GeoXp: an R package for exploratory spatial data analysis

T. Laurent*, A. Ruiz-Gazen† and C. Thomas-Agnan‡

Toulouse School of Economics

GREMAQ, 21 allée de Brienne 31042 Toulouse, FRANCE

September 18, 2008

Abstract.

We present GeoXp, an R package implementing interactive graphics for exploratory spatial data analysis. We use a data basis 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, etc. 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.

Key Words. 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 are very basic tools from descriptive statistics (boxplots, histograms, barcharts, etc.) but none of the state of the art tools specific to spatial data from geostatistics or spatial econometrics. Openshaw (1994) and Anselin

*e-mail: thibault.laurent@univ-tlse1.fr

†e-mail: ruiz@cict.fr

‡e-mail: cthomas@cict.fr

(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 reversely the selection of a portion of the graph results in the automatic highlighting of the corresponding points on the map.
Haslett et al. (1991) link histograms, double histograms, scatterplot matrices, and varioclouds (see section 4) with the maps using the PASCAL language. Anselin and Bao (1995) 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 the XLISP-STAT language. Haining et al. (1998) and Wise et al. (2001) develop SAGE, a software system held in the ARC-INFO GIS, with very similar capabilities as those quoted above. Let us mention also the linkage of ArcView and XGobi by Cook et al. (1996) and Symanzik et al. (2000) and the cartographic data visualizer (cdv) of Dykes (1998) based on the Tcl/Tk language. The ArcGIS Geostatistical Analyst extension [1] includes extensive kriging capabilities and exploratory tools but is mainly oriented towards geostatistics and requires the expensive ArcGIS software. Mondrian [2], written in JAVA, features interactive descriptive tools such as mosaic plots, scatter plots, bar charts, histograms and parallel coordinates plots.
MANET (Unwin et al, 1996), preceded by SPIDER (Haslett et al. 1990) and REGARD, 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^{TM} 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 et al., 2006), GeoDa is a "closed box" which does not benefit from the tremendous expansion of the R project and has to be considered as an introductory tool to spatial data analysis.
Wise et al. (2001) [3] evaluate and compare cdv, MANET, SAGE and SpaceStat.
LeSage and Pace (2004) develop C/C++ code to export polygons and data information from ArcView shapefiles into Matlab and a GUI interface as well as mapping functions

[1] http://www.esri.com/software/arcgis/extensions/geostatistical/index.html

[2] http://stats.math.uni-augsburg.de/Mondrian/

[3] The visualization of area-based spatial data, In: Case Studies of Visualization in the Social Sciences online. (ed.) D. Unwin, P. Fisher. Advisory Group on Computer Graphics (AGOCG). Available online via http://www.agocg.ac.uk/reports/visual/casestud/contents.htm

to link a map with a histogram and a Moran scatterplot 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. 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 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 up to date as possible. The final map for a publication can always be produced by a more sophisticated mapping tool if necessary.

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. 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 amount of modelling tools available in this environment. GeoXp includes spatial econometrics as well as geostatistics tools. 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. Some unique features are present 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 the version 1.2 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 10000, 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 1000 (resp. 3mn50s for size 2500). However, beyond a data set of size 4000, 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 geostatistics 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 web site[4].

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[5] of the Midi-Pyrénées region of France during the 2003-2004 school year. These schools are located at the centroids of the "communes"[6] 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, the cost per student and the occupancy rate which is the number of students in the school divided by the number of students the school has been designed for.

We also have a measure of rurality of the "communes" where the schools are located. This measure has been defined by INSEE[7] (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. For Figures 1 and for the econometric functions (section 5), we use a second version of the data set which is at the "pseudo-canton"[8] 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

[4] http://gremaq.univ-tlse1.fr/stat/Chrisweb/SiteGeoXp/Index.htm

[5] collège in French

[6] The "commune" is the smallest french administrative subdivision

[7] Institut National de la Statistique et des Études Économiques

[8] 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.

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. 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 *read.ShapePoly* of the R package *maptools* or the function *readOGR* of the R package *rgdal*, 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 three GeoXp functions *map2list, polylist2list* and *spdf2list* allow to convert respectively the *maptools*, *sp* and *rgdal* formats into the format of GeoXp contours.

