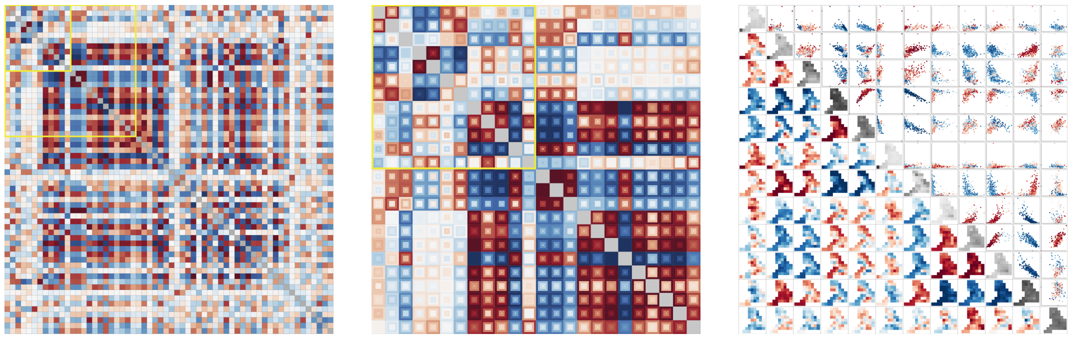
Visualizing Multiple Variables Across Scale and Geography

Sarah Goodwin, Jason Dykes, Aidan Slingsby and Cagatay Turkay



(a) Global Correlation Matrix (b) Scale Mosaic Matrix (c) Local Correlation Asymmetrical Matrix

Fig. 1: Multivariate comparison across scale and geography showing correlation through a red (positive) to blue (negative) diverging color scheme: *(a)* global correlation matrix; *(b)* scale mosaic matrix showing four levels of scale resolution (SR) within a global correlation matrix (for a subset of variables), *(c)* geographical and statistical views in an asymmetrical correlation matrix for a further subset reveals geographic variation in local correlation values (adaptive moving window – $N = 25$).

Abstract— Comparing multiple variables to select those that effectively characterize complex entities is important in a wide variety of domains – geodemographics for example. Identifying variables that correlate is a common practice to remove redundancy, but correlation varies across space, with scale and over time, and the frequently used global statistics hide potentially important differentiating local variation. For more comprehensive and robust insights into multivariate relations, these local correlations need to be assessed through various means of defining locality. We explore the geography of this issue, and use novel interactive visualization to identify interdependencies in multivariate data sets to support geographically informed multivariate analysis. We offer terminology for considering scale and locality, visual techniques for establishing the effects of scale on correlation and a theoretical framework through which variation in geographic correlation with scale and locality are addressed explicitly. Prototype software demonstrates how these contributions act together. These techniques enable multiple variables and their geographic characteristics to be considered concurrently as we extend visual parameter space analysis (vPSA) to the spatial domain. We find variable correlations to be sensitive to scale and geography to varying degrees in the context of energy-based geodemographics. This sensitivity depends upon the calculation of locality as well as the geographical and statistical structure of the variable.

Index Terms—Scale, Geography, Multivariate, Sensitivity Analysis, Variable Selection, Local Statistics, Geodemographics, Energy

1 Introduction

The various ways that geographical phenomena interrelate with location and scale [1, 3, 11, 26] are at the core of geographical analysis. Despite this, multivariate geographical phenomena are frequently studied using *global* summary statistics that do not take geography into account [13, 50]. Although these make results and analysis manageable, they hide important local variations in space, time, and scale.

We investigate the effects of varying scale in multivariate comparison, establish a complex parameter space for geographic analysis and present a new theoretical framework to manage these complexities. The framework allows multiple variables, and the relationships between them to be visualized concurrently in the contexts of geography and scale. Just as visual parameter space analysis (vPSA) [38] explores the effects of varying parameter values in a model parameter space, we explore the parameters of geography and scale. By applying existing vPSA techniques and terminology to this context, we establish the construct of geo-visual PSA (gvPSA).

• *Sarah Goodwin is with Monash University and the giCentre, City University London. E-mail: sarah.goodwin@monash.edu.*
• *Jason Dykes, Aidan Slingsby, Cagatay Turkay are with the giCentre, City University London. E-mail: {J.Dykes, A.Slingsby, Cagatay.Turkay.1}@city.ac.uk.*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx Aug. 2015; date of current version 25 Oct. 2015. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

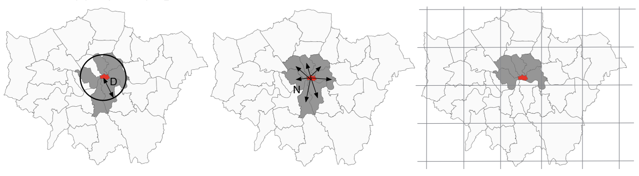
To demonstrate the applicability of the framework we apply it to the selection of variables for use in a geodemographic classifier focused on UK domestic energy consumption. Household energy is particularly relevant in this context as consumption is known to be highly geographical, but varies with the socio-economic characteristics of the population [8]. Our work with energy analysts identifies a need for better energy consumer profiling in the UK as well as the benefit of specifically designed visualization [15].

Appropriate variable selection is important in achieving valid and discriminating geographical profiles [49]. Standard practice is to avoid variables that may bias clustering results – such as those that strongly correlate or are heavily skewed. Variable selection is a well researched topic in the visualization and machine learning literature [17, 25, 29, 32, 39, 45]. However, geographical aspects of these variables and their interactions with each other are not usually considered and where variables interact not only with each other but also with underlying geography, an algorithmic approach is not always sufficient [33].

Supporting analysts in identifying patterns and making decisions using their knowledge of the domain and geography is essential [21]. Our approach uses interactive visual exploration to identify variables whose correlations vary geographically and those whose distributions

(a) Filtering spatial scale extent: Great Britain, England and London

(b) Aggregating spatial scale resolution: London at LSOA, LAU and NUTS2



(c) Locality definitions using '*City of London*' (red) as a starting location: *fixed* (by distance D) and *adaptive* (by N neighbors) *moving window* or regular *partitioning*

Fig. 2: Spatial scale extent, resolution and types of locality definition.

and correlations vary at different spatial scales. It complements a call for geodemographics to follow more specific, domain centred approaches utilizing the advances in geographically informed statistics, data exploration and visualization [40]. Our approaches use new visual designs that enable us to consider *where*, *how* and at *what scale* these variations occur concurrently so that geography can be used as a means of selecting discriminating variables.

We claim three contributions to information visualization:

i. a *theoretical framework* for visually comparing multivariate data across scale and geography;

ii. a *series of visual designs* for variable selection in this context;

iii. the consideration of geography as an input in visual parameter space analysis (vPSA) to establish *geo-visual PSA* (gvPSA).

Additionally, applying the framework to our work on energy consumption and socio-economic characteristics results in:

iv. *exploration* and *sensitivity analysis* of variables relating to energy usage relevant to the UK energy industry.

2 Geography and Scale in Multivariate Analysis

Standard summary statistics describe the distributions of variables and the relationships between them. Where phenomena have strong geographical variation such comparisons should take spatial variation and scale into account. The sensitivity of scale in analysis has been observed in general kernel density estimation tasks with visualization shown to be effective, for example in displaying curve features over a range of bandwidths [4]. We use visualization to explore two important aspects of scale in multivariate spatial analysis that can account for scale and geography in variables and their relationships: *Scale-dependent aggregation* and spatially-weighted *local statistics* [3, 11].

2.1 Scale-dependent Aggregation

We can study variables that are scale-dependent by varying the aggregation used in summary statistics. Scale can apply to attributes and time as well as space. In each context we can distinguish between *scale resolution* (SR) and *scale extent* (SE) [26, 46]. *Scale resolution* is the degree of precision used to define individual measurements and is determined either directly by the sampling intervals used or imposed by subsequent aggregations, e.g. aggregating age-group attributes, aggregating variable values into grid cells or aggregating time based data into days of the week. *Scale extent* refers to the scope of focus of analysis, e.g. the breadth of categorical information being aggregated, the geographical boundary, or the total length of a period of time.

