Journal of Statistical Software

April 2012, Volume 47, Issue 2. *http://www.jstatsoft.org/*

GeoXp: An R Package for Exploratory Spatial Data Analysis

Thibault Laurent
Toulouse School of Economics

Anne Ruiz-Gazen
Toulouse School of Economics

Christine Thomas-Agnan
Toulouse School of Economics

Abstract

We present **GeoXp**, an R package implementing interactive graphics for exploratory spatial data analysis. We use a data set concerning public schools of the French Midi-Pyrénées region to illustrate the use of these exploratory techniques based on the coupling between a statistical graph and a map. Besides elementary plots like boxplots, histograms or simple scatterplots, **GeoXp** also couples maps with Moran scatterplots, variogram clouds, Lorenz curves and other graphical tools. In order to make the most of the multidimensionality of the data, **GeoXp** includes dimension reduction techniques such as principal components analysis and cluster analysis whose results are also linked to the map.

Keywords: exploratory analysis, spatial econometrics, spatial statistics, interactive graphics, brushing and linking, dimension reduction.

1. Introduction

Exploratory analysis of georeferenced data must take into account their spatial nature. The aims of exploratory spatial data analysis include describing geographical distributions, identifying spatial outliers, discovering trends or heterogeneity, regimes of spatial association, validating models. Geographic information systems (GIS) are very elaborate cartographic tools but their statistical analysis capabilities are generally limited. When they include statistical techniques, they often are very basic tools from descriptive statistics (boxplots, histograms, barcharts, etc.). Some GIS include scripts runnable from a graphical user interface (GUI) that can calculate the local Moran's I and other local indicators of spatial association (LISA) but they are poorly integrated. Openshaw (1994) and Anselin (1994, 1998) attempt to define the type of exploratory data analysis techniques that GIS should try to incorporate. Anselin

(1994) advocates the integration in the GIS of local measures of spatial association, spatial lag pies, spatial lag scatterplots, Moran scatterplots as well as variogram clouds and pocket plots. Wilhelm and Steck (1998) and Unwin and Unwin (1998) also argue for the use of local measures of spatial association.

The use of the coupling between a map and a statistical graph such as a histogram, a boxplot or a scattermatrix has already been advocated in the literature (see detailed references below). The coupling is the fact that the selection of a zone on the map results in the automatic highlighting of the corresponding points on the statistical graph or conversely the selection of a portion of the graph results in the automatic highlighting of the corresponding points on the map.

Haslett, Bradley, Craig, Unwin, and Wills (1991) link histograms, double histograms, scatterplot matrices, and varioclouds (see Section 4) with the maps using the Pascal language. Anselin and Bao (1997) implement the methods advocated in Anselin (1994) linking **ArcView** and **SpaceStat**. Brundson (1998) implements the scatterplot matrix, the neighbour plot and the angle plot (see Section 4) plus some spatial smoothing of maps for trend detection with XLISP-STAT. Haining, Wise, and Ma (1998) and Wise, Haining, and Ma (2001) develop **SAGE**, a software system held in the **ArcInfo** GIS, with very similar capabilities as those quoted above. Let us mention also the linkage of **ArcView** and **XGobi** by Cook, Majure, Symanzik, and Cressie (1996) and Symanzik, Cook, Lewin-Koh, Majure, and Megretskaia (2000) and the cartographic data visualizer (**CDV**) of Dykes (1998) based on the Tcl/Tk language. The **ArcGIS** (Esri 2011) geostatistical analyst extension (see `http://www.esri.com/software/arcgis/extensions/geostatistical/`) includes extensive kriging capabilities and exploratory tools but is mainly oriented towards geostatistics and requires the expensive **ArcGIS** software. **Mondrian** (Theus and Urbanek 2008), written in Java, features interactive descriptive tools such as mosaic plots, scatter plots, bar charts, histograms and parallel coordinates plots.

MANET (Unwin, Hawkins, Hofman, and Siegl 1996), preceded by **SPIDER** (Haslett, Wills, and Unwin 1990) and **REGARD** (Unwin, Wills, and Hasslett 1990), also contains a number of interactive descriptive tools with a central objective of dealing with missing values, but does not contain any tool from spatial statistics.

GeoDa is a free specialized software for spatial data analysis developed by Anselin (2003) and combines maps with statistical graphs dynamically. It offers many functionalities for exploratory data analysis and spatial regression and its main strength is extensive mapping with full linking and brushing possibilities. In contrast (see Anselin, Syabri, and Kho 2006), **GeoDa** is a "closed box" which does not benefit from the tremendous expansion of the R project (R Development Core Team 2011) and has to be considered as an introductory tool to spatial data analysis.

Wise, Haining, and Signoretta (1998) evaluate and compare **CDV**, **MANET**, **SAGE** and **SpaceStat**.

LeSage and Pace (2004) and Liu and LeSage (2010) develop C/C++ code to export polygons and data information from **ArcView** shapefiles into MATLAB (The MathWorks Inc. 2007) and a GUI as well as mapping functions to link a map with a histogram, a Moran scatterplot, a parallel coordinate plot and a distribution density plot with the possibility of zooming (see also LeSage 1998).

The need for a more adaptable, comprehensive and unified tool motivated us to start the

development of a set of statistical routines adapted to the exploration of georeferenced data called **GeoXp**. At an earlier stage of this project, some exploratory functions had first been implemented in both the MATLAB and S-PLUS environments (Heba, Malin, and Thomas-Agnan 2002). **GeoXp** is now available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=GeoXp`. It is mainly an exploratory tool for researchers and experienced users in spatial statistics, spatial econometrics, geography, ecology, epidemiology, etc. **GeoXp** is a stand-alone (free-standing) package independent of a GIS and this is certainly an advantage. Its functions allow coupling between statistical plots and elementary maps as defined before. The routines are user friendly. The user does not need to write a lot of R code except for loading the data and calling a function in the command window: After entering some parameters as arguments of the function (usual inputs are at least a `Spatial` object as defined by Pebesma and Bivand 2005 and the name of the variables concerned by the graph), the user needs to execute it. He is then asked to perform the selections by mouse clicking.

The quality of the cartographic display is not a priority for the exploration itself and this is why the emphasis in **GeoXp** is rather in the implementation of spatial statistics tools as numerous and as up to date as possible. The final map for a publication can always be produced by a more sophisticated mapping tool if necessary. With R, one may for example use the function `map` from the **maps** package (Brownrigg 2012) or the function `spplot` from the **sp** package (Pebesma and Bivand 2005; Bivand, Pebesma, and Gómez-Rubio 2008).

GeoXp is based on R: This choice of language is motivated by the flexibility of R and the existence of many statistical packages developed in this language. Moreover, the number of R packages dealing with the analysis of spatial data is now significant (see Bivand 2011a). The flexibility and adaptability of **GeoXp** comes from the fact that R is an open source software and thus the user who is familiar with R can customize **GeoXp** with its own routines and benefit from the large number of modelling tools available in this environment. The advantage over approaches linking a computer engine for statistical computations and a cartographic device such as **ArcView** is that **GeoXp** is not specific to an operating system and it avoids file transfers. **GeoXp** is meant to be used in association with other R packages in a modeling strategy. Since **GeoXp** is mainly concerned by interactive exploratory analysis, it should be used in the beginning at the exploratory phase for outlier and trend detection, autocorrelation assesment, and later on at the confirmatory phase as a diagnostic tool for residuals. **GeoXp** includes spatial econometrics as well as geostatistics tools so that it can be coupled with **spdep** (Bivand 2011b) in the analysis of areal data as well as with **gstat** (Pebesma 2004) or **geoR** (Jr and Diggle 2001) for example in the analysis of continuous random fields, allowing an interactive analysis of specific subsets of points. Some unique features are also present in **GeoXp** such as linking a map with a Lorenz curve (see Section 2) or with generalized principal components analysis (see Section 6). **GeoXp** offers also some rare tools such as the angle plot of Brundson (1998) (see Section 4) and the neighbor plot (see Section 5).

As far as timing performances are concerned, we ran some tests on an Optiplex GX745 2 duo 2.13GHz under Windows Vista and using version 1.5.0 of **GeoXp**. With a function like `histomap`, the time required to make a selection is under 1 second for a data set of size less than 5,000. With a data set of size 10,000, the time required is about 1.5 seconds and for size 50,000, it is about 6.5 seconds. For functions which involve selections on couples of points like for example the `moranplotmap` function, the call takes about 19 seconds for size 1,000 (resp. 3 minutes 50 seconds for size 2,500). However, beyond a data set of size 4,000,

an allocation memory problem arises and we should be able to improve on this in the next version of **GeoXp**.

Section 2 describes the basic functionalities of **GeoXp** illustrated through an example. In Section 3, we present briefly descriptive functions which link simple univariate or bivariate graphs to maps. In Section 4, we focus on geostatistical functions such as the variocloud and the drift plot while, in Section 5, we describe econometrics functions such as the Moran plot and the neighbor plot. The multivariate functions such as generalized principal component analysis are presented in Section 6.

For this paper, we have chosen to illustrate only a selection of the different routines and the reader will find a comprehensive list of the **GeoXp** functions in the annex and more illustrations on the **GeoXp** web site `http://gremaq.univ-tlse1.fr/stat/Chrisweb/SiteGeoXp/`.

