Visualizing the Effects of Scale and Geography in Multivariate Comparison

Sarah Goodwin* Jason Dykes† Aidan Slingsby‡

giCentre, City University London

Abstract

Our research investigates the sensitivities and complexities of visualizing multivariate data over multiple scales with the consideration of local geography. We investigate this in the context of creating geodemographic classifications, where multivariate comparison for the variable selection process is an important, yet time-consuming and intensive process. We propose a visual interactive approach which allows skewed variables and those with strong correlations to be quickly identified and investigated and the geography of multi-scale correlation to be explored. Our objective is to present comprehensive documentation of the parameter space prior to the development of the visualization tools to help explore it.

Index Terms: D.2.2 [Software Engineering]: Design Tools and Techniques; I.5.2 [Pattern Recognition]: Design Methodology—Feature Evaluation and Selection

1 Introduction

The comparison of geographically varying phenomena is both position and scale dependent. We investigate this in the context of creating and visualizing geodemographic classifications. Geodemographics group geographical areas by similar population characteristics and are used by academics, governments and professionals to identify typical population or customer characteristics [5].

The selection of variables through comparison is an important part of building the classifier and variables should be independent, of near-normal distribution and have little or no correlation to one another [5]. The variable selection (known in clustering as 'feature selection' [6]) is a time consuming and intensive process [5, 13], which may be subjective to user interpretation. We propose a visual interactive approach to aid the process, allowing skewed and strongly correlating variables to be quickly identified and investigated and the geography of multi-scale correlation to be explored.

Scale and geography are of particular importance in our proposal as knowledge of local variations may influence variable selection and classifications can be created at multiple scales with each likely to produce very different outcomes. There is limited research in the area of spatially weighted geodemographics [1] or varying geodemographic scales. Our research investigates the sensitivities and complexities of visualizing multiple data variables over multiple scales with the consideration of local geography.

2 Data Sources

This research follows previous work on investigating domain specific geodemographic visualization and creation in the context of energy consumption [3]. We use small-area summary statistics from the 2011 UK Census [9], based on the open geodemographic methodology [13], combined with energy consumption data for gas and electricity from the 'Department of Energy and Climate Change' (DECC) [2].

*e-mail:Sarah.Goodwin.1@city.ac.uk
†e-mail:J.Dykes@city.ac.uk
‡e-mail:Aidan.Slingsby.1@city.ac.uk

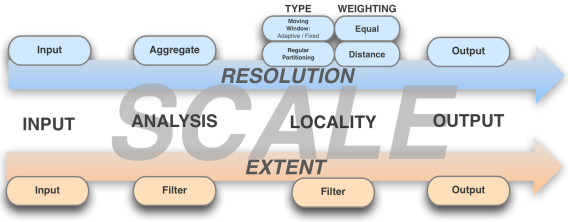

Figure 1: Four stages of the process: Input, Analysis, Locality and Output, each with two dimensions of Scale: Resolution and Extent

3 Scale

Figure 1 identifies four stages of the variable selection process in which scale can be varied: *Input*, *Analysis*, *Locality* and *Output*. Adjusting the scale particularly at the two central stages allows the associated sensitivities to be explored. At each stage there are two dimensions: *Scale Resolution* and *Scale Extent* [7, 12], which are defined as:

Scale Resolution (SR) - the level of aggregation used to make comparisons. When data is aggregated the nature of the summaries used to describe areas at each scale, and relationships between them, can vary. Aggregation of data can remove outliers and is associated with the modifiable areal unit problem (MAUP) [11]. The use of visualization to illustrate how different variables react to changes in resolution may help to identify the optimal resolution for analysis as well as illustrate the effects of MAUP.

Scale Extent (SE) - the geographical extent of the data; for example selecting the whole of the dataset or a subset (a geographic filter) of the data can lead to entirely different results.

The four stages introduced above, can be defined as follows:

Input - resolution (IR) and extent (IE) refers to the smallest areal unit and full extent of the 'raw data'. For our data sources this is Output Area [10] for the Census variables and Lower Super Output Area [10] for DECC. Both sources have an IE that covers England and Wales.

Analysis - resolution (AR) and extent (AE) refers to the scale for the chosen analysis. The IR maybe aggregated to a larger areal unit for example Local Authority region (AR) and/or the IE can be filtered to a specific geographical area of interest (AE), such as Wales or Greater London.

Locality - resolution (LR) and extent (LE) allows for the calculation of summary statistics at varying local as well as global scales. Such local summary statistics can be calculated in various ways as indicated by *Type* in Fig. 1. These include using a *Moving Window* technique with a *Fixed* (number of areas) or *Adaptive* (using a distance measurement) kernel or by using *Regular Partitioning*, where a grid (of a certain distance) is overlaid on to the data (size $>$AR).

