ps.2

Interactive displays for the visual exploration of point data

Carolina Toban, Centre for Advanced Spatial Analysis (CASA), University College London

Introduction

The "explosion of observational and model-based data and the development of computer visualisation tools" (MacEachren and Monmonier, 1992) has made visualisation an effective way of analysing vast amounts of information and multivariate datasets. When embedded in interactive and dynamic computing environments, visualisation techniques can facilitate the formulation of hypotheses about their relations, as will be presented here.[1] The capacity of graphics to reveal information such as patterns, outliers and trends from datasets has been acknowledged at least since the second half of the 1970s.[2] However hypotheses or insights obtained through visual exploration and analysis frequently need to be confirmed in some other way. This short paper suggests that the visual interrogation of multivariate datasets can be very effective to inform further statistical modelling of (spatial) variables.

The problem in hand was to analyse a so-called "lifestyles" database. This type of commercially available data is highly multivariate and usually built from a disparate set of data sources that are concatenated by means of a common field, usually their postcode. Although not devoid of problems and limitations, the dataset used was of interest as it contains information about income at the unit postcode level and it is updated on a yearly basis. The next section briefly discusses the general characteristics of lifestyles datasets. This is followed by a description of some of the insights obtained from the dataset when explored with an interactive visualisation tool. Sections 4 and 5 show results from the (spatial) statistical analysis of the data aimed at confirming (or rejecting) such observations. The final section reflects on the results and poses some directions for future research.

1 LIFESTYLES DATA

"Lifestyle databases contain non-aggregated individual-level data (or household data) that can be accessed and statistically manipulated to gain knowledge about people, particularly their consumer habits and behaviours. Lifestyles data have been predominantly used for direct marketing, especially to target promotional literature through the letter-boxes of existing or potential consumers ('mail targeting')" (Harris, 1999). Lifestyles data sets are generally very diverse and frequently concatenated from disparate data sources using a common field, the residential address, to form a profile of individual households. Most lifestyles datasets are obtained from responses to postal consumer surveys or questionnaires (Harris, 1999). Each survey usually consists of a few hundred questions on varied "lifestyles" topics ranging from income and occupation to shopping habits, holiday and travel preferences.

The lifestyles dataset used for this study was available at the address level and variables were aggregated to unit postcode. 844 observation were available for the Bristol area. The main drivers for using this dataset have been that: i) it is updated every year (as opposed to once every ten years as in the case of the UK Census of Population), and ii) it contains information on income at the postcode level. One of our main

[1] For examples of such computer environments for different data types see Ferreira and Wiggins (1990), Unwin, *et al* (1996), Adrienko and Adrienko (1999), Dykes (1999), Wise, *et al* (1999), or Tobon (2001).
[2] See for instance Tukey (1977), McCormick, *et al* (1987), Buja, et.al. (1991), MacEahern and Taylor (1994), Dykes (1999).

interests was to model how income (as well as other available proxy variables) is distributed across space using data at this fine level of resolution. The income variable in the lifestyles database is nevertheless available in value ranges. These had to be transformed to approximate a series of real values that could be modelled statistically.

This illustrates one of the many limitations of lifestyles databases. Since the data is collected for marketing purposes, answers to each question must be chosen from a set of pre-defined responses that are not always suitable for statistical analysis. A second limitation is that there is no obligation to complete the questionnaires. Hence, the respondents are self-selecting and constitute a non-random sampling of the population (Harris, 1998). These limitations make lifestyles data "fundamentally more biased and unscientific" (Longley, 1998). However, the increased geographic coverage, detail and periodicity of lifestyles datasets offer an alternative source of information that can challenge the traditional use of Census-based geodemographic data both for marketing applications and academic research.[3]

2 DYNAMIC VISUAL EXPLORATION OF THE DATA

The lifestyles dataset variables were initially represented as points and visually explored using a Point Visualisation Tool (PVT) that facilitates the detection of patterns, trends and outliers in the variables (Tobon, 2001). In PVT, multiple representations of the same variable can be interrogated simultaneously through "brushing" (Monmonier, 1989), or dynamically linking such representations as illustrated in Fig. 1 by highlighting points in a view. In every view, each point represents an attribute values and its associated postcode.[4] Although each view conveys only a subset of the characteristics of the data, this can be compensated by linking them "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 enables the linking of various representations of the dataset on screen allowing their simultaneous inspection.