The names of the main GeoXp functions reflect their functionality and always end with "map" (example: *moranplotmap, scattermap*). As one can see on 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 delete its selection. Upon exit, each function returns a zero-one selection vector (one for selected units and zero for the rest) allowing further analysis of the selection's characteristics.
Selected units are marked with a different color or alternatively with a different symbol and an option allows a chosen label to be printed too. Polygons representing the boundaries of the spatial units can be added easily if available. 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

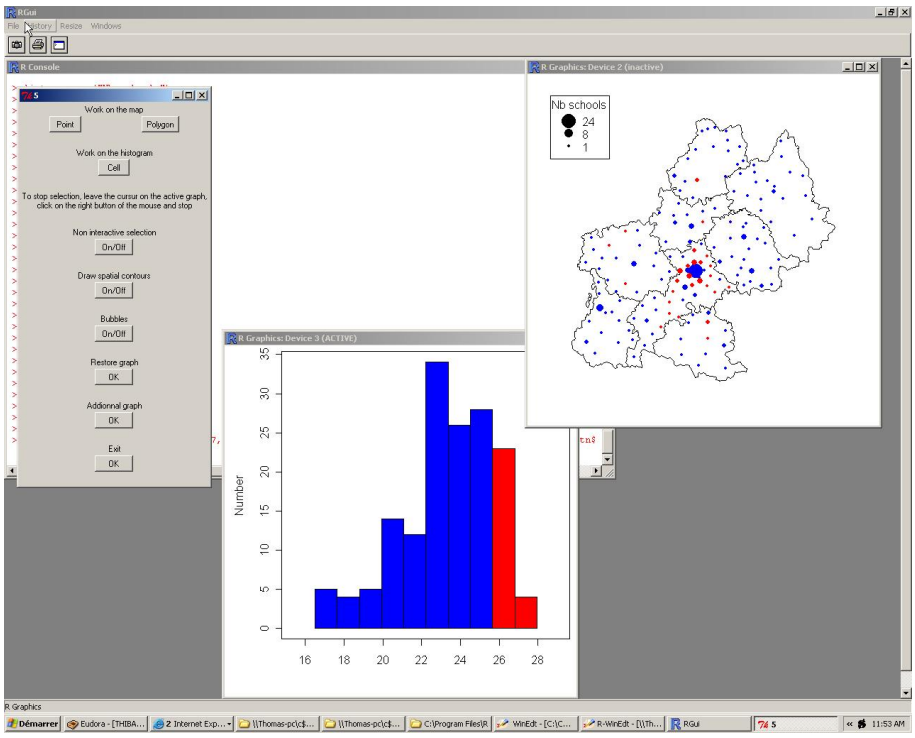

Figure 1: Example of GeoXp display

to the map, proportional to a given variable with a legend for their size.
For most functions, an additional statistical graph can be added in a given choice list, using variables specified in an option. 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.

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 load the data set and produce these plots for version 1.2 of GeoXp.