Weighting refers to whether the areal units within the moving window or partition are given equal or distance weighting to the cal-

Table 1: Table identifies the ability to make comparisons when visualizing multiple Scale Resolutions (SR) and Extents (SE) with increasing numbers of variables (V) and local summaries (L)

	Distribution	Correlation	
	V=1	V=small	V=large
L=1	SR: Many SE: Many	SR: Some SE: Limited	SR: Limited SE: None
L=small	SR: Many SE: Many	SR: Some/Limited SE: Some/Limited	SR: Limited/None SE: Limited/None
L=large	SR: Many SE: Many	SR: Limited/None SE: Limited/None	SR: None SE: None

culation of the local statistic. This framework is based on the principles of Geographically Weighted Modelling [4]. LE is changed from AE only if locally weighted statistics are needed in a subset of the analysis, for example to investigate locally weighted statistics in London compared to elsewhere.

Output - resolution (OR) and extent (OE) refers to the dimensions of the data once it has been through the previous stages and is ready for spatial aggregation to a lower resolution. OR = AR unless *Partitioning* has been choosen in *Locality* then OR will take the size of the partition. OE = AE, unless LE has been utilized.

4 Visual Comparison

Through the utilization of *Locality* we can calculate local as well as global summary statistics for each variable and with this the complexity of the visualization options increase. The visual representation of such a complex set of scales can be simplified by considering scale in three broad and loosely delimited bands: global (as used in cases where local variations are not considered), macro and micro. Where L = 1 for Global, L = small (but >1) for macro and L = large for micro. The point at which macro becomes micro depends upon the number of variables being shown (V), the number of data points in the comparison, the visualization represented and the users' experience and display possibilities. The ability to make comparisons when exploring the parameter space reduces with increased V and L, as shown in Table 1. This ability to explore the data must be reflected in an adaption of the visual representation at these thresholds. Possibilities for visually encoding these data are multifarious. Given the need to compare skewness of variables and strong correlations both globally and locally we propose two types of visual representation: Statistical and Spatial as shown in Table 2.

4.1 Statistical and Spatial Views

As shown in Table 2 when V and L are large presenting a detailed comparison visually becomes difficult and here we rely on color encoding of the correlation coefficient (or other descriptive statistics in the case of V=1) for a space efficient representation. Matrices in which cells represent pairs of variables can be useful in the layout - whether this is through multiple scatterplots [8], maps showing the geographies of correlation of all pairs of variables or a color encoded grid cell showing the global level of association between each pair. Asymetrical matrices have been identified as a possible way to compare two differing datasets: for example before and after a data transformation.

5 Conclusion

Having established the need for visual representation to support the sensitive and time-consuming issue of variable selection we have produced a framework for considering and visualizing the multiple dimensions of scale and the effects of geography in this process. An interactive application through which these effects can be explored through this framework is in development with novel candidate designs established. Our poster uses the framework to present these designs graphically, describe the prototype through which the framework is explored and offer reflection and a discussion of opportunities for improvement and future work.

Table 2: Table identifying Statistical (top) and Spatial (bottom) visualization possibilities when considering a balance between number of variables (V) and number of local summaries (L). Characteristics of display, user, task and data will be influential in establishing appropriate methods in specific cases

	Distribution	Correlation	
	V=1	V=small	V=large
L=1	Histogram with dot plot	Matrix of Scatterplots	Color encoding
	Choropleth Map (Cartogram or Treemap)	Series of Choropleth Maps	Color encoding
L= Small	Boxplots or Histograms	Matrix of Scatterplots (showing L)	Color encoding
	Choropleth Map	Matrix of Correlation Maps	Matrix of Correlation Maps
L= Large	Color encoding	Color encoding	Color encoding
	Choropleth Map	Matrix of Correlation Maps	Color encoding

Acknowledgements

This PhD research is funded by a Vice Chancellor's Scholarship from City University London and undertaken through collaboration with the g2Lab, Hafencity University.

References

[1] M. Adnan, A. Singleton, and P. Longley. Spatially weighted geodemographics. In *GIS Research UK 21st Annual Conference*, Liverpool University, Apr. 2013.

[2] DECC. Sub-National Electricity and Gas Consumption Statistics - http://bit.ly/1bCqsb9, 2013.

[3] S. Goodwin and J. Dykes. Visualising variations in household energy consumption. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 217–218. IEEE, 2012.

[4] P. Harris, C. Brunsdon, and M. Charlton. Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, 25(10):1717–1736, 2011.

[5] R. Harris, P. Sleight, and W. R. *Geodemographics: GIS and Neighbourhood Targeting*. Wiley-Blackwell, 2005.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[7] N. Lamand and D. A. Quattrochi. On the issues of scale, resolution and fractal analysis in the mapping sciences. *The Professional Geographer*, 44(1):88–98, 1992.

[8] M. Monmonier. Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis*, 21(1):81–84, 1989.

[9] ONS. Census Data: http://bit.ly/onsCen11, 2011.

[10] ONS. Census Geographies: http://bit.ly/cenGeog, 2011.

[11] S. Openshaw and P. Taylor. *The modifiable unit areal problem*. Norwich:Geobooks, 1984.

[12] C. Turkay, A. Slingsby, H. Hauser, J. Wood, and J. Dykes. Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data. *IEEE Transactions on Visualization and Computer Graphics*, Dec 2014.

[13] D. Vickers and P. Rees. Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):379–403, 2007.