Clusters of similar and outlier values were found and their location in space identified using the simple interactive visualisation technique of brushing. In Fig. 1, note the set of points highlighted in yellow in the Scatterplot where the income variable (labelled as "V2") is plotted against the number of households with two cars (variable labelled as "V1"). Note that points with corresponding postcodes are also highlighted in yellow in the window labelled "Map", where data points are plotted according to their UK grid coordinate. The points highlighted in yellow correspond to combinations of the aforementioned variables that behave rather differently than the rest of their counterparts. It can be seen in the Map how most of these points are clustered towards the northwest of Bristol. Note also that a few points of interest have been highlighted in a different colour (red) as the spatial location of the relative extreme values of income was of interest.[5]

3 LISA

The previous section illustrates one of many other discoveries about relationships between variables of the lifestyles dataset that were made using a visual analysis tool such as PVT. Simple techniques such as brushing various graphical representations of the same data gave some indication of the correlation of the variables across space. This prompted the use of Local Indicators of Spatial Association (LISA), which measure the spatial dependence (or independence) between locations and provide evidence of whether spatial modelling of a variable is appropriate. LISA identify the association between a single value of a variable on one location and its neighbours. Two of such indicators—local Moran's *I* and local Geary's *c*—which take a variable's location explicitly into account, were used to test for spatial autocorrelation. Following Anselin (1995), they can be formally defined as:

$$I_i = z_i \sum_j w_{ij} z_j$$

[3] For details see Openshaw and Turton (1998).

[4] The only exception is the Map view where each point represents a postcode plotted according to its UK coordinates.

[5] Similar analysis was done for all of the lifestyles database variables using these and other tools available in PVT.

$$c_i = \frac{n}{\sum_i z_i^2} \sum_j w_{ij} \left(z_i - z_j\right)^2$$

where i and j refer to the spatial units of which there are n; $z_i = x_i - \bar{x}$, where $\bar{x}$ is the mean of x or the attribute being measured; w_{ij} is the weight or degree of connection between zones i and j and it is defined according to some distance threshold or a degree of adjacency between the two locations. Positive values of Moran's *I* indicate "spatial clustering of similar values (either high or low), and negative values a clustering of dissimilar values (for example, a location with high values surrounded by neighbours with low values)" (Anselin, 1995). Hence, low (negative values) of Moran's *I* indicate negative spatial autocorrelation. Geary's *c* is a similar statistic but it measures the squared difference of deviations from the mean of a value in a location and deviations from the mean of the values of its neighbours. By contrast to Moran's *I*, low values of Geary's *c* indicate positive autocorrelation and high values indicate negative spatial autocorrelation.[6]

This part of the analysis was performed in the SpatialStats module of S-Plus 4.5. Results were visualised in ArcView GIS 3.2a using its link to S-Plus, S+SpatialStats. LISA in this module can only be estimated for lattice data so the variables had to be aggregated by Enumeration District (ED). Fig. 2 shows the results of the local Moran's *I* for the income variable colour coded by I_i value.[7] This gives an indication of local spatial clusters[8] and hence evidence that spatial modelling of the income variable might be appropriate.

Going back to PVT, Fig. 2 was used as a backdrop in the Map view to give further context to the visual understanding of the data. In Fig. 3, note how the points initially identified as having an extreme two-dimensional value (or 'outlier' combination of income against households with two cars), and detected as being spatially clustered, do correspond to EDs of high positive autocorrelation. The backdrop image allowed further interrogation of values by location and an understanding of where certain attribute values (such as outliers) occurred.

4 SPATIAL STATISTICAL ANALYSIS

Once evidence was gathered about spatial autocorrelation of the income variable, it was appropriate to go one step further and attempt its statistical modelling. Two types of spatial regression models were used: Autoregressive (AR) and Moving Average (MA). The first assumes correlation in the dependent variable (income in the present case). Following Fotheringham *et al* (2000), it can be formally expressed as:

$$y = X\beta + \rho W y + \varepsilon$$

where y is the vector of values for the dependent variable, X is the matrix containing all independent or explanatory variables from the lifestyles dataset, β is a vector of regression coefficients, W is an adjacency matrix,[9] ρ the vector of the corresponding regression coefficients, and ε is a vector of random and independently distributed errors. The Moving Average model on the other hand assumes the correlation occurs in the error terms and can be expressed as:

$$y = X\beta + u$$

where:

