Spatial multivariate analysis of Guerry's data in R

Stéphane Dray

Université de Lyon ; université Lyon 1 ; CNRS
UMR 5558
Laboratoire de Biométrie et Biologie Evolutive
43 boulevard du 11 novembre 1918
Villeurbanne F-69622, France
`dray@biomserv.univ-lyon1.fr`
`http://pbil.univ-lyon1.fr/members/dray`

December 1, 2009

Abstract

This vignette indicates how to perform the analyses described in Dray and Jombart (submitted) of data in the `Guerry` package, derived from André-Michel Guerry's (1833) *Essai sur la Statistique Morale de la France*, using R. It demonstrates some classical methods for analysis of multivariate spatial data that focus *either* on the multivariate aspect or on the spatial one, as well as some more modern methods that attempt to integrate geographical and multivariate aspects *simultaneously*.

Contents

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 dataset. Dray and Jombart (submitted) reanalyzed Guerry's data using classical tools that focus on either the multivariate or spatial aspect of the data, as well as a variety of modern tools of spatial multivariate analysis that integrate both aspects. Here, we show how these analyses can be carried out in ℝ. Commands are written in red and outputs are written in blue.

1.1 Installation and loading of required packages

Several packages must be installed to run the different analyses:

```
pkg <- c("maptools", "spdep", "ade4", "Guerry", "spacemakeR")
inst.pkg <- row.names(installed.packages())
pkg2inst <- pmatch(pkg, inst.pkg)
if (any(is.na(pkg2inst[1:4]))) install.packages(pkg[which(is.na(pkg2inst[1:4]))],
    repos = "http://cran.at.r-project.org")
if (is.na(pkg2inst[5])) install.packages("spacemakeR", repos = "http://R-Forge.R-project.org")
library(maptools)
library(ade4)
library(spdep)
library(spacemakeR)
library(Guerry)
```

1.2 Preliminary steps

We use the dataset `gfrance85`. 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).

```
data(gfrance85)
df <- data.frame(gfrance85)[, 7:12]
xy <- coordinates(gfrance85)
dep.names <- data.frame(gfrance85)[, 6]
region.names <- data.frame(gfrance85)[, 5]
col.region <- colors()[c(149, 254, 468, 552, 26)]
```

2 Standard approaches

In this section, we focus on classical approaches that consider either the multivariate or the spatial aspect of the data.

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.

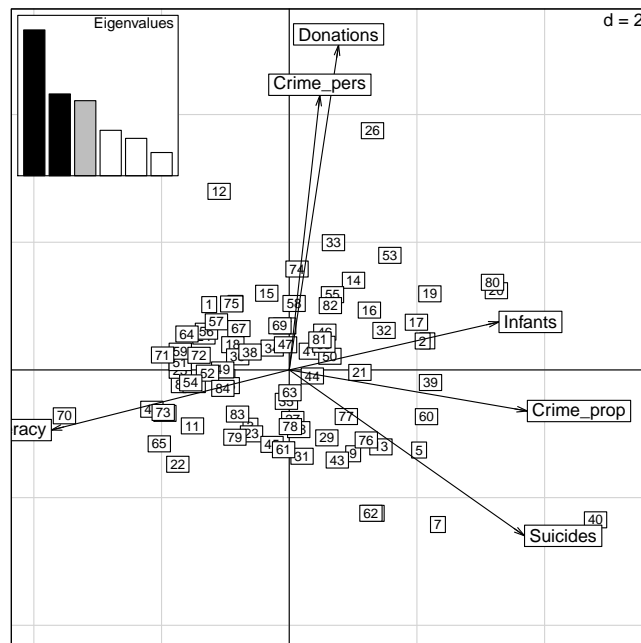| Label	Description
Crime_pers	Population per crime against persons
Crime_prop	Population per crime against property
Literacy	Percent of military conscripts who can read and write
Donations	Donations to the poor
Infants	Population per illegitimate birth
Suicides	Population per suicide

2.1 Multivariate analysis

Here we consider $p = 6$ variables measured for $n = 85$ individuals (départements of France). As only quantitative variables have been recorded, principal component analysis (PCA, Hotelling, 1933) is well adapted. PCA summarizes the data by maximizing simultaneously the variance of the projection of the individuals onto the principal axes and the sum of the squared correlations between the principal component and the variables.

```
pca <- dudi.pca(df, scannf = FALSE, nf = 3)
```

The biplot is simply obtained by:
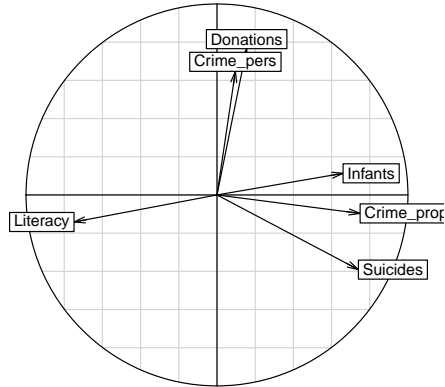
```
scatter(pca)
```

