

A new regionalization method for spatially dependent functional data based on local variogram models: an application on environmental data

Elvira Romano, Antonio Balzanella, and Rosanna Verde

Abstract Regionalization is to divide a set of spatial objects into a number of spatially contiguous regions while optimizing an objective function, which is usually a homogeneity (or eterogeneity) measure of the derived regions. In this paper we propose a new regionalization method for spatial functional data by means of a Dynamic Clustering Algorithm. The method optimizes a criterion of spatial association among functional data, it is such that the prototype is a trace-semivariogram model chosen from a set of mathematical functions that describes spatial relationships. The appropriate model is chosen by matching the shape of the curve of the experimental variogram to the shape of the curve of the mathematical function. Performances of these methods are checked through an application on real data.

Key words: Regionalization, clustering, spatial functional data, trace-semivariogram function, spatial association

1 Introduction and Problematic

Nowadays many phenomena are monitored by sensor networks, let's think for instance, to security, transport, environment (pollutant and radioactivity monitoring), health (patient care, epidemiological monitoring), etc..

Usually these sensed data are correlated in space. With this in mind, recently, in spatial data analysis, there has been an increasing interest to the development of methods that attempt to exploit such correlation, among them, regionalization methods.

Regionalization methods aim to aggregate basic spatial units into larger units (regions) in order to preserve confidentiality, to minimize population differences, to

Dipartimento di Studi Europei e Mediterranei, Facoltà di Studi Politici, Seconda Università degli Studi di Napoli, Via del Setificio 15, 81100 Caserta
elvira.romano@unina2.it, balzanella2@alice.it, rosanna.verde@unina2.it

reduce the effects of outliers or inaccuracies in the data, or simply, to facilitate the visualization and interpretation of information in maps [10].

Essentially regionalization is a special form of classification where the spatial units are grouped together according to a particular criteria and a set of contiguity or adjacency constraints [2]. The wide range of fields in which these methods can be used makes it difficult to give a single definition.

The main common characteristics to any method are that:

- they aggregate geographical areas into a predefined number of regions while optimizing an aggregation criterion;
- they assume a prior knowledge about the aggregation process about the relevant variables for the aggregation (if are several), the number of cluster to be designed, the spatial contiguity constraint and the existence of an aggregation criterion.

These methods can be distinguished into two groups: methods which apply, at first, conventional clustering techniques and then revise the resulting clusters in terms of spatial contiguity, and methods which force spatial contiguity by maximizing regional compactness [4].

Our proposal merges both classes of methods since it is a method for exploring and discovering spatial correlation structures by means of a clustering algorithm but also it forces the spatial contiguity by maximizing the spatial association. We propose it for spatial functional data, but it is a new method that can be applied at the same way for spatial data.

The main motivation of our approach is that whenever the analyzed region is spatially wide, global measure of correlation are not locally representative. To deal with this issue, we propose to partition the whole spatial region into clusters which are internally homogeneous in terms of their spatial correlation structure.

A measure of spatial association ables to emphasize the average spatial dependence over the area of the study is the variogram.

We propose to use a Dynamic Clustering Algorithm which finds the best partition of the spatial region and the best representation of the clusters minimizing as criterion the fitting between the empirical and the teoretical trace-semivariogram function for each cluster.

According to the variability structure of the regions a suitable trace-semivariogram model (linear, exponential, spherical, Gaussian, Mathern) is chosen.

The estimated trace-semivariogram models are assumed as the prototypes of the clusters of the partition. The allocation of the curves to clusters is performed according to the minimum of the mean squared errors with respect to the estimated trace-semivariogram model of the several clusters.

The paper is organized as follows. Section 2 presents a review of the related works, section 3 includes details on the spatial functional data, section 4 introduces the method we propose, and the algorithm on which our procedure is based on. Section 5 includes an application of the proposed methodology on a real dataset.

2 State of art: related works

Regionalization is one of the most frequently treated problems in the field of constrained clustering. A wide number of approaches to this problem have been taken.

The main aim of these approaches is to search for an optimal solution within a confined or reduced search space. A good review of these methods can be found in [7], it classifies these method in four groups: non-spatial clustering followed by spatial processing; non-spatial clustering with a spatially weighted dissimilarity measure; trial-and-error search and optimization; and finally spatially constrained clustering and partitioning.

