{2}\right)^2 \qquad (4)$$

with
Z : the number of seeds,
A : the area of the field under consideration,
Sz : the average area of the within-field zones,
r : the range of the semi-variogram,

b. Seed location

To make sure that the seeds cover the range of variation of the data, a constraint is introduced on two of the *Z* seeds. The first seed is randomly chosen among points whose value belongs to the first 20% percentiles of data distribution. A similar procedure is used for the second seed which is chosen among the last 80 % percentiles. For the remaining seeds, their relative positioning is constrained by two conditions. On the one hand, distance between two seeds location has to be higher than the range of the semi-variogram to maximize the chance that each seed is positioned on an independent zone of the field. On the other hand, the distance must not greatly exceed the range of the semi-variogram in order to be able to position the *Z* seeds. As a result, optimal distance between two seeds is chosen as the range of the semi-variogram (Equation 5).

$$\forall L_i, L_j \in F \text{ with } i \neq j,\ d(L_i, L_j) > r \tag{5}$$

where
$L_{i;\,i=1,2,\ldots,Z,}$ the location of the seeds,
F : set of possible locations corresponding to the field,
r : the range of the semi-variogram,
$d(L_i,L_j)$: Euclidian distance between locations L_i and L_j,

Region growing is conditioned to the value of the variance between the region and its neighbours. Positioning a seed on an outlier would necessarily entail a significant variance (V) between the seed and its neighbors and the inability for this seed to grow to form a zone. To avoid outliers, a maximum variance value (Vs) between a seed and its neighbours was introduced. When V is higher than Vs, neighbours are alternatively considered as seed until a variance lower than Vs is observed. Vs is determined from the semi-variogram and the data resolution following equation 6.

$$V_s = \gamma(d_a) \text{ with } d_a = 2.\sqrt{\frac{1}{\pi.R}} \tag{6}$$

where
d_a : average distance between two points over the fields,
R : data resolution over the field (number of points per square meter),
γ : theoretical (model) semi-variogram of the field,

Objective evaluation criterion

A criterion is necessary to provide an objective evaluation of the zoning, which is repeated 10 times to obtain a fair evaluation, the zoning results being dependent on the initial seed location, which is subject to a random effect.
Let I be the number of repetitions of the zoning procedure. Each repetition generates Z zones, one zone for each seed, denoted by $Z_1,Z_2,\ldots Z_Z$.
The criterion is the R^2 coefficient, averaged over the I repetitions. For a given zoning, R^2= 1-SSL/SST, where SST is the total inertia of the variable of interest X, and SSL is the inertia of X explained by the Z-zone model. It represents the amount of explained variance.

Data

Two types of data were used to test the proposed segmentation procedure:

- Yield monitor data obtained from a sensor mounted on a grape-harvesting machine (Pellenc S.A). The field is 1.4 ha of Bourboulenc variety harvested in 2001 in Provence (France). The average sampling rate is about 2400 measurements per ha. Yield data were not kriged.
- Hypothetical fields of known spatial variability obtained from a simulated annealing procedure (Goovaerts, 1997). Parameters of the hypothetical fields have been chosen according to grape fields already harvested with real-time yield monitors (Taylor et al., 2005). For all the fields of our data-base, the theoretical semi-variogram is an exponential model where the nugget effect is approximately one third of the sill. We decided to apply a nugget effect of 5 and a sill of 16 (arbitrary unit); that means the

different fields differ only by the range of their semi-variogram. These theoretical fields were presented in a previous work (Tisseyre and McBratney, 2008). Three fields were used with practical ranges for the semi-variogram model of 27, 36 and 45 m. All of the fields have an area of 1 ha (100 × 100 m) and an original resolution grid of 1 m. In order to fit with classical monitored data (i.e. yield data) resolution was decreased to 2000 points ha^{-1} by a random selection. Data are not regularly distributed over the fields.

Results and discussion

Numerical values derived from semi-variogram and data distribution

Table 1 shows, for each of the fields under study, the main characteristics extracted from both data distribution and semi-variogram analysis. Data are assumed to be second order stationary variables, and the exponential semi-variogram model is used. The range is reported in the third column. A low value indicates a poor data structure and is associated with a high number of potential zones, given in column number 4. These two values are used by the segmentation algorithm: the maximum number of zones is the number of seeds while the range is the minimum distance between seeds. Table 1 illustrates the diversity of these data: the number of zones ranges from 3 to 18, and the range is between 27 and 80 m. The seeds have to fill another constraint: their local variance should not be higher than *Vs*, defined in Equation 6 and reported in column number 5 of Table 1. The distribution tails, columns 6 and 7, used to select two of the seeds, are, as expected, the same for all of the hypothetical fields.

