Exploration spatio-temporelle de l'Information statistique territoriale avec ses métadonnées

Christine Plumejeaud

▶ **To cite this version:**

Christine Plumejeaud. Exploration spatio-temporelle de l'Information statistique territoriale avec ses métadonnées. CIST2011 - Fonder les sciences du territoire, Collège international des sciences du territoire (CIST), Nov 2011, Paris, France. pp.354-360. hal-01353200

HAL Id: hal-01353200
https://hal.science/hal-01353200

Submitted on 10 Aug 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Exploration Spatio-Temporelle de l'Information statistique territoriale avec ses Métadonnées

AUTEURS
Christine PLUMEJEAUD, COGIT (France)

RÉSUMÉ
L'information statistique territoriale présente un haut niveau d'hétérogénéité, du fait de la multiplicité des producteurs, des échelles de mesure, et des fréquences de collecte. La compréhension de cette information sous-tend de disposer de modes de représentation intuitifs (cartes, courbes), et d'avoir des moyens efficaces pour comparer ces données entre elles, et ceci dans toutes leurs dimensions : spatiale, temporelle et thématique.
En réponse à cette problématique, nous avons réalisé un outil de géovisualisation de cette information, permettant de détecter des valeurs exceptionnelles. Nous détaillons ici les avantages de cette solution, qui implique la mise en œuvre de méthodes géostatistiques, et l'accès pour l'utilisateur aux métadonnées dans une interface spatio-temporelle. Egalement, les difficultés soulevées sont discutées en tant que nouvelle perspective de recherche.

ABSTRACT
Territorial statistical information can be very heterogeneous and hard to analyze, due to the multiplicity of data sources. This paper presents a platform that has been developed to find and analyze outliers. This exploration combines the spatial and temporal dimensions of a large set of thematic data, and shows their associated metadata. Dedicated to outliers' detection and their visualization, this platform makes a large use of geostatistical methods that have been written with R. Basing on this experiment, we draw some conclusions concerning some aspects dealing with the selection of the right spatio-temporal context for analysis in time-enabled tools.

MOTS CLÉS
Information statistique territoriale, hétérogénéité, métadonnées, géostatistique, géovisualisation.

INTRODUCTION
L'abondance et la diversité de l'information statistique territoriale aujourd'hui disponible représentent un potentiel riche pour l'aménagement du territoire. Cependant cette information présente un haut niveau d'hétérogénéité, du fait de la multiplicité des producteurs, des échelles de mesure, et des fréquences de collecte. Dans ces conditions, l'analyse et la comparaison de ces données est rendue plus difficile.
L'analyse exploratoire de données, (*Exploratory Data Analysis* en anglais) propose des méthodes et des outils qui peuvent répondre à cette problématique. En effet, de nombreux produits issus de ce domaine permettent d'activer des méthodes d'analyses statistiques sur des données, et la visualisation des résultats de ces analyses dans des interfaces de géovisualisation interactives. Cependant, ces solutions, en se concentrant sur les capacités statistiques et exploratoires des outils, ont le plus souvent ignoré les métadonnées. Or les métadonnées comportent des informations extrêmement utiles

concernant les modalités de production de l'information, et permettent de mieux prendre en compte l'hétérogénéité de l'information. Par exemple, si l'analyse s'oriente vers la recherche de valeurs exceptionnelles, les métadonnées peuvent certainement contribuer à expliquer les comportements atypiques de certaines valeurs.
Notre proposition vise donc à souligner l'intérêt d'un accès simultané aux données et aux métadonnées dans un outil d'analyse et d'exploration spatio-temporelle, s'interfaçant avec une base de données d'information statistique territoriale. L'outil qui a été développé en Java, offre un couplage avec des méthodes géostatistiques développées avec R, et permet de repérer des valeurs exceptionnelles dans les jeux de données statistiques. La première section de cet article expose en détails les motivations de cette proposition, en liaison avec la critique des travaux existants dans le domaine de l'ESDA. La seconde section détaille la démarche et les possibilités d'analyse offertes par cet outil. La troisième section discute des nouveaux problèmes que soulèvent l'interrogation et la combinaison d'une information aussi hétérogène.