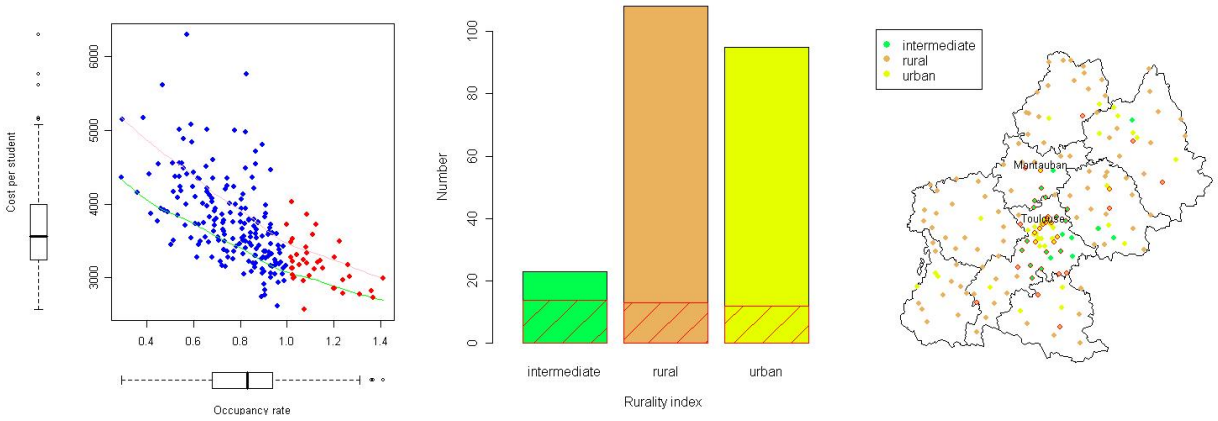

Figure 2: Scatterplot of cost per student versus occupancy rate and barplot of rurality index: selection of schools with occupancy rate greater than 1.

```
load("C:/PATH/tcmidipyr.Rdata")

load("C:/PATH/coordd.Rdata")

scattermap(tcmidipyr$latitude,tcmidipyr$longitude,Occupancy_rate,cost_per_student,
carte=coordd,label=label,opt=3, quantile=cbind(0.25,0.75),
labvar=c("Occupancy rate","Cost per student"),listvar=tcmidipyr$rurality,
listnomvar="Rurality index",color=1)
```

3 Descriptive functions

The descriptive functions are called *barmap, boxplotmap, histomap, densitymap, histobarmap, dblehistomap, dbledensitymap, polyboxplotmap, ginimap* 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 Haining et al., 1998). 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. On 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

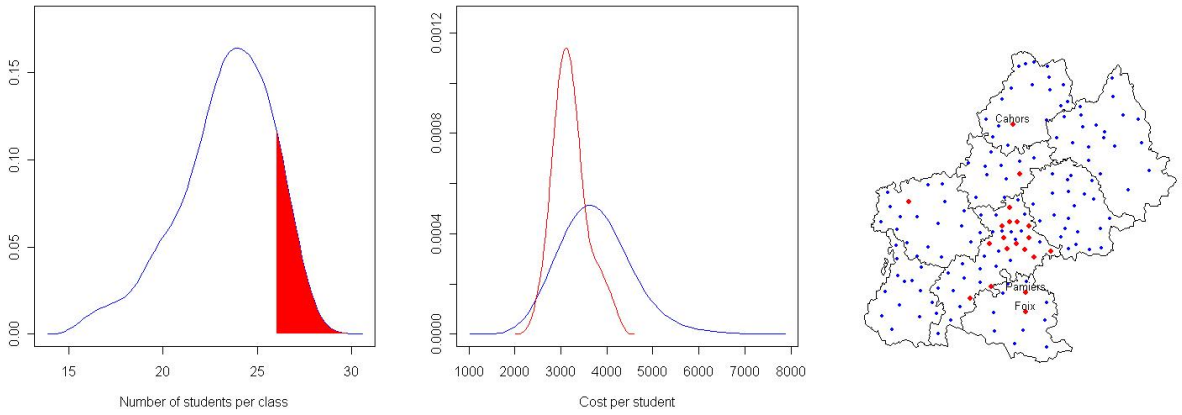

Figure 3: Density of cost per student and number of students per class: selection of cantons with more than 26 students per class.

