Editorial Introduction

The date used by social scientists are frequently multivariate. In part, this is a consequence of a need to characterise objects of interest, such as people, houses and so on, as fully as possible, but it is also often a result of a desire to capture concepts such as social class or intelligence and overcrowding that do not permit easy measurement along one axis of variation. In consequence, quantitative social science has a long history of using statistical and mathematical transforms of data matrices such as factor and principal component analysis to reduce the dimensionality of these data and perhaps suggest appropriate constructs that might also be used to describe individuals.

These techniques are not intrinsically visual, although the reprojection of individual cases onto axes that define these constructs (for examples as component scores) may well create data that can be visualized by any of the standard techniques. There remains a need to develop appropriate alternative visualizations for multidimensional data that are efficient in allowing the detection of patterns in the multivariate data space. In this Case Study, Chris Brunsdon, Stweart Fotheringham and Martin Charlton develop and illustrate three alternative projections that can be applied to multivariate data.

It is interresting to note that although the static displays produces are in themselves useful, they gain maximum utility when visualized in an interactive environment.

An Investigation of Methods for Visualising Highly Multivariate Datasets

C. Brunsdon
Department of Town
and Country Planning
University of Newcastle

A. S. Fotheringham
Department of Geography
University of Newcastle

M. E. Charlton
Department of Geography
University of Newcastle

Abstract

Although visualisation has become a ‘hot topic’ in the social sciences, the majority of visualisation studies and techniques apply only to one or two dimensional datasets. Relatively little headway has been made into visualising higher dimensional data although, paradoxically, most social science datasets are highly multivariate. Investigating multivariate data, whether it be done visually or not, in just one or two dimensions can be highly misleading. Two well-known examples of this are the use of a correlation coefficient instead of a regression parameter as an indicator of the relationship between two variables and the use of scatterplots instead of leverage plots as indicators of relationships.

This project has therefore investigated several methods for visualising aspects of higher dimensional (i.e. multivariate) datasets. Although some techniques are quite well-established for this purpose, such as Andrews Plots and Chernov Faces, we have ignored these because of their well-know problems. In the case of Andrews plots the functions used are subjective and the plots become very difficult to read when the number of observations rises beyond 30. In the case of Chernov faces, variables which are attached to certain attributes of the face, for example, the eyes, receive more weight in the subjective determination of ‘unusual’ cases.

Instead we have examined the use of four newer techniques for visualising aspects of higher dimensional data sets: **projection pursuit; Geographically Weighted Regression; RADVIZ;** and **Parallel Co-ordinates**. In projection pursuit the objective is to project an m-dimensional set of points onto a two-dimensional plane (or a three-dimensional volume) by constrained optimisation. The choice of function to be optimised depends on what aspect of the data are the focus of investigation. The technique therefore offers a great deal of flexibility from identifying clusters of similar cases to identify outliers in multivariate space. A problem with projection pursuit though is that it is difficult to interpret because the projection plots produced are of indices produced by linear combinations of variables which might not have any obvious meaning.

The technique of Geographical Weighted Regression usefully allows the visualisation of spatial non-stationarity in regression parameter estimates. The output from the technique consists of maps of the spatial drift in parameter estimates which can be used to investigate spatial variations in relationships or for model development because the maps can indicate the effects of missing variables. Relatively little mention is made of Geographically Weighted Regression here because the authors have developed this technique and have written about it in a number of other sources.

The RADVIZ approach essentially involves calculating the resultant vector, for each case, of a series of m forces which are the m variables measured for that case. A plot of the locations of these resultants depicts the similarity in the overall measurements across the cases. It is particularly useful for compositional data, such as percentage shares of votes in elections. One drawback of the technique is that it is possible to get similar looking projections from quite different basic data properties and so the interpretation of RADVIZ needs some caution.

Finally, the parallel co-ordinates approach is perhaps the most intuitive of the four techniques we examined in that it is essentially a multidimensional variation on the scatterplot. Instead of two axes though, in parallel co-ordinates you can draw relationships between m axes which are depicted as parallel lines. However, the choice of ordering of the axes is influential to the depiction of relationships within the dataset and care must therefore be taken in selecting a particular ordering and the depiction of the data in parallel co-ordinates can get rather messy when large numbers of cases are involved.

1. Introduction

Suppose we have a set of m continuous observed variables for each of n cases, and denote the *j*th observation on the *i*th case by x_{ij}. Such a situation frequently arises when examining social data. For example, the cases might be census wards, and the observations might be rates computed from census variables, such as the percentage of households without cars, the percentage of households without central heating and so on. Before calibrating models based on these variables, it is generally useful to apply exploratory techniques to the data. Many of these techniques are graphical in nature - for example histograms, box plots or scatter plots may be drawn. However, these approaches are limited by the fact that they can only represent the relationship between at most two variables at any one time. In fact, apart from the scatter plot, the methods above only provide graphical representations of a *single* variable.

In order to decide how useful a representation is, one needs to consider the kind of feature in data that one wishes to detect. Three common possibilities in social science data are *clusters*, *outliers* and *geographical trends*. Clusters are distinct groupings in the data points, usually corresponding to multimodality in the underlying probability distribution for the data. Outliers are one-off cases that have unusual combinations of observed values, when compared to the remainder of the sample. Geographical trends are fairly self explanatory, but it is worth noting that as well as univariate trends, such as house prices increasing in certain areas, there could be trends in the *relationships* between some variables. It is also worth noting that these trends are rarely linear.

For most types of feature, there is variability in subtlety. For example an extremely high or low value of one particular variable would be a fairly crude type of outlier. This could be detected using a well-established univariate graphical tool such as a box-and-whisker plot (Velleman and Hoaglin, 1981). On the other hand, a more subtle outlier might be a point in the centre of a sphere, when all of the other points are close to its surface. The problem here is that none of the three coordinate values (x_1, x_2, x_3) defining the central point are unusual in their own right, and even worse, none of the possible coordinate value pairs such as (x_1, x_2) are unusual. Thus, no simple univariate or bivariate representation could detect this outlier. The problem would become even worse if instead of a sphere, a ten dimensional hypersphere were substituted in the previous example! Generally, more subtlety tends to imply a greater degree of sophistication required in the graphical representation. This leads to the statement of a fundamental problem: "*How can the interactions between large numbers of variables be represented in a managable number of dimensions?*".

In this Case Study, two data sets will be used to demonstrate a number of ways of addressing the above problem. The two data sets are described in detail in Appendix A, but, briefly, the first is a set of six socio-economic variables for northern England measured at census ward level, taken from the 1991 census, and the second is a simulated data set designed to have a `pathological' outlier, as discussed above. The following sections each describe a particular approach to visualisation, giving examples using the census data set. After these sections, a number of specific issues are considered, including a comparison of the way each method responds to the synthesised data.

2. The Projection Pursuit Approach to Visualisation

2.1 Context

Suppose, for a set of cases, m variables are recorded. Then each case can be thought of as a point in m-dimensional space. Unfortunately, unless m £• 3 it is not possible to view these points directly. However, it is possible to *project* an m-dimensional set of points onto a two-dimensional plane, or a three-dimensional volume. Here we will restrict the problem to projections onto two-dimensional planes. To visualise the concept of projection, figures 1 and 2 should be considered. In both figures a rectangular `screen' is shown, either above or to the right of a set of three dimensional data. Imagine a very bright light on the other side of the data points. The shadows thrown on the screen from the data points are the projection. The dotted lines in the diagram link the data points to their projected images.

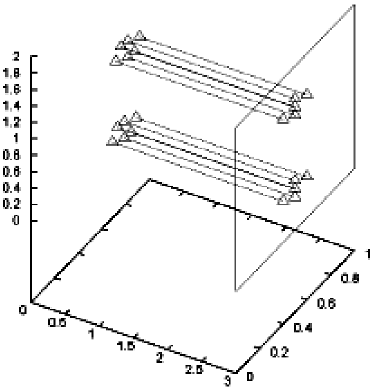


Figure 1: Example of point projection (1)

In figure 1 the data points are projected onto a plane to the right. Here the projected image shows two distinct clusters of points. In figure 2 the same data points are projected onto a plane above. Here the projected image shows only a single cluster of points. Obviously in this case the projection is from R^3 to R^2, but similar (and sometimes more complex) phenomena occur when the projection is from R^m dimensions and $m > 3$.