2. Description of the basic functionalities

2.1. Description of the data set

The data set we consider concerns the 226 public junior high schools (collège in French) of the Midi-Pyrénées region of France during the 2003–2004 school year. These schools are located at the centroids of the "communes" (the "commune" is the smallest french administrative subdivision) they belong to, since it is the most precise geographical information we have. The contours of the eight departments of the region (Ariège, Gers, Haute-Garonne, Hautes-Pyrénées, Lot, Tarn, Tarn-et-Garonne) are displayed on the subsequent maps. For each school, we consider the following characteristics: The number of students per class (`Nb.students.per.class`), the cost per student (`Cost.per.student`) and the occupancy rate (`Occupancy.rate`) which is the number of students (`Nb.students`) in the school divided by the number of students the school has been designed for.

We also have the mean age of the teachers in the school (`Teachers.age`), the frequency of certifiés teachers[1] (`Freq.certifies`), the frequency of agrégés teachers (`Freq.agreges`), the frequency of students who repeated a class (`Freq.rep.stud`) and the number of specialities offered to students in the school (`Nb.specialties`).

Finally, we have a measure of rurality (`index.rurality`) of the "communes" where the schools are located. This measure has been defined by INSEE (Institut National de la Statistique et des Études Économiques, see Bessy-Pietri and Sicamois 2001). Following this classification which is based on demographic and economic criteria, the "communes" with at least one public school may be `urban`, `intermediate` or `rural`. Among the 226 public schools, there are 95 schools which are located in urban "communes", 23 in intermediate ones and 108 in rural ones.

In order to illustrate the descriptive (Section 3), the geostatistical (Section 4) and the multivariate functions (Section 6), we use a first version of the data set which is at the school level. We thus have 226 observations corresponding to 175 "communes" with at least one school on the Midi-Pyrénées map. We also use this data for Figure 2.

This data set is included in a `data.frame` object of **GeoXp** and contains both the coordinates of the school (`longitude` and `latitude` variables) and some observed characteristics:

[1] Schematically, certifiés are tenured teachers with a Bachelor level while agrégés are tenured teachers with a Master level.

```
R> library("GeoXp")
R> data("mp.school")
R> names(mp.school)

 [1] "longitude"             "latitude"              "name.city"
 [4] "index.rurality"        "Nb.students"           "Occupancy.rate"
 [7] "Cost.per.student"      "Nb.students.per.class" "Freq.certifies"
[10] "Freq.agreges"          "Freq.rep.stud"         "Nb.specialties"
[13] "Teachers.age"
```

Note that we did our analysis on the original data but in the **GeoXp** version of the data, random permutations have been applied to the last three variables for confidentiality reasons so that one should not try to give meaning to exploratory results drawn from its analysis. For Figure 3 and for the econometric functions (Section 5), we use a second version of the data set which is at the "pseudo-canton"[2] level. The data set has been aggregated by pseudo-cantons with 155 pseudo-cantons with at least one public school. The variables we consider for these pseudo-cantons are the mean number of students per class, the mean cost per student and the mean occupancy rate together with the number of schools in the pseudo-cantons and a rurality index. The rurality index takes the value 1 if the ratio of the number of rural communes in the pseudo-canton to the number of communes is larger than 1/2, and 0 otherwise.

2.2. General principles

The **GeoXp** functions apply to the analysis of any data set of variables measured at geographical sites or on geographical zones such as cities, counties, countries, etc. called basic spatial units. For each site (for each zone), the data set must contain the cartesian coordinates of the site (respectively of the centroid of the zone). Variables can be continuous or categorical.

Since version 1.5.0 of **GeoXp**, we adopt the use of `Spatial` classes as described in Pebesma and Bivand (2005) and available in the **sp** package, for a larger compatibility with other packages. Bivand *et al.* (2008) enumerate the packages which directly or indirectly depend on or import **sp** objects, like **geoR**, **gstat**, **spdep**, etc. Besides, it is quite simple to create a `Spatial` object. For example, in our data set, the spatial units are represented by points. The two steps consist in first creating a `SpatialPoints` object which only contains the coordinates of the spatial units and second in creating a `SpatialPointsDataFrame` object which contains both the `SpatialPoints` object and the `data.frame`:

```
R> mp.school_coord <- cbind(mp.school$longitude, mp.school$latitude)
R> mp.school_sp <- SpatialPoints(mp.school_coord)
R> mp.school_spdf <- SpatialPointsDataFrame(mp.school_sp, mp.school)
```

In most of the **GeoXp** functions, the first argument is a `Spatial` object[3] as created above and the second argument is the name of the variables concerned by the graph (or their column number).

[2] A "canton" is a french administrative subdivision which usually is an aggregate of several communes. However, large "communes" may be divided into several cantons and in that case, a pseudo-canton corresponds simply to the commune. In the other cases, pseudo-cantons correspond to cantons.

[3] It could be a `SpatialPointsDataFrame` for points, a `SpatialLinesDataFrame` for lines, a `SpatialPixelsDataFrame` for pixels or a `SpatialPolygonsDataFrame` for polygons.

In the case of geographical zones, one may use additionally the coordinates of polygonal spatial contours to improve the map quality and to help identifying locations.

As far as format is concerned, any format that can be imported in R can be used as long as it contains the geographical coordinates. For example one can import a shapefile format from **ArcView** using the function `readShapePoly` of the R package **maptools** (Lewin-Koh and Bivand 2012) or the function `readOGR` of the R package **rgdal** (Keitt, Bivand, Pebesma, and Rowlingson 2012), and a MIF/MID format from **MapInfo** using the function `readOGR` of the R package **rgdal**. The geographical contours have their own format (coordinates of vertices separated from one unit to the next by the missing value symbol NA). The two **GeoXp** functions `polylist2list` and `spdf2list` allow to convert some `Spatial` object into the format of **GeoXp** contours. In the following illustrative graphics, we also use the spatial contours of the departments of the Midi-Pyrénées region, using the following code:

```
R> shp.file <- system.file("shapes/school.shp", package = "GeoXp")[1]
R> mp_map <- readShapePoly(shp.file)
R> mp.contour <- spdf2list(mp_map)$poly
```

The names of the main **GeoXp** functions reflect their functionality and always end with "map" (example: `moranplotmap`, `scattermap`). As one can see in Figure 1, a call to a **GeoXp** function generally opens three windows: Two graphical R windows for the statistical graph and the map respectively, and one Tk window for the menu. The user then selects on the menu the graph on which he wants to select points first. This graph then becomes active and the selection by mouse clicking begins.

For selecting the points either on a statistical graph or on a map, the user can choose between selecting individual points (centroids) or selecting points inside a given polygon. For the selection on the statistical graph, there are several cases. In the case of a histogram or a bar plot, it is possible to select several non necessarily contiguous bars. In the case of a density plot, one can select one or several intervals on the x-axis by mouse clicking or by specifying its endpoints. In the case of a boxplot, one can select outliers or inter-quartiles ranges. In the case of the Lorenz curve, it is possible to select either a given percentage of spatial units on the first axis or a given threshold value of the variable.

The selection of an already selected unit deletes its selection. Upon exit, if the user clicks on the `Save results` button, it creates an object called `last.select`, a vector of integers with the numbers of the spatial units selected at the last selection step, allowing further analysis of the selection's characteristics.

Selected units are marked with a different color or alternatively with a different. Polygons representing the boundaries of the spatial units can be added easily if available[4]. Names of the variables can be specified for use in the graph axes labels.

As in cartographic devices, proportional symbol maps can be produced by adding bubbles to the map, proportional to a given variable with a legend for their size.

For most functions, an additional statistical graph can be added. This additional graph is only interactive in one direction though: Selections made on the first graph or the map will appear on this additional graph but one cannot select from the additional graph.

[4]If the `Spatial` object given is a `SpatialPolygonsDataFrame`, the polygons are automatically proposed with `Spatial contours` button.