The first two PCA dimensions account for 35.7 % and 20 % ,respectively, of the total variance.

```
pca$eig/sum(pca$eig) * 100
```

```
[1] 35.675 20.014 18.367 11.116  9.145  5.684
```
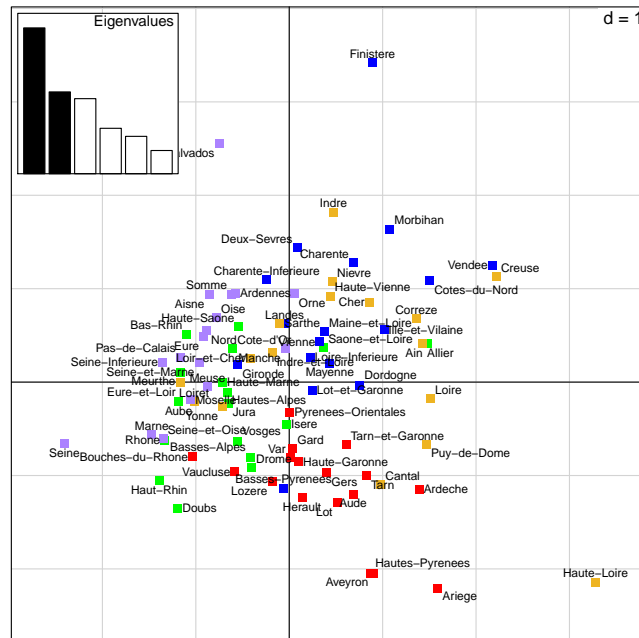
Correlations between variables and principal components can be represented on a correlation circle. The first axis is negatively correlated to literacy and positively correlated to property crime, suicides and illegitimate births. The second axis is aligned mainly with personal crime and donations to the poor.

```
s.corcircle(pca$co, clabel = 0.8)
```

We can add spatial information on the map representing the projections of départements on principal axes (with the barplot of eigenvalues):

```r
s.class(pca$l1, fac = region.names, col = col.region, cellipse = 0,
    cstar = 0, clab = 0, cpoint = 0)
par(mar = rep(0.1, 4))
points(pca$l1, col = col.region[region.names], pch = 15)
pointLabel(pca$l1[, 1:2], as.character(dep.names), cex = 0.7)
add.scatter.eig(pca$eig, xax = 1, yax = 2, posi = "topleft", ratio = 0.25)
```
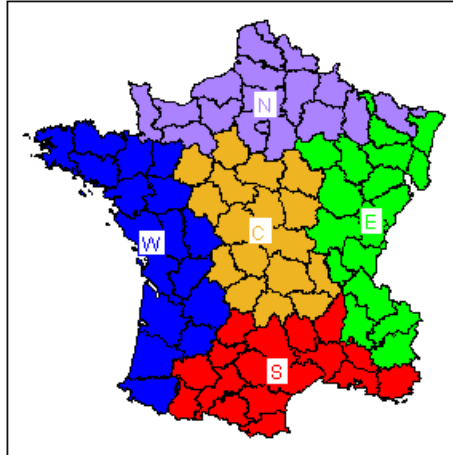

The colors represent the different regions of France.

```
par(mar = rep(0.1, 4))
plot(gfrance85, col = col.region[region.names])
s.class(xy, region.names, cellipse = 0, cstar = 0, add.plot = T, cpoint = 0,
    col = col.region)
```


For the first axis, the North and East are characterized by negative scores, corresponding to high levels of literacy and high rates 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

Spatial autocorrelation statistics, such as Moran (1948) Coefficient (MC) and 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 spatial weighting matrix $\mathbf{W} = [w_{ij}]$. In the case of Guerry's data, we simply defined a binary neighborhood where two départements are considered as neighbors if they share a common border. The spatial weighting matrix is then obtained after row-standardization (`style = "W"`):

```
nb <- poly2nb(gfrance85)
lw <- nb2listw(nb, style = "W")
```
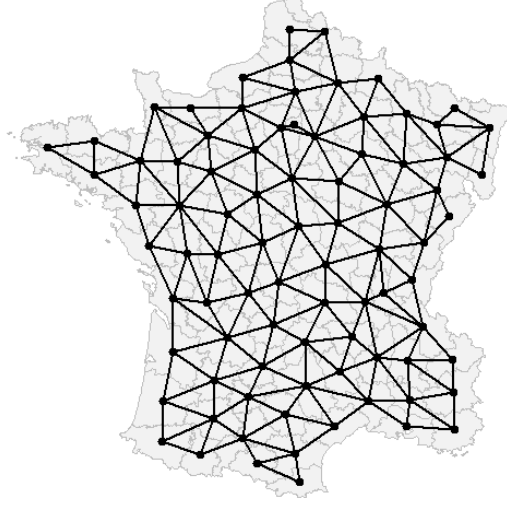
We can represent this neighborhood on the geographical map:

```
png(file = "figs/fig-fig2.png", width = 7, height = 7, units = "in",
    res = 72)
par(mar = rep(0.1, 4))
plot(gfrance85, col = "grey95", border = "grey")
plot(lw, coordinates(gfrance85), add = TRUE, pch = 20, lwd = 2, cex = 2)
dev.off()
```

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

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

MC can be rewritten using matrix notation:

$$MC(\mathbf{x}) = \frac{n}{\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1}} \frac{\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\mathbf{z}}, \quad (2)$$

where $\mathbf{z} = \left(\mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}/n\right)\mathbf{x}$ 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 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 the Guerry's dataset. A positive and significant autocorrelation is identified for each of the six variables. Thus, the values of literacy are the most covariant in adjacent departments, while illegitimate births (Infants) covary least.

```
l1 <- lapply(df, moran.mc, lw, 999)
l1
```

