



**CENTRE FOR ADVANCED
SPATIAL ANALYSIS
Working Paper Series**



Paper 31

**Visual and
Interactive
Exploration of Point
Data**

Carolina Tobon



Centre for Advanced Spatial Analysis
University College London
1-19 Torrington Place
Gower Street
London WC1E 6BT

[t] +44 (0) 20 7679 1782
[f] +44 (0) 20 7813 2843
[e] casa@ucl.ac.uk
[w] www.casa.ucl.ac.uk

<http://www.casa.ucl.ac.uk/paper31.pdf>

Date: March 2001

ISSN: 1467-1298

© Copyright CASA, UCL.

Visual and Interactive Exploration of Point Data

Carolina Tobon
Department of Geography, University College London (UCL)
Centre for Advanced Spatial Analysis (CASA), UCL
Email: c.tobon@ucl.ac.uk
Last revised March 2001

ABSTRACT

Point data, such as Unit Postcodes (UPC), can provide very detailed information at fine scales of resolution. For instance, socio-economic attributes are commonly assigned to UPC. Hence, they can be represented as points and observable at the postcode level. Using UPC as a common field allows the concatenation of variables from disparate data sources that can potentially support sophisticated spatial analysis. However, visualising UPC in urban areas has at least three limitations. First, at small scales UPC occurrences can be very dense making their visualisation as points difficult. On the other hand, patterns in the associated attribute values are often hardly recognisable at large scales. Secondly, UPC can be used as a common field to allow the concatenation of highly multivariate data sets with an associated postcode. Finally, socio-economic variables assigned to UPC (such as the ones used here) can be non-Normal in their distributions as a result of a large presence of zero values and high variances which constrain their analysis using traditional statistics.

This paper discusses a Point Visualisation Tool (PVT), a proof-of-concept system developed to visually explore point data. Various well-known visualisation techniques were implemented to enable their interactive and dynamic interrogation. PVT provides multiple representations of point data to facilitate the understanding of the relations between attributes or variables as well as their spatial characteristics. Brushing between alternative views is used to link several representations of a single attribute, as well as to simultaneously explore more than one variable. PVT's functionality shows how the use of visual techniques embedded in an interactive environment enable the exploration of large amounts of multivariate point data.

KEY WORDS: Visualisation, Point Data, Exploratory Spatial Data Analysis, Interactive and Dynamic Display.

1 Introduction

Unit postcodes (UPC) can provide detailed information at a very fine level of resolution since a common postcode field can act as a link between data from disparate sources. Data concatenated in this manner can thus be represented as points. However, the analysis of UPC and the assigned data is not straightforward in urban areas for various reasons. First, UPC occurrences in urban areas are very frequent which frustrates the visual inspection of individual points at small scales. On the other hand, patterns in the associated attribute values can are not easily distinguished from point data depicted at larger scales. Secondly, the large presence of zero values and high variances of the socio-economic variables used here make their distribution non-Normal posing restrictions to their analysis via traditional statistical methods.

Postcodes can be highly multivariate when associated to several attribute values. A usual approach to handle and explore multivariate point data sets in a Geographic Information System (GIS) is to transform them, for instance, into density surfaces, contour lines or DEMs. GIS environments traditionally treat visual display as an output or endpoint of an analysis process and therefore do not facilitate the use of graphics for the visual exploration of data. Hence, the aforementioned transformations allow both a form of visualisation of attributes and the possibility to perform operations on them to explore their relations. However, it is argued here that exploring data at the point level of resolution—before making transformations that make assumptions or generalisations about their characteristics—can provide detailed knowledge to inform further processes of data manipulation. Hence, transformations into other representations and the choice of operations to perform on the original data sets would be based on a better knowledge of the data.

The use of visualisation techniques can provide an effective way to learn about the data and facilitate the formulation of hypotheses about their relations when embedded in interactive and dynamic software environments. PVT aims to provide such functionality for spatial point data using an interface that responds to users' actions and their interrogation of data on a screen. PVT's purpose was also to extend existing knowledge of how the point representation of geographical data and their analysis via visualisation techniques can be used to expand their understanding. PVT can therefore be considered as a proof-of-concept or prototype system that demonstrates the type of functionality that software aiming to investigate spatial point data by means of visualisation techniques should provide.

The paper is organised as follows. Section 2 introduces a short review of the literature, which addresses the issue of how the visual interaction with data on screen can aid users to gain information from it. Also, a sample of available software and tools that have implemented visualisation techniques for both point and areal data are reviewed. Section 3 discusses the particular statistical characteristics of the data sets used here and some limitations of using traditional statistical analysis. Section 4 discusses PVT's functionality and how exploring point data using the techniques described in Section 2 was found to be both appropriate and illuminating to understand the characteristics of the data and the relations between different sets of attribute variables. Section 5 offers some conclusions and directions for future research.

2 Visualisation

Visualisation has primarily been “used as an informal way to understanding” (Unwin et al., 1994) until recently. It was not considered as an acceptable method of scientific practice when compared to mathematical or statistical analysis. However, the “explosion of observational and model-based data and the development of computer visualisation tools” (MacEachren and Monmonier, 1992) have made visualisation an effective way of analysing vast amounts of information. In this context, visualisation can be defined as the use of graphic methods to explore data in transient and innovative ways with which to obtain information that may otherwise remain hidden.

The capacity of graphics to reveal information and patterns from data sets using Exploratory Data Analysis (EDA) has been acknowledged at least since the second half of the 1970s by authors such as Tukey. However, not until 1987 was there a formal initiative, known as Visualisation in Scientific Computing (ViSC) (McCormick et al., 1987), to investigate the use of visualisation for data analysis. The initiative addressed the need of “a whole variety of broader fields in the natural and environmental sciences and statistics” (Dykes, 1999) for new methods of analysing complex and often multivariate data and the provision of more flexible environments that allowed data exploration. It was also responding to the need for processing the large amounts of data being produced by computer models and automatic observational instruments.

Visualisation since then has been a crucial point in the research agenda of various disciplines. For instance, from 1995 to 1999, the International Cartographic Association (ICA) charged the Commission on Visualization to address problems “associated with extending cartographic methods into an increasingly dynamic technological environment” (MacEachren and Kraak, 2000a). Hence, the concern was using novel technology to graphically represent and interact with geographical data to develop ideas, as well as reveal aspects that could “form the basis for formal hypothesis” (Fisher et al., 1993). The Commission was also interested in identifying points of complementary research between disciplines interested on the visual exploration of data such as computer graphics, information visualisation, or exploratory data analysis (MacEachren and Kraak, 2000a).

Advances in multimedia and computing technology increasingly enabled ViSC to be used as a tool kit to discover information from data. Software and hardware developments have made computer processing faster and better graphics can now be displayed at greater speed using high-fidelity screens. These facts combined with the availability of ‘ware-houses’ or ‘fire-hoses’ of data due to

improvements in data capture equipment, have made the use of graphics a practical way to explore, assimilate and process large amounts of information.¹ New scripting and prototyping tools also now allow developers to deliver products that are interactive and flexible for users to explore data (Dykes, 1999). ViSC has therefore been “recognised as a method, and product, that integrated the power of digital computers and human vision and directed the result towards facilitating insight across the sciences” (McCormick et al., 1987).

It has been argued that visualisation of geographical data offers nothing new since it has long been a main issue in the discipline of cartography. However, cartography has traditionally been concerned with producing maps that communicate a particular message using some predefined design. Its main focus has been the production of static maps printed on paper and devised for communicating specific messages to average users.² Visualisation on the other hand, is aimed at “[helping] individuals (or small groups of individuals) think spatially” (MacEachren, 1994b) by providing interactive and dynamic displays and multiple representations of the data to allow its exploration. Therefore, promoting visual thinking and decision making as well as a renewed map use are some of ViSC concerns when providing “maps that present answers [as opposed] to maps that foster a search for questions” (MacEachren and Monmonier, 1992).

Various software packages have been developed for the visual and interactive exploration of data on screen. In order to be effective, such systems should be designed having the users and the tasks they will perform in mind. Hence, interface design is a key issue in creating environments suitable for visualisation as it influences the way users interact with a system and understand it. Since 1999, a second ICA Commission on Visualization and Virtual Environments has tackled some of these concerns. Hence, its research agenda on ‘geovisualisation’ addressed four themes: “representation of geospatial information, integration of visual with computational methods of knowledge construction, interface design for geovisualization environments, and cognitive/usability aspects of geovisualization” (MacEachren and Kraak, 2000a).

Software for data exploration that provided guidelines for the development of PVT are discussed below. Various well known visualisation techniques such as ‘brushing’ and ‘dynamic linking’ were

¹ Two examples are referenced here for illustration. First, the use of graphics to interpret the results of simulations and non-linear models (see Unwin et al. (1994)). Second, the use of visualisation to inform NASA how well an orbiting satellite can measure properties of particles in the earth’s atmosphere (Kahn et al., 1998).