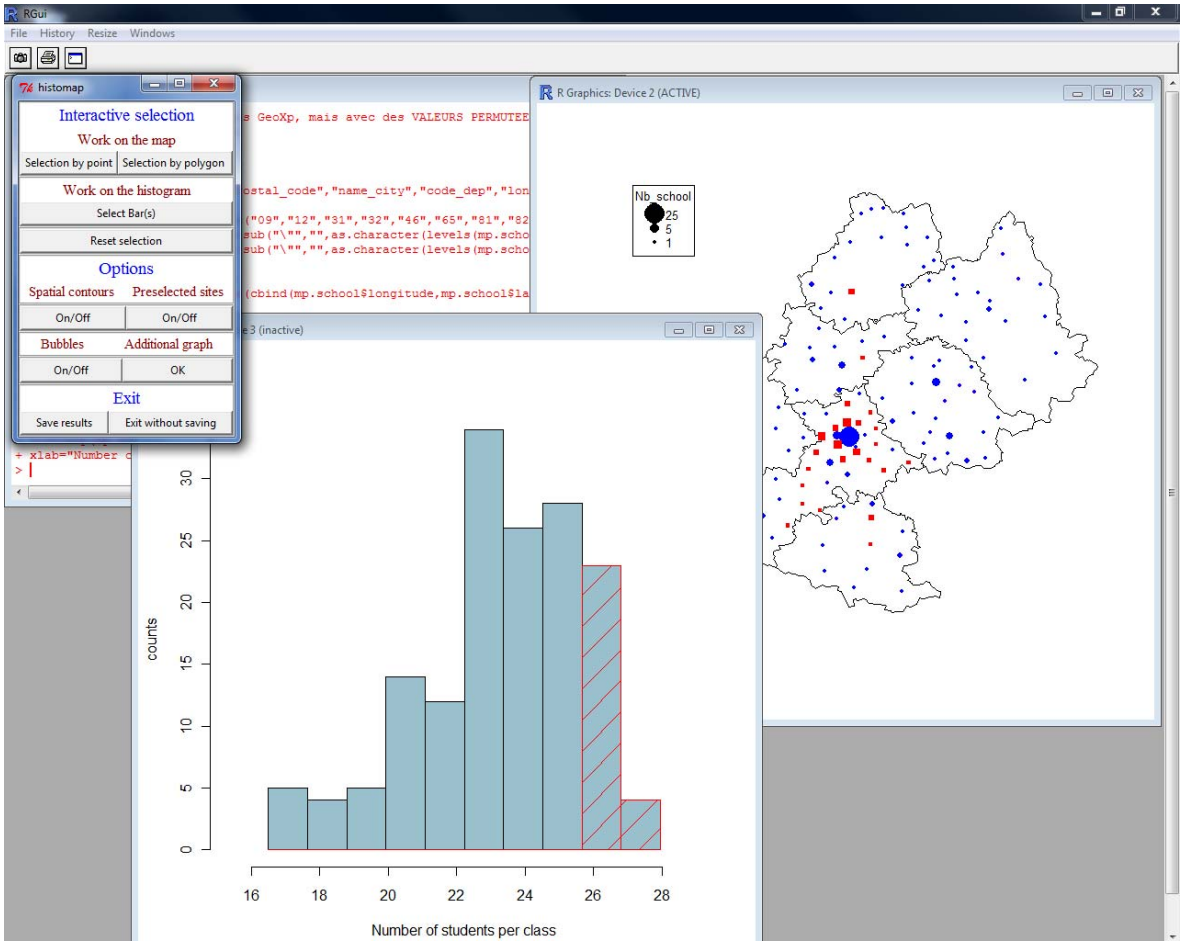

Figure 1: Example of **GeoXp** display.

The reader can get more details about the use of the options in the vignette obtained with the code `vignette("presentation_geoxp")`.

2.3. Example

Figure 2 displays a scatterplot of the cost per student of each school versus its occupancy rate with conditional quartile curves. An additional graph shows the bar plot of the rurality index of the school's "commune". A selection on the scatterplot of schools with an occupancy rate greater than 1, in red on the plot, shows that they belong to rural as well as intermediate and urban areas but that they represent a high proportion of schools in the intermediate areas. The map reveals that they are mainly located in the surroundings of Toulouse. To underline the simplicity of the code, you will find below the code used to produce these plots for version 1.5.6 of **GeoXp**.

```
R> scattermap(mp.school_spdf, c("Occupancy.rate", "Cost.per.student"),
+    quantiles = c(0.25, 0.75), carte = mp.contour, pch = 15, cex = 0.9,
+    xlab = c("Occupancy rate", "Cost per student"))
```

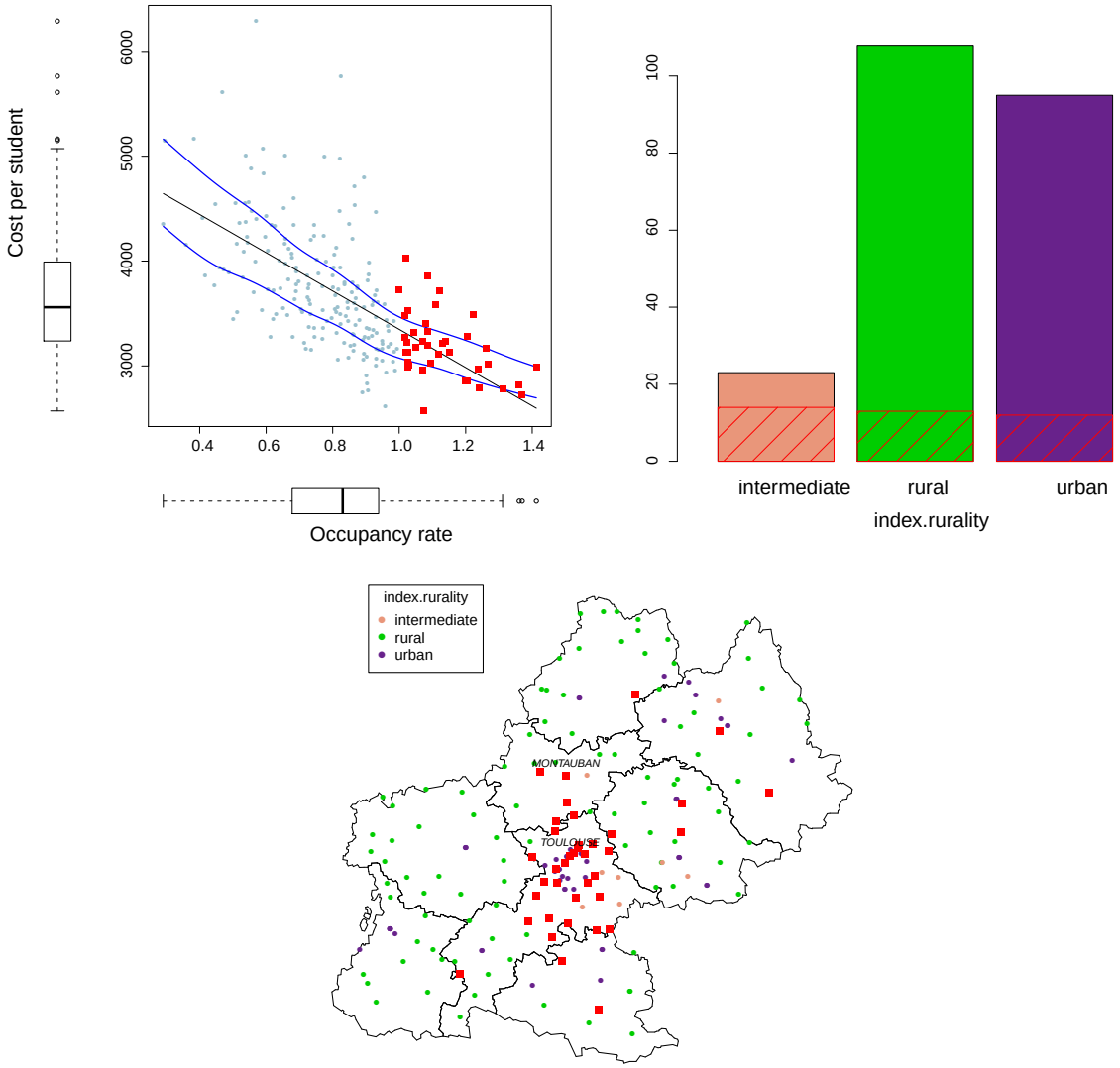

Figure 2: Scatterplot of cost per student versus occupancy rate and barplot of rurality index: Selection of schools with occupancy rate greater than 1.

2.4. Package dependencies

GeoXp depends on several packages. The **spdep** package contains classes for spatial weight matrices which are used in **GeoXp** for its econometrics functions. **spdep** depends itself on the **sp** and **maptools** packages described above, necessary for the Spatial class definition and for importing Spatial format files like shapefiles. The `qsreg` function of the **fields** package (Furrer, Nychka, and Sain 2012) is necessary for drawing a quantile spline regression which is an option common to scatterplot graphics. The `bkde` function of the **KernSmooth** package (Wand 2011) is used for distribution density plot. Finally, the `inout` function of the **splancs** package (Rowlingson, Diggle, and Bivand 2012) is used for selection, to test point inclusion in a polygon.

3. Descriptive functions

The descriptive functions are called `barmap`, `boxplotmap`, `histomap`, `densitymap`, `histobarmap`, `dblehistomap`, `dbledensitymap`, `polyboxplotmap`, `ginimap`, `plot3dmap` and `scattermap`.

In the case of a simple histogram, the selection of some bars of this histogram will show the corresponding zones on the map, which is just a more elaborate variant of the previous tool as in Haslett *et al.* (1991). In the other direction, a selection of a subregion of the map produces the subhistogram of the distribution of the variable in this subregion. Since the goal is then to compare the distribution of the variable on the whole map to its subdistribution on the selected zone, it is not optimal to use histograms based on counts as most packages do, so we have introduced an alternative function allowing the user to produce two kernel density estimators instead of two histograms. The user can choose the bandwidth or use a default option for this choice. He can also change the initial bandwidth selection with a ruler displayed in the Tk window, resulting in an automatic updating of the graphs. For discrete variables, it is also possible to link a bar plot to the map.

When the statistical graph is a simple boxplot, only the selection on the boxplot is implemented and allows the user to display the zones corresponding to lower or upper quartiles as well as to outliers (as in Haslett *et al.* 1990). The same information is conveyed by choropleth maps in a GIS.

For a couple of variables, a double histogram or a double kernel density estimator can be graphed and linked to the map. Selection is then possible on the map as well as on one of the histograms or density graphs. In Figure 3, the `dbledensitymap` function displays the density of the number of students per class and of the cost per student. A selection of the pseudo-cantons with more than 26 students per class is made on the first density and produces on the second plot the graph of the corresponding subdensity for the cost per student in these pseudo-cantons. The subdensity appears to be shifted to the left revealing a lower cost per student in these pseudo-cantons, which are mainly located in the surroundings of Toulouse.

A simple scatterplot of a couple of variables can also be linked to the map and selection is again possible in both directions as in Brundson (1998). A kernel smoother can be added to the scatterplot for convenience with a flexible choice of bandwidth. An option allows the user to overlay conditional quantile estimates instead of the kernel smoother which estimates the conditional mean, thus allowing a more precise exploration of the cloud when one is interested for example in the extreme rather than the average behaviour.