$$u = \rho W u + \varepsilon$$

[6] Only results for the Moran's *I* are showed for succinctness but all the results are shown in Table 1.
[7] Note that the Moran's *I* indicator shown was estimated using a neighbourhood definition that follows an adjacency criterion. The software's procedure is to build a spatial weights matrix based on the adjacency of spatial units or polygons (this corresponds to the first order adjacency). This is then adjusted to include all other polygons whose centroid-to-centroid distance is less than the average of all other polygons considered as neighbours by the first criterion (which is called an adjusted first order criterion).
[8] Although this should be interpreted with care (see Anselin, 1995).
[9] Various definitions of this matrix were explored and used in the estimations as in the case of the estimation of the LISA. However, only results for the Adjusted First Order adjacency criteria are shown here as models using it had a better fit.

Results for these models are shown in Table 2. The MA model was chosen to better fit the data as: i) ρ had the highest value "suggesting a strong degree of correlation in the error term" Fotheringham *et al* (2000); ii) the R-squared (r^2 in Table 2) is the highest showing that about 66.2% of the variation in income is explained by the regression; and iii) the Akaike Information Criterion (AIC), which models the model fitting, is the smallest for the MA model indicating its superiority over the AR.

5 FURTHER THOUGHTS

"The increased availability of large spatially referenced data sets and the sophisticated capabilities for visualization, rapid data retrieval, and manipulation in geographic information systems (GIS) have created a demand for new techniques for spatial data analysis of both an exploratory and a confirmatory nature" (Anselin, 1995). This paper has presented evidence of how visualisation environments, which allow the dynamic and interactive exploration of datasets, allow fast and efficient information discovery and hypotheses formulation. Interesting findings can then inform further routes data investigation and analysis, for instance using statistical modelling.

However, one obvious observation is that the analysis performed here could not be carried out using a single package. The data was initially explored using a visualisation tool (PVT), the statistical analysis performed in a statistical package (S-Plus), and the visualisation of the Moran's *I* and other statistical results was partly done in ArcView and partly back in PVT. Unfortunately, there is no package that integrates exploratory visual techniques with statistical analysis and hence the power of this two complementary forms of data analysis. Geographic Information Systems (GIS) 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. Statistical packages usually support multiple representations of a dataset, but they do not facilitate their simultaneous exploration or linking. Conversely, visualisation packages provide little statistical functionality and usually support a limited set of data structures.[10] Although efforts have been done to integrate these two forms of data analysis (see for instance Anselin (1998), Wise *et al* (1999)), this is still an area with much room for research and development.

REFERENCES

ANSELIN, L. (1995). "Local Indicators of Spatial Assosiation." *Geographical Analysis*, 27(2), 93-115.

ANSELIN, L. (1998). "Exploratory Spatial Data Analysis in a Geocomputational Environment," in P. A. Longley, S. M. Brooks, R. McDonell, and B. Macmillan (eds) *Geocomputation: A Primer*, Chichester: John Wiley and Sons Ltd, 278.

ADRIENKO, G.L. and ADRIENKO, N.V. (1999). "Interactive Maps for Visual Data Exploration." *International Journal of Geographical Information Science*. 13(4), 355-374.

BUJA, A., MCDONALD, J.A., MICHALAK, J. and STUETZLE, W. (1991). "Interactive Data Visualisation." *Journal of Computational and Graphical Statistics*. 5, 78-99.

MACEACHERN, A.M. and TAYLOR, D.R.F. (1994), *Visualization in Modern Cartography: Setting the Agenda,* Pergamon: Oxford, pp. 287-312.

DYKES, J. (1999). Interactive Maps for Exploratory Spatial Data Analysis: Cartographic Visualization; Approach, Implementation and Application. *PhD Thesis* (Unpublished). University of Leicester, 186 p.

FERREIRA, J. JR. and WIGGINS, L.L. (1990). "The Density Dial: A Visualization Tool for Thematic Mapping." *Geo Info Systems*, 10, 69-71.

FOTHERINGHAM, A.S., BRUNSDON, C. and CHARLTON, M.E. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis*. SAGE: London. 270 p.

GETIS, A., AND ORD, J. K. (1996), "Local Spatial Statistics: An Overview," in P. A. Longley, and M. Batty (eds) *Spatial Analysis: Modelling in a GIS Environment*, Cambridge: GeoInformation International, 261-277.

[10] See Anselin (1998) for a good discussion of implemented solutions to this problem and their shortcomings and limitations.