$Crime_pers

	Monte-Carlo simulation of Moran's I

data: X[[1L]]
weights: lw
number of simulations + 1: 1000

statistic = 0.4115, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater

$Crime_prop

	Monte-Carlo simulation of Moran's I

data: X[[2L]]
weights: lw
number of simulations + 1: 1000

statistic = 0.2636, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater

$Literacy

	Monte-Carlo simulation of Moran's I

data: X[[3L]]
weights: lw
number of simulations + 1: 1000

statistic = 0.7176, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater

$Donations

	Monte-Carlo simulation of Moran's I

data: X[[4L]]
weights: lw
number of simulations + 1: 1000

statistic = 0.3534, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater

$Infants

	Monte-Carlo simulation of Moran's I

data: X[[5L]]
weights: lw
number of simulations + 1: 1000

statistic = 0.2287, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater

$Suicides

	Monte-Carlo simulation of Moran's I

data: X[[6L]]
weights: lw
number of simulations + 1: 1000

statistic = 0.4017, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater

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. Thus, we have:

$$MC(\mathbf{x}) = \frac{\mathbf{z}^{\mathrm{T}}\tilde{\mathbf{z}}}{\mathbf{z}^{\mathrm{T}}\mathbf{z}}, \tag{3}$$

since in this case $\mathbf{1}^{\mathrm{T}}\mathbf{W}\mathbf{1} = n$. This 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.
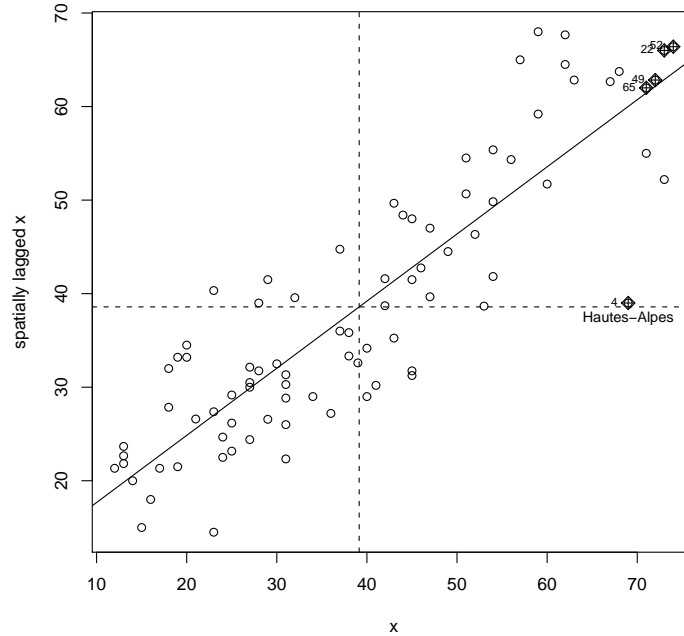
Considering the Literacy variable of Guerry's data, the Moran scatterplot clearly shows strong autocorrelation. It also shows that the Hautes-Alpes département has a slightly outlying position characterized by a high value of Literacy compared to its neighbors.

```
x <- df[, 3]
x.lag <- lag.listw(lw, df[, 3])
moran.plot(x, lw)


Potentially influential observations of
        lm(formula = wx ~ x) :

   dfb.1_ dfb.x dffit   cov.r   cook.d hat
4    0.43 -0.64 -0.74_* 0.83_*  0.24   0.05
22 -0.07   0.10  0.11   1.08_*  0.01   0.06
49 -0.01   0.02  0.02   1.08_*  0.00   0.05
52 -0.07   0.09  0.11   1.08_*  0.01   0.06
65 -0.01   0.02  0.02   1.08_*  0.00   0.05


 text(x[5], x.lag[5], dep.names[5], pos = 1, cex = 0.8)
```
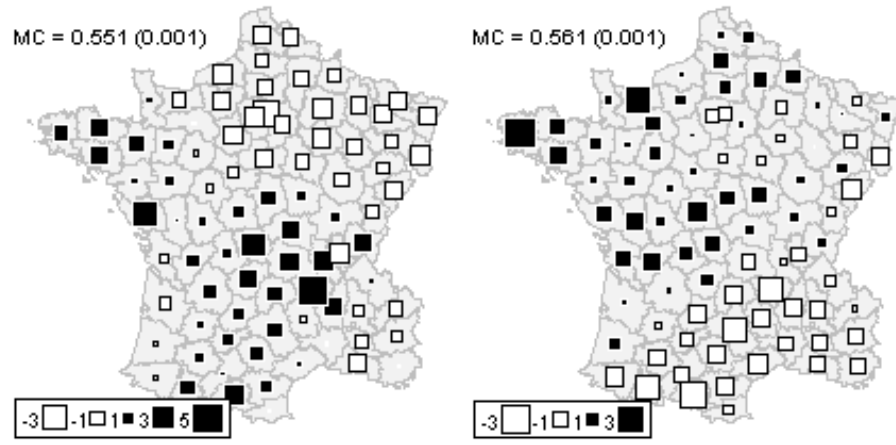
2.3 Toward an integration of multivariate and geographical aspects

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. 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. We mapped scores of the départements for the first two axes of the PCA of Guerry's data. 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*.

```
mc.pca <- lapply(pca$l1, moran.mc, lw, 999)
par(mar = rep(0.1, 4))
par(mfrow = c(1, 2))
plot(gfrance85, col = "grey95", border = "grey")
s.value(xy, pca$li[, 1], add.plot = TRUE)
text(240699, 2607012, paste("MC = ", round(mc.pca[[1]]$statistic, 3),
    " (", mc.pca[[1]]$p.value, ")", sep = ""), cex = 0.8)
plot(gfrance85, col = "grey95", border = "grey")
s.value(xy, pca$li[, 2], add.plot = TRUE)
text(240699, 2607012, paste("MC = ", round(mc.pca[[2]]$statistic, 3),
    " (", mc.pca[[2]]$p.value, ")", sep = ""), cex = 0.8)
```

11

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.

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. For instance, Guerry's data contained a partition of France into 5 regions.