The possibility of linking the map with a Lorenz curve allows the study of the geographical component of the concentration or inequality measured by the Gini index (see Gastwirth 1972). The Lorenz curve is a scatterplot of the relative mass of a given variable X due to the sites with a value of X less than or equal to x versus the relative frequency of such sites. The Gini index (area between the Lorenz curve and the diagonal of the unit square) measures the inequality in the distribution of X.

The selection of a given frequency F on the frequency axis results in the printing of the corresponding relative mass G on the other axis, the corresponding quantile (value x such that the cumulative distribution function of X at x equals to F) as well as the selection of the corresponding points on the map (spatial units such that X is less than or equal to x). For example Figure 4 shows the Lorenz curve of the number of students and the bar plot of the rurality index. The Lorenz curve, which is away from the diagonal with a Gini index

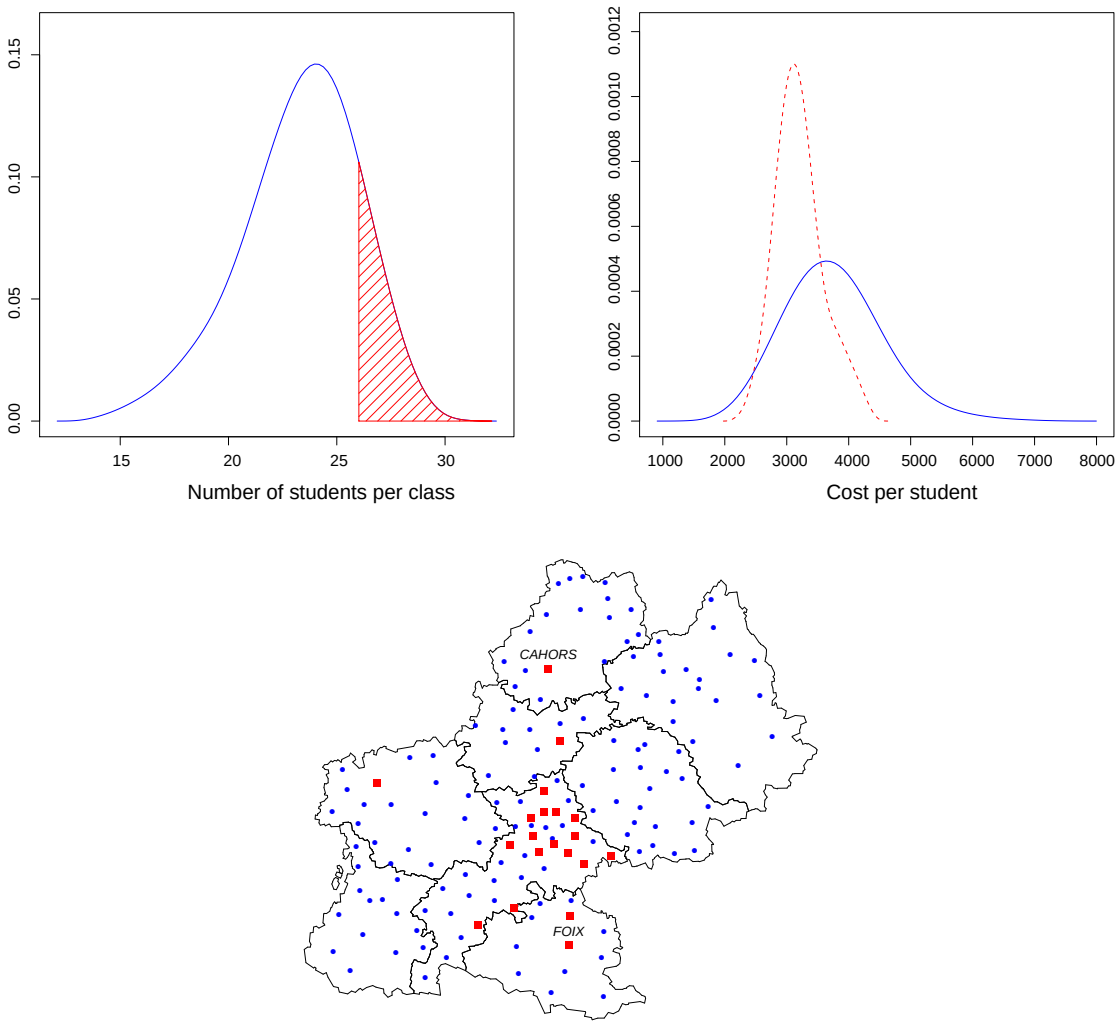

Figure 3: Density of cost per student and number of students per class: Selection of pseudo-cantons with more than 26 students per class.

of 0.28, shows that a small number of schools concentrate a large number of students. A selection of the first 20 % of schools sorted by increasing number of students (corresponding to a number of students less than 362) is reflected on the bar plot which shows that they are mainly located in rural areas.

4. Geostatistical functions

The geostatistical functions are called `angleplotmap`, `driftmap` and `variocloudmap`.

As in Cressie (1993, p. 37), in order to examine trends in one variable, **GeoXp** creates a grid of a given fineness and for each square of the grid computes the mean of the variable for all basic units intersecting the square. It is then easy to produce row and column means and medians,

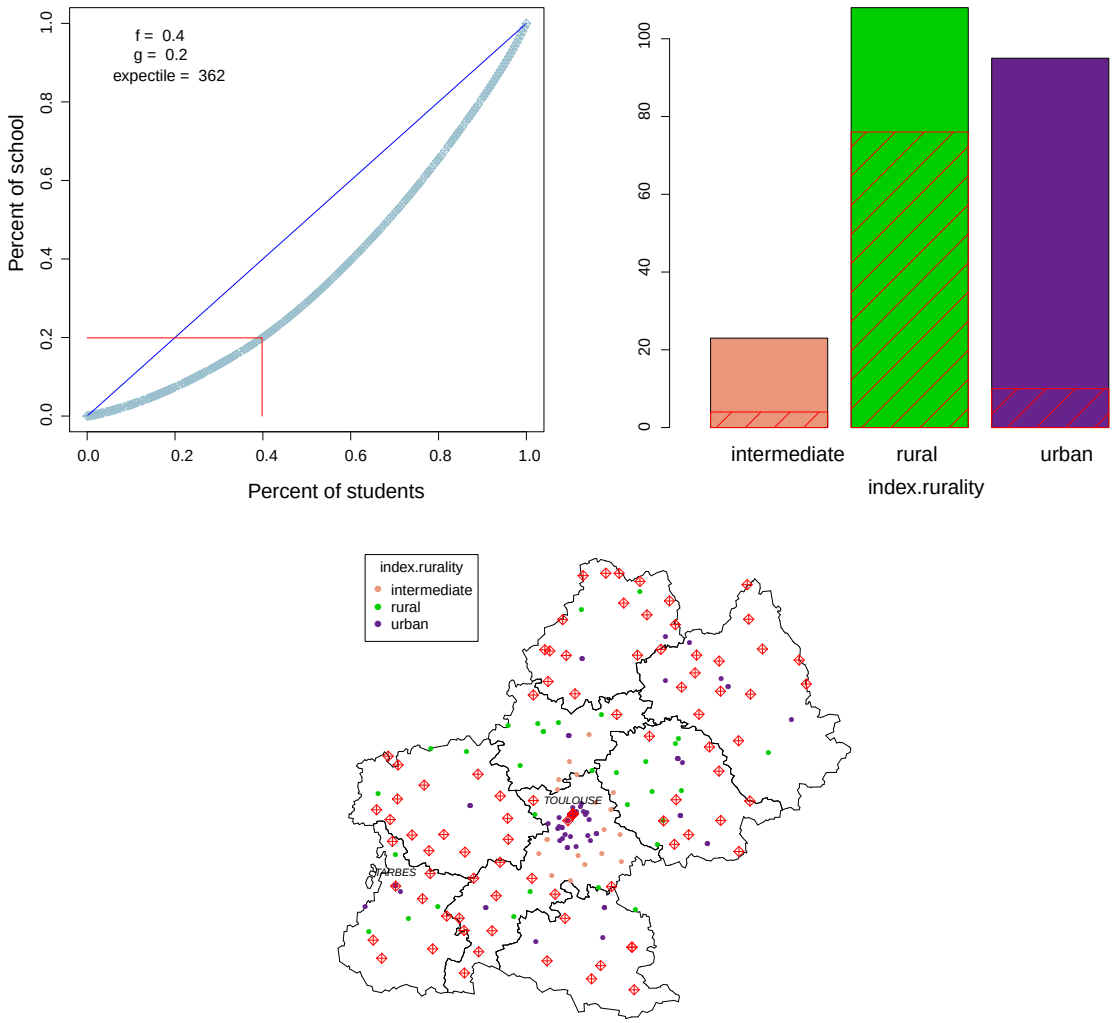

Figure 4: Lorenz curve and Gini index for the number of students: Selection of the first 20 % of schools sorted by increasing number of students.

and plot the row means and medians to the right of the map as well as the column means and medians below the map. The study of the variation of the row means with longitude and column means with latitude brings out the north-south and east-west trends if present. An option allows the user to rotate the map by a given angle and thus study trends in any direction. Discrepancies between means and corresponding medians detect the presence of outliers in a given row or column. Generally, the user may have no prior idea of the directions of the main trends.

