d' Aubigny, Gérard

Article

A statistical toolbox for mining and modeling spatial data

Comparative Economic Research. Central and Eastern Europe

Provided in Cooperation with:
Institute of Economics, University of Łódź

Suggested Citation: d' Aubigny, Gérard (2016) : A statistical toolbox for mining and modeling spatial data, Comparative Economic Research. Central and Eastern Europe, ISSN 2082-6737, De Gruyter, Warsaw, Vol. 19, Iss. 5, pp. 5-24, https://doi.org/10.1515/cer-2016-0035

This Version is available at:
https://hdl.handle.net/10419/184412

Comparative Economic Research, Volume 19, Number 5, 2016
10.1515/cer-2016-0035

GÉRARD D'AUBIGNY*

A Statistical Toolbox For Mining And Modeling Spatial Data

Abstract

Most data mining projects in spatial economics start with an evaluation of a set of attribute variables on a sample of spatial entities, looking for the existence and strength of spatial autocorrelation, based on the Moran's and the Geary's coefficients, the adequacy of which is rarely challenged, despite the fact that when reporting on their properties, many users seem likely to make mistakes and to foster confusion. My paper begins by a critical appraisal of the classical definition and rational of these indices. I argue that while intuitively founded, they are plagued by an inconsistency in their conception. Then, I propose a principled small change leading to corrected spatial autocorrelation coefficients, which strongly simplifies their relationship, and opens the way to an augmented toolbox of statistical methods of dimension reduction and data visualization, also useful for modeling purposes. A second section presents a formal framework, adapted from recent work in statistical learning, which gives theoretical support to our definition of corrected spatial autocorrelation coefficients. More specifically, the multivariate data mining methods presented here, are easily implementable on the existing (free) software, yield methods useful to exploit the proposed corrections in spatial data analysis practice, and, from a mathematical point of view, whose asymptotic behavior, already studied in a series of papers by Belkin & Niyogi, suggests that they own qualities of robustness and a limited sensitivity to the Modifiable Areal Unit Problem (MAUP), valuable in exploratory spatial data analysis.

Keywords: *duality diagram, spatial autocorrelation, Moran'index, Moran's Eigenvector Maps, Laplace operator, spatial eigenfunction filtering*

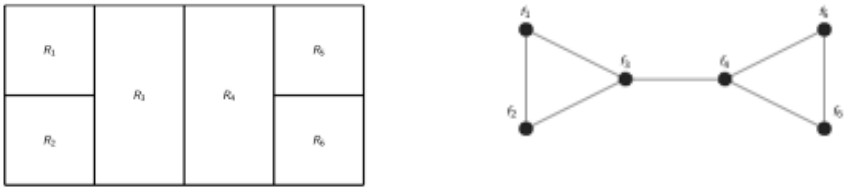
* Professor at the University of Grenoble-Alpes, France; Jean-Kuntzmann Laboratory, e-mail: Gerard.d-Aubigny@univ-grenoble-alpes.fr

1. Introduction

The results presented in this paper were established at the occasion of an exploratory data analysis project aimed to identify potential links existing between the partisan choices of French voters and a number of socio-demographic and economic attributes describing their living environment, through geo-referenced measurements: first, we had the scores achieved by each candidate to each of a series of polls, during the period 1980–2012, including French district and presidential elections and elections to the French and the European Parliaments, and measured at the commune level (source: the French Ministry of Interior Affairs); second, we had extractions of the 2010 census (source INSEE) at the commune level; and last, we had limited tax information, disaggregated at two scales: irregularly located rectangles with surface 1 km^2 and a regular grid of squares of size 200 *m* × 200 *m* (source INSEE), see d'Aubigny (2012) for more details.

This clue of problems can be approached by adopting a Graph Data Mining formalism which expresses spatial dependence information through a (weighted) binary relation, and uses (weighted) graphs to describe the inter-areas relationships. Let us illustrate the approach on a toy example.

Figure 1. Toy example of a domain partitionned in six areas. The left figure shows the geographic shape of this domain, while the right one represents its topological structure in the form of a graph of neighboorhood (or contiguity)



Source: Own calculation

Let $\mathbf{D} = \bigcup_{s=1}^{n} R_s$ denote a domain in the plane, which is partitioned in n disjoined areas $\boldsymbol{R_i}$ ($R_s \cap R_t = \emptyset$ if $s \neq t$) as illustrated by the left panel of Figure 1. We drew in the right panel of this Figure, its translation in the language of graphs: here, we get a graph ***G*=(*V*, *E*)**, with n=6 nodes s and m=7 edges (s, t). The topology of a graph ***G*=(*V*, *E*)** is classically described in algebraic terms by its adjacency (*aka* contiguity) matrix,

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

$\boldsymbol{A}$ is a symmetric $n \times n$ matrix, with general term a_{st}=1 if nodes s and t are linked by an edge, and 0 otherwise. Two other symmetric $n \times n$ matrices, are deduced from the adjacency matrix $\boldsymbol{A}$, namely

$$\mathbf{D}_{\mathbf{a}+} = \begin{pmatrix} 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix} \quad \boldsymbol{L_A} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & -1 & 0 & 2 \end{pmatrix}$$

Here, the diagonal matrix $\mathbf{D}_{\mathbf{a}+} = \boldsymbol{diag}\,(\boldsymbol{a}_{s+})$ has for general diagonal term $\boldsymbol{d}_{ss} = \sum_{t=1}^{n} \boldsymbol{a}_{st} = \boldsymbol{a}_{s+}$ where $\boldsymbol{a}_{s+}$ is the *outer degree* weight of the node s of the graph $\boldsymbol{G}$. As a consequence, $\boldsymbol{a}_{++} = \sum_{t=1}^{n} \boldsymbol{a}_{s+} = \sum_{s=1}^{n} \sum_{t=1}^{n} \boldsymbol{a}_{st}$ counts the total number of observed relations of neighbourhood (counted twice since A is symmetric) and is called the *volume* of the graph G. The matrix $\mathbf{L_A}$ is known as the *combinatorial Laplacian* (*aka Graph Laplacian*) matrix associated to the graph $\boldsymbol{G}$. By definition, $\mathbf{L_A} = \mathbf{D}_{\mathbf{a}+} - \boldsymbol{A}$ see *e.g.* Bollobas (1990) or Bapat (2010) for more details.