² This however has started to change due to the influence of ViSC and EDA. This is seen for instance in a recent special issue of the journal of Computers and Geosciences dedicated to “Exploratory Cartographic Visualisation”.

implemented in PVT. However, they were adapted for the visual inspection of point data which poses a different set to lattice data for which most of the systems discussed next were designed.

2.1 Visualisation Software: Examples

Flexible environments that enable users to interrogate data have been implemented to some extent as software or tools for visualising data. Their common characteristics are allowing high levels of real-time interaction with the data, as well as innovative representations that foster insight about the information depicted on screen. The nine systems described in Table 1 are not necessarily a comprehensive review of available tools for Exploratory Spatial Data Analysis (ESDA). However they illuminated the process of defining the conceptual design for PVT, which is described in a later section, and point towards useful and effective characteristics that a system of this nature should provide.

The first four systems were particularly relevant in the development of PVT as they were created as software packages for ESDA. They aim at fostering ideas and discovery by providing dynamic displays that enable different combinations and representations of the data. They all implement well known visualisation techniques such as brushing and dynamic linking but they do so in creative ways to address their target users and the tasks they are likely to perform with a defined type of data. The conceptual design of Spotfire.net seems to be based on its users' tasks as well as their previous knowledge and experience as it easily and promptly communicates to the user what the system can do through its interface. The learning curve of the software package is arguably the steepest from all the systems reviewed. This is not surprising, as Spotfire.net is the product of six years of research within the Human Computer Interaction department of the University of Maryland. CDV on the other hand was a breaking ground cartographic visualisation software when it was released. Although other systems had implemented tools to investigate data on-screen (such as the Density Dial, RADVIZ, or MANET reviewed below), CDV provided a complete environment to visually explore geographical areal data with a degree of interactivity that was not previously available. Descartes has taken some of these ideas further to provide functionality suited for the type of data used.

All of the systems provided guidelines of useful tools for a visualisation software. However, with the exception of Spotfire.net which is not intended to handle geographical data, all the systems are concerned with areal data. PVT's main purpose was to apply these know visualisation techniques to spatial point data. Section 4 shows results in this direction. But before describing PVT and its

possible applications, a brief discussion about the limitations of using traditional statistics for the analysis of geographical data is included below. The next section describes the special characteristics of spatial data, which justify the research and use of innovative ways of exploring it.

Software	Description
Spotfire.net	This web-enabled software is aimed at visually exploring and analysing point data that is non-spatial. It is a flexible tool in terms of how different views can be modified by the user. It provides an easy to use environment for the exploration of data and for the rapid identification of trends, anomalies, outliers and patterns. One of its salient features is interactively filtering (excluding or including) ranges of data while still showing the user the whole range of the information. This form of interaction with the data is provided in 2D and 3D scatter plots, histograms, bar charts, pie charts, and parallel coordinate plots. Spotfire.net wide range of users (pharmaceuticals, biotechnology, manufacturing, living sciences and semiconductors) shows the ample number of fields that benefit from exploring and analysing point data to acquire information that can support further research or decision making. <i>Developed by C. Ahlberg at Chalmers University and B. Shneiderman at the University of Maryland. A downloadable demo is available from http://www.spotfire.com/.</i>
CDV (Cartographic Data Visualiser)	CDV implements several re-expression and brushing techniques that are very effective for exploring relations among area-based data sets. It provides alternative views (dotplots, boxplots, circle maps, cartograms, scatterplots and parallel coordinate plots) of one, two or three data sets simultaneously. Transient symbolism, multiple views, variation of colour schemes, symbol scaling, and various forms of brushing to link the views are some of the salient features provided by means of an interface that is interactive and tailored to a specific set of users. <i>Developed by J. Dykes as part of his doctoral thesis (University of Leicester) on cartographic visualisation and exploratory spatial analysis. Available from http://www.geog.le.ac.uk/jad7/.</i>
Descartes	Descartes' main concern is providing traditional cartographic presentation methods, which are interactive and dynamic for the exploration of areal data. It uses multiple representations of the information, each being best suited for certain types of analysis, to explore the data. Dynamic choropleth maps, pie charts, bar diagrams, dotplots, boxplots, and scatterplots are available to the user. Two interesting features can be mentioned. First, the use of sliders to represent the range of a numeric variable displayed on the map. They can be moved to interactively change maps' classifications and colour coding. Secondly, the use of a colour matrix created from dividing the range of two variables into intervals. " $M \times N$ classes, [are defined] where M and N are the numbers of intervals for the first and for the second variable, respectively" (Adrienko and Adrienko, 1999). A colour scale is then selected for each variable. This generates a colour matrix on top of which a scatterplot is laid. Colours in the matrix thus create a visual classification of the data by value ranges. <i>Developed at the German National Research Centre for Information Technology (GMD) by G. Adrienko and N. Adrienko. Functionality is described in Adrienko and Adrienko (1997, 1999) and a web supplement of the 1999 article with demos in http://allanon.gmd.de/and/java/iris.</i>
GeoVista Studio 1.0	This system is a Java-based experimental environment, designed to allow the integration of modules for the analysis of geographic data. It draws from advances in software engineering, geocomputation, and visualisation to provide a programming environment that permits data visualisation. It is said to combine visual approaches as interactive parallel coordinate plots and 3D rendering with a spreadsheet and statistics package to enhance the analysis. The parallel coordinate plot is of particular interest since it allows a high degree of interaction by the user with a large set of variables (For a demo see http://www.geovista.psu.edu/products/demos/edsall/Tclets072799/peptcl.html). The variable allocated to each axis can be interactively chosen as well as the criteria to classify data into ranges. Different ways to colour the lines in the plot are provided to easily distinguish data ranges. <i>Developed at the GeoVista Centre, Penn State University (http://www.geovista.psu.edu/products/studio/index.htm).</i>
Density Dial	Its most salient feature is providing users with a tool to interactively vary the statistical value at which intervals between two classes occur (Ferreira and Wiggins, 1990).
RADVIZ	"Maps a set of m -dimensional points onto two dimensional space" (Fotheringham et al., 2000). Points are arranged to be equally spaced around a circumference or on the vertices of a regular polygon. Any point inside the polygon is a projection or representation in 2D space of a point in m -D space. This method is particularly useful for the identification of outliers, extreme values and correlation among variables. <i>Commercially available from http://www.anvilinformatics.com/tools-radviz.html.</i>
MANET	Designed for area-based spatial data to implement interactive graphics tools for data sets with missing values. Interesting features include allowing users to experiment with the sensitivity of a histogram to the width of its bins (Unwin et al., 1996). By dragging their edges, the data are reclassified into new bins and the user can see the effect on the resulting distribution. <i>Developed by Unwin, A and Hofmann, H. Available from http://www1.math.uni-augsburg.de/Manet/.</i>
SpaceStat	Developed for the analysis of spatial lattice data through the use of spatial statistics and spatial econometrics. Non-linear optimisation routines are implemented in the system to obtain maximum likelihood estimates for spatial regression models, facilitates testing for the presence of spatial autocorrelation and estimates models that incorporate this form of dependence. Results can be visualised in ArcView GIS from ESRI. <i>Developed by Anselin, L. Commercially available from http://www.spacestat.com/.</i>
PINstudent (Postcoded Information Navigation)	Designed to explain the geography of the UK postcode system by exploring the locational, numerical and textual aspects of postcodes and their role in geographical analysis. Descriptive statistics, distances and densities of postcodes are some of its main features. <i>Developed by J. Shepherd, J. Raper and D. Rhind, in collaboration with D. Ming (Birkbeck College). Available from http://cb180.geog.bb.ac.uk/pinweb/.</i>

Table 1 Example Software

3 The Data

Socio-economic data was provided through CASA (Centre for Advanced Spatial Analysis, University College London) to explore using PVT. The main sources of information were two government agencies: the Valuation Office Agency (VOA) and the Office of National Statistics (ONS).³ The data sets are highly confidential which is one of the reasons why they are aggregated to the UPC level and only a portion of the data (that for the London boroughs of Wandsworth and Fulham) is used here to demonstrate PVT's functionality. Three data sets were thus used: Retail Floorspace, Retail Employment and Entertainment Employment (or RF2, R2, and E2, respectively). Each data set contains a postcode, a corresponding UK grid coordinate, and a value for the variable at each UPC.

“It is generally helpful to *look* at the data set before any models are fitted or hypothesis formally tested” (Fotheringham et al., 2000) as it gives indications of possible traits or relations between the data sets. Hence, the purpose of this section is to provide an initial analysis to show the particular characteristics of these data when investigated geographically. Descriptive statistics, G-functions and Principal Components Analysis (PCA) were all used for this purpose.

3.1 Descriptive Statistics

The socio-economic data used for this study have a particular set of characteristics: they are very dense, highly multivariate, and are non-Normally distributed with high variances and a predominance of zero values. Evidence of these characteristics was obtained through the analysis of the series descriptive statistics⁴ are shown in Table 2 for the three data sets.

