Heba, Ines; Malin, Eric; Thomas-Agnan, Christine

Conference Paper

Exploratory spatial data analysis with GEOXP

42nd Congress of the European Regional Science Association: "From Industry to Advanced Services - Perspectives of European Metropolitan Regions", August 27th - 31st, 2002, Dortmund, Germany

Provided in Cooperation with:

European Regional Science Association (ERSA)

Exploratory spatial data analysis with GEOχP

Inès Heba[*], Eric Malin[†] and C.Thomas-Agnan[‡]
Université des Sciences Sociales, Toulouse 1
GREMAQ, 21 allée de Brienne 31042 Toulouse, FRANCE
Université de La Réunion
Saint Denis de LA REUNION

June 21, 2002

Abstract. We present GEOχP, a computer package of Splus and Matlab routines implementing interactive graphics methods for exploring spatial data. We use a data basis from the regional public health insurance agency concerning physicians' activity in the French Midi-Pyrénées region to illustrate the use of these exploratory techniques based on the coupling between a statistical graph and a map. This coupling has been exploited in the literature for elementary plots like boxplots, histograms or simple scatterplots. In order to make the most of the multidimensionality of the data and thus produce more informative graphs, we suggest using a preliminary dimension reduction technique.

Key Words. exploratory analysis, spatial econometrics, interactive graphics, dimension reduction, sliced inverse regression.

1 Introduction

Exploratory analysis of georeferenced data must take into account their spatial nature. Geographic Information Systems are very elaborate cartographic tools but their statistical analysis capabilities are limited and they generally do not incorporate up to date statistical techniques with a spatial component. Openshaw (1994) and Anselin (1998) attemp 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

[*]e-mail : inaalb@yahoo.com
[†]e-mail : eric.malin@univ-reunion.fr
[‡]e-mail : cthomas@cict.fr

(1998) also argue for the use of local measures of spatial association. The aims of exploratory spatial data analysis include describing geographical distributions, identifying spatial outliers, discovering trends or heterogeneity, regimes 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 idea of the coupling is that the user can select a zone on the map and the program will automatically highlight the corresponding points on the statistical graph or reversely select a portion of the graph to highlight the corresponding points on the map.

Haslett et al. (1991) links histograms, double histograms, scatterplot matrices, and varioclouds (to be described below) with the maps using the PASCAL language. Anselin and Bao (1995) implement the methods advocated in Anselin (1994) linking SpaceStat and ArcView. Brundson (1998) implements the scatterplot matrix, the neighbour plot and the angle plot (to be described below) plus some spatial smoothing of maps for trend detection with the XLISP-STAT language. Haining et al. (1998) 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 the cartographic data visualizer (cdv) of Dykes (1998) based on the Tcl/Tk language.

But the need for a more comprehensive and unified tool motivated us to start the development of a set of statistical routines in the Splus and in the Matlab languages adapted to the exploration of georeferenced data. The choice of Matlab is motivated by the existence of the econometric toolbox of Le Sage (1998) and the choice of Splus by its Splus-Spatial toolbox (Kaluzny et al., 1998). For the moment, the emphasis in GEOχP has not been in the quality of the cartographic display but rather in implementing as many different tools as possible and as up to date as possible. We hope that the current low quality of the maps will improve over time. GEOχP has been written by a team of students under the supervision of faculty of the university of Toulouse I. For this paper, we have choosen to present the graphs produced by the Matlab version of GEOχP and to illustrate only a small selection of the different routines because of space limitations.

The database we use to illustrate our description was provided to us by the regional public health insurance agency of the French Midi-Pyrénées region. It contains 80 variables concerning physician's activity for each of 268 zones (called "canton") of the region, together with socio-economic information such as income, unemployment, demographic distribution, education, ...

