The Cartographic journal *Vol. 45 No. 3 pp. 182–192 Geovisualisation Special Issue, August 2008*

REFEREED PAPER

Combining Geovisual Analytics with Spatial Statistics: the Example of Geographically Weighted Regression

Urška Demšar, A. Stewart Fotheringham and Martin Charlton

National Centre for Geocomputation, National University of Ireland, Maynooth, Co. Kildare, Ireland
urska.demsar@nuim.ie, stewart.fotheringham@nuim.ie, martin.charlton@nuim.ie

An attempt is made to facilitate interpretation of the results of a spatial statistical method – Geographically Weighted Regression (GWR) – using a geovisual exploratory approach. The GWR parameter space is treated as a multivariate dataset and explored in a geovisual exploratory environment with the goal to identify spatial and multivariate patterns that describe the spatial variability of the parameters and underlying spatial processes.

INTRODUCTION

Traditional regression analysis describes a modelled relationship between a dependent variable and a set of independent variables. When applied to spatial data, the regression analysis often assumes that the modelled relationship is stationary over space and produces a global model which is supposed to describe the relationship at every location in the study area. This would be misleading, however, if relationships being modelled are intrinsically different across space. One of the spatial statistical methods that attempts to solve this problem and explain local variation in complex relationships is Geographically Weighted Regression (GWR) (Fotheringham *et al.*, 2000; 2002).

In a global regression model, the dependent variable is often modelled as a linear combination of independent variables, where a parameter belonging to each variable is assumed to be stationary over the whole area (i.e. the model returns one value for each parameter). GWR extends this framework by dropping the stationarity assumption: the parameters are assumed to be continuous functions of location. The result of the GWR analysis is a set of continuous localised parameter estimate surfaces, which describe the geography of the parameter space (Fotheringham *et al.*, 2002). These estimates are usually mapped or analysed statistically to examine the plausibility of the stationarity assumption of the traditional regression and different possible causes of non-stationarity (Fotheringham *et al.*, 2002). However, questions which are not currently addressed in the GWR literature are: 'Do there exist areas of stability where all the parameters keep relatively constant values?' and 'Are there any predominant groupings of parameters that behave in a similar way everywhere in the area of investigation?'. These and similar questions relate to the structure and patterns in the GWR parameter space, which a typical presentation of results of a statistical analysis cannot answer. In this paper we suggest a geovisual exploratory post-analysis of the GWR parameter space using an automatic-visual data exploratory environment in order to attempt to answer such questions. The goal is to facilitate interpretation of the GWR results and to raise and answer new questions about the spatial variability of the parameters and the underlying spatial processes.

The remainder of the paper is structured as follows: the next section introduces GWR and presents the geovisual exploratory environment used in the post-analysis of the GWR results. Then, a small case study, on which our suggested approach is tested, is described together with observations made on the resulting visual representations of the GWR parameter sets. Finally, some conclusions and ideas for further development are presented.

METHODOLOGY

This section introduces the statistical method (GWR) and the geovisual exploratory environment used in this study.

Geographically Weighted Regression

Consider a standard linear regression model in which a dependent variable y_i whose value is recorded at location i is regressed on a set of independent variables x_{1i}, x_{2i}, … x_{ni}:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} \quad (1)$$

where β_0 … β_n are parameters to be estimated. The model need not be linear for what follows but this provides a

DOI: 10.1179/000870408X311378

useful example and represents the most commonly applied form of regression model. An assumption implicitly made in using this model form is that the parameters of the model are stationary over space – meaning that the same relationships are found everywhere in the study area and, hence, that the processes producing these relationships are spatially invariant. Whilst this might be a reasonable assumption to make in some circumstances, in others it is questionable. For instance, suppose the above general model were replaced by a specific example in which the ys represented the price of a house and the xs represented features of the house or its neighbourhood that were thought to affect the price of the house. Suppose further that one of the $\mathbf{x}$ attributes was the presence or absence of a garage. In some areas, where garages are relatively rare, the presence of a garage would be expected to generate a higher increase in the price than it would in areas where garages were plentiful. That is, local variations in supply and demand would induce locally varying relationships. Such local variations in relationships might be more common than we imagine and are at least worth exploring. It would seem strange to continue with the assumption that all processes are spatially invariant – why not question this assumption and allow parameter estimates to vary locally if indeed processes might vary over space?

This is the simple idea behind GWR. The above 'global' model is replaced by a 'local' version in which parameter estimates β are allowed to vary over space:

$$y_i = \beta_{0i} + \beta_{1i}x_{1i} + \cdots + \beta_{ni}x_{ni} \quad (2)$$

The parameters $\beta_{0i} \ldots \beta_{ni}$ now have a subscript i denoting that they can vary over space. In the form given here, there will be one set of parameter estimates generated for every location at which data are recorded, but this is not a requirement of the method and parameter estimates can be generated for any location. Details on the calibration of this model are given in Fotheringham *et al.* (2002) but essentially the data around each regression point are weighted according to their distance from this regression point with data at smaller distances being given higher weights. Each time the regression point is moved around the study region, all the weights are recalculated and the estimator re-run. Hence, at each regression point, a set of local parameter estimates is obtained which represents the processes operating around that regression point. By moving the regression point around the whole study area, a local parameter surface can be constructed showing visually the spatial variation in the process being described by the parameter.

The estimator for GWR is shown below (eq. 3). The notation can be generalised such that $\mathbf{u}$ represents any location in the study area. Parameter estimates can be obtained at locations at which the data used to fit the model have not been collected. This may seem eccentric, but these locations might be the members of a control set where the data have been divided into training and control sets.

$$\hat{\beta}(\boldsymbol{u}) = (X^T W(\boldsymbol{u}) X)^{-1} X^T W(\boldsymbol{u}) y \quad (3)$$

$W(\mathbf{u})$ is a diagonal matrix of weights. Details of this matrix and the associated weighting functions are given in Fotheringham *et al.* (2002).

Not only local versions of parameter estimates are obtained in the calibration of models by GWR; any diagnostic from a standard global regression model will have its local counterpart. Hence, we can generate local standard errors, local t values, local goodness-of-fit measures and so forth. Of course, this can produce a large volume of results which is why the visualisation of GWR outputs is important. To date, this visualisation has been confined largely to producing 2D and pseudo-3D univariate surfaces of local parameter estimates and their t values but we now take this one step further.

Using a geovisual exploratory post-analysis to interpret GWR results

Typically, the results of a regression analysis are presented in the form of parameter estimates and some univariate summary statistics such as the r^2 statistic or the F-test. These can be thought of as 'whole map' statistics – there is only one r^2 for the entire model, and the relationship between the dependent and independent variables is assumed to be spatially stationary. It is possible, therefore, that interesting patterns and structures in the data remain unexplored. Typical visual data presentations on the other hand generate rich visual, animated, interactive displays in multiple coordinated views which support user-controlled exploration. Such visual presentations are powerful in revealing trends and showing clusters and other trends, but have their limitations in terms of the objectivity of observations produced. A combination of statistical and visual approaches therefore incorporates the advantages of both and supports faster and more effective analytical reasoning (Shneiderman, 2001; Theus, 2005).

In the case of GWR, the variability and behaviour of the parameter estimates are influenced by the underlying spatial processes. To examine the spatial variability of the parameters and, thereby, the processes, GWR provides statistical summaries for each parameter, which indicate if non-stationarity is present or not. Additionally, Monte Carlo tests can be performed to determine the existence of the spatial variability of each parameter (Fotheringham *et al.*, 2002). Another way to examine the spatial variability of the parameters is to visualise parameter surfaces – this is usually done by univariate mapping, such as producing a choropleth map of each separate parameter surface. These visualisations serve as an informal inference tool (Fotheringham and Brunsdon, 2004) for interpretation of the GWR results.