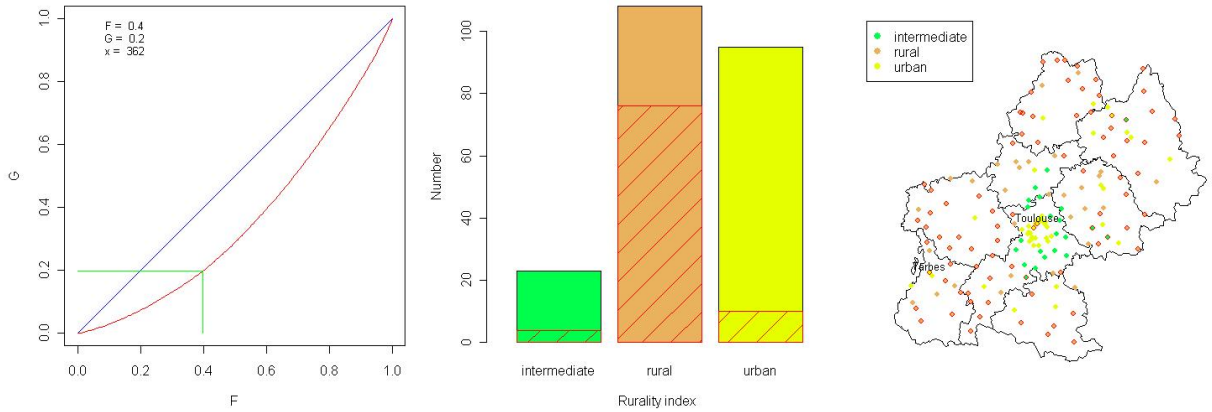

Figure 4: Lorenz curve and Gini index for the number of students: selection of the first 40 % of schools sorted by increasing number of students.

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 of 0.28, shows that a small number of schools concentrate a large number of students. A selection of the first 40 % 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 Geostatistic functions

The geostatistic functions are called *angleplotmap, driftmap* and *variocloudmap*.
As in Cressie (1993, page 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, and plot the row means and medians to the right of the map as well as the column means and medians below the map. No selection is possible here but 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 or degrees). The *driftmap* of the number of students per class on 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.

On 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

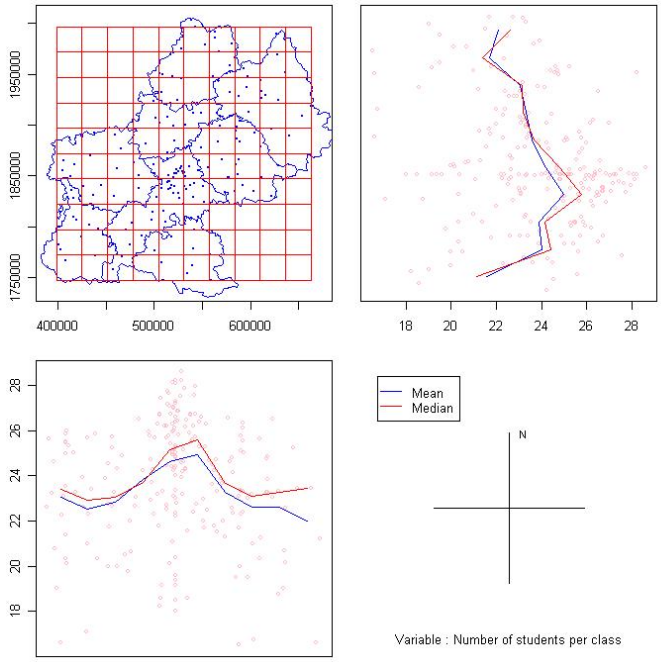

Figure 5: *Drift map* for number of students per class

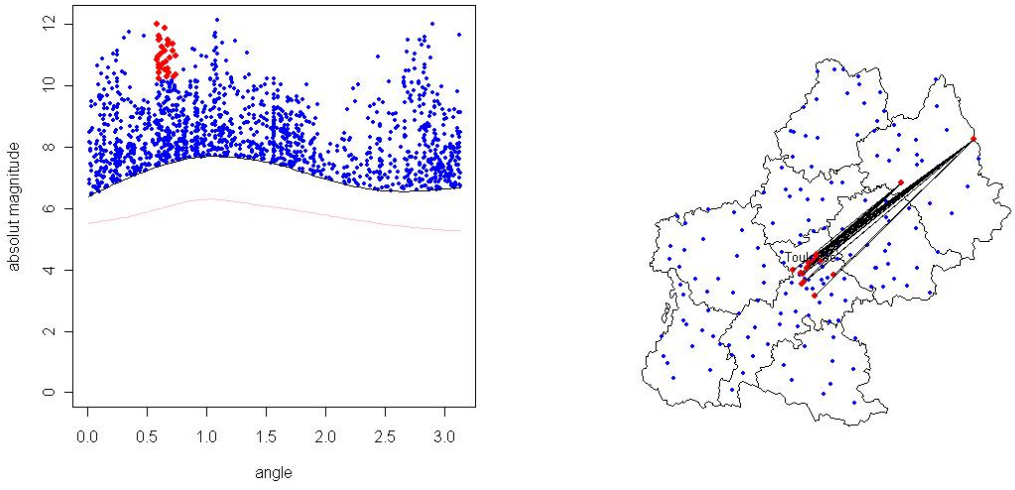

Figure 6: Angle plot for the number of students per class: selection of large absolute differences for a given angle.