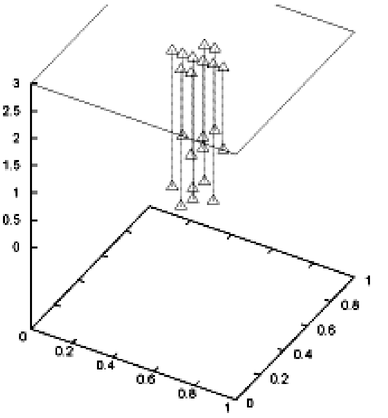


Figure 2: Example of point projection (2)

The above example demonstrates that different projections of the same data set can reveal different aspects of the data structure - indeed some projections can fail to reveal any structure. There are in fact an infinite number of possible projections to choose from, so which one should be used? *Projection pursuit* (Jones and Sibson, 1987} is concerned with resolving this kind of problem.

2.2 The Projection Pursuit Method

To see how this technique operates, it is first worth noting that projections from R^m to R^2 are linear mappings. Thus, if $\mathbf{X} = \{x_{ij}\}$ is a matrix whose (i,j)th element is the jth variable for observation i, we can write the general projection from R^m to R^2 as $(\mathbf{z}_1, \mathbf{z}_2) = (\mathbf{Xa'}, \mathbf{Xb'})$. Here **a** and **b** are m-dimensional row vectors defining the linear transform, and $\mathbf{z}_1$ and $\mathbf{z}_2$ are n-dimensional column vectors representing the points on the projection screen. The prime denotes transposition. Choosing a projection is now a matter of choosing **a** and **b**.

The next problem is to decide what kind of feature one wishes to detect. When this decision is made, one attempts to measure the degree to which this feature is exhibited in $(\mathbf{z}_1, \mathbf{z}_2)$. Call this measure $I(\mathbf{z}_1, \mathbf{z}_2)$ It is sometimes called the *index function*. For example, suppose one wishes to detect clusters. A common test statistic for clustering in two dimensional data is the *mean nearest neighbour distance* (MNND). Lower values of this statistic indicate greater clustering. Thus, here $I(\mathbf{z}_1, \mathbf{z}_2)$ is the MNND for the data set $(\mathbf{z}_1, \mathbf{z}_2)$. Noting that the expression $I(..,..)$ can be written in the form $I(\mathbf{Xa'}, \mathbf{Xb'})$, the projection choice problem can be thought of as an optimisation problem in which **a** and **b** must be chosen to minimise I. Essentially, this is the projection pursuit process.

At this stage, however, some careful thought should take place. Clearly the nearest neighbour distance index is scale dependent. Multiplying **a** and **b** by a constant will also multiply the MNND by this factor - so that one can make I as small as one likes with an appropriate choice of constant. This problem can be avoided by adding the constraint that $\mathbf{z}_1$ and $\mathbf{z}_2$ are standardized - that is that they both have a mean of zero and a variance of one. Also, it is helpful to add the constraint that $\mathbf{z}_1$ and $\mathbf{z}_2$ are not correlated. This ensures that maximum information is given in the two dimensional plot, in the sense that if two variables are correlated, they are `sharing' some underlying one-dimensional feature pattern rather than each representing different patterns.

The projection pursuit algorithm is thus equivalent to a constrained optimisation problem. The difficulty with the specification given is that the constraints are given in terms of $\mathbf{z}_1$ and $\mathbf{z}_2$ rather than **a** and **b**. However, suppose each variable in **X** is mean-centred and then transformed to its principal components - and the principal components are standardised so that each has a variance of one, giving a transformed matrix **Q**. **Q** is a linear transform of **X**, say **XP**, where **P** is an m by m matrix, so that a linear mapping of Q is also a linear mapping of **X**. Thus, we can re-write the index in the form $I(\mathbf{z}_1, \mathbf{z}_2) = I(\mathbf{Qc'}, \mathbf{Qd'})$. In this case, the column vectors **c** and **d** take the same form as **a** and **b**. In fact $\mathbf{Pc'} = \mathbf{a}'$ and $\mathbf{Pd'} = \mathbf{b}'$. The advantage of the re-stated problem is the fact that if $\mathbf{c'c} = 1$, $\mathbf{d'd} = 1$ and $\mathbf{c'd} = 0$ then $\mathbf{z}_1$ and $\mathbf{z}_2$ will be uncorrelated, have zero mean and variance of one. Thus, if constraints are imposed on **c** and **d** then $\mathbf{z}_1$ and $\mathbf{z}_2$ automatically satisfy the constraints proposed above. Thus, the projection pursuit problem may be stated:

$$
\begin{array}{ll}
\text{Minimise} & I(\mathbf{Qc'}, \mathbf{Qd'}) \\
\text{Subject to} & \mathbf{c'c} = 1 \text{ and} \\
 & \mathbf{d'd} = 1 \text{ and} \\
 & \mathbf{c'd} = 0
\end{array}
$$

This is a standard form for a constrained optimisation problem. Computationally, the difficulty of this problem depends on the index function I. Applying the method with I as the MNND is fairly intensive, mainly because it involves finding the nearest neighbour of every point in the projected data set.

In figure 3 the result of applying this technique to the census data is shown. Due to the nature of the index function, the rotation of the point pattern obtained is arbitrary, so there is no clear interpretation of the individual axes in the plot.

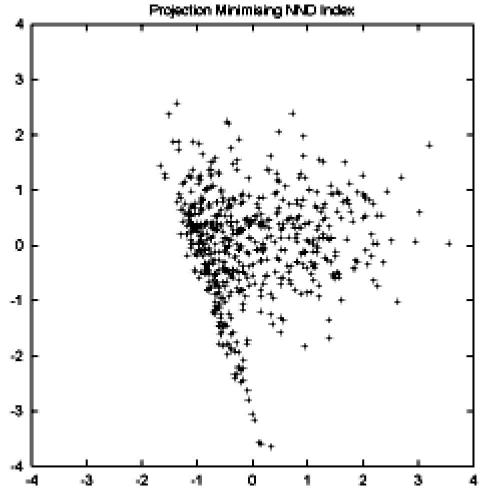


Figure 3: Minimised MNND projection of census data

No obvious clusters exist in the plot, suggesting perhaps that the data is not bimodal in any way detectable by projecting onto a two dimensional plane. However, some features are very clear, most notably a `spur' in the lower part of the plot.

2.3 Interpretation

Having obtained an optimal projection, it is essential that this can be easily interpreted. Since the projection is a linear mapping, interpretation is fairly straighforward. Having optimised in terms of **c** and **d**, one may work backwards to obtain **a** and **b**. If *j*th individual elements of these vectors are a_j and b_j, then a unit change in the *j*th original variable causes a change (a_j, b_j) in the projection space. Since the projection is linear, this statement is independent of the values of other variables. Also due to linearity, a change by an amount k in the *j*th variable leads to a change (ka_j, kb_j) in the projection space. Using this fact, one can plot the `change vectors' for a given point in the plot in projection space when each of the initial variables changes by one standard deviation. This is illustrated in figure 4.

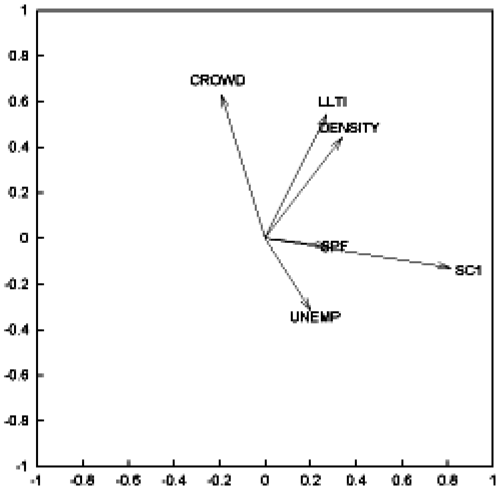


Figure 4: Minimised MNND projection of census data - Interpretation Plot

This gives some clues as to the variables causing the `spur' in the projection shown in figure 3. Although a number of possible variable combinations could cause this, figure 4 suggests that very high unemployment levels or low crowding could cause this, perhaps with low levels of the other variables.