The first methods uses clustering algorithm to obtain a partition of the data based on similarity. The obtained clusters are then divided or merged to form regions in the geographic space. The second type of regionalization methods uses clustering algorithms with a similarity (or dissimilarity) measure to incorporate spatial information explicitly, for example, using a distance-weighted attribute similarity. The third group of methods, starts with an initial random regionalization and iteratively refines the solution by reassigning objects to neighboring regions for better solutions. The fourth group of methods explores a reduced search space, which is confined by the clustering method employed.

In the spatial functional framework, to our knowledge, the existent methods properly dealing with regionalization problems do not exist. The only two existent approaches can be considered constrained clustering methods [6], [9].

The first is a hierarchical clustering methods for both geographically referenced data and functional data developed to give a solution to the problem of classifying spatially correlated curves. The methodology allows to find groups of curves which are spatially homogeneous, by weighting the dissimilarities between curves through the trace-semivariogram [5] and through the multivariate variogram of the coefficients of basis functions used to smooth the observed data.

The second is our earlier proposal in this framework. It is a clustering algorithm based on an optimization problem that minimizes the spatial variability among the curves in each cluster and a prototype that is a spatial functional linear model in a point of the space. The main feature of this approach is that the prototypes obtained by an ordinary kriging prediction have a representative spatial location in the clusters.

Following the distinction in the spatial framework, we have given before, they can be located in the second group of methods.

Inspired by the third groups of methods, this research proposes, and implements a regionalization method based on a Dynamic Clustering Algorithm. The optimized criteria is such to incorporate the spatial information in the data by using the fitting of a predefined theoretic variogram model to an empirical variogram.

In the following we explain in which way. Before going on, we recall some main notions for spatial functional data.

3 Centrality Notions for Spatial Functional data

Spatially dependent functional data may be defined as the data for which the measurements on each observation, that is a curve, are part of a single underlying continuous Spatial functional process defined as:

$$\Xi = \left\{ \chi_s : s \in D \subseteq \mathbb{R}^d \right\} \quad (1)$$

Where s is a generic data location in the d -dimensional Euclidean space (d is usually equal to 2), the set $D \subseteq \mathbb{R}^d$ can be fixed or random and χ_s are functional random variables, defined as random elements taking values in an infinite dimensional space. The nature of the set D allows to classify the kind of Spatial Functional Data.

Following Giraldo et. al [3] these can be distinguished in geostatistical functional data, functional marked point pattern and functional areal data.

Geostatistical functional data appear when D is a fixed subset of $\mathbb{R}^d$ with positive volume and n points $s_1, \dots, s_n$ in D are chosen to observe a random functions $\chi_{s_i}, i = 1, \dots, n$.

Functional marked point pattern appear when a complete function is observed at each point generated by a standard point process.

Functional areal data (or functional data in lattice) correspond to the case of D being a fixed and countable set. The aims in spatial functional data analysis are similar to those concerning univariate or multivariate spatial analysis: detection of the spatial dependence, identification of spatial clusters, and modeling the spatial dependence. This corresponds to the case of D being a fixed and countable set.

We focus on geostatistical functional data, when samples of functions are observed in different sites of a region (spatially correlated functional data).

Let us consider a sample of curves $\chi_{s_1}(t), \dots, \chi_{s_i}(t), \dots, \chi_{s_n}(t)$ for $t \in T$ where s_i is a generic data location in the d -dimensional Euclidean space.

We assume for each t that we have a second order stationary and isotropic random process, that is, the mean and variance functions are constant and the covariance depends only on the distance between sampling sites.

Formally, we assume that:

- $E(\chi_s(t)) = m(t)$, for all $t \in T, s \in D$.
- $V(\chi_s(t)) = \sigma^2$, for all $t \in T, s \in D$.
- $Cov(\chi_{s_i}(t), \chi_{s_j}(t)) = C(h, t)$
- $\frac{1}{2}V(\chi_{s_i}(t) - \chi_{s_j}(t)) = \gamma(h, t) = \gamma_{s_i s_j}(t)$ where $h_{ij} = \|s_i - s_j\|$ and all $s_i, s_j \in D$.

The function $\gamma(h, t)$ as function of h is called semivariogram of $\chi(t)$.