In a way to simplify the formulas used in the remainder of this paper, and without loss of generality, we have considered the matrix $\boldsymbol{Q}$, obtained by the normalization $q_{st} = \dfrac{a_{st}}{a_{++}} \in [0,1]$. Since the generic term $q_{s+} = \dfrac{a_{s+}}{a_{++}}$ of the $\mathbf{D}_{q+} = \boldsymbol{diag}\,(\boldsymbol{q}_{s+})$ is also smaller than 1 and satisfies $\sum_{s=1}^{n} \boldsymbol{q}_{s+} = 1$, $\mathbf{D}_{q+}$ gets interpretable as a probability measure on the nodes of $\boldsymbol{G}$. Notice too that $\mathbf{L}_0 = \mathbf{D}_{\mathbf{q}+} - \mathbf{Q} = \mathbf{D}_{\mathbf{q}+}(\mathbf{I} - \mathbf{D}_{\mathbf{q}+}^{-1}\mathbf{Q}) = \mathbf{D}_{\mathbf{q}+}(\mathbf{I} - \mathbf{W})$, $\mathbf{W} \overset{def}{=} \mathbf{D}_{\mathbf{q}+}^{-1}\mathbf{Q}$, so that the generic term $w_t^s = q_{st} / q_{s+}$ of $\mathbf{W} = (\boldsymbol{w}_t^s)_{st}$ is a positive quantity, such that $\boldsymbol{w}_+^s = \sum_{t=1}^{n} \boldsymbol{w}_t^s = 1$ for all s in {1:n}. So, each row s of $\mathbf{W}$ is interpretable as a conditional probability measure on the set of nodes {1: n}, and is invariant by any renormalisation of the $\boldsymbol{A}$ matrix.

Definition (d'Aubigny 2006): *Let* $G=(V,E)$ *denote the contiguity graph associated to a partitioning of a spatial domain in n sub-areas s, and let X denote an attribute measured on these areas. Then, we call*

a) local mean of X in the neighbourhood of the area s, the quantity

$$\widetilde{X}_s \overset{def}{=} \sum_{t=1,\, t\neq s}^{n} w_t^s X_t ;$$

b) (local) image of X *the variate* $\widetilde{X}$ *taking the value* $\widetilde{X}_s$ *at site s.*

c) (local) anti-image of X the variate $\breve{X} \overset{def}{=} X - \widetilde{X}$.

So, the local image of X on the graph G: $\widetilde{X} = WX$, associates to each area s the mean value of X on its neighborhood.

Now, for any function f defined on D, and any embedding of the representation graph G in a n-dimensional real vector space F, let us denote by x_i the representative of R_i in F and by $f: R_i \rightarrow R_i \in \mathbb{R}$ a functional which affects the value f_i to the node i of graph G. One way to control the smoothness of a function f consists in making as small as possible the squared differences $(f_s - f_t)^2$ between adjacent nodes $s \sim t$ and globally, to minimize their sum over the m existing edges of the graph. But elementary algebra shows that:

$$S(f) = \sum_{s\sim t} (f_s - f_t)^2 = 1/2^t \, f L_A f$$

So, $S(f)$ is an indicator of roughness of the functional f defined on the Graph G, which measures the variability of values of f on neighbor points, *id est* an index of local variation.

1.1. Classical Moran and Geary spatial dependence coefficients

The autocorrelation coefficient, proposed originally by Moran (1948) in the case of a sample of observations of a random variable X in n areas, spatially structured with a topology which is described by an adjacency matrice A writes in our notations:

$$I_M = \frac{\frac{1}{2\, a_{++}} \sum_{st} a_{st} \dot{X}_s \dot{X}_t}{\frac{1}{n} \sum_s (\dot{X}_s)^2} = \frac{1}{2} \sum_{st} q_{st} Z_s Z_t$$

where $\boldsymbol{Z}$ is a scaled version of X with s-th coordinate $\boldsymbol{Z_s = \dot{X}_s / S(X)}$, using the classical standardization calculated with help of the empirical moments

$$\bar{X} = \frac{1}{n}\sum_s X_s, \qquad S^2(X) = \frac{1}{n}\sum_s (\dot{X}_s)^2, \qquad \dot{X}_s = X_s - \bar{X}$$

From a geometric point of view, the n observations may be interpreted as forming the coordinates of a vector $\boldsymbol{x}$ of the Euclidean space $F = (\Re^n, N)$ with metric $N = \frac{1}{n} I_n$, $\bar{\boldsymbol{x}}$ is interpreted as the length of the orthogonal projection of $\boldsymbol{x}$ on the support line of the vector of constants ${}^t\iota=(1,1, \ldots, 1)$, and the sample variance of X may be written:

$$S^2(X) = \frac{1}{n}\sum_s (\dot{X}_s)^2 = {}^t\boldsymbol{x}\ {}^t\boldsymbol{H_1}\boldsymbol{H_1}\boldsymbol{x} = {}^t\boldsymbol{x}\,\boldsymbol{H_1}\boldsymbol{x}$$

with $H_1 = \left(I_n - \frac{1}{n}\iota^t\iota\right) : x \mapsto H_1 x = x - \bar{x}\iota$ the centered vector induced by the metric $\boldsymbol{N}$. These formulas show that both the mean and the variance of the variate X are evaluated by Moran under two implicit postulates about the process or the sampling design which generated the data: one of spatial independence, and one of equiprobability (or equal weighting of nodes). The same assumptions are in action in the a challenger index proposed by Geary (1954), and called the *contiguity ratio*, which is defined as

$$C = \frac{\frac{(n-1)}{4n\, a_{++}}\sum_{st} a_{st}(X_s - X_t)^2}{\frac{1}{n}\sum_s (\dot{X}_s)^2} = \frac{(n-1)}{2n}\times\frac{1}{2}\sum_{st} q_{st}(Z_s - Z_t)^2$$

Cliff and Ord (1981) extended both indices to the case where the weights q_{st} are more general than mere descriptors of the topology of the graph $\boldsymbol{G}$, and is supposed to measure the relative strength of any notion of proximity of areas s and t. Elementary algebra shows that I_M writes with the preceding notations:

$$I_M = \frac{\sum_{s=1}^n q_{s+}\dot{X}_s\tilde{\dot{X}}_s}{\frac{1}{n}\sum_{s=1}^n (\tilde{\dot{X}}_s)^2}$$

This expression reveals the nature of difficulties met by this index in practice as well as in theory. First, except when $\boldsymbol{G}$ is a complete and regular graph, and contrary to what is often said, I_M is not a linear correlation coefficient – it is not a cosine, and it does not vary in [-1,1] -. Second, it does not take the form of the ordinary least squares estimator of a slope parameter in linear regression. These denials result from the fact that the two systems of weights q_{s+}

and $1/n$, respectively used in the numerator and the denominator of $I_{M,}$ are usually different. In fact, both I_M and C suffer the same sources of troubles as can be made explicit by establishing the following monotone decreasing relation that exists between Moran's I and Geary's C indices:

$$\boldsymbol{C} = \frac{n-1}{n}\left\{\frac{\sum_{s=1}^{n} q_{s+}(\tilde{\boldsymbol{X}}_s)^2}{\frac{1}{n}\sum_{s=1}^{n}(\tilde{\boldsymbol{X}}_s)^2} - I_M\right\}$$

In fact, both indices I_M and C suffer one and the same incoherence in the way the two first moments of the repartition of $\mathbf{X}$ are calculated in their numerator and their denominator. While less explicit in the formula, this inconsistency affects also all their distributional properties (used in statistical inference) as well as their geometric ones used in Exploratory Data Analysis (*aka* EDA) studies.

