Haining, Robert; Wise, Stephen; Signoretta, Paola

Conference Paper

Providing scientific visualisation for spatial data analysis: criteria and an assessment of SAGE

38th Congress of the European Regional Science Association: "Europe Quo Vadis? - Regional Questions at the Turn of the Century", 28 August - 1 September 1998, Vienna, Austria

Provided in Cooperation with:
European Regional Science Association (ERSA)

PROVIDING SCIENTIFIC VISUALIZATION FOR SPATIAL DATA ANALYSIS: CRITERIA AND AN ASSESSMENT OF *SAGE*

Robert Haining and Stephen Wise

Dept of Geography, and the Sheffield Centre for Geographic Information and Spatial Analysis; University of Sheffield

Paola Signoretta

Sheffield Centre for Geographic Information and Spatial Analysis; University of Sheffield

Abstract:

A consistent theme in recent work on developing exploratory spatial data analysis (ESDA) has been the importance attached to visualization techniques, particularly following the pioneering development of packages such as SPIDER and REGARD (Haslett et al 1990).

The focus on visual techniques is often justified in two ways: (a) the power of modern graphical interfaces means that graphics is no longer a way of simply presenting results in the form of maps or graphs, but a tool for the extraction of information from data; (b) graphical, exploratory methods are felt to be more intuitive for non-specialists to use than methods of numerical spatial statistics enabling wider participation in the process of getting data insights.

This paper briefly reviews a theoretical framework that has been suggested for developing visualization tools for ESDA that comprises two elements: (a) a data model, based on the distinction between rough and smooth properties of spatial data, that defines what an analyst is looking for in data (Haining et al 1998) and (b) a theoretical model for assessing the quality of visualisation tools (Cleveland 1994). The emphasis of this paper is the use of the theoretical framework to structure an assessment of SAGE, a software system that has been written for the spatial statistical analysis (including both exploratory and confirmatory data analysis) of area based data linked to a GIS. The aim of the assessment is to identify and illustrate what appear to be desirable features of the system (that might be employed in other systems) whilst also highlighting what the additional requirements are thereby contributing to the development of systems for ESDA that contain good quality scientific visualization tools for exploratory spatial data analysis.

Key words: data rough; data smooth; table look up; pattern perception; interactive graphics.

1. Introduction

Exploratory spatial data analysis (ESDA) extends exploratory data analysis to spatial data (Fotheringham et al 1994). The aims of ESDA are descriptive, rather than confirmatory and seek to detect patterns in spatial data, to formulate hypotheses which are based on, or which are about, the geography of the data and to assess spatial models (Haining et al 1998). (The latter is associated with the later stage of data analysis when the analyst is exploring model fits.) These aims are in addition to the general aims of exploratory data analysis which include identifying interesting or unusual features in the data (including detecting possible data errors) and distinguishing accidental from important features (Tukey, 1997; Hoaglin et al 1985). The techniques that are employed in ESDA are both visual and resistant. Resistant techniques are those where results are not greatly affected by a small number of unusual or aberrant values. Visual techniques include those that employ charts, graphs, figures and in the case of ESDA, crucially, maps. ESDA includes EDA in the sense that spatial data, which comprise attribute values with associated locational identifiers, at one level can be explored without reference to where data values occur on a map. As a final quality of ESDA techniques, they usually "stay close" to the original data in the sense of either working with the original data or only employing simple intuitive transformations of the data (Unwin 1996).

Visualization plays an important role in ESDA. The availability of a map enables the analyst to ask the question "where are those cases on the map?" or "where do attribute values from this part of the map lie in the data summary?" Visual tools are usually easier to interpret so that they make it possible for a wider group of researchers (not just specialist data analysts) to participate in the process of drawing out insights from data. The emphasis here is on scientific visualisation, as opposed to presentation graphics. The later, as its name implies, is for the presentation of data and usually focuses on the best way of presenting a single, static view of data to a user who is not necessarily familiar with the data. Scientific visualisation, by contrast is concerned with providing multiple, dynamic views of the data where the user may already know much about the data (they may have collected it or been responsible for its collection) and are progressively learning more about the data as they use the visualisation tools.

Wise et al (1998) proposed a conceptual framework for implementing scientific visualization in ESDA. This framework included a conceptual model for ESDA (Haining et al 1998) and a conceptual model for evaluating visualization tools. The conceptual model for ESDA drew on the conceptual model for EDA that is:

$$\text{DATA} = \text{ROUGH} + \text{SMOOTH}$$

The "smooth" and "rough" elements of the model can refer to just the set of attribute values of the dataset. In this case a non-spatial "smooth" property includes the central tendency of the distribution (measured by the median), its dispersion (measured by the inter-quartile range) whilst an overall representation of the data can be captured by a boxplot. Non-spatial "rough" properties are cases which are a certain distance from a defined "smooth" element such as the median. An outlier is a case with a particularly high level of "rough" as measured by distance above the upper or below the lower quartile.

When the locational reference is attached to each case then "rough" and smooth" can be defined in terms of *where* on the map the cases are found. "Smooth" spatial properties include spatial trend and spatial autocorrelation which are global (or whole map) spatial

properties. "Rough" spatial properties are local (case specific) properties such as spatial outliers, cases that are very different from their neighbouring values, local spatial clusterings of high or low values, or even lines of discontinuity. Local "rough" properties might be identified by applying techniques that sweep through all defined subsets of data, one at a time, or by working on a sub area of the map. Note that cases that are "rough" by non-spatial criteria need not be "spatially rough". For example if we define all cases more than a certain distance above the best fit regression line as possessing some element of "roughness", the geography of these positive outliers may be smooth (for example if they are all located in one area of the map).

The purpose of this conceptual model is to specify what it is the analyst might be looking for in a dataset. It is against this yardstick that the set of available tools that are provided in any piece of software might be judged.