2.4 Choice of *I*

At this point, some further discussion about the choice of the index function, *I*, might be appropriate. The example above was chosen to maximise clustering in the projected data image. However, other functions could be chosen to reflect other desired properties of the projection. Another important data feature is the presence of outliers. One way of `rewarding' projections that produce outliers is to negate the previous measure, or equivalently to *maximise* the MNND subject to the previous constraints. To see this, note that outliers are a long way from their nearest neighbours. When there are a large number of outliers, or one or two very extreme ones, the MNND will tend to be large.

The result of this approach is shown in figure 5, together with an interpretation plot in figure 6. The spur feature has now completely disappeared, and the projected points now form a more symmetrical pattern, but a number of outliers are visible in many directions around the outside of the cloud. The interpretation plot should help in identifying the nature of the outliers as in the previous example.

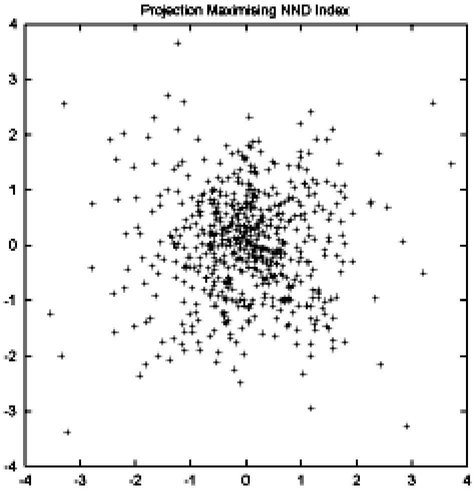


Figure 5: Maximised MNND Projection of Census Data

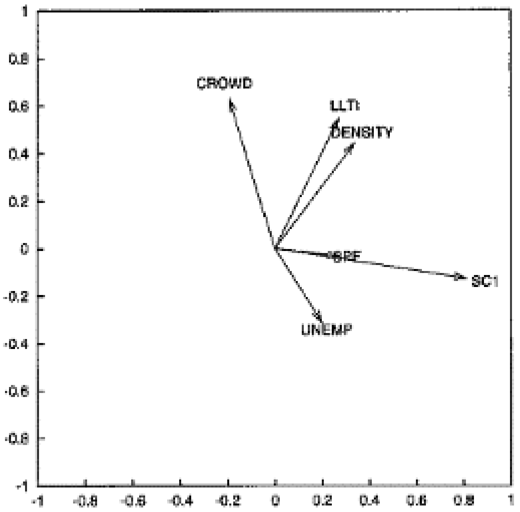


Figure 6: Maximised MNND Projection of Census Data - Interpretation Plot

A further possibility is to consider means of highlighting geographical trends. In this case, the idea of `projection' takes a different form. Here, a one-dimensional projection is used, *z* say, but the value of this projection is shown on a choropleth map. If this approach is taken, the projection should be chosen to emphasize geographical trends. One way of making this choice is to consider the *spatial autocorrelation* of the trends. To find trends which vary smoothly over the geographical study area, one needs to maximise the degree of spatial autocorrelation of the index. Similarly, to highlight *local differences* in data, one needs to *minimise* the spatial autocorrelation. Each of these goals could be achieved by defining *I* in terms of spatial autocorrelation. Morans (Moran, 1948} measure of spatial autocorrelation may be written as

$$\frac{n \quad \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} \quad (z_i - \bar{z}) \quad (z_j - \bar{z})}{\left(\sum_{i=1}^{n} (z_i - \bar{z})^2 \right) \left(\sum \sum_{i \neq j} s_{ij} \right)}$$

where s_{ij} is one if zones i and j are neighbours, and zero if they are not. Neighbourhood may be defined in a number of ways. Typically zones are neighbours if they share a common boundary, or if their centroids are less than some distance apart. If the z values are standardised to have variance one and mean zero, then the above expression simplifies to

$$\frac{\sum_{i-1}^{n}\sum_{j-1}^{n} s_{ij} z_i z_j}{\left(\sum\sum_{i \neq j} s_{ij}\right)}$$

If we are attempting to maximise or minimise this expression, the denominator may be ignored, since it is a positive constant. Thus, for projection pursuit designed to high geographical relationships, a suitable I is given by

$$\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} z_i z_j$$

Table 1:Projection Persuit Coefficients - Autocorrelation

Variable	Maximum *I*	Minimum *I*
CROWD	0.516	0.952
DENSITY	-0.296	-0.652
LLTI	0.555	-0.396
SCI	0.362	-0.630
SPF	-0.048	-0.317
UNEMP	1.098	0.111

Thus, here the projection pursuit problem can be stated as

Minimise or Maximise	$I(\mathbf{Qc}')$
Subject to	$\mathbf{c}'\mathbf{c} = 1$

since in this case I is simply a quadratic expression, and there is just one constraint, the computational overheads are much lower than for the MNND-based problem.

Interpretation of the single-dimensional projection is probably best done by tabulating the elements of **a**, possibly scaling by the standard deviation of each variable. This shows the degree and direction of change that would be seen in z if a given variable were to increase by one standard deviation.

Applying the method to the census data gives the maps in Appendix B, which show indices for maximum and minimum spatial autocorrelation respectively. The coefficients of projection (adjusted for scale) are given in table 1. Here it can be seen that the maximising map mostly picks up an urban/rural trend, whereas more subtle differences are picked out in the minimising map. In particular it highlights the way some nearby rural areas differ. The strongest contributing variables in the maximising case are CROWD, LLTI and UNEMP. It is suggested that this linear combination of variables is perhaps a useful indicator of 'urbanness' in the sense that high values tend to coincide with inner cities and low values with rural areas. On the other hand, the coefficients for the minimising case give a very different index. This index is useful for differentiating between nearby places, and is more strongly influenced by variables that are more spatially variable. A good example is SPF which has a much greater weighting in the minimising index. Although there is no strong geographical trend in the proportion of single parent families, it can be used as a means of differentiating between nearby places. Another variable that does this is CROWD which is possibly a differentiator between affluent and poor rural communities.

2.5 Geographically Weighted Regression

At this point, another trend-based method of analysis should be considered briefly. This is the method of *Geographically Weighted Regression* (GWR), see for example Brunsdon *et al.* (1996). In this approach, a multivariate regression is carried out, but instead of a *global* model, localised models are fitted around a number of points in the study area. For example, using the LLTI data set from the previous section, a number of sample points in northern England are chosen, and, taking a circle drawn around each point, a `local' regression is calibrated. Typically, this is a weighted regression, and the weight given to each observation, in this case a Census ward, decays with the distance from the sample point. Thus, eventually, a different regression is calibrated for each sample point. Mapping the way the regression coefficients change for a series of sample points spread throughout the study region shows geographical changes in the relationship between the variables. Typically the sample points are placed on a regular grid, or centred on the areal unit centroids. Note that it does not matter if the circles centred on the points overlap; indeed this allows smoother trends in the regression coefficients to be mapped.

Note that this differs from the projection pursuit using Moran's-I in two major ways. First, whereas the projection pursuit method produces just one map, (or two if Moran's I is both minimised and maximised), GWR produces a map for each regression coefficient, plus one for the intercept coefficient. Secondly, project pursuit treats all variables identically, whereas GWR requires that one variable has to be `singled out' as the dependent variable. A comprehensive example of the technique is given in Brunsdon *et al.* (1996). Although this method does not fall directly into the projection pursuit category, one way of viewing the regression model is as a `best fit' linear projection, and this is a useful approach to finding geographical trends in such projections.

3. The RADVIZ Approach to Visualisation}

Like the previous approach, the RADVIZ method (Ankerst et al., 1996) maps a set of *m*-dimensional points onto two dimensional space. However, in this case the mapping is nonlinear.