The statistical summaries, Monte Carlo tests and univariate visualisations serve their purpose if the aim is to investigate each parameter surface separately. However, since these approaches focus on one parameter surface at a time, they are not sufficient to discover multivariate spatial and non-spatial relationships and patterns in the parameter space. To approach this problem, we suggest treating the GWR parameter space as a multivariate dataset and exploring it using a geovisual exploratory environment, thereby combining statistical analysis with post-analysis visual exploration. The goal is to uncover information related to the spatial variability of the parameter estimates, such as finding areas of stability where all parameters behave

in a similar way or identifying groups of parameters that behave similarly everywhere in the study area.

Visual data exploration of spatial data is a part of exploratory spatial data analysis (Unwin and Unwin, 1998) and is essential to prompt ideas and generate hypotheses through the creation, inspection and interpretation of visual representations of spatial data. The goal is to transform complex data into visual displays, which allow an analyst to look for patterns, trends, relationships and structure that describe the significant aspects and characteristics of the data (Keim and Ward, 2003). Discovered patterns serve to infer knowledge not only about the data but also about the geographical processes that generated the data. The perceptual-cognitive process of alternatively interpreting and analytically reasoning about georeferenced visual displays is explored in the discipline of Geovisual Analytics. Geovisual Analytics, 'the science of analytical reasoning and decision-making with geospatial information, facilitated by interactive visual interfaces, computational methods, and knowledge construction, representation, and management strategies' (MacEachren, 2008), falls under the recent new discipline Visual Analytics (NVAC, 2005) and has evolved from geovisualisation (MacEachren *et al.*, 1999; MacEachren and Kraak, 2001). Geovisual analytical data exploration is typically used to derive knowledge from large and highly dimensional geospatial data and to discover the unexpected. Replacing the traditional univariate visualisation of the GWR result space with a multivariate geovisual analytical exploration is therefore a logical step towards better understanding of the GWR results.

While there exist several visualisation environments that support development of geovisual exploration systems for spatial data (such as for example the Common GIS (Andrienko G. *et al.*, 2003; Andrienko N. and Andrienko G., 2006) and GeoDa – Geodata Analysis Software (Anselin *et al.*, 2004)), we have decided to recycle an existing geovisual exploratory environment. This environment was originally developed by one of the authors for exploration of a multivariate environmental dataset (Demšar, 2007) and was built using GeoVISTA Studio, a collection of various geographic and other visualisations as well as computational data mining methods (Gahegan *et al.*, 2002; Takatsuka and Gahegan, 2002).

The system consists of the following visualisations: a Self-Organising Map (SOM); two parallel coordinates plots (PCP) – one ordinary and one linked to the SOM; a multiform bivariate matrix with scatterplots, spaceFill visualisations and histograms; and a bivariate geoMap. A brief description of each visualisation and the integrated computational data mining method, the SOM, follows.

In a PCP, each parallel vertical axis represents one dimension/variable of the input data space. In the exploratory environment employed here, there are two PCPs. One is ordinary, where each data instance is displayed as a polygonal line intersecting each of the axes at the point which corresponds to the respective attribute value for this data instance (Inselberg, 2002). In the second PCP, each line represents a group of data instances which were assigned to one SOM cell (Guo, 2003; Guo *et al.*, 2005), as described below. The reason for including two such plots is that the ordinary GeoVISTA PCP has many more interaction possibilities and shows statistical descriptions of the data, while cluster analysis is easier in the PCP linked to the SOM because of the grouping of the data elements according to their respective SOM cells.

A multiform bivariate matrix is a generalisation of a scatterplot matrix and consists of univariate visualisations – histograms on the diagonal and bivariate visualisations at other positions in the matrix. In the matrix in this particular exploratory environment, scatterplots of each corresponding pair of variables are located above the diagonal and spaceFills below the diagonal. In spaceFills, each data vector is represented by a grid square. The first of the two display variables defines the colour of each square, while the second defines the order of the squares inside the rectangular display (Gahegan *et al.*, 2002).

The geographical visualisation in the exploratory environment is the geoMap from GeoVISTA Studio, which is a choropleth map, whose colour scheme is either defined by a cross-tabulation of the two display attributes (Gahegan *et al.*, 2002) or can alternatively be inherited from other visualisations, such as the SOM lattice, as described below.

The SOM is an unsupervised neural network. The algorithm projects the multidimensional data onto a two-dimensional lattice of cells while preserving the topology and the probability density of the input data space. The result of this is that similar input data vectors will be mapped to neighbouring cells. This ensures that the similarity patterns that exist in the higher dimensional space correspond to patterns in the SOM lattice (Kohonen, 1997; Silipo, 2003) – this characteristic produces a very visualisable result due to its two-dimensionality (Vesanto, 1999). Because it preserves both the topology and the distribution of data vectors in the high dimensional input space, the SOM is considered a useful method for knowledge discovery tool from spatial data, as demonstrated in a number of recent studies (for example Takatsuka, 2001; Gahegan *et al.*, 2002; Jiang and Harrie, 2004; Koua and Kraak, 2004; Guo *et al.*, 2005; Skupin and Hagelman, 2005; Demšar, 2007; Špatenková *et al.*, 2007; to list a few).

The GeoVISTA version of SOM used in our exploration environment implements the original Kohonen algorithm (Takatsuka, 2001; Guo *et al.*, 2005) and does not take into account the geographical location as for example a GeoSOM (Bação *et al.*, 2005) does. The reason for choosing the original non-spatial Kohonen SOM over a GeoSOM in our case is because one of the potentially interesting patterns for interpretation of GWR results is similarity in attribute space. The task in question is to try to identify groups of data elements with similar behaviour of several parameter estimates and only when this identification has been done explore the spatial distribution of such groups. This can be achieved by transferring the similarity pattern discovered in the attribute space by the original Kohonen SOM to a geographical visualisation, i.e. the geoMap through colour brushing as described below.

The SOM visualisation in GeoVISTA (Guo *et al.*, 2005) is a hexagonal U-matrix, consisting of two types of cells: node cells, which contain circles and represent nodes of the SOM and distance cells, which are dispersed between node hexagons and whose grey shade represents multivariate

dissimilarity between two neighbouring node hexagons. The grey shade of each node cell is also calculated according to the cell's distance to its neighbours. Light areas in the lattice therefore indicate areas with very similar cells and represent clusters. Dark areas indicate borders between clusters. In total, there are 13 6 13 node cells in the GeoVISTA SOM that we used. The distribution of data vectors in the SOM lattice is represented by the size of the circles that are projected over the node cells. The larger the circle the more data vectors have been mapped to the cell that the circle belongs to. Circles are linearly scaled so that the largest circle touches the border of its respective hexagon. The groupings of data vectors marked with the circles are transferred to one of the two PCPs where each polygonal line represents one cell and the width of the line the number of the input data vectors that have been mapped to the cell. The second visual variable transferred from the SOM lattice into all other visualisations, not just one of the two PCPs, is the colour of the circles. This colour is originally defined by draping a smooth 2D colour map over the circles in the lattice and then the hue of each circle is inherited by graphic entities in other visualisations for visual brushing. More information on how the 2D colour map is derived and on other characteristics of this particular SOM visualisation can be found in Guo *et al.* (2005).