2 Description of the basic functionalities

This set of functions applies to the analysis of a data set of variables measured on geographical zones such as cities, counties, countries, etc ... called basic spatial units. Moreover for each basic unit, the data set must contain the latitude and longitude of its centroid. For selecting the points either on a statistical graph or on a map, the user can choose between selecting individual points or selecting points inside a given polygon.

Each function returns the indices of the selected points.

2.1 Univariate tools

When the statistical graph is a simple boxplot (as in Haining, 1998), 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. The same information is conveyed by cloropleth maps in a GIS.

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. (1998). 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, we have introduced an option allowing the user to produce two kernel density estimators instead of two histograms. Figure 1 illustrate this method with the variable: female physician's rate. The subregion selected is the surroundings of Toulouse, the main city in Midi-Pyrénées and one sees on the corresponding density graph that the subdistribution of the female physician's rate in this highly populated area is more concentrated than in the whole region and is also shifted to the right with a mode around 3 physicians per thousand inhabitants rather than 1 in the whole Midi-Pyrénées. Figure 2 (resp: 3) illustrate the same method for the variable: physician's fees per patient with a selection of the Haute-Pyrénées region (resp: Ariège region). One of the subdensity is shifted to the right and the other to the left, showing evidence of two different physician's behaviour with respect to fees. The bimodality in the case of Ariège suggest the presence of two subpopulations.

As in Cressie (1993), in order to examine trends in one variable, GEOχP creates a grid of a given fineness and for each square of the grid compute the means of the variable of any basic unit 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. Figure 4 illustrate this method on a demographic variable (rate of aged 15-59) showing that the largest deviations arise between Toulouse and the peripheral areas.

To examine spatial autocorrelation, given a spatial binary weight matrix (Bavaud,

1998) containing information about the neighbouring 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 neighbouring units. Points far away from the diagonal on this plot identify outliers and selection is again possible on the plot as well as on the map.

The variogram cloud is another tool inspired by geostatistics to study autocorrelation. It is a simple scatterplot of the 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. Selection on the map is also possible. An option allows the user to overlay a smooth to this scatterplot thus estimating the variogram function.

2.2 Multivariate tools

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 graph. A simple scatterplot of this couple of variables can also be linked to the map and selection is again possible in both directions as in Brundson (1998). A kernel smooth has been added to the scatterplot for convenience with a flexible choice of bandwidth as can be seen on figure 5. An option allows the user to overlay conditional quantile estimates instead of the kernel smooth which estimates the conditional mean, thus allowing a more precise exploration of the cloud when one is interested in the extreme rather than the average behaviour. For example, figure 7 displays the 0.15 and 0.85 conditional quantiles of the drug prescription per client given the physician's rate. The three selected points correspond to zones with a low number of physician per inhabitant and a high prescription.

For several variables, a scatterplot matrix can be linked to the map but selection of points on one of the scatterplots becomes cumbersome if the number of variables is too large.

3 More advanced functionalities

The simple scatterplot linked to the map has potentials for more advanced investigations if one applies it to transformations of the raw variables.

For example, the scatterplot of a variable X against its spatial lag variable WX for a given weight matrix W is the classical Moran scatterplot (Anselin, 1995). GEOχP links the 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. 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.

Less classically we suggest using a preliminary dimension reduction technique such as principal components analysis or sliced inverse regression 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.

In the case of principal components analysis, one can do a scatterplot of the projection of the cloud for any couple of factorial axes and 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. 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.

