The Annals of Applied Statistics
2011, Vol. 5, No. 4, 2278–2299
DOI: 10.1214/10-AOAS356

REVISITING GUERRY'S DATA: INTRODUCING SPATIAL CONSTRAINTS IN MULTIVARIATE ANALYSIS

BY STÉPHANE DRAY AND THIBAUT JOMBART

Université Lyon 1 and Imperial College

Standard multivariate analysis methods aim to identify and summarize the main structures in large data sets containing the description of a number of observations by several variables. In many cases, spatial information is also available for each observation, so that a map can be associated to the multivariate data set. Two main objectives are relevant in the analysis of spatial multivariate data: summarizing covariation structures and identifying spatial patterns. In practice, achieving both goals simultaneously is a statistical challenge, and a range of methods have been developed that offer trade-offs between these two objectives. In an applied context, this methodological question has been and remains a major issue in community ecology, where species assemblages (i.e., covariation between species abundances) are often driven by spatial processes (and thus exhibit spatial patterns).

In this paper we review a variety of methods developed in community ecology to investigate multivariate spatial patterns. We present different ways of incorporating spatial constraints in multivariate analysis and illustrate these different approaches using the famous data set on moral statistics in France published by André-Michel Guerry in 1833. We discuss and compare the properties of these different approaches both from a practical and theoretical viewpoint.

1. Introduction. A recent study [Friendly (2007)] revived André-Michel Guerry's (1833) *Essai sur la Statistique Morale de la France*. Guerry gathered data on crimes, suicide, literacy and other "moral statistics" for various départements (i.e., counties) in France. He provided the first real social data analysis, using graphics and maps to summarize this georeferenced multivariate data set. The work of Friendly (2007) contained a historical part describing Guerry's life and work in detail. In a second part, Friendly reanalyzed Guerry's data using a variety of modern tools of multivariate and spatial analysis. He considered two main approaches to analyzing a data set involving both multivariate and geographical aspects: data-centric (multivariate analysis) and map-centric (multivariate mapping) displays. In the first approach, the multivariate structure is first summarized using standard analysis methods [e.g., principal component analysis, Hotelling (1933)] and visualization methods [e.g., biplot, Gabriel (1971)]. The geographic information is only added a posteriori to the graphs, using colors or other visual attributes.

Received November 2009; revised April 2010.
Key words and phrases. Autocorrelation, duality diagram, multivariate analysis, spatial weighting matrix.

This approach thus favors the display of multivariate structures over spatial patterns. On the other hand, multivariate mapping (i.e., the representation of several variables on a single map using multivariate graphs) emphasizes the geographical context but fails to provide a relevant summary of the covariations between the variables. Moreover, multivariate mapping raises several technical issues such as the lack of readability of multivariate symbols (e.g., Chernoff faces), which can only be used to represent a few variables and are sometimes difficult for non-specialists to interpret. Friendly (2007) stated that Guerry's *questions, methods and data still present challenges for multivariate and spatial visualization today*. While he acknowledged progress in both exploratory spatial data analysis and multivariate methods, he also suggested that *the integration of these data-centric and map-centric visualization and analysis is still incomplete*. He concluded his paper with a motivating question: *Who will rise to Guerry's challenge?*.

This challenge has been one of the major methodological concerns in community ecology (and in other disciplines, e.g., public health) over the last few decades. Community ecology is a subdiscipline of ecology that aims to understand the organization and causes of species associations. As community data are essentially multivariate (many species, many sites, many environmental factors and complex spatio-temporal sampling designs), questions about the structure and drivers of ecological communities have traditionally been addressed through multivariate analyses [Legendre and Legendre (1998)]. Hence, it has been and remains a very fertile field for the development and the application of multivariate techniques. One of the most active research goals in ecology today is to understand the relative importance of processes that determine the spatial organization of biodiversity at multiple scales [Legendre (1993)]. As a consequence, the last decade has seen efforts in the methodological domain to render the multivariate analysis of community data more spatially explicit or, conversely, to generalize analyses of spatial distributions to handle the covariation of many species. These methods allow us to identify the main spatial patterns by considering simultaneously both multivariate and geographical aspects of the data. They thus represent a first step toward the integration of data-centric and map-centric visualizations into a single method.

In this paper we take up Friendly's challenge by demonstrating how several spatially-explicit multivariate methods developed initially in the context of community ecology could also be of benefit to other fields. We present different ways of incorporating the spatial information into multivariate analysis, using the duality diagram framework [Escoufier (1987)] to describe the mathematical properties of these methods. We illustrate these different methodological alternatives by reanalyzing Guerry's data.

2. Standard approaches. We use the data set compiled by Michael Friendly and available at http://www.math.yorku.ca/SCS/Gallery/guerry/. This data set has

TABLE 1
Variable names, labels and descriptions. Note that four variables have been recorded in the form of "Population per..." so that low values correspond to high rates, whereas high values correspond to low rates. Hence, for all of the variables, more (larger numbers) is "morally" better

Label	Description
Crime_pers	Population per crime against persons
Crime_prop	Population per crime against property
Literacy	Percent of military conscripts who can read and write
Donations	Donations to the poor
Infants	Population per illegitimate birth
Suicides	Population per suicide

been recently analyzed by Dykes and Brunsdon (2007) to illustrate a new interactive visualization tool and is now distributed in the form of an R package [see Dray and Jombart (2010) for details]. We consider six key quantitative variables (Table 1) for each of the 85 départements of France in 1830 (Corsica, an island and often an outlier, was excluded). In this section we focus on classical approaches that consider either the multivariate or the spatial aspect of the data. In the next sections we will present methods that consider both aspects simultaneously.

2.1. *Multivariate analysis*. Multivariate analysis allows us to identify and summarize the primary underlying structures in large data sets by removing any redundancy in the data. It aims to construct a low-dimensional space (e.g., 2 or 3 dimensions) that retains most of the original variability of the data. The classical output consists of graphical summaries of observations and variables that are interpreted for the first few dimensions.

2.1.1. *The duality diagram theory.* Multivariate data are usually recorded in a matrix $\mathbf{X}$ with n rows (observations) and p columns (variables). The duality diagram is a mathematical framework that defines a multivariate analysis setup using a set of three matrices. We can consider the (possibly transformed) data matrix $\mathbf{X}$ ($n \times p$) as a part of a statistical triplet ($\mathbf{X}, \mathbf{Q}, \mathbf{D}$), where $\mathbf{Q}$ ($p \times p$) and $\mathbf{D}$ ($n \times n$) are usually symmetric positive definite matrices used as metrics [i.e., $\mathbf{Q}$ provides a metric for the variables (columns of $\mathbf{X}$) and $\mathbf{D}$ provides a metric for the observations (rows of $\mathbf{X}$)]. This unifying mathematical framework encompasses very general properties, which will be described, to the analysis of a triplet. For more details, the reader should consult Escoufier (1987), Holmes (2006) or Dray and Dufour (2007). The mathematical properties of each particular method (corresponding to a particular choice of matrices $\mathbf{X}$, $\mathbf{Q}$ and $\mathbf{D}$) can then be derived from the general properties of the diagram. Note that the analysis of the duality diagram associated to the triplet ($\mathbf{X}, \mathbf{Q}, \mathbf{D}$) is equivalent to the generalized singular value

decomposition [GSVD, e.g., Greenacre (1984), Appendix A] of $\mathbf{X}$ with the metrics $\mathbf{Q}$ and $\mathbf{D}$.

The analysis of the diagram consists of the eigen-decomposition of the operators $\mathbf{XQX}^{\mathrm{T}}\mathbf{D}$ or $\mathbf{X}^{\mathrm{T}}\mathbf{DXQ}$. These two eigen-decompositions are related to each other (*dual*) and have the same eigenvalues. Thus, we have

$$\mathbf{XQX}^{\mathrm{T}}\mathbf{DK} = \mathbf{K}\boldsymbol{\Lambda}_{[r]},$$
$$\mathbf{X}^{\mathrm{T}}\mathbf{DXQA} = \mathbf{A}\boldsymbol{\Lambda}_{[r]}.$$

r is called the rank of the diagram, and the nonzero eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$ are stored in the diagonal matrix $\boldsymbol{\Lambda}_{[r]}$.