Aside of colour brushing, all visualisations in GeoVISTA-based systems are also connected by the interactive selection and brushing through mouse-over operation (Gahegan *et al.*, 2002; Takatsuka and Gahegan, 2002; Guo *et al.*, 2005).

Even though the exploratory environment presented here was not specifically designed for exploration of the GWR result space, it can be efficiently used for this purpose. The similarity groupings produced by the SOM can help discover groups of areas where several parameters behave in the same way. Following the trajectories of these groups in the PCP gives an idea of how the parameters behave. Spatial variability of the parameters can be examined in the map, not only for one parameter at a time, but in combination with others, so that multivariate spatial variability patterns can be discovered. The map in combination with the SOM also allows comparison between spatial similarity and parameter similarity, while the visualisations in the bivariate matrix indicate if any of the pairs of parameter estimates are correlated or not.

CASE STUDY AND RESULTS

To examine the utility of viewing the GWR results in a geovisual exploratory environment, we performed a small case study on a spatial dataset concerning educational attainment in the US state of Georgia. The dataset consisted of records for the 174 counties in the state and had the following seven variables: percentage of inhabitants with at least a bachelor degree, total population in 1990, percentage of rural population, percentage of elderly, percentage of foreign-born inhabitants, percentage of inhabitants living below the poverty level and percentage of African-Americans. GWR was run on this dataset to model the relationship between the educational variable (i.e. the percentage of inhabitants with at least a bachelor degree) and the other six variables and to determine if there were any geographical variations in the relationships between educational attainment and these variables.

The result of the GWR analysis was an output file consisting of 28 localised variables: seven parameter estimates for the intercept and the six independent variables; the localised versions of the standard errors of each local parameter estimate and the associated t-values; the observed y value; the predicted y value; and several other statistical measures, such as the residuals, standardised residuals; various influence statistics and the local r^2 statistic, which describes the local goodness-of-fit of the model.

This dataset was then transferred into a shape file and imported into the geovisual exploratory environment. Figure 1 shows the visualisations of the exploratory environment immediately after the SOM clustering, but before any visual exploration has been performed and before any of the visualisations have been manipulated. The rainbow colour scheme is inherited from the SOM.

The SOM clustering was based on the seven parameter estimates (i.e. for the intercept and the six independent variables), using equal weights for each of the variables. These seven variables were then further visually explored together with the local r^2 statistic in order to uncover spatial or other patterns. Table 1 gives a list of visual variables and respective parameter estimates for independent regression variables. Further exploration of the other 20 output variables would of course be possible, but was considered beyond the scope for this paper and is something we plan to look at in the future.

The pattern appearing in both PCPs in Figure 1 is fairly jagged, i.e. the bluish and the reddish lines cross many times. In both these PCPs, the order of the variable axes corresponds to the order in which they were listed in the data file, starting from the intercept, through six independent variables in both PCPs with an added r^2 axis as the last one on the right in the PCP. Such arbitrary ordering usually produces irregular and jagged patterns that are difficult to explore visually. If the axes are instead permuted so that similar variables are positioned adjacently in the graphical display, it is much easier to discover correlations between variables and groupings of data elements (Hurley, 2004). The first step of our exploration was therefore to permute the axis in the ordinary PCP, which is interactive and allows this operation (the SOM PCP does not allow it). This interactivity was one of the reasons why there are two PCPs in the exploratory environment (which might otherwise seem redundant). While there is some limited interactivity implemented in the SOM PCP, such as interactive selection of lines representing the SOM cells and a union or intersection combination selection (Guo, 2003), it does

Table 1. A list of visual variables and respective parameter estimates for independent regression variables

Visual variable	Parameter estimate for
PARM_1	Intercept
PARM_2	Total population
PARM_3	Percentage of rural population
PARM_4	Percentage of elderly
PARM_5	Percentage of foreign-born inhabitants
PARM_6	Percentage of inhabitants living below the poverty line
PARM_7	Percentage of African-Americans
LOCRSQ	Local r^2 statistic

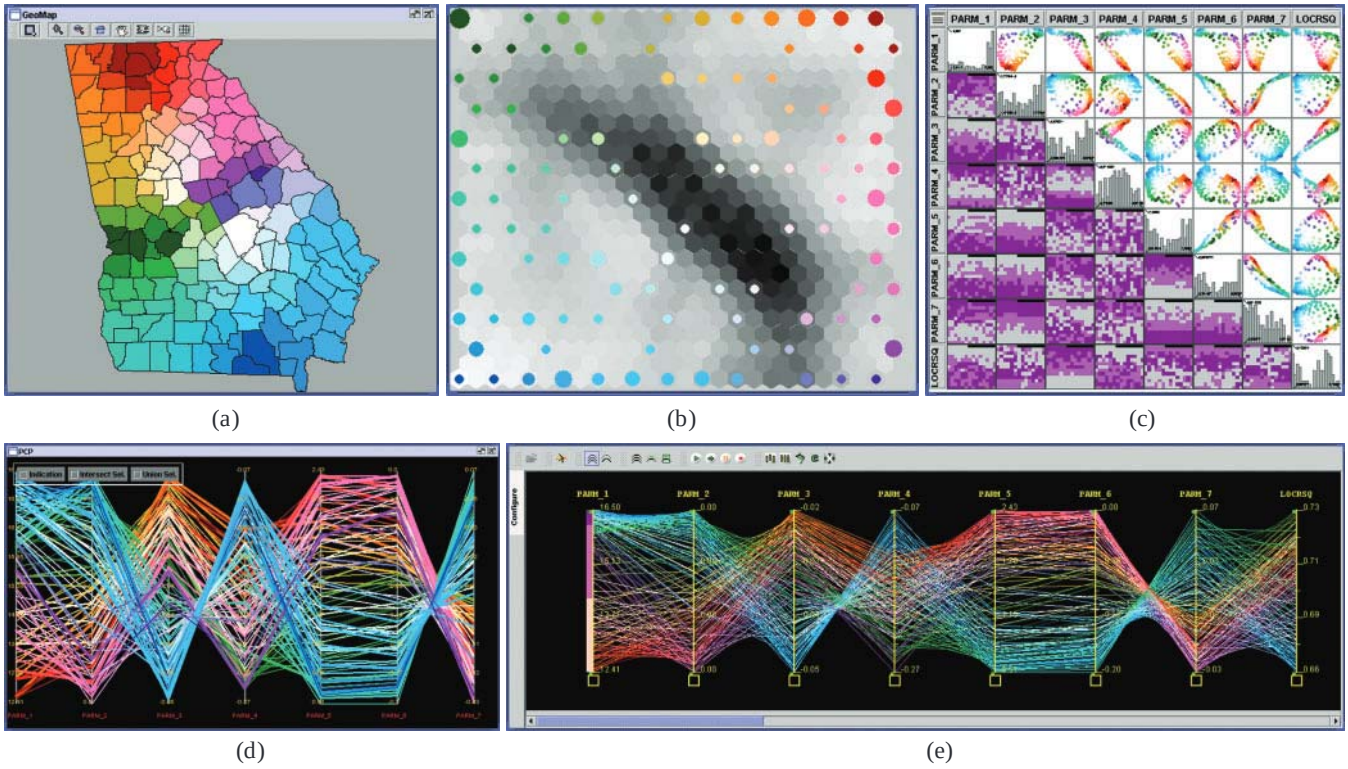


Figure 1. Examining the GWR parameter estimate space in an integrated automatic-visual exploratory environment: a) the geoMap, b) the SOM, c) the bivariate matrix with the seven parameter estimates and the additional r^2 variable, d) the SOMPCP showing the seven parameter estimates upon which the SOM clustering was based, e) the PCP, showing the seven parameter estimates and the additional r^2 variable. The colour in all visualisations, except the spaceFills and the histograms in the matrix, is inherited from the SOM

not allow operations such as add, remove, permute or scale the axes or change the colour scheme according to a chosen variable, nor does it provide the ability to display additional statistical measures, such as boxplots on each axis, for example, all of which the ordinary PCP does.