The sequel of this paper is devoted to the proposal of fixes of the alleged incoherence in the definition of the Moran's and Geary's indices. This leads us first to highlight the special role played by the combinatorial Laplacian in the analysis of geo-referenced data. The next subsection presents our corrected versions of both coefficients: we show their strict statistical equivalence, and how the Laplacian plays a key tool in the elaboration of dimension reduction and visualization of spatial data. We report on the following subsection on the usage of methods developed by specialists of machine learning for the problem of approximation of points clouds by Riemannian manifolds, on the existence of infill convergence of the combinatorial Laplacian on a discrete set of n points to the Laplace Beltrami operator associated with the underlying Riemannian manifold, when n tends to infinity. We close the paper by a short discussion.

2. Correction of the Moran's and Geary's coefficients

The proposed correction applies the same geometric principle underlying the analysis of the algebraic duality that binds the nodes and either edges of a non-oriented graph or arcs in the case of oriented graphs. In both cases, we need to choose arbitrarily an orientation of edges, but this choice is mandatory only for technical reasons, and has no consequence on the results.

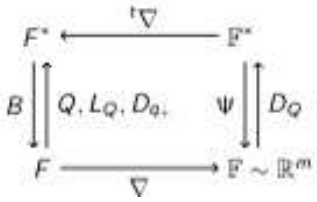
Let us consider a network or weighted graph $N=(G, Q)$ where $G=(V, E)$ is supposed a simple graph (at most one edge can link two nodes), and contains no loop (no edge from one node to itself). The valuation q is defined on the set of existing edges $E \subseteq V \times V$, and supposed real, positive and symmetric $q_{st} = q_{ts}$ Moreover, in

all this text ***Q*** is supposed normalized. As is usual in graph theory, see Bapat (2010), we associate to the graph *G* its incidence matrix ∇. This is a rectangular matrix of order $m \times n$ if $card(V) = n$ and $card(E) = m$, whose generic term is $\nabla_s^e = +1 \; if \; e = (s,t) \; and \; -1 \; if \; e = (t,s)$. As a consequence, m denotes the number of edges defining the neighborhood links existing in the graph G and so, $m \leqq \widetilde{m} \stackrel{def}{=} n(n-1)/2$ where $\widetilde{m}$ is the number of edges of a complete graph having *n* nodes.

Now, as classical in the french school of Multivariate Data Analysis we interpret any data matrix, thus the incidence matrix ∇ here, as inducing an algebraic duality between two representation spaces: one real vector space $\mathbb{F}$ of dimension m for columns of ∇ (the n nodes here) and containing real functions operating on E(G); and one real vector space ***F*** of dimension n for rows of ∇ (the m edges here) and containing real functions operating on V(G), see *e.g.* Cailliez & Pages (1975), Escoufier (1987), d'Aubigny (1989). For any ***x*** in ***F***:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_s \\ \vdots \\ x_n \end{pmatrix} \hookrightarrow \nabla x = \begin{pmatrix} x_1 - x_2 \\ x_1 - x_3 \\ \vdots \\ x_1 - x_n \\ \vdots \\ x_s - x_t \\ \vdots \\ x_{n-1} - x_n \end{pmatrix} \in \mathbb{F}$$

∇ is called the *difference matrix* in mathematics, the *incidence matrix* in Graph theory, and the *simple contrasts matrix* in statistics.



But, any visualization of points in a vector space requires the definition (implicit or explicit) of a geometry in a way to allow one to measure proximities between elements of this space, and the simplest one is the Euclidean one, defined by some symmetric matrix which is positive and semi-definite. So, the *duality diagram* above illustrates this setting where the definition of a geometrical model useful to represent the structure of *G, as* expressed in ∇, necessitates to specify two *Euclidean metrics* (distances): first, one on *F**, in a way to measure proximities

between nodes – denoted B in the duality diagram given above – and a second one on $\mathbf{F}$ in a way to measure proximities between elements of this space, that is edges here. Most often, specialists of spatial econometrics ignore B as a modeling opportunity and so, implicitly take it equal to the identity ($B=I_n$). When a weight matrix Q - symmetric and of order $n\times n$ – is given, one natural choice of metric on $\mathbf{F}$ is the diagonal matrix of order $m\times m$, denoted by D_O on the diagram above. This setting may be summarized by writing the so-called *statistical triple* (∇, B, D_O), from which is all the elements of the duality diagram are deducible: ∇ is fixed by the data, while B and D_O are specified by the analyst. Their specification relates to modeling activities.

2.1. Connection between the Laplacian and Spatial Regression Models

Note first that under the hypothesis of independence of the sample of areas, one has $Q = p\otimes p$, for any system p of weights of these areas, and then,

$$q_{s+} = \sum_{t=1,t\neq s}^{n} p_s p_t = p_s(1-p_s) \;\Rightarrow\; D_{q+} = D_p - D_p^2$$

But this definition of Q violates the postulate of absence of loops in simple graphs. It becomes verified when we change Q in $Q = p\otimes p - D_p^2$. As a consequence, under the spatial independence hypothesis but without imposing a constraint of uniformity of weights, the Laplacian becomes:

$$L_Q = D_{q+} - Q = \left(D_p - D_p^2\right) - \left(p\otimes p - D_p^2\right) = D_p - p\otimes p = D_p - p\,{}^t p$$

Thus, the combinatorial Laplacian which corresponds to independence is $L_Q = D_p P_l^{\perp}$, where $P_l = \left(I_n - \iota^t \iota D_p\right)$ is the projector D_p-orthogonal on the space orthogonal to the subspace of constants on F. A natural choice of p is $p=q_+$ and we shall adopt the simplified notation $H_q = (I_n - \iota^t \iota D_{q+})$ in the remainder of our text. By substitution and elementary algebra, we have proved the following

Lemma 1: *For any set of standardized outer degrees q_+ associated to a graph, we have*

(E1): $H_q = H_q H_q$ (property of a D_{q+}-orthogonal projector);

(E2): ${}^t H_q D_{q+} H_q = {}^t H_q D_{q+} = D_{q+} H_q = D_{q+} - q_+ {}^t q_+$