We used the between-class analysis (BCA, Dolédec and Chessel, 1987), to investigate differences between regions. BCA maximizes the variance between groups.

```
bet <- between(pca, region.names, scannf = FALSE, nf = 2)
```
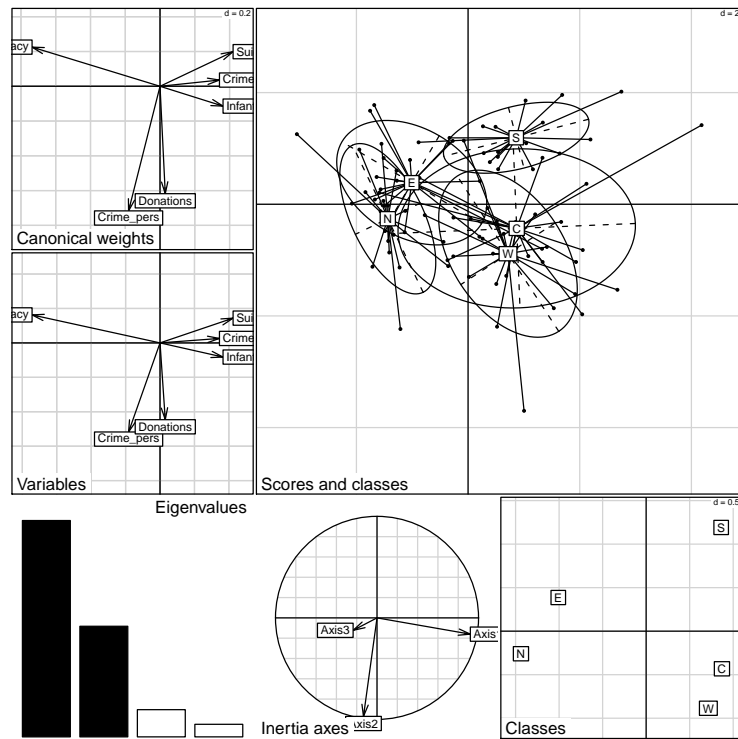
Here, 28.8% of the total variance (sum of eigenvalues of PCA) corresponds to the between-regions variance (sum of the eigenvalues of BCA).

```
bet$ratio
```

```
[1] 0.2881
```

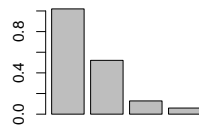The main graphical outputs are obtained by the generic `plot` function:

```
plot(bet)
```

12

The barplot of eigenvalues indicates that two axes should be interpreted. The first two BCA dimensions account for 59 % and 30.2 %, respectively, of the between-regions variance.
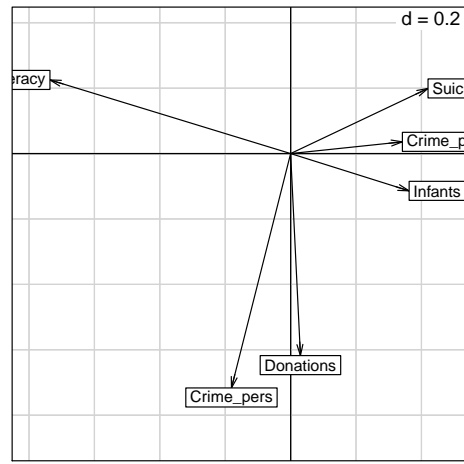
```
barplot(bet$eig)
bet$eig/sum(bet$eig) * 100
```

```
[1] 58.996 30.160  7.417  3.427
```


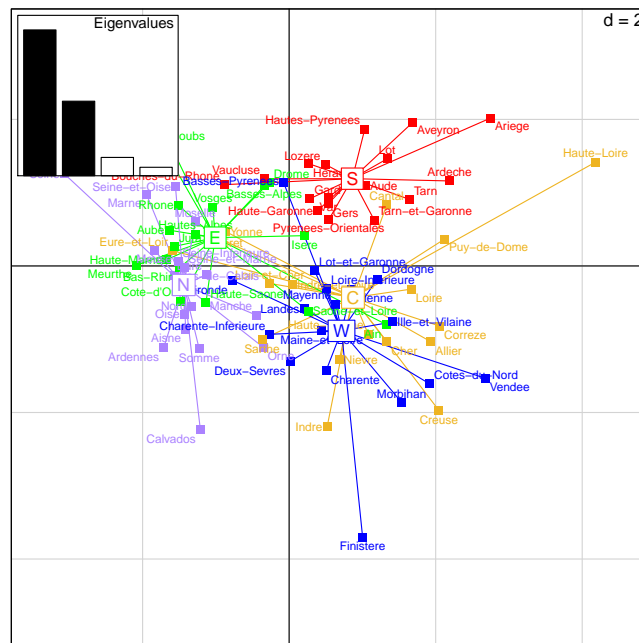The coefficients used to construct the linear combinations of variables are represented:

```
s.arrow(bet$c1, clabel = 0.8)
```

13

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.
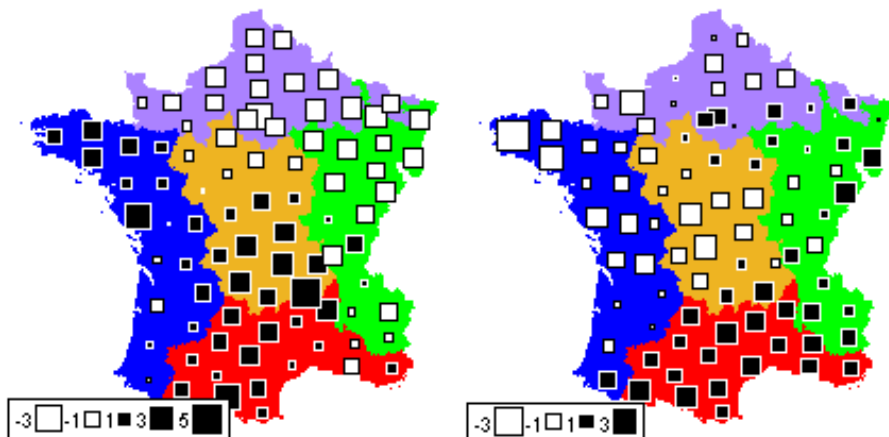
Projections of départements on the BCA axes can be represented on the factorial map:

```
s.class(bet$ls, fac = region.names, col = col.region, cellipse = 0,
    cstar = 1, clab = 0, cpoint = 0)
par(mar = rep(0.1, 4))
points(bet$ls, col = col.region[region.names], pch = 15)
pointLabel(bet$ls[, 1:2], as.character(dep.names), cex = 0.7, col = col.region[(region.names)])
s.class(bet$ls, fac = region.names, col = col.region, cellipse = 0,
    cstar = 0, clab = 1, cpoint = 0, add.plot = T)
add.scatter.eig(bet$eig, xax = 1, yax = 2, posi = "topleft", ratio = 0.25)
```