Another difference between the two PCPs is the scaling of the axes. The axes in the ordinary PCP are linearly scaled from the minimum to the maximum value of each dimension. In the SOM PCP, which is based on the Hierarchical Density cluster viewer (Guo, 2003) however, the axes are scaled using nested means scaling. This means that each axis is recursively being divided into two sub-intervals where the mean of the data is assigned to the central point of the axis and splits the data into two subsets. The procedure is then repeated on each of the two subsets of data, those data items that are larger than the mean and those that are smaller, until the entire axis has been divided into eight sub-intervals. Each of these eight sub-intervals is then linearly scaled (Guo, 2003). While the primary aim of the nested scaling is to reduce overprinting (Guo *et al.*, 2005), it distorts the real statistical distribution of data at each axis. This can on the other hand be easily visually analysed in the ordinary PCP through display of boxplots on each axis (Gahegan *et al.*, 2002).

In highly dimensional PCPs, the permutation of the axes has to be automated and is usually based on a similarity clustering of the variables according to some interestingness measure (Hurley, 2004). In our case, the dimensionality was low enough to do this manually. The permutation is shown in Figure 2 and the ordering of the axes is as follows: PARM_1, PARM_2, PARM_4, PARM_7, PARM_3, PARM_5, PARM_6 and LOCRSQ, where PARM_1 represents the parameter estimate for the intercept, PARM_i the parameter estimate for the independent variable i–1 and LOCRSQ the r^2 value (Table 1). The pattern in the permuted PCP in Figure 2 is much less jagged than the one in the original PCP in Figure 1e.

After this initial step, the exploration continued through interactive manipulation of various visualisations. The remainder of this section presents some of the more interesting observations.

One fairly obvious pattern that catches the eye is the distribution of clusters in the SOM visualisation (Figure 1b). There are four lighter areas in the SOM, which represent four clusters. Each of these areas is located in one of the corners of the lattice and they roughly correspond to the following colours: a blue–turquoise cluster, a green cluster, a red cluster and a violet cluster. There is a large dark boundary area in the centre of the SOM, which indicates that the blue and red and the violet and green clusters are, respectively, very different from each other. However, data elements in the violet and red clusters are not that very different from each other, as the boundary between the two clusters is of a lighter shade of grey than the boundary areas between other clusters. The map (Figure 1a) reveals the spatial distribution of the elements belonging to the four clusters: the red and the blue clusters are situated far from each other, in the north and the south of the study area. The violet and green clusters are located in close proximity on the map.

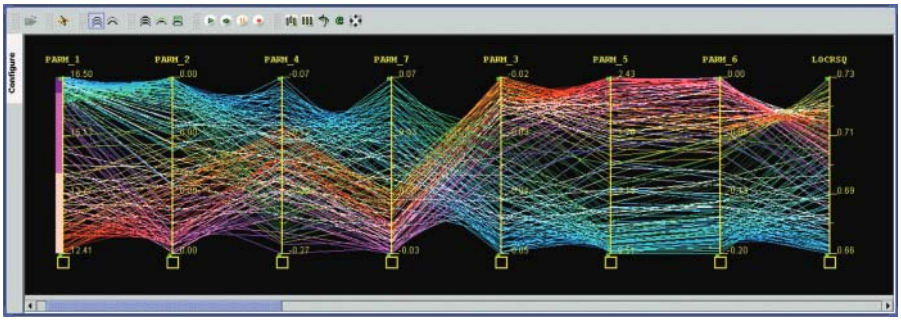
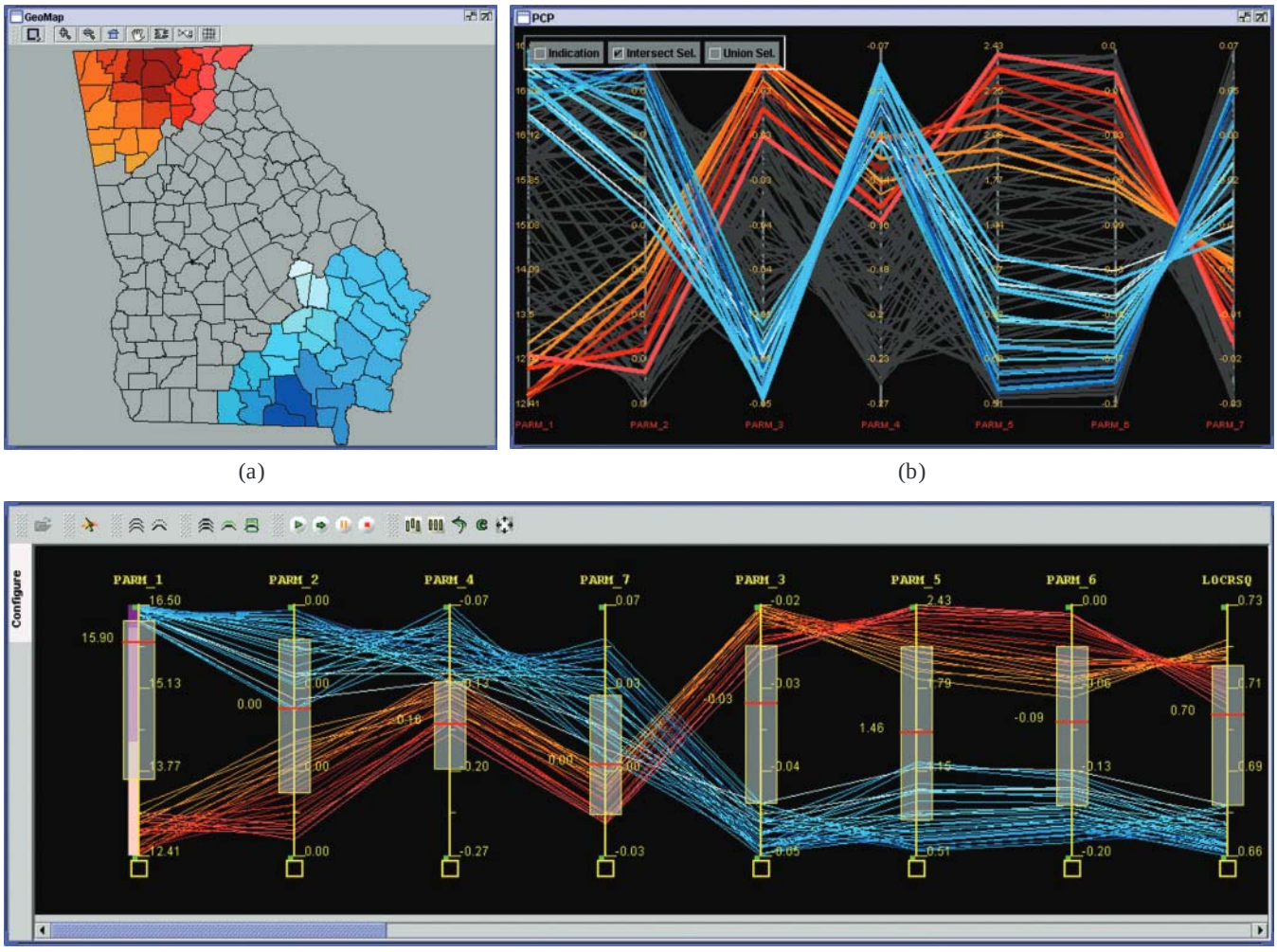


Figure 2. Permuted PCP with seven parameters and the r^2 variable

Figure 3 shows a selection of the blue and the red cluster. The fact that these two areas are far from each other in the SOM as well as in the map produces a not too surprising observation, namely that areas which are far from each other spatially may also be very different from each other in terms of the spatial processes operating within them (although this is not always the case). The different characteristics of these two clusters are fairly obvious also in the SOM PCP (Figure 3b) and the ordinary PCP (Figure 3c), where the groups of blue and red lines, respectively, have completely different trajectories through the display. There are additional boxplots displayed on each axis of the ordinary PCP in Figure 3c. Interestingly, the selected blue and red lines mostly fall out of the boxes at both extremes and different extremes at that, except for the boxplot at the axis of parameter PARM_4. There, both groups of lines are located above the mean, the red lines inside the boxplot (i.e. with values between the mean and the upper quartile) and the blue lines outside the boxplot (i.e. with values higher than the upper quartile).