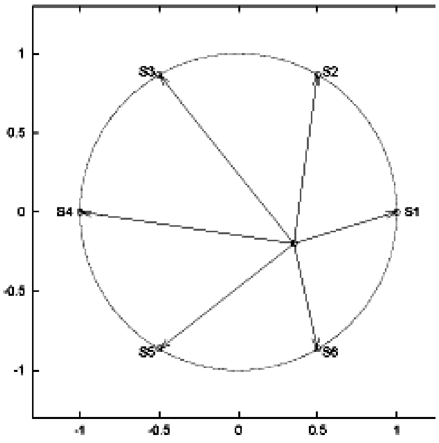


Figure 7: The Physical System Basis for RADVIZ

To explain the RADVIZ approach, it is helpful to imaging a physical situation. Suppose *m* points are arranged to be equally spaced around the circumference of the unit circle. Call these points S_1 to S_m. Now suppose a set of *m* springs are fixed at one end to each of these points, and that all of the springs are attached to the other end to a puck, as in figure 7.

Finally, assume the stiffness constant (in terms of Hooke's law) of the *j*th string is x_{ij} for one of the data points *i*. If the puck is released and allowed to reach an equilibrium position, the coordinates of this position, $(u_i, v_i)^T$ say, are the projection in two dimensional space of the point $(x_{i1}, \ldots, x_{im})^{\mathrm{T}}$ in *m*-dimensional space. This, if $(u_i, v_i)^T$ is computed for $i = 1 \ldots \mathrm{n}$, and these points are plotted, a visualisation of the *m*-dimensional data set in two dimensions is achieved.

To discover more about the projection from $R^m \rightarrow R^2$, consider the forces acting on the puck. For a given spring, the force acting on the puck is the product of the vector spring extension and the scalar stiffness constant. The resultant force acting on the puck for all *m* springs will be the sum of these individual forces. When the puck is in equilibrium there are no resultant forces acting on it and this sum will be zero. Denoting the position vectors of S_1 to S_m by $\mathbf{S}_1$ to $\mathbf{S}_{\mathrm{m}}$, and putting $\mathbf{u_i} = (u_i, v_i)^T$ we have

$$\sum_{j=1,m}(\mathbf{S}_j - \mathbf{u}_i)x_{ij} \quad = \quad 0$$

which may be solved for $\mathbf{u_i}$ by

$$\mathbf{u}_i \quad = \quad \sum_{j=1,m} w_{ij}\mathbf{S}_j$$

where

$$w_{ij} \quad = \quad \left(\sum_{j=1,m} x_{ij}\right)^{-1} \quad x_{ij}$$

Thus, for each case *i*, $\mathbf{u}_i$ is simply a weighted mean of the $\mathbf{S}_j$'s whose weights are the *m* variables for case *i* normalised to sum to one. Note that this normalisation operation makes the mapping from $R^m \rightarrow R^2$ nonlinear.

Viewing the projection in this explicit form allows several of its properties to be deduced. First, assuming that the x_{ij} values are all non-negative, each $\mathbf{u}_i$ lies within the convex hull of the points $\mathbf{S}_1$ to $\mathbf{S}_{\mathrm{m}}$,. Due to the regular spacing of these points, this convex hull will be an *m*-sided regular polygon. Note that if some of the x_{ij} values are negative this property need not hold, but that often each variable is re-scaled to avoid negative values. Two typical methods of doing this are the local metric (L-metric) rescaling, in which the minimum and maximum values of x_{ij} for each *j* are respectively mapped onto zero and one respectively:

$$x_{ij}^{*} \quad = \quad \frac{x_{ij} - \min(x_{ik} \mid k = j)}{\max(x_{ik} \mid k = j) - \min(x_{ij} \mid k = j)}$$

and the global metric (G-metric), in which the rescaling is applied to the data set as a whole, rather than on a variable by variable basis:

$$x_{ij}^{*} = \frac{x_{ij} - \min(x_{ik})}{\max(x_{ik}) - \min(x_{ij})}$$

in each case, the rescaled x_{ij} values will all lie in the interval [0,1].

The weighted centroid interpretation of the projection also allows some other properties to become apparent. If, for a given *i*, the values of x_{ij} are constant, $\mathbf{u}_i$ will be the zero vector. This is a rather strange property, since it implies that observations in which all variables take on a very high constant value (once re-scaled) will be projected onto the same point as observations in which all variables take on a very low constant value. More generally, it suggested that observations which take on very similar values for re-scaled data will be mapped into regions close to the origin.

A RADVIZ projection for the census data is shown in figure 8. The result is superficially similar to the maximised MNND projection pursuit, showing a circular cluster of points and identifying outliers around this.

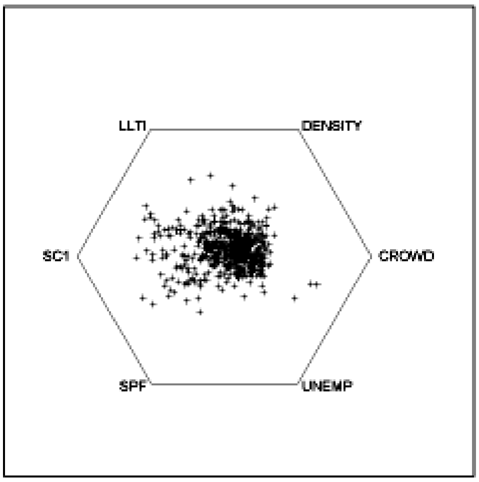


Figure 8: A RADVIZ Projection of the Census Data

For general data sets this property could lead to difficulties in interpreting the plots, but it is particularly useful when considering *compositional* geographical data. Suppose the population of a geographical region is classified into *m* categories, for example, those aged under 18, those aged 18 to 65 and those aged over 65. Another example would be voting data for an electoral ward where the categories are the parties which each constituent voted for. A compositional data set is one in which each of the variables represents the proportion (or percentage) of each category for each area. Numerically, the most notable property of such data is that for each case the variables sum to 1 (or 100 if percentages are used). This constraint suggests that the only way that all *m* variables can be equal is when they all take the value 1/*m*. Note, however that even in this case, the fact that $\mathbf{u}_i = 0$ does not imply that all proportions are equal. For example, if a pair of variables are represented by diametrically opposite points on the circle, and the proportions are 0.5 in each of these, then this will also give $\mathbf{u}_i = 0$. Another aid to interpretation for compositional data is that if an area consists entirely of one category then the corresponding variable will take the value 1, while the others will take the value zero. This implies that $\mathbf{u}_i$ will lie on the vertex of the regular *m*-sided polygon corresponding to that category.

In the case $m = 3$ for compositional data the RADVIZ procedure produces the compositional triangular plot used for various purposes - notably by Dorling (1990) and Coombes *et al.* (1996) to illustrate voting patterns in Britain in a three-party system. If we were to extend the analysis to look at more than three parties (for example by considering the nationalist vote as a fourth option) then

RADVIZ provides a natural extension of this concept. The main difficulty when moving beyond m = 3 for compositional data is that points on a RADVIZ plot no longer correspond *uniquely* to (x_{i1}, … ,x_{ij}) for a given case, more than one composition can project onto the same $\mathbf{u}_i$, as discussed above.

It is also interesting to note that for a given set of variables, there are several possible RADVIZ projections, since the m initial *m* variables could be assigned to $\mathbf{S}_1$ … $\mathbf{S}_\mathrm{m}$ in *m*! different ways. If we are mostly interested in identifying clusters and outliers, a number of possible projections will be essentially equivalent, since they will be identical up to a rotation or a mirror image. To see how many non-trivially different permutations there are, we need first to note that any permutation can be rotated *m* ways (i.e. rotation through 360/*m* degrees, by 2(260/*m*) degrees and so on up to (*m* - 1)(360/*m*) degrees, and of course the identity rotation through zero degrees), and so we need to divide the *m*! by *m*. We then note that any permutation can be reflected in two ways (i.e. mirror imaged or left alone) so the figure of (*m* - 1)! must be halved. Thus, if *m* is the number of variables, there are effectively (*m* - 1)! / 2 possible RADVIZ projections.