$\mathbf{K} = [\mathbf{k}^1, \ldots, \mathbf{k}^r]$ is a $n \times r$ matrix containing the r nonzero associated eigenvectors (in columns). These vectors are $\mathbf{D}$-orthonormalized (i.e., $\mathbf{K}^{\mathrm{T}}\mathbf{DK} = \mathbf{I}_r$) and are usually called the *principal components*.

$\mathbf{A} = [\mathbf{a}^1, \ldots, \mathbf{a}^r]$ is a $p \times r$ matrix containing the r nonzero eigenvectors (in columns). These vectors are $\mathbf{Q}$-orthonormalized (i.e., $\mathbf{A}^{\mathrm{T}}\mathbf{QA} = \mathbf{I}_r$) and are usually called the *principal axes*.

The row scores $\mathbf{R} = \mathbf{XQA}$ are obtained by projection of the observations (rows of $\mathbf{X}$) onto the principal axes. The vectors $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^r$ successively maximize, under the $\mathbf{Q}$-orthogonality constraint, the following quadratic form:

$$Q(\mathbf{a}) = \mathbf{a}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{DXQa}. \tag{1}$$

If $\mathbf{D}$ defines a scalar product, then we have $Q(\mathbf{a}) = \|\mathbf{XQa}\|_{\mathbf{D}}^2$.

The column scores $\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{DK}$ are obtained by projection of the variables (columns of $\mathbf{X}$) onto the principal components. The vectors $\mathbf{k}^1, \mathbf{k}^2, \ldots, \mathbf{k}^r$ successively maximize, under the $\mathbf{D}$-orthogonality constraint, the following quadratic form:

$$S(\mathbf{k}) = \mathbf{k}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{XQX}^{\mathrm{T}}\mathbf{Dk}. \tag{2}$$

If $\mathbf{Q}$ defines a scalar product, then we have $S(\mathbf{k}) = \|\mathbf{X}^{\mathrm{T}}\mathbf{Dk}\|_{\mathbf{Q}}^2$. Usually, the outputs (column and row scores) are only interpreted for the first few axes (dimensions).

2.1.2. *Application to Guerry's data.* Here we consider $p = 6$ variables measured for $n = 85$ observations (départements of France). As only quantitative variables have been recorded, principal component analysis [PCA, Hotelling (1933)] is well adapted. Applying PCA to the correlation matrix where $\mathbf{Q} = \mathbf{I}_p$, $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$ and $\mathbf{X}$ contains z-scores, we obtain $Q(\mathbf{a}) = \|\mathbf{XQa}\|_{\mathbf{D}}^2 = \mathrm{var}(\mathbf{XQa})$ and $S(\mathbf{k}) = \|\mathbf{X}^{\mathrm{T}}\mathbf{Dk}\|_{\mathbf{Q}}^2 = \sum_{j=1}^{p} \mathrm{cor}^2(\mathbf{k}, \mathbf{x}^j)$ from equations (1) and (2). Hence, this PCA summarizes the data by maximizing simultaneously the variance of the projection of the observations onto the principal axes and the sum of the squared correlations between the principal component and the variables.

For didactic purposes, following Friendly (2007), we interpret two dimensions, while the barplot of eigenvalues (Figure 1A) would rather suggest a 1-D or a 3-D

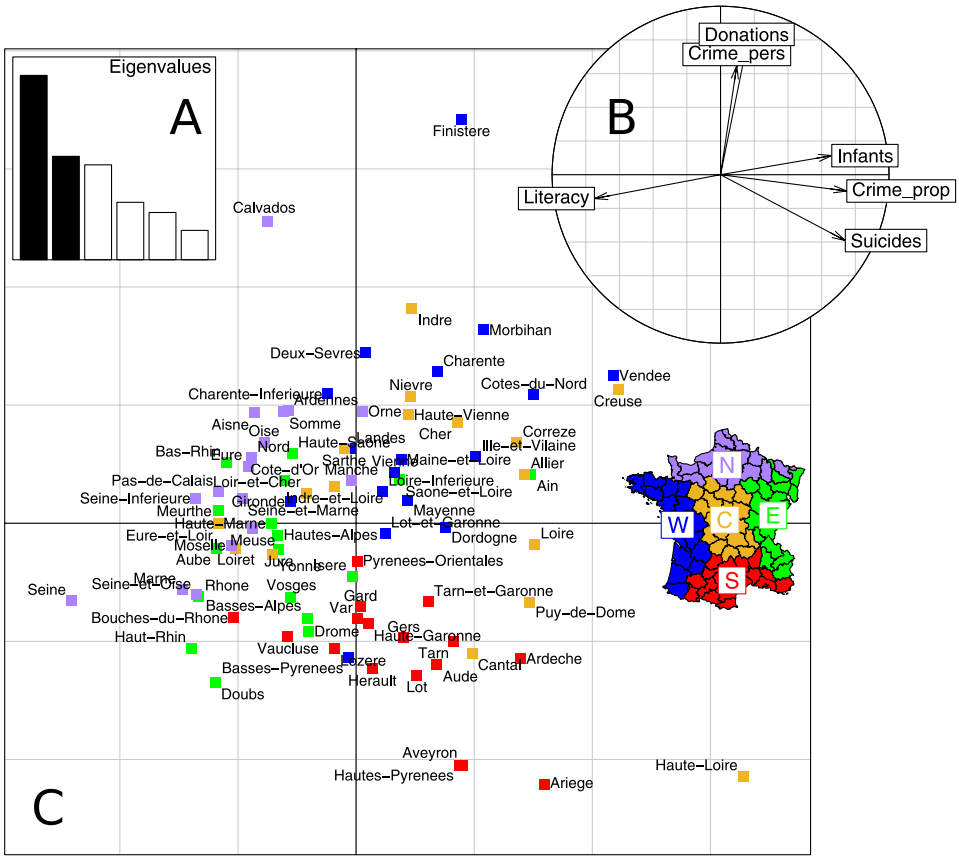

FIG. 1. *Principal component analysis of Guerry's data.* (A) *Barplot of eigenvalues.* (B) *Correlation between variables and principal components.* (C) *Projections of départements on principal axes. The color of each square corresponds to a region of France.*

solution. The first two PCA dimensions account for 35.7% and 20%, respectively, of the total variance. The correlations between variables and principal components are represented on the correlation circle in Figure 1B. As we have excluded Corsica (an outlier) in the present paper, the results are slightly different from those reported in Friendly (2007). The first axis is negatively correlated to literacy and positively correlated to property crime, suicides and illegitimate births. The second axis is aligned mainly with personal crime and donations to the poor. As we are also interested in spatial patterns, we have added geographical information in the form of color symbols on the factorial map of départements (Figure 1C). Each color corresponds to one of five regions of France. The results are quite difficult to interpret, but some general patterns can be reported. For the first axis, the North and East are characterized by negative scores, corresponding to high levels of literacy and high numbers of suicides, crimes against property and illegitimate births.

The second axis mainly contrasts the West (high donations to the the poor and low levels of crime against persons) to the South.

2.2. *Spatial autocorrelation.* Exploratory spatial data analysis (ESDA) is a subset of exploratory data analysis [EDA, Tukey (1977)] that focuses on detecting spatial patterns in data [Haining (1990)]. In this context, spatial autocorrelation statistics, such as Moran (1948)'s Coefficient (MC) and the Geary (1954) Ratio, aim to measure and analyze the degree of dependency among observations in a geographical context [Cliff and Ord (1973)].

2.2.1. *The spatial weighting matrix.* The first step of spatial autocorrelation analysis is to define a $n \times n$ spatial weighting matrix, usually denoted $\mathbf{W}$. This matrix is a mathematical representation of the geographical layout of the region under study [Bivand (2008)]. The spatial weights reflect a priori the absence ($w_{ij} = 0$), presence or intensity ($w_{ij} > 0$) of the spatial relationships between the locations concerned. Spatial weighting matrices can be usefully represented as graphs (neighborhood graphs), where nodes correspond to spatial units (départements) and edges to nonnull spatial weights.