All the series are skewed to the left with both the median and the mode lying to the left of the mean. This is clearly seen in Figure 1 - Figure 3 which show the histogram and cumulative density functions of the three series. Skewness is always positive however due to absence of negative values. The high positive kurtosis values indicate very peaked distributions. Therefore it was not surprising that the distributions were non-Normal as showed by small probability value of the Jarque-Bera statistic.

³ The VOA collects floorspace information on individual properties and the ONS collects employment and turnover data on individual business.

⁴ They were derived using E-Views Version 3.1.

Apart from a characterisation of the data sets, the relations between the variables were of particular interest. Since the data were aggregated to the UPC level, clustering between the attribute values was anticipated and formally tested using G-functions. Further characterisation of the clusters was explored through PCA.

	RF2	R2	E2
Mean	622.84	13.21	12.38
Median	292.01	4	6
Mode	29.6	1	3
Minimum	2.4	0	0
Maximum	19,534.63	592	295
Count	652	964	587
Sum	406,090.2	12,735	7,266
Std Deviation	1,293.0	38.93	23.86
Sample	1'671,849.0	1,515.55	569.30
Kurtosis	84.22	100.88	59.26
Skewness	7.60	8.99	6.62
Jarque-Bera	195,942.9	417,488.60	88,670.88
Probability	0.000	0.000	0.000

Table 2 Descriptive Statistics

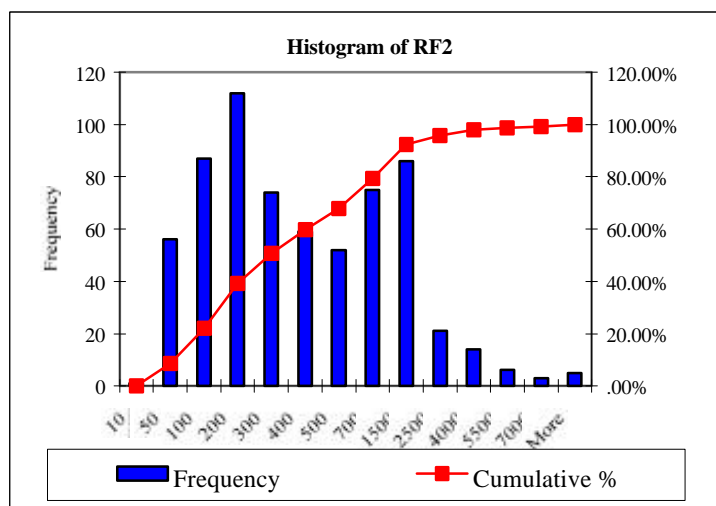


Figure 1 Histogram and Frequency distribution of RF2

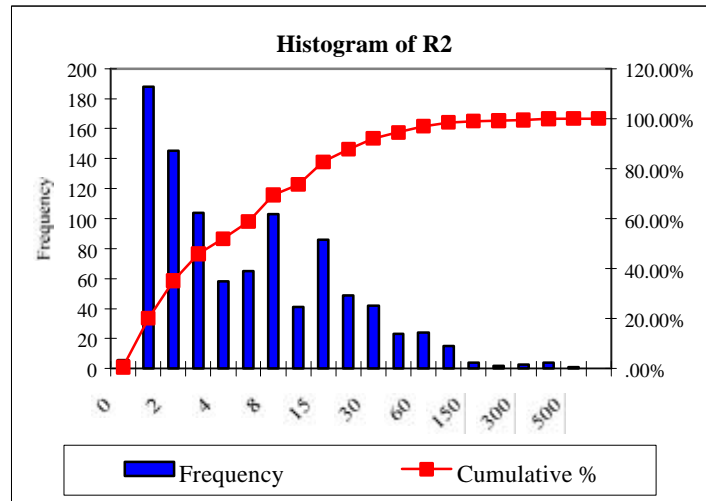


Figure 2 Histogram and Frequency Distribution of R2

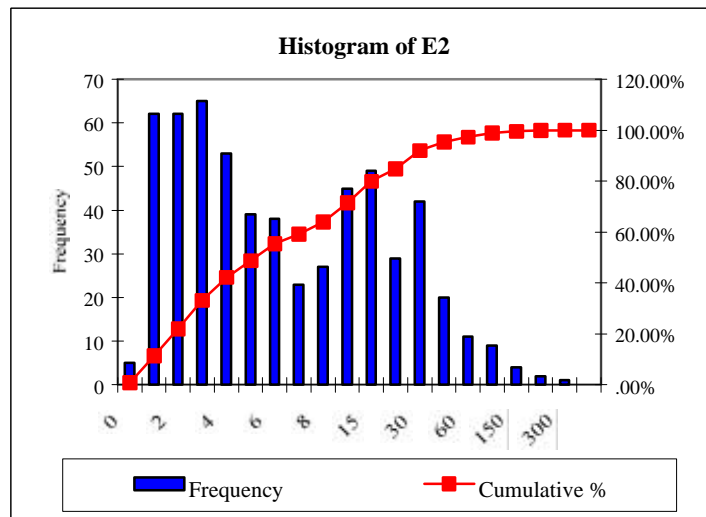


Figure 3 Histogram and Frequency Distribution of E2

3.2 The G-Functions

G-functions⁵ (Figure 4 - Figure 6) characterise the distribution of a point pattern within a study area since they “result from the spatial correlation structure, or the spatial dependence in the process” (Bailey and Gatrell, 1995) using nearest neighbour distances between events in the study area⁶. In all cases, the cumulative probability distribution function climbs steeply in its early range and then flattens out which indicates an observed high probability of clustering. This can be inferred from

⁵ Defined as $G(w) = \#(w_i \leq w) / n$ where w_i is the minimum average Nearest Neighbour inter-event distance, w is some chosen nearest neighbour distance, and $\#$ is the number of counts obeying the condition (Bailey and Gatrell, 1995). The G functions are plotted against values of w for the three series.

the concave shape of the function resulting from a high frequency of short nearest neighbour distances.

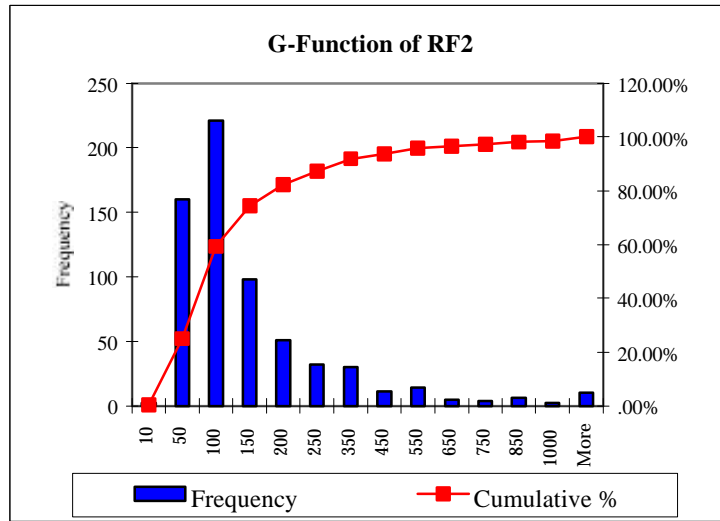


Figure 4 G-Function of RF2

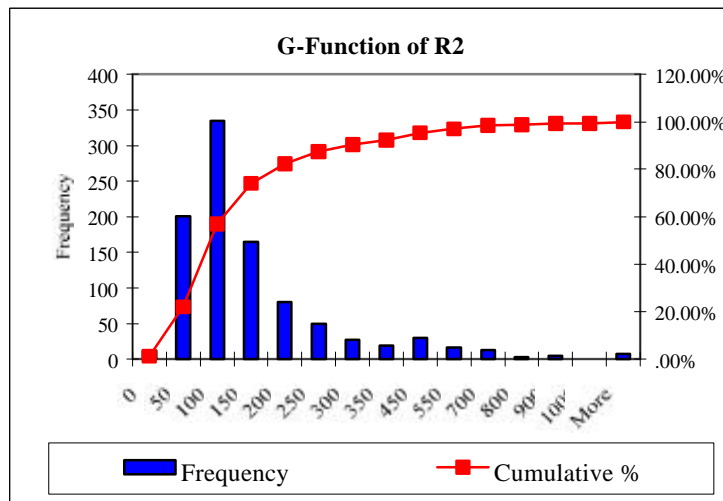


Figure 5 G-Function of R2

⁶ G-functions for the 3 series were calculated in GAMS version 225 due to the size of the series.