In a regression situation where one variable Y is singled out as to be explained by a set of other explanatory variables, sliced inverse regression (SIR) is a technique that estimates a set of linear combinations, called the indices, of the explanatory variables that best explain Y in a possibly nonlinear fashion. The vectors of coefficients of these indices are called e.d.r. directions. A bivariate plot of Y against each of the indices can then be linked to the map. In the case of multivariate Y, SIR also produces linear combinations of Y that are best explained by the e.d.r. directions thus allowing to use the same bivariate scatterplot linked to the map to investigate the relationships of Y and X and its geographic component. Figure 5 illustrate this method on the study of the amount of drug prescription per patient as a function of mean income of the zone, density of population, mean age of physicians in the zone and proportion of patients aged 70 years or more. The corresponding SIR analysis exhibits a first eigenvalue of 0.3 and relatively negligible further eigenvalues favouring a model with a unique e.d.r. Figure 6 shows the scatterplots of the index versus all the explanatory variables. On the scatterplot of the prescription per patient versus the index, we can for example select the points corresponding to low values of the index as is shown on figure 8 and we see the corresponding zones on the map. It is then interesting and fast to describe how the selected group differs from the rest by just comparing means of each variable on the database: compared to the whole region, the selected group exhibits lower income, lower density of population, lower number of physicians, much lower rate of female physicians and higher amount of prescription per patient. From the map and the description, it corresponds to some rural areas.

Other dimension reduction techniques could similarly be coupled with the map, for example correspondence analysis or projection pursuit.

This set of routines will soon be downloadable from our website: http://www.univ-tlse1.fr/GREMAQ/Statistique/index.htm.

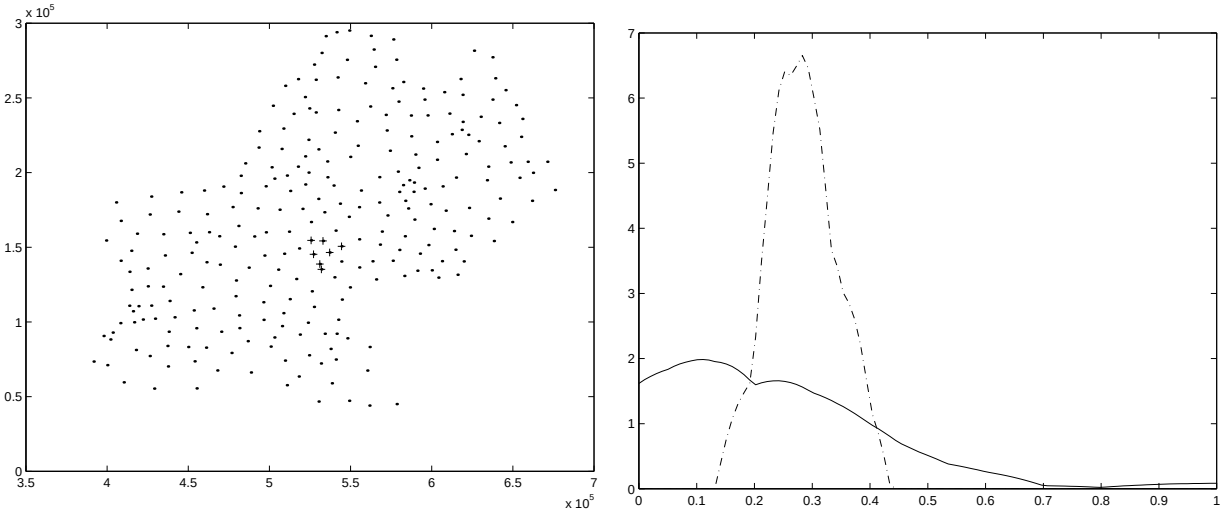

Figure 1: Map of Midi-Pyrénées: selection of the Toulouse region with the corresponding density and subdensity of the female physician's rate