Wise et al (1998) drew on the conceptual model of Cleveland (1994) for evaluating visualization tools. According to Cleveland there are two main activities associated with reading a statistical graph and each of these involve three tasks. Cleveland's model attempts to classify the mental processes that are undertaken by the reader. The first activity is *table look up* which is the process by which an individual retrieves from the graph, for any individual case, the data about the real world that has been encoded by the graph. This involves one or more of the following tasks: scanning (relating the case to the axis), interpolating (estimating the value of the case from the tick marks on the axis) and matching (linking the case symbol back to the key). The second activity is *pattern perception* which is the process by which collections of cases and in particular the geometry of those assemblages of cases encoded on the graph are read in order to identify patterns in the data and draw out useful information. The three tasks associated with pattern perception are, for any given visual tool, detection (recognizing what the graphical object is showing in terms of the symbols used and the geometric segments contained within the graph), assembly (grouping geometric segments) and estimation (identifying the properties of the segments which ranges from identifying that two or more graph segments are different through to estimating quantitative measures of the difference). Cleveland provides illustrations of these tasks on a wide range of graphs (Cleveland 1994). The success of any particular graphic tool often rests on how effective it is in terms of helping the user to pick out particular types of information ("fitness for purpose") rather than being a generic statement. If *table look up* makes the link back from the graph to the real world, *pattern perception* makes the link forward from the graph to the extraction of information from the data according to what types of information the analyst is looking for - in the case here, the patterns identified in the data model for ESDA.

SAGE (**S**patial **A**nalysis in a **G**IS **E**nvironment) is a software system that was developed to provide spatial statistical analysis tools in a GIS (Arc/Info) environment with particular reference to health data. It includes both exploratory and confirmatory tools (Ma et al 1996). Although some of the tools are particularly appropriate to the analysis of health data, most tools are appropriate for general forms of area-based data analysis - what Cressie (1991 p7-10) classifies as "lattice" data where the regions that partition the map may be irregular in shape. SAGE was built, wherever possible, using existing well tested, software. The processes of data input, data management and data analysis are provided within the system. The fundamental input to SAGE is a polygon coverage containing the locations of area boundaries, area centroids, topological relationships between areas and attribute values for each area. Figure 1 shows SAGE with all four types of window open: the table window displaying the current set of data and any new variables created during the session; the map window; the graph window (more than one can be opened at any one time) and the text output

window that returns statistical output such as model parameters. Note that the linked window facility is being used. The outlier cases in the graph window have been brushed and they are highlighted in the table and map windows. If other graph windows were open these brushed cases would be highlighted in those windows also.

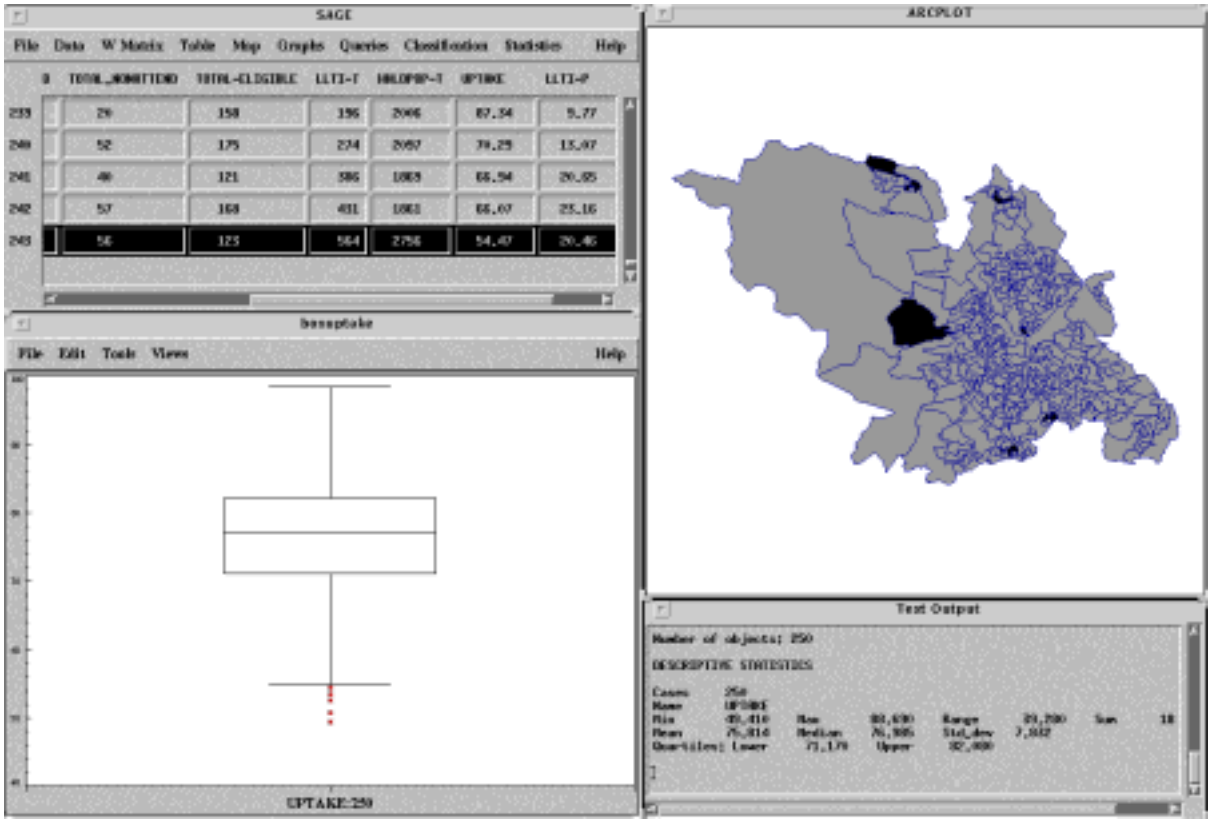

Figure 1. SAGE displaying the four types of windows and the linked windows facility.

The rest of the paper is an evaluation of SAGE against the conceptual framework. We might term this an evaluation of the software in terms of "fitness for purpose". In section 2 there is an assessment of the range of visual tools provided by SAGE that are relevant to ESDA whilst in section 3 there is an assessment of the effectiveness of the visual tools in SAGE in terms of table look up and pattern perception. The purpose of the paper is to contribute to the development of software for effective spatial data analysis focusing here on the visual elements of such software for ESDA. This contribution is made by displaying and evaluating the particular features of SAGE - its strengths and weaknesses. Note that SAGE contains features other than tools for ESDA and details for accessing the software can be found in the appendix.

The illustrations in section 3 are based on data relating to the uptake of the breast cancer screening service in Sheffield. Enumeration district level data (there are 1159 EDs in Sheffield) have been aggregated into approximately 300 areas so that the illustrations can be seen in the prints here. The grouping (implemented in SAGE) was done on the basis of grouping EDs according to similarity of Townsend deprivation score whilst also trying to create areas of similar population size and with a secondary requirement of areal compactness (for details see Wise et al 1997). The speed of SAGE in processing the operations described here is very rapid and remains so even at the ED level.