Table 1 : Estimated parameters for seed number and location for each field

Field	Area (m^2)	Range (m)	*Z* (number of zones)	*Vs*	20 %	80 %
a	10000	45	7	7.5	< 96.87	> 103.08
b	10000	36	10	7.9	< 96.87	> 103.08
c	10000	27	18	8.5	< 96.87	> 103.08
d	9520	80	3	19	<13.83	>23.78

Results of the zoning

Let us take *Field a* as example, and compare the zoning resulting from one repetition, when expert-guided zoning is performed (Figure 1-a) and when random zoning is used (Figure 1-b). As the algorithm does not require the data to be on a regular grid, it is run with the raw data, without any pre-processing step, such as kriging which would introduce artificial information. There is the same number of seeds, therefore the same number of zones, so the only difference in the two maps comes from the seed location. Figure 1 shows the seed location impact on the zoning results. Some of these zones are very similar, zone #5 in (a) and zone #4 in (b). This zone makes sense, and when a seed is included in the area, the algorithm lets it grow. In some cases, a zone in one map corresponds to a set of zones in the other. The eastern part of map, zone #5 in (b) is divided in to three zones, #2, #4 and #6 in (a). Zones #2, #3 and #4 of (b) are part of zone #3 in (a).

From a qualitative point of view, let us point out that our algorithm allows, by introducing some knowledge related to the data spatial distribution, to spread the seeds in function of the observed within field variability. In particular, it avoids the choice of seeds that would be too close to each other, and would result in the formation of small zones, which would have no relationship with the field spatial structure. Such an undesirable behaviour is obvious in Figure 1b, when examining zones 2 and 4 on the one hand, and zones 1 and 7 on the other hand.

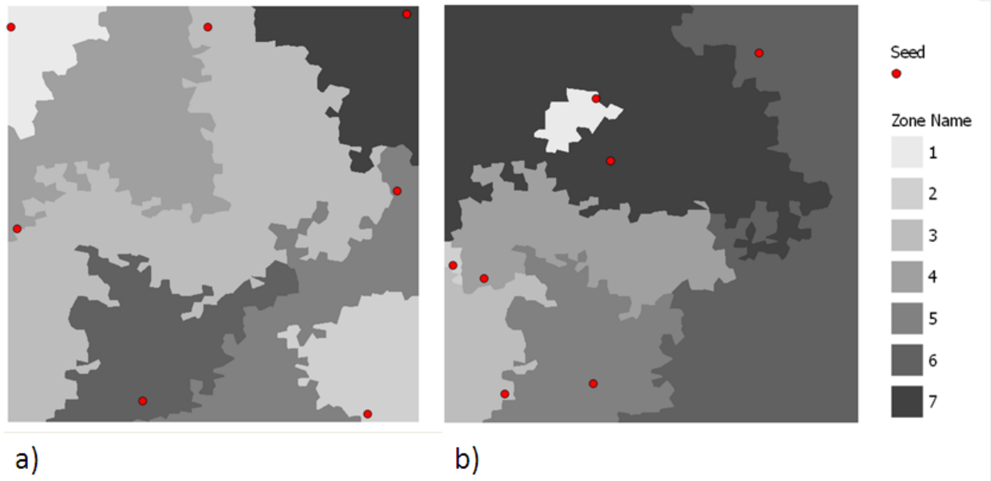

Figure 1 : Result of the segmentation of the field a) in 7 zones a) with seed locations defined by criteria derived from the semi-variogram and the data distribution and b) with random seed positions.

Table 2 : Mean R^2 over ten repetitions for each field and both segmentation methods: seed randomly positioned over the fields, and seed location constrained by criteria defined from semi-variogram and data distribution.

field	Zones number	mean R^2 (with random seed locations)	mean R^2 (with constrained seeds locations)
a	7	0.39	0.45
b	10	0.24	0.27
c	18	0.27	0.26
d	3	0.22	0.27

Table 2 shows the mean over 10 repetitions of the computed criterion, R^2, for the 4 fields and both approaches. In all cases except *Field c*, the R^2 is higher when the seed locations are guided by the data distribution and the semi-variogram analysis. Field c is the one with the weakest spatial structure, so it has the highest number of zones. Note that it is only a general trend, which allows us to demonstrate the interest of the proposed approach.

The R^2 distribution shows a high dispersion, especially when the seed location results from a random selection. Therefore the means presented in Table 2 are not significantly different from a statistical point of view.

