Interactive Generation of Visual Summaries for Multivariate Geographical Data Analysis

Cagatay Turkay, Aidan Slingsby, Helwig Hauser, Jo Wood, and Jason Dykes