2. Visual statistical tools in SAGE relevant to ESDA

The assessment of software for statistical analysis is only partly in terms of the range of tools provided, what is more important is whether they enable the user to carry out a coherent analysis. Here "coherent" means that the software provides a set of tools that enables the user to explore the data for the presence of the various elements of "rough" and "smooth". This in turn raises two questions. First, are sufficient tools provided for identifying the different elements and second, are they the best tools or are better ones available? It is unlikely that there will be full agreement on what the best tools are for any particular purpose but some general observations can be made particularly with respect to what is provided in SAGE. First we identify the tools that are provided in SAGE and then comment on them.

Table 1 shows the visual tools provided by SAGE for displaying properties of the data and which are traditionally used to identify "rough" and "smooth" properties of data disregarding the locational identifier. All SAGE tools can be implemented on the entire data set or on user defined subsets. These subsets can be specified by reference to spatial (all areas within or overlapping a box, circle, polygon, or user defined area-by-area selection) and/or logical rules. The histogram plot can be implemented by dividing the data into equal intervals or using the mean around which to pivot the plot in which case the intervals refer to half standard deviations. The user can select the number of bars. Variables that have been aggregated (using one of the classification or regionalization tools implemented in SAGE (Wise et al 1997)) can be plotted as aggregates using any one of several aggregation methods (mean, median, standard deviation, inter-quartile range or sum) where these are valid operations to perform on the raw data. In the case of boxplots, box widths are proportional to the number of cases used to construct the plot and the number is also recorded in the legend below each plot. Several plots can be placed side by side and at the same scale in any one window. If an entire boxplot is highlighted all cases are highlighted giving the impression of a density plot of cases. The rankit plot is available as a visual test for normality. The matrix scatterplot displays multiple pairwise plots within a single graph window. This plot is used in conjunction with the brushing facility in order to visually explore conditional associations between variables.

Table 1: Visual tools for non-spatial analyses in SAGE

Univariate visual tools	**Bivariate visual tools**
Histogram; Boxplot; rankit plot.	scatterplot; matrix scatterplot

Several tools are provided in SAGE to help the analyst identify spatial "rough" and "smooth" properties of a single attribute with a locational identifier and these are listed in table 2. First order, or trend properties of a data set can be explored through any of several tools. SAGE provides a "trellis-like" plot in which the categorical variable is spatial lag order (as defined by any one of the automatically or manually constructed spatial connectivity matrices in SAGE (Haining 1990)). After the user has selected an origin zone SAGE automatically generates a sequence of box plots (one for each lag order) at increasing distance from the origin zone up to a user specified maximum. We term this a spatially lagged boxplot. Anisotropy can be explored by selecting spatial subsets of the data, as described above, from the map window. A suite of smoothing operators provided in SAGE can also be

used to look for trend. Then, because results from these smoothing operators are stored as new columns in the SAGE table window they can then be used to mathematically extract the smooth element from the original attribute data to leave the second order and rough properties of the data. Three smoothers are provided in SAGE, a mean smoother, a median smoother and a relative-risk smoother (Haining et al 1998). The smoothed values can of course be displayed in the map window and since they are new variables in the table window can be subject to any of the other techniques available in SAGE.

Table 2: Visual tools for spatial analyses in SAGE

First Order Properties	**Second Order Properties**
spatially lagged boxplot; median smoothed map; mean smoothed map; relative risk map	Moran plot

Second order properties of spatial data include spatial autocorrelation and concentration (Cliff and Ord 1981; Getis and Ord 1992). The global and local Moran and Getis-Ord statistics are not (in the definition used here) exploratory tools and not, of course, visual tools. The Moran *plot,* however, is available in SAGE. This is a plot of attribute value on the vertical axis against the average of the attribute values in the adjacent areas using a row standardised form of a selected connectivity matrix. A scatter of values sloping upward to the right is indicative of positive spatial autocorrelation whilst if the scatter is downward to the right it is indicative of negative spatial autocorrelation(Haining, 1990). This tool can be used either on the raw data or on data processed in ways described above.

The suite of visual tools provided in SAGE can be broadly grouped into three categories: those that prioritise attribute value similarity in the construction of the graphic (like the histogram or the boxplot); those that prioritise spatial proximity (the map); and those that are hybrids of these two forms (the spatially lagged boxplots and Moran plot). There appears to be a sufficient suite of visual tools to explore spatial data for the purpose of evaluating rough and smooth properties of data. There are however a number of deficiencies.

First, there are a number of tools that could usefully extend the set provided in SAGE (some of which are available for example in other software systems (Wise et al 1998). There is no effective facility to handle missing values, as for example in MANET (Unwin et al 1996). There is no facility for a general trellis plot that would allow the user to explore differences in an attribute with respect to a second categorical variable (also available in MANET for example). Whilst there is a facility for constructing cross-tabulations for categorical variables there is no mosaic plot for visualising such a table (Riedwyl et. al. 1994; Friendly, 1995). Whilst resistant statistics (medians, quartiles) are generated in the text window and resistant smoothers are provided (the spatial median smoother) no resistant tools are provided to summarize bivariate relationships such as resistant best fit lines through scatterplots or Moran plots and instead least squares fits are provided. Finally there are a number of plots that have been proposed to explore spatial data properties that are not included here (for example Chauvet's cloud plots and Cressie's square root difference plot) which can be used to explore second order properties, spatial discontinuities and non-stationarities (Cressie 1984, Haining 1990).

Second, the effectiveness of some of the tools would be enhanced if they could be made dynamic (Haslett et al 1991). The shape of a histogram depends on the selected bin

size. The ability to dynamically vary bin size and observe the changing behaviour of the histogram is a facility that is available in MANET for example (Unwin et al 1996, 1997). A related question is to examine whether the same feature could be made available for the regionalisation and classification tools provided in SAGE (Wise et al 1997) that is, with respect to the raw area-based data itself. This may be too ambitious given the time that it takes to implement even relatively simple regionalisation algorithms but if an areal aggregation tool could be provided and made dynamic then as areas were grouped the user would be able to see the effects of this on various graphs and summary statistics - effectively allowing the user to dynamically explore the influence of the areal framework. This could be combined perhaps with other modifications to the visual tools that indicate the (changing) robustness of different values as the areal framework changes. Whilst brushing is available, dynamic brushing (dragging a window or transect over a map and observing the changing behaviour of statistical summaries) is not available. In fact dynamic brushing is impossible to implement with adequate speed of response in SAGE's client-server architecture. Dynamic brushing using an area (circle, box, polygon) or a point would help in the exploration of local map properties and local changes in these properties (Craig et al 1989). These comments are illustrative of a general observation which is that scientific visualisation tools for exploring the spatial properties of spatial data are still in their infancy.