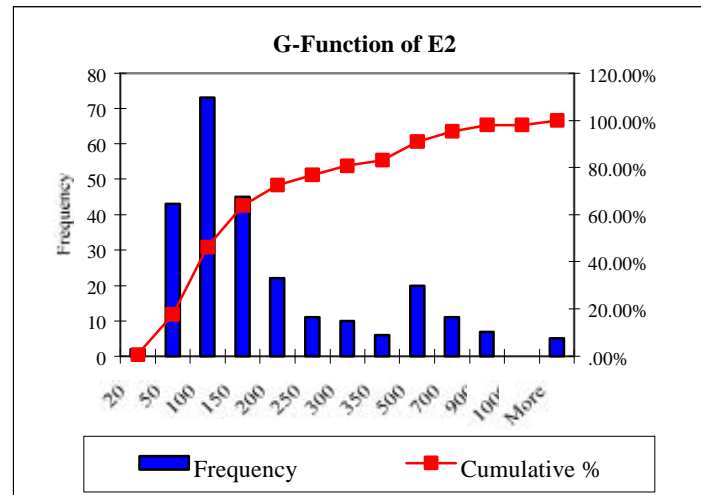


Figure 6 G-Function E2

3.3 Principal Component Analysis

Once spatial clustering was formally identified, relations between the series were investigated through Principal Component Analysis (PCA). This is a standard technique for extracting the main relations in data of high dimensionality. It is suitable for finding optimal linear combinations of the variables and identifying underlying or latent ones. It can be defined as “a method to combine variables into a smaller number of variables (the principal components) that are statistically independent of one another and ‘explain’ as much of the co-variation between elements as possible” (Bonham-Carter, 1994).

Componen	Eigen	Difference	Proportion	Cumulati
1	1.4391	0.5093	0.4797	0.4797
2	0.9299	0.2988	0.3099	0.7897
3	0.6311	-	0.2103	1

Table 3 PCA for 3 Series

Table 3 shows the result of the PCA for RF2, R2, and E2⁷ where all components contribute considerably to the variation.⁸ In a similar test with socio-economic data for Andover, Thurstain-

⁷ The Principal Components were estimated using Intercooled STATA version 6.0. It is important to note that only postcodes containing these three attribute values were considered in the analysis. Distributions of the sub-samples were still skewed and non-Normal and highly variable.

⁸ In Table 3, take the first component’s Eigen Value (1.4391) and divide it by the number of components (3). This will give the value reported in the column named Proportion: $1.4391/3 = 0.4797$ or 47.97% which is that component’s contribution to the variation.

Goodwin and Unwin (2000) found that although two of their variables “[correlated] quite well, within either set the correlations [were] weak to non-existent”. This was also the case in this exercise as RF2 and R2 are correlated (Table 4) but no evidence could be found of correlation within the series.

	R2	RF2	E2
R2	1		
RF2	0.3657	1	
E2	0.1499	0.1037	1

Table 4 Correlation Matrix

To corroborate the results of this first PCA, a second one was performed on the complete range of the three series (RF2, R2, and E2) plus six more available from the VOA and ONS.⁹ Results for this second exercise are reported in Table 5. In this case, the first component accounts for 94.64% of the variance. Although “the point [of PCA] is to use a small number of principal components so that the sum of their contribution is satisfactorily close to one” (Greene, 1994), the first component accounts for almost all of the variation. Therefore, the extraction of principal components in the two exercises just described did not provide information with which to further characterise the relations between the clusters in the data sets.

Componen	Eigen	Difference	Proportion	Cumulati
1	8.5172	8.2039	0.9464	0.9464
2	0.3133	0.2273	0.0348	0.9812
3	0.0860	0.0367	0.0096	0.9907
4	0.0492	0.0304	0.0055	0.9962
5	0.0189	0.0092	0.0021	0.9983
6	0.0097	0.0068	0.0011	0.9994
7	0.0029	0.0012	0.0003	0.9997
8	0.0017	0.0007	0.0002	0.9999
9	0.0010	-	0.0001	1

Table 5 PCA for 9 Series

The large variances and non-Normal distributions of the data sets pose limitations to performing traditional statistical analysis such as PCA. The question remains if it is possible to learn from the

⁹ Due to strict confidentiality reasons, no further description of the data sets can be provided.

visual exploration of point data which are heavily clustered, as in the case of UPC in urban areas. The sections to follow suggest how the visualisation techniques described previously can be efficient and effective in deriving information from such data sets and in discerning their relations.

4 Dynamic Display in PVT

Dynamic display is used in PVT to facilitate the visual exploration of point data. The term is used here in the same sense as in Adrienko and Adrienko (1999) to denote “display changes in real time in response to user’s actions rather than to denote animated presentations of time-series data”. It also refers to the depiction of movement and change, flexible user interaction, multiple data representations and their linking, to allow the simultaneous exploration of one or more attribute variables at the same time. When embedded in an environment that allows manipulation by the viewer, dynamic display allows users to experiment with “multiple perspectives of the same data [...] until patterns emerge that are particularly appropriate” (Dykes, 1999). It can therefore provide an environment suitable for visual thinking which is the main purpose of PVT.

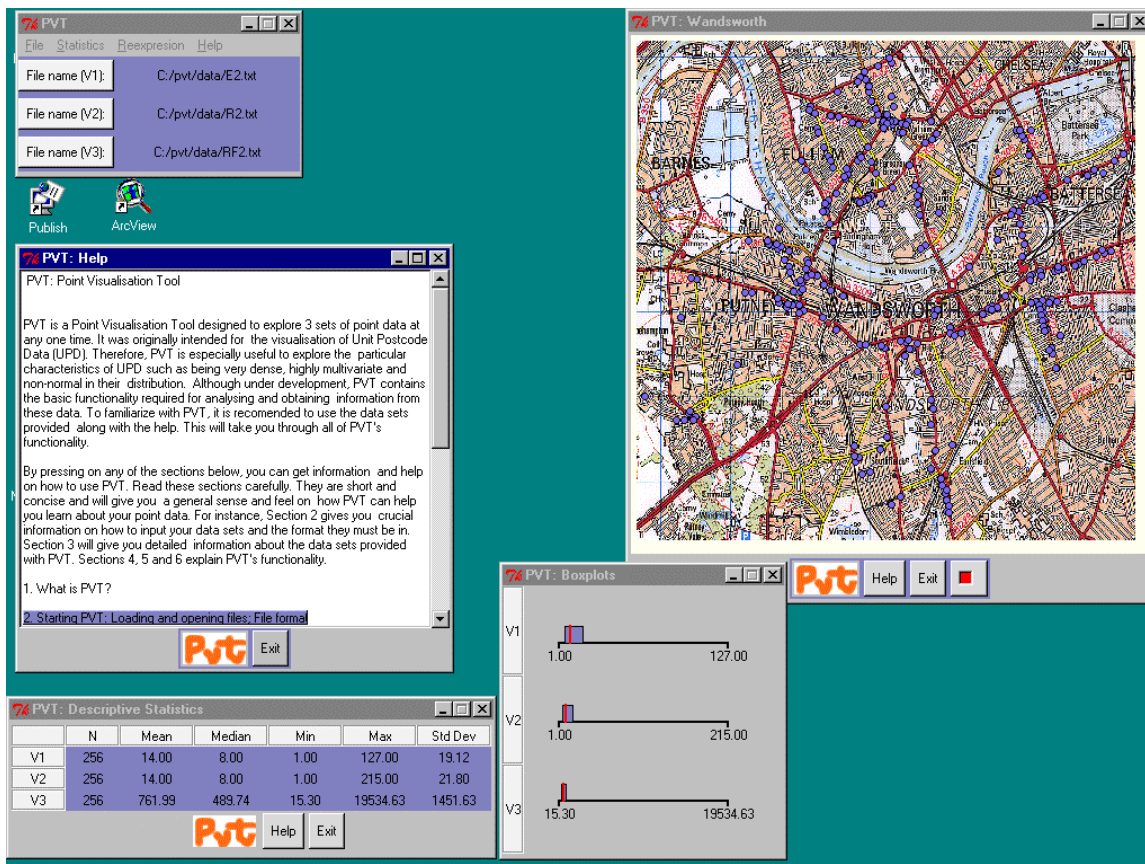


Figure 7 PVT

PVT (Figure 7) was created using the Tcl/Tk toolkit (Ousterhout, 1994) for the purpose of exploring clustered and highly dense spatial data at the point level of resolution. The dynamic display used in PVT implies some form of temporal variation of the data represented on screen as suggested by Shepherd (1995). Different forms of variation provide particular services which aid the user in investigating the data. Emphasis is given to variation prompted by the user's behaviour or interaction with the data, and by changes in the representation of the displayed symbolism.

PVT provides multiple representations of the data as requested by the user such as scatterplots, dot plots, boxplots, parallel coordinate plots, map and descriptive statistics, each focusing on certain information about it. They enhance the observer's perception of pattern by providing different graphical versions of the same data. The fact that each view will convey only a subset of characteristics of the data can be compensated by linking the views "so that the information contained in individual views can be integrated into a coherent image of the data as a whole" (Buja et al., 1991). Brushing (Monmonier, 1989) was implemented in PVT for this purpose as a method to enable the linking of more than one representation of the data set on screen. Its most common manifestations are blinking, flashing or highlighting all representations of the data when an item is selected in any of the views. Therefore, each point can behave "in relation to actions taken by the observer and hence serve as an interpretative aid" (Shepherd, 1995). This functionality is explored next in greater detail using the data described in section 3.