The simplest neighborhood specification is a connectivity matrix $\mathbf{C}$, in which $c_{ij} = 1$ if spatial units i and j are neighbors and $c_{ij} = 0$ otherwise. More sophisticated definitions [Getis and Aldstadt (2004); Dray, Legendre and Peres-Neto (2006)] are able to take into account the distances between the spatial units or the length of the common boundary between the regions for areal data. In the case of Guerry's data, we simply defined a binary neighborhood where two départements i and j are considered as neighbors ($c_{ij} = 1$) if they share a common border (Figure 2).

The connectivity matrix $\mathbf{C}$ is usually scaled to obtain a spatial weighting matrix $\mathbf{W}$, most often with zero diagonal. The row-sum standardization (elements sum to 1 in each row) is generally preferred; it is obtained by

$$w_{ij} = \frac{c_{ij}}{\sum_{j=1}^{n} c_{ij}}.$$

Alternative standardizations are discussed in Tiefelsdorf, Griffith and Boots (1999).

2.2.2. *Moran's coefficient.* Once the spatial weights have been defined, the spatial autocorrelation statistics can then be computed. Let us consider the n-by-1 vector $\mathbf{x} = [x_1 \cdots x_n]^{\mathrm{T}}$ containing measurements of a quantitative variable for n spatial units. The usual formulation for Moran's coefficient of spatial autocorrelation [Cliff and Ord (1973); Moran (1948)] is

$$\text{(3)} \quad \mathrm{MC}(\mathbf{x}) = \frac{n \sum_{(2)} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{(2)} w_{ij} \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \text{where } \sum_{(2)} = \sum_{i=1}^{n} \sum_{j=1}^{n} \text{ with } i \neq j.$$

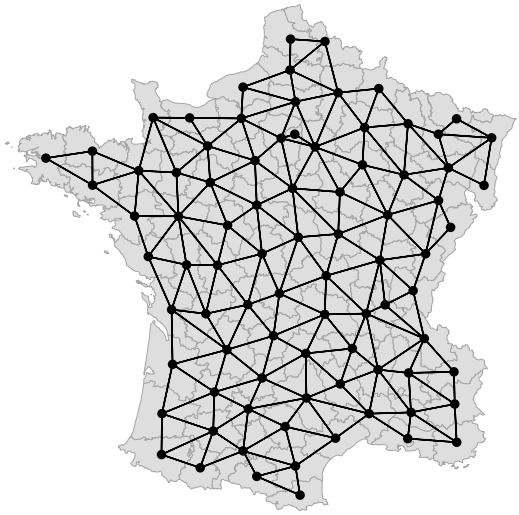

FIG. 2. *Neighborhood relationships between départements of France.*

MC can be rewritten using matrix notation:

$$\text{MC}(\mathbf{x}) = \frac{n}{\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1}} \frac{\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\mathbf{z}}, \tag{4}$$

where $\mathbf{z} = (\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}/n)\mathbf{x}$ is the vector of centered values ($z_i = x_i - \bar{x}$) and $\mathbf{1}_n$ is a vector of ones (of length n).

The numerator of MC corresponds to the covariation between contiguous observations. This covariation is standardized by the denominator, which measures the variance among the observations. The significance of the observed value of MC can be tested by a Monte Carlo procedure, in which locations are permuted to obtain a distribution of MC under the null hypothesis of random distribution. An observed value of MC that is greater than that expected at random indicates the clustering of similar values across space (positive spatial autocorrelation), while a significant negative value of MC indicates that neighboring values are more dissimilar than expected by chance (negative spatial autocorrelation).

We computed MC for Guerry's data set using the row-standardized definition of the spatial weighting matrix associated with the neighborhood graph presented in Figure 2. A positive and significant autocorrelation is identified for each of the six variables (Table 2). Thus, the values of literacy are the most covariant in adjacent departments, while illegitimate births (Infants) covary least.

2.2.3. *Moran scatterplot.* If the spatial weighting matrix is row-standardized, we can define the lag vector $\tilde{\mathbf{z}} = \mathbf{W}\mathbf{z}$ (i.e., $\tilde{z}_i = \sum_{j=1}^{n} w_{ij}x_j$) composed of the weighted (by the spatial weighting matrix) averages of the neighboring values.

TABLE 2
Values of Moran's coefficient for the six variables. P-values obtained by a randomization testing procedure (999 permutations) are given in parentheses

	MC
Crime_pers	0.411 (0.001)
Crime_prop	0.264 (0.001)
Literacy	0.718 (0.001)
Donations	0.353 (0.001)
Infants	0.229 (0.001)
Suicides	0.402 (0.001)

Equation (4) can then be rewritten as

$$\mathrm{MC}(\mathbf{x}) = \frac{\mathbf{z}^{\mathrm{T}}\tilde{\mathbf{z}}}{\mathbf{z}^{\mathrm{T}}\mathbf{z}}, \tag{5}$$

since in this case $\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1} = n$. Equation (5) shows clearly that MC measures the autocorrelation by giving an indication of the intensity of the linear association between the vector of observed values $\mathbf{z}$ and the vector of weighted averages of neighboring values $\tilde{\mathbf{z}}$. Anselin (1996) proposed to visualize MC in the form of a bivariate scatterplot of $\tilde{\mathbf{z}}$ against $\mathbf{z}$. A linear regression can be added to this *Moran scatterplot*, with slope equal to MC. The Moran scatterplot is a very nice graphical tool to evaluate and represent the degree of spatial autocorrelation, the presence of outliers or local pockets of nonstationarity [Anselin (1995)].

Considering the Literacy variable of Guerry's data, the Moran scatterplot (Figure 3) clearly shows strong autocorrelation. It also shows that the Hautes-Alpes département has a slightly outlying position characterized by a high value of Literacy compared to its neighbors. This département can be considered as a leverage point that drags down the assessment of the link between Literacy and spatial-lagged literacy (i.e., MC). This is confirmed by different diagnostic tools [DFFITS, Cook's D, e.g., Chatterjee and Hadi (1986)] adapted to the linear model.

2.3. *Toward an integration of multivariate and geographical aspects.* The integration of multivariate and spatial information has a long history in ecology. The simplest approach considered a two-step procedure where the data are first summarized with multivariate analysis such as PCA. In a second step, univariate spatial statistics or mapping techniques are applied to PCA scores for each axis separately. Goodall (1954) was the first to apply multivariate analysis in ecology, and he integrated spatial information a posteriori by mapping PCA scores onto the geographical space using contour lines. One can also test for the presence of spatial autocorrelation for the first few scores of the analysis, with univariate autocorrelation statistics such as MC. For instance, we mapped scores of the départements

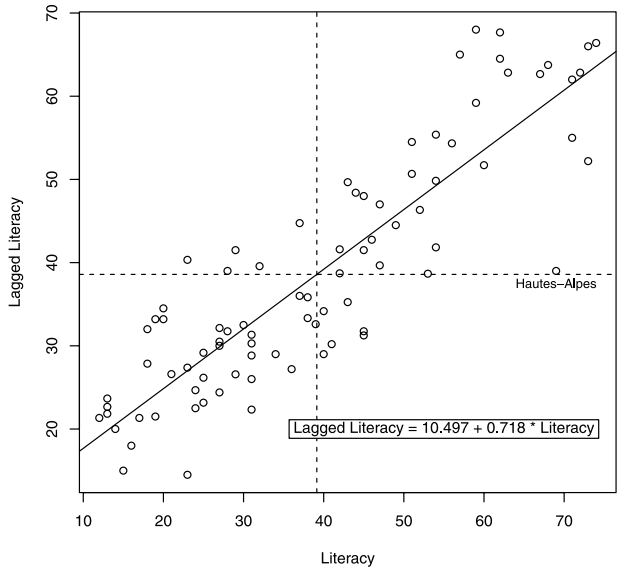

FIG. 3. *Moran scatterplot for Literacy. Dotted lines corresponds to means.*

for the first two axes of the PCA of Guerry's data (Figure 4). Even if PCA maximizes only the variance of these scores, there is also a clear spatial structure, as the scores are highly autocorrelated. The map for the first axis corresponds closely to the split between *la France éclairée* (North-East characterized by an higher level of Literacy) and *la France obscure*.