(E3): The empirical total (D_{q+})-variance of a variable X is

$$\widetilde{S^2}(X) = \sum_s q_{s+}\,(x_s - \hat{x})^2 = \| H_q x \|^2_{D_{q+}} = \| x \|^2_{D_{q+} - q_+ {}^t q_+}$$

where $\tilde{x} = \sum_s q_{s+} x_s$ *is the weighted mean of X in the observed sample.*

Observe too that one well known specification of models of spatial regression, that takes into account spatial auto-correlation is

$$Y = \alpha \boldsymbol{W} Y + \boldsymbol{X}\beta + \epsilon \Leftrightarrow (I_n - \alpha \boldsymbol{W})Y = X\beta + \varepsilon$$

where as said above, $W = D_{q+}^{-1} Q$. But, for $\alpha \in \mathbb{R}$,

$$\alpha\left(D_{q+} - Q\right) + (1-\alpha)D_{q+} = \left(D_{q+} - \alpha Q\right) = D_{q+}(I_n - \alpha W)$$

and when $\alpha \in [0,1]$ this identity shows that α locates the data on a continuum between two baselines: the perfect spatial dependence induced by the (combinatorial Laplacian of the) graph G corresponding to α near to 1, and independence corresponding to α near to 0.

2.1. Moran and Geary revisited

One way to correct the inconsistency noticed above consists substituting the two first weighted moments to the unweighted ones in the scaling transformation of the observations Xin the numerator as well as in the denominator of both indices:

$$\tilde{X} = \sum_s q_{s+} X_s, \quad \widetilde{S^2}(X) = \sum_s q_{s+}(X_s - \tilde{X})^2, \quad Z_s = \frac{X_s - \tilde{X}}{S(X)}$$

Modifying the repartition of weights changes in particular the centering process: the centered vector becomes the image of X by the Q-centering operator $H_{q+} : (I_n - \imath^t \imath D_{q+}) : x \mapsto H_{q+} x = x - \tilde{x}\imath$. The Q-centering operator $\boldsymbol{H_{q+}}$ is symmetric while the usual $\boldsymbol{H_1}$ is I_n – symmetric only, and its role is to center the points cloud at its weighted barycenter $\tilde{X} = \sum_s q_{s+} X_s$. $\tilde{X}$ reduces to a weighted mean in the case $p = 1$, and no more an arithmetic (unweighted) mean $\bar{x}\imath$. The sample variance induced by the differential weights is:

$$\widetilde{S^2}(X) = \sum_s q_{s+}(X_s - \tilde{X})^2 = {}^t x\, {}^t H_{q_+} D_{q+}\, H_{q_+} x = {}^t x\, D_{q+}\, H_{q_+} x = {}^t x\,(D_{q+} - q_+ {}^t q_+) x$$

This introduction of differential weights changes the centering and the scaling, and so it yields two new indices varying in $[0,1]$:

$$\textbf{Moran: } \widetilde{I_M} = \frac{\sum_{st} q_{st}(X_s - \tilde{X})(X_t - \tilde{X})}{\sum_s q_{s+}(X_s - \tilde{X})^2} = \sum_{st} q_{st} Z_s Z_t$$

$$\textbf{Geary: } \tilde{C} = \frac{\frac{1}{2}\sum_{st} q_{st}(X_s - X_t)^2}{\sum_s q_{s+}(X_s - \tilde{X})^2} = \frac{1}{2}\sum_{st} q_{st}(Z_s - Z_t)^2$$

Notice that some problems generated by Q remain: since Q is usually not semi-definite positive, the quadratic form ${}^t\mathbf{x}\mathbf{Q}\mathbf{x}$ does not define an Euclidean norm on *F*, but the additive decomposition $\mathbf{D}_{q+} = \mathbf{L}_Q + \mathbf{Q}$ induces for all $x \in F$: $\tilde{C} + \tilde{I}_M = 1$ since

$$\| x \|^2_{D_{q+}} = \| x \|^2_{L_Q} + {}^t x Q x \Rightarrow 1 = \frac{\|x\|^2_{L_Q}}{\|x\|^2_{D_{q+}}} + \frac{{}^t x Q x}{\|x\|^2_{D_{q+}}} \qquad (1)$$

where for any symmetric and positive semi-definite matrix N, $\| x \|^2_N = {}^t\mathbf{x}\mathbf{N}\mathbf{x}$ designates the associated squared Euclidean norm of x specified by the metric N. This property is satisfied in spatial statistics for $\mathbf{D}_{q+}$ and $\mathbf{L}_Q$, but not for Q !

Finally, we get a more meaningful formula of decomposition of the total variance by applying equation (1) to any q_+ −centered vector $\tilde{x} = H_q x$.

Proposition1:

Any q_+ −centered vector $\tilde{x} = H_q x$ verifies:

$$\| H_q x \|^2_{D_{q+}} = \| H_q x \|^2_{L_Q} + {}^t x C_Q x \text{ , where } \quad C_Q = {}^t H_q Q H_q \qquad (2)$$

The proof is immediate since the right term is a direct application of equation (1), and the first term of this addition results by substitution of $H_q = I_n - q_+ {}^t\iota$ in the semi-metric $N = {}^t H_q L_Q H_q$ and use of $L_Q \iota = 1$■

The equation (2), proves that the two corrected spatial autocorrelation coefficients are basically complementary, since $\tilde{C} = 1 - \tilde{I}_M$

Note also that when the q_{st} are interpretable as proximity indices, $\| x \|^2_{L_Q}$ may be interpreted as a local variance coefficient (after Lebart (1969)), $\| x \|^2_{D_{q+}}$ is a measure of *total variance*, and ${}^t x C_Q x$ is a measure of *global variability*, interpretable as a variance only when Q is semi-definite and positive. Moreover, the second term of this addition is ${}^t x C_Q x$, which generalizes the proposal of Griffith (2000): it is directly linked to the numerator of the classical Moran coefficient, while here, the centering is more general than in Griffith (2000, 2003) since we use the centering projector H_q in place of H_1 which constrains q to be uniform: $q_s = 1/n$, for $s = 1:n$.

2.2. Connections to the Laplace Beltrami operator and to Machine Learning

The Statistical and Machine Learning communities develop powerful statistical methods useful for data mining under the assumption that the data lies on a manifold. Example going, research domains like image analysis often use sources of high-dimensional data, where the number of (redundant) features available is much higher than the intrinsic dimensionality of their support, while this one is highly nonlinear. In such a case, the analysis is complicated by the fact than in high dimension, one can trust only local distances, but not global ones. In the following, we show the relation existing between the statistical analysis based on the New Moran's coefficient and the existing statistical learning methods which look for optimal decompositions of the type:

Data = Riemannian Manifolds with a measure + Noise

This goal generated a renewed interest for Non Linear Dimension Reduction (*aka* NLDR) methods which developed at the turn of the 21-th century, motivated by the following question: How to built faithful and low dimensional representations of data obtained by sampling a probability law distributed over a manifold? Given a set of n points $\{x_s \in R^p,\ s=1:n\}$ the problem is to find a set of n points $\{y_s \in R^d,\ s=1:n\}$ such that $d \ll p$ and each y_s represents x_s with small residuals.