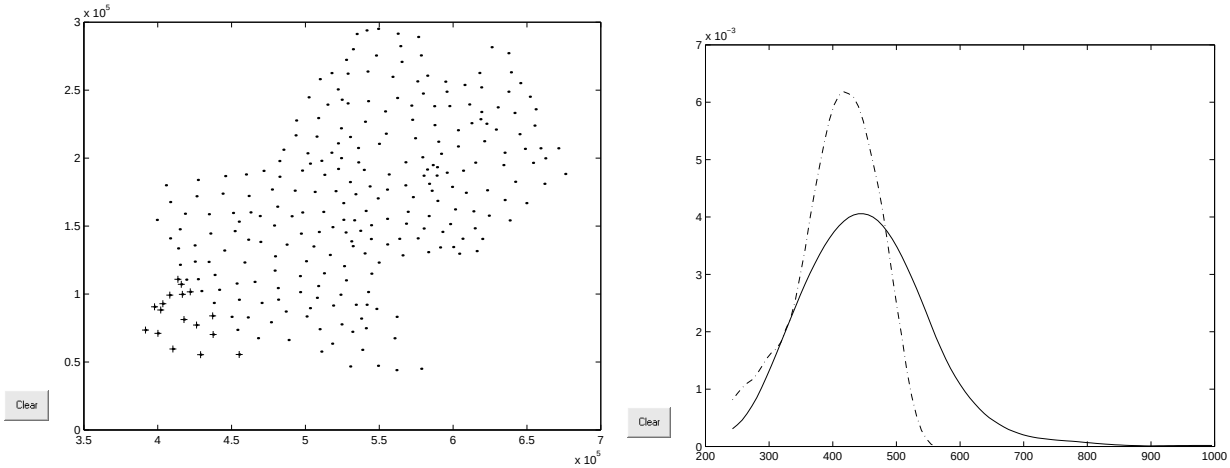

Figure 2: Map of Midi-Pyrénées: selection of Hautes-Pyrénées with the corresponding density and subdensity of the physician's fees per patient

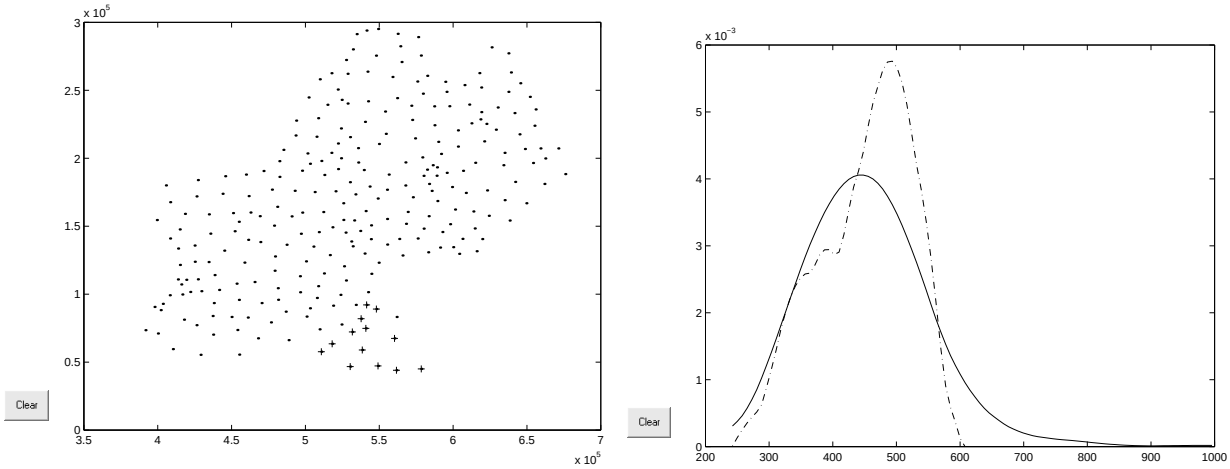

Figure 3: Map of Midi-Pyrénées: selection of Ariège with the corresponding density and subdensity of the physician's fees per patient