Our focus is on *spatial* scale in areal based data with Fig. 2 illustrating relevant examples of spatial SR and SE. Spatial data can be aggregated in many different ways and are subject to the *modifiable areal unit problem* (MAUP) [37]. We argue that by exploring the effects of different methods and scales of aggregation, some of the effects of MAUP may be mitigated, helping us interpret data more reliably.

2.2 Local Statistics

Local subsets of data can be used to produce a multitude of spatially-weighted statistics. We use the term "*locality*" to define the spatial subset used. The resulting statistics depend on the scale (SR and SE) of the data as well as the *type* and *size* of the *locality definition*.

2.2.1 Locality Definitions

Based on established theoretical and applied literature [11, 18, 19, 20], three types of definition are identified. In Fig. 2c each *locality* (in grey) is based on different areas around the '*City of London*' area depending on the method used. These are: *Fixed moving window*: whereby the size of the defined locality is based on a fixed distance *D* – scale is consistent; *Adaptive moving window*: as above, but based on *N* nearest neighbors – ensures minimum sample size; *Partition*: local summary value for imposed geography – can be *regular*, usually grid squares (as in Fig. 2c), or *irregular*, usually administrative areas.

There are multiple ways of allocating variable values to localities to weight each statistic[1] adding to the complexity of the parameter space associated with this type of analysis.

2.2.2 Sensitivity of Locality Definition

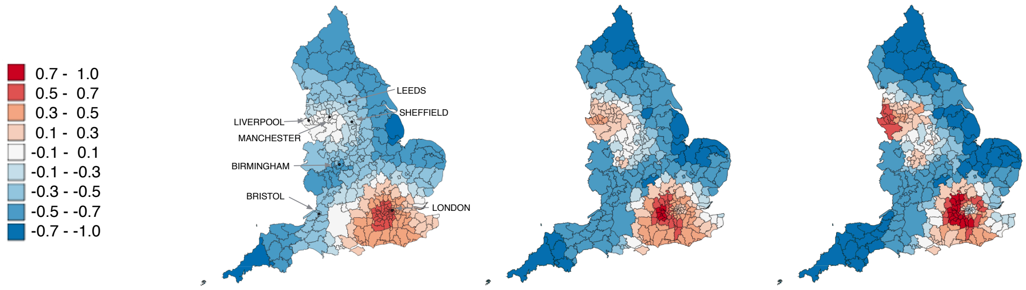
The values of *D*, *N* and the number of partitioning regions have a strong effect on the statistical outputs. Fig. 3 demonstrates the sensitivity of varying *N* in the distance weighted adaptive moving window approach. Here, a local correlation coefficient (Pearson's *r*) is calculated for 326 local authority units (LAU) in England to establish the relationships between '*gas consumption*' and '*electricity consumption*'. The global correlation coefficient is -0.32, yet this negative correlation is not consistent throughout the country. The maps of local statistics show increasingly positive correlations in more densely populated areas and increasingly negative correlations in more rural isolated locations. This strong spatial structure is to be expected given the lower levels of gas supply in remote areas of the country and to the apartment blocks that dominate residential living in inner London. It demonstrates that correlation is geographically and scale variant and that variation in particular phenomena are detectable at certain scales.

2.3 Comparing the Effects of Spatial Scale and Locality

The fact that local statistics vary with the locality resolution (N = 100, 50 or 25) is apparent in Fig. 3. Although these local correlation coefficients are calculated for data at LAU resolution, the source data are first aggregated from smaller geographic entities (see Section 4.3). Whilst these geographical units form part of an established hierarchy, the effects that aggregation have on such statistics are an important consideration. Varying the spatial scale (SR and/or SE) at which global and local statistics are calculated and the parameters used in locality definition results in a multitude of alternative outputs.

The visual comparison [47] of such outputs could contribute to the kind of multi-scale "*special view*" described by Lam and Quattrochi

[1] We allocated spatially-varying variable values by intersecting the population-weighted centroid of the small area it referred to with the locality extent. Distance weighted local statistics are produced using the R package *GWModel* [27] based on the adaptive and fixed moving window approaches.



(a) Legend (b) N = 100 (c) N = 50 (d) N = 25

Fig. 3: Local correlation coefficient of '*gas consumption*' compared to '*electricity consumption*' for the 326 LAUs in England using an adaptive moving window approach where *N* nearest neighbors is varied from 100 to 50 to 25. Consumption figures are *annual averages*.

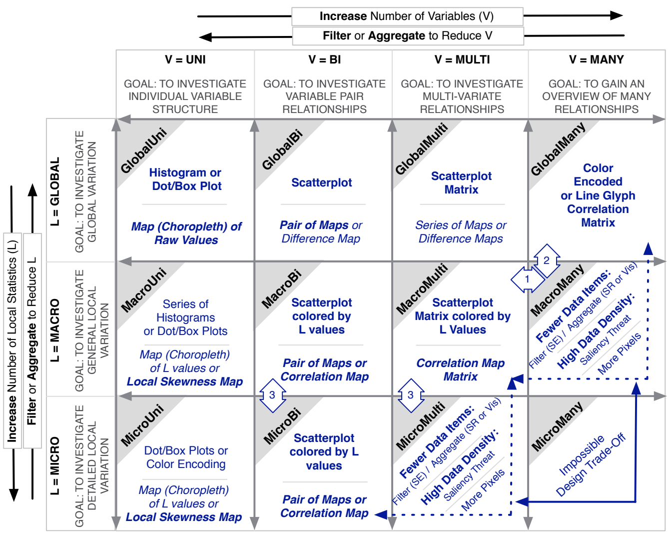


Fig. 4: The layout of the framework with cell names and goals (grey). Rows represent the number *Global, Macro, Micro* of local statistics (*L*) used to summarize *Uni* or compare *Bi, Multi, Many* of the variables (*V*) considered. Possible statistical and spatial (italicized) visual design options are shown in blue. Those in bold are demonstrated in the prototype. Arrows 1-3 identify discussed transitions.

[26]. Previous efforts include the use of interactive graphics to visually explore the effects of scale [9] and the results of geographically-weighted regression [7]. But managing the effects and parameters of comparison in which spatial scale and locality vary concurrently in workflows for multivariate geographical analysis requires a broader and more structured approach. The visual exploration of this complex *parameter space* can draw from existing work on vPSA, which aims to enable analysts faced with such spaces through visualization and interaction [38]. As such, we propose a framework for considering scale and geography in multivariate analysis with visual means and apply techniques from vPSA to investigate the effects of varying parameters associated with geography, scale and locality.

3 A Framework for Multivariate Visual Comparison Across Scale and Geography

Comparisons across scale and geography are increasingly challenging as the numbers of variables (*V*) and/or local summary statistics (*L*) increase. The framework is designed to structure and guide this process.

3.1 Framework Structure and Terminology

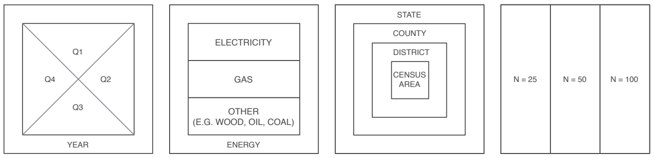
The structure of the framework is illustrated in Fig. 4 (in grey/black). Consider the top row: 'Global'. Here, the number of variables (*V*) under consideration increases along four columns divided into loosely defined bands: *Uni* for single variables – comparison is not an issue here; *Bi* where bi-variate comparison is important; *Multi* involving small numbers of variables, and; *Many* when large numbers of variables are involved. Subsequent rows describe the consideration of local variation with the number of local statistics (*L*) to be represented categorized into two equally loosely defined bands termed *Macro* and *Micro* – relating to larger and smaller SR and thus involving smaller and larger numbers of *L* respectively at any SE. *L* increases from the top of the figure to the bottom, where higher levels of spatial resolution usually result in more local geographical phenomena being identified. The downside of increasing *L* is greater numbers of summary statistics and the possibility of these failing to detect large scale phenomena.