One way of deciding which of these should be used is to use an index, in a similar manner to projection pursuit in the previous section. In fact, the same indices could be used - for example maximising the variance of the $\mathbf{u}_i$'s or using one of the nearest-neighbour distance based criteria. In this case, the optimisation is a discrete search over a finite number of possibilities, rather than a continuous multivariate optimisation problem as in projection pursuit. It should be noted that the number of options to be search increases very rapidly with *m* (worse than m^2), and this has implications for computation. Clearly investigation into optimisation heuristics for this problem is necessary if it is to be applied in cases where *m* is very large.

4. The Parallel Coordinates Approach to Visualisation

In this final section on approaches to visualising multidimensional data sets, a very different approach is taken. In both of the previous techniques, a point in *m*-dimensional space was mapped onto a point in 2-dimensional space. In this approach, a point in *m*-dimensional space is represented as a series of *m*-1 line segments (Inselberg *et al.*, 1987) in 2-dimensional space. Thus, if the original data observation is written as (x_1, x_2, … x_m) then, its parallel coordinate representation is the *m*-1 line segments connecting the points (1,x_1), (2,x_2), … (m,x_m}. Each set of line segments could be thought of as a `profile' of a given case. The shape of the segments conveys information about the levels of the *m* variables. This is illustrated in figure 9. Typically, continuous variables will be standardised before a parallel coordinate plot is drawn.

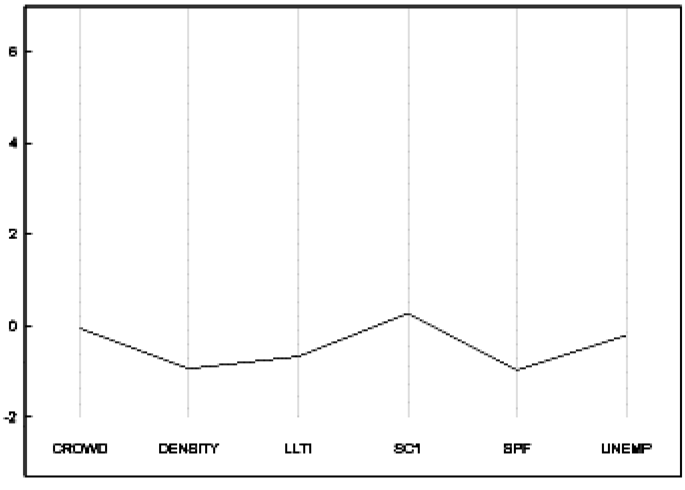


Figure 9: A Parallel Coordinate Representation of One Case

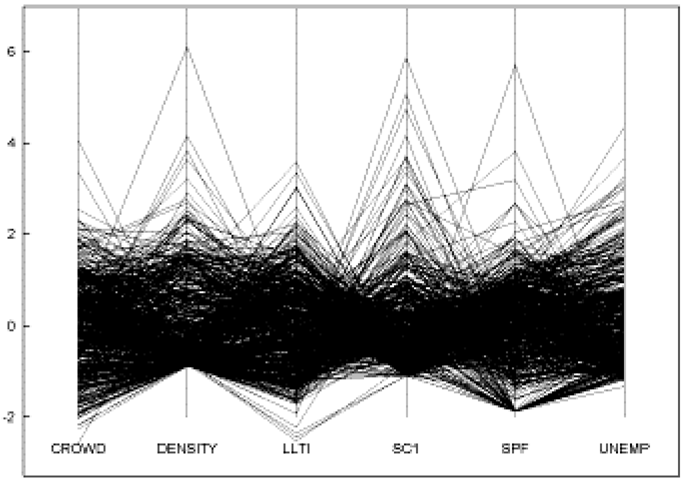


Figure 10: A Parallel Coordinates Plot

To view an entire *m*-dimensional data set one simply plots *all* such profiles on the same graph. This is illustrated in figure 10. For large data sets, the appearance of such a plot appears confusing, but can be used to highlight outliers. However, the real strength of the technique can be seen when subsets of the data are selected, usually on the basis of one particular variable.

To see this, consider figure 11. Here, the subset of the data in the lowest decile of the variable **LLTI** is shown in black, and the remainder of the dataset in grey. Looking at the relative locations of the black and grey lines shows the distribution of the data values in the subset in relation to the entire data set. Obviously, all of the black lines pass through the lowest section of the **LLTI** axis. However, looking at the locations of the black lines on the other axes shows whether the low values of this variable tend to be accompanied by any notable distributional patterns in the other variables. From the plot, it may be seen that often there are also low values of **DENSITY** and **UNEMP**.

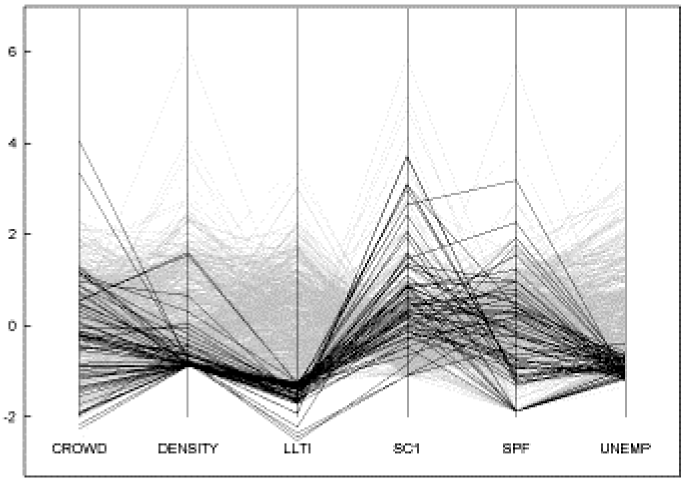


Figure 11: Parallel Coordinates Plot, Lowest Decile of LLTI Highlighted

Parallel plots may also be used to detect outliers in two dimensions. Again looking at figure 11, there are a few cases in the subset where **DENSITY** is unusually high, *given* the low value of **LLTI**. It is also apparent that this phenomenon does not occur with the variable **UNEMP**. This technique can also be used, at least sometimes, to detect three-dimensional outliers. For example, the black line joining a high(ish) value of **SC1** to a similar value of **SPF** is unusual: first in a two-dimensional sense because it appears unusual that the two variables *both* have high values, and second in three dimensions because we can also see that this line is black and therefore associated with the lowest decile of **LLTI**.

Outliers detected in terms of the lines connnecting pairs of axes in the parallel system pose an interesting problem. Although the method provides a striking image of outliers between two variables, it only works if the two variables have neighbouring parallel axes. For m variables, there are only (m-1) such neighbours possible, but there are m(m-1)/2 possible variable pairs. Thus, (m-1)(m-2)/2 pairs cannot be displayed. The problem is similar to the ordering problem in RADVIZ, that is the patterns that parallel coordinate plots yield depend on the ordering of the axes. In this case, there are m! possible orderings, although if we assume that reversing the order of the axes generates equivalent patterns, this leaves m!/2 possibilities. Again, as with RADVIZ, we are left with a combinatorial computational problem.

One approach to this might be to maximise the variability of the centre points of lines between pairs of variables. If these are well-separated then this makes patterens or outliers easier to detect. Suppose v_1 and v_2 are variables with neighbouring axes, then the midpoint on the plot will have a height of (v_1 + v_2)/2. The horizontal coordinate is not of interest, as it will be fixed for all values of v_1 and v_2. Suppose also that v_1 and v_2 are standardised, and so have variance of one. Then, the variance of the height will be

$$\frac{1+\rho}{2}$$

where ρ the the correlation between v_1 and v_2. If these quantities are added together for each pair of adjacent axes, an index score is created. Choosing a suitable ordering is then a matter of maximising this quantity. In fact, the problem may be simplified by replacing the above expression with ρ, or changed to a minimisation problem by replacing ρ with 1- ρ. In fact, this problem is equivalent to the travelling salesman problem. To see this, regard 1-ρ 's between pairs of variables as lengths of trips between towns, and axis ordering as visit ordering for the towns. Total distance is then equivalent to total 1-ρ 's, which is minimised in the travelling salesmen problem. A large amount of research into this mathematical problem has been carried out, and, although solutions are possible they require large amounts of computational effort. It is also worth noting that other indices besides the correlation sum could be used to choose an ordering, so that, for example, clustering of centre points into multimodal groups could be rewarded, in a similar manner to projection pursuit. So long as the measure used is a sum of two-way interactions between the variables, the equivalence to the travelling salesman problem applies.