The scores can be mapped to show the spatial aspects:

```
png(file = "figs/fig-fig5d.png", width = 6, height = 3, units = "in",
    res = 72)
par(mar = rep(0.1, 4))
par(mfrow = c(1, 2))
plot(gfrance85, col = col.region[region.names], border = "transparent")
s.value(xy, bet$ls[, 1], add.plot = TRUE)
plot(gfrance85, col = col.region[region.names], border = "transparent")
s.value(xy, bet$ls[, 2], add.plot = TRUE)
dev.off()
```

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 Centre is also noticeable. In contrast, the South is very homogeneous.
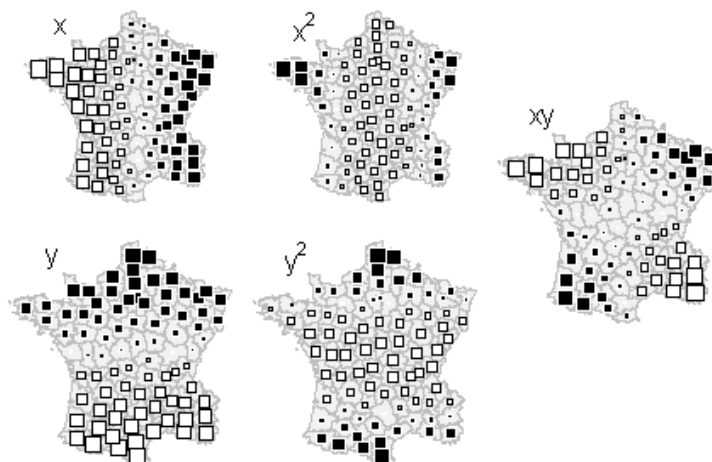
3.2 Spatial explanatory variables

Principal component analysis with respect to the instrumental variables (PCAIV, Rao, 1964), and related methods, have been often used in community ecology to identify spatial relationships. The spatial information is introduced in the form of spatial predictors and the analysis maximized 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

Student (1914) proposed to express observed values in time series as a polynomial function of time, and mentioned that this could be done for spatial data as well. Borcard et al. (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.

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

```
png(file = "figs/fig-fig6.png", width = 6, height = 4, units = "in",
    res = 72)
poly.xy <- poly(xy, degree = 2)
layout(matrix(c(1, 2, 0, 1, 2, 4, 3, 5, 4, 3, 5, 0), byrow = T, nrow = 4),
    width = 1, height = 1/2)
par(mar = rep(0.1, 4))
lab <- c(expression(x), expression(x^2), expression(y), expression(xy),
    expression(y^2))
for (i in 1:ncol(poly.xy)) {
    plot(gfrance85, col = "grey95", border = "grey")
    s.value(xy, poly.xy[, i], add.plot = TRUE, clegend = 0)
    text(240699, 2607012, lab[i], cex = 2)
}
dev.off()
```

PCAIV is then performed using the `pcaiv` function:

```
pcaiv.xy <- pcaiv(pca, poly.xy, scannf = FALSE, nf = 2)
```

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.

```
sum(pcaiv.xy$eig)/sum(pca$eig) * 100
```
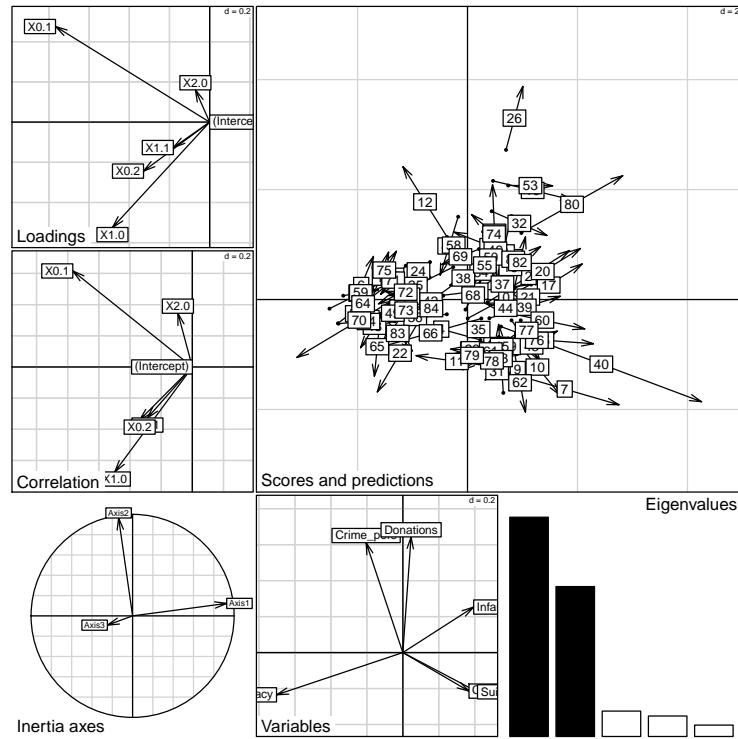
```
[1] 32.36
```

```
pcaiv.xy$eig/sum(pcaiv.xy$eig) * 100
```

```
[1] 51.423 35.152  5.954  4.799  2.671
```