Originally, research carried on the case when the $\{x_s \in R^p,\ s=1:n\}$ were directly observed (*aka* manifest): then, prototypical examples of LDR dimension reduction methods are Principal Component Analysis (*aka* PCA) and Multiple Correspondence Analysis (*aka* MCA) respectively for the case where the measurements of p numerical (resp. categorical) variables are available. In both cases, the $\{y_s \in R^d,\ s=1:n\}$ derive from the d first eigenvectors of the Gramian matrix induced by the measurements $\{x_s \in R^p,\ s=1:n\}$.

Then, the researchers got interested in the case when the $\{y_s \in R^d,\ s=1:n\}$ are latent, and known in an indirect way through sufficient statistics or maximal invariants, such as scalar products (kernel methods), Euclidean distances (distance-based methods) and more generally by some form of proximity measure. In the last two cases, one often used way to operate – especially in ecology – consists in substituting a Principal Coordinate Analysis (*aka* PCoA or Classical Scaling) to PCA in a way to built the $\{y_s \in R^d,\ s=1:n\}$, see e.g. Torgerson (1952), Gower (1966), or d'Aubigny (1989, 2009).

These methods offer three examples of the Linear Dimension Reduction (*aka* LDR) approach, while during the last 20 years, the challenge for Statistical Learning specialists has consisted in extending these methods to the curvilinear manifold setting. Such developments required the adoption of mathematical formulations relevant to differential geometry, and we shall show in the following that one such formalization may be useful for spatial statistics because it is related to the new (corrected) Moran's and Geary's coefficients: it consists in using graph theory to express the neighborhood information contained either in the adjacency matrix or its close relative the Laplacian of a graph to explore or integrate the topology of this graph, in modeling spatial dependencies. Moreover, this approach may be extended to the case when topology does not exhaust the available information, because the analyst also disposes of inter-area proximity measurements, in the form of an edge-weighting function.

The approach promoted by Belkin et al. (2001, 2003) seems to us to provide an illuminating framework looked for, which bears tite links with the foundations of our new measures of spatial dependence for spatial modeling.

While dominant for spatial analysis in ecology and in spatial econometrics, the *Spatial Filtering* method is not devoid of difficulties, and weaknesses which seem generally underestimated by its proponents, e.g. Griffith (2003) or Legendre & Legendre (2012). Its general principle is based on the spectral analysis of the connectivity matrix $C = {}^{t}H_1 \, Q \, H_1$ which is the source of opacity of the results since it results in a liberal application of PCoA to an often non semi-definite matrix Q, see d'Aubigny (1989, 2009). As a matter of fact we are still looking for explicit optimality criteria for this method, more generally satisfied than in the very special case where Q satisfies the constraints ensuring the Euclideanicity of a metric defined on **F**.

d'Aubigny (1989, 2009) considers methods of analysis of proximity data based on an analogy to the modelling of electric networks adopted by Doyle & Snell (1984), and expressed in terms of the combinatorial Laplacian of a graph, see also Bollobas (1991). But, while the presentation of Belkin & Niyogi (2001, 2003) leads to the same formalism, it is mathematically much more grounded and more detailed. Belkin and co-workers call this formalism the Laplacian Filtering approach. Its three main qualities are: first, it preserves optimally (in a given sense) the local neighborhood information; Second, the representation of spatial entities obtained by the algorithm may be interpreted as a discrete approximation of a smooth map derived from the intrinsic geometry of the underlying manifold; and this approximated map is the solution of a classical Heat equation problem, expressed with help of the Laplace Beltrami Operator (LBO) to provide an optimal embedding of the manifold. The connection between the LBO and the

Laplacian of a Graph is well known to geometers and specialists of spectral graph theory, see Chung (1997).

Let us go back to the minimization of the penalty criterion $S(f)$ usable for smoothing a functional $\boldsymbol{f}$ defined on the set of nodes of a graph. A way to control the smoothness of a function **f** defined on a discrete set of points which belong to some Riemannian manifold $\mathcal{M}$ consists in minimizing a roughness penalty criterion, *such as* the sum of squared differences $(\boldsymbol{f_s} - \boldsymbol{f_t})^2$ between neighbor points $\boldsymbol{s} \sim \boldsymbol{t}$, over the m existing edges of the graph:

$$\boldsymbol{S(f)} = \sum_{s\sim t} (\boldsymbol{f_s} - \boldsymbol{f_t})^2 = \boldsymbol{1/2}\ {}^t\boldsymbol{f}\boldsymbol{L_A}\boldsymbol{f}\ .$$

So, the Roughness $\boldsymbol{S(f)}$ of a functional f on $\mathcal{M}$ is controlled by the combinatorial Laplacian $\boldsymbol{L_A}$. The properties of $\boldsymbol{L_A}$ and of its close relatives are discussed in detail in Chung (1997). Belkin & Niyogi (2003), inspired by the past working in mathematical physics, see *e.g.* Rosenberg (1997), consider the formal analogue of this problem in a continuum of points describing a differentiable manifold $\mathcal{M}$ when the analyst ultimately wants to embed this smooth compact Riemannian manifold $\mathcal{M}$ in a linear d-dimensional vector space F. In such a case, the Riemannian structure on $\mathcal{M}$ is induced by the one on F, and the authors notice that if one attaches a roughness penalty in any point $x_i \in F$ defined in its vicinity specified by a small ball of radius δ, the gradient ∇f of the function f (supposed twice differentiable) satisfies:

$$\boldsymbol{1/\delta^n} \int_{\delta} \left(\boldsymbol{f}(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x}+\boldsymbol{\delta})\right)^2 \boldsymbol{p}(\boldsymbol{x})\boldsymbol{d\delta} \approx \| \boldsymbol{\nabla f} \|^2\, \boldsymbol{p}(\boldsymbol{x})$$

But, differential geometry theory says that for any point x of $\mathcal{M}$, the tangent space of $\mathcal{M}$ at point x, noted $\boldsymbol{T\mathcal{M}_x}$, is an Euclidean linear subspace of F equipped with a natural scalar product $<\boldsymbol{u}, \boldsymbol{v}>_{\mathcal{M}}$. The gradient $\nabla f(x)$ is a vector of $\boldsymbol{T\mathcal{M}_x}$ such that given another vector $v \in \boldsymbol{T\mathcal{M}_x}$, one has $df(v) = <\nabla f(x), v>_{\mathcal{M}}$. Thus, the total penalty in F is shown to be a function of the *Laplace Beltrami operator* (*aka* LBO), which is defined as

$$\mathcal{L}(f) = \boldsymbol{\Delta f} \triangleq -\sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^{\,2}}$$

Namely, one has:

$$\int_{\mathcal{M}} \| \boldsymbol{\nabla f} \|^2\, \boldsymbol{p}(\boldsymbol{x})\boldsymbol{dx} = < \boldsymbol{f}, \boldsymbol{\Delta f} >_{\mathcal{M}} \qquad (3)$$

Equation (3) shows that the LBO L is a symmetric semidefinite operator whose spectrum is discrete: its eigenvalues are conventionally numbered in increasing order $0 = \lambda_0 \leqq \lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n$, and we shall note f_i the

eigenfunction associated to λ_i Specifically, one may verify that $f_0=\imath$ is the constant function of coordinates uniformly equal to 1.