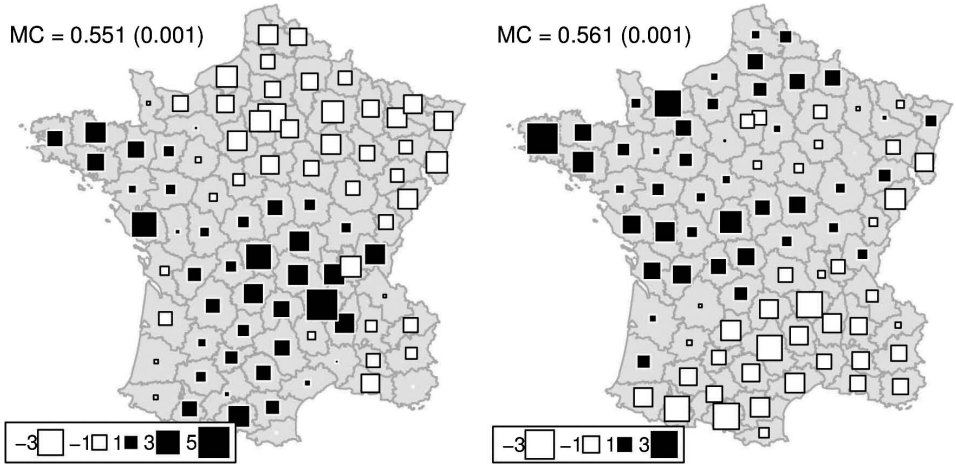

FIG. 4. *Principal component analysis of Guerry's data. Map of départements' scores for the first* (*left*) *and second* (*right*) *PCA axes. Values of Moran's coefficient and associated P-values obtained by a randomization testing procedure* (999 *permutations*) *are given.*

It is very simple to carry out this two-step approach but it has the major disadvantage of being indirect, as it considers the spatial pattern only after summarizing the main structures of the multivariate data set. Anselin, Syabri and Smirnov (2002) proposed a more direct approach by extending the Moran scatterplot to the bivariate case. If we consider two centered variables $\mathbf{z}_1$ and $\mathbf{z}_2$, the bivariate Moran scatterplot represents $\tilde{\mathbf{z}}_2 = \mathbf{W}\mathbf{z}_2$ on the vertical axis and $\mathbf{z}_1$ on the horizontal axis. In a case with more than two variables, one can produce bivariate Moran scatterplots for all combinations of pairs of variables. However, this approach becomes difficult to use when the number of variables increases. In the next section we present several approaches that go one step further by considering the identification of spatial structures and the dimensionality reduction simultaneously.

3. Spatial multivariate analysis. Over the last two decades, several approaches have been developed to consider both geographical and multivariate information simultaneously. The multivariate aspect is usually treated by techniques of dimensionality reduction similar to PCA. On the other hand, several alternatives have been proposed to integrate the spatial information. We review various alternatives in the following sections.

3.1. *Spatial partition*. One alternative is to consider a spatial partition of the study area. In this case, the spatial information is coded as a categorical variable, and each category corresponds to a region of the whole study area. This partitioning can be inherent to the data set (e.g., administrative units) or can be constructed using geographic information systems [e.g., grids of varying cell size in Dray, Pettorelli and Chessel (2003)]. For instance, Guerry's data contained a partition of France into 5 regions (Figure 1).

In this context, searching for multivariate spatial structures would lead us to look for a low-dimensional view that maximizes the difference between the regions. To this end, Friendly (2007) used discriminant analysis, a widely-used method providing linear combinations of variables that maximize the separation between groups as measured by an univariate F statistic. However, this method suffers from some limitations: it requires the number of variables to be smaller than the number of observations, and it is impaired by collinearity among variables. Here we used an alternative and lesser known approach, the between-class analysis [BCA, Dolédec and Chessel (1987)], to investigate differences between regions. Unlike discriminant analysis, BCA maximizes the variance between groups (without accounting for the variance within groups) and is not subject to the restrictions applying to the former method.

BCA associates a triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ to a $n \times g$ matrix $\mathbf{Y}$ of dummy variables indicating group membership. Let $\mathbf{A}$ be the $g \times p$ matrix of group means for the p variables and $\mathbf{D}_\mathbf{Y}$ be the $g \times g$ diagonal matrix of group weights derived from the

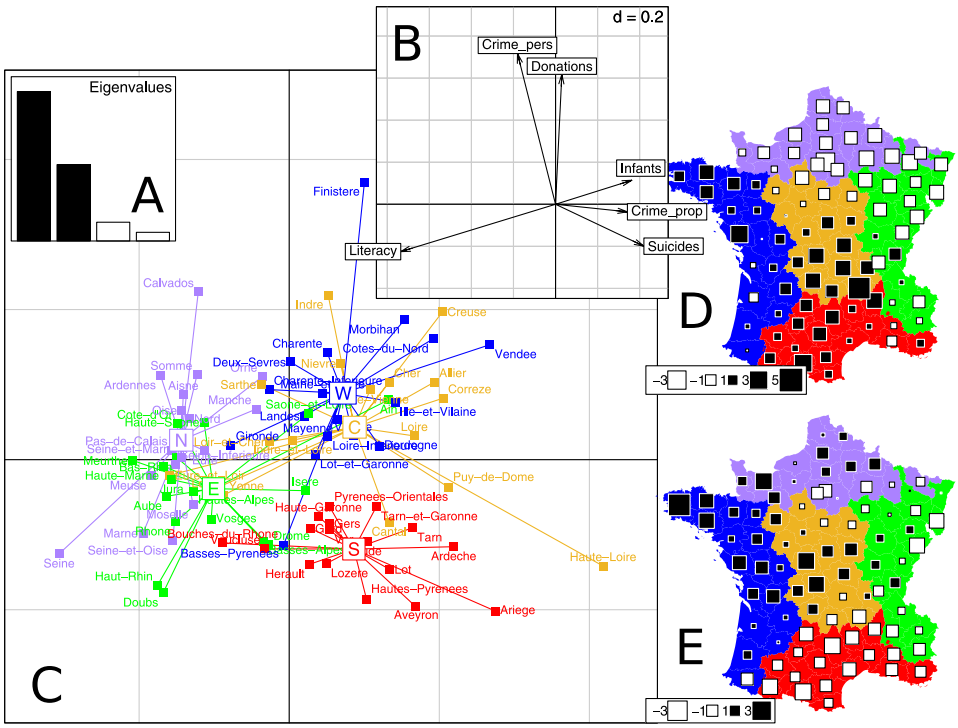

FIG. 5. *Between-class analysis of Guerry's data.* (A) *Barplot of eigenvalues.* (B) *Coefficients of variables.* (C) *Projections of départements on the BCA axes. Map of départements scores for the first* (D) *and second* (E) *axes. The different colors correspond to regions of France.*

matrix $\mathbf{D}$ of observation weights. By definition, we have $\mathbf{A} = (\mathbf{Y}^{\mathrm{T}}\mathbf{D}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{D}\mathbf{X}$ and $\mathbf{D}_{\mathbf{Y}} = (\mathbf{Y}^{\mathrm{T}}\mathbf{D}\mathbf{Y})$. BCA corresponds to the analysis of $(\mathbf{A}, \mathbf{Q}, \mathbf{D}_{\mathbf{Y}})$ and diagonalizes the between-groups covariance matrix $\mathbf{A}^{\mathrm{T}}\mathbf{D}_{\mathbf{Y}}\mathbf{A}\mathbf{Q}$.

Here, 28.8% of the total variance (sum of eigenvalues of PCA) corresponds to the between-regions variance (sum of the eigenvalues of BCA). The barplot of eigenvalues indicates that two axes should be interpreted (Figure 5A). The first two BCA dimensions account for 59% and 30.2%, respectively, of the between-regions variance. The coefficients used to construct the linear combinations of variables are represented on Figure 5B. The first axis opposed literacy to property crime, suicides and illegitimate births. The second axis is mainly aligned with personal crime and donations to the poor. The factorial map of départements (Figure 5C) and the maps of the scores (Figure 5D, E) show the spatial aspects. The results are very close to those obtained by PCA: the first axis contrasted the North and the East (*la France éclairée*) to the other regions, while the South is separated from the other regions by the second axis. The high variability of the region Center is also noticeable. In contrast, the South is very homogeneous.