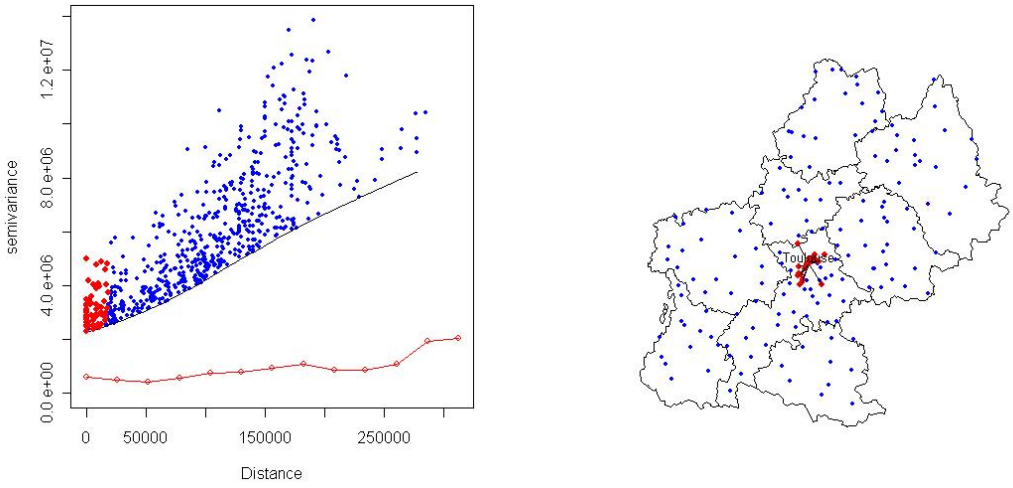

Figure 7: Variocloud for cost per student above the 95^{th} percentile : selection of large absolute differences and small distance.

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 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. On 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*. These two functions use neighborhood or weight matrices of several types, which can be constructed using the auxiliary functions *makeneighborsw* and *makedistancew*, or using other similar functions in the R package *spdep* of Bivand [9]. The functions *makeneighborsw* and *makedistancew* create a weighting scheme based on a given number of nearest neighbors for the first one and a given distance threshold for the second one. The function *normW* performs row standardizing of these matrices if necessary.
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

[9]see the site http://cran.r-project.org/web/views/Spatial.html

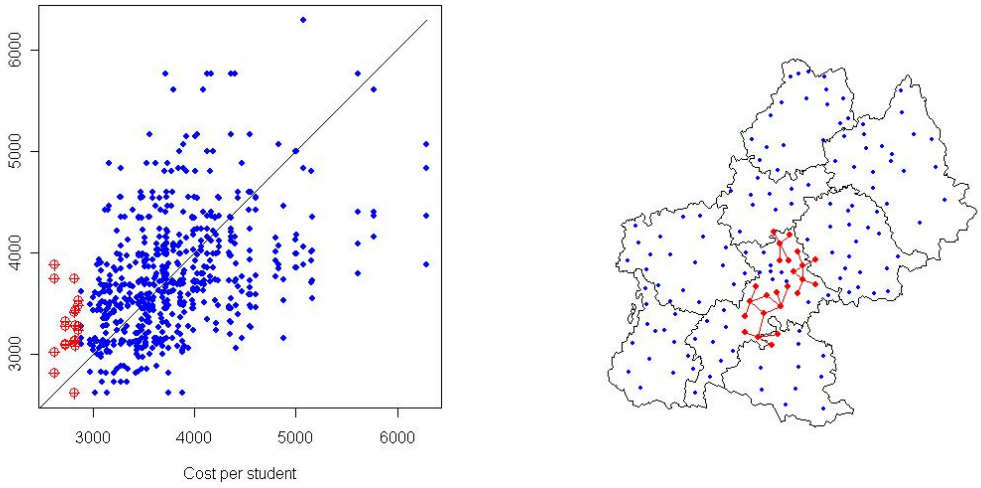

Figure 8: Neighbor plot for cost per student: selection of small costs.