We deliberately offer no numbers to define the thresholds between *Macro* and *Micro*, and *Multi* and *Many*. They are conceptual and adapt according to what is achievable with design in response to technology, data, task and user. Issues such as the number of pixels available and other characteristics of the device being used, the complexity of the data being shown and the intrinsic dimensionality of the phenomena, as well as the complexity of the task being addressed and knowledge, experience and needs of the analyst are defining factors here. The framework aims to draw attention to options and issues that might be considered when designing and analysing in these contexts. Each cell in Fig. 4 is named to help navigation, and the analytical goals for each are displayed. Moving from left to right increases the number of variables and potentially the information to display, with some visualization challenges likely as information quantities increase. Moving from right to left (through, for example, aggregation or filtering) has the benefit of reducing the visualization challenge at the cost of omitting information. Moving from top to bottom is likely to increase the number of measurements by refining resolution, from bottom to top has the opposite effect, usually making visualization and interpretation more straightforward at the cost of omitting important local variations.

In addition to varying *V* and *L* we can visualize the effect of scale by comparing multiple datasets in one view. Fig. 1a shows a *GlobalMany* view for one SR and Fig. 1b a *GlobalMulti* view for four SR. We describe multiple scales in our terminology with bracketed items, e.g. *GlobalMulti(SR4)*.

Where there are too many data items for the available screen space we visually aggregate and represent the aggregate with a statistical summary [10]. This can refer to filtering the size of the SE or through aggregation of the SR – whether by geographical area, a particular attribute or a period of time. Data items can also be reduced through larger partitioning in the locality calculation in place of the moving window option (Section 2.2.1). Alternatively the visual encoding can be aggregated on the fly for visualization purposes [10].

Where data reduction does not occur, particularly for the *MacroMany* and *MicroMulti* cases, discrimination will be challenging and occlusion is likely, hindering the ability to estimate values and make comparisons. A context in which more screen space is available may offer a means of representing all the data items concurrently, but this may be at the cost of ease of interpretation. Judicious visual design is important here. Smooth and fluid transitions between the cells and



(a) Temporal (SR5) (b) Attribute (SR4) (c) Spatial (SR4) (d) Locality (SR3)

Fig. 5: Scale mosaic designs use containment and layout to reflect relationships between scales – SRs in this context: e.g. seasons and yearly total (circular / hierarchical); types of and total household energy consumption (nominal / hierarchical); statistical boundaries (hierarchical) and ordered number of neighbors in locality (ordinal).

across the framework are likely to be important [23] in allowing analysts to manage the complexities and volumes of data.

3.2 Possible Visual Designs

The possibilities for visually encoding the parameters framed through this structure are vast. Viable encodings can draw upon existing idioms, and in places the framework suggests that more novelty is required. We investigate visual design options through our experience of dealing with geographic data (e.g. [41, 43, 46]), discussions with those using multivariate geographic data in their analysis [13, 15] and relevant literature on variable space observations (e.g. [45, 51]), local variations in multivariate comparison (e.g. [32]) and variable/feature selection (e.g. [25, 29, 39]). In Fig. 4 we present a collection of possible visual design options for each cell of the framework, chosen to encourage fluid movement between cells. We separate visual representations to emphasize the statistical and spatial (*italicized*) relationships in each case. Visualizing many variables at the *Micro* level (*MicroMany*) is deemed to be near impossible in many cases due to the volume of data – remembering that the thresholds between these conceptual states of affairs are defined to an extent by what is feasible in a given context – while three proposals are expressed for the *MacroMany* and *MicroMulti* situations. We investigate the use of correlation matrices for variable comparison, especially for the *M* cases – *Macro-Micro* / *Multi-Many*. Their compact nature often provides space into which alternative local statistics computed with different parameters can be visually represented in novel ways for comparison [14]. One novel example is the *scale mosaic*, designed for the comparison of global values for temporal-, attribute- or spatial-based SR or SE, or variations in locality definition. In Fig. 5 we split the correlation matrix cell using juxtaposition or containment [14]. Candidate designs reflect the variation in the data, with cyclical, linear and hierarchical arrangements shown. Frameworks for configuring hierarchical layouts may be usefully applied here [41]. Arranging cells in juxtaposition enables the number of variables that are under comparison to be increased from *GlobalBi* (e.g. Figs. 5, 8a) to *GlobalMulti* (Figs. 1b, 8b). As *V* increases to *GlobalMany* the limited visual space requires a single statistic be presented through color encoding. The variance, rank or the range of the values associated (e.g. Fig. 8c) are good candidates.

3.3 Geo-visual Parameter Space Analysis (gvPSA)

Given the fact that we treat the effects of geography, scale and locality as *parameters* within multivariate geographical data analysis, our framework can be characterized in terms of the established vPSA [38] model. vPSA is defined over three components: data flow model, navigation strategies, and analysis tasks. In terms of the *data flow model*, we introduce geographical location and scale as inputs to any multivariate analysis algorithm, with geodemographic classification the primary focus of our work (Section 4). A second component of the data flow model is the *derived outputs*. Our current focus is on variable selection, with correlation and skewness coefficients computed and visualized as *derived output*. Due to the focus on simulation in vPSA our overlap in terms of *navigation strategies* is small but the framework offers a structure to transition from *local-to-global* or *global-to-local*, which we demonstrate through a prototype (Section 5). Additionally, our approach considers an *informed trial and error process* – to support variable selection in our geodemographic context. Here we engage in some tasks described in vPSA. The first of these is *optimization*, where we support analysts in finding a suitable subset of the variables by taking geography and scale into account through well-informed, robust decisions in variable selection. The second relates to the *sensitivity* of the observations. We support this task by including visual designs that inform on the scale and location dependency of multivariate relationships. Considering geography, scale and locality as introduced in the previous section and the overlap with vPSA enables us to introduce our work as Geo-visual Parameter Space Analysis (gvPSA).

4 Applied Context: Energy Geodemographics

The framework is inspired by and demonstrated in the context of our ongoing work in variable selection for energy geodemographics [16].

Geodemographic approaches classify *areas* through clustering based on measured characteristics of the *characteristics of the resident population* [21]. Geodemographics are widely used for inferring characteristics about neighbourhoods, linking other relevant data, and making marketing and service delivery decisions. They are appropriate for energy consumer profiling in the UK as domestic energy consumption varies with geographic location [8] and is shown to correlate with many socio-economic characteristics of the population [8, 31]. Each area is allocated to the statistical cluster that best describes it. Membership can be mapped [42, 48] and used in modeling [21], with the uncertainties associated with the classification process explored through interactive graphics [43].

Generating a geodemographic classification is time-intensive and complex [21, 49]. The methodology used in the open-geodemographic classification developed by the UK Office of National Statistics (ONS) is well documented – the original Output Area Classification (OAC) used data from the 2001 UK Census [49] with amended variables and methodology for the second release for the 2011 Census [13].

4.1 Variable Selection: OAC Process

Suitable variables are selected for clustering from a pool of candidates, e.g. 167 variables were considered for OAC 2011 and reduced to 60 for the final classification [13]. Possible candidate variables for OAC 2001 and 2011 were initially identified based on a standard source SE and SR (Stage 1 of Fig. 6). Only variables representing all of the UK's SE were considered at the Output Areas (OA)[2] SR – hence 'OAC'.

Standardization (stage 2 of Fig. 6) ensures all variables are at the same scale for comparison and clustering [21]. Once standardized, global statistics – for example the correlation coefficients – are used to assess variable suitability for the classification [13, 50]. Variable selection decisions are made based on variable distribution and multivariate relationships as heavily skewed and strongly correlated variables can bias the classification results [21]. Multiple variables are compared to assess their correlation, skewness and similarities. Variables with little or no geographical variation also have little impact in producing geographically discriminating profiles [49]. As comparing the geographical variation of multivariate data is such a difficult task, the geographical variation is currently only explored as a final check when deciding whether to retain one variable over another [13, 49].

4.2 Variable Selection: Scale, Tasks & vPSA

The priorities and methods that guide domain specific and local classification differ depending on the use-case. The effects of scale and geography on the variables under consideration are important factors throughout the process as documented in the literature [40, 50] and illustrated in Section 2. Our reading of the literature (e.g. [13, 21, 33, 40, 49, 50]), experience of geographical analysis and initial explorations of the effects of geography and scale on multivariate correlation lead us to suggest: 1) *four stages* of the variable selection process where scale (SE and SR) decisions are made and may have an effect (Fig. 6); 2) *five analytical tasks* relating to correlation, geography and scale that

[2] UK statistical boundaries created for the release of the Census

can support geodemographic variable selection. Geodemographic analysts confirm a need to be able to determine *whether and where* [12]:

T1 *variables correlate or are heavily skewed*;
T2 *local correlation or skewness differ from global values*;
T3 *globally correlating variables show geographical differentiation*;
T4 *global correlation or skewness are sensitive to changes in SR*;
T5 *variables are effected by the determination of locality*.