5. Making Use of User Interaction

All of the above methods may be enhanced by the introduction of *user interaction*. In particular, the use of *linked plots*, where the output from any of these techniques could be linked to another view of the data, using the technique of *data brushing* - see for example Tierney (1990). Two particularly useful techniques are those of *linked maps* and *slicing*. The first of these is documented in Brunsdon and Charlton (1996), and in short involves highlighting zones on a map corresponding to selected points on a scatterplot, or *vice versa*. This is particularly useful if one of the projection-based techniques is used. For example, one can check whether the spur in the minimising MNND projection pursuit corresponds to any particular geographical pattern, as in the screenshot in figure 12.

Another useful interactive approach is *slicing*. In this case, points in a scatterplot are selected according to the value of an auxiliary variable. This value is controlled by a slider button, as in the second screenshot, figure 13. The value shown in the slider is the central point of a decile `slice' of the data, based on the values of the variable **LLTI** - clearly any other variable could also be used. Moving the slider causes the highlighted points in the scatterplot to change - so one can see which regions of the projection correspond to high and low values of the slicing variable. This method helps to interpret the patterns seen in projection-based methods such as projection pursuit and RADVIZ.

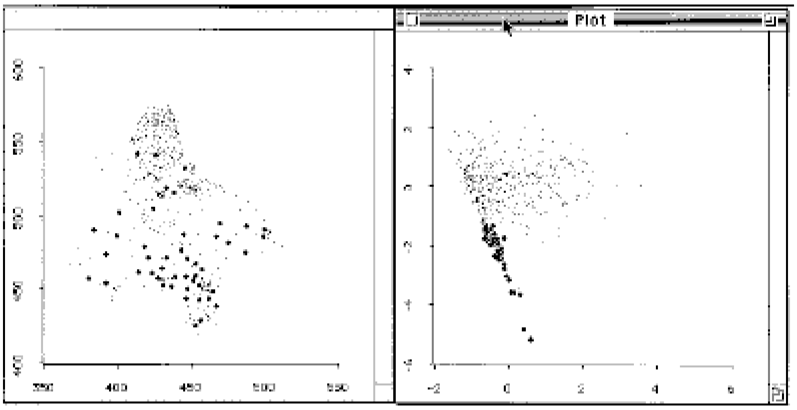


Figure 12: A Map-Linking Screenshot

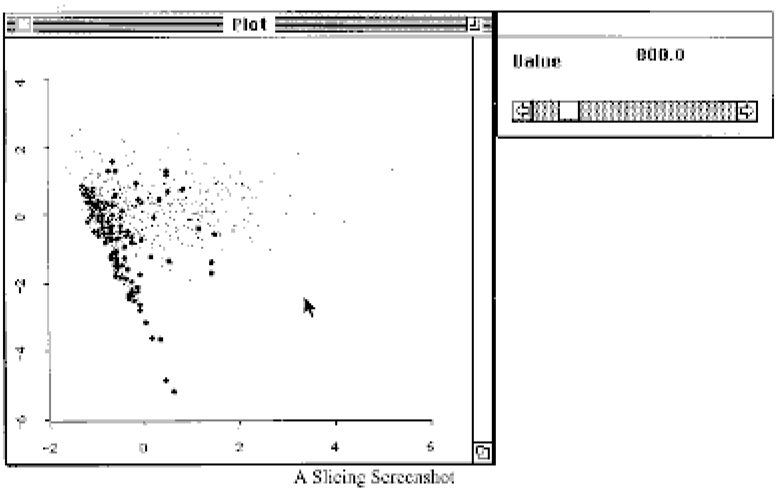


Figure 13: A Slicing Screenshot

6. Finding Outliers - the Synthesised Data Set

Each of the methods of visualisation outlined above was applied to the synthesised data set described in the appendix. As suggested in the appendix, this data set was deliberately chosen to exhibit `pathological' behaviour, in terms of an outlier. This outlier can be thought of as a sixth order outlier, in that no five-way combinations of the variables V_1 to V_6 appear to have any unusual observations. The outlier lies on the six-dimensional point (0,0,0,0,0,0), whilst all other observations satisfy $V_1^2{+}V_2^2{+}V_3^2{+}V_4^2+V_5^2{+}V_6^2=1$.

In figure 14, the parallel coordinates plot of the data is shown. As is typical of the technique, it is subset selection and highlighting that brings out patterns in the data. Here, the darker lines correspond to cases where $|V_2|<0.1$. From this, it is clear that an unusual value of V_1 occurs. In fact, selecting this line only would then reveal the straight line through the zero point of all six parallel axes. Thus, the parallel coordinates plot has shown reasonable success in detecting the outlier.

Note that the parallel axes are calibrated in terms of *standardised* variables, so that the ranges for V_1 to V_6 extend beyond the range from -1 to 1.

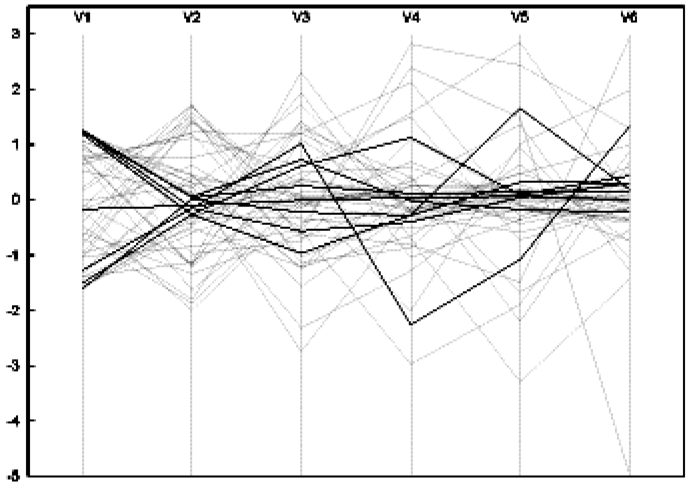


Figure 14: A Parallel Coordinates Plot of the 'Pathological' Data Set

Next, consider the projection pursuit approach to the data set. The results are shown in figures 15 and 16. Clearly, these do not show any obvious outliers. It is possible that the problem here is that, due to the nature of this particular data set, there are no linear projections that are able to identify the point at the origin. If one considers the case in three dimensions, it is possible to envision the difficulty.

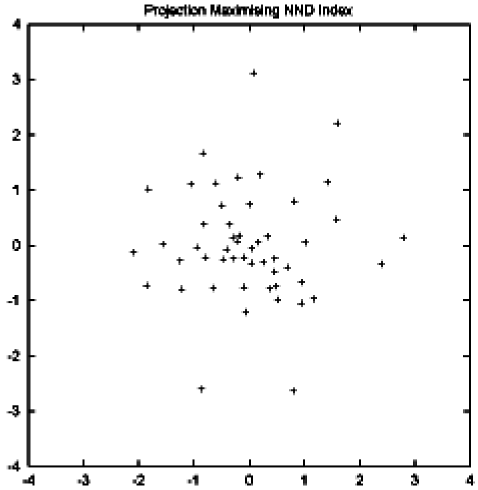


Figure 15: Maximum MNND Projection Plot of the 'Pathological' Data Set

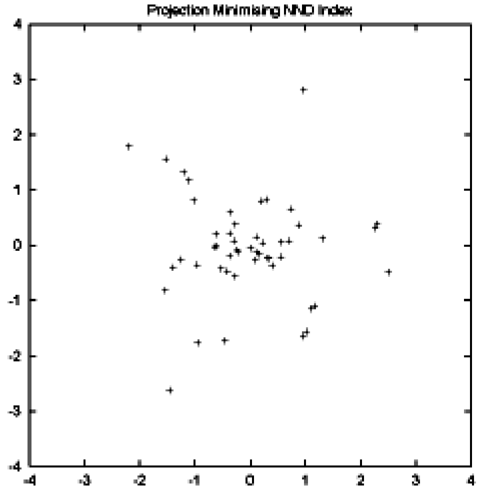


Figure 16: Minimum MNND Projection Plot of the 'Pathological' Data Set

Similarly disapointing results are experienced with RADVIZ (see figure 17). It is likely that similar arguments occur here (although the mapping from six to two dimensions is no longer linear).