LBO L has been often studied in mathematics and physics, because of its role in the modeling of the Heat flows, i.e. the diffusion of heat over space and time. Let $u(x,t)$ be the heat distribution at time $t\in\mathbb{R}$, with initial distribution $\boldsymbol{u(x,0)=f(x), x\in F}$. Then, the heat equation writes $\Delta u(x,t)=\frac{d}{dt}u(x,t)$ and its solution is obtained by convolution of $\boldsymbol{f}$ with the heat kernel :

$$\boldsymbol{u(x,t)=\int_F f(y)\,h(x,y)dy,\qquad h(x,y)=(4\pi t)^{-n/2}\exp\left(-\frac{\|x-y\|^2}{4t}\right)}$$

When we take the limit of the derivative of the solution of the heat equation for $\boldsymbol{t\mapsto 0}$, we get:

$$\boldsymbol{\Delta u(x,t)=\frac{d}{dt}\left[\int_F f(y)\,h(x,y)dy\right]_0=-\frac{1}{t}\{f(x)h(x,x)-\int_F f(y)\,h(x,y)dy\}}$$

$\Delta u(x,t)$ may be approximated by $-\frac{1}{t}\{f(x)h(x,x)-\sum_{s=1}^{n}f(x_s)\,h(x,x_s)\}$ from any sample of n empirical data. Belkin & Niyogi (2001, 2003) demonstrated the convergence (in the infill asymptotics sense) of the structure of the induced approximate combinatorial Laplacian L_Q^n to the structure of the underlying manifold $\mathcal{M}$ induced by its latent Laplace Beltrami operator, when its number n of sampled vertices grows to infinity under various sampling schemas and Belkin et al. (2009) extend these results to the case of points clouds. From a practical point of view, these theorems are important because they provide objective arguments to assert the existence of a relative stability of the empirical embedding obtained by the spectral analysis of the combinatorial Laplacian and its expectable robustness against the Modifiable Areal Unit Problem (MAUP).

2.3. How to extract manifold structures from data?

The link between LB0 and the combinatorial Laplacian may be explicited in taking:

$$\boldsymbol{Q=(q_{ij}),\ q_{ij}=\exp\left(-\frac{\|x_i-x_j\|^2}{4t}\right),\ q_{i+}=\sum_{j=1}^{n}q_{ij}} \qquad (4)$$

This choice gives $L_Q f(x_i) = q_{i+} f(x_i) - \sum_{j=1}^{n} q_{ij} f(x_j)$ and induces the index of spatial smoothness we looked for: ${}^t f L_O f = 2\sum_{j\sim 1}^{n} q_{ij}(f(x_i) - f(x_j))^2$. This is precisely the numerator of the modified Moran's coefficient.

Of practical interest for the analyst, Belkin & Niyogi (2003) point out that it is necessary to restrain consideration to pairs of points close enough (say less than a fixed ε) in order to ensure the positive semi-definiteness of the approximation matrix $\boldsymbol{L_Q}$. So they propose to compute the graph Laplacian by the local formula:

$$\boldsymbol{q_{ij} = \exp\left(-\frac{\| x_i - x_j \|^2}{4t}\right) \; if \; \| x_i - x_j \|^2 < \varepsilon, and \; 0 \; otherwise}$$

This restriction of the retained edges to close neighbors has been progressively advised in ecology, as a statement of experience and the object of simulation results. In fact, ecologists often extend the use of the Gaussian kernel (4) to $q_{ij} = \exp(-d_{ij}^2 / 4t)$ where d_{ij} is any dissimilarity index which is chosen on thematic arguments. As far as I know, the theoretical consequences of this relaxation on optimality properties of the method are not known yet.

Whatever this choice is, the evaluation and analysis of the spatial dependence constitutes only a preliminary step, which, in econometrics, is most often followed by the specification and validation studies of some linear model controlled for some form of spatial autocorrelation operating, either on the residuals – as in the Simultaneous Autoregressive Error model (SAR), *cf.* Whittle 1954) – or on the response – as in the Conditional Autoregressive model (CAR), *cf.* Besag 1974) – see also Anselin (1988, 1995, 2014).

With some variants induced by the specificities of the questionings central to both disciplines, an analogous methodology has developed in ecology, with greater dynamism and under various names: by now it is promoted under the name of *distance-based Moran's Eigenvector Map* (db-MEM), see *e.g.* Borcard & Legendre (2002), Dray *et al.* (2006), Legendre & Legendre (2012, Ch. 14).

One technical question remains in practice: how to choose the threshold ε? The fashionable practice in ecology consists in a minimum spanning tree of the graph G, in a way to fix $\varepsilon = l_t$, where l_t is the length of the longest edge of this spanning tree. In the next steps of the analysis process, one can retain only edges closer than $\varepsilon = l_t$, see Legendre & Legendre (2012, Ch. 14).

A challenger methodology may be practiced in coherence with the use of the corrected autocorrelation coefficients presented in this paper since,

maximizing the corrected Moran's coefficient is equivalent to minimizing the corrected Geary's. This minimum is realized by v^1, the smallest solution D_{q+}-orthogonal of the generalized eigenvector problem defining the spectral analysis of the combinatorial Laplacian $\mathbf{L_Q}$, associated to a non-null eigenvalue:

$$\max {}_{tvD_{q+}v} \widetilde{I_M}(v) = \min_{{}_{tvD_{q+}v}} \frac{\|\mathbf{v}\|^2_{\mathbf{L_Q}}}{\|\mathbf{v}\|^2_{\mathbf{D_{q+}}}} \Leftrightarrow \mathbf{L_Q v^1} = \boldsymbol{\lambda_1} \mathbf{D_{q+}} \mathbf{v^1} \qquad (5)$$

This process may be repeated sequentially under the constraints of D_{q+}-orthogonality of the solution vectors and as is classical in multivariate analysis, the solution of dimensionality k is given by the D_{q+} -orthogonal generating system of eigenvectors $\left\{v^1, v^2, ..., v^k; k \leq (n-1)\right\}$ satisfying $\left\{{}^t v^j D_{q+} v^k = 0 \; if \;\; j \neq k\right\}$ and associated to the corresponding eigenvalues ordered as the corresponding eigenvalues: $\lambda_0 = 0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. The introduction of the metric $N=D_{q+}$ on F is noticeable here, since then a point i accounts especially in the spectral decomposition of $\mathbf{L_Q}$ if it has more (and closer) neighbors, in proportion to its participation in the definition of spatial structures.