HARRIS, R.J. (1999). Geodemographics and the Analysis of Urban Lifestyles. *PhD Thesis* (Unpublished). University of Bristol. 383 p.

LONGLEY, P.A. (1998). "Foundations." In Longley, P.A., Brooks, S.M. and McDonnell, R. (eds) *Geocomputation: A Primer*. Chichester: John Wiley and sons, 3-15.

MCCORMICK, B.H., DEFANTI, T.A. and BROWN, M.D. (1987). "Visualization in Scientific Computing." *Computer Graphics*, 21(6).

MONMONIER, M. (1989). "Geographic Brushing: Enhancing Exploratory Analysis of the Scatterplot Matrix." *Geographical Analysis*, 21(1), 81-84.

OPENSHAW, S. and TURTON, I. (1998). "Geographical Research using Lyfestyles databases." http://www.geog.leeds.ac.uk/staff/s.openshaw

TOBON, C, (2001). "Interactive Displays for the Visual Exploration of Point Data." *Proceedings of the GIS Research in the UK Annual Conference, GISRUK 2001*, University of Glamorgan, 614-620.

TUKEY, J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley, 688 p.

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), 113-122.

WISE, S., HAINING, R. and SIGNORETTA, P. (1999). "Scientific Visualisation and the exploratory analysis of area data." *Environment and Planning A*. 31(10), 1825-38.

Carolina Tobon
Centre for Advanced Spatial Analysis (CASA)[11]
University College London (UCL)
and
Department of Geography, UCL[12]

Email: c.tobon@ucl.ac.uk

[11] 1-19 Torrington Place, Gower Street, London, WC1E 6BT; Tel: (+44) (0)20 7679 4260, Fax: (+44) (0)20 7679 4293; URL: http://www.casa.ucl.ac.uk/

[12] 26 Bedford Way, London, WC1H 0AP; Tel: (+44) (0)20 7679 5500, Fax: (+44) (0)20 7679 7565; URL: http://www.geog.ucl.ac.uk

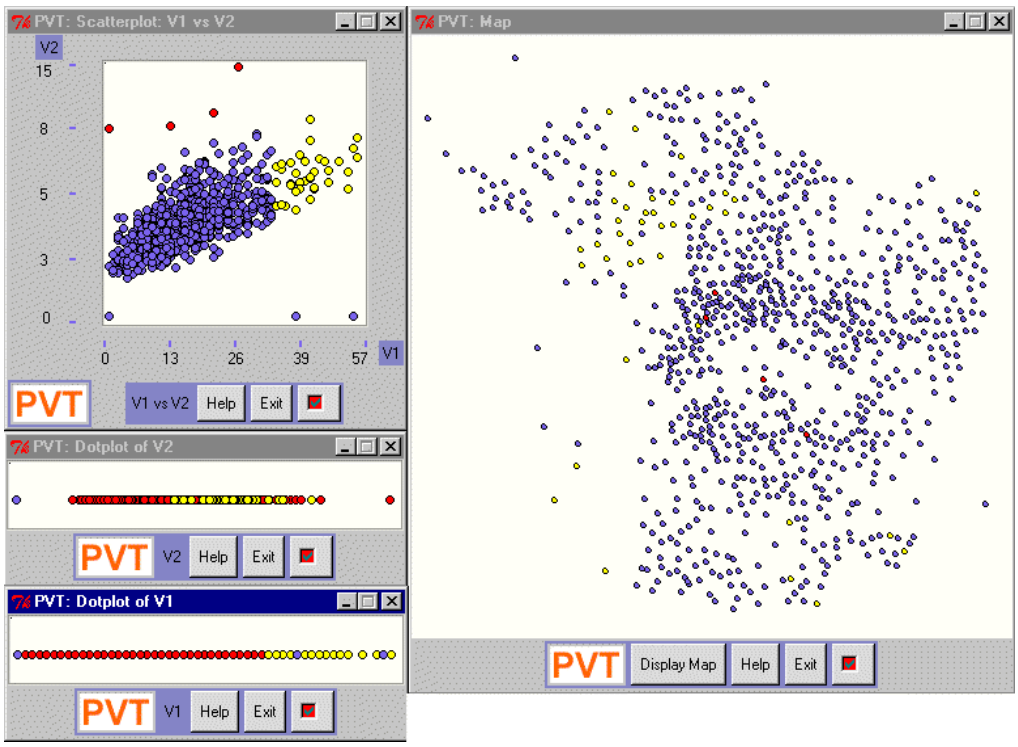

Figure 1: Visually exploring the lifestyles dataset. Note the use of colour to investigate a point in different views and detect spatial patterns or outliers. Note how the highlighted points in the Scatterplot correspond to a spatial cluster of points as seen in the "Map" view on the right.