The outputs of PCAIV-POLY (coefficients of variables, maps of départements scores, etc.) are very similar to those obtained by BCA. They can be represented easily by the generic `plot` function:
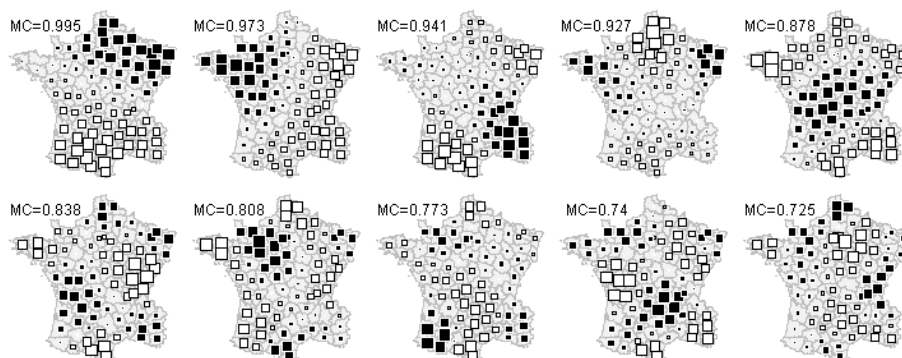
```
plot(pcaiv.xy)
```

3.2.2 Moran's eigenvector maps

An alternative way to build spatial predictors is by the diagonalization of the spatial weighting matrix $\mathbf{W}$. Moran's eigenvector maps (MEM, Dray et al., 2006) are the $n - 1$ eigenvectors of the doubly-centered matriw $\mathbf{W}$. 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.

We used the spatial weighting matrix defined above to construct MEM. The first ten MEM, corresponding to the highest levels of spatial autocorrelation, have been mapped:

```
png(file = "figs/fig-fig7.png", width = 10, height = 4, units = "in",
    res = 72)
mem <- scores.listw(lw)
par(mfrow = c(2, 5), mar = rep(0.1, 4))
for (i in 1:10) {
    plot(gfrance85, col = "grey95", border = "grey")
    s.value(xy, mem$vectors[, i], add.plot = TRUE, clegend = 0)
    text(270000, 2600000, bquote(paste("MC=", .(round(mem$values[i],
        3)))), cex = 1.5)
}
dev.off()
```

18

We introduced the first ten MEM as spatial explanatory variables in PCAIV. We call this approach PCAIV-MEM.

Here, 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.

```
sum(pcaiv.mem$eig)/sum(pca$eig) * 100
```
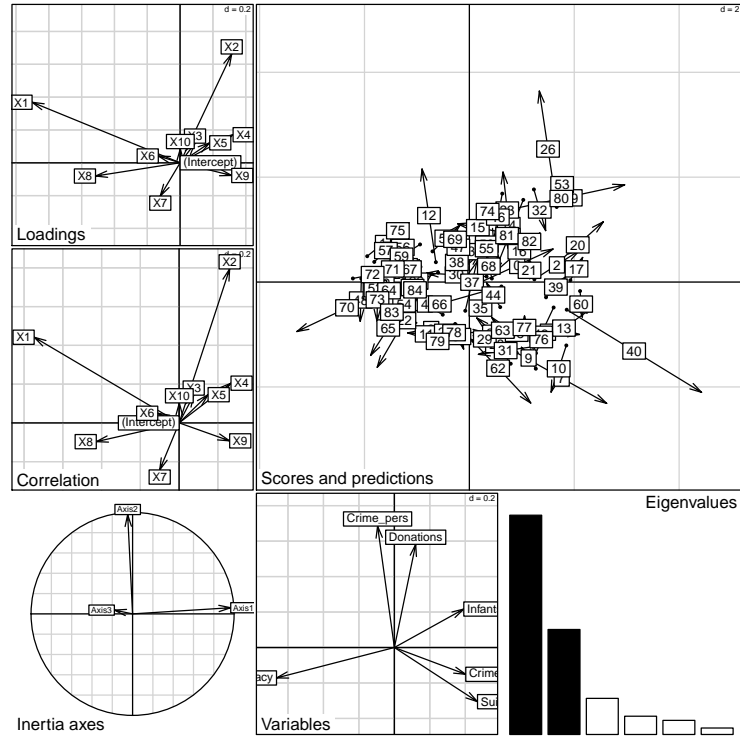
```
[1] 44.12
```

```
pcaiv.mem$eig/sum(pcaiv.mem$eig) * 100
```

```
[1] 54.929 26.300  9.042  4.574  3.533  1.622
```

The outputs of PCAIV-MEM (coefficients of variables, maps of départements scores, etc.) are very similar to those obtained by BCA. They can be represented easily by the generic `plot` function:

```
plot(pcaiv.mem)
```
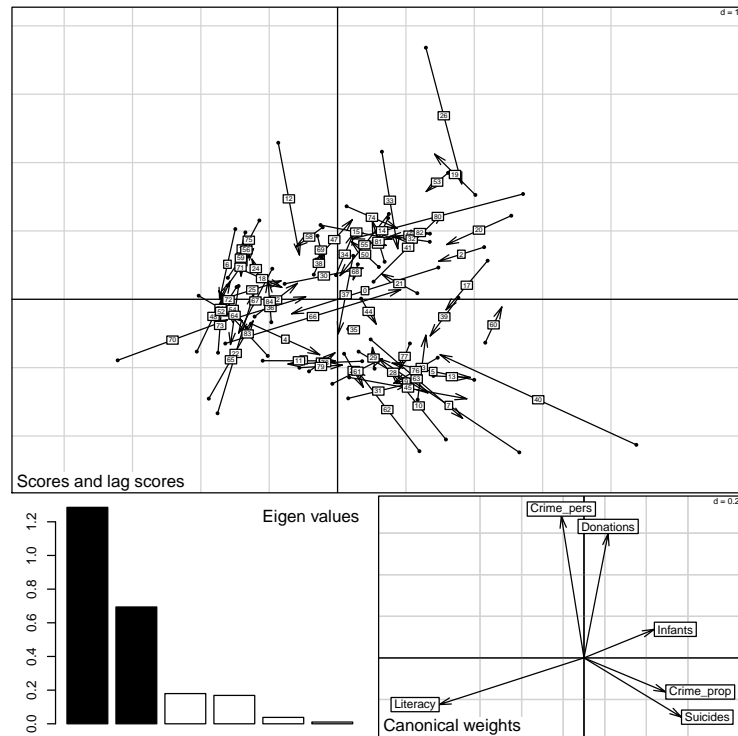
3.3 Spatial graph and weighting matrix