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 on 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 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 on Figure 9. For the same 4 nearest 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 centered 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

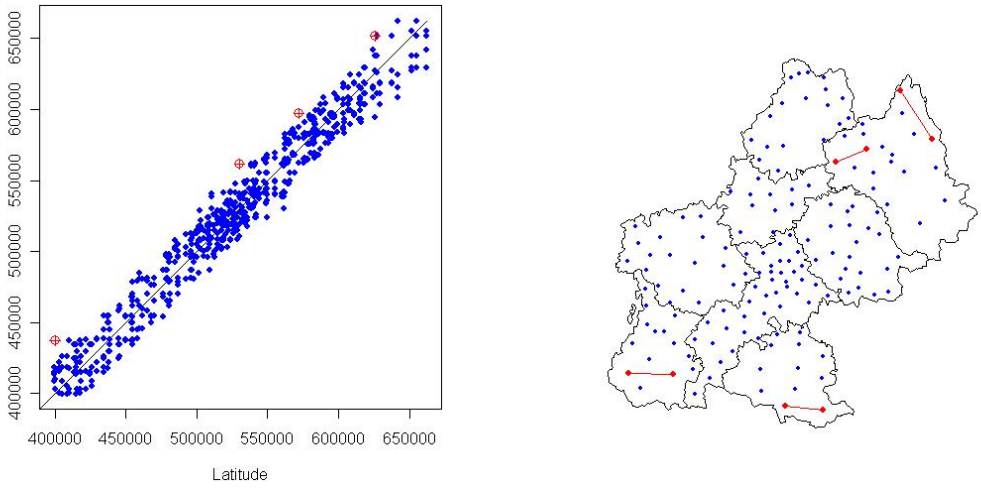

Figure 9: Neighbor plot for latitude: selection of large differences in latitude.

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 cantons with a number of students per class higher than average as well as their neighbors, shows that these are mainly urban 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 et al. (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

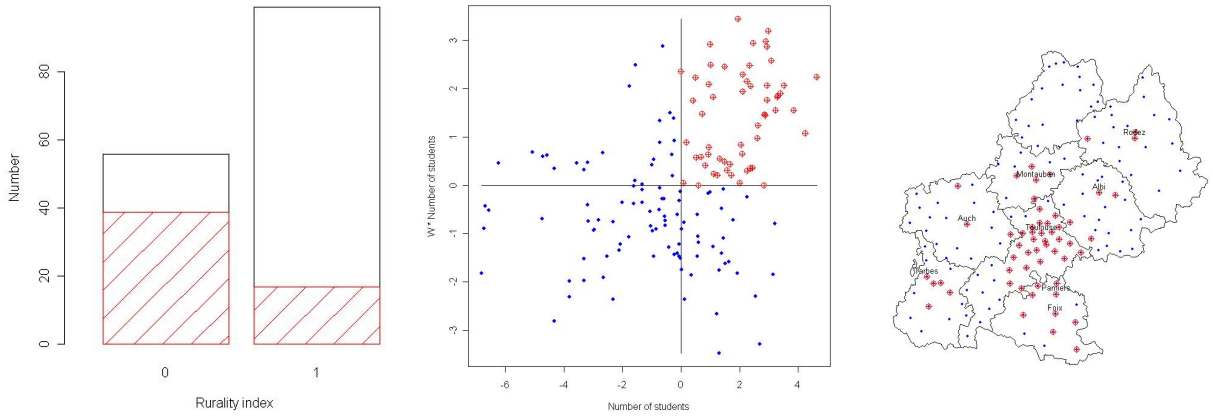

Figure 10: Moran scatterplot of the number of students per class: selection of first quadrant.