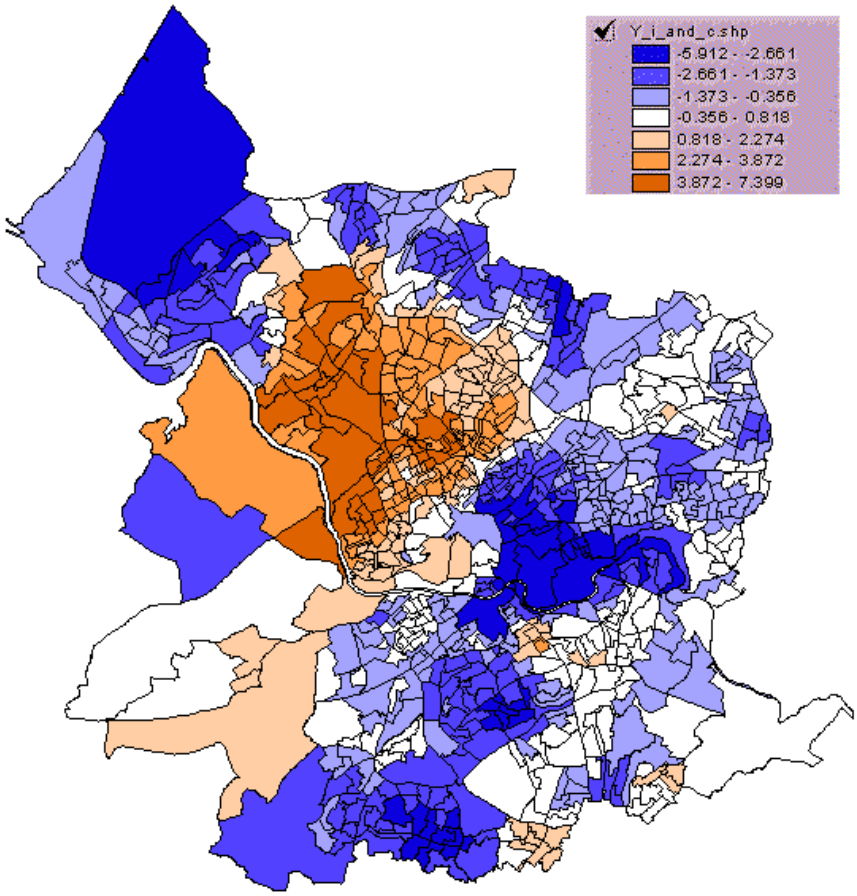

Figure 2: Local Moran's I for the lifestyles' income score variable. Note that shades of blue indicate negative spatial autocorrelation while shades of orange positive correlation.