The MEM framework introduced the spatial information into multivariate analysis through the eigendecomposition 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.

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 et al. (2008) generalized Wartenberg's method by introducing a row-standardized spatial weighting matrix in the analysis of a statistical triplet. 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. MULTISPATI finds coefficients to obtain a linear combination of variables that maximizes a compromise between the classical multivariate analysis and a generalized version of Moran's coefficient.

```
ms <- multispati(pca, lw, scannf = FALSE)
```

The main outputs of MULTISPATI can be represented easily by the generic `plot` function:

```
plot(ms)
```


The barplot of eigenvalues suggests two main spatial structures. Eigenvalues of MULTISPATI are the product between the variance and the spatial autocorrelation of the scores, while PCA maximizes only the variance. The differences between the two methods are computed by the `summary` function:

```
sum.ms <- summary(ms)
```

```
Multivariate Spatial Analysis
Call: multispati(dudi = pca, listw = lw, scannf = FALSE)

Scores from the initial duality diagramm:
     var   cum   ratio  moran
RS1 2.140 2.140 0.3567 0.5506
RS2 1.201 3.341 0.5569 0.5614
RS3 1.102 4.443 0.7406 0.1806

Multispati eigenvalues decomposition:
     eig   var   moran
CS1 1.286 2.017 0.6375
CS2 0.694 1.177 0.5898
```
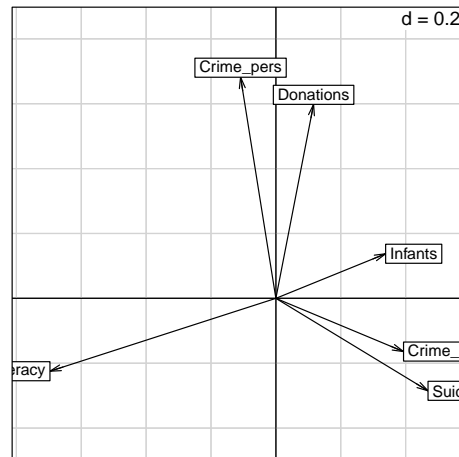
```
sum.ms
```

```
        eig     var    moran
CS1 1.28587 2.0172 0.63746
CS2 0.69399 1.1766 0.58985
CS3 0.17948 1.0072 0.17821
CS4 0.16856 0.6645 0.25366
CS5 0.03803 0.7531 0.05050
CS6 0.01045 0.3815 0.02739
```

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).

Coefficients of variables allow to interpret the structures:
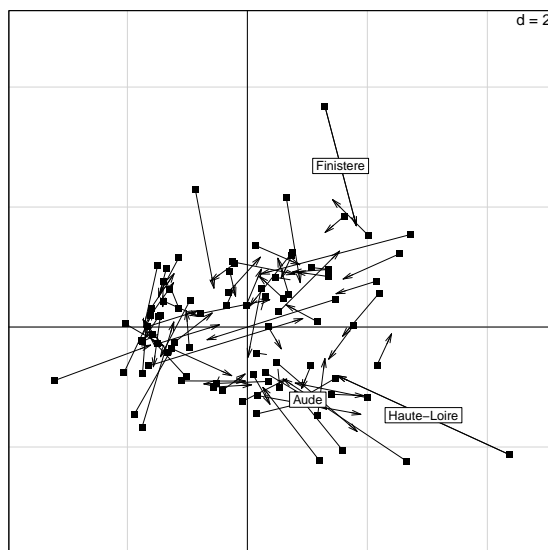
```
s.arrow(ms$c1, clabel = 0.8)
```


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 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.

Spatial autocorrelation can be seen as the link between one variable and the lagged vector. 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 and lagged scores:

```
s.match(ms$li, ms$ls, clabel = 0, pch = 15)
s.match(ms$li[c(10, 41, 27), ], ms$ls[c(10, 41, 27), ], label = dep.names[c(10,
    41, 27)], clabel = 0.8, add.plot = TRUE, pch = 15)
```
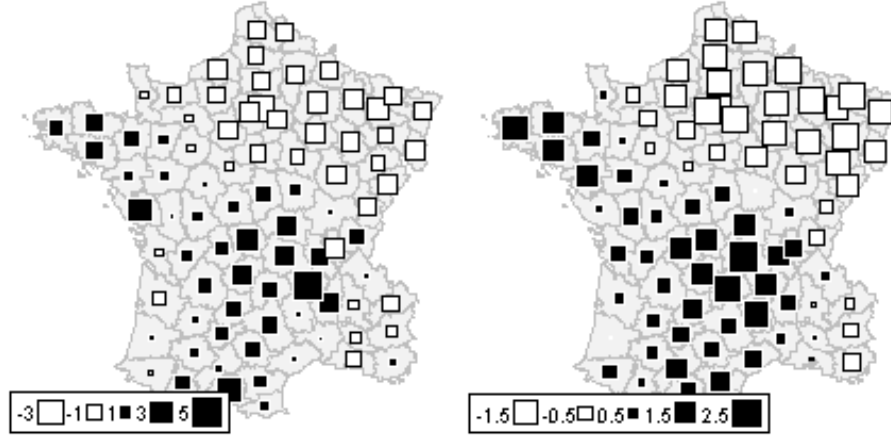
Each département can be represented on the factorial map by an arrow (the bottom corresponds to its score, the head corresponds to its lagged score. A short arrow reveals a local spatial similarity (between one plot and its neighbors) while a long arrow reveals a spatial discrepancy. This viewpoint can 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).