It is then interesting to use an angle plot prior to the trend graphic (see Brundson 1998) that may reveal unknown spatial heterogeneity. The angle plot implemented here is a scatterplot of the square root of the absolute differences between the values of the variable at two given zones as a function of the bearing of a line joining the centroids of the two zones (in radians

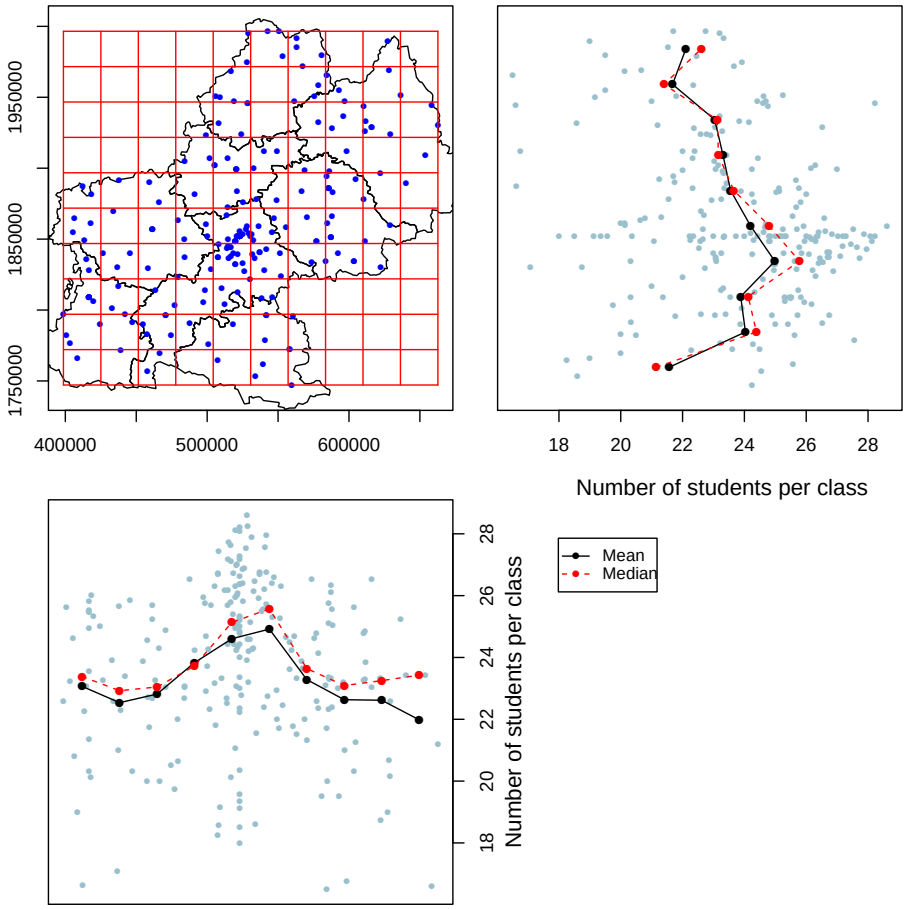

Figure 5: *Drift map* for the number of students per class.

or degrees). The `driftmap` of the number of students per class in Figure 5 shows that the central region (area of Toulouse) corresponds to the highest levels of students per class and that there is no outlier.

In Figure 6, the selection of the couples of schools with a bearing of $\pi/4$ radians and with large absolute differences in the number of students per class reveals a disparity between the area of Toulouse and the north-east of the region. It is interesting to train oneself in the interpretation of angle plots by applying them to deterministic trends such as latitude and longitude.

The variogram cloud is another tool inspired by geostatistics to study autocorrelation (Chauvet 1982). It is a simple scatterplot of the half square of the difference between the value of the variable at two locations against the distance between these points. As in Haslett *et al.* (1991), outliers may be mapped by highlighting those points on this graph which have a high value of the second coordinate. An option allows the user to overlay an empirical variogram or a smooth of this scatterplot thus estimating the variogram function (with the possibility of a robust alternative Cressie 1993). This option is important to represent the bulk of the cloud since, because of the high number of couples of positions with a low value on the vertical axis, it is often desirable to combine this with another option allowing to represent only

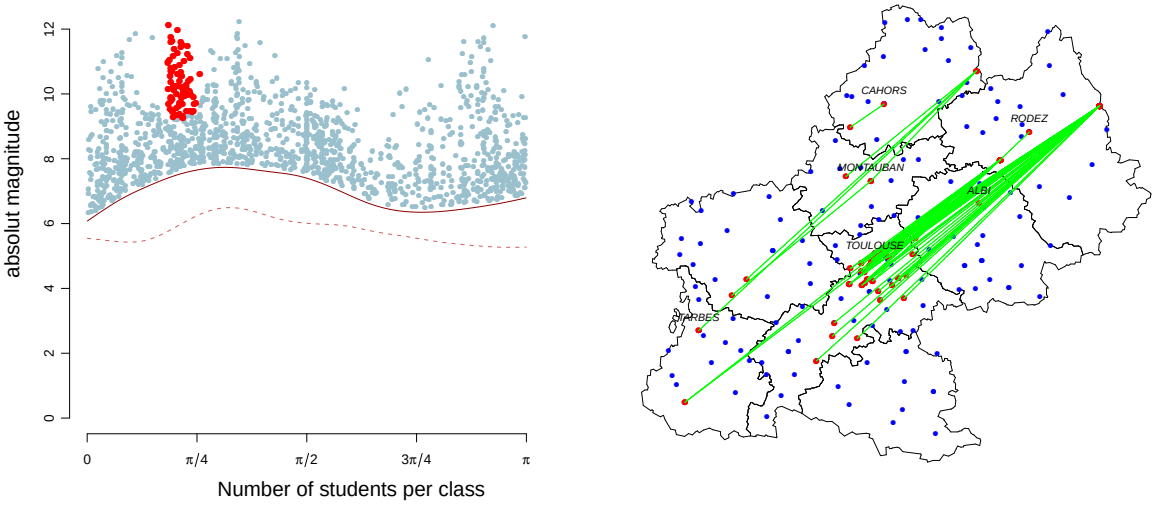

Figure 6: Angle plot for the number of students per class: Selection of large absolute differences for a given angle.

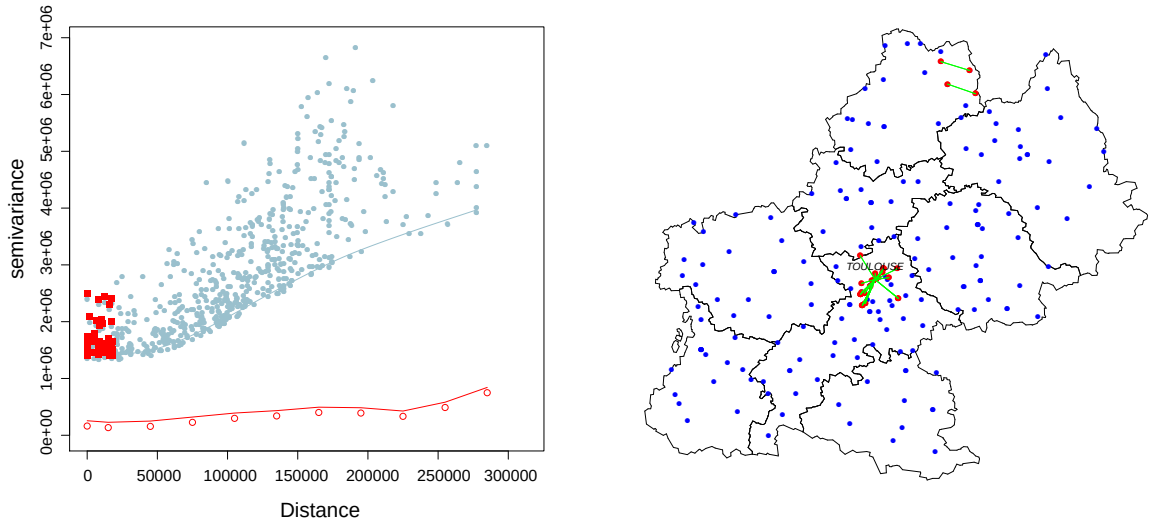

Figure 7: Variocloud for the cost per student above the 95th percentile : Selection of large absolute differences and small distance.

those couples with a value above a chosen threshold percentile (conditional on the value of the horizontal coordinate). Finally, another option allows to concentrate on couples of points in a given direction (with a tolerance) and overlay a directional variogram. In Figure 7, one can see that high differences in the cost per student for neighboring schools appear between schools located in Toulouse and schools located in the suburban areas of Toulouse. A threshold of 95 percent has been chosen for representing the points.

5. Econometric functions

The econometric functions are called `moranplotmap` and `neighbourmap`. For specifying the spatial weight matrix, the first function uses a weight list object (`listw` object) whereas the second function uses a neighbor list (`nb` object). A neighbor list can be constructed using the function `knearneigh` (based on the K nearest neighbours) or using other similar functions in the R package **spdep** as `tri2nb` (based on a Delaunay triangulation). The **GeoXp** function `makeneighborsw` creates a `matrix` object based on a given number of nearest neighbors or a given distance threshold or both. Conversion functions from one class of weight matrix to another are found in **spdep** such as `nb2listw`, `listw2mat`, `mat2listw`, etc. The **GeoXp** function `normw` performs row standardizing of a `matrix` object. A `matrix` object can then be converted into a `listw` object, which contains itself a `nb` object. Since version 1.4, **GeoXp** contains two functions called `barnbmap` and `histnbmap` which make interactive exploratory analysis of a neighbors list. They allow for example the detection of spatial units with a large number of neighbors or neighbors with a high euclidian distance.