Since we are assuming that the mean function is constant over D the function:

$$\gamma(h, t) = \gamma_{s_i s_j}(t) = \frac{1}{2}V(\chi_{s_i}(t) - \chi_{s_j}(t)) = \frac{1}{2}E[\chi_{s_i}(t) - \chi_{s_j}(t)]^2 \quad (2)$$

as function of h is called semivariogram and the function

$$\gamma(h) = \int \gamma_{s_i s_j}(t) dt \quad h = \|s_i - s_j\| \quad (3)$$

is called trace-semivariogram. This function can be estimated by the classical methods of the moments by means of [3]:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt \quad (4)$$

where $N(h) = \{(s_i; s_j) : \|s_i - s_j\| = h\}$ for regular spaced data and $|N(h)|$ is the number of distinct elements in $N(h)$. When data are irregularly spaced the $N(h)$ become $N(h) = \{(s_i; s_j) : \|s_i - s_j\| \in (h - \varepsilon, h + \varepsilon)\}$ with $\varepsilon \geq 0$ being a small value.

In computing the empirical semivariogram, we consider that the functions are expanded in terms of some basis functions by

$$\chi_{s_i}(t) = \sum_{l=1}^Z a_{il} B_l(t) = \mathbf{a}_i^T \mathbf{B}(t), \quad i = 1, \dots, n \quad (5)$$

The coefficients of the curves can be consequently organized in a matrix as follows:

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,Z} \\ a_{2,1} & a_{2,2} & \dots & a_{2,Z} \\ \vdots & \ddots & \dots & \dots \\ a_{n,1} & a_{n,2} & \dots & a_{n,Z} \end{pmatrix}_{n \times Z}$$

then the empirical semivariogram function can be expressed by

$$\gamma_{l,q}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \left[(\mathbf{a}_i - \mathbf{a}_j)^T \mathbf{W} (\mathbf{a}_i - \mathbf{a}_j) \right] \quad \forall i, j \mid \|s_i - s_j\| = h$$

where $\gamma_{l,q}(h)$ is the generic element of a trace-semivariogram matrix with $l, q = 1, \dots, Z$ and, $\mathbf{a}_i, \mathbf{a}_j$ are vectors of the basis coefficients for the χ_{s_i} and χ_{s_j} .

Note that $\mathbf{W} = \int_T \mathbf{B}(t) \mathbf{B}(t)^T dt$ is the Gram matrix that is the identity matrix for any orthonormal basis while for other basis as B-Spline basis function, $\mathbf{W}$ is computed by numerical integration.

The empirical trace-semivariogram cannot be computed at every lag distance h and due to variation in the estimation it is not ensured that it is a valid variogram.

Such as in applied geostatistics the empirical trace-semivariograms are thus approximated (by ordinary least squares (OLS) or weighted least squares (WLS)) by model function ensuring validity [1]. Some widely used models are [2]: Spherical, Gaussian, exponential or Mathern.

The trace-semivariogram model is the key function of our approach as we will see in the following section.

4 A Dynamic Clustering Algorithm for spatially dependent functional data

Dynamic clustering algorithm (DCA) or *Nueés Dynamiques* is an unsupervised batch training algorithm. Like in the classical clustering techniques the aim is to find groups that are internally dense and sparsely connected with the others.

Let E be a set of n objects, the Dynamic Clustering Algorithm finds a partition $P^* = (C_1, \dots, C_k, \dots, C_K)$ of E in K non empty clusters and a set of representative prototypes $L^* = (G_1, \dots, G_k, \dots, G_K)$ for each C_k cluster of P so that both P^* and L^* optimize a criterion Δ :

$$\Delta(P^*, L^*) = \text{Min} \{ \Delta(P, L) / P \in P_k, L \in \Lambda^k \} \quad (6)$$

with P_k the set of all the K -clusters partitions of E and Λ^k the representation space of the prototypes. $\Delta(P, L)$ is a function which measures how well the prototype G_k represents the characteristics of objects of the cluster and it can be usually interpreted as an heterogeneity or a dissimilarity measure of goodness of fit between G_k and C_k .

Our proposal is to partition a sample of curves $\{\chi_{s_1}(t), \dots, \chi_{s_i}(t), \dots, \chi_{s_n}(t)\}$ for $t \in T$ into a set of K clusters such that the obtained clusters can be described by a trace-semivariogram model.