(a) (b)

(c)

Figure 3. The selection of the blue and red areas in the a) geoMap, b) the SOM PCP and c) the PCP

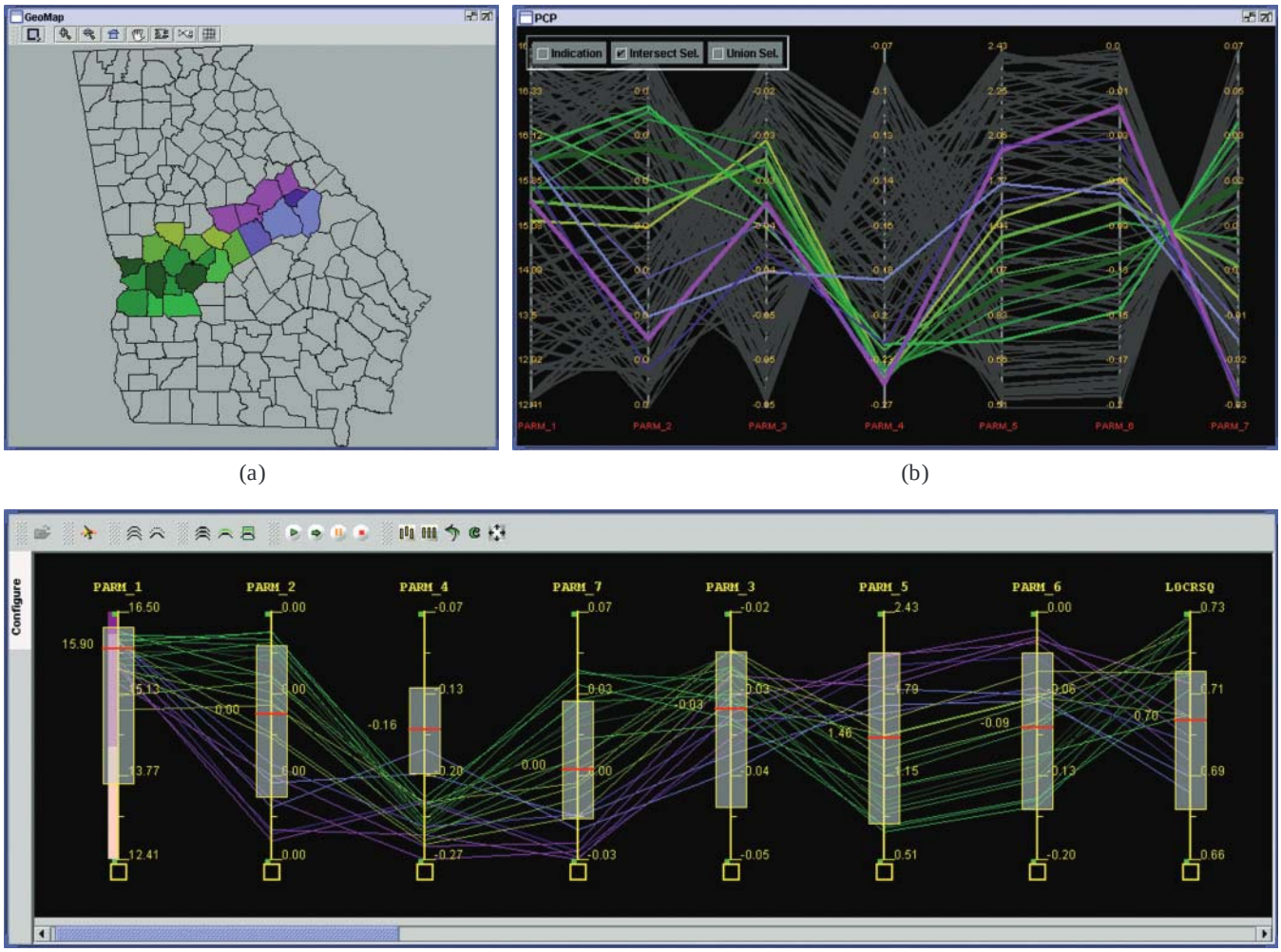


Figure 4. The selection of the green and violet areas in the a) geoMap, b) the SOM PCP and c) the PCP

Of interest are PARM_5 and PARM_7 (foreign-born and Afro-American), both of which exhibit significant parameter variation in the Monte Carlo test. Counties in which the influence of the foreign-born variable is strongly positive (shown in orange and red) are also associated with negative values of the Afro-American variable, and least influence of the rural and poverty variables. By contrast, areas where the foreign-born variable is least influential (shown in blue) are associated with positive influences in the Afro-American variable. The former set of relationships is in the north-west of the state, and the latter in the south-east. Visualisation analysis has drawn attention to two very different regimes in the state of Georgia in which different social processes appear to be leading to variation in educational attainment. The fit of the models is least good in the south-east suggesting that perhaps other social phenomena might be included in the analysis.

The green and violet clusters are also located far from each other in the SOM and are therefore very dissimilar from each other – yet they are located in the close proximity on the map (Figure 4a). Even though the trajectories of lines representing these two clusters in the PCPs (Figures 4b, 4c) are different (although not as different as the blue and red ones), the values of parameters mostly fall inside the boxplots (Figure 4c). Green and violet lines fit into boxplots at the different sides of the mean, but this difference is less than with the red and blue trajectories.

As in the red–blue case, the only exception is PARM_4, the percentage of elderly population, where both trajectories intersect the axis below the mean (fig 4c). Does this tell us something about this particular parameter? Why are estimates behaving differently for this variable than for any of the others? Perhaps a test of local significance is necessary.

Looking at the two significant variables again, PARM_5 (foreign-born) and PARM_7 (Afro-American), we can see that in those counties where the influence of the Afro-American variable is most strongly negative, the influence of the foreign-born variable is positive.