To examine spatial autocorrelation, given a spatial binary weight matrix (Bavaud 1998) containing information about the neighboring relationships of the basic spatial units, one can simply make a scatterplot of the value of the variable on each unit versus the value of the same variable on the neighboring units (neighbor plot). Points far away from the diagonal on this plot identify local outliers and selection is again possible on the plot as well as on the map. When a point is selected on the map, its neighbors are shown connected by lines to this point. For the variable cost per student, we draw in Figure 8 a neighbor plot with a weight matrix based on 4 nearest neighbors. This graph shows some amount of spatial autocorrelation with points not too far from the diagonal and we notice the asymmetry due to the corresponding asymmetry of the weight matrix. The selection of the pseudo-cantons with the smallest costs reveals that their neighbors have small to medium cost per student and that they are exclusively located in the surroundings of Toulouse. This tool is also interesting for investigating a chosen spatial weight matrix as is shown in Figure 9. For the same 4 nearest

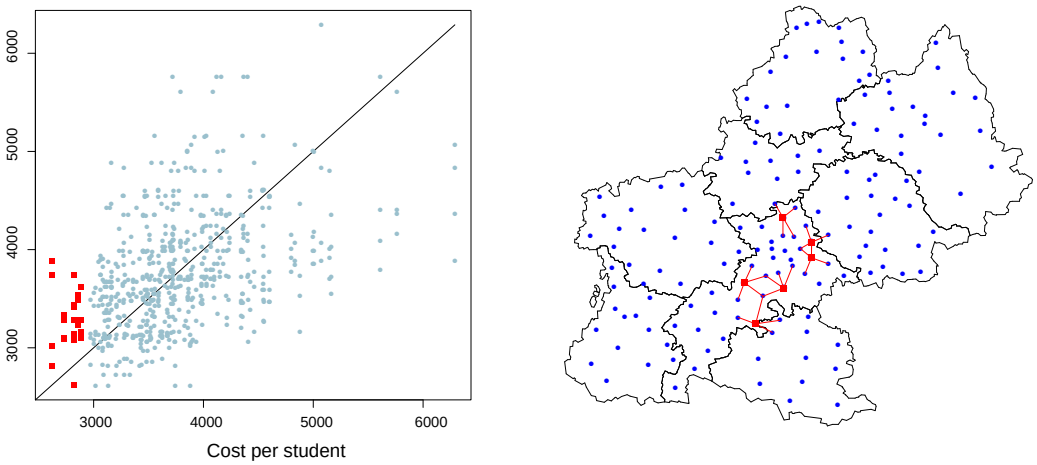

Figure 8: Neighbor plot for the cost per student: Selection of small costs.

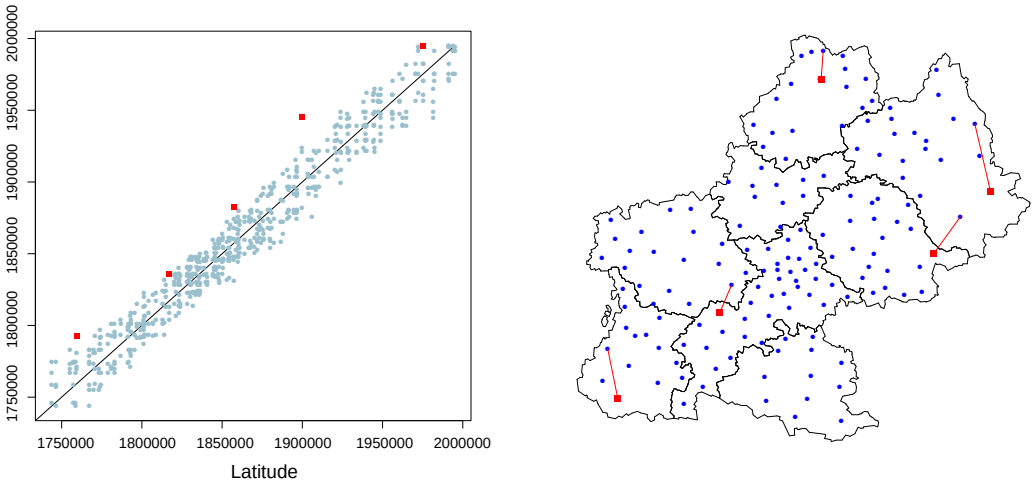

Figure 9: Neighbor plot for latitude: Selection of large differences in latitude.

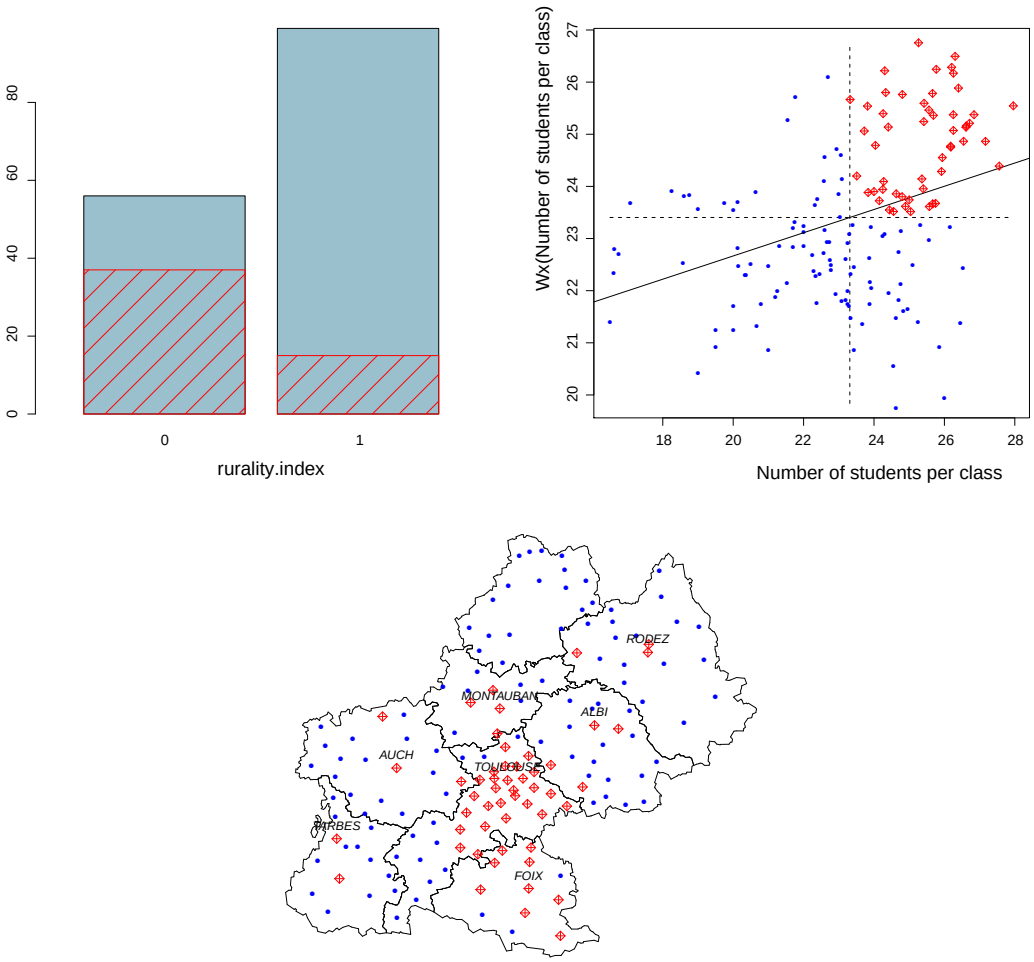

Figure 10: Moran scatterplot of the number of students per class: Selection of first quadrant.

neighbors matrix, the couples of points with a large difference in latitude are selected. Large distances between neighbors may arise for some weight matrices (for example those based on a Delaunay triangulation) and this type of graph points out at these inappropriate neighbors.

A simple scatterplot linked to the map has potentials for more advanced investigations if one applies it to transformations of the raw variables. For example, for a variable X and for a given weight matrix W, the classical Moran scatterplot (Anselin 1995) is the scatterplot of the spatial lag variable WX against X. The function `moranplotmap` of **GeoXp** links this scatterplot to the map and exhibits the regression line whose slope is the Moran index indicating the strength and nature of the spatial autocorrelation. But the observation of the cloud itself conveys more information about changes in spatial autocorrelation regimes and also outliers (see Anselin 1995 for details). The selection of each quadrant on the plot exhibits zones of positive and negative autocorrelation on the map. An option allows the computation of the local Moran statistic for the selected points. The p value of the Moran gaussian test for spatial autocorrelation is displayed by default and the p value of the permutation test based on a chosen number of simulations can also be obtained.

Figure 10 displays the Moran scatterplot of the number of students per class. The Moran index of 0.22 has a p value of 0.0001 for the gaussian and the permutation tests (with 500 permutations). The selection of the first quadrant, corresponding to pseudo-cantons with a number of students per class higher than average as well as their neighbors, shows that these are mainly urban pseudo-cantons. Besides the north of the Haute-Garonne department, they correspond to the main cities of other deparments, except for the Lot department in the north west of Midi-Pyrénées.

6. Multivariate functions

GeoXp includes the possibility of linking the results of a clustering algorithm (k-means from the R function `kmeans` or hierarchical clustering from the R function `hclust`) to the map. We suggest using a preliminary dimension reduction technique such as principal components analysis to produce bivariate plots of relevant linear combinations of the variables linked to the map. Exploratory analysis becomes rapidly cumbersome with large numbers of variables hence it is essential to use devices that select interesting projections of the data. The multivariate functions are called `clustermap` and `pcamap`. The function `pcamap` implements the generalized principal components analysis (PCA) as it is described in Caussinus, Fekri, Hakam, and Ruiz-Gazen (2003). Note that using the link between map and scatterplot, users can rapidly customize **GeoXp** to any other dimension reduction method.