In our context we assume as criterion function:

$$\Delta(P, L) = \sum_{k=1}^K \sum_{h \in N_k(h)} (\gamma_k(h, t) - \gamma_k^*(h, t))^2 \quad (7)$$

where $N_k(h) \subseteq N(h)$ is the set of ordered $h_{ij} = \|s_i - s_j\|$ for all $\chi_{s_j}(t), \chi_{s_i}(t) \in C_k$.

This criterion is based on the best fit function $\phi(\gamma_k(h, t), \gamma_k^*(h, t))$ between the empirical variogram $\gamma_k(h, t)$ and the theoretical variogram $\gamma_k^*(h, t)$, for each cluster C_k , according to a chosen model $\gamma^*(h, t)$ of variogram.

In our proposal we assume the estimated variogram functions $\gamma_k^*(h, t)$ (for $k = 1, \dots, K$) as prototypes of the clusters.

Starting from a random initialization partition P of the of n functions in K clusters, this algorithm alternatively performs a *representation* and an *allocation* step.

The *representation* step estimates the theoretical variogram $\gamma_k^*(h, t)$, for each cluster C_k , by Ordinary Least Square method.

The *allocation* step is performed computing for each $\chi_{s_m}(t)$ with respect to each cluster C_k the values $\delta_{m,k}(h, t) = V(\chi_{s_m}(t) - \chi_{s_i}(t))$

where:

$$\chi_{s_i}(t) \in C_k \text{ and } \|s_m - s_i\| = h_{m,i} \in N_{m,k}(h).$$

Then, $\chi_{s_m}(t)$ is assigned to a cluster C_k according to the minimum mean squared distances between $\delta_{m,k}(h, t)$ and $\gamma_k^*(h, t)$ (for $k = 1, \dots, K$):

$$\sum_{h \in N_{m,k}(h)} (\delta_{m,k}(h,t) - \gamma_k^*(h,t))^2 \leq \sum_{h \in N_{m,k'}(h)} (\delta_{m,k'}(h,t) - \gamma_{k'}^*(h,t))^2 \quad k' \neq k. \quad (8)$$

It is worth of notice that the consistence between the optimized criterion and the allocation function is ensured by:

$$\gamma_k(h,t) = \frac{1}{\sum_{m=1}^{|C_k|} |N_m(h)|} \sum_m^{|C_k|} \delta_{m,k}(h,t) * |N_{m,k}(h)| \quad (9)$$

In the following we introduce the algorithm scheme.

Algorithm 1 Dynamic Clustering Algorithm for spatio-functional data

Initialization:

Start from a random partition $P = (C_1, \dots, C_k, \dots, C_K)$

Representation step:

for $C_k \in P$ **do**

 Compute the prototypes $\gamma_k^*(h,t)$ which optimize the best fit criterion:

$$\min \sum_{h \in N_k(h)} (\gamma_k(h,t) - \gamma_k^*(h,t))^2 \quad (10)$$

end for

allocation step:

for all $\chi_{s_i}(t) \in E$ **do**

 find the cluster index k^* such that:

$$k^* = \arg \min_{k=1, \dots, K} \sum_{h \in N_m(h)} (\delta_{N_{m,k}(h,t)} - \gamma_k^*(h,t))^2$$

end for

5 Application

In order to evaluate the effectiveness of the proposed method we have performed several experimental tests on real data.

The studied dataset comes from the European Climate Assessment Dataset (ECAD) project, the data are available at <http://eca.knmi.nl/>. It provides climate monitoring series from a wide set of meteorological stations throughout Europe and Mediterranean.

Especially we focus on maximum temperature data recorded from 1, January, 2000 to 28, February, 2010.

Even though the whole dataset is made by 1110 series, we performed our tests on 730 monitoring stations. This is because the observations of 380 series are not available on the whole time period. Each time series records a new observations each day, whenever single measurements are not available, these are obtained by infilling from nearby stations.

Due to the extension of the spatial region involved in the monitoring activity, several spatial variability structures are present in the data.

Thus, we evaluate how the proposed strategy is able to discover such variability structures and their associated spatial regions. With this aim, we analyze, for each cluster, the fitting between the empirical trace-semivariogram and the representative theoretical trace-semivariogram according the optimized criterion in (8).