Notice also that the solutions of the eigenvector problem (5) differ from those of db-MEM and of Spatial Filtering in two respects: i) the centering are not the same, and ii) here, orthogonality must be understood in the metric D_{q+} that is as D_{q+} - orthogonality. Finally, when we explicit the role of $\mathbf{Q}$ in (3), we get an eigenequation different from the db-MEM and Spatial Filtering ones:

$$\mathbf{L_Q} v^s = \boldsymbol{\lambda_s} \mathbf{D_{q+}} v^s \Leftrightarrow \mathbf{Q}\, v^s = (1 - \boldsymbol{\lambda_s})\, \mathbf{D_{q+}}\, v^s \qquad (6)$$

This equation defines a *Laplacian Filtering Analysis*, whose resulting eigenvectors and eigenvalues are in fact solutions of the *Eigenmap Algorithm* due to Belkin & Niyogi (2001). These specialists of machine learning do not make reference to spatial data analysis, but their work justifies the use of the Laplacian in this context by an argument of approximation – through the spectral decomposition (5) of the weighted Laplacian - of the geometric structure of the underlying manifold $\mathcal{M}$,attached to the corresponding Laplace Beltrami operator L. Let us also point out that Lebart (1969) initiated related work motivated by applications to spatial data analysis, and named *Local Analysis*.

3. Conclusion and discussion

In starting the work reported here, I wanted to deepen our understanding of formulas giving the Moran's and Geary's coefficients of spatial autocorrelation, and their relationships to other indices discussed in Getis & Ord (1992). Over time, I found that part of their complexity is due to an incoherence in the distributional

assumptions made in defining their numerator and their denominator. As explained above, we did use geometric arguments to correct the problem and then, I discovered that it simplified considerably their relationship derived in books like Cliff & Ord (1980) or Tiefelsdorf (2000), in a way which makes closer the ties between geostatistics and statistical methods adapted to the analysis of areal data. In terms of methodological contributions, the geometric justification used for the proposed changes, formalized in the framework of the of Multivariate Exploratory Data Analysis in the French tradition – see Cailliez & Pages (1976) or Escoufier (1987) – also resulted in the proposal of series of new methods – not discussed here – that I consider as close relatives of the dominant existing ones, like db-MEM in ecology and Spatial Filtering proposed by Griffith in geography and econometrics. Their advantage when compared to these established method lies in the fact that they are free of the problems induced by the use of non definite-positive weight matrices Q, and that, as proved in the machine learning literature, they are supported by a well known theoretical model of physics: the Heat Equation. I must add that the spectral decomposition used in the Laplacian Filtering method presented above, did generate in the twenty years a huge literature in Machine learning, mainly motivated by applications in image analysis. Its use in clustering is becoming dominant in that setting because of repeated successes in applications – in particular for big data – and good and specific theoretical qualities, see *e.g.* Shi & Malik (2000), Qiu & Hancock (2007), or Saerens et al. (2004). It is likely to spread in the community of geographers and economists formed to multivariate spatial statistics.

A second possible generalization of the methods presented in this paper is the coupling of two (or several) data tables. It has been the subject of numerous studies in ecology because of the combinatorics of possible variants, see *e.g.* Dray *et al.* (2003). In this setting one considers two triples (X, M_{XX}, N) and (Y, M_{YY}, N) where X and Y are two data tables containing the measurements of the two sorts of variables on the same set of geographical entities. Then X (resp. Y) permit to represent the spatial entities in an Euclidean space E_X (resp. E_X). One can analyze separately each triple by one of the methods discussed above, but we may also want to study them simultaneously. This is in fact easily possible: Chessel & Mercier (1993) show that one can find two vectors $u_1 \in E_X$ and such that for any distribution $N=D_v$, the covariance $COV(a,b)={}^taNb$ be maximum for $a^1=XM_{XX}u^1$ and $b^1=XM_{XX}v^1$. In fact, an elementary proof shows that the u^s and v^s are eigenvector solutions (s=1:n) of the analysis of the triple (S, M_{XX}, M_{YY}) , where $S={}^tYNX$. The method is very general as argued by Dray et al. (2003) the solutions optimize $COV(a,b)=COR(a,b)\times STD(a) \times STD(b)$, a mixed criterion. The question Here is how to take space into account? What arguments justifies the choice by the authors of $N=D_{q+}$ here? I noticed too that the same statistical arguments can apply to the triples of simple contrasts (∇X, M_{XX}, $N=D_O$) and (∇Y, M_{YY}, $N=D_O$), with equal formal justification, and the advantage to explicitly introduce the spatial

structure in the formulation of the model. I observed with interest that in this case, the solution is built by a PCA of the triple, (T, M_{XX}, M_{YY}) where $T={}^t(\nabla Y)D_O\nabla X={}^tYL_OX$. So, the solution depends on the combinatorial Laplacian.

Clearly, this line of research has a lot to teach us yet!

Acknowlegements: The present research benefited from a financial support by the french ministery of research and education, as relevant to the Action Concertée Incitative (ACI) program named Terrains, Techniques, Théorie*: "travail interdisciplinaire en Sciences Humaines et Sociales*".

References

Anselin L. (1988), *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht, The Netherland.

Anselin L. (1995), *Local indicators of spatial association - LISA*. Geographical Systems, 3: 1–13.

Anselin L. & Rey S.J. (2014), *Modern Spatial Econometrics in Practice*, GeoDa Press LLC, Chicago IL, USA.

Aubigny (Drouet d') G. (1989), *L'Analyse Multidimensionnelle des Données de Dissimilarité*, Thèse de Doctorat d'état es Sciences Mathématiques, Université Joseph Fourier – Grenoble I, France.

Aubigny (d') G. (2006), *Dépendance spatiale et auto-corrélation*, in J.~J. Droesbeke, M. Lejeune & G. Saporta (Eds.), *Analyse Statistique des Données Spatiales*, Editions TECHNIP, Paris, France: Chap 2: 17–45.

Aubigny (d') G. (2009), *The Analysis of Proximity Data*, in Govaert G. (Ed.), *Data Analysis*, John Wiley & sons Inc., Hoboken, USA: Chap 4: 93–147.