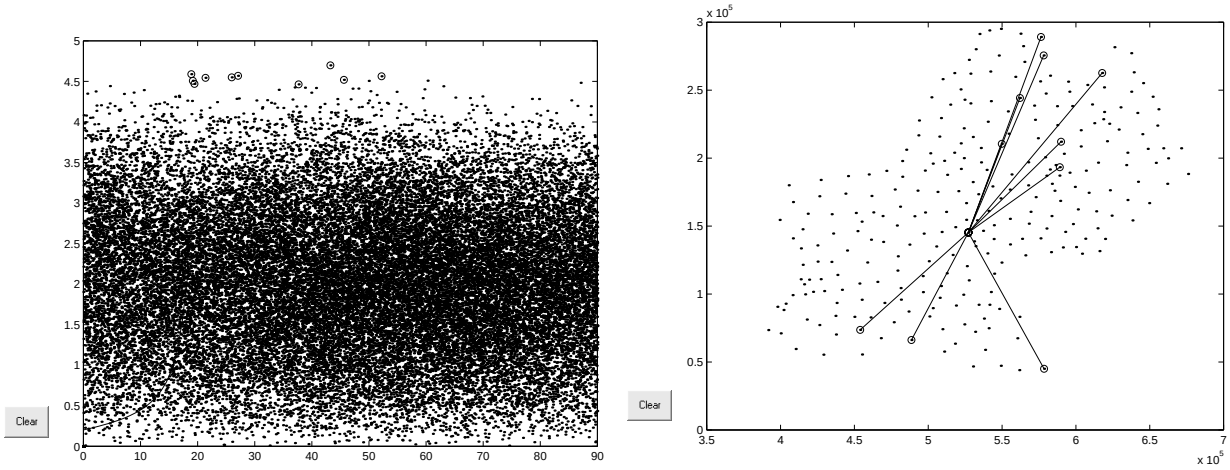

Figure 4: Angle plot of the age 15-59 rate: selection of highest deviations with corresponding points on the map

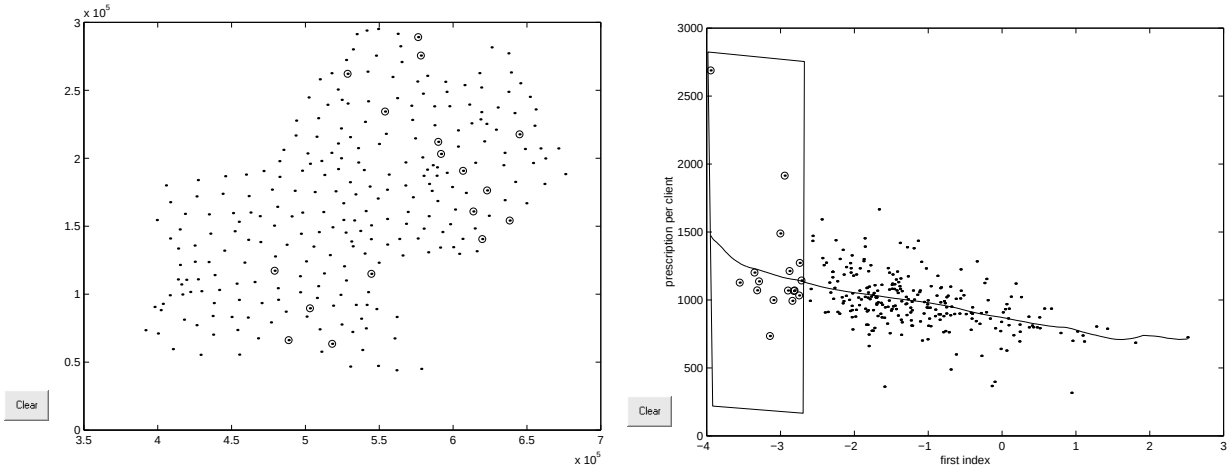

Figure 5: Scatterplot of the prescription per patient versus the first SIR index: selection of points with a low index and corresponding points on the map