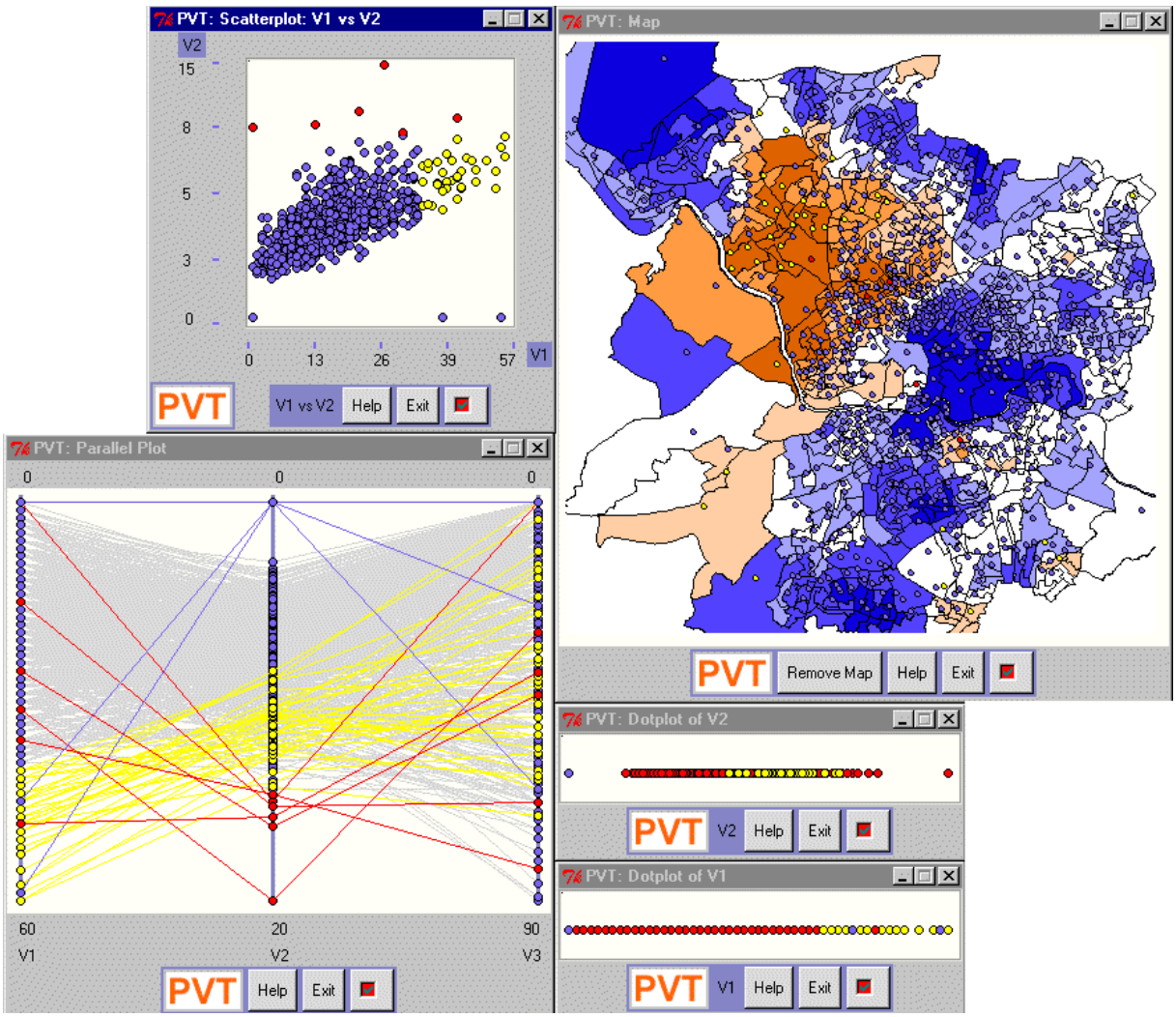

Figure 3: Providing context to the visualisation in PVT by including an ED map of Bristol. Colour coding corresponds to Moran's I values of spatial autocorrelation of the (lifestyles) income variable as explained in Fig.2.