In order to provide the results, the following choices have to be performed:

- The basis functions to use for series representation
- The parameters to use in the computation of the empirical trace-semivariogram
- The number of clusters K
- The trace-semivariogram model function $\gamma_k^*(h, t)$

The first step is to construct the set of functions (in 5) expanded in terms of B -Spline Basis functions. We need to choose an appropriate order of expansion Z , taking into account that a large Z causes overfitting and a K too small we may miss important aspects of the function that we are estimating [8]. We consider a procedure based on a classical non-parametric cross-validation analysis. Especially we evaluate cubic splines for each series, and get a collection of smooth curves that is able to take into account the variability of the data.

Whenever the data are detected on irregularly spaced locations, the computation of the empirical trace-semivariogram may be performed considering only one pair of locations that is h apart. Averages based on only few points are poor estimates with large uncertainties. It is usual to increase the accuracy of the point estimates by defining tolerance regions and group the sample pairs into these regions prior to averaging. Each region has to be small enough so that we retain enough spatial resolution to define the structure of the semivariogram, but also large enough so that we base each point estimate on a relative large number of paired differences.

On the analyzed dataset we impose that the empirical trace-semivariogram has to be computed using 10 regions. Such value has allowed to get a reasonable number of curves in the averaging process. The choice of the number of clusters is based

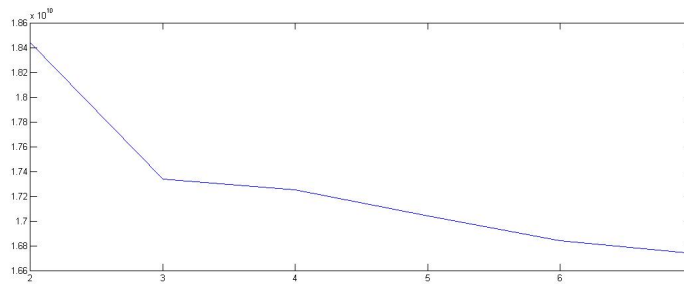


Fig. 1 Value of the optimized criterion for $K = 2, \dots, 7$

on tests which evaluate the optimized criterion for several values of K . As shown in

Fig.1, the optimal number of clusters is $K = 3$ since it provides the highest decrease of $\Delta(P, L)$.

A further preliminary decision concerns the trace-semivariogram model function to be fitted in the clustering procedure. Since we have no a-priori information on what is the best fitting trace-semivariogram model function, we tested the value of the criterion for the Exponential model, the Spherical model and the Gaussian model. The results, which are available in table 1, highlight that the Exponential model is the most suitable in our context.

Trace-semivariogram model	$\Delta(P, L)$
Exponential	$1.73e^{+10}$
Gaussian	$1.75e^{+10}$
Spherical	$1.76e^{+10}$

Table 1 Criterion evaluation for several Trace-semivariogram model functions

Starting from the chosen input parameters, the algorithm run on the dataset, detects the spatial regions available in Fig. 2. The value of the optimized criterion is $\Delta(P, L) = 1.73e^{+10}$, the number of iterations until convergence has been 7.

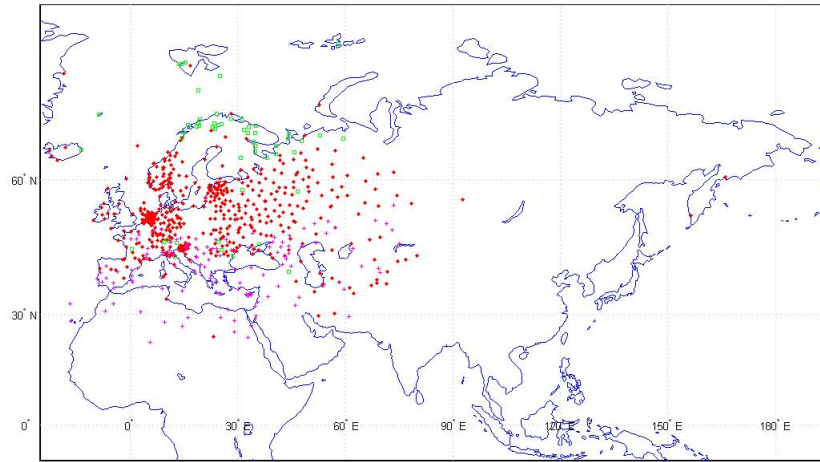


Fig. 2 Discovered clusters plotted on the map.