3 Effectiveness of the visual tools in SAGE

This section addresses the question of the implementation of visual tools in SAGE and in particular how effectively different visual tools have been implemented in order to facilitate table look up and pattern perception.

3.1 Table look up.

The tabular and graphical SAGE windows both contribute to the table look up operation. In addition, the linkage between these two types of window using the brushing facility, greatly enhances the usefulness of these tools for this operation.

The table window contains all the current data, both the original data input at the start of the session and any new data generated during the session and which can be saved at the end of the session. Only new data generated during a session can be modified during a SAGE session to ensure that the analyst does not inadvertantly generate modified versions of the original dataset. Areas on the map or cases on a graph that are brushed are highlighted in the table enabling the analyst to immediately see the original data values. The whole row is highlighted so the individual value that may have been brushed in a graph is not picked out in isolation from the other data for that case. Scrolling may be necessary to see all the highlighted cases and there is no quick facility to see all the values that have been highlighted. The table facility used in conjunction with brushing provides the most immediate facility for linking back to the source data which is the fundamental objective of the table look up operation.

The graphical window contains a variety of options that can be turned on or off and which facilitate other aspects of table look up when it is sufficient for the user to identify cases (perhaps many cases) rapidly but with less accuracy. SAGE allows the user to switch grid lines on or off to ease value identification, graph windows can be re-sized which rescales the plot, the user can zoom into or out from all or any segment of a graph window to assist the task of interpolation or to assist the process of identifying individual cases so they can be brushed and their exact values (from the table) or exact locations (on the map) can be determined. It is also possible, as noted above to zoom in and out within the map window to

help identify areas. Figure 2 shows a boxplot window on the left and then the effects of zooming in on the bottom segment of the boxplot where there are outliers. (A second window has been created to illustrate the effect of this.) The lowest value has been highlighted on the boxplot so that the exact case and its location have been highlighted in the table and map windows. The map zoom operation has also been used.

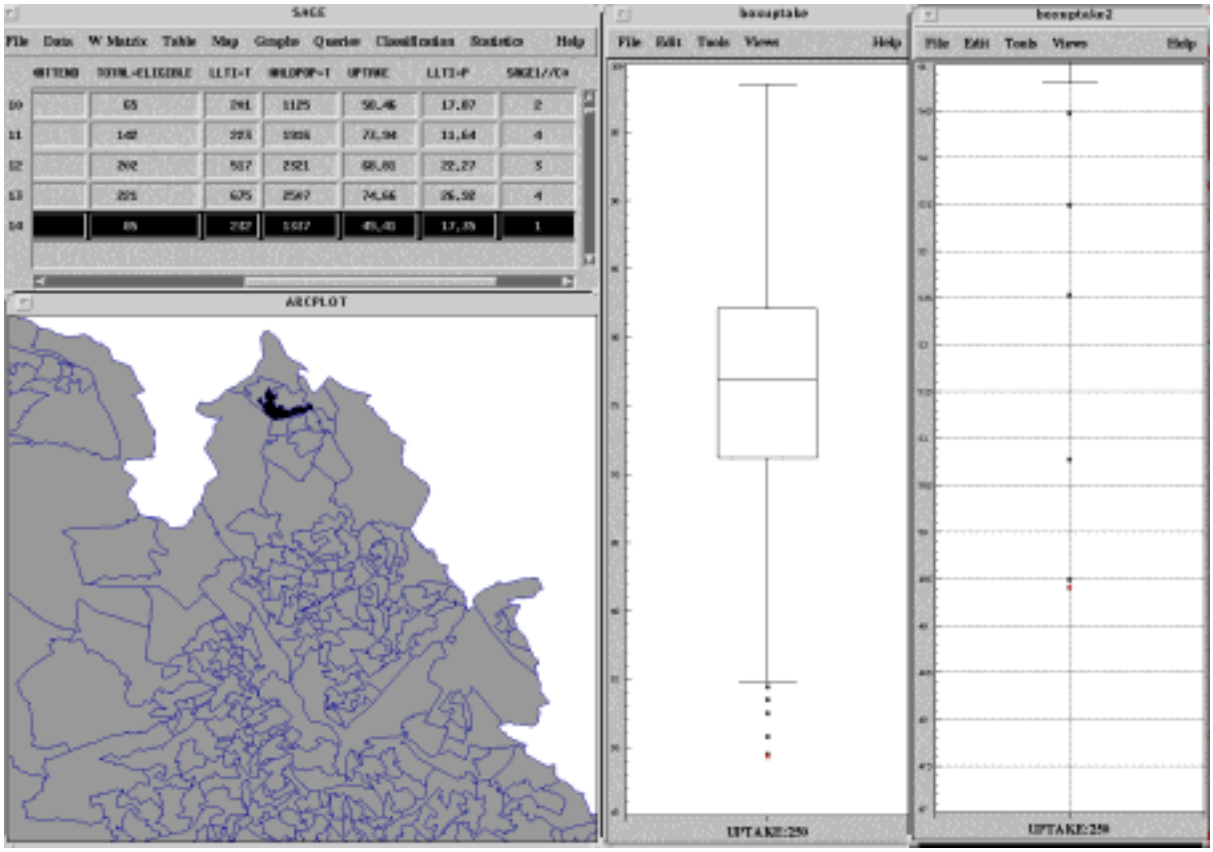

Figure 2. Boxplot features that facilitate table look up

Figure 3 shows the options for the scatterplot. The plot is area uptake rate on the horizontal axis against percentage suffering limiting long term illness on the horizontal axis. There is wider evidence of a (negative) relationship between general levels of health and the extent to which individuals adopt preventative behaviours and this is broadly supported in the case of this dataset. The graph on the right has been implemented with the grid lines switched on and the legend switched on to indicate which variables were used for the plot (although from the design of this it may not be immediately clear to the user which of the variables is on the vertical axis and which is on the horizontal axis). It is important to stress that in scientific visualisation, unlike presentation graphics, there may be times when these facilities are helpful and other times when they are unnecessary hence the desirability of being able to switch them on or off.