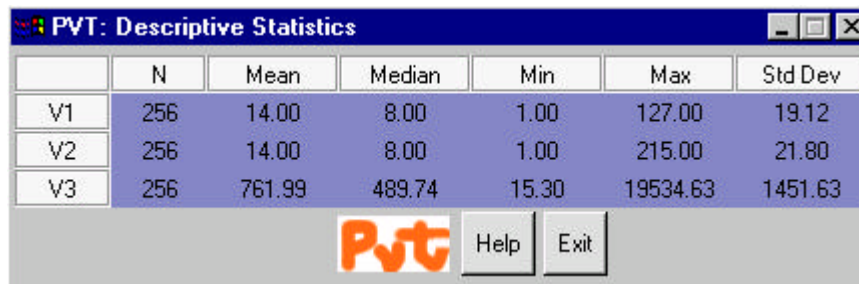
4.1 Alternative views of point data in PVT

Section 3 showed the importance and effectiveness of descriptive statistics for initially characterising data. PVT provides two distinctively different forms of presenting statistics: a Descriptive Statistics Table¹⁰ and Boxplots¹¹ both of which are useful to learn about the distribution and spread of the data. Minimum and maximum values, when compared to the mean, can give an indication of outliers or extreme values. For instance, the median of variable "V3"¹² lies to the left of the mean making the distribution of that attribute skewed towards the smaller values of the data set (Figure 8). The maximum value however appears to be very large when compared to these two statistics. This gave an indication of an unusual realisation of the variable. When the data was explored, a value of 19534.63 ft² was not possible for the retail establishment associated to that UPC and was designated as an error in the data. Excluding such values from the mean would make

¹⁰ Which includes the number of observations, minimum, maximum, mean, median and standard deviation.

¹¹ Which contain the so-called five number summary: minimum, maximum, median, lower and upper quartiles.

it drop to 688.38 ft². Therefore, a simple comparison between statistics can give an indication of outliers, or unusually high/low values that affect the distribution of the data.



	N	Mean	Median	Min	Max	Std Dev
V1	256	14.00	8.00	1.00	127.00	19.12
V2	256	14.00	8.00	1.00	215.00	21.80
V3	256	761.99	489.74	15.30	19534.63	1451.63

Figure 8 PVT: Descriptive Statistics Table

Boxplots (Figure 9) are also fast and effective for learning about the spread of a series. The black horizontal line delimits the range of the data and the whiskers at either end show the minimum and maximum values. The edges of the box are given by the lower and upper quartiles of the series.¹³ The vertical line inside the box is given by the median of the series. This five number summary provides a general description of the data distribution by providing the user with a visual representation of the information which is most effective to determine skeweness.

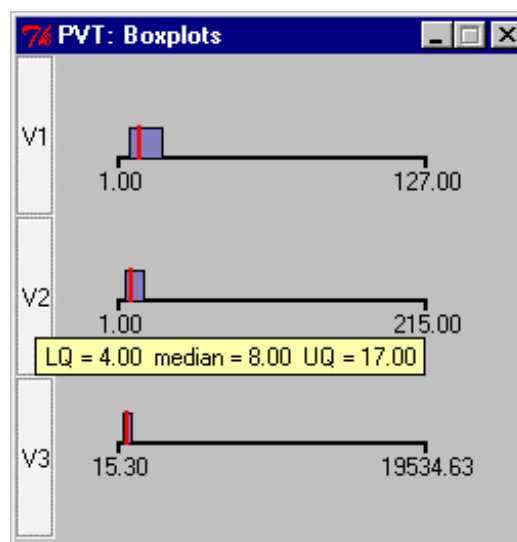


Figure 9 PVT: Boxplot

¹² “V1” corresponds to Entertainment Employment, “V2” to Retail Employment, and “V3” to Retail Floorspace.

¹³ The lower quartile indicates the value below which the smallest 25% of the data can be found. The upper quartile marks the value after which the largest 75% of the data can be found.

The actual numerical values of these statistics are shown when the user leaves the cursor on top of any of the labels (marked “V1”, “V2”, and “V3”) to the left of the Boxplot view. This action prompts a balloon with the lower and upper quartiles (LQ and UQ, respectively) and the median as shown in Figure 9. The balloon for variable “V2” is showing that 75% of the data values are higher than 17 but still the median has a smaller value of 8. This is indicating both skewness and possibly the presence of an unusually which value (215).¹⁴ The boxplot of “V3” is also giving visual indication of skewness possibly caused by the high value identified as a typing error (19534.63 ft²). Both the Descriptive Statistics Table and the Boxplot views give information on the three series simultaneously. This allows the user to inspect each variable’s characteristics as well as compare them relative to other variables.

As it was mentioned previously, each postcode can be associated with multiple attribute values. Therefore, a main concern in the development of PVT was to enable the investigation of variables associated to each point or UPC both independently and in sets that would allow their comparison. This was achieved through the use of dotplots, scatterplots, and parallel coordinate plots that allow the exploration of data in one, two or multiple dimensions, respectively. Views in PVT are intended to be transient to assist the user in highlighting specific information and exploring alternative relations among the attribute data. An example follows of how to use them to detect clusters and outliers by displaying combinations of these views or even by displaying them independently. Identifying clusters is of interest as they are groupings in the data points commonly associated with multimodality in the underlying distribution. Outliers, on the other hand, can point to rare univariate or multivariate cases of unusual data combinations when compared to the rest of the series or, in some cases, errors in the data.

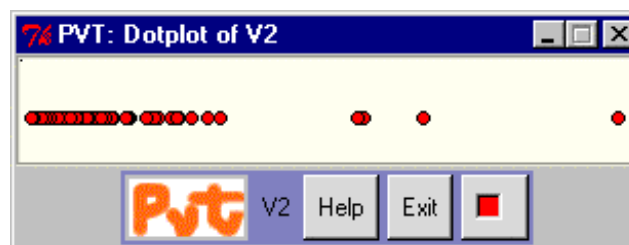


Figure 10 PVT: Dotplot

A dotplot is a way to represent the statistical distribution of a single variable using symbols located along an axis. The Dotplot view (Figure 10) shows clustering towards small positive values

¹⁴ These 2 facts were corroborated as can be seen from previous discussions.

for “V2”.¹⁵ This clustering is in part due to the presence of outliers or unusual realisations of the series visible to the right of the view. Scatterplots on the other hand, display each observation on a graph where the axes are two of the variables associated to the UPC. For this reason, the scatterplots can show outliers and clusters in two dimensions. Figure 11 shows the observations previously identified as outliers in “V3” (see the point to the upper left corner) and “V2” (see the point to the far right). They are in part responsible for the clustering of the data in this view. When removed, the positive correlation between the two series is more easily identifiable. It is also showing at least three more points that appear to be two-dimensional outliers that have not been identified in previous views.

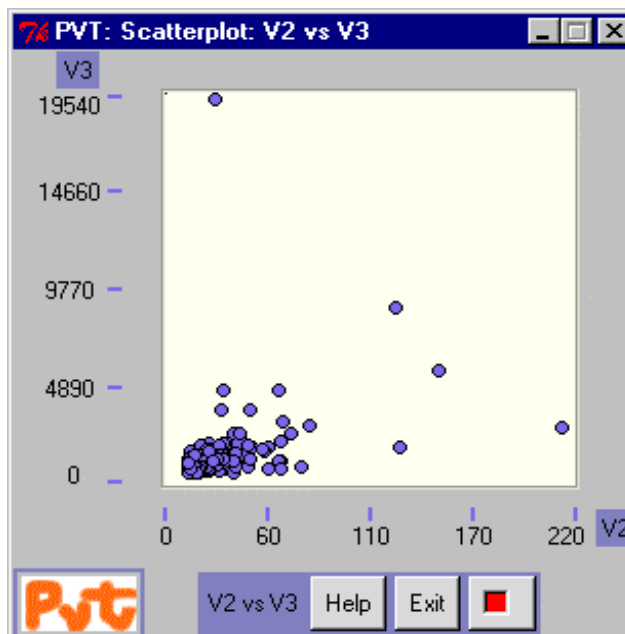


Figure 11 PVT: Scatterplot

A combination of scatterplots, dotplots and boxplots is provided in the Scatterplot Matrix (Figure 12) to aid in the analysis of the data in more than one or two dimensions. In such a matrix, all of the possible scatterplots and dotplots are arranged into a table. In each row, the y variable is common to all the plots and in every column the x variable is common to all the scatterplots. This allows all two-way patterns in a data set to be viewed simultaneously. Plotting multiple dotplots is useful for identifying multi-modal distributions, clusters and outliers as explained before. Note that the corresponding boxplots are also included with the dotplots to provide additional information on the distribution of each data set. When the three scatterplots are put together the strongest positive

¹⁵ Low values of the variable are displayed to the left of the dotplot axis and high ones to the right.

correlation between “V2” and “V3” is made more obvious, even when the aforementioned outliers have not been removed. The other two scatterplots do not suggest as strong relations between the variables.¹⁶