It is possible to note that the three discovered variability structures split the studied area into three spatial regions which include:

- the Northern Europe
- the most of the continental Europe and the Great Britain
- the Northern Africa, the Spain, the Balkans

The discovered trace-semivariogram model functions which are representative of each clusters are shown in Fig.3

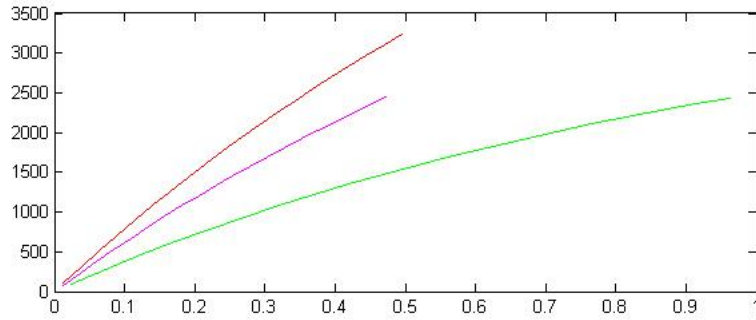


Fig. 3 Theoretical trace-semivariogram models for each cluster.

Looking at the plots, it is possible to highlight that the variability in the first cluster rises at an higher rate, when it is compared to the two other clusters. This is indicative of an high spatial dependence. At the opposite, the third cluster, shows a reduced spatial dependence due to the low rate of growth of the spatial variability. It is still possible to note that the third cluster includes locations at an higher spatial distance when compared with the two other clusters.

6 Conclusion

In this paper we have introduced a new regionalization method.

We have focused on spatially related curves using the main notions of spatial functional data analysis. The method can be applied to a wide range of domains where temporally data are observed. The method is able to discover both the spatial partition of the data and the spatial variability structures representative of each cluster. The main novelty is the use of the trace-semivariogram model function as prototype of each cluster along the iterations of the clustering process. We have analyzed an environmental dataset to asses the performance of the method.

We have focused on OLS techniques as classical estimation method, however further attention will be given to alternative methods parameter and nonparametric, like weighted ordinary least squares, generalized least squares and maximum likelihood estimation [2].

References

1. Chiles, J.P., P. Delfiner: *Geostatistics, Modelling Spatial Uncertainty*, Wiley-Interscience.(1999)
2. Cressie, N.: *Statistics for spatial data*, Wiley Interscience. (1993)
3. Delicado, P., Giraldo, R., Comas, C., Mateu, J.: *Statistics for Spatial Functional data: Some recent contribution*. *Environmetrics*. Forthcoming. Technical Report, <http://hdl.handle.net/2117/2446>, Universitat Politecnica de Catalunya (2009).
4. Duque, J. C., Ramos, R., and Surinach, J.: *Supervised regionalization methods: a survey*. *International Regional Science Review*, 29. Forthcoming.(2006)
5. Giraldo, R., P. Delicado, and J. Mateu: *Geostatistics for functional data: An ordinary kriging approach*. Technical Report <http://hdl.handle.net/2117/1099>, Universitat Politecnica de Catalunya. Submitted to *Environmental and Statistics*.(2007)
6. Giraldo, R., Delicado, P., Comas, C., Mateu, J.: *Hierarchical clustering of spatially correlated functional data*. Technical Report, <http://www.ciencias.unal.edu.co/estadistica/reporte02.pdf>. (2009).
7. Guo, D.: *Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP)*. *International Journal of Geographical Information Science* **22**, No. 7, July 2008, 801-823
8. Ramsay, J.E., Silverman, B.W.. *Functional Data Analysis* (Second ed.). Springer, 2005.
9. Romano E., Balzanella A., Verde R. *Clustering Spatio-functional data: a model based approach*. *Studies in Classification, Data Analysis, and Knowledge Organization* Springer, Berlin-Heidelberg, New York,(2010).
10. Wise, S.M., Haining, R.P. and MA, J.: *Regionalization tools for the exploratory spatial analysis of health data*. In *Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling and Neuro-Computing*, M. Fischer and A. Getis (Eds), pp. 83-100 (Berlin: Springer)(1997).