3.2. *Spatial explanatory variables*. Principal component analysis with respect to the instrumental variables [PCAIV, Rao (1964)], also known as redundancy analysis [van den Wollenberg (1977)], is a direct extension of PCA and multiple regression adapted to the case of multivariate response data. The analysis associates a $n \times q$ matrix $\mathbf{Z}$ of explanatory variables to the triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ where the matrix $\mathbf{X}$ contains the response variables. The $\mathbf{D}$-orthogonal projector $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}^T\mathbf{D}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{D}$ is first used in a multivariate regression step to compute a matrix of predicted values $\hat{\mathbf{X}} = \mathbf{P_Z}\mathbf{X}$. The second step of PCAIV consists of the PCA of this matrix of predicted values and corresponds then to the analysis of the triplet $(\hat{\mathbf{X}}, \mathbf{Q}, \mathbf{D})$. Whereas PCA maximizes the variance of the projection of the observations onto the principal axes, PCAIV maximizes the variance explained by $\mathbf{Z}$.

PCAIV and related methods, such as canonical correspondence analysis [ter Braak (1986)], have been often used in community ecology to identify spatial relationships. The spatial information is introduced in the matrix $\mathbf{Z}$ under the form of spatial predictors and the analysis maximized then the "spatial variance" (i.e., the variance explained by spatial predictors). Note that BCA can also be considered as a particular case of PCAIV, where the explanatory variables are dummy variables indicating group membership.

3.2.1. *Trend surface of geographic coordinates*. From the EDA point of view, the data exploration has been conceptualized by Tukey (1977) in the quasi-mathematical form $DATA = SMOOTH + ROUGH$. Trend surface analysis is the oldest procedure for separating large-scale structure (*SMOOTH*) from random variation (*ROUGH*). Student (1914) proposed expressing observed values in time series as a polynomial function of time, and mentioned that this could be done for spatial data as well. Borcard, Legendre and Drapeau (1992) extended this approach to the spatial and multivariate case by introducing polynomial functions of geographic coordinates as predictors in PCAIV. We call this approach PCAIV-POLY in the rest of the paper. Usually, polynomials of degree 2 or 3 are used; spurious correlations between these spatial predictors can be removed using an orthogonalization procedure to obtain orthogonal polynomials.

The centroids of départements of France were used to construct a second-degree orthogonal polynomial (Figure 6).

Here, 32.4% of the total variance (sum of eigenvalues of PCA) is explained by the second-degree polynomial (sum of eigenvalues of PCAIV). The first two dimensions account for 51.4% and 35.2%, respectively, of the explained variance. The outputs of PCAIV-POLY (coefficients of variables, maps of départements scores, etc.) are not presented, as they are very similar to those obtained by BCA.

3.2.2. *Moran's eigenvector maps*. An alternative way to build spatial predictors is by the diagonalization of the spatial weighting matrix $\mathbf{W}$. de Jong, Sprenger

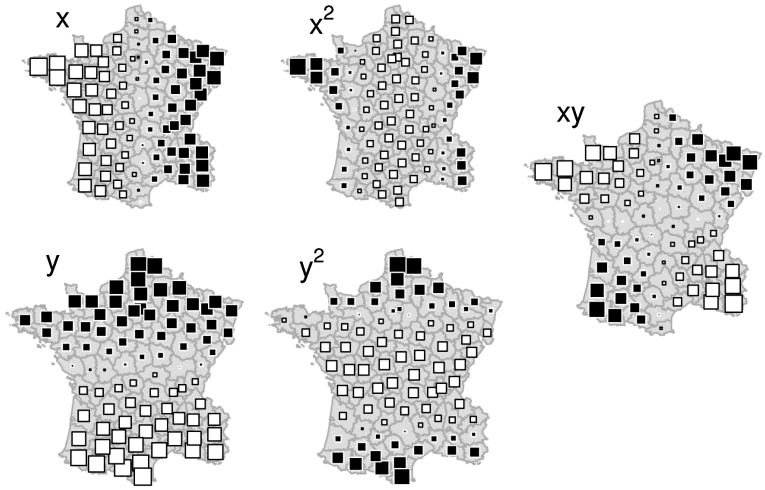

FIG. 6. *Maps of the terms of a second-degree orthogonal polynomial. Centroids of départements have been used as original coordinates to construct the polynomial.*

and van Veen (1984) have shown that the upper and lower bounds of MC for a given spatial weighting matrix $\mathbf{W}$ are equal to $\lambda_{\max}(n/\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1})$ and $\lambda_{\min}(n/\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1})$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the extreme eigenvalues of $\mathbf{\Omega} = \mathbf{HWH}$ where $\mathbf{H} = (\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n)$ is a centering operator. If a nonsymmetric spatial weighting matrix $\mathbf{W}^*$ has been defined, the results can be generalized using $\mathbf{W} = (\mathbf{W}^* + \mathbf{W}^{*\mathrm{T}})/2$.

Moran's eigenvector maps [MEM, Dray, Legendre and Peres-Neto (2006)] are the $n - 1$ eigenvectors of $\mathbf{\Omega}$. They are orthogonal vectors with a unit norm maximizing MC [Griffith (1996)]. MEM associated with high positive (or negative) eigenvalues have high positive (or negative) autocorrelation. MEM associated with eigenvalues with small absolute values correspond to low spatial autocorrelation, and are not suitable for defining spatial structures [Dray, Legendre and Peres-Neto (2006)]. Unlike polynomial functions, MEM have the ability to capture various spatial structures at multiple scales (coarse to fine scales). MEM have been used for spatial filtering purposes [Griffith (2003); Getis and Griffith (2002)] and introduced as spatial predictors in linear models [Griffith (1996, 2000)], generalized linear models [Griffith (2002, 2004)] and multivariate analysis [Dray, Legendre and Peres-Neto (2006); Jombart, Dray and Dufour (2009)].

We used the spatial weighting matrix associated to the neighborhood graph presented on Figure 2 to construct MEM. The first ten MEM, corresponding to the highest levels of spatial autocorrelation, have been mapped in Figure 7 and introduced as spatial explanatory variables in PCAIV. We call this approach PCAIV-MEM in the rest of the paper. 44.1% of the total variance (sum of eigenvalues of PCA) is explained by the first ten MEM (sum of eigenvalues of PCAIV). The first two dimensions account for 54.9% and 26.3%, respectively, of the explained variance. The outputs of PCAIV-MEM (coefficients of variables, maps of département

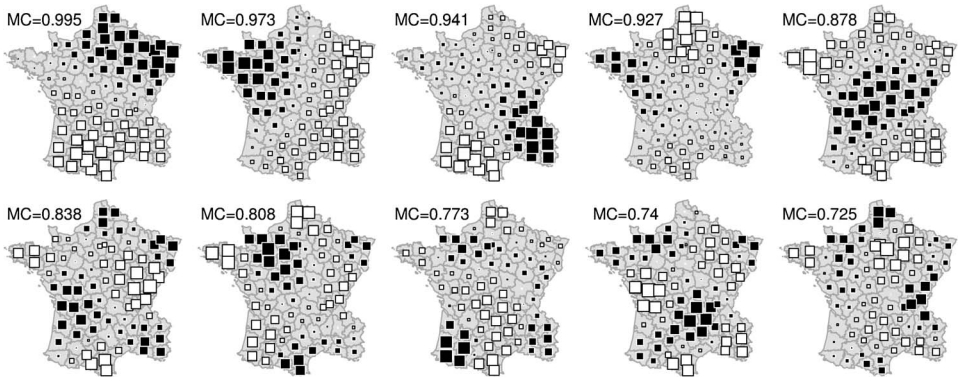

FIG. 7. *Maps of the first ten MEM of the spatial weighting matrix associated to the neighborhood graph presented on Figure 2. By definition, MEM are orthogonal vectors maximizing the values of Moran's coefficient.*

scores, etc.) are not presented, as they are very similar to those obtained by the previous analyses.