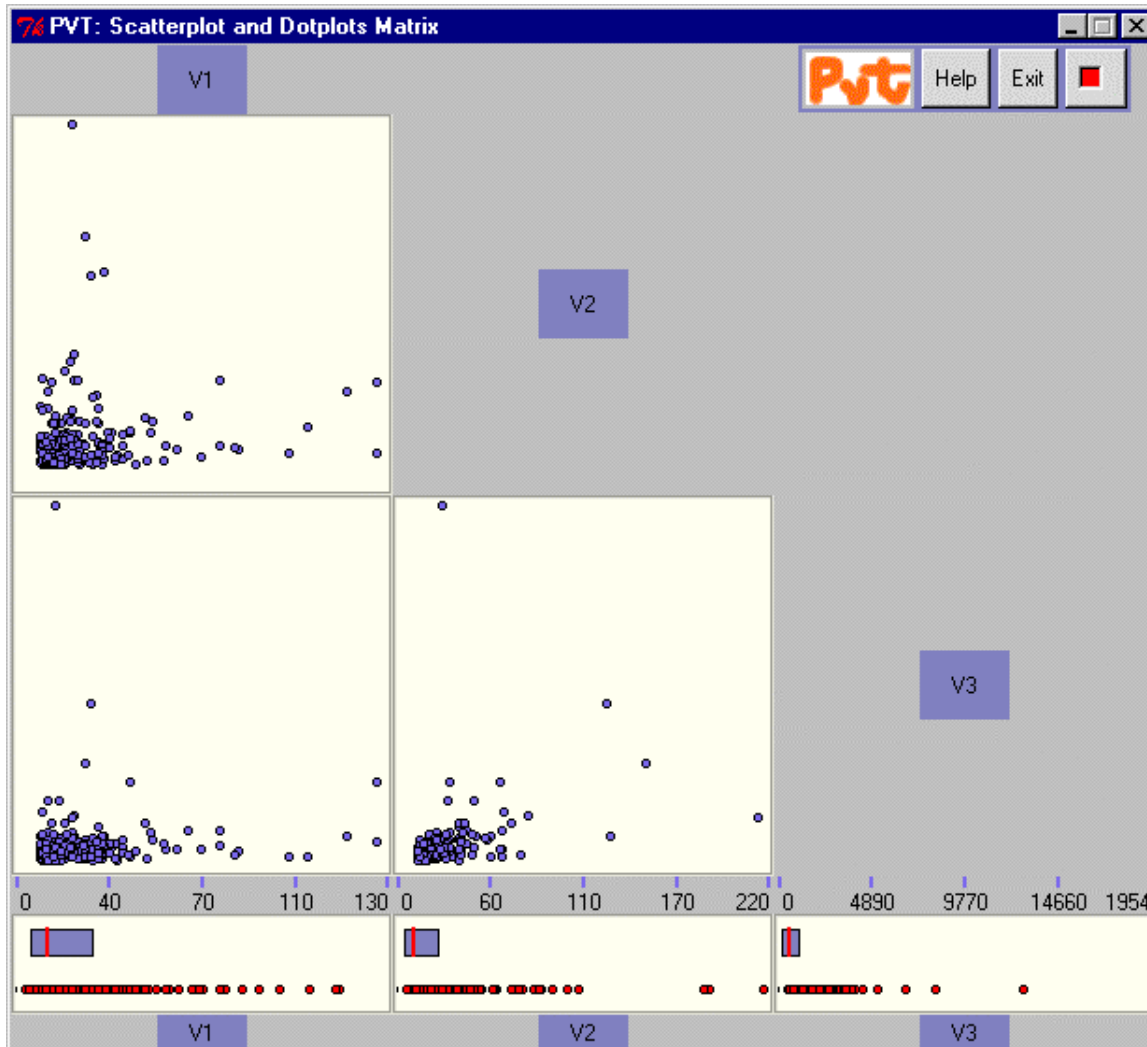


Figure 12 PVT: Scatterplot Matrix

Although a scatterplot matrix is one approach to exploring data in more than one or two dimensions simultaneously, scatterplots still only show interactions between pairs of variables and dotplots describe the distribution of a single data set. In a parallel coordinate plot, a point can be explored in many more dimensions. Such representation of the data is in effect a set of parallel dotplots, where lines link the three data values associated to each UPC. When one of these lines is touched, it changes colour and the region from which these data were extracted is highlighted in the map. In other words, each UPC used in this exercise contained three attribute values (E2, R2 and RF2), each

¹⁶ Alternatively, scatterplots and dotplots of all the variables can be displayed simultaneously at any time.

of which is plotted along one of the three lines in the parallel coordinate plot in Figure 13. The highlighted line shows the three attribute values populating one UPC and it can be prompted when a user touches any of the points. Each one of these lines can be understood as a profile of UPC where the shape of the line gives information on the levels of each variable. Figure 13 shows that for a large data set, lines overlap and many are not clearly distinguishable. Lines with similar patterns tend to show clusters in the data. However, it is those lines which do not have a common pattern the most interesting. They tend to show outliers in the data in two or more dimensions.

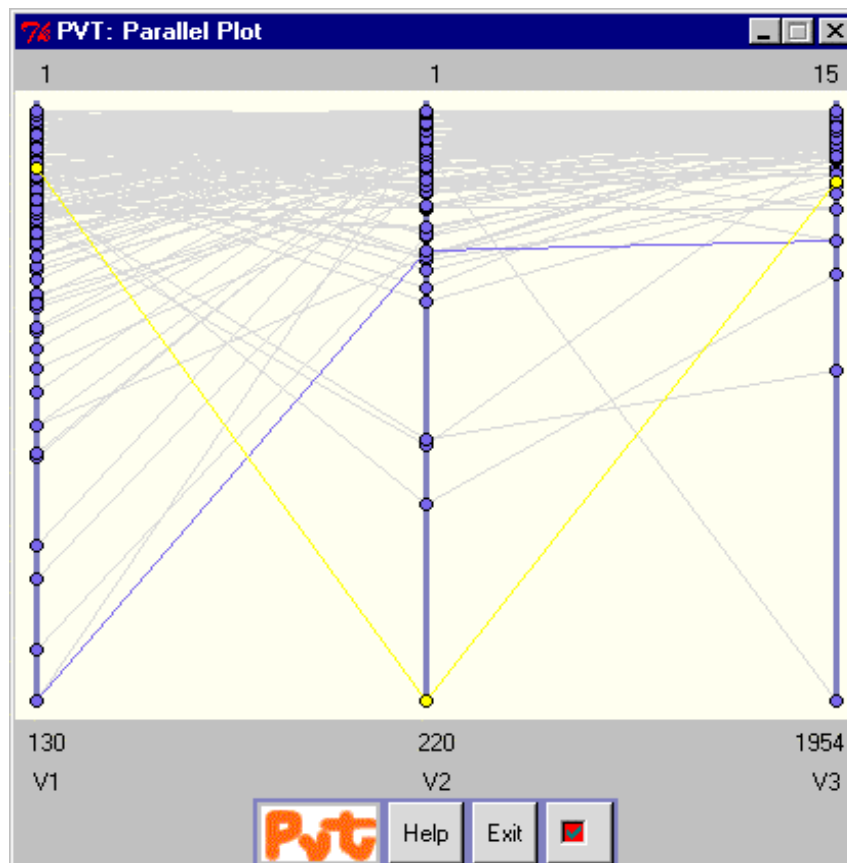


Figure 13 PVT: Parallel Coordinate Plot

4.2 Point Data Analysis with PVT: Brushing and Linking

The previous discussion illustrated how this traditional forms of representing data can aid in the exploration of massive point data sets. However, their utility is greatly enhanced when they are linked through brushing. This technique enables the user to link different representations of the same data, including that with the UPC on a map of the area (Figure 14) which permits the understanding of their spatial distribution.

By ticking the box next to the “Exit” buttons, a view can be linked to any other where the same box has been checked and thus prompt the brushing between views. When the user touches a point in any view, it will temporarily change colour to yellow in all the linked ones. Note that points in the scatterplot, parallel coordinate plot and map views are originally light blue. Therefore, after a point is queried in these views, it will return to its original colour. However, points in the dotplot views are red. Hence, if a point is queried in a dotplot it will turn red in all other views. Alternatively, if a point is queried in a scatterplot, parallel plot or in the map, points in a linked dotplot will permanently turn to light blue.

This simple feature of having three colours (yellow, red and light blue) to brush between views is very useful for analysing clusters or querying outliers. See for instance the parallel plot¹⁷ in Figure 14 where one of the lines has been permanently turned red because that point was originally queried in a dotplot. Alternatively, note that the dotplot has some blue points which were queried in one of the other views. This provides the user with a tool to select and analyse more than one cluster at any time by taking advantage of this form of dynamic linking which employs a change in the objects’ colour to relate various of its representations.

A third way of brushing the data is by drawing a box on any view by pressing the right mouse button and dragging it to enclose the points of interest. Notice the black box in the dotplot. Using this dragging-box feature, the enclosed points in all linked views will permanently turn yellow. By using these three colours i.e. red, yellow and light blue, the user can select and compare at least three clusters of data or three single points at the same time.