1. POUR UNE EXPLORATION CONJOINTE DES MÉTADONNÉES ET DES VALEURS EXCEPTIONNELLES

Notre objectif vise à proposer des modèles et des méthodes pour analyser et comparer une information statistiques territoriales très hétérogène, issue de sources diverses telles que les producteurs nationaux, comme l'INSEE, ou supra-nationaux comme l'ONU ou EUROSTAT, à toutes les échelles géographiques, et sur des temporalités variables. La comparaison de ces valeurs est rendue difficile par l'usage de méthodes de mesure et de transformation des données différentes, même lorsque la définition des indicateurs dont elles rendent compte est partagée. La découverte de valeurs exceptionnelles, c'est-à-dire qui divergent fortement des valeurs de leur voisinage (géographique, temporel, thématique) présente un intérêt double. D'une part, des valeurs thématiquement intéressantes peuvent être identifiées plus vite, d'autre part, les erreurs de mesure peuvent être filtrées.
Cependant, pour discerner ce qui relève de l'erreur, il est nécessaire d'accéder à une description complète de la source des données, c'est-à-dire à leurs métadonnées. En ce basant sur un profil de la norme ISO 19115 spécifique pour l'information statistique territoriale (Plumejeaud et al. 2010), qui décrit les informations relatives à un jeu de données, à chacun des indicateurs, et à chacune des valeurs, il est possible de connaître le lignage des valeurs, mais également d'accéder à des informations relatives à leur fiabilité supposée.
Notre objectif s'inscrit dans celui de l'EDA, une discipline établie par Tukey (1977), qui vise à détecter et décrire des formes, des tendances et des relations entre les données. Le processus d'exploration des données est interactif, itératif et dynamique, c'est-à-dire que l'utilisateur prend une place centrale dans ce processus car il raffine son questionnement au fur et à mesure de ses interactions avec le système. Les capacités statistiques sont primordiales pour un outil d'analyse spatio-temporelle exploratoire. La reconnaissance, l'analyse et la mesure des formes d'association spatiale par le calcul de l'autocorrélation spatiale est une des fonctionnalités les plus classiques (Anselin,1993). Il s'agit également de disposer de méthodes pour la comparaison de différentes évolutions temporelles en vue d'identifier les différentes formes d'évolution (Andrienko, 2005). Concernant l'interface, l'EDA défend également le concept des vues multiples (des cartes, des courbes, des graphiques par exemple) et synchronisées pour un même sous-ensemble de variables (Monmonier, 1989).
Ces principes d'interactivité, de vues multiples, et la mise en œuvre de méthodes statistiques sont repris dans les principaux outils d'EDA qui existent à l'heure actuelle. La plupart peuvent être réutilisés, soit dans leur ensemble, soit comme des

composants, pour la recherche de valeurs exceptionnelles. Des outils d'analyse spatiale comme SADA, Geoda, CrimeStat, QuantumGis, TerraLib, GRASS GIS proposent des fonctions d'analyse statistique spatiale, couplées à des fonctions de visualisation et d'exploration de données. Certain de ces logiciels offrent la possibilité d'intégrer des scripts pour l'analyse statistique, programmés avec R (www.r-project.org), un langage libre privilégié par de nombreux statisticiens.
Ce domaine fertile a produit de nombreux outils et méthodes pour l'exploration des données, mais sans tenir compte de l'hétérogénéité de ces données. En effet, aucun de ces outils ne fournit d'informations sur les métadonnées sous un format non textuel, par exemple, au moyen de cartes ou de représentations interactives qui permettraient à l'utilisateur de mettre facilement en relation les informations collectées sur le jeu de données qu'il analyse avec les résultats calculés. Ces logiciels ignorent tout à fait la présentation des métadonnées associées aux données, puisque le schéma d'importation des données n'intègre pas l'import des métadonnées.