Aubigny (d') G. (2012), *Analyse contextuelle et modélisations multiniveaux des Données Electorales.* Coordinateur principal, Action Concertée Incitative '*Terrains, Techniques, Theorie: travail interdisciplinaire en Sciences Humaines et Sociales*'. Rapport de fin de projet, Grenoble, France. 148 pages.

Aubigny (d') C. & Aubigny (d') G. (2009), *New LISA indices for spatio-temporal Data Mining*, XVI-èmes Rencontres de la Société Francophone de Classification, Grenoble, 2–4 Septembre, France.

Bapat R.B. (2010), *Graphs and Matrices*, Springer, New York, USA.

Belkin M. & Niyogi P. (2001), *Laplacian Eigenmaps and Spectral techniques for Embedding and Clustering.* Advances in Neural Information Processing Systems, 595–591.

Belkin M. & Niyogi P. (2003), *Laplacian Eigenmaps for Dimensionality Reduction and Data.* Neural Computation, Vol. 15, No 6: 1373–1396.

Belkin M., Sun J. & Wang Y. (2009), *Constructing Laplace Operator from Point Clouds in* $\mathbb{R}^d$. In Proceedings of the Symposium on Discrete Algorithms, 1031–1040.

Besag J. (1974), *Spatial interaction and the statistical analysis of lattice systems.* Journal of the Royal Statistical Society Series B 36:192–236.

Bollobas B. (1990), *Modern Graph Theory,* Springer, New-York, USA.

Borcard D. & Legendre P. (2002), *All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices.* Ecological Modelling 153 : 51–68.

Cailliez F. & Pages A.J. (1976), *Introduction à l'Analyse des Données,* SMASH, Paris, France.

Chessel D. & Mercier P. (1993), *Couplage de triplets statistiques et liaisons espèce environnement.* In: *Biométrie et environnement.* J.D. Lebreton et B. Asselin (Eds.), Masson, Paris, France, 1993.

Chung F.R.K. (1997), *Spectral Graphs Theory,* American math. Society Ed., CBMS 92, USA.

Cliff A.D. & Ord J.K. (1981), *Spatial Processses: Models and Applications,* Pion Limited, London, UK.

Doyle P.G. & Snell J.L. (1984), *Random Walks and Electric Networks,* Carus Mathematical Monographs Number 22, The Mathematical Association of America, Washington D.C, USA.

Dray S., Chessel D. & Thioulouse J. (2003), *Co-inertia Analysis and the Linking of Ecological Data Tables*. Ecology 84(11):3078–3089.

Dray S., Legendre P. & Peres-Neto P.R. (2006), *Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM).* Ecological Modeling 196: 483–493.

Escoufier Y. (1987). *The Duality Diagram: a means for better practical applications,* in Legendre, P. & Legendre L. (Eds.), *Developments in Numerical Ecology:* NATO ASI Series, Series G: Ecological Sciences, Vol 14. Springer, New-York, USA: 139–156.

Geary, R.C. (1954), *The Contiguity Ratio and Statistical Mapping*. The Incorporated Statistician 5: 115–145.

Getis A. and J.K. Ord (1992). *The analysis of spatial association by use of distance statistics.* Geographical Analysis, 24: 189-206.

Gower J.C. (1966), *Some distance properties of latent root and vector methods used in multivariate Analysis.* Biometrika, 55: 325–388.

Griffith D.A. (2000), *A linear regression solution to the spatial autocorrelation problem.* Journal of Geographical Systems 2: 141–156.

Griffith D.A. (2003), *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization (Second Edition),* Springer, New-York, USA.

Lebart L. (1969), *Analyse statistique de la contiguité*. Publications de l'Institut de Statistique de l'Université de Paris, 28, pp. 81–112.

Legendre P. & Legendre L (2012), *Numerical Ecology* (Third English Edition), ELSEVIER, Amsterdam, The Netherland.

Moran P.A.P. (1950), *Notes on continuous stochastic phenomena*. Biometrika 37:17–23.

Qiu H. & Hancock E.R. (2007), *Clustering and Embedding Using Commute Times.* IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 22, No 8: 888–905.

Rosenberg S. (1997), *The Laplacian on a Riemannian Manifold*, Cambridge University Press, Cambridge, USA.

Saerens M., Fouss F., Yen L. & Dupont P. (2004), *The Principal Components Analysis of a Graph, and its relationships to Spectral Clustering.* Proc. 15th European Conference in Machine Learning Vol. 3201: 371–383.

Shi J. & Malik J. (2000), *Normalized Cuts and Image Segmentation*. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, No 8: 888–905.

Tiefelsdorf M. (2000), *Modeling Spatial Processes: The identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I*, Springer, New-York, USA.

Torgerson W.S. (1952), *Multidimensional Scaling, 1: Theory and Methods.* Psychometrika, 17: 401–417.

Whittle P. (1954), *On stationary processes in the plane*. Biometrika 41: 434–449.

Streszczenie

NARZĘDZIE STATYSTYCZNE DO ANALIZY EKSPLORACYJNEJ ORAZ MODELOWANIA DANYCH PRZESTRZENNYCH

Wiekszość analiz eksploracyjnych danych przestrzennych rozpoczyna się od oceny próby jednostek przestrzennych, pod względem występowania oraz siły autokorelacji przestrzennej dla zbioru zmiennych, stanowiących atrybuty jednostek przestrzennych. Trafność aplikacji najbardziej cenionych narzędzi weryfikacji autokorelacji przestrzennej – współczynników Morana oraz Geary'ego jest rzadko kwestionowana, pomimo faktu, że w przypadku opisywania ich własności wielu użytkowników zdaje się popełniać błędy oraz wprowadzać nieład. Artykuł rozpoczyna się od krytycznej oceny klasycznej definicji indeksów. Założono, że pomimo intuicyjnej konstrukcji, koncepcja indeksów boryka się z brakiem spójności w przypadku wielu ich składowych. Następnie zaproponowano korektę współczynników autokorelacji przestrzennej, która upraszcza ich relacje, i otwiera drogę do włączenia statystyk do zestawu narzędzi statystycznych, modelowania oraz wizualizacji. W drugiej części zaprezentowano teoretyczne przesłanki konstruowania wielowymiarowych narzedzi statystycznych, uwzględniających skorygowane definicje współczynników autokorelacji przestrzennej, zaczerpnięte z ostatnich prac w dziedzinie statystyki. Przedstawione metody eksploracyjnej wielowymiarowej analizy danych charakteryzują się łatwością zastosowania oraz oprogramowania z wykorzystaniem dostępnych, darmowych pakietów.

Słowa kluczowe: *analiza wielowymiarowa, graf dualności, autokorelacja przestrzenna, współczynnik Morana, mapa wektora własnego statystyki Morana, operator Laplace'a, funkcja własna filtracji przestrzennej*