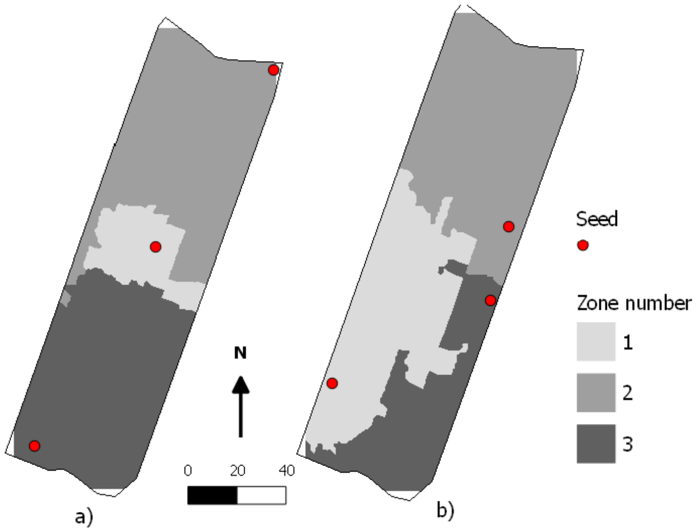

Figure 2 : Result of the segmentation of the field with real data in 3 zones a) with seed locations defined by constraints derived from the semi-variogram and the data distribution and b) with random seed locations.

Commentaire [J3]: Mark units on length scale

Other observations (vigour, resistivity, soil depth, etc.) on the field presented in Figure 2 have shown that the variation in yield was due mainly to variation in the soil and soil-water availability (Tisseyre and McBratney 2008). The southern end of the field has strong growth of the vines, a large yield and deep soil, whereas the northern end has a small yield, plants with weak growth (low vigour) and light soil. For this particular field, yield maps should be a good indication of how to perform site-specific management. Expert-guided segmentation (Figure 2-a) leads to seeds spread along the field and three zones with a yield mean of 15, 22 and 25 Mg ha^{-1} for zones 2, 1 and 3 respectively. This segmentation fits with the known north-south gradient of variability of this field. Random seed locations (Figure 2-b) lead to a different segmentation which is less convergent with the known spatial variability of the field, two zones (3 and 1) present yield means which are not different from a statistical point of view: respectively 23 and 24 Mg ha^{-1} for zone 1 and zone 3.

In this particular case, expert-guided segmentation better corresponds to the site-specific management performed by the grower. This site-specific management consisted in i) grassing zone 3 to decrease the water availability to the vine in order to reduce the yield and the vigour of the plants, consequently, this would increase the quality (sugar content, flavour compounds, etc.), ii) decrease the fruit load after pruning on zone 2 and iii) leave zone 1 as it was. Zone 1 was considered as a buffer zone between zones 2 and 3.

Conclusion

This paper deals with a first attempt to include expert knowledge, derived from data distribution and geostatistical analysis into a segmentation algorithm for management zone delineation. Let us note that an interesting contribution of the presented approach appears to be the proposal of a consistent number of zones depending on the field spatial variability. A rigorous study of that contribution is necessary, and it requires the comparison of the proposed zoning with other ones, for various numbers of zones. That comparison is far from being trivial. It implies the definition of more sophisticated evaluation criteria than the simple one used here, based on the variance explained by the zoning procedure. The first results on hypothetical and real world data sets are encouraging, but should be confirmed on other data. Quantitative and qualitative analysis of the results have been reported. Qualitative analysis shows a better behaviour of the expert-guided zoning. Even so, there are a lot of random effects even in the expert-guided procedure. Nevertheless, it is difficult to find criteria that support this subjective assessment. In this paper, we proposed a simple analysis based on the part of variance explained by the zoning (R^2). This work also highlights the need for new map comparison indices, which take into account not only the R^2 criterion but also the zone shapes and sizes in expert-guided and random maps.

References

Commentaire [J4]: Re-arrange refs to be in alphabetical order of first author

Goovaerts P., 1997. Geostatistics for Natural Ressources Evaluation, Applied Geostatistics Series, Oxford University Press, New York.

Taylor, J., Tisseyre B., Bramley R., Reid A., 2005. A comparisaon of the spatial variability of vineyrad yield in European and Australian production systems. In: Proceedings of the 5th European Conference on Precision Agriculture, Ed. J. Stafford, Wageningen Academic Publishers, The Netherlands, pp 907-915.

Tisseyre B., McBratney A.B., 2008. A technical opportunity index based on mathematical morphology for site-specific management using yield monitor data : application to viticulture. Precision Agriculture 9 (1-2) 101-113.

Adams R.and Bischof L. 1994. Seeded Region Growing. IEEE Transactions on Pattern Analysis and Machine Intelligence archive 16 6.

Mehnert A. and Jackway V. 1997. Improved seeded region growing algorithm. Pattern Recognition Letters 18 (10) 1065–1071.

Vrindts, E., Mouazen, A., Reyniers, M., Maertens, K., Maleki, M., Ramon, H., and de Baerdemaeker, J. (2005). Management zones based on correlation between soil compaction, yield and crop data. Biosystems Engineering 92(4) 419-428.

Taylor J., McBratney A.B., and Whelan B., 2007. Establishing management classes for broadacre agricultural production. Agronomy Journal 99 1366–137.

Commentaire [J5]: Not quoted in text