In the case of usual PCA, which is a specific case of generalized PCA, one can do a scatterplot of the projection of the cloud for any couple of factorial axes and one can link it to the map. If outliers or groups appear on one of these plots, it is interesting to locate them on the map and explore their relative spatial position. Conversely, the positions on the scatterplot of a selected subregion of the map may provide information about its specificities with respect to the principal axes. The interpretation of the principal axes is guided by the representation of the variables on a separate non interactive plot. In the case of standardized PCA, correlations between the original variables and the principal components are plotted inside a correlations circle.

Figure 11 illustrates this method for the schools of Midi-Pyrénées on the following set of seven

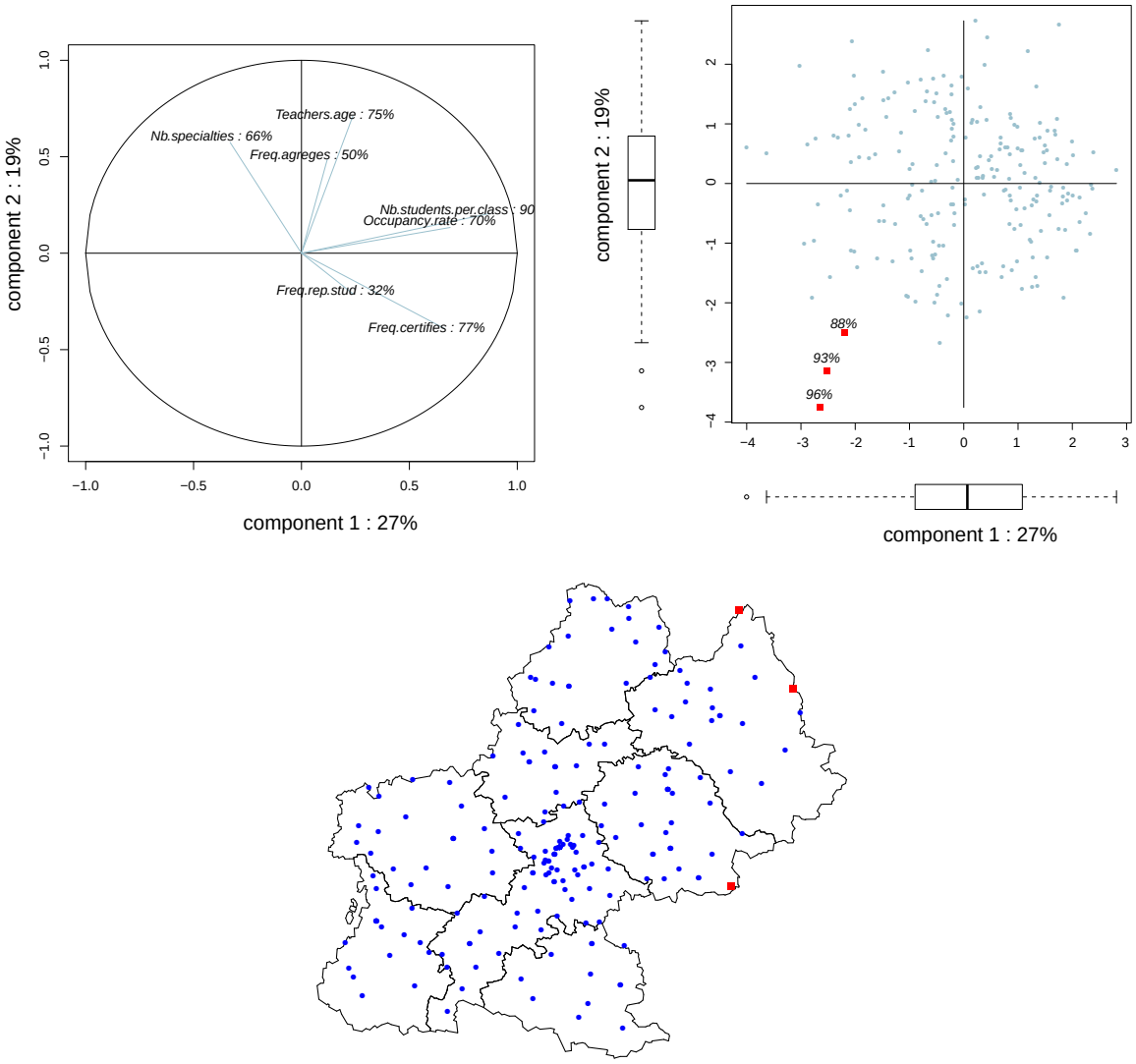

Figure 11: Principal components analysis: Selection of the three schools on the left bottom part of the first principal plane.

variables: The mean age of the teachers in the school, the frequency of certifiés teachers, the frequency of agrégés teachers, the frequency of students who repeated a class, the number of specialities offered to students in the school, the number of students per class and the occupancy rate of the school. The left plot of Figure 11 shows that the first axis is positively correlated to the number of students per class, the occupancy rate, the frequency of certifiés and more moderately to the frequency of students repeating a year. The second axis is positively correlated to the age of the teachers, the number of specialties, the frequency of agrégés and moderately negatively correlated with the number of students repeating a year. The labels on the axes indicate the percentages of inertia associated with each principal axis (which is nearly 50% for the first two axes of this example) while the percentages for the variables indicate their quality of representation on the principal plane. Three schools

have been selected on the extreme left bottom part of the scatterplot. The quality of their representation on the factorial plane is given on the scatterplot and is high for the three schools (more than 88% of their norm is accounted for by the first two principal coordinates). They differ from the other schools in the region because they have low numbers of students per class and low occupancy rates, young and not highly qualified teachers and a small number of specialties. As displayed by the map, the three of them are located at the east boundary of the region.

7. Conclusion

The project **GeoXp** started before 2000 and has known many different versions. A 1998 MATLAB version (working with MATLAB 6) still is on the site of the econometrics toolbox of LeSage (1998) and contains tools which have not yet been translated to R. It is now an R package downloadable from CRAN.

For applications oriented purposes, this set of routines has also been translated into C++ in the context of a contract with the Midi-Pyrénées region council. There are a lot of new tools that we plan to include in **GeoXp** such as a weighted version of `ginimap`, a Moran scatter plot for residuals of an OLS model, a micromap display (see Symanzik and Carr 2008), an `Apleplotmap` based on Li, Calder, and Cressie (2007), etc. More structural changes will involve in the near future the use of R classes for handling large weight matrices as sparse matrices and the use of `SpatialPolygonsDataFrame` objects which would allow coloring differently the inside of a polygon after selection, instead of only coloring the centroids of the selected polygons. The interactivity of **GeoXp** is achieved thanks to the `locator` function of R. Alternatives to `locator` have been considered such as using the interactive package **playwith** (Andrews 2010) but it does not meet all our needs yet and we will watch its evolution. Alternatives for the graphic devices such as the Qt toolbox or R-to-Java interface (as used in the **iplots**, Urbanek and Wichtrey 2011, package) still are under study.

Acknowledgments

The authors would like to thank the two anonymous reviewers and Roger Bivand for their valuable comments and suggestions. We thank the team of students who participated in the writing of the several versions of **GeoXp**. We thank as well our faculty colleagues from the university of Toulouse 1 Capitole for their research assistance and many colleagues for their remarks and contributions (E. Malin, I. Héba, J. LeSage, J. Symanzik., etc.). This work was supported by the agence nationale de la recherche through the ModULand project (ANR-11-BSH1-005).

References

Andrews F (2010). ***playwith**: A GUI for Interactive Plots Using **GTK+***. R package version 0.9-53, URL http://CRAN.R-project.org/package=playwith.

Anselin L (1994). "Exploratory Spatial Data Analysis and Geographic Information Systems." In M Painho (ed.), *New Tools for Spatial Data Analysis*, pp. 45–54. Eurostat, Luxembourg.

Anselin L (1995). "Local Indicators of Spatial Association – LISA." *Geographical Analysis*, **27**, 93–115.

Anselin L (1998). "Exploratory Spatial Data Analysis in a Geocomputational Environment." In P Longley, S Brooks, B Macmillan, R McDonnell (eds.), *GeoComputation, A Primer*. John Wiley & Sons, New York.

Anselin L (2003). ***GeoDa*** *0.9. User's Guide*. Spatial Analysis Laboratory (SAL), Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.

Anselin L, Bao S (1997). "Exploratory Spatial Data Analysis Linking **SpaceStat** and **Arcview**." In *Recent Developments in Spatial Analysis*, pp. 35–39. Springer-Verlag, Berlin.

Anselin L, Syabri I, Kho Y (2006). "**GeoDa**: An Introduction to Spatial Data Analysis." *Geographical Analysis*, **38 (1)**, 5–22.

Bavaud F (1998). "Models for Spatial Weights: A Systematic Look." *Geographical Analysis*, **30 (2)**, 153–171.

Bessy-Pietri P, Sicamois Y (2001). "Le Zonage en Aires Urbaines en 1999. 4 Millions d'Habitants en plus dans les Aires Urbaines." *INSEE-Première*, **765**, 1–4.

Bivand R (2011a). "CRAN Task View: Analysis of Spatial Data." Version 2011-12-22, URL http://CRAN.R-project.org/view=Spatial.