There is another eye-catching pattern in the original visualisation in Figure 1, namely the correlations that exist between several of the parameter estimates – these can be observed in the bivariate matrix (Figure 1c) as well as in the permuted ordinary PCP (Figure 2). Both these visualisations indicate that there are two groups of parameter estimates where variables seem to be fairly well correlated. The first group consists of the following variables: PARM_1, PARM_2, PARM_4 and PARM_7, and the second one of the other four variables, namely PARM_3,

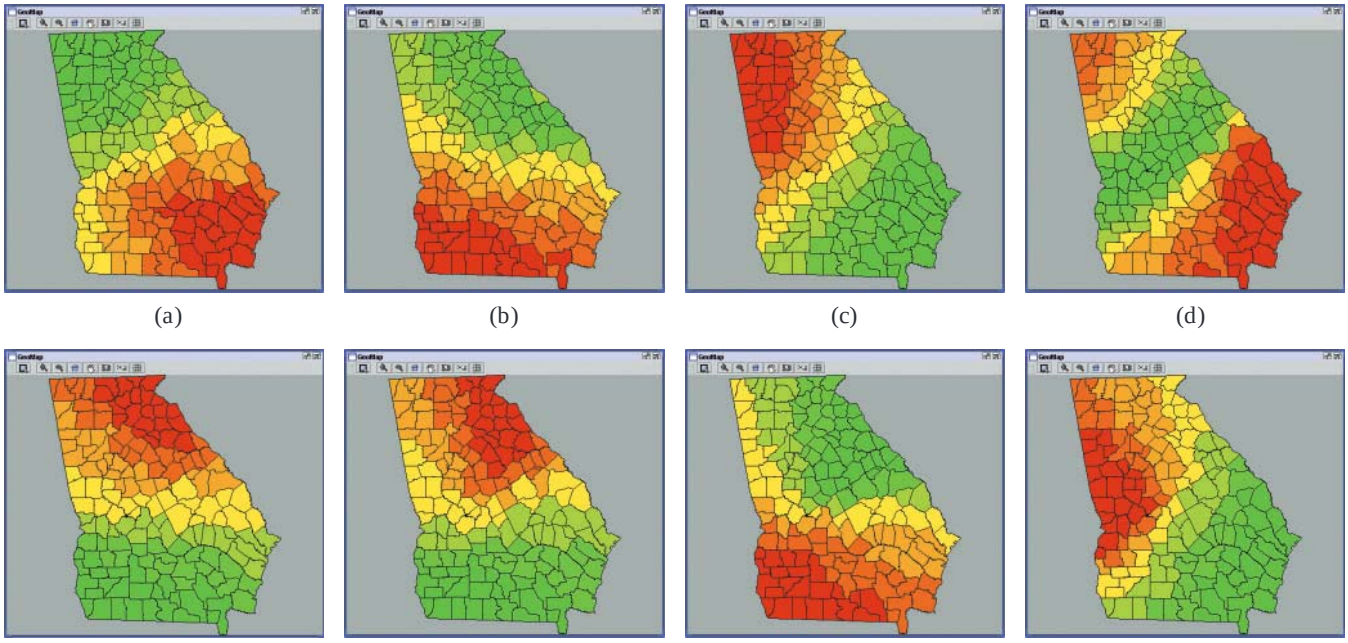


Figure 5. Graduate colour maps for all seven parameters and the r^2 variable: a) PARM_1, b) PARM_2, c) PARM_3, d) PARM_4, e) PARM_5, f) PARM_6, g) PARM_7, h) r^2

PARM_5, PARM_6 and LOCRSQ. There is an inverse correlation between these two groups – this can be seen in the PCP in Figure 2, where the lines cross each other in-between the two central axes. A more detailed analysis of the scatterplots in the bivariate matrix (Figure 1c) shows that some parameter estimates are fairly well correlated or inversely correlated, but points in some scatterplots are more randomly dispersed and in some scatterplots there are double trends (for example, in the scatterplots of PARM_1 vs. PARM_3 and PARM_4).

Earlier in this section, the parameter estimate PARM_4 was identified as the one that is behaving differently from all other parameter estimates. Further investigation of this parameter estimate in the bivariate matrix (Figure 1c) shows that this is the only parameter with a relatively Gaussian distribution in its histogram. The distributions of all the other parameters are either skewed or have two peaks.

The next step in the exploration was to produce a graduated colour map of each parameter estimate by assigning the colour according to each variable – the maps are shown in Figure 5. All maps and figures from here on employ the quantile classification into seven classes. The colour runs from green for low values through yellow for middle values to red for high values of each variable.

Looking at the maps in Figure 5, it is fairly obvious that while all parameter estimate surfaces have a general horizontal, vertical or diagonal increasing trend, it is again PARM_4, i.e. the percentage of elderly population, which stands out and is different from all other parameters. Instead of showing a monotonic increasing/decreasing trend, PARM_4 has a 'ditch' running from southwest to northeast through the centre of the study area.

Producing such univariate maps as those in Figure 5 is a traditional way to analyse the GWR results and we do not actually need the exploratory environment for this. However, the advantage of using the environment is that it allows us to make an interactive selection of the areas where the anomalous PARM_4 has either high or low values and see what happens to the other parameters in such areas. Figure 6 shows such a selection.

Figure 6a shows the map of the lowest values of PARM_4 (values lower than the lower quartile, selected in the boxplot of the PARM_4 axis in the ordinary PCP). After the selection, the colour scheme of the map was iteratively defined according to each variable and the map scrutinised for patterns each time a colour change was made. Nothing special could be seen at any of the other parameter estimates, but the selected area only has higher than average values of r^2, including some of the highest ones indicating areas where the local model is fitting the data particularly well, as the map in Figure 6b shows. This can also be seen in the last axis in the PCP in Figure 6c, where most of the green lines cross the last axis at the mean (indicated by the red line in the relevant boxplot) or higher.

A selection of the highest values of PARM_4 (values higher than the upper quartile in the relevant PCP axis) produces a more interesting result. The map of these areas is shown in Figure 7a. The ordinary PCP of these areas, coloured according to the PARM_4 variable (Figure 7e), shows two distinct trajectories in the PCP, which are particularly well separated at PARM_1, PARM_3, PARM_5, PARM_6 and r^2. These separations can also be seen in the maps of this selection, for PARM_1 (Figure 7b), PARM_3 (Figure 7c) and r^2 (Figure 7d) or in the PCP coloured according to PARM_3 (Figure 7f). The trends suggest a separation between predominantly urban counties and predominantly rural counties.