Figure 14 shows three points of interest which were permanently highlighted in yellow as explained above. The relationship between the three values stood out from the rest as can be seen in all the views. In the scatterplot and dotplot they appear as outliers from the rest of the data. Note the angles of the yellow lines joining the attribute values which populate the three postcodes. Two points are immediately noticeable. First, the one to the far right, which shows a relation of high retail employment (V2) to low floorspace (V3) and low entertainment employment. When analysed on the map view, this point corresponded to the postcode of a shopping centre warehouse. Also in the parallel plot, a second point shows a relationship of high V2 to high V3. When inspected on the

¹⁷ The parallel coordinate plot in Figure 14 is horizontal as opposed to that previously shown in Figure 13 since both options are available. Although they contain exactly the same information, one perspective may be preferred specially when several views are opened simultaneously on the screen.

map view, it corresponded to the shopping centre next to the warehouse. This example shows how simple visualisation techniques can allow a rapid and effective detection of pattern and clusters in the data sets.

These basic brushing techniques proved to be powerful for the users who tested PVT to discover spatial relations between the variables that had not be identified before. In particular, spatial relations between V2 and V3, not previously identified through the use of statistics or by transforming the data into density surfaces, were revealed by the simultaneous use of the various views provided by PVT. Locations where these relations took place were easily spotted using the map view provided.

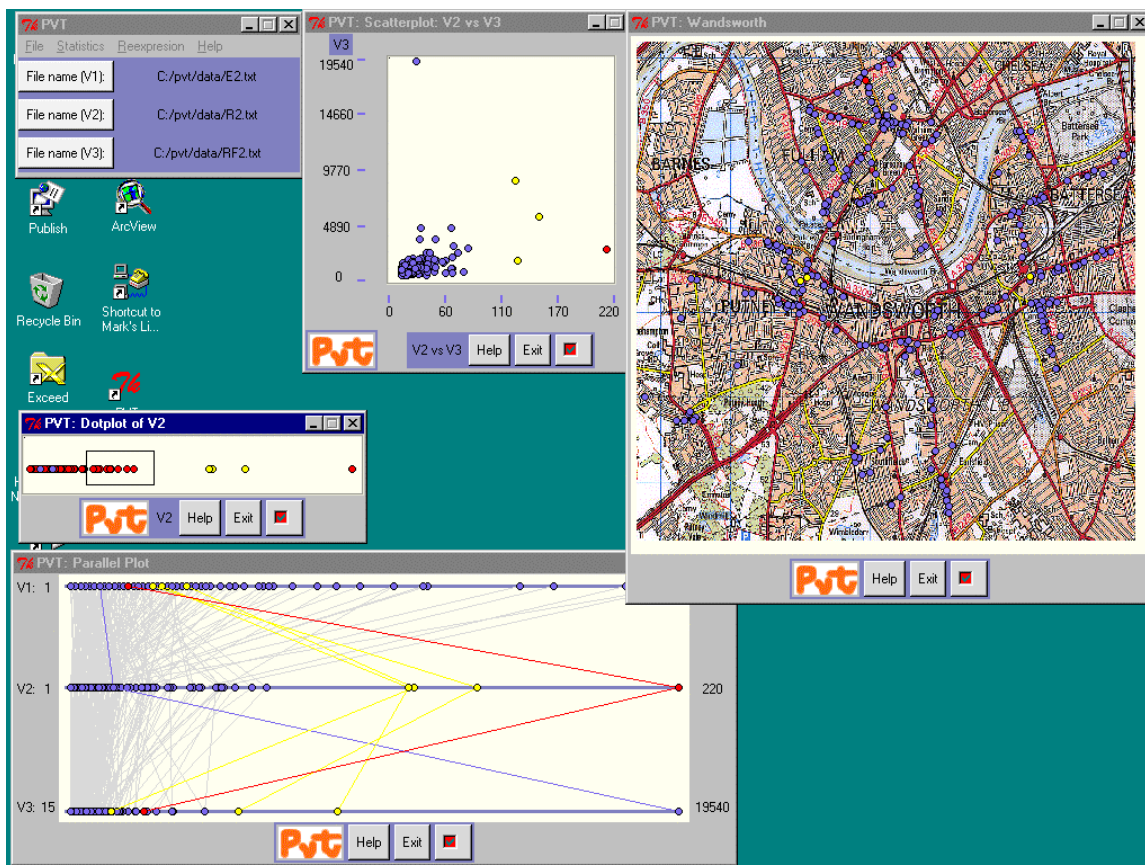


Figure 14 PVT: 3 Forms of Brushing

These facts point towards the usefulness of a tool that explores spatial point data at such level of resolution, before performing any operations on it, to learn about traits and to elaborate hypothesis that can be further tested. The ease of identification and exploration of clusters was the most useful

feature to users acquainted with the data since their identification is particularly problematic through the use of statistics¹⁸ and highly elaborate in the case of surfaces¹⁹.

4.3 Further Functionality

PVT was developed using Tcl/Tk which allowed the fast development of an interactive user interface. Therefore PVT can be thought of as a prototype or proof-of-concept system which shows some of the functionality that a tool of this nature should have. PVT's main objective was to enhance the understanding of how visualisation techniques and the use of dynamic mapping and brushing could serve the purpose of interrogating and learning from large amounts of spatial point data. The previous section showed results that indicate how such information can be obtained through the use of a tool which provides an interactive environment and can support the visualisation of multivariate point data. However, PVT can be more flexible and interactive in terms of allowing users more control of the views and forms of classifying data in the different representations such as the ones discussed below.

A desirable feature in scatterplots is that provided very successfully by Spotfire.net: filtering data and redefining the x and y scales interactively. It was previously discussed how the presence of outliers made the series appear more highly skewed and clustered than they would otherwise. Data filtering in a scatterplot would affect the display as it acts as a zoom on the data points. Outliers could also be permanently removed but statistical evidence should help decide whether a value is an unusual realisation of the data.²⁰ This could be accomplished by either linking PVT to a statistical package or developing the statistical functionality within it.

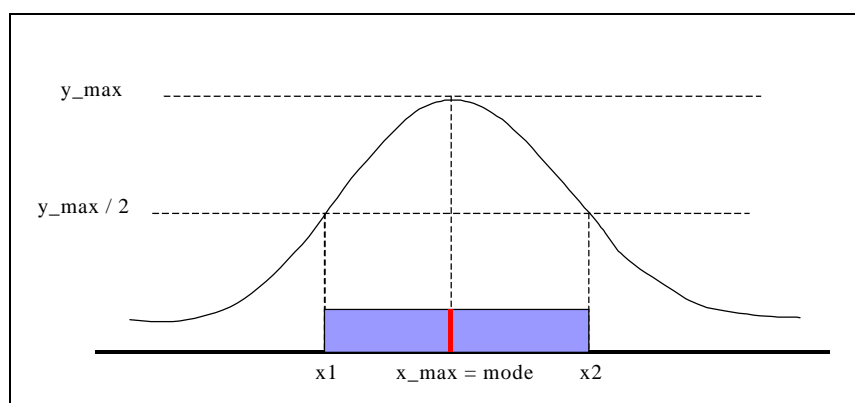


Figure 15 A Different 5 Number Summary

¹⁸ See for instance the discussion in Sarle (1996), Milligan and Cooper (1985) or Silverman (1986).

¹⁹ For example, Thurstain-Goodwin and Unwin (2000) developed what they call an Index of Towncentredness to identify, through the use of surfaces, clusters of points that met certain criteria.

²⁰ See for instance Wei (1990) for an example of an algorithm for the detection of outliers.

The five number summary provided by the Boxplot view proved to be a fast way to learn about the distribution of a series.²¹ However, other statistics could be used to define the box. Figure 15 shows a different set of statistics that use the mode instead of the median as a the measure of central tendency of the data. These statistics are relevant for unimodal distributions where a global y_{\max} can be clearly identified. Allowing the user to alternate between boxplots showing different statistics or five number summaries can provide views that are best suited for different types of data sets. In this same spirit, providing functionality to calculate and plot mathematical transformations of the variables (such as their logarithm) would allow a better understanding of the data distribution.

Histograms were not included in this version of PVT due to time constraints in its development. However, histograms a la Spotfire.net or MANET, where the users can interactively adjust the bins to see how their alteration affects the shape of the resulting distribution. Although histograms are one of the most common views for analysing the spread of data, they have been heavily criticised because of the “arbitrary nature of the bins used to categorise the continuous data values” (Orford et al., 2000) which either exaggerate or hide local variation. Therefore, other representations regarded as less arbitrary such as quintile plots and percentile plots,²² can be as effective for showing the variable’s distribution.

Parallel coordinate plots benefit from allowing the interactive addition and removal of variables or axes, as well as changing the order in which variables are assigned to the axes. This functionality would allow for the exploration of relations between various combinations of variables. At the present stage of development, PVT plots the data on their corresponding axes from their minimum to their maximum absolute values. Data could be classified in some other way, for instance by correlation or according to the bins decided for the histogram. GeoVista Studio uses colour coding of lines coming from different classification slots to further inform the relation between variables

²¹ In the present version of PVT, the minimum, maximum, median, lower and upper quartiles were chosen.