Bivand R (2011b). ***spdep****: Spatial Dependence: Weighting Schemes, Statistics and Models*. R package version 0.5-43, URL http://CRAN.R-project.org/package=spdep.

Bivand RS, Pebesma EJ, Gómez-Rubio V (2008). *Applied Spatial Data Analysis with R*. Springer-Verlag, New York. URL http://www.asdar-book.org/.

Brownrigg R (2012). ***maps****: Draw Geographical Maps*. R package version 2.2-5, URL http://CRAN.R-project.org/package=maps.

Brundson C (1998). "Exploratory Spatial Data Analysis and Local Indicators of Spatial Association with XLISP-STAT." *The Statistician*, **47**, 471–484.

Caussinus H, Fekri M, Hakam S, Ruiz-Gazen A (2003). "A Monitoring Display of Multivariate Outliers." *Computational Statististics and Data Analysis*, **44**, 237–252.

Chauvet P (1982). "The Variogram Cloud." In *Proceedings of the 17th APCOM International Symposium, Golden, Colorado*.

Cook D, Majure JJ, Symanzik J, Cressie N (1996). "Dynamic Graphics in a GIS: Exploring and Analysing Multivariate Spatial Data Using Linked Software." *Computational Statistics*, **11**, 467–480.

Cressie N (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Dykes J (1998). "Cartographic Visualization: Exploratory Spatial Data Analysis with Local Indicators of Spatial Association Using Tcl/Tk and **CDV**." *The Statistician*, **47**, 485–497.

Esri (2011). ***ArcGIS*** *Desktop: Release 10*. Environmental Systems Research Institute, Redlands, CA. URL http://www.arcgis.com/.

Furrer R, Nychka D, Sain S (2012). ***fields***: *Tools for Spatial Data*. R package version 6.6.3, URL http://CRAN.R-project.org/package=fields.

Gastwirth JL (1972). "The Estimation of the Lorenz Curve and Gini Index." *The Review of Economics and Statistics*, **54 (3)**, 306–16.

Haining R, Wise S, Ma J (1998). "Exploratory Spatial Data Analysis in a Geographic Information System Environment." *The Statistician*, **47**, 457–469.

Haslett J, Bradley R, Craig P, Unwin A, Wills G (1991). "Dynamic Graphics for Exploring Spatial Data with Application to Locating Global and Local Anomalies." *The American Statistician*, **45**, 234–242.

Haslett J, Wills G, Unwin AR (1990). "**SPIDER** – an Interactive Statistical Tool for the Analysis of Spatially Distributed Data." *International Journal of Geographical Information Systems*, **4 (3)**, 285–296.

Heba I, Malin E, Thomas-Agnan C (2002). "Exploratory Spatial Data Analysis with **GeoXp**." *ERSA conference papers ersa02p498*, European Regional Science Association.

Jr PJR, Diggle PJ (2001). "**geoR**: A Package for Geostatistical Analysis." *R News*, **1**(2), 14–18. URL http://CRAN.R-project.org/doc/Rnews/.

Keitt TH, Bivand R, Pebesma E, Rowlingson B (2012). ***rgdal***: *Bindings for the Geospatial Data Abstraction Library*. R package version 0.7-8, URL http://CRAN.R-project.org/package=rgdal.

LeSage J (1998). *Spatial Econometrics*. URL http://www.spatial-econometrics.com/.

LeSage J, Pace K (2004). "**Arc_Mat**, a Toolbox for Using **ArcView** Shape Files for Spatial Econometrics and Statistics." In MJ Egenhofer, C Freksa, HJ Miller (eds.), *Geographic Information Science, Proceedings of the Third International Conference*, pp. 179–190. Springer-Verlag, Berlin. Lecture Notes in Computer Science.

Lewin-Koh NJ, Bivand R (2012). ***maptools***: *Tools for Reading and Handling Spatial Objects*. R package version 0.8-14, URL http://CRAN.R-project.org/package=maptools.

Li H, Calder CA, Cressie N (2007). "Beyond Moran's I: Testing for Spatial Dependence Based on the SAR Model." *Geographical Analysis*, **39**, 357–375.

Liu X, LeSage J (2010). "**Arc_Mat**: a MATLAB-Based Spatial Data Analysis Toolbox." *Journal of Geographical Systems*, **12**(1), 69–87.

Openshaw S (1994). "What Is a Gisable Spatial Analysis." In M Painho (ed.), *New Tools for Spatial Data Analysis*, pp. 36–44. Eurostat, Luxembourg.

Pebesma EJ (2004). "Multivariable Geostatistics in S: the **gstat** Package." *Computers & Geosciences*, **30**, 683–691.

Pebesma EJ, Bivand RS (2005). "Classes and Methods for Spatial Data in R." *R News*, **5**(2), 9–13. URL http://CRAN.R-project.org/doc/Rnews/.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rowlingson B, Diggle P, Bivand R (2012). ***splancs**: Spatial and Space-Time Point Pattern Analysis*. R package version 2.01-31, URL http://CRAN.R-project.org/package=splancs.

Symanzik J, Carr DB (2008). "Interactive Linked Micromap Plots for the Display of Geographically Referenced Statistical Data." In Chen, C, Härdle, W, Unwin, A (eds.), *Handbook of Data Visualization*, pp. 267–294. Springer-Verlag, Berlin/Heidelberg.

Symanzik J, Cook D, Lewin-Koh N, Majure JJ, Megretskaia I (2000). "Linking **Arcview** and **XGobi**: Insight Behind the Front End." *Journal of Computational and Graphical Statistics*, **9**, 470–490.

The MathWorks Inc (2007). *MATLAB – The Language of Technical Computing, Version 7.5*. The MathWorks, Inc., Natick, Massachusetts. URL http://www.mathworks.com/products/matlab/.

Theus M, Urbanek S (2008). *Interactive Graphics for Data Analysis: Principles and Examples*. Chapman & Hall/CRC.

Unwin A, Unwin D (1998). "Exploratory Spatial Data Analysis with Local Statistics." *The Statistician*, **47**, 415–421.

Unwin A, Wills G, Hasslett J (1990). "**REGARD**, Graphical Analysis of Regional Data." In *ASA Proceedings of the Section on Statistical Graphics*, pp. 36–41. American Statistical Association.

Unwin AR, Hawkins G, Hofman H, Siegl B (1996). "Interactive Graphics for Data Sets with Missing Values – **MANET**." *Journal of Computational and Graphical Statistics*, **5**, 113–122.

Urbanek S, Wichtrey T (2011). ***iPlots**: Interactive Graphics for R*. R package version 1.1-4, URL http://CRAN.R-project.org/package=iplots.

Wand M (2011). ***KernSmooth**: Functions for Kernel Smoothing for Wand & Jones (1995)*. R package version 2.23-7, URL http://CRAN.R-project.org/package=KernSmooth.

Wilhelm A, Steck R (1998). "Exploring Spatial Data with Interactive Graphics and Local Statistics." *The Statistician*, **47**, 423–430.

Wise S, Haining R, Ma J (2001). "Providing Spatial Data Analysis Functionality for the GIS User: The **SAGE** Project." *International Journal of Geographical Information Science*, **15 (3)**, 239–254.

Wise S, Haining R, Signoretta P (1998). "The Visualisation of Area-Based Spatial Data." In D Unwin, P Fisher (eds.), *Case Studies of Visualization in the Social Sciences*. URL http://www.agocg.ac.uk/reports/visual/casestud/contents.htm.

A. Main functions in GeoXp

GeoXp includes two types of functions: The main functions and the auxiliary ones. The auxiliary functions are just routines called by the main functions. The list of main functions can be found below.

1. `angleplotmap`: Links a map and an angle plot (only the angle plot is active).
2. `barmap`: Links a map and a bar plot.
3. `barnbmap`: Links a map and a bar plot of the number of neighbors.
4. `boxplotmap`: Links a map and a box and whiskers plot.
5. `clustermap`: Links a map and a bar map of a clustering variable (k-means method).
6. `dbledensitymap`: Links a map and two density estimators.
7. `dblehistomap`: Links a map and two histograms.
8. `densitymap`: Links a map and a density estimator.
9. `driftmap`: This function is meant for detecting trends (non interactive).
10. `ginimap`: Links a map and a Gini plot (Lorenz curve).
11. `histobarmap`: Links a map to an histogram and a bar plot.
12. `histomap`: Links a map and an histogram.
13. `histnbmap`: Links a map and a histogram of the distances between neighbors.
14. `moranplotmap`: Links a map and a Moran scatterplot.
15. `neighbourmap`: Links a map and a neighbor plot (scatterplot of variable against variable for the neighboring sites).
16. `pcamap`: Links a map and a scatterplot of principal axes of principal components analysis.
17. `plot3dmap`: A 3-d version of the `scattermap`.
18. `polyboxplotmap`: Links a map and a box and whiskers plot.
19. `scattermap`: Links a map and a two-dimensional scatterplot.
20. `variocloudmap`: Links a map and a variogram cloud (only the variogram cloud is active).

Affiliation:

Thibault Laurent, Anne Ruiz-Gazen, Christine Thomas-Agnan
Toulouse School of Economics
Gremaq, 21 allée de Brienne
31042 Toulouse, France
E-mail: thibault.laurent@univ-tlse1.fr, ruiz@cict.fr, Christine.Thomas@tse-fr.eu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 47, Issue 2 *Submitted:* 2009-08-26
April 2012 *Accepted:* 2012-01-25