

Permanent City Research Online URL: http://openaccess.city.ac.uk/3876/

Copyright & reuse
City University London has developed City Research Online so that its users may access the research outputs of City University London's staff. Copyright © and Moral Rights for this paper are retained by the individual author(s) and/ or other copyright holders. All material in City Research Online is checked for eligibility for copyright before being made available in the live archive. URLs from City Research Online may be freely distributed and linked to from other web pages.

Versions of research
The version in City Research Online may differ from the final published version. Users are advised to check the Permanent City Research Online URL above for the status of the paper.

Enquiries
If you have any enquiries about any aspect of City Research Online, or if you wish to make contact with the author(s) of this paper, please email the team at publications@city.ac.uk.
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].

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].

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) 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.

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

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

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.

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