²² “In a quantile plot, the data are ranked from lowest to highest and the relative position of each observation is determined by its quantile value. A quantile value of a particular observation is simply the proportion of the data that have lower values than the observation itself. If the proportions are expressed as percentages, then the plot is called a percentile plot” (Orford et al., 2000).

5 Conclusions

The use of computer visualisation as a means to analyse complex geographic point data sets is a valuable tool for conducting ESDA. It makes use of the human eye's ability to recognise structure and relationships that may be inherent within the data. Traditional GIS are very poor for visualising point data since they offer limited functionality to explore it in a meaningful way besides transforming them into alternative representations or performing limited statistical analysis. Also, traditional GIS environments treat visual display as an output or endpoint of an analysis process rather than a means for hypothesis generation. PVT provides a way to explore the data as points which has various advantages. First, much geographical information comes in this form and the visualisation techniques described here allow for their meaningful, effective and fast exploration. Second, the data need not be transformed in any way that implies the making of assumptions before knowing about their characteristics, as is the case with density surfaces. Third, problems such as the dependency of areal units, or the visual dominance large areas in choropleth maps are not an issue at the point level of resolution. All this makes PVT an innovative tool for analysing data at an exploratory level, which is very informative in terms of identifying relations between variables such as clustering and correlation, as well as outlier values in one or more dimensions.

There is "no single package [which] combines the full range of tools necessary to support ESDA in a highly interactive graphical environment" (Wise et al., 1999). PVT, as well as the tools and software systems reviewed in this article, point towards the use of visualisation techniques as a successful way to break down and understand complex and large sets of multivariate data. There is not yet a unified effort even less a tool that integrates these developments. The initiative remains as a disjoint set of tools that tackle specific data and problems. However visualisation is a necessary first step in all spatial data analysis, simply because the position of particular attribute values on a map induces associative processes in the analyst, drawing upon analogies, possible prior information, or memory (Unwin et al., 1994). This can only be attained in an appropriate environment is designed which empowers users to flexibly interact and interrogate the data.

The visualisation techniques implemented in PVT are well known and have been successfully used in other software packages particularly for areal or lattice data. Their adaptation to investigate spatial point data showed its usefulness to identify pattern and spatial clustering. These results are being extended through research on further use of animation to explore this type of data. However, that is the topic of a forthcoming paper.

6 References

- Adrienko, G.L. and Adrienko N.V. (1997). Intelligent Cartographic Visualisation for Supporting Data Exploration in the IRIS System. *Programming and Computer Software*, 23, pp. 268-282.
- Adrienko, G.L. and Adrienko N.V. (1999). Interactive Maps for Visual Data Exploration. *International Journal of Geographical Information Science*. 13(4), pp. 355-374.
- Bailey, T.C., and Gatrell, A.C. (1995). *Interactive Spatial Analysis*. John Wiley & Sons, Inc.
- Bonham-Carter, G.F. (1994). *Geographical Information Systems for Geoscientists: Modelling with GIS*. Pergamon: Oxford, 398 pp.
- Buja, A., McDonald, J.A., Michalak, J. and Stuetzle, W.(1991). Interactive Data Visualisation. *Journal of Computational and Graphical Statistics*. 5, pp. 78-99.
- DiBiase, D., MacEachren, A.M., Krygier, J.B. and Reeves, C. (1992). Animation and the Role of Map Design in Scientific Visualization. *Cartography and Geographic Information Systems*, 19 (4), pp. 204-214.
- DiBiase, D., Reeves, C., MacEachren, A.M., von Wyss, M., Krygier, J.B. Sloan, J., and Detweiler, M.C. (1992). Multivariate Display of Geographic Data: Applications in Earth System Science. In MacEachern, A.M. and Taylor, D.R.F. (1994), *Visualization in Modern Cartography*, Pergamon: Oxford, pp. 287-312.
- Dykes, J. (1999). Interactive Maps for Exploratory Spatial Data Analysis: Cartographic Visualization; Approach, Implementation and Application. *Ph.D. Thesis*. University of Leicester, 186 pp.
- Dykes, J. and Unwin, D. (1998). Maps of the Census: A Rough Guide. <http://www.geog.le.ac.uk/jad7/AGOCG/>. Last Accessed March 2001.
- Fairbain, D., Adrienko, G., Adrienko, N., Buziek, G., Dykes, J. (1999). Representation and its Relationship with Cartographic Visualisation: A Research Agenda. *Cartography and Geographic Information Science*. 28(1).
- Ferreira, J. Jr. and Wiggins, L.L. (1990). The Density Dial: A Visualization Tool for Thematic Mapping. *Geo Info Systems*, 10, pp. 69-71.
- Fisher, P.F., Dykes, J. and Wood, J. (1993). Map Design and Visualization. *The Cartographic Journal*, 30, pp. 136-142.
- Fotheringham, A.S., Brunson, C. and Charlton, M.E. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis*. SAGE: London. 270 pp.
- Greene, W.H. (1991). *Econometric Analysis*. Mawell MacMillan: New York. 783 pp.
- Kahn et al. (1998). *Journal of Geophysical Research*, 103, pp. 32195-32213.

- Ousterhout, J. (1994). *Tcl and the Tk Toolkit*. Addison-Wesley: Massachusetts. 458 pp.
- MacEachren, A.M. and Kraak, M-J. (2000). *ICA Commission on Visualization and Virtual Environments: Research Agenda*. <http://www.geovista.psu.edu/icavis/pdf/visagenda.pdf>. Accessed March 2001.
- MacEachren, A.M. and Kraak, M-J. (2000a). *Research Agenda Development Process*. <http://www.geovista.psu.edu/icavis/agenda/>. Accessed March 2001.
- MacEachren, A.M. (1994b). Time as a Cartographic Variable. In Hearnshaw, H. and Unwin, D. (Eds.), *Visualization in Geographical Information Systems*, Wiley: Chichester, pp. 115-130.
- MacEachren, A.M. (1994). Visualization in Modern Cartography: Setting the Agenda. In MacEachren, A.M. and Taylor, D.R.F. (1994), *Visualization in Modern Cartography*, Pergamon: Oxford, pp. 1-12.
- MacEachren, A.M. and Monmonier, M.S. (1992). Introduction. *Cartography and Geographical Information Systems*, 19(4), pp. 197-200.
- McCormick, B.H., DeFanti, T.A. and Brown, M.D. (1987). Visualization in Scientific Computing. *Computer Graphics*, 21(6).
- Milligan, G.W. and M.C. Cooper (1985). "An Examination of Procedures for Determining the Number of Clusters in a Data Set". *Psychometrika*, 50, 159-179.
- Monmonier, M. (1989). Geographic Brushing: Enhancing Exploratory Analysis of the Scatterplot Matrix. *Geographical Analysis*, 21(1), pp. 81-84.
- Sarle, W. (1996). The Number of Clusters. <http://www.pitt.edu/~wpilib/clusfaq.html>. Last Accessed March 2001.
- Shepherd, I.D.H. (1995). Putting Time on the Map: Dynamic Displays in Data Visualization and GIS. In Fisher, P.F. (Ed.), *Innovations in GIS 2*, Taylor & Francis: London, pp. 165-187.
- Silverman, B.W. (1986). *Density Estimation*, NY: Chapman and Hall.
- Thurstain-Goodwin, M. and Unwin, D. (2000). Defining and Delineating the Central Areas of Towns for Statistical Monitoring using Continuous Surface Representations. *CASA Working Papers Series*, 18.
- Tomlin, C.D. (1990). *Geographic Information Systems and Cartographic Modeling*. Prentice Hall: NJ, pp. 249.
- Unwin, D.J., Dykes, J.A., Fisher P.F., Stynes K., & Wood, J.D. (1994). WYSIWYG? Visualization in the Spatial Sciences. *AGI Conference*. Birmingham.

- Unwin, A.R., Hawkins, G., Hoffman, H. and Siegl, B. (1996). Interactive graphics for Data Sets with Missing Values – MANET. *Journal of Computational and Graphical Statistics*, 5(2), pp. 113-122.
- Wei, W.S. and D.P. Reilly (1990). *Time Series Analysis: Univariate and Multivariate Methods*. Addison-Wesley: Massachusets.
- Wise, S., R. Haining and P. Signoretta (1999). Scientific Visualisation and the exploratory analysis of area data. *Environment and Planning A*. 31(10), pp.1825-38.

All web resources last accessed March 2001:

<http://allanon.gmd.de/and/IcaVisApplet/>

<http://www.anvilinformatics.com/tools-radviz.html/>

<http://cbl80.geog.bbk.ac.uk/pinweb/>

<http://www.geog.le.ac.uk/jad7/cdv/>

<http://www.geovista.psu.edu/icavis/>

<http://www.geovista.psu.edu/>

<http://www.spacestat.com/>

<http://www.spotfire.com/>