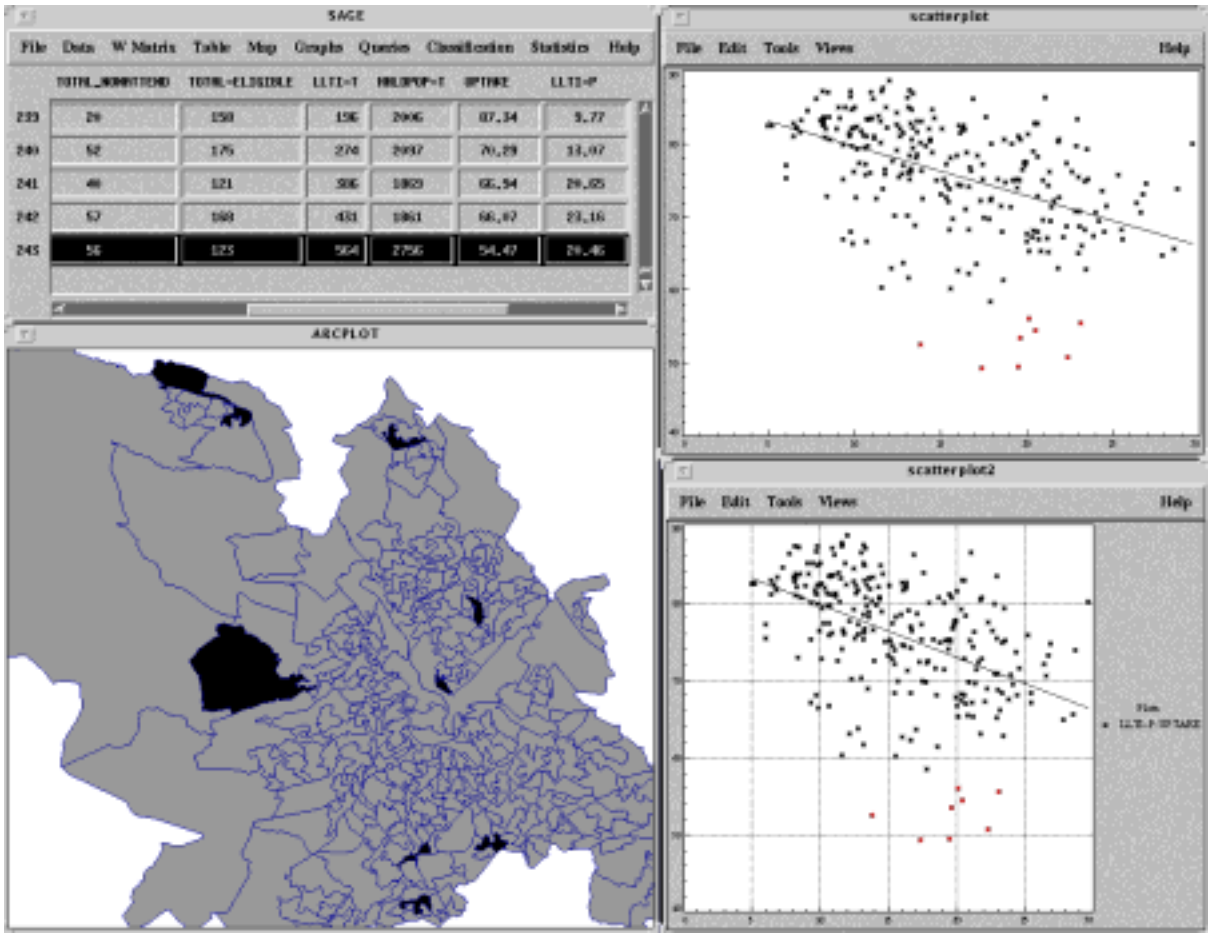

Figure 3 Scatterplot facilities in SAGE with highlighting of the areas with low uptake rates

3.2 Pattern Perception

We focus here on the effectiveness of the visual tools for detecting spatial patterns. The facilities provided by SAGE involve the use of the map window and certain graphic tools described above. The issue here is how effectively have the tools in SAGE been implemented for the purpose of extracting information (in the form of rough and smooth properties of the data) from the encoded data where the encoding is in the form of a graph or a map? We illustrate these operations using boxplot and scatterplot tools. Both of these tools are familiar for pattern perception with non spatial data and their general appropriateness for a wide range of information gathering tasks is not in question. Researchers are generally familiar with what is involved in extracting meaning from these plots. The new interpretative roles here concern what information they can give about the rough and smooth *geography* of data values.

Figure 4 illustrates the use of the spatially lagged boxplot centred on the location of the single breast cancer screening unit in Sheffield. Nine spatial lags have been chosen and the aim is to explore whether there is any tendency for uptake rates to decline with increasing distance from the unit. Grid lines have been switched on to assist comparison of segments across the box plots. Since box widths are proportional to the number of areas at the given lag from the area containing the screening unit this further helps the user to decide on the reliability of comparisons. Because of the nature of the areal framework the link between spatial distance and spatial lag may become tenuous at some lag order. The highlight facility is useful in addressing this concern. The user can select and highlight any boxplot and see the

geography of the cases that are included in the construction of that particular plot in the map window. In the illustration the lag three boxplot has been highlighted and suggests that there is still a reasonably close relationship between order and distance. By the time lag 5 is reached the scatter of areas is much greater because some of the larger areas that stretch out to the edges of Sheffield start to influence the construction of the plots. There is little or no evidence that median uptake rates fall with distance although there are issues of comparability which are discussed below.

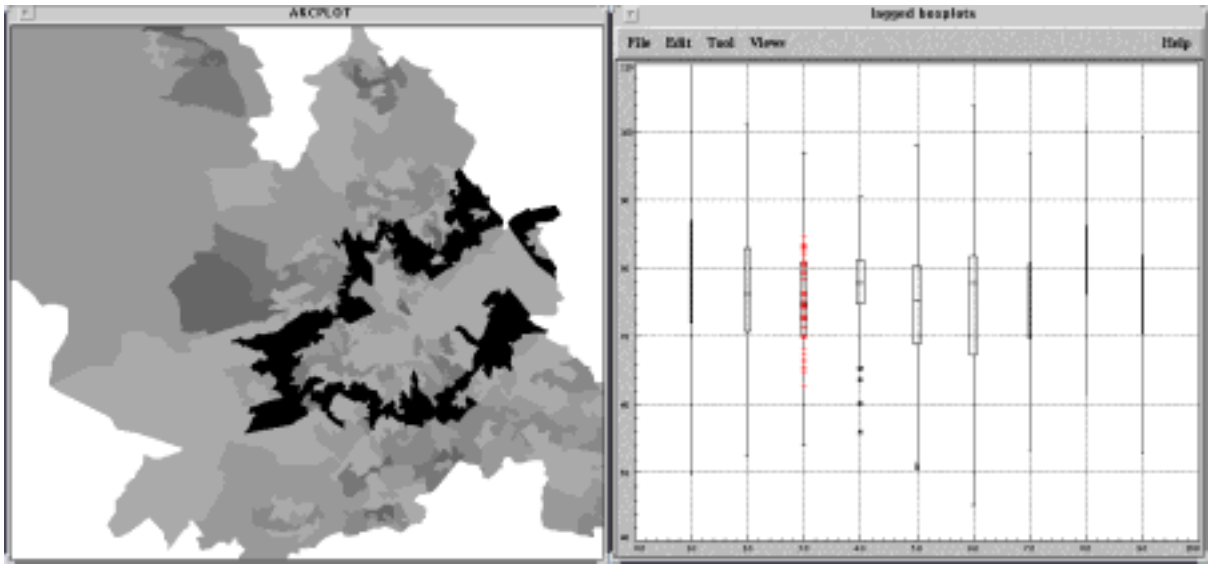

Figure 4 The lagged boxplot facility in SAGE with areas at lag order 3 from the screening unit highlighted