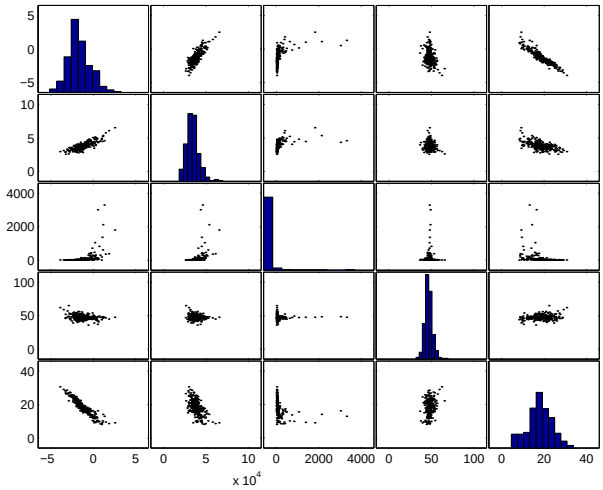

Figure 6: Scatterplot matrix of the first SIR index and the explanatory variables

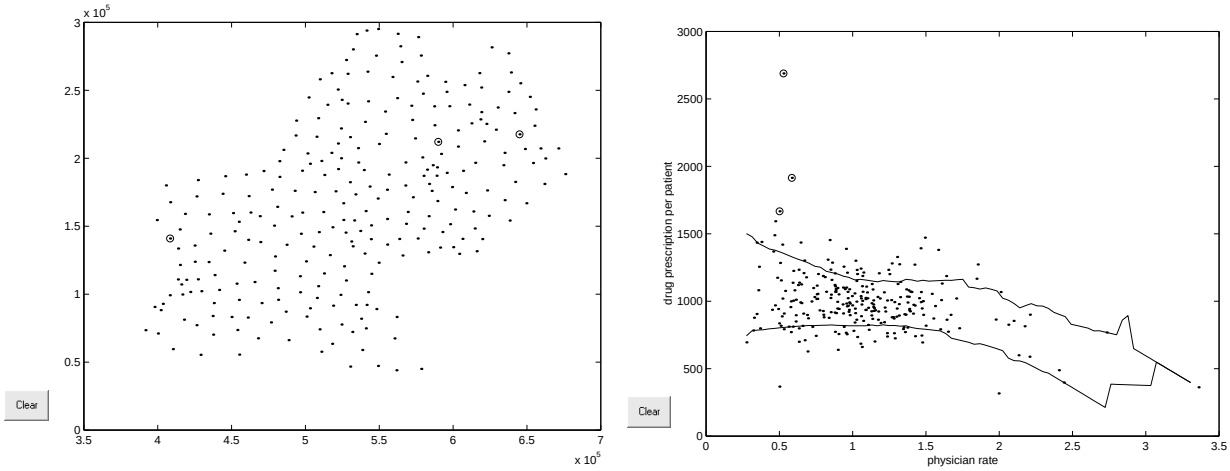

Figure 7: Scatterplot of the drug prescription per patient versus the physician's rate with 0.15 and 0.85 conditional quantiles

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, pp. 45-54.

[2] Anselin L. (1995), Local indicators of spatial association-LISA. Geographical Analysis 27, pp. 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, 1998: 7794.

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

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

[7] Cook , D., Majure J.J., Symanzik J. and Cressie N. (1996), Dynamic graphics in a GIS: exploring and analysing multivariate spatial data using linked software, Comput. Statist., 11, pp. 467-480.

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

[9] Dykes J. (1998), Cartographic visualization: exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv. The Statistician 47, pp. 4851-497.

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

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

[12] Kaluzny, S., P., Vega, S., C., Cardoso, T., P. et Shelly, A. A. (1998). *S+SpatialStats, User's Manual for Windows and Unix*, Springer.

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

[14] Li K.C. (1991) Sliced inverse regression for dimension reduction, with discussions, J.A.S.A. 86, pp. 316-342.

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

[16] Openshaw S. (1994), A framework for research on spatial analysis relevant to geostatistical information systems. in M. Painho ed., New tools for spatial data analysis, Luxembourg, Eurostat, pp. 157-162.

[17] Unwin A. and D. Unwin (1998), Exploratory spatial data analysis with local statistics. The Statistician 47, pp. 415-421.

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