The link between the scores and the lagged scores (averages of neighbors weighted by the spatial connection matrix) can be mapped in the geographical space. For the first axis, we have:

```
png(file = "figs/fig-fig8e.png", width = 6, height = 3, units = "in",
    res = 72)
par(mar = rep(0.1, 4))
par(mfrow = c(1, 2))
plot(gfrance85, , col = "grey95", border = "grey")
s.value(xy, ms$li[, 1], add.plot = TRUE)
plot(gfrance85, , col = "grey95", border = "grey")
s.value(xy, ms$ls[, 1], add.plot = TRUE)
dev.off()
```

4 Conclusions

Even if the methods presented are quite different in their theoretical and practical viewpoints, their applications to Guerry's dataset yield very similar results. We provided a quantitative measure of this similarity by computing Procrustes statistics (Peres-Neto and Jackson, 2001; Dray et al., 2003) between the scores of the départements onto the first two axes for the different analyses. All the values of the statistic are very high and significant; this confirms the high concordance between the outputs of the different methods.

```
mat <- matrix(NA, 4, 4)
mat.names <- c("PCA", "BCA", "PCAIV-POLY", "PCAIV-MEM", "MULTISPATI")
colnames(mat) <- mat.names[-5]
rownames(mat) <- mat.names[-1]
test1 <- procuste.randtest(pca$li[, 1:2], bet$ls[, 1:2])
test2 <- procuste.randtest(pca$li[, 1:2], pcaiv.xy$ls[, 1:2])
test3 <- procuste.randtest(pca$li[, 1:2], pcaiv.mem$ls[, 1:2])
test4 <- procuste.randtest(pca$li[, 1:2], ms$li[, 1:2])
test5 <- procuste.randtest(bet$ls[, 1:2], pcaiv.xy$ls[, 1:2])
test6 <- procuste.randtest(bet$ls[, 1:2], pcaiv.mem$ls[, 1:2])
test7 <- procuste.randtest(bet$ls[, 1:2], ms$li[, 1:2])
test8 <- procuste.randtest(pcaiv.xy$ls[, 1:2], pcaiv.mem$ls[, 1:2])
test9 <- procuste.randtest(pcaiv.xy$ls[, 1:2], ms$li[, 1:2])
test10 <- procuste.randtest(pcaiv.mem$ls[, 1:2], ms$li[, 1:2])
mat[1:4, 1] <- c(test1$obs, test2$obs, test3$obs, test4$obs)
mat[2:4, 2] <- c(test5$obs, test6$obs, test7$obs)
mat[3:4, 3] <- c(test8$obs, test9$obs)
mat[4, 4] <- test10$obs
mat
```

```
                 PCA    BCA PCAIV-POLY PCAIV-MEM
BCA           0.9789     NA         NA        NA
PCAIV-POLY    0.9792 0.9897         NA        NA
PCAIV-MEM     0.9886 0.9936     0.9954        NA
MULTISPATI    0.9869 0.9954     0.9951    0.9986
```

References

L. Anselin. Local indicators of spatial association. *Geographical Analysis*, 27: 93–115, 1995.

L. Anselin. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In M. Fischer, H. Scholten, and D. Unwin, editors, *Spatial analytical perspectives on GIS*, pages 111–125. Taylor and Francis, London, 1996.

D. Borcard, P. Legendre, and P. Drapeau. Partialling out the spatial component of ecological variation. *Ecology*, 73:1045–1055, 1992.

A. Cliff and J. Ord. *Spatial autocorrelation*. Pion, London, 1973.

S. Dolédec and D. Chessel. 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(3):403–426, 1987.

S. Dray and T. Jombart. Revisiting guerry's data: introducing spatial constraints in multivariate analysis. *Annals of Applied Statistics*, submitted.

S. Dray, D. Chessel, and J. Thioulouse. Procrustean co-inertia analysis for the linking of multivariate data sets. *Ecoscience*, 10(1):110–119, 2003.

S. Dray, P. Legendre, and P. Peres-Neto. Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). *Ecological Modelling*, 196:483–493, 2006.

S. Dray, S. Saïd, and F. Débias. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science*, 19:45–56, 2008.

M. Friendly. A.-M. Guerry's moral statistics of France: challenges for multivariable spatial analysis. *Statistical Science*, 22:368–399, 2007.

R. Geary. The contiguity ratio and statistical mapping. *The incorporated Statistician*, 5(3):115–145, 1954.

D. A. Griffith. Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Canadian Geographer*, 40 (4):351–367, 1996.

A. Guérry. *Essai sur la Statistique Morale de la France*. Crochard, Paris, 1833.

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

P. Moran. The interpretation of statistical maps. *Journal of the Royal Statistical Society Series B-Methodological*, 10:243–251, 1948.

P. Peres-Neto and D. Jackson. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, 129:169–178, 2001.

C. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya A*, 26:329–359, 1964.

W. Student. The elimination of spurious correlation due to position in time or space. *Biometrika*, 10:179–180, 1914.

D. Wartenberg. Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, 17(4):263–283, 1985.