3.3. *Spatial graph and weighting matrix.* The MEM framework introduced the spatial information into multivariate analysis through the eigen-decomposition of the spatial weighting matrix. Usually, we consider only a part of the information contained in this matrix because only a subset of MEM are used as regressors in PCAIV. In this section we focus on multivariate methods that consider the spatial weighting matrix under its original form.

Lebart (1969) was the first to introduce a neighborhood graph into a multivariate analysis. Following this initial work, many methods have been mainly developed by the French school of statisticians [Le Foll (1982); Benali and Escofier (1990); Méot, Chessel and Sabatier (1993)]. These contributions were important from a methodological point of view, but have been rarely used for applied problems. Indeed, they have a major drawback in their objectives: they maximize the local variance (i.e., the difference between neighbors), while users more often want to minimize this quantity and maximize the spatial correlation (i.e., the *SMOOTH*).

Wartenberg (1985) was the first to develop a multivariate analysis based on MC. His work considered only normed and centered variables (i.e., normed PCA) for the multivariate part and a binary symmetric connectivity matrix for the spatial aspect. Dray, Saïd and Débias (2008) generalized Wartenberg's method by introducing a row-standardized spatial weighting matrix in the analysis of a statistical triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. Hence, this approach is very general and allows us to define spatially-constrained versions of various methods (corresponding to different triplets) such as correspondence analysis or multiple correspondence analysis.

By extension of the lag vector, a lag matrix $\tilde{\mathbf{X}} = \mathbf{WX}$ can be defined. The two tables $\tilde{\mathbf{X}}$ and $\mathbf{X}$ are fully matched, that is, they have the same columns (variables) and

rows (observations). MULTISPATI (Multivariate spatial analysis based on Moran's index) aims to identify multivariate spatial structures by studying the link between $\tilde{\mathbf{X}}$ and $\mathbf{X}$ using the coinertia analysis [Dolédec and Chessel (1994); Dray, Chessel and Thioulouse (2003a)] of a pair of fully matched tables [Torre and Chessel (1995); Dray, Chessel and Thioulouse (2003b)]. It corresponds to the analysis of the statistical triplet $(\mathbf{X}, \mathbf{Q}, \frac{1}{2}(\mathbf{W}^{\mathrm{T}}\mathbf{D} + \mathbf{DW}))$. The objective of the analysis is to find a vector $\mathbf{a}$ (with $\|\mathbf{a}\|_{\mathbf{Q}}^2$) maximizing the quantity defined in equation (1):

$$\begin{aligned} Q(\mathbf{a}) &= \mathbf{a}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\tfrac{1}{2}(\mathbf{W}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}} + \mathbf{DW})\mathbf{XQa} \\ &= \tfrac{1}{2}(\mathbf{a}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{XQa} + \mathbf{a}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{DWXQa}) \\ &= \tfrac{1}{2}\langle \mathbf{XQa}, \mathbf{WXQa}\rangle_{\mathbf{D}} + \langle \mathbf{WXQa}, \mathbf{XQa}\rangle_{\mathbf{D}} \\ &= \mathbf{a}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{DWXQa} = \mathbf{r}^{\mathrm{T}}\mathbf{DWr} = \mathbf{r}^{\mathrm{T}}\mathbf{D}\tilde{\mathbf{r}}. \end{aligned} \tag{6}$$

This analysis maximizes the scalar product between a linear combination of original variables ($\mathbf{r} = \mathbf{XQa}$) and a linear combination of lagged variables ($\tilde{\mathbf{r}} = \mathbf{WXQa}$). Equation (6) can be rewritten as

$$\begin{aligned} Q(\mathbf{a}) &= \frac{\mathbf{a}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{DWXQa}}{\mathbf{a}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{DXQa}}\mathbf{a}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{DXQa} \\ &= \mathrm{MC}_{\mathbf{D}}(\mathbf{XQa}) \cdot \|\mathbf{XQa}\|_{\mathbf{D}}^2 = \mathrm{MC}_{\mathbf{D}}(\mathbf{r}) \cdot \|\mathbf{r}\|_{\mathbf{D}}^2. \end{aligned} \tag{7}$$

MULTISPATI finds coefficients ($\mathbf{a}$) to obtain a linear combination of variables ($\mathbf{r} = \mathbf{XQa}$) that maximizes a compromise between the classical multivariate analysis ($\|\mathbf{r}\|_{\mathbf{D}}^2$) and a generalized version of Moran's coefficient [$\mathrm{MC}_{\mathbf{D}}(\mathbf{r})$]. The only difference between the classical Moran's coefficient [equation (4)] and its generalized version $\mathrm{MC}_{\mathbf{D}}$ is that the second one used a general matrix of weights $\mathbf{D}$, while the first considers only the usual case of uniform weights ($\mathbf{D} = \frac{1}{n}\mathbf{I}_n$).

In practice, the maximum of equation (7) is obtained for $\mathbf{a} = \mathbf{a}^1$, where $\mathbf{a}^1$ is the first eigenvector of the $\mathbf{Q}$-symmetric matrix $\frac{1}{2}\mathbf{X}^{\mathrm{T}}(\mathbf{W}^{\mathrm{T}}\mathbf{D} + \mathbf{DW})\mathbf{Q}$. This maximal value is equal to the associated eigenvalue λ_1. Further eigenvectors maximize the same quantity with the additional constraint of orthogonality.

MULTISPATI has been applied to Guerry's data (Figure 8). The barplot of eigenvalues (Figure 8A) suggests two main spatial structures. The coefficients used to construct the linear combinations of variables are represented in Figure 8B. The first axis opposes literacy to property crime, suicides and illegitimate births. The second axis is aligned mainly with personal crime and donations to the poor. The maps of the scores (Figure 8C, E) show that the spatial structures are very close to those identified by PCA. The similarity of results between PCA and its spatially optimized version confirm that the main structures of Guerry's data are spatial.

MULTISPATI maximizes the product between the variance and the spatial autocorrelation of the scores, while PCA (Figure 1) maximizes only the variance.

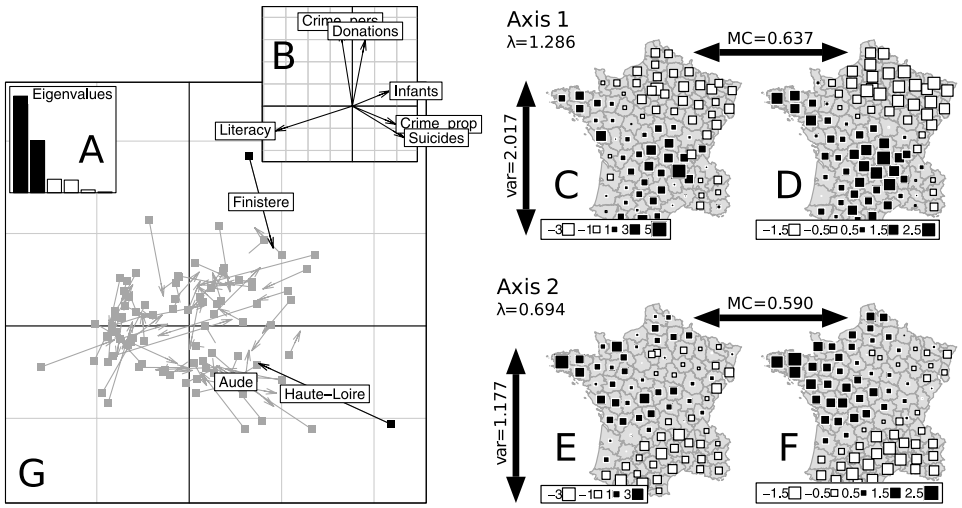

FIG. 8. *MULTISPATI of Guerry's data*. (A) *Barplot of eigenvalues*. (B) *Coefficients of variables*. *Mapping of scores of plots on the first* (C) *and second* (E) *axis and of lagged scores* (*averages of neighbors weighted by the spatial connection matrix*) *for the first* (D) *and second* (F) *axis*. *Representation of scores and lagged scores* (G) *of plots* (*for each département*, *the arrow links the score to the lagged score*). *Only the départements discussed in the text are indicated by their labels*.