Figure 5 shows the application of the Moran plot with the grid and legend facilities switched on. Uptake rate is plotted on the vertical axis whilst the weighted average of the rates in the adjacent areas (using a row standardised binary connectivity matrix) is plotted on the horizontal axis. There has been no attempt to remove trend (the evidence from figure 4 is not sufficiently persuasive) so the plot is based on the original uptake rates. The evidence from the scatter, without any further analysis is that there is some evidence of positive spatial autocorrelation in uptake rates. The plot includes a least squares best fit line through the scatter to visually assist the user although there are some technical problems with this line (see Haining 1990 p.214) and the line cannot be switched off - which means it can be hard to ignore it!

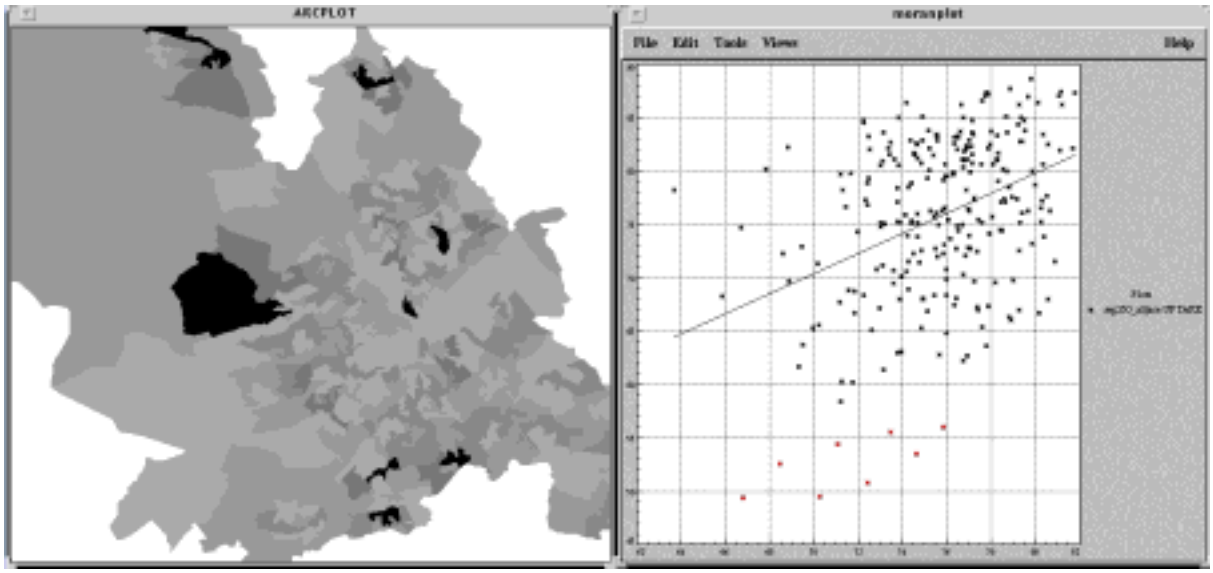

Figure 5 Moran plot facility in SAGE with areas with low uptake rates relative to rates in adjacent areas ("spatial outliers") highlighted.

There is some evidence of areas with low uptake rates that are adjacent to encircling areas that on average have high uptake rates. There are two questions addressed by the illustration. First, where are these areas and second are they widely scattered or are they relatively close together? There is a set of eight areas selected by the brushing tool in the lower portion of the graph that can be seen in the map window. They are widely scattered over the map. (Note that the user has again employed the zoom facility to help with identification.) The areas may deserve closer investigation as to why they have relatively low rates (they are low in distributional terms - outliers - as well as low relative to their neighbours - spatial outliers) but there is no evidence that such areas have any underlying spatial distribution that needs to be explained.

Figure 6 demonstrates the use of the option to select a subset of cases. A geographically defined section of the city has been chosen. Selected are areas lying in a part of Sheffield that embraces some of the most deprived council estates (suffering from a general mix of high levels of deprivation, poor health as measured by various mortality and incidence rates, and high levels of crime). They have been selected by the polygon option including only those areas lying entirely inside the polygon which gives the user greater control of which areas to include and which to leave out. An alternative selection method would have been to pick areas one at a time. The selected area is highlighted on the map window and the boxplot of uptake rates for the selected areas is added to the existing all-areas uptake rate boxplot. Note that the boxplot is beside the all-cases boxplot, an alternative approach would be to directly overlay the two but this is not possible in SAGE. (Note that this overlay facility is possible if multiple scatterplots are produced and there is a colour pallette and symbol option to help differentiate the two.) Grid lines have been switched on for ease of comparison and the graph window re-sized. Interestingly there is no clear evidence on the basis of this exploration of the data of a substantial discrepancy between the selected area and the city as a whole which suggests that the groups of women residents of this area appear to be no less inclined to use the breast screening service than groups of women across the city as a whole.