2. RECHERCHE ET ANALYSE DE VALEURS EXCEPTIONNELLES AVEC ESTIM

L'outil que nous proposons, ESTIM, développé en Java, a pour objectif l'identification de valeurs exceptionnelles, via la combinaison de plusieurs types d'analyses (géo)statistiques. Il propose en outre de confronter la multiplicité de ces analyses aux métadonnées, dans un mode interactif, permettant ainsi de mieux discerner si le caractère exceptionnel des valeurs s'explique thématiquement, ou serait du à une anomalie de production des données.
Notre proposition repose sur la mise en place d'un cycle itératif d'analyse basé sur l'approche du mantra « *Overview, Zoom and Filter, Details on demand* » de Schneiderman (1996). Il s'agit d'abord de se donner une vue générale de l'ensemble des données, de pouvoir concentrer son attention sur des sous-ensembles, et de filtrer l'information selon certains critères, et enfin de demander des informations supplémentaires sur certaines données ainsi repérées.
Durant la première étape de ce processus, l'utilisateur choisit le jeu de données qu'il souhaite analyser, via une interface qui lui permet d'interroger la base de données d'information statistique. Un premier affichage cartographique avec curseur temporel lui permet d'avoir un aperçu de la distribution des données, de la quantité de valeurs manquantes (Fig.1). Il est encore dans une vue générale (*Overview*).
Ensuite, l'utilisateur peut s'intéresser à un sous-ensemble de valeurs qui sont mises en évidence par l'usage de méthodes de recherche de valeurs exceptionnelles (Tabl.1). Ces méthodes, développées avec R, ont été mises à disposition gracieusement par le *National Centre for Geocomputation* dans le cadre du projet *ESPON 2013 database* (Harris et Charlton, 2010). Dans cette phase de filtrage (*Filter*), il choisit une méthode qu'il paramètre, et demande son exécution. En retour, les unités dont les valeurs sont considérées comme exceptionnelles sont surlignées en rouge dans la carte choroplèthe (Fig.2). Le rapport d'analyse est affiché sous la forme de cartes et de diagrammes.
Enfin, l'utilisateur peut demander plus de détails (*Details on demand*) sur la provenance de valeurs qui semblent exceptionnelles pour une ou plusieurs méthodes : en cliquant sur une unité, les métadonnées correspondant à cette unité, cet indicateur et le jeu de données sont affichées (Fig.3).
Ce processus de filtrage peut être réitéré. Il permet de combiner plusieurs résultats de méthodes dans une même vue (Fig.4), qui met en évidence les valeurs considérées comme exceptionnelles sous différents point de vue (par rapport au voisinage spatial, temporel et thématique).
L'exécution de chaque méthode sélectionnée et paramétrée par l'utilisateur produit un rapport d'analyse, conforme à la norme ISO 19115, qui pourra être exporté, et servir à

enrichir les métadonnées existantes.

Figure 1. Interface d'ESTIM : distribution spatiale du taux d'accroissement du PIB entre 2000 et 2005 en Europe, niveau NUTS3.

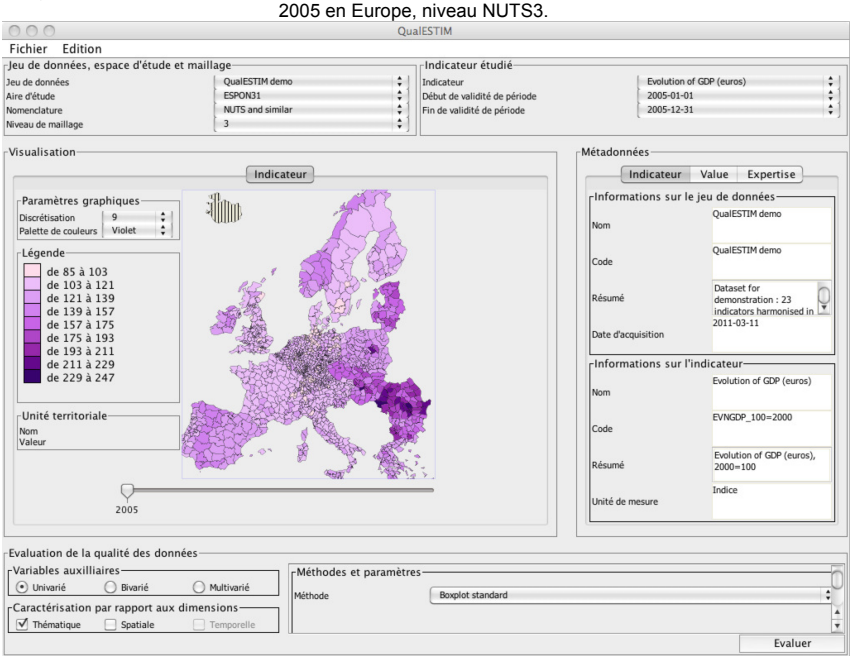


Tableau 1. Liste des méthodes géostatistiques disponibles dans ESTIM.