Hence, there is a loss of variance compared to PCA (2.14 versus 2.017 for axis 1; 1.201 versus 1.177 for axis 2) but a gain of spatial autocorrelation (0.551 versus 0.637 for axis 1; 0.561 versus 0.59 for axis 2).

Spatial autocorrelation can be seen as the link between one variable and the lagged vector [equation (5)]. This interpretation is used to construct the Moran scatterplot and can be extended to the multivariate case in MULTISPATI by analyzing the link between scores (Figure 8C, E) and lagged scores (Figure 8D, F). Each département can be represented on the factorial map by an arrow (the bottom corresponds to its score, the head corresponds to its lagged score, Figure 8G). A short arrow reveals a local spatial similarity (between one plot and its neighbors), while a long arrow reveals a spatial discrepancy. This viewpoint can be interpreted as a multivariate extension of the local index of spatial association [Anselin (1995)]. For instance, Aude has a very small arrow, indicating that this département is very similar to its neighbors. On the other hand, the arrow for Haute-Loire has a long horizontal length which reflects its high values for the variables Infants (31017), Suicides (163241) and Crime_prop (18043) compared to the average values over its neighbors (27032.4, 60097.8 and 10540.8 for these three variables). Finistère corresponds to an arrow with a long vertical length which is due to its high values compared to its neighbors for Donations (23945 versus 12563) and Crime_pers (29872 versus 25962).

4. Conclusions. We have presented different ways of incorporating the spatial information in multivariate analysis methods. While PCA is not constrained, spatial information can be introduced as a partition (BCA), a polynomial of geographic coordinates (PCAIV-POLY), a subset of Moran's eigenvector maps (PCAIV-MEM) or a spatial neighborhood graph (MULTISPATI). This variety of constraints induces a diversity of criteria to be maximized by each method: variance (PCA), variance explained by a spatial partition (BCA) or by spatial predictors (PCAIV-POLY, PCAIV-MEM), product of the variance by the spatial autocorrelation (MULTISPATI). By presenting these methods in the duality diagram framework, we have shown that these approaches are very general, and can be applied to virtually any multivariate analysis.

These theoretical considerations have practical implications concerning the use of these methods in applied studies. PCA is a very general method allowing one to identify the main spatial and nonspatial structures. BCA maximally separates the groups corresponding to a spatial partition. It is thus adapted when a study focuses on spatial structures induced by a partitioning defined a priori (e.g., administrative units, etc.). If such an a priori partitioning does not exist, one can easily define such a partition albeit introducing some element of subjectivity in the consideration of the spatial information. This problem is solved by PCAIV-POLY, which uses polynomials to incorporate the spatial information. Polynomials are easily constructed, but their use is only satisfactory when the sampling area is roughly homogeneous and the sampling design is nearly regular [Norcliffe (1969)]. Other limitations to their use have been reported in the literature such as the arbitrary choice of the degree and their ability to account only for smooth broad-scale spatial patterns [Dray, Legendre and Peres-Neto (2006)].

The use of graphs and spatial weighting matrices allows the construction of more efficient and flexible representations of space. Binary spatial weighting matrices can be constructed using distance criteria or tools derived from graph theory [Jaromczyk and Toussaint (1992)]; they may also describe spatial discontinuities, boundaries or physical barriers in the landscape. Spatial weights can be associated to the binary links to represent the spatial heterogeneity of the landscape using functions of geographic distances or least-cost links between sampling locations [Fall et al. (2007)] or any other proxies/measures of the potential strength of connection between the locations. MEM are obtained by the eigen-decomposition of the spatial weighting matrix $\mathbf{W}$. For a data set with n observations, this eigen-decomposition produces $n - 1$ MEM. Hence, a subset of these spatial predictors must be selected to avoid overfitting in the multivariate regression step of PCAIV. Concerning Guerry's data set, we choose the first ten MEM arbitrarily. Other objective selection procedures have been proposed in the literature. For instance, the criteria can be based on the minimization of the autocorrelation in residuals [Tiefelsdorf and Griffith (2007)] or on the maximization of the fit of the model [Blanchet, Legendre and Borcard (2008)]. Hence, only a part of the spatial information contained in $\mathbf{W}$ (corresponding to the subset of MEM retained by the selection procedure) is considered in PCAIV. In MULTISPATI, the spatial weighting

TABLE 3
Procrustes statistics measuring the concordance between the scores of the départements on the first two axes of the different analyses. A value of 1 *indicates a perfect match between two configurations of département scores. Randomization procedures with* 999 *permutations have been used to test the significance of the concordance. All the statistics are significant* ($p = 0.001$)

	PCA	**BCA**	**PCAIV-POLY**	**PCAIV-MEM**
BCA	0.979			
PCAIV-POLY	0.979	0.990		
PCAIV-MEM	0.989	0.994	0.995	
MULTISPATI	0.987	0.995	0.995	0.999

matrix is used in its original form, so that the whole spatial information contained in it is taken into account in the multivariate analysis.

Even if the methods presented are quite different in their theoretical and practical viewpoints, their applications to Guerry's data set yield very similar results. We provided a quantitative measure of this similarity by computing Procrustes statistics [Peres-Neto and Jackson (2001); Dray, Chessel and Thioulouse (2003b)] between the scores of the départements on the first two axes for the different analyses (Table 3). All the values of the statistics are very high and significant; this confirms the high concordance between the outputs of the different methods. This similarity of results is due to the very clear structures of the data set and to the high level of autocorrelation of these structures (Figure 4). In this example the main advantage of the spatially-constrained methods is in the choice of the number of dimensions to interpret; while the barplot of eigenvalues of PCA can be difficult to interpret (see above and Figure 1A), it is clear that two spatial dimensions must be interpreted in BCA (Figure 5A) or MULTISPATI (Figure 8A).

In the case of Guerry's data, the very simple and clear-cut structures seem to be recovered by all the approaches presented here. In more complex data sets, spatially constrained methods prove superior to standard approaches for detecting spatial multivariate patterns. Dray, Saïd and Débias (2008) presented an example where a standard multivariate method was unable to identify any structure and is outperformed by MULTISPATI, which allows us to discover interesting spatial patterns. In general, if the objective of a study is to detect spatial patterns, it would be preferable to use a spatially-constrained method. PCA could also be useful, but it is designed to identify the main structures that can or cannot be spatialized. On the other hand, spatial multivariate methods are optimized to focus on the spatial aspect and would generally produce clearer and smoother results. The outputs and interpretation tools of these methods are also more adapted to visualizing and quantifying the main multivariate spatial structures.

From a methodological viewpoint, these approaches provide new ways of taking into account the complexity of sampling designs in the framework of multivariate methods. Following the famous paper of Legendre (1993), the analysis of spatial structures has been a major issue in community ecology and originated several methodological developments in the field of spatial multivariate analysis. To date, the most integrated and flexible approaches have used a spatial weighting matrix which can be seen as a general way to consider spatial proximities. Potential methodological perspectives are important, as these approaches could easily be extended to any other sampling constraints that can be expressed in the form of a matrix of similarities between the observations.

Acknowledgments. We would like to warmly thank Michael Friendly for freely distributing Guerry's data set and for providing constructive comments on an earlier version of the manuscript. We thank Susan Holmes for her invitation to participate in this special issue.

SUPPLEMENTARY MATERIAL

Implementation in R (DOI: 10.1214/10-AOAS356SUPP; .zip). This website hosts an R package (`Guerry`) containing the Guerry's data set (maps and data). The package contains also a tutorial (vignette) showing how to reproduce the analyses and the graphics presented in this paper using mainly the package `ade4` [Dray and Dufour (2007)]. The package `Guerry` is also available on CRAN and can be installed using the `install.packages("Guerry")` command in a R session.

REFERENCES

ANSELIN, L. (1995). Local indicators of spatial association. *Geographical Analysis* **27** 93–115.

ANSELIN, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In *Spatial Analytical Perspectives on GIS* (M. M. Fischer, H. J. Scholten and D. Unwin, eds.) 111–125. Taylor and Francis, London.

ANSELIN, L., SYABRI, I. and SMIRNOV, O. (2002). Visualizing multivariate spatial correlation with dynamically linked windows. In *New Tools for Spatial Data Analysis: Proceedings of a Workshop* (L. Anselin and S. Rey, eds.). CSISS, Santa-Barbara, CA.