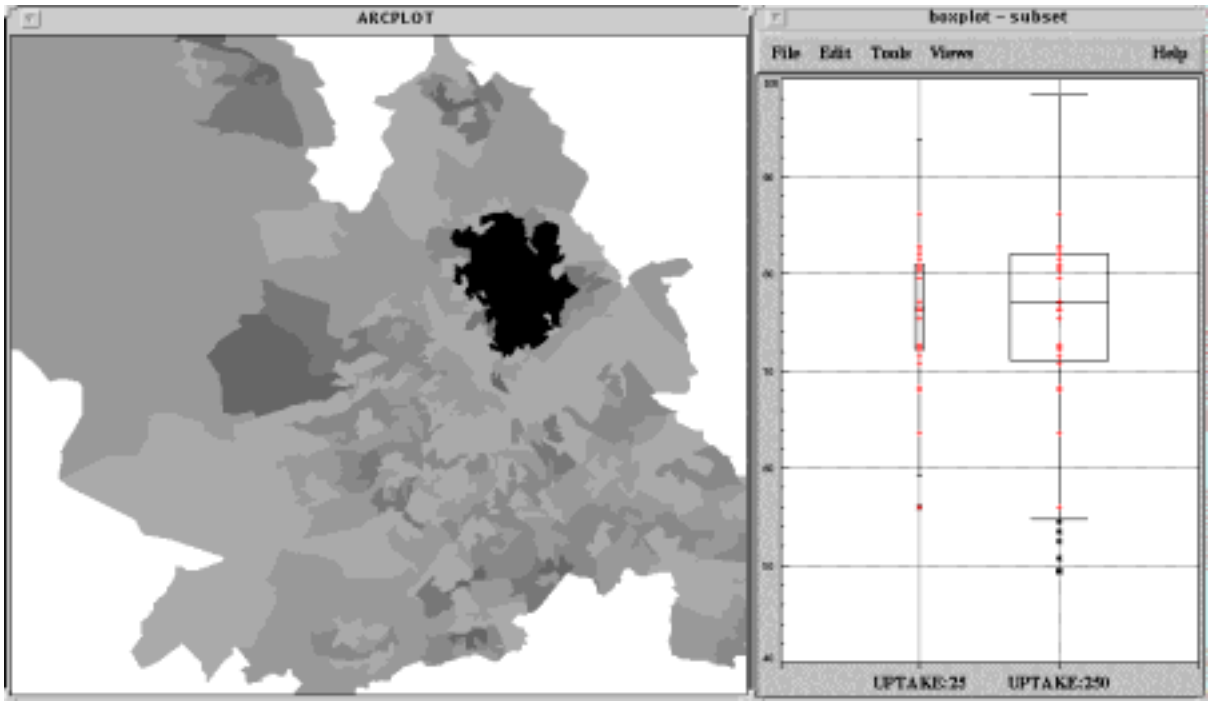

Figure 6 Closer analysis of a particular area in Sheffield in order to compare with Sheffield as a whole

We now comment briefly on ways in which the visual tools in SAGE might be significantly enhanced for the purposes of table look up and pattern perception. First the mapping tool in SAGE is cumbersome - paricularly so in comparison with a system like cdv (Dykes 1996). Although the user can zoom in and out of any portion of the map to investigate the mapped data more closely, in a software system for the visual and exploratory statistical analysis of spatial data a flexible mapping facility would seem to be of major importance and this is an area where SAGE is disappointingly slow and cumbersome particularly in those situations where the user wants to re-shade the map and it may be both tedious and difficult to obtain a satisfactory colour palette for maping an attribute. Perhaps surprisingly this is a consequence of using Arc/Info whose mapping capability does not seem to be well suited to the sort of flexible mapping role required in ESDA. Whilst the mapping capability is adequately suited to the operation of table look up (and in particular the task of matching data cases to areas), for the reason given above it is less suited to the detection, assembly and estimation tasks involved in pattern perception.

Second, even though graph windows can be re-scaled and it is possible to zoom in to a graph there may still be cases where it is difficult to identify all cases at a point on the graph which in turn makes brushing individual cases difficult. The possibility of switching on and off a “jittering” operation would overcome this and would seem to be particularly appropriate for the table look up aspects of scientific visualisation (Cleveland 1993, 1994).

Third, whilst Wise et al (1998) express the view that the operations of table look up and pattern perception (and the three associated tasks) are as appropriate to the development of scientific visualisation tools for spatial data as for non spatial data they do argue that there are some special problems with area data. In many areas associated with the development of scientific visualisation tools each case is “equivalent” and for any given category can be represented by the same symbol. Data values associated with areas are often far from being “equivalent”. Rates based on large populations are more robust than rates based on small

populations for example (Clayton and Kaldor 1987). A scatterplot of values, for example, that reveals obvious outliers may be highlighting areas where the rates are particularly non-robust (because the associated areas are small or have a small population). This issue shares common ground with the use of ESDA tools (visual and non visual) to detect data errors but in this case the data values may not be wrong but rather affected by the way the data has been collected. In figure 7, the graph appearing in figure 3 has been modified to highlight those cases that are based on small populations (using the logical query in SAGE) representing the cases that are the least robust. Interestingly these are the cases that are extreme on the scatterplot. The development of tools for visualizing spatial data should recognize this property of area based data values.