Méthode	**Dimension**	**Nombre de variables auxilliaires**
Boxplot standard	thématique	0
Boxplot ajusté	thématique	0
Bagplot	thématique	1
Distance de Mahalanobis	thématique	1 ou +
Analyse en composantes principales	thématique	1 ou +
Régression linéaire multiple	thématique	1 ou +
Test de Hawkins	spatiale	0
Moyenne Locale	spatiale	0
Regression locale	spatiale	0 ou +
Régression géographiquement pondérée	spatiale	0 ou +

Figure 2. Carte des valeurs exceptionnelles et rapports d'analyse produits par la méthode « boîte à moustache » sur le taux d'accroissement du PIB entre 2000 et 2005 en Europe, niveau NUTS3.

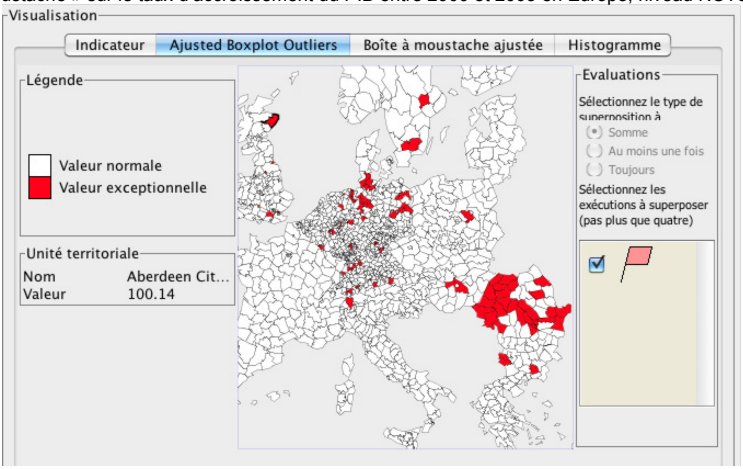


Figure 3. Détails sur la provenance d'une valeur, dans l'onglet « Métadonnées » d'ESTIM.

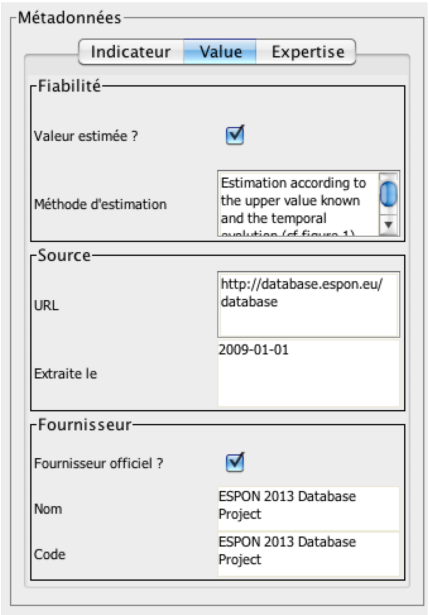
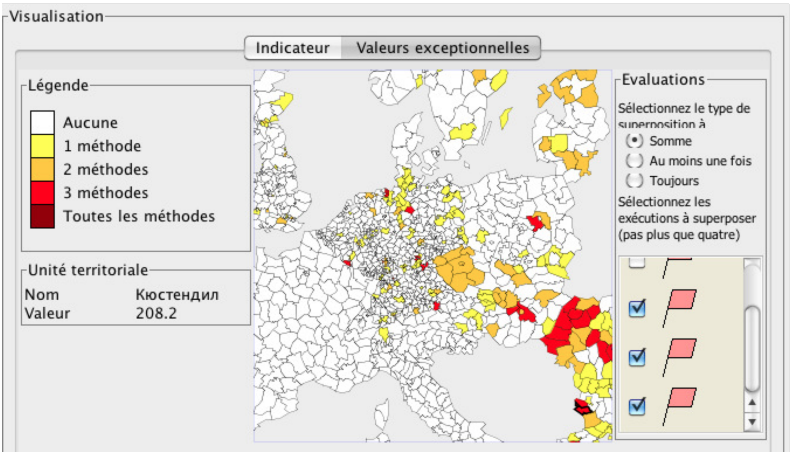


Figure 4. Combinaison des résultats de plusieurs méthodes géostatistiques.