BENALI, H. and ESCOFIER, B. (1990). Analyse factorielle lissée et analyse factorielle des différences locales. *Rev. Statist. Appl.* **38** 55–76.

BIVAND, R. (2008). Implementing representations of space in economic geography. *Journal of Regional Science* **48** 1–27.

BLANCHET, F. G., LEGENDRE, P. and BORCARD, D. (2008). Forward selection of explanatory variables. *Ecology* **89** 2623–2632.

BORCARD, D., LEGENDRE, P. and DRAPEAU, P. (1992). Partialling out the spatial component of ecological variation. *Ecology* **73** 1045–1055.

CHATTERJEE, S. and HADI, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statist. Sci.* **1** 379–393. MR0858516

CLIFF, A. D. and ORD, J. K. (1973). *Spatial Autocorrelation*. Pion, London.

DE JONG, P., SPRENGER, C. and VAN VEEN, F. (1984). On extreme values of Moran's I and Geary's c. *Geographical Analysis* **16** 17–24.

DOLÉDEC, S. and CHESSEL, D. (1987). Rythmes saisonniers et composantes stationnelles en milieu aquatique I—Description d'un plan d'observations complet par projection de variables. *Acta Oecologica—Oecologia Generalis* **8** 403–426.

DOLÉDEC, S. and CHESSEL, D. (1994). Co-inertia analysis: An alternative method for studying species-environment relationships. *Freshwater Biology* **31** 277–294.

DRAY, S. and JOMBORT, T. (2010). Suplement to "Revisiting Guerry's data: Introducing spatial constraints in multivariate analysis." DOI:10.1214/10-AOAS356SUPP.

DRAY, S., CHESSEL, D. and THIOULOUSE, J. (2003a). Co-inertia analysis and the linking of ecological data tables. *Ecology* **84** 3078–3089.

DRAY, S., CHESSEL, D. and THIOULOUSE, J. (2003b). Procrustean co-inertia analysis for the linking of multivariate data sets. *Ecoscience* **10** 110–119.

DRAY, S. and DUFOUR, A. B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *J. Statist. Soft.* **22** 1–20.

DRAY, S., LEGENDRE, P. and PERES-NETO, P. R. (2006). Spatial modeling: A comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). *Ecological Modelling* **196** 483–493.

DRAY, S., PETTORELLI, N. and CHESSEL, D. (2003). Multivariate analysis of incomplete mapped data. *Transactions in GIS* **7** 411–422.

DRAY, S., SAÏD, S. and DÉBIAS, F. (2008). Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science* **19** 45–56.

DYKES, J. and BRUNSDON, C. (2007). Geographically weighted visualization: Interactive graphics for scale-varying exploratory analysis. *IEEE Transactions on Visualization and Computer Graphics* **13** 1161–1168.

ESCOUFIER, Y. (1987). The duality diagram: A means of better practical applications. In *Developments in Numerical Ecology* (P. Legendre and L. Legendre, eds.) **14** 139–156. Springer, Berlin. MR0913539

FALL, A., FORTIN, M. J., MANSEAU, M. and O'BRIEN, D. (2007). Spatial graphs: Principles and applications for habitat connectivity. *Ecosystems* **10** 448–461.

FRIENDLY, M. (2007). A.-M. Guerry's moral statistics of France: Challenges for multivariable spatial analysis. *Statist. Sci.* **22** 368–399. MR2399897

GABRIEL, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58** 453–467. MR0312645

GEARY, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician* **5** 115–145.

GETIS, A. and ALDSTADT, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis* **36** 90–104.

GETIS, A. and GRIFFITH, D. A. (2002). Comparative spatial filtering in regression analysis. *Geographical Analysis* **34** 130–140.

GOODALL, D. W. (1954). Objective methods for the classification of vegetation III. An essay on the use of factor analysis. *Australian Journal of Botany* **2** 304–324.

GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London. MR0767260

GRIFFITH, D. A. (1996). Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Canadian Geographer* **40** 351–367.

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

GRIFFITH, D. A. (2002). A spatial filtering specification for the auto-Poisson model. *Statist. Probab. Lett.* **58** 245–251. MR1920751

GRIFFITH, D. A. (2003). *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization*. Springer, Berlin.

GRIFFITH, D. A. (2004). A spatial filtering specification for the autologistic model. *Environment and Planning A* **36** 1791–1811.

GUÉRRY, A. M. (1833). *Essai sur la Statistique Morale de la France*. Crochard, Paris.

HAINING, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge Univ. Press.

HOLMES, S. (2006). Multivariate analysis: The French way. In *Festschrift for David Freedman* (D. Nolan and T. Speed, eds.). IMS, Beachwood, OH.

HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24** 417–441.

JAROMCZYK, J. W. and TOUSSAINT, G. T. (1992). Relative neighborhood graphs and their relatives. *Proceedings of the IEEE* **80** 1502–1517.

JOMBART, T., DRAY, S. and DUFOUR, A. B. (2009). Finding essential scales of spatial variation in ecological data: A multivariate approach. *Ecography* **32** 161–168.

LE FOLL, Y. (1982). Pondération des distances en analyse factorielle. *Statistique et Analyse des données* **7** 13–31.

LEBART, L. (1969). Analyse statistique de la contiguïté. *Publication de l'Institut de Statistiques de l'Université de Paris* **28** 81–112.

LEGENDRE, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology* **74** 1659–1673.

LEGENDRE, P. and LEGENDRE, L. (1998). *Numerical Ecology*, 2nd ed. Elsevier, Amsterdam.

MORAN, P. A. P. (1948). The interpretation of statistical maps. *J. Roy. Statist. Soc. Ser. B* **10** 243–251. MR0029115

MÉOT, A., CHESSEL, D. and SABATIER, R. (1993). Opérateurs de voisinage et analyse des données spatio-temporelles. In *Biométrie et environnement* (J. D. Lebreton and B. Asselain, eds.) 45–72. Masson, Paris.

NORCLIFFE, G. B. (1969). On the use and limitations of trend surface models. *Canadian Geographer* **13** 338–348.

PERES-NETO, P. R. and JACKSON, D. A. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia* **129** 169–178.

RAO, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā Ser. A* **26** 329–359. MR0184375

STUDENT (1914). The elimination of spurious correlation due to position in time or space. *Biometrika* **10** 179–180.

TER BRAAK, C. J. F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67** 1167–1179.

TIEFELSDORF, M., GRIFFITH, D. A. and BOOTS, B. (1999). A variance-stabilizing coding scheme for spatial link matrices. *Environment and Planning A* **31** 165–180.

TIEFELSDORF, M. and GRIFFITH, D. A. (2007). Semi-parametric filtering of spatial autocorrelation: The eigenvector approach. *Environment and Planning A* **39** 1193–1221.

TORRE, F. and CHESSEL, D. (1995). Co-structure de deux tableaux totalement appariés. *Revue de Statistique Appliquée* **43** 109–121.

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

VAN DEN WOLLENBERG, A. L. (1977). Redundancy analysis, an alternative for canonical analysis. *Psychometrika* **42** 207–219.

WARTENBERG, D. (1985). Multivariate spatial correlation: A method for exploratory geographical analysis. *Geographical Analysis* **17** 263–283.

UMR 5558
LABORATOIRE DE BIOMÉTRIE ET BIOLOGIE EVOLUTIVE
UNIVERSITÉ DE LYON; UNIVERSITÉ LYON 1; CNRS
43 BOULEVARD DU 11 NOVEMBRE 1918
VILLEURBANNE F-69622
FRANCE
E-MAIL: dray@biomserv.univ-lyon1.fr
URL: http://pbil.univ-lyon1.fr/members/dray

MRC CENTRE FOR OUTBREAK ANALYSIS & MODELLING
DEPARTMENT OF INFECTIOUS DISEASE EPIDEMIOLOGY
FACULTY OF MEDICINE
IMPERIAL COLLEGE LONDON
NORFOLK PLACE
LONDON W2 1PG
UNITED KINGDOM
E-MAIL: t.jombart@imperial.ac.uk