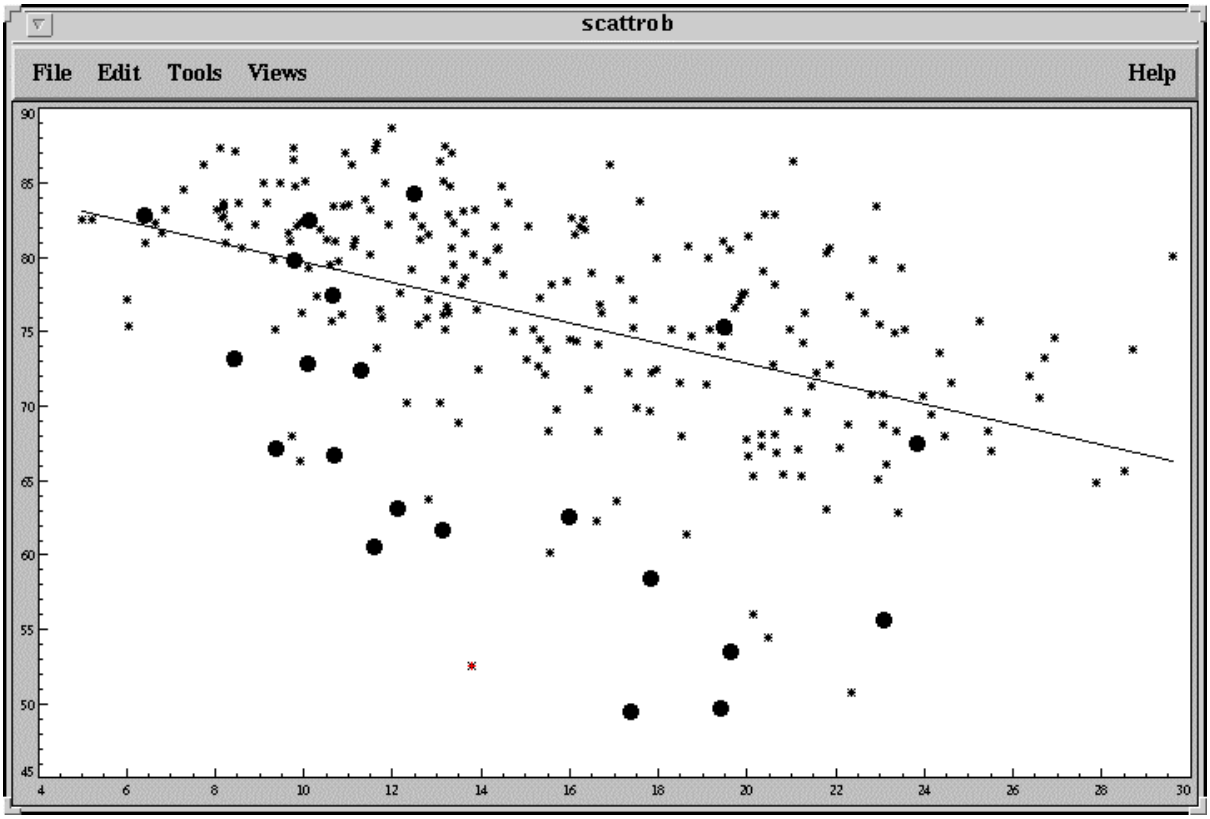

Figure 7: Scatterplot with areas whose rates are based on small populations highlighted.

4. Conclusions

This paper has summarized a possible framework for the development of scientific visualisation tools to support ESDA. The SAGE software system has been used to provide a focus for the identification of what is needed in the way of a practical system in terms of tools and how they should be implemented. Note that confirmatory spatial data analysis (CSDA) has not been touched on here. Since SAGE stores variables generated during a session, (for example since SAGE has a modelling capability it is able to fit various regression models and store, amongst other variables, regression residuals) all the tools discussed here can be used to explore aspects of model fit.

This paper has demonstrated that some simple but effective visual tools can be implemented for exploring spatial data. If there are areas that warrant particular attention it is

in the development of certain types of dynamic facilities that recognize the special attributes of spatial data. These special attributes include the need to incorporate into visualisation tools the spatial relationships between cases and the “non-equivalence” of data values that come from different areal units with markedly different underlying properties that could influence data values.

References

Clayton, D. and J. Kaldor (1987) Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics*, Vol 43, 671-681.

Cleveland W.S. (1993) Visualizing data. AT&T Bell Laboratories, Murray Hill NJ.

Cleveland W.S. (1994) The elements of graphing data. AT&T Bell Laboratories, Murray Hill NJ.

Cliff, A.D. and J.K. Ord (1981) *Spatial Processes.* Pion, London.

Craig P., Haslett J., Unwin A.R. and Wills G. (1989) Moving Statistics - An Extension of Brushing for Spatial Data. In *Computing Science and Statistics, Proceedings of the 21st Symposium on the Interface* p170-174

Cressie N. (1984) Towards Resistant Geostatistics. In G.Verly et al *Geostatistics for Natural Resources Characterization* Dordrecht, Reidel. p. 21-44.

Cressie, N (1991) *Statistics for Spatial Data.* Wiley, New York.

Dykes J. (1996) Dynamic maps for spatial science: a unified approach to cartographic isualization. In D.Parker (ed), *Innovation in GIS 3*, 177-187. Taylor and Francis, London.

Fotheringham, A.S. and M.E. Charlton (1994) "GIS and exploratory spatial analysis: an overview of some research issues" *Geographical Systems,* Vol 1, 315-328.

Friendly M. Conceptual and Visual Models for Categorical Data. *The American Statistician* 49, 153-160.

Getis A. and J.K.Ord (1992) "The analysis of spatial association by use of distance statistics". *Geographical Analysis.*,Vol 24, 189-206.

Haining, R.P (1993) *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge University Press.

Haining, R.P., Wise, S.M and Ma, J. (1998) Exploratory spatial data analysis in a Geographic Information System Environment. *The Statistician* (in press).

Haslett J., Wills G. and Uniwn A.R. (1990) SPIDER - an interactive statistical tool for the analysis of spatially distributed data. *Int.J.Geographical Information Systems* 4(3),285-296.

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

Hoaglin D.C., Mosteller F. Tukey J.W (1985) *Exploring Data Tables, Trends, and Shapes* Wiley, New York.

Ma, J., R.P. Haining and S.M. Wise (1997) *SAGE users guide.* Available within the SAGEV01.tar file at ftp://hippo.shef.ac.uk/pub/uni/academic/D-H/g/sage/sagehtm/sage.htm

Riedwyl H. and Schuepbach M. (1994) Parquet diagram to plot contingency tables. In F.Faulbaum (ed) *Advances in Statistical Software, 4,* Gustav Fischer, Stuttgart p293-299.

Tukey,J.W. (1977) *Exploratory Data Analysis.* Reading, Mass. Addison Wesley.

Unwin A.R. (1996) "Exploratory spatial analysis and local statistics" *Computational Statistics*, Vol 11, 387-400

Unwin A., Hawkins G. Hofman H and Siegl B. (1996) Interactive graphics for data sets with missing values - MANET. *J.Computational and Graphical Statistics*,5,113-122.

Unwin A. and Hofman H. (1997) New Interactive Graphics Tools for Exploratory Analysis of Spatial Data. In S.Carver (ed), *Innovations in GIS 5.* (Other information at http://www1.math.uni-augsburg.de/Manet/)

Wise, S.M., Haining, R.P. and Signoretta, P (1998) Scientific visualization and the exploratory analysis of area based data. (Submitted for publication)

Wise, S.M, R.P. Haining and J.Ma (1997) "Regionalisation tools for the exploratory spatial analysis of health data". In M.Fischer and A.Getis (Eds) *Recent Developments in Spatial Analysis: Spatial statistics, behavioural modelling and neuro-computing.* Berlin, Springer-Verlag p83-100.

Acknowledgements. The first two authors wish to acknowledge receipt of ESRC grant number R000234470 “Developing spatial statistical software for the analysis of area based health data linked to a GIS” which enabled the development of SAGE. All three authors wish to acknowledge the receipt of a grant from the Joint Information Services Committee (JISC) and the ESRC which made possible some of the work reported here.

In addition the authors wish to thank Jingsheng Ma for the work he undertook on the development of SAGE including the visualisation tools and Dawn Thompson, a PhD student in the Department of Geography of The University of Sheffield, who kindly allowed us to use the breast cancer screening uptake data.

Appendix:

For further details on SAGE including access to the software, this is available from the following web site: http://www.shef.ac.uk/~scgisa