LOCAL MORAN	STD MORAN	LOCAL GEARY	STD GEARY
0.46964210	-2.96835473	0.25376104	-1.76709896
0.11786247	-0.73750789	0.05673962	-1.21091547
0.52484895	1.01287730	0.16120503	-1.21448137
0.17676890	-0.44570092	0.77315771	-0.79755829
0.11759074	0.42971668	0.56715842	-0.73079883
0.70513403	-1.67714555	0.64639121	-1.31732316
0.11329480	-1.26224983	0.26172848	-0.85162845
-0.23060082	-1.72446354	1.43317956	0.40305801
0.03852975	-0.57434035	0.06816063	-0.95054035
0.17044153	-0.89661416	0.09580423	-1.07562014
-0.03846063	-0.33993055	0.60640815	-0.67602462
0.80019877	-1.85704449	0.54816208	-1.23252006
0.04377593	-0.61548519	0.06786481	-1.06286423
-0.01750287	0.04281469	2.15837948	0.22016180
0.38473077	-1.38831145	1.35487274	-0.57527386
0.20409279	-0.83334439	0.17717439	-1.30485522
-0.03061855	-0.05896041	2.73454130	0.42876789
0.38410438	-0.99624067	1.54252494	-0.44179763
0.48716861	-1.95047712	0.24005873	-1.37161231
0.00050023	-1.63925753	1.14913110	0.23070752
1.28841414	-1.59940617	1.00894096	-1.28064813
0.21615481	-1.25871536	0.12846751	-1.45070459
0.00022118	-0.80225479	0.48246345	-0.89535825
0.05175776	-0.41151651	0.68669247	-0.52190892
0.34584940	-1.10764807	0.23239928	-1.23589372
0.27967690	-1.46847408	0.20114667	-1.08283399
0.28298529	0.49861571	3.46413715	-0.45828634
1.77463896	-3.35720649	0.69586586	-2.63041315
0.53383615	-1.54176155	0.32865131	-1.21669283
0.27975162	-1.52304482	0.04483718	-1.24277497
0.21331742	-1.14064528	0.05910619	-1.23056033
1.25474733	-2.71922036	0.30564522	-1.71095658
0.61640237	-2.20854255	0.05850295	-1.63089661
0.28417194	-1.41179176	0.00722897	-1.42112941
0.63104951	-1.27678358	0.71844823	-1.05370782
-0.14439768	1.68694228	1.67160605	0.83219189
0.91440940	-2.80460989	0.25836482	-1.54958909
0.97111828	-2.53218384	0.19858288	-1.21920136
-0.03146434	-1.55359719	0.76965659	-0.34875894
0.10575337	0.96456391	1.21375823	0.13597202
0.47020742	-2.05086270	0.06049539	-1.52615545
0.55535361	-1.56547869	0.24294112	-1.47793580
0.25917854	-1.31160610	0.04060752	-1.13529779
0.81132506	1.35537767	2.31220639	-1.28540575
-0.04878404	0.19353876	1.45265502	-0.27888984
0.56121973	-3.19249114	0.24522689	-1.94115499
0.86032053	-3.31042299	0.19815133	-1.69622638
0.40257831	-1.49995688	0.01737803	-1.36636471
1.01280032	-1.85343166	0.21762679	-1.50198315
-0.04527758	-1.02221065	0.43568514	-0.67797833
0.63641465	-1.14273027	0.87755120	-1.09678769
0.53018101	-1.85495320	0.40498240	-1.55620528
0.32057122	-1.67728121	0.35381059	-1.11459029
0.20132207	-1.47823498	0.26823422	-0.95658382
0.28064971	-2.19934928	0.25364039	-1.21557463
0.05557280	-0.85580387	0.07249354	-1.34038492
0.63482010	2.24939527	0.60258982	-1.16505346
-0.14192651	-1.83675886	1.15908460	0.16508147
1.56882727	2.62448471	0.19603534	-1.96831843
0.39025057	-1.54483869	0.14528290	-1.48068901
1.53625073	4.14444622	0.53828480	-2.98917299
0.32070284	-0.99503825	0.51588859	-1.00053248
1.00431467	-2.26767454	0.40530084	-1.78730878
0.50296447	-1.73107299	0.31792751	-0.97737648
1.29991707	-2.11994237	0.35431697	-1.55374663
-0.01087197	0.05782039	0.98319583	-0.34344732
0.10580727	-1.24096531	0.06675113	-1.59589216
1.20792735	-2.82383503	0.05066079	-1.61391711
-0.22732826	-0.76512588	1.83157657	0.04388623
1.18439813	-2.18326162	0.56875354	-1.51321320
0.72526687	-2.75738167	0.07062581	-1.71752497
-0.03318218	-1.09995618	0.28004386	-1.15483158
0.31965446	-1.37417329	0.43758635	-1.00596289
0.07084599	-0.27667047	0.91639871	-0.56925173

Table 1: Moran's *I* and Geary's *c* values and standardised values. The images presented in this paper correspond to the standardised values which are simply:

$$StdLISA = \frac{LISA - E[LISA]}{StdDev_{LISA}}$$

	AR		MA	
Intercept	2.2499	25.4153	2.2523	25.4153
BIGACC	0.0237	7.4836	0.0237	7.4794
Buying	0.0224	7.3428	0.0224	7.3337
DETCH	0.0138	7.2862	-0.0114	-3.5582
SEMIDET	0.0309	4.8319	-0.0063	-4.1302
TERRCE	-0.0117	-3.6888	-0.0061	-4.0408
HHTCAR	-0.0064	-4.1831	0.0306	4.7749
QUALM	-0.0061	-4.0625	0.0136	7.2094
RHO	0.03018		**0.0425**	
LOG-LIKELIHOOD	-2451		**-2567**	
R^2	0.659813		**0.66159**	
AIC	0.391497		**0.299014**	

Table 2: Results of spatial autoregressive (AR) and moving average (MA) models of income. Out of 59 possible explaining variables, these seven (plus the intercept) were the most significant explaining variables. Numbers in bold show that the MA model fits the data better.