It seems that unmodified projection-based approaches do not work well with the data set generated here. However, it is possible that some more flexible non-linear approaches might be helpful. For example, if one were to analyse the *squares* of the variables, the outlier would be much more easily detected. To see this note that $V_1^2+V_2^2+V_3^2+V_4^2+V_5^2+V_6^2$ is equal to one for all points except the outlier, when the expression is equal to zero. However, one would have to have very strong prior knowledge to consider using this particular transform!

It should also be noted that the slicing technique discussed in the previous section might have helped to highlight the outlier in the projective methods. It is significant that the parallel coordinates plot showed few patterns until a subset was selected and highlighted.

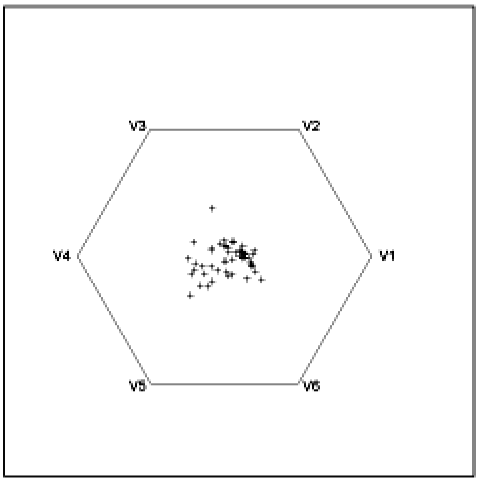


Figure 17: RADVIZ Plot of the 'Pathological' Data Set

7. Conclusions

This study has shown that there are several possible ways to visualise multidimensional data and that each has merits and pitfalls. The projection pursuit approach, particularly in the autocorrelation form, is able to identify patterns in linear combinations of the data that perhaps would not be unearthed with any of the other techniques. This is particularly true of geographical patterns. These can either be discovered in a very direct way (the autocorrelation approach) or by using packages such as XLISP-STAT to link two-dimensional projections with maps. The technique also offers a great deal of flexibility. The index function *I* can be chosen in many different ways, and in each case an optimal projection for serving some very specific purpose can be found.

On the negative side, the technique is perhaps one of the hardest to interpret. This is mainly due to the fact that the projection plots produced are of indices produced by linear combinations of variables - and one has the problem of assigning meaning to these indices. Interpretation plots such as figues 4 and 6 are of some help, but they still leave some ambiguity. This is perhaps inevitable, as any linear projection from a higher dimensional space onto a lower one is likely to map several points in the domain space onto the same point in the image. For the method to come into its own, it is usually necessary to use the projection as a single view in a system of linked views in an interactive system, as suggested above.

The RADVIZ approach is also a projective technique, and so many of the comments that might be applied to projection pursuit also apply here, particularly those relating to the interpretation of plotted patterns. However, when exploring compositional data RADVIZ comes into its own. In this instance, the patterns have a very intuitive interpretation in that points near to the vertices of the regular polygon are correspond to observations dominated by a particular component of the compositional breakdown. If the vertices are labelled by their corresponding variables, as in figure 8, then it becomes immediately clear which variable it is. It is also worth noting that the RADVIZ projection of the census data was quite similar to the maximising MNND projection, but required considerably less computational effort to achieve.

The parallel coordinates approach is perhaps the most intuitive. The labelling of the axes makes it very clear exactly what values individual variables take - a property which none of the other approaches have. As with RADVIZ, the choice of representation for a given data set is not unique, and the problem of choosing an *optimal* representation is a difficult one. In this case, a choice of the ordering of the parallel axes must be made. However, in the author's experience, non-optimal choices of parallel coordinate axes can work reasonably well such that in many cases sub-optimality may not imply unacceptably poor quality. Perhaps a useful compromise is to allow the user to swap the axes interactively, and explore more than one possible axis ordering. Indeed, a similar approach could be applied with RADVIZ.

8. References

Ankerst, M., Keim, D., and H.-P., K. 1996. Circle segments: A technique for visually exploring lagre dimensional data sets. In *Proceedings of the IEEE Visualization Conference*.

Brunsdon, C. and Charlton, M., 1996. A spatial analysis development system using Lisp. In Parker, D., editor, *Innovations in GIS 3*, Taylor & Francis, London.

Brunsdon, C., Fotheringham, A. and Charlton, M., 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis* 28:281-289.

Inselberg, A., Tuval, C. and Reif, M., 1987. Convexity algorithms in parallel coodinates. *Journal of the ACM* 34: 765-801.

Jones, M. and Sibson, R., 1987. What is projection pursuit (with discussions). *Journal of the Royal Statistical Society* 150: 1-36.

Moran, P., 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society (B)* 10: 243-251.

Tierney, L., 1990. *LISP-STAT: An Object Oriented Environment for Statistical Computing and Dynamic Graphics*. Wiley, Chichester.

Velleman, P.F., and Hoaglin, D., 1981. *Applications, Basics and Computing of Exploratory Data Analysis*. Duxbury Press, Boston.

9. Appendix A - Data Sets Used in the Study

In this study, two datasets are used. The first of these is derived from the 1991 UK census, at ward level for the Northern region of England, using variables as follows:

LLTI The percentage of persons in households in each ward where a member of the household has some limiting long-term illness. This is the response variable. Note that to control for different age profiles in areas, this is only computed for 45-65 year olds - an age category that is perhaps most at likely to suffer LLTI s a result of working in the extractive industries.

CROWDING This is the proportion of households in each census ward having an average of more than one person per room. This is an attempt to measure the level of cramped housing conditions in each ward.

DENSITY This is the housing density of each ward, measured in millions per square kilometer. This is intended to measure `Rurality' of areas. Note the differences between this and the previous variable - a remote village with poor housing conditions may well score low in this variable, but high in the previous.

UNEMP The proportion of male unemployment in an area. This is generally regarded as a measure of economic well-being for an area.

SC1 The proportion of heads of households whose jobs are classed in *social class I* in the Census. These are professional and managerial occupations. Whilst the previous variable measures general well-being, this measures affluence.

SP-FAM The proportion of single parent families in an area. This is an attempt to measure the *nature* of household composition in areas.

The second dataset is a synthesised, six-variable data set. The variables are named V1 to V6. Each data point lies on the surface of a six-dimensional hypersphere of radius one, with the exception of one outlier, which lies at the centroid of the hypersphere. This outlier is particularly `pathological' in that in any five-dimensional subset of the six variables, the value of this outlier is not particularly unusual. While it is uncertain how often this situation will happen with `real life' social science data, it does provide a yardstick for assessing each visualisation method in a worst case scenario.