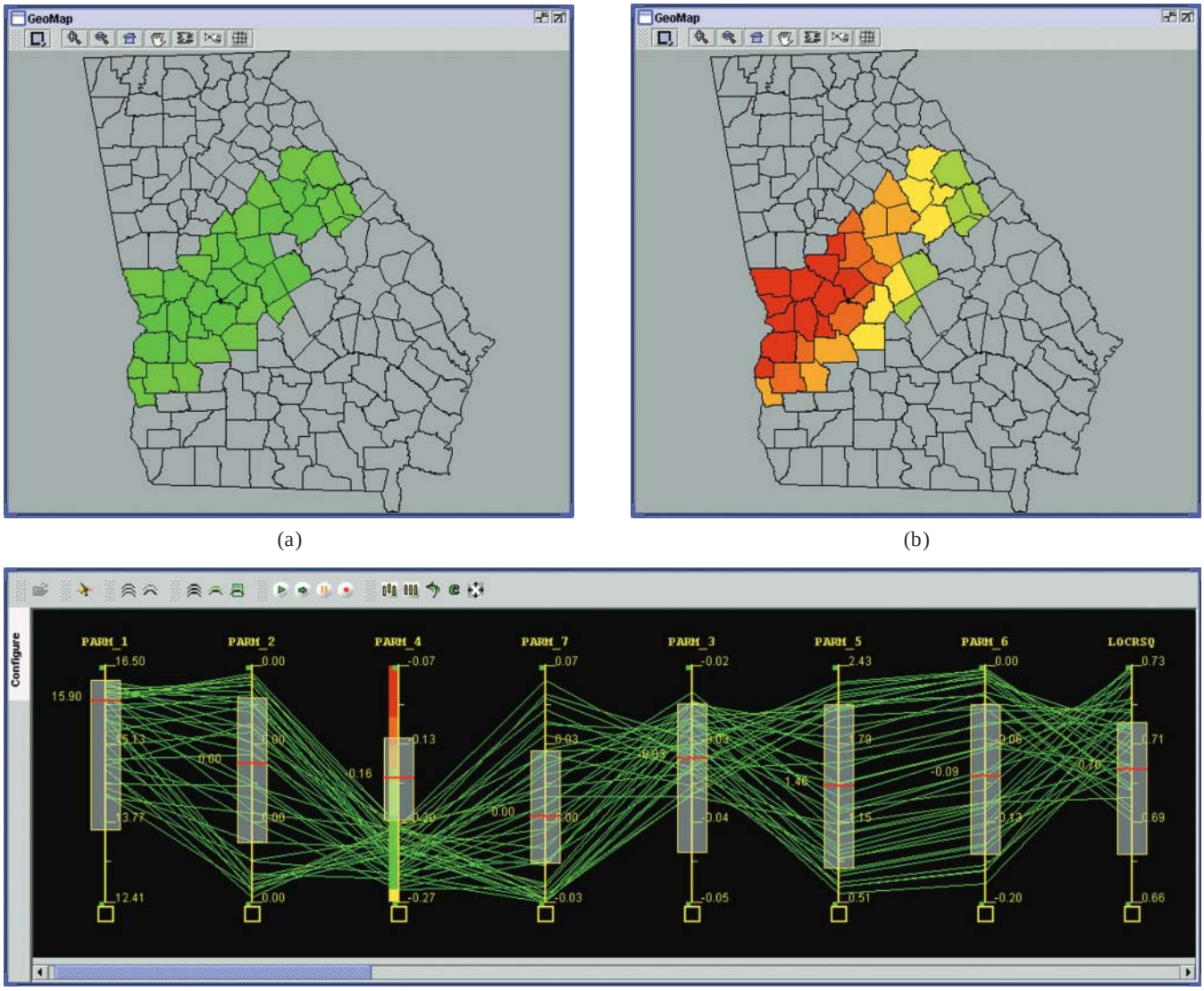


Figure 6. Selection of areas with low values of parameter 4: a) the PARM_4 map, b) the r^2 map, c) the PCP

CONCLUSIONS

In this paper, we presented an example of how a geovisual exploratory post-analysis could be employed to obtain insights into the result space of a statistical method that would otherwise remain unnoticed. While it seems that the approach facilitates interpretation of the GWR results and supports analytical reasoning about the underlying spatial processes, this is only a preliminary attempt based on a existing visual exploration environment. In the future, we plan to conduct a systematic comparison and evaluation of possible visualisation methods that could contribute to improve the understanding of the GWR results. We also plan to evaluate the performance of the combined GWR and geovisual exploration on a synthetic controlled spatio-temporal dataset in order to see if the method is really performing in the way it is expected to and investigate what pre-set patterns (spatial, non-spatial, spatio-temporal, temporal, etc.) can be discovered. Finally, if the controlled case proves viable, we plan to conduct a similar exploration of a large spatio-temporal dataset of housing prices to investigate the dynamics of the pricing process in the real estate market in London (Crespo *et al.*, 2007).

The geographical weighting principle of GWR can be easily adapted for use with multivariate statistical techniques other than regression, such as for example principal components analysis and factor analysis (Fotheringham *et al.*, 2000) or discriminant analysis (Brunsdon *et al.*, 2007). In all these methods, the model coefficients are usually treated as stationary. The trouble with adopting geographical weighting for these methods is that the interpretation of the results becomes very difficult. For example, how does one deal with spatially varying eigenvalues or what do the spatially varying discriminant functions represent? A geovisual exploratory investigation of the results could perhaps provide one way to elucidate more easily the geometrical and geographical meaning behind the spatially varying model coefficients.

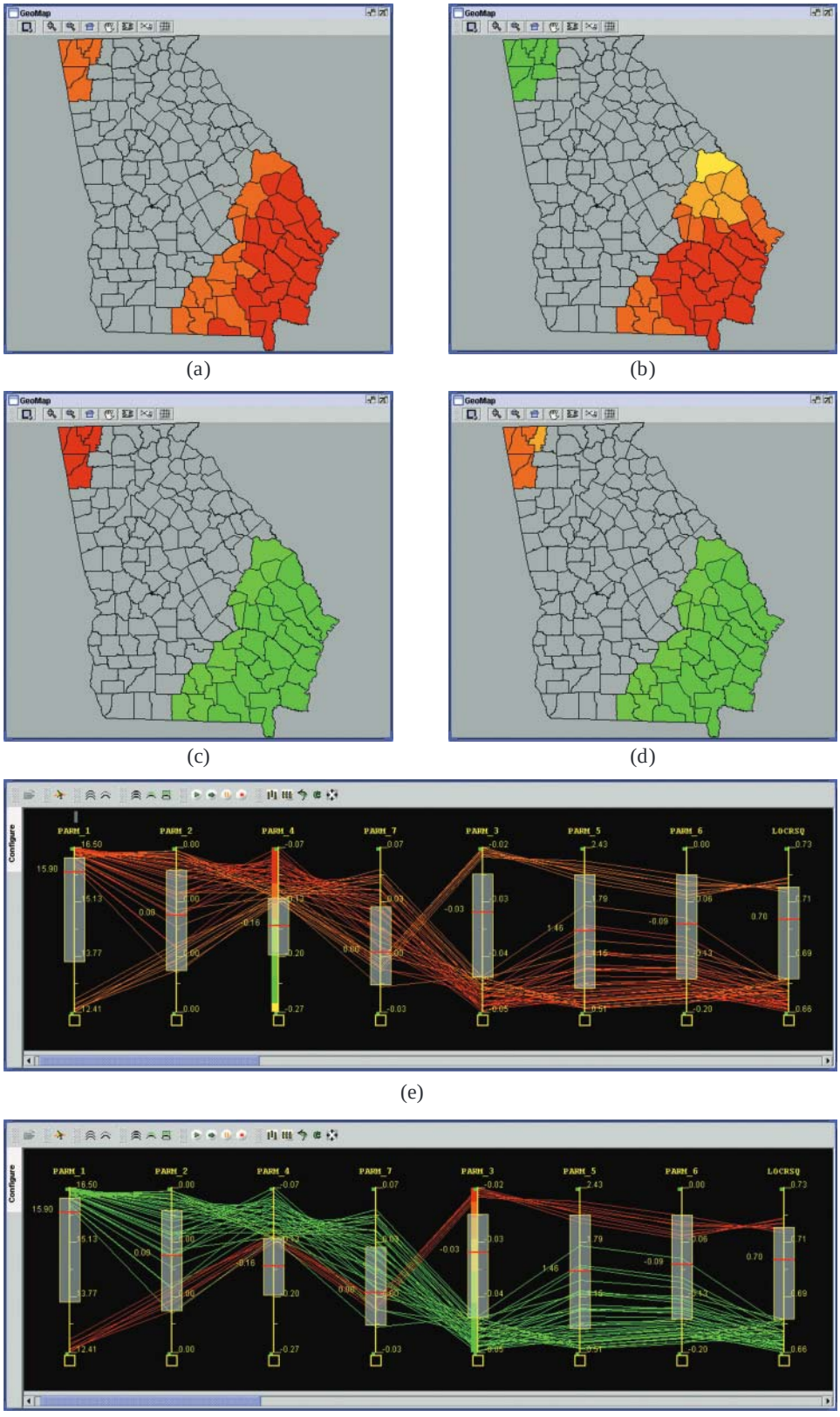


Figure 7. Selection of areas with high values of parameter 4: a) the PARM_4 map, b) the PARM_1 map, c) the PARM_3 map, d) the r^2 map, e) the PCP with colours of PARM_4, f) the same PCP with colours of PARM_3

BIOGRAPHICAL NOTES

Urška Demšar is a Postdoctoral Researcher at the National Centre for Geocomputation at the National University of Ireland, Maynooth. She has a PhD in Geoinformatics from the Royal Institute of Technology (KTH), Stockholm, Sweden and two degrees in Applied Mathematics from the University of Ljubljana, Slovenia. Previously she worked as a researcher and teacher at the Geoinformatics Department of the Royal Institute of Technology in Stockholm and as a teaching asistant in Mathematics at the Faculty of Electrical Engineering at the University of Ljubljana.