them on the map and explore their relative spatial position. Reversely, 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 variables: the mean age of the teachers in the school (*Teachers_age*), the frequency of certifiés teachers[10] (*Freq_certifies*), the frequency of agrégés teachers (*Freq_agreges*), the frequency of students who repeated a class (*Freq_rep_stud*), the number of specialties offered to students in the school (*Nb_specialties*), the number of students per class (*Nb_students_per_class*) and the occupancy rate of the school (*Occupancy_rate*). 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

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

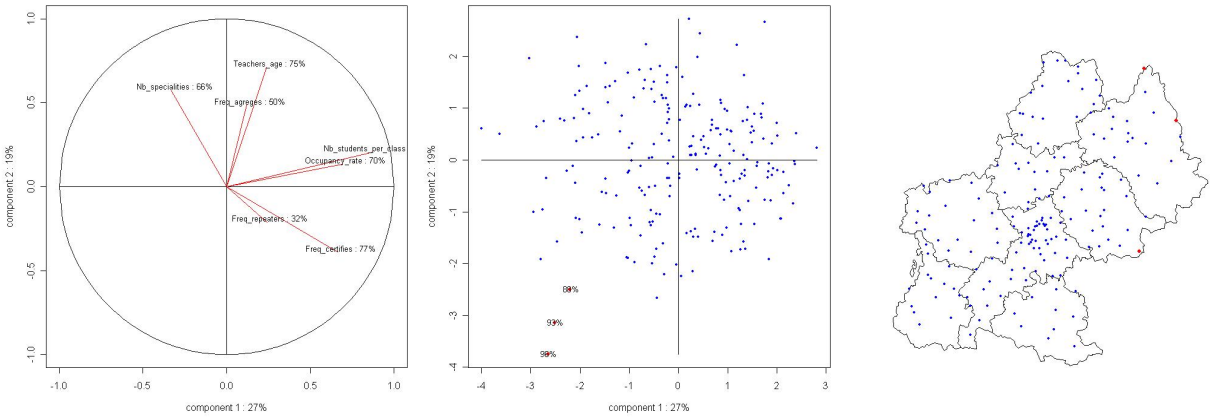

Figure 11: Principal components analysis: selection of the three schools on the left bottom part of the first principal plane.

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 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 a OLS model, a 3-d version of the *scattermap*, a micromap display (see Symanzik and Carr, 2008), an *Apleplotmap* based on Li et al. (2007), etc. More structural changes will involve in the near future the use of R classes for a larger compatibility of formats with the *sp* R package and also of sparse matrices for handling large weight matrices.

Acknowledgments 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 I for their research assistance and many colleagues for their remarks (E. Malin, I. Héba, J. Symanzik., etc.)

8 Annex

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. boxplotmap : links a map and a box and whiskers plot.
4. clustermap : links a map and a bar map of a clustering variable (kmeans method).
5. dbledensitymap : links a map and two density estimators.
6. dblehistomap : links a map and two histograms.
7. densitymap : links a map and a density estimator.
8. driftmap : this function is meant for detecting trends (non interactive)
9. ginimap : links a map and a Gini plot (Lorenz curve).
10. histobarmap : links a map to an histogram and a bar plot.
11. histomap : links a map and an histogram.
12. moranplotmap : links a map and a Moran scatterplot.
13. neighbourmap : links a map and a neighbor plot (scatterplot of variable against variable for the neighboring sites)
14. pcamap: links a map and a scatterplot of principal axes of Principal Components Analysis
15. polyboxplotmap : links a map and a box and whiskers plot.
16. scattermap : links a map and a two-dimensional scatterplot.
17. variocloudmap : links a map and a variogram cloud (only the variogram cloud is active).

References

[1] Anselin, L (1994). "Exploratory spatial data analysis and geographic information systems", in M. Painho ed., *New tools for spatial data analysis*. Luxembourg: Eurostat: 45-54.

[2] Anselin, L (1995). "Local indicators of spatial association-LISA", *Geographical Analysis* 27: 93-115.

[3] Anselin, L and Bao, S (1995). "Exploratory spatial data analysis linking SpaceStat and Arcview". Regional Research Institute, West Virginia University.