Introducing local statistics[3] (as represented by the 'Locality' stage of Fig. 6) as *parameters* to the process enhances variable selection by exposing geographic variation and parameter sensitivities associated with the variables. In vPSA [38] terms, the variable values and summary statistics from the standardize and locality stages of our process form the output values for our multivariate (variable selection) analysis (stage 4 of Fig. 6). The different options available at both stages, relating both to the resolution and extent considerations, form the *parameter* space that, as in vPSA, affects the results of the analysis. We use our framework to guide the design of graphics for our software prototype, which shows correlations between multiple variables and their variation across scale and geography. Pathways through the framework guide the interactions through which the variable selection process is graphically augmented. We investigate the benefit of these methods informally in light of our work in multivariate geographical analysis through visual exploration.

4.3 Demonstrative Dataset

To demonstrate the process we use the 71 unique variables from the 2011 UK Census [34] that generate discriminating profiles at the national level in OAC (41 were used in OAC 2001 [49] and 60 in OAC 2011 [35]). We augment these for energy geodemographics with a further 7 energy-related variables. Gas consumption, electricity consumption and fuel poverty levels come from the UK Department of Energy and Climate Change (DECC) [5, 6] and the 2011 UK Census [34] provides the further 4 variables on usage of the different central heating types: electric, gas, other (e.g. wood, coal or oil) or none.

In terms of scale, each variable from the Census is available at an OA SR, whilst DECC variables are available at LSOA 2001[4]. Both of these Census units are not only relatively small in area but are specifically designed to produce an optimal arrangement in terms of social homogeneity, thereby reducing the impact of MAUP [28]. In terms of SE, the Census 2011 and consumption data are available for England and Wales[5] and the fuel poverty indicator for England only. This leads us to investigate multiple levels of spatial SR for the dataset of 78 variables, with SE fixed and focusing on England.

The source data scales are illustrated in Fig. 7, which uses the model shown in Fig. 6. Here, we show that values for the LSOA 2001 data are first disaggregated to OA 2011 units (171,372) for England, before being aggregated to three common levels of SR (see Fig. 7). These levels are: LSOA 2011 (32,844 units), LAU (326) and NUTS2 (32) from the *Nomenclature of Territorial Units for Statistics* [36].

In terms of localities, local summaries were calculated for NUTS2 and LAU using the adaptive moving window approach for three differing values of N (see Fig. 7). LSOA and OA were too numerous to perform the calculations given the computing resources available. The output scale (Fig. 7) that is available for representation in the final prototype provides four data sets – forming a geographical hierarchy. Each contains data for the 78 variables and are summarized as global statistics allowing the four SR to be compared globally, while NUTS2 and LAU offer additional local statistics and the opportunity to investigate the sensitivity of locality, through varying the value of N.

4.4 Geodemographics and the Framework

In reference to our framework, current practice for geodemographic variable selection [13, 21, 49] takes place in the first row *Global*, where variable structure and relationships are compared from V = *Uni* through to V = *Many* – 167 variables in the case of OAC 2011 [13]. The introduction of locality into the process (Figs. 6 and 7) allows us to augment this analysis by exploring the *Macro* and *Micro* rows through which we can investigate geographical variation. From our demonstrative data set (Section 4.3) the four levels of SR provide the ability to vary scale at the *Global* level of the framework, through which the geographic variation of a single variable is considered. Additionally the local statistics calculated at two SRs (with 32 and 326 values) can be associated with the *Macro* and *Micro* levels of the framework. Graphical techniques for producing data dense relational graphics [44] associated with the combinations of V and L across the framework are required to enable us to explore the effects of varying both scale and geography in our analysis.

[3]We use Pearson's r Correlation Coefficient and Skewness because spatially-weighted versions of these exists in the *GWModel* [27] R package.

[4]Second tier census boundaries aggregated from OA

[5]The Census 2011 has since been released for the whole of the UK

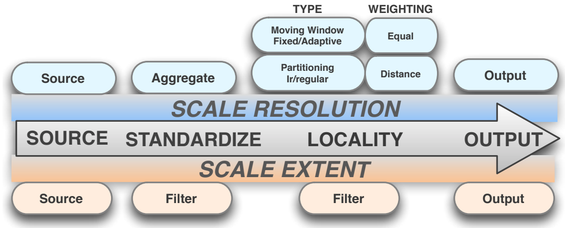


Fig. 6: Adding locality into the variable preparation process for selection for geodemographic classification. Each stage – source, standardize, locality and output – involves both dimensions of SR and SE.

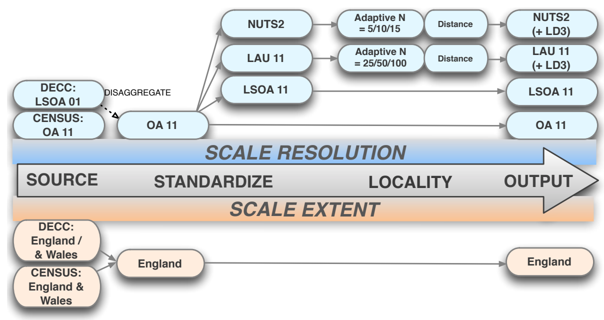


Fig. 7: The data scale resolution and locality definition (LD) options associated with the demonstrative dataset.

5 Prototype Visual Design

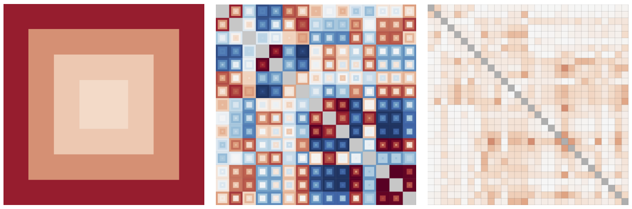
We instantiated the framework in prototype software [24] by implementing the visual design options shown in bold in Fig. 4. These are described here and in the accompanying video[6].

5.1 Prototype Layout

The prototype features three panels: overview, comparison and detail (see Fig. 9). The *overview* panel allows all 78 variables to be ranked by four global measures describing complimentary characteristics: theme (variable category), skewness (as an indication of distribution), variance of correlation (as an indication of how correlation varies across all variables) and Moran's I (to establish geographical dependencies [2]). The *comparison* panel is an adaptable correlation matrix (Fig. 1a-1c) suitable for *Multi*-to-*Many* variables. Rows and columns represent variables ordered according to the overview panel. Each cell thus represents a pair of variables other than those on the diagonal, which relate to a single variable.

Cells in the matrix show pairwise correlation statistically and geographically (Section 5.2) through scatterplots and maps colored by global or local correlation. This view adapts to color encoded cells or line glyphs (representing the scatterplot shape as a diagonal line) as the number of variables increase from *Multi* to *Many*. The central diagonal

[6]http://vimeo.com/123730484



(a) GlobalBi (b) GlobalMulti (c) GlobalMany

Fig. 8: Spatial scale mosaic with superimposed color encoded global correlation coefficients for SR4: *(a) V = Bi* and *(a) V = Multi*. *(c)* For *V = Many* the variance of the four correlation values is encoded – the darker the more variation across SR for the variable pair.

shows the distribution of the variable in the row/column through interchangeable views depending on the size of the matrix: histograms, variable distribution maps, local skewness maps or cell color encoded by global skewness from positive (purple) to negative (green). We use diverging and sequential ColorBrewer schemes [22] consistently to show different forms of variation: e.g. *RdBu* for correlation (+1 red to -1 blue), *PrGn* for skewness and *YlOrBr* for geographic distribution. Alternative ColorBrewer schemes are used for the global measures in the overview panel.