The data may be generated as follows:

Note that a point on the circumference of a unit circle may be parametrised in terms of a single variable θ by the expression

$$(\sin(\theta),\cos(\theta))$$

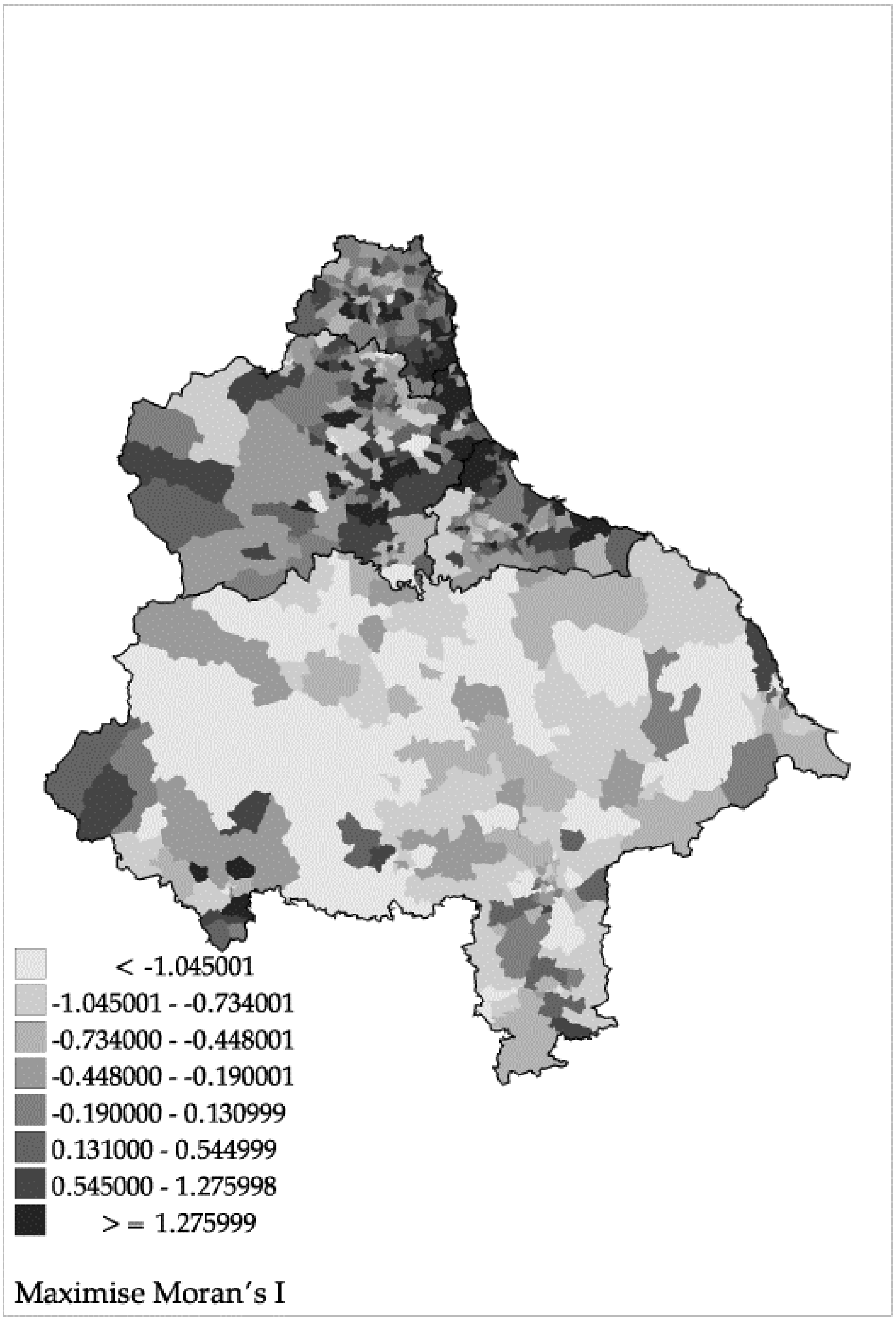
If theta is a uniform random variable, then random points on the circumference may be generated from this expression. Call this expression $C_2(\theta)$. Now let $C_n\ (\theta_{1\ \ldots}\ \theta_{n-1})$ be a point on the surface of an n-dimensional unit hypersphere. For example, if n=3, then $\mathrm{C}_3(\theta_1,\ \theta_2)$ is a point on the surface of a sphere. Recursively, we can parametrise C_{n+1} by

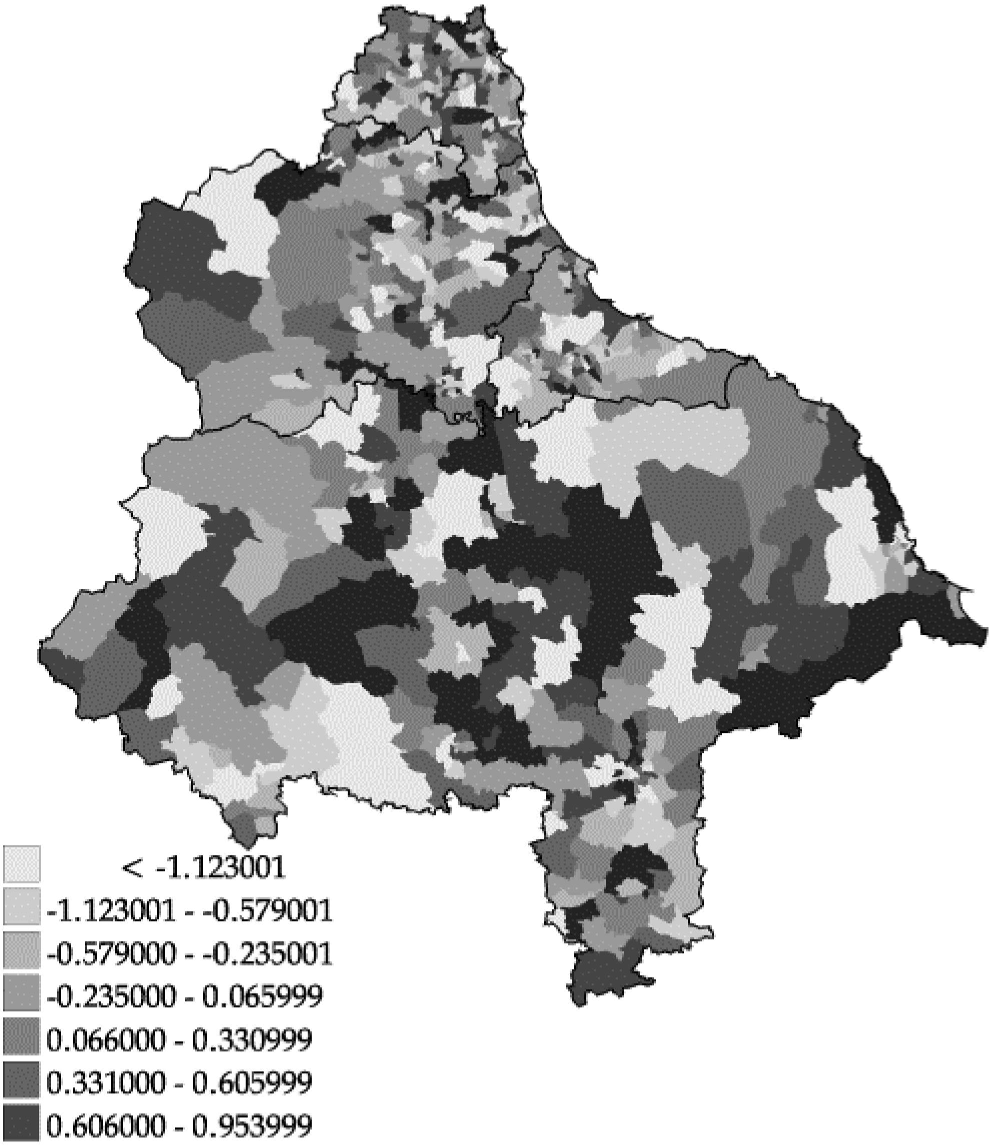
$$C_{n+1}(\theta_{1\ \ldots}\ \theta_n) = \sin(\theta_n)\ C_n(\theta_{1\ \ldots}\ \theta_{n-1})*\cos(\theta_n)$$

where * is a vector concatenation operator such that (x,y)*z = (x,y,z). One can check inductively that if the squared elements of C_n sum to one, then the squared elements of C_{n+1} also sum to one. Since this can be checked directly for C_2, it is true for all n > 2 also.

Thus, the surface on an n-dimensional sphere can be parametrised by a vector $(\theta_{1\ \ldots}\ \theta_{n-1})$. By generating uniform random numbers for the elements of this vector and applying this transform, it is possible to generate points on the surface of the hypersphere. The simulation is then finalised by adding the origin point as the outlier in the data set.

10. Appendix B





Minimise Moran's I