[4] Anselin, L (1998). "Exploratory spatial data analysis in a geocomputational environment", in Paul Longley, Sue Brooks, Bill Macmillan and Rachel McDonnell (eds) *GeoComputation, a Primer*. New York: Wiley.

[5] Anselin, L (2003). "GeoDa 0.9. User's guide". Urbana-Champaign, IL: Spatial Analysis Laboratory (SAL), Department of Agricultural and Consumer Economics, University of Illinois.

[6] Anselin, L, Syabri I, Kho Y (2006). "GeoDa: An Introduction to Spatial Data Analysis", *Geographical Analysis* 38 (1), 522.

[7] Bavaud, F (1998). "Models for spatial weights: a systematic look", *Geographical Analysis* 30 (2): 153-171.

[8] 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.

[9] Brundson, C (1998). "Exploratory spatial data analysis and local indicators of spatial association with XLISP-STAT", *The Statistician* 47: 471-484.

[10] Chauvet, P, (1982). "The variogram cloud", Proceedings of the 17th APCOM International Symposium, Golden, Colorado.

[11] Cook, D, Majure, JJ, Symanzik, J and Cressie, N (1996). "Dynamic graphics in a GIS: exploring and analysing multivariate spatial data using linked software", *Computational Statistics* 11: 467-480.

[12] Cressie, N (1993). *Statistics for spatial data*. Wiley.

[13] 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.

[14] Gastwirth, J L, (1972). "The Estimation of the Lorenz Curve and Gini Index", The Review of Economics and Statistics, MIT Press, vol. 54(3), pages 306-16, August.

[15] Haining, R , Wise S, and Ma, J (1998). "Exploratory spatial data analysis in a geographic information system environment", *The Statistician* 47: 457-469.

[16] Haslett, J ,Wills, G and 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.

[17] Haslett, J , Bradley, R, Craig, P, Unwin A and Wills G (1991). "Dynamic graphics for exploring spatial data with application to locating global and local anomalies", *The American Statistician* 45: 234-242.

[18] LeSage J and Pace K, (2004), "Arc Mat, a Toolbox for Using ArcView Shape Files for Spatial Econometrics and Statistics," in Geographic Information Science, Proceedings of the Third International Conference, Max J. Egenhofer, Christian Freksa and Harvey J. Miller (eds.), Lecture Notes in Computer Science, (Springer-Verlag: Berlin), pp. 179-190.

[19] LeSage J (1998). Spatial Econometrics, http://www.econ.utoledo.edu/faculty/lesage/lesage.html

[20] Li H, Calder C A, and Cressie N, (2007)," Beyond Moran's I: Testing for spatial dependence based on the SAR model." Geographical Analysis, 39, 357-375.

[21] Openshaw, S (1994). "What is a gisable spatial analysis", in M. Painho (ed.) *New tools for spatial data analysis*. Luxembourg: Eurostat: 36-44.

[22] Symanzik J, Cook D, Lewin-Koh N, Majure JJ and Megretskaia I (2000). "Linking Arcview and XGobi: Insight behind the front end". *Journal of Computational and Graphical Statistics* 9, pp. 470-490.

[23] Symanzik J, and Carr D. B. (2008). Interactive Linked Micromap Plots for the Display of Geographically Referenced Statistical Data, In: Chen C, Haerdle W., Unwin A (Eds.), Handbook of Data Visualization, Springer, Berlin/Heidelberg, 267-294 & 2 Color Plates.

[24] Unwin AR, Hawkins, G, Hofman H and Siegl, B (1996). "Interactive graphics for data sets with missing values - MANET", *J. Comput. Graph. Statist.* 5: 113-122.

[25] Unwin, A, Unwin, D (1998). "Exploratory spatial data analysis with local statistics", *The Statistician* 47: 415-421.

[26] Wilhelm, A, Steck, R (1998). "Exploring spatial data with interactive graphics and local statistics", *The Statistician* 47: 423-430.

[27] Wise S, Haining R and Ma J (2001). "Providing spatial data analysis functionality for the GIS user: the SAGE project", International Journal of Geographical Information Science 15 (3), pp. 239-254.