3. LIMITES DE CETTE PROPOSITION

L'usage de ce prototype qui interroge une base de données d'information territoriale complexe, hiérarchique et évolutive, mais selon des modalités assez basiques (par région géographique, par niveau de maillage, par indicateur et plage de validité de l'indicateur dans la base), met en évidence de nouveaux besoins pour l'interrogation spatio-temporelle des données.

Il serait nécessaire d'offrir des méthodes pour combiner des indicateurs collectés sur des versions de maillages différentes afin d'étendre les possibilités d'analyse. Ainsi, les indicateurs mesurés entre 1980 et 1990 en Europe sont généralement associés à la version de maillage de 1980 ou 1988, et ne peuvent pas être aisément combinés avec des données connues dans les versions 2003 ou 2006 de la NUTS. Il serait très intéressant de proposer l'activation « à la demande » de méthodes de transferts de certains indicateurs vers la version du maillage d'étude.

Il s'agit aussi de réfléchir en termes de *voisinages temporels*, et de combinaison de variables ayant une fréquence de mesure et une *inertie temporelle* différentes. Par exemple, les variables démographiques peuvent être utilisées avec une large plage de tolérance (une dizaine d'années par exemple), car les évolutions de ce type de variable sont lentes (20 ans), alors qu'à l'opposé, le prix moyen du baril du pétrole, qui varie d'une semaine à l'autre dans des proportions importantes, devrait faire l'objet d'une restriction à un mois. Il est donc nécessaire de mener une réflexion sur les échelles temporelles, afin de spécifier ce qui est comparable et à quel rythme dans le temps.

CONCLUSION

Cette proposition souligne l'intérêt de l'intégration de métadonnées dans un outil d'analyse exploratoire spatio-temporelle de l'information statistique territoriale ayant pour finalité la détection et l'analyse de valeurs exceptionnelles. Dans notre prototype, nous avons suggéré de combiner les résultats de plusieurs méthodes (géo)statistiques et d'associer leur analyse à celle de métadonnées adaptées de la norme ISO 19115 pour l'information statistique territoriale.

Une de nos perspectives concerne le développement de l'analyse d'évolutions exceptionnelles contextualisées, prenant en compte la structure hiérarchique de l'information territoriale statistique (Plumejeaud, 2011), et permettant de mettre en œuvre des méthodes d'analyses plus originales que celles proposées dans ce prototype. Enfin, nous visons la conception d'un nouveau modèle d'interrogation prenant en compte la diversité des vitesses d'évolution des indicateurs, ainsi que les problèmes liés à l'exploitation de données associées à des versions de maillage non compatibles.

REFERENCES

Andrienko N., Andrienko G., 2005, Exploratory analysis of spatial and temporal data. Springer-Verlag, 715 p.

Anselin, L., 1993, « Exploratory spatial data analysis and geographic information systems ». New tools for spatial analysis, pp. 45–54. Eurostat, Luxembourg.

Harris P. and Charlton M., 2010, Spatial analysis for quality control, phase 1: The identification of logical input errors and statistical outliers, ESPON, Technical Report.

Monmonier M., 1989, Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. Geographical Analysis, vol. 21, pp. 81–84,

Plumejeaud, C., Gensel, J., Villanova-Oliver, M., 2010, « Opérationnalisation d'un profil ISO 19115 pour des métadonnées socio-économiques », INFORSID Marseille, May 25-28 2010.

Plumejeaud C., Mathian H., Gensel J., Grasland C., 2011, Spatio-temporal analysis of territorial changes from a multi-scale perspective, International Journal of Geographical Information Science, 23-08-2011,1-16.

Schneiderman B., 1996, « The Eyes Have It : A Task by Data Type Taxonomy for Information Visualizations » Proceedings of the 1996 IEEE Symposium on Visual Languages, pp. 336-344, Washington, DC, USA.

Tukkey J., 1977, Exploratory data analysis, Addison Wesley Longman Publishing Co., Inc., 688 p.

LES AUTEURS

Christine **Plumejeaud**
COGIT, Institut Géographique National
christine.plumejeaud@ign.fr