Her primary research interests are in Geovisual Analytics and geovisualisation. She is combining computational and statistical methods with visualisation for knowledge discovery from geospatial data. She is also interested in spatial analysis and mathematical modelling and has an established cooperation on spatial analysis of networks for crisis management with researchers at the Helsinki University of Technology, Finland.

ACKNOWLEDGEMENTS

Research presented in this paper was funded by a Research Professorship (03/RPI/1382) awarded to Professor Fotheringham by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support without which the work would not have been undertaken. The authors would also like to thank the two anonymous reviewers, whose comments helped to significantly improve this paper.

REFERENCES

Andrienko, G., Andrienko, N. and Voss, H. (2003). 'GIS for everyone: the CommonGIS project and beyond', in: **Maps and the Internet**, ed. by Peterson, M., pp. 131–146, Elsevier, Amsterdam.

Andrienko, N. and Andrienko, G. (2006). **Exploratory Analysis of Spatial and Temporal Data**, Springer Verlag, Berlin–Heidelberg.

Anselin, L., Syabri, I. and Youngihn, K. (2004). 'GeoDa: An Introduction to Spatial Data Analysis', **Geographical Analysis**, 38, 5–22.

Bação, F., Lobo, V. and Painho, M. (2005). 'The self-organizing map, the Geo-SOM and relevant variants for geosciences'. **Computers & Geosciences**, 31, 155–163.

Brunsdon, C., Fotheringham, A. S. and Charlton, M. (2007). 'Geographically Weighted Discriminant Analysis', **Geographical Analysis**, 39(4), 376–396.

Crespo, R., Fotheringham, A. S. and Charlton, M. (2007). 'Application of Geographically Weighted Regression to a 19-year set of house price data in London to calibrate local hedonic price models', in **Proceedings of the 9th International Conference on Geocomputation**, Maynooth, Ireland.

Demšar, U. (2007). 'Knowledge discovery in environmental sciences: visual and automatic data mining for radon problems in groundwater', **Transactions in GIS**, 11(2), 255–281.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2000). **Quantitative Geography – Perspectives on Spatial Data Analysis**, Sage Publications.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002). **Geographically Weighted Regression – the Analysis of Spatially Varying Relationships**, John Wiley & Sons Inc., Hoboken, New Jersey.

Fotheringham, A. S. and Brunsdon, C. (2004). 'Some thoughts on inference in the analysis of spatial data', **International Journal of Geographic Information Science**, 18(5), 447–457.

Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F. (2002). 'Introducing Geo-VISTA Studio: an integrated suite of visualisation and computational methods for exploration and knowledge construction in geography', **Computers, Environment and Urban Systems**, 26, 267–292.

Guo, D. (2003). 'Coordinating computational and visual approaches for interactive feature selection and multivariate clustering', **Information Visualisation**, 2, 232–246.

Guo, D., Gahegan, M., MacEachren, A. M. and Zhou, B. (2005). 'Multivariate Analysis and Geovisualisation with an Integrated Geographic Knowledge Discovery Approach', **Cartography and Geographic Information Science**, 32(2), 113–132.

Hurley, C. (2004). 'Clustering Visualisations of Multidimensional Data'. **Journal of Computational and Graphical Statistics**, 13(4), 788–806.

Inselberg, A. (2002). 'Visualisation and data mining of high-dimensional data', **Chemometrics and Intelligent Laboratory Systems**, 60, 147–159.

Jiang, B. and Harrie, L. (2004). 'Selection of streets from a network using self-organizing maps', **Transactions in GIS**, 8, 335–350.

Keim, D.A. and Ward, M. (2003). 'Visualisation', in: **Intelligent Data Analysis**, ed. by Berthold, M. and Hand, D. J., 2nd edition, pp. 403–428, Springer Verlag, Berlin–Heidelberg.

Kohonen, T. (1997). **Self-Organizing Maps**, 2nd edition, Springer Verlag, Berlin–Heidelberg.

Koua, E. L. and Kraak, M.-J. (2004). 'Alternative visualisation of large geospatial datasets', ***The* Cartographic Journal**, 41, 217–228.

MacEachren, A. M., Wachowicz, M., Edsall, R., Haug, D. and Masters, R. (1999). 'Constructing knowledge from multivariate spatio-temporal data: integrating geographical visualisation with knowledge discovery in database methods', **International Journal of Geographic Information Science**, 13(4), 311–334.

MacEachren, A. M. and Kraak, M.-J. (2001). 'Research Challenges in Geovisualisation', **Cartography and Geographic Information Science**, 28(1), 3–12.

MacEachren, A. M. (2008). Invited paper in preparation on Geovisual Analytics (title to be decided), **Information Visualisation**.

National Visualisation and Analytics Center (NVAC) (2005). **Illuminating the Path: Creating the R&D Agenda for Visual Analytics**. Available at: http://nvac.pnl.gov/agenda.stm.

Shneiderman, B. (2001). 'Inventing Discovery Tools: Combining Information Visualisation with Data Mining', in **Proceedings of the 12th International Conference on Algorithmic Learning Theory**, **Lecture Notes in Artificial Intelligence**, 2226, pp. 17–28, Springer Verlag, Berlin–Heidelberg.

Silipo, R. (2003). 'Neural Networks', in **Intelligent Data Analysis**, ed. by Berthold, M. and Hand, D. J., 2nd edition, pp. 269–320, Springer Verlag, Berlin–Heidelberg.

Skupin, A. and Hagelman, R. (2005). 'Visualizing Demographic Trajectories with Self-Organizing Maps', **Geoinformatica**, 9(2), 159–179.

Špatenková, O., Demšar, U. and Krisp, J. M. (2007). 'Self-Organising Maps for exploration of spatio-temporal emergency response data', in **Proceedings of the 9th International Conference on Geocomputation**, Maynooth, Ireland.

Takatsuka, M. (2001). 'An application of the self-organizing map and interactive 3-D visualisation to geospatial data', in **Proceedings of the 6th International Conference on Geocomputation**, Brisbane, Australia.

Takatsuka, M. and Gahegan, M. (2002). 'GeoVISTA Studio: a codeless visual programming environment for geoscientific data analysis and visualisation', **Computers & Geosciences**, 28, 1131–1144.

Theus, M. (2005). 'Statistical Data Exploration and Geographical Information Visualisation', in **Exploring Geovisualisation**, ed. by Dykes, J., MacEachren, A. M. and Kraak, M.-J., pp. 127–142, Elsevier, Amsterdam.

Unwin, A. and Unwin, D. (1998). 'Exploratory Spatial Data Analysis with Local Statistics', **The Statistician**, 47(3), 415–421.

Vesanto, J. (1999). 'SOM-based data visualisation methods', **Intelligent Data Analysis**, 3, 111–126.