The *detail* panel represents the *Bi* and *Uni* columns of the framework, where the structure and pairwise correlation of any pair of variables selected from the other panels are displayed (see Example 4 in Fig. 9). This consists of enlarged and enhanced views from the correlation matrix including: maps showing the geographical distribution of the individual variable or the local skewness, and pairwise local correlation, as well as a scatterplot presenting the local correlation.

In using different screen real estate to show relationships between different numbers of variables (*V*) and local statistics (*L*) the three panels populate distinct cells in the framework under contrasting constraints.

5.2 Concurrent Geographical and Statistical Views

To represent the local values we use an *asymmetrical matrix* in the comparison panel. Concurrent and complimentary views of local variation draw upon the statistical and geographical design spaces above and below the diagonal respectively (Fig. 1c, Fig. 9). For these views in *MacroMany* and *MicroMulti* we use different reduction techniques. For the maps in the matrix we aggregate the data spatially on the fly and present the average value as a colored square. We chose not to aggregate the statistical space but allow linked views in the detail panel to enable the aggregated geographical space to be explored in the non-aggregated statistical space. As our main focus is with LAU and NUTS2 we can present the data for *Multi* variables in this way. Some of the local relationships evident in the scatterplots require resolution to establish the good continuation through which trends can be detected (e.g. Fig. 9, scatterplot 1). As we move further into the *Micro* space (e.g. the use of LSOA in this example) statistical aggregation is likely to be necessary.

5.3 Spatial Scale Mosaics

We use our scale mosaic design to present multiple spatial scale (SR4) in our comparison and detail panels for *GlobalBi* (Fig. 8a) and *GlobalMulti* (Fig. 8b). Global correlation is shown in the matrix with global skewness filling the central diagonal. As our spatial SR is hierarchical we use containment to reflect the scale relationships where space allows. As we move to *GlobalMany* the number of pixels available within each cell decreases and the ability to visually detect the differences between the scales reduces. We therefore revert to a single global summary (Fig. 8c). We use variance here as we are predominantly interested in discovering whether a variable is scale dependent and by how much (alternatives in Section 6.2.2). Interaction between the three panels of the prototype allows for the detailed investigation of the variation associated with the four SR in our data set. Clear effects are identifiable for many variables, as discussed in Section 6.2.2 and shown in the accompanying video.

5.4 Navigating the Framework

An important aspect of the framework is that the visual representations adapt as we shift between the framework cells. In developing our prototype we have implemented transitions in which both *L* and *V* vary, paying particular attention to transitions between some of the more challenging cells: in *L = Macro* from *Multi-Many*; in *V = Many* from *Macro-Global*; in *V = Bi* and *Multi* from *Macro-Micro*. These transitions are shown by arrows 1-3 in Fig. 4. The transition from *Macro-Multi* to *MacroMany* (Transition 1 in Fig. 4) occurs through spatial aggregation and over-plotting, which gradually reduces the amount of local detail shown on the screen. When the screen space becomes too limited to show *Macro*, the visual representation shifts from *Macro-Many* to *GlobalMany* (Transition 2) and the cell is color encoded. Transition 3 from *Micro-Macro* is shown in both the *Bi* (detail panel) and *Multi* (comparison panel) views by varying the number of cells used to create the maps. These interactive and dynamic features showing framework in action are demonstrated in the supplementary video.

6 Variable Exploration and Sensitivity Analysis

The prototoype provides us with the capability to compare, select and filter variables, reorder the matrix according to key (global) features, automatically highlight strongly correlated or skewed variables, and focus on the details of selected variable pairs through linked larger visuals provided in the detail panel. This enables us to explore geographical variations in energy and OAC variables (***T1***-***T3*** from Section 4.1) as well as their sensitivities to scale and geography (***T4***-***T5***). The sensitivities associated with geography, scale and locality can be explored by varying the SR and the number of *N* in the locality calculation. The alternative local summary statistics that result can be depicted visually – perhaps as scale mosaics (see Fig. 5d) within maps or as graphics in juxtaposition as used in our exploration (see Fig. 10).

Values of *N*, *L*, *V* and level of SR can be varied interactively in our prototype with visual depictions updating appropriately. The rapid interactive filtering, reordering and parameter variation supported by the prototype allow us to engage in exploration through *informed trial and error* that supports gvPSA.

6.1 Energy Variable Exploration

For example, visual exploration of global and local correlation and skewness reveals the energy variables to be highly geographical and sensitive to changes in SR (***T3***, ***T5***). As the SR increases from fine (OA) to coarser resolution (NUTS2) the magnitude of global correlation increases, with extreme values more evident at the larger SRs. We explore the benefit of adding locality to our interpretation of correlation by discussing the local variation evident in the energy and other related variables through use of the prototype.

We consider a *MacroMulti* case in Fig. 9, with local correlation for the data at LAU SR for each of the seven energy variables. Four numbered variable pairs of particular interest. Example 1 relates '*electricity consumption*' and '*gas consumption*', showing both strongly positive and strongly negative local correlations (as discussed in Section 2.2.2 and Fig. 3). Scatterplot 1 suggests two relationships amongst LAUs, with map 1 showing the positive correlation to be characteristic of London and the North West and negative correlations elsewhere. Example 2 ('*electricity consumption*' and '*% in fuel poverty*') has a near zero (0.03) global correlation coefficient, yet the local correlation varies substantially across England. Strong positive and negative local correlations occur (e.g. >0.8 in the far South West, <-0.5 in the North West). Alternatively Examples 3 and 4 show variable pairs with very strong positive global correlation (***T1***), with the local correlation map and scatterplots showing strength of positive correlation to vary locally (***T2***). Both Example 3 ('*gas consumption*' and '*gas central heating*') and Example 4 ('*electricity consumption*' and '*other central heating*') are less strongly correlated in the more densely populated North West and London. Revealing the geography of these relationships may be

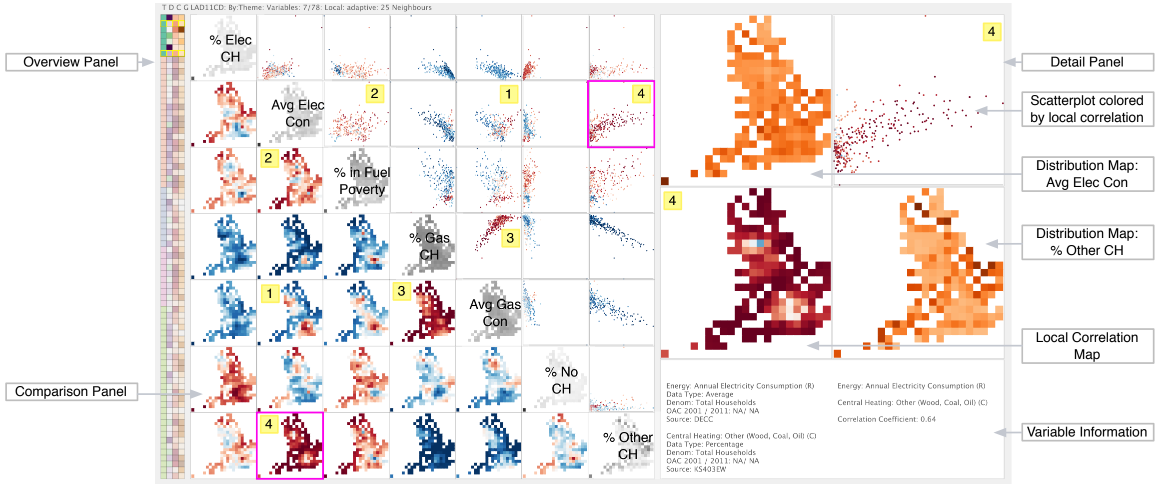


Fig. 9: Software prototype showing all three panels and *MicroMulti* information in an asymmetrical matrix. Seven energy – consumption (Con), fuel poverty and central heating (CH) – variables displaying locality information, based on an adaptive moving window with 25 neighbors (N=25). Pairwise correlation examples 1-4 are highlighted and discussed in the text.

helpful in identifying key variables that differentiate populations and behaviors despite global correlation. For instance, in Example 4 '*electricity consumption*' and '*other central heating*' correlate at the national scale (*Global*) and so one variable could be considered redundant if geography is not considered. Our maps however show that this pair of variables allows us to discriminate between characteristics in the North West and South East that would not be captured if one of the variables were omitted from analysis (***T3***).

When expanding the comparison to the non-energy variables and skewness, local differences are also evident (***T2***). This is particularly the case with gas consumption and central heating as the availability of gas is so geographically variant in the UK. Variables that are heavily skewed at the global level are often only extremely skewed in certain locations (***T3***). For example '*electric central heating*' has a global skewness value of 4.3 yet the local skewness map shows extreme values to be mainly located around London with some rural areas having a negative skew (***T2***, ***T3***). As the resolution of *locality* is enlarged (by increasing N) the local patterns disappear and the skewness of London dominates the global statistic. Local skewness evident in other variables frequently reveals a different skewness in densely populated urban areas such as London and the North West of England than the rural areas particularly where N is low. This effect reduces as the number of neighbors (N) increases and the values converge to the mean of the full geographic extent (SE). The prototype shows where these differences occur in multiple variables concurrently (***T1***, ***T2***, ***T3***). This information can help to identify which variables differentiate which localities (***T3***) and which variables are affected by SR (***T4***) as well as changes in SE. Whilst SE is not accommodated in the software prototype we can make some inferences – examples with strong correlation or heavy skewness in London for instance may be useful for a nationwide profile but might not be suitable for a London (SE) based profile.

6.2 Geo-Visual Parameter Sensitivity Analysis

Having explored some of the ways in which correlations vary with scale and geography we next use the prototype to explore the sensitivities associated with definitions of locality when varying the number of neighbors (N) (***T5***) and when changing the SR (***T4***) at which global statistics are calculated.

6.2.1 Varying Neighbourhood Parameterization (N)

In order to see which variables are most affected by varying the value of N (25, 50 and 100), we subdivide the cells of the comparison panel and display the resulting graphics for each value of N in juxtaposition. A small section of the matrix that results is shown in Fig. 10, with vertical juxtaposition in this instance. The full figure with more variables and showing local skewness is supplied as supplementary material.

In Fig. 10 geographical variation in local correlation between household variables and electricity or gas consumption is shown. In all examples we can see differences, particularly in the map view, as the locality definition varies by N (***T5***). These differences depend on the type of correlation pattern. When displayed in this manner and interpreted with a little local knowledge we can detect three types of geographical correlation pattern (***T3***): those that have largely urban/rural correlations (e.g. '*gas consumption*' and '*home owned*'), those revealing little difference across the country (e.g. '*electricity consumption*' and '*home owned*') and those with a distinctive North-South divide (e.g. '*electricity consumption*' and '*semi-detached housing*' or '*private-rented housing*'). In Fig. 10 the correlation with gas largely shows an urban/rural trend; however, the other patterns are present in the full supplementary figure – for example '*gas*' and '*no central heating*' shows a largely North-South divide, perhaps reflecting cultural differences that may be important in energy geodemographics (***T3***).

In Fig. 10 there is also a clear difference between patterns of gas and of electricity correlation when comparing local and global values (***T2***, ***T3***). Globally, gas is shown to not correlate with the other variables in Fig. 10 as the values are near zero– the shapes of the scatterplots and the values reveal this. However, all examples show clear geographical differences with most returning both highly negative and positive correlations locally (***T2***). Visualizing data in this manner enables us to relate local knowledge to the scale and locality dependent information and should result in more informed decision-making. Global correlations with '*electricity consumption*' are more varied; The two variables with the weakest global correlations show clear and different local patterns ('*rent private*' and '*semi-detached*'), whilst '*detached*' has a very strong global correlation and shows little variation locally (***T3***). Expanding this analysis to other variables (see supplementary material) enables us to visually demonstrate that local statistical summaries are not only dependant on the value of N but also heavily influenced by the structure of the variables themselves – in terms of both their statistical and geographical distribution. Variables with distinct geographical patterns, such as gas consumption, show greater variation in their correlations when explored at the local level but these vary depending on the comparator. Heavily skewed variables also have more local skewness variability than those that are more normally distributed.

These observations lead to a more robust understanding of the energy variables through the investigation of correlation relations. Further development is required to extend this analysis of varying N to the comparison of locality definitions (Section 2.2.1). Whilst our implementation does not enable this at present it demonstrates ways in which structured juxtaposed graphics can be used to explore sensitivi-

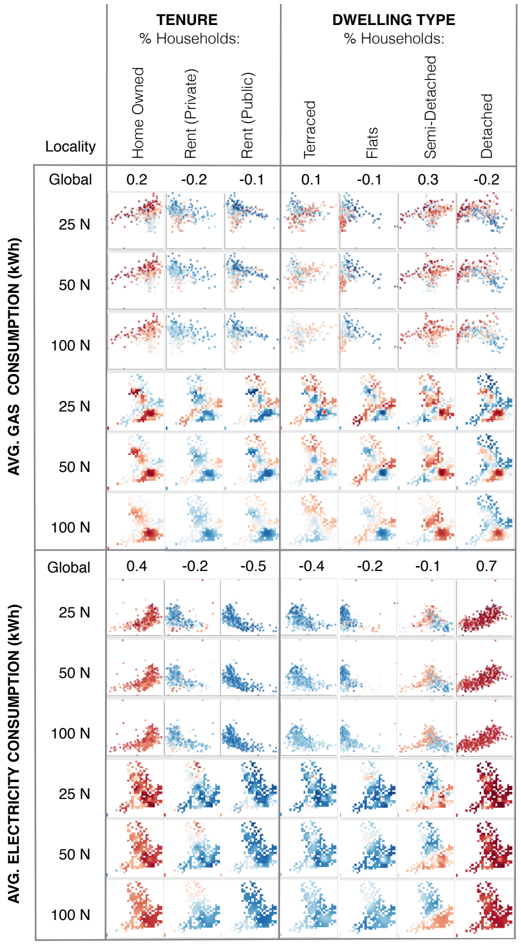


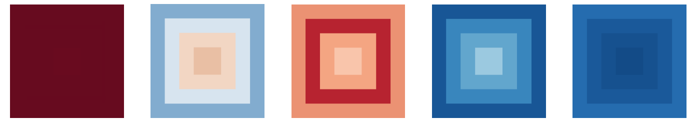
Fig. 10: A selection of household related variables in comparison with gas and electricity consumption when number of neighbors (N) in the locality is varied (25, 50 and 100) for LAU. Dark red shows positive correlation and dark blue represents negative correlation in statistical (top) and geographic (bottom) views for each variable pair.

ties associated with the definition of locality across multiple variables. The framework identifies and relates parameters for gvPSA and views through which gvPSA tasks can be accomplished.

6.2.2 Varying Scale Resolution (SR)

Visual exploration of the effects of varying scale (***T5***) draws upon the scale mosaic view in the prototype. We use it to represent global correlation at a number of scales. Five forms of sensitivity are identified in our multivariate data set, as shown in Fig. 11. Of the 78 variables in the prototype, the majority *strengthen* with scale with the global correlation getting stronger as the SR increases from fine resolution (e.g. OA) to coarse (e.g. NUTS2). Pairs of variables that are associated with regional trends, and local variability, are likely to exhibit these characteristics. Examples include Fig. 11d, which shows how the correlation between '*gas central heating*' and '*electricity consumption*' strengthens in a negative sense from -0.5 to -0.8 as the resolution increases. *Constant* shows variables in which correlation is scale invariant, an unusual pattern in our analysis of OAC variables that are either very similar or the direct inverse of each other. For example, Fig. 11a, shows '*Aged 65+*' from OAC 2001 with '*Aged 65-89*' from OAC 2011[7]. That the strong correlation is independent of scale suggests limited discriminatory value here, unless geographic (regional) differences are encountered. *Polarity* relates to change in the sign of the correlation value, e.g. Fig. 11b shows a positive correlation between

[7]Further discrimination of age group variables were added to OAC 2011.



(a) Constant (b) Polarity (c) Fluctuates (d) Strengthens (e) Weakens

Fig. 11: Five forms of scale sensitivity as SR increases from fine resolution to generalized using the scale mosaic design. Color refers to the degree of positive (red) or negative (blue) correlation.

'*flats*' and '*separated/divorced*' at OA, but a negative correlation at NUTS2. *Weakens* occurs rarely in our analysis, but Fig. 11e shows an example where '*average house size*' and '*single status*' has a correlation value of -0.8 at OA level and this decreases to -0.7 at NUTS2. *Fluctuates* is particularly sensitive to increases in scale and levels of correlation both *strengthen* and *weaken* as scale changes. Fig. 11c shows '*electricity consumption*' with '*people who bicycle or walk to work*', with correlation at LAU notably higher (0.7) than both NUTS2 (0.4) and LSOA (0.4).

In our prototype design we use variance to encode the *GlobalMany (SR4)* view. Whilst it provides a useful metric of the dependency each variable pair has on scale, this single number summary is not sufficient to identify all of five types of scale sensitivity in Fig.11. For example, it is not possible to distinguish weakening and strengthening trends. Alternative global measures depending on the analysts' interests and foci, may be appropriate for the *GlobalMany* view. For example, rank correlation or the maximal information coefficient may be more appropriate for identifying monotonically increasing or decreasing relations. They are applicable summaries within our broad framework.

In general, the mosaics reveal that correlations at the SRs of OA and LSOA are less strong than those at the two coarser SRs of LAU and NUTS2. The parameters are likely to be sensitive to processes that operate at different scales – with differences within and between regions depending upon the scales and localities used, thus justifying the consideration of resolution and extent as parameters in gvPSA.

7 Discussion

We provide a framework with candidate designs, interactions and guidance for comparing and making sense of geographic aspects of multiple variables. We reflect on this framework, prototype designs and the applied context of geodemographic variable selection.

7.1 Applied Context

Using variables with geographical variation helps produce more discriminating profiles when creating geodemographic classifiers [49]. The addition of the locality stage (see Fig. 6) in the variable selection process helps determine geographic variability. Whilst this increases the complexity of the process, our framework offers guidelines to allow many variables to be visualized concurrently with the consideration of local variation. The prototype shows how this might be achieved – exploration that unearths interesting patterns and fundamental differences when visualizing local, rather than global statistics. The fluid transition from *Global* to *Macro* to *Micro* and the three panel prototype layout design allow strongly correlated variables to be discovered at the global level and then investigated through exploration of their geographical variation at the local level.

The asymmetrical correlation matrix allows local statistics to be visualized in spatial and statistical spaces concurrently – supporting comparison. Local patterns of similarity (or dissimilarity) can be identified and data points that are statistically similar compared – for example, different locations in densely populated areas with similar correlations (Fig. 9). The linked views and different data reduction techniques for the two views allow such patterns to be discovered. Partitioning correlation coefficient cells, as in the scale mosaic design, offers a new way of visualizing the effects of scale, allowing multiple variables across four scales to be represented in a single view. We do not order variables in the matrix by scale effect, but could do so.

We continue to develop our design suggestions as we apply our framework to other application areas. We are currently working to improve the quality of future survey data by identifying geographic and demographic patterns in response rates. This involves a large number of census and related variables to investigate *model rates of response to household surveys*[8]. In this context both time and attribute-based SR and SE are important in addition to spatial scale. The appropriate and informed selection of a limited set of discriminating variables is an important process in this analysis and one that can be supported by the framework, techniques and prototype software.

7.2 Feedback

Demonstrating the prototype to the creator of OAC 2011 [12] resulted in a positive reaction to the framework and its implementation. Although the prototype design itself was considered too complex for current practice due to the sheer amount of new information being presented, the approach was seen as relevant and visual designs regarded as being potentially useful.

The scale mosaic and local correlation matrices (Figs. 1b and 1c) were considered as a big improvement over current practices, though some may have been a little too data dense. We acknowledge that considering local statistics and the parameters of geography and scale increases the amount of information to which one must attend whilst working with multiple variables. However, our framework accommodates this with reductions in V and/or L and associated graphical updates as required enabling users to manage the information being presented. These could be permanent or temporary, whilst the views and their meanings are learned. We cannot provide evidence that this learning will occur, but we can accommodate richer graphics that contain information deemed to be important should the need arise.

Even at the global level, the prototype interaction and reordering possibilities were seen as more useful than the static matrices used to select variables for OAC 2011. The scale mosaics and our comparison of four SR were considered useful for local or domain specific geodemographics, where particular data are likely to be available at different SR and SE. In reality, value of the framework be determined by the detail of its instantiation. This is highly likely to be dependent on characteristics of data, task, technology and user. It provides a structure through which use-case specific workflows may be generated to help users deal with large amounts of information. It is sufficiently flexible to accommodate many of these in a range of contexts.

7.3 Limitations

A number of limitations from a user-perspective were identified in Section 7.2. Here, we reflect on some other limitations.

Our *framework* is general, partially populated and contains only broad design guidelines. *Specific designs* are untested and only evaluated internally though our ongoing work in this domain. We employed a correlation matrix as the predominant layout for comparison (e.g. in the comparison panel of the prototype), but whilst further design work would be needed to adapt these techniques to multiple scales, more space-efficient visualization techniques [30] could be used.

The nature of local statistics means that some summaries are based upon small samples, perhaps leading to instability. Some patterns detected (or missed) may be dependent on artefacts of the graphics – orderings of matrices, layering of scatterplots and visual aggregation (as in the case of the maps in the prototype). We have not accounted for these effects in the current designs.

Whilst the *prototype* enables us to efficiently explore sensitivities in SR and locality definition (***T1–T5***) it does not implement the framework fully. For example, our consideration of the problem of alternative locality definitions is incomplete, involving the results of just one approach. Local statistics were pre-calculated for NUTS2 and LAU only with small numbers of N. Statistics were not calculated at the higher resolution LSOA and OA levels due to limited resources. It elicited feedback and demonstrated potential, but more work is needed to further study the effects of scale on multivariate correlation for more SRs, different SEs and different locality definitions; e.g. by combining the mosaic scale design within the map squares. As such, we are only able to address certain aspects of the vPSA model and apply them to geodemographics. We will need to make calculations more efficient to demonstrate the full effects of varying the parameters of the geodemographic classification model in real time so that gvPSA can be undertaken on the modeling as well as on the variable selection task through *informed trial and error*.

These issues can be addressed through further design, development and experimentation. Initial findings suggest that such work will be worthwhile. Whilst we do not yet have usable software to make the visualization of multiple variables across scale and geography manageable in the manner that is our aim the prototype has enabled us to develop and evaluate approaches, identify effects and establish needs.

8 Conclusion and Further Work

Our framework for visualizing multivariate data across scale and geography is the fundamental contribution of our research. It allows some of the sensitivities and complexities of multivariate geographic data to be investigated through visual means and provides a basis for structuring this activity. By instantiating key aspects of the framework in software, we offer design proposals and visualization techniques to support such work. In doing so we have applied known statistics, geographically weighted statistics, new *visual designs* (scale mosaics; asymmetrical correlation matrices) that show correlation, scale and geography, and interactions that established a new construct – gvPSA (geographical vPSA). The approach has supported tasks that we deem important in multivariate geographical analysis and enabled us to identify notable local variations in geography and scale that were hidden by global statistics in our own geographic work. We demonstrate that these approaches have clear application through analytical exploration (Section 6) in which geographical variation and scale are shown to be important considerations in multivariate correlation.

Although we argue that these approaches are beneficial to the geodemographic variable selection process and have revealed variables that are variously geographically correlated, further work is needed to determine the exact effects that variables with a '*weak global and strong local*' or a '*weak local and strong global*' correlation may have on the final classification. Further work, where the *data flow model* [38] is realized in its entirety, with geodemographic classifications as the *output* and classifiers as the *model* could demonstrate how the variable decisions informed by geography and scale affect the classifier, increasing the need for effective gvPSA *navigation strategies*.

We would like to see the framework populated with further effective encodings that achieve a usable balance between information density and pragmatism. We hope to develop pathways through it to support sophisticated multivariate geographical analysis. But the true value of the ideas implemented in the prototype will only be established through the addition of robust and effective functionality to support navigation, selection and sense-making in applied contexts. A system that applies the framework fully would allow analysts to record and select the various geographies that are distinguished through local analysis and their relationships with correlating variables. Whilst feedback received from colleagues developing geodemographic classifiers showed the potential of the framework and prototype it also demonstrated the challenges associated with presenting the rich information associated with gvPSA and thus of implementing usable systems for this activity. Iterative user-centred approaches would be key in designing a system that built on the contributions presented here to allow interactive variable selection and geodemographic classification through which the effects of scale and geography on clustering outputs could be assessed in real time in their geographic contexts.

Acknowledgments

This work was partially funded by a 'Vice Chancellor's Scholarship' from City University London. Thanks to Tamara Munzner, IEEE VIS 2014 participants and members of the giCentre and g2Lab (HCU Hamburg) for their feedback and support. We also thank four anonymous reviewers whose comments helped improve the manuscript.

[8] https://blogs.city.ac.uk/addresponse/

REFERENCES

[1] G. Andrienko, N. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. I. Fabrikant, M. Jern, M.-J. Kraak, H. Schumann, and C. Tominski. Space, time and visual analytics. *IJGIS*, 24(10):1577–1600, 2010.

[2] L. Anselin, I. Syabri, and O. Smirnov. Visualizing multivariate spatial correlation with dynamically linked windows. In *University of California, Santa Barbara. CD-ROM*, 2002.

[3] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4):281–298, 1996.

[4] P. Chaudhuri and J. S. Marron. Sizer for exploration of structures in curves. *JASA*, 94(447):807–823, 1999.

[5] DECC. Fuel poverty: a framework for future action. Technical report, Dept. Energy & Climate Change, London, July 2013. http://bit.ly/fuelPoverty.

[6] DECC. MSOA/IGZ and LSOA gas and electricity statistics: Methodology and Guidance. Technical report, Dept. Energy & Climate Change, Mar 2013. http://bit.ly/DECCEnergy.

[7] U. Demšar, A. S. Fotheringham, and M. Charlton. Exploring the spatio-temporal dynamics of geographical processes with geographically weighted regression and geovisual analytics. *Information Visualization*, 7(3-4):181–197, 2008.

[8] A. Druckman and T. Jackson. Household energy consumption in the UK: A highly geographically and socio-economically disaggregated model. *Energy Policy*, 36:3177–3192, Aug. 2008.

[9] J. Dykes and C. Brunsdon. Geographically weighted visualization: interactive graphics for scale-varying exploratory analysis. *IEEE TVCG*, 13(6):1161–1168, 2007.

[10] N. Elmqvist and J.-D. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE TVCG*, 16(3):439–454, 2010.

[11] A. S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, Chichester, England; Hoboken, NJ, USA, 1st edition, Oct. 2002.

[12] C. G. Gale. Personal Communication, 2014.

[13] C. G. Gale. *Creating an Open Geodemographic Classification using the UK Census of the Population*. PhD thesis, UCL (University College London), May 2014.

[14] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, Oct. 2011.

[15] S. Goodwin, J. Dykes, S. Jones, I. Dillingham, G. Dove, A. Duffy, A. Kachkaev, A. Slingsby, and J. Wood. Creative user-centered visualization design for energy analysts and modelers. *IEEE TVCG*, 19:2516–2525, 2013.

[16] S. Goodwin, J. Dykes, and A. Slingsby. Visualizing the effects of scale and geography in multivariate comparison. In *Conference on VAST (Poster)*, Paris, France, Nov. 2014.

[17] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[18] P. Harris, C. Brunsdon, and M. Charlton. Geographically weighted principal components analysis. *IJGIS*, 25(10):1717–1736, 2011.

[19] P. Harris, C. Brunsdon, M. Charlton, S. Juggins, and A. Clarke. Multivariate spatial outlier detection using robust geographically weighted methods. *Mathematical Geosciences*, 46(1):1–31, Jan. 2014.

[20] R. Harris, A. Singleton, D. Grose, C. Brunsdon, and P. Longley. Grid-enabling geographically weighted regression: A case study of participation in higher education in England. *Trans. in GIS*, 14(1):43–61, 2010.

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

[22] M. Harrower and C. A. Brewer. ColorBrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, June 2003.

[23] J. Heer and G. G. Robertson. Animated transitions in statistical data graphics. *IEEE TVCG*, 13(6):1240–1247, 2007.

[24] A. Hevner and S. Ram. Design science in information systems research. *MIS Quarterly*, 28(1):75–105, Mar. 2004.

[25] J. Krause, A. Perer, and E. Bertini. INFUSE: Interactive feature selection for predictive modeling of high dimensional data. *IEEE TVCG*, 20(12):1614–1623, Dec. 2014.

[26] N. S.-N. Lam 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.

[27] B. Lu, P. Harris, M. Charlton, C. Brunsdon, T. Nakaya, and I. Gollini. Package GWmodel, 2014. http://bit.ly/GWModel.

[28] D. Martin. Towards the geographies of the 2001 UK census of population. *Trans. of the Institute of British Geographers*, 25(3):321–332, Jan. 2000.

[29] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guiding feature subset selection with an interactive visualization. In *IEEE Conference on VAST*, pages 111–120, Oct. 2011.

[30] S. McKenna, M. Meyer, C. Gregg, and S. Gerber. s-corrplot: An interactive scatterplot for exploring correlation. *Journal of Computational and Graphical Statistics*, 2015. DOI: 10.1080/10618600.2015.1021926.

[31] F. McLoughlin, A. Duffy, and M. Conlon. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. *Energy and Buildings*, 48:240–248, 2012.

[32] T. Muhlbacher and H. Piringer. A partition-based framework for building and validating regression models. *IEEE TVCG*, 19(12):1962–1971, 2013.

[33] S. Murphy and M. Smith. Geodemographic model variable selection spacial data mining of the 2011 Irish census. In *IEEE International Advance Computing Conference*, pages 613–622, Feb. 2014.

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

[35] ONS. Methodology note for the 2011 area classification for output areas, 2014. http://bit.ly/OAC11Meth.

[36] ONS. A Beginner's Guide to UK Geography, 2015. http://bit.ly/UKGeogs.

[37] S. Openshaw and P. Taylor. *The Modifiable Areal Unit Problem*. Geobooks, 1984.

[38] M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller. Visual parameter space analysis: A conceptual framework. *IEEE TVCG*, 20(12):2161–2170, Dec. 2014.

[39] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, July 2005.

[40] A. D. Singleton and P. A. Longley. Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*, 29(3):289–298, July 2009.

[41] A. Slingsby, J. Dykes, and J. Wood. Configuring hierarchical layouts to address research questions. *IEEE TVCG*, 15(6):977–984, Dec. 2009.

[42] A. Slingsby, J. Dykes, and J. Wood. Rectangular hierarchical cartograms for socio-economic data. *Journal of Maps*, 6(1):330–345, 2010.

[43] A. Slingsby, J. Dykes, and J. Wood. Exploring uncertainty in geodemographics with interactive graphics. *IEEE TVCG*, 17(12):2545–2554, Dec. 2011.

[44] E. R. Tufte. *The Visual Display of Quantitative Information*, volume 2. Graphics Press Cheshire, CT, 1983.

[45] C. Turkay, A. Lundervold, A. Lundervold, and H. Hauser. Representative factor generation for the interactive visual analysis of high-dimensional data. *IEEE TVCG*, 18(12):2621–2630, Dec. 2012.

[46] C. Turkay, A. Slingsby, H. Hauser, J. Wood, and J. Dykes. Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data. *IEEE TVCG*, Dec 2014.

[47] A. Unwin and D. Unwin. Exploratory spatial data analysis with local statistics. *The Statistician*, 47(3):pp. 415–421, 1998.

[48] D. Vickers and J. Pritchard. Visualising the output area classification. *Journal of Maps*, 6(1):410–416, Jan. 2010.

[49] D. Vickers and P. Rees. Creating the UK national statistics 2001 output area classification. *Statistics in Society*, 170(2):379–403, 2007.

[50] D. W. Vickers. *Multi-Level Integrated Classifications Based on the 2001 Census*. PhD thesis, University of Leeds, Jan. 2006.

[51] J. Yang, D. Hubball, M. O. Ward, E. A. Rundensteiner, and W. Ribarsky. Value and relation display: Interactive visual exploration of large data sets with hundreds of dimensions. *IEEE TVCG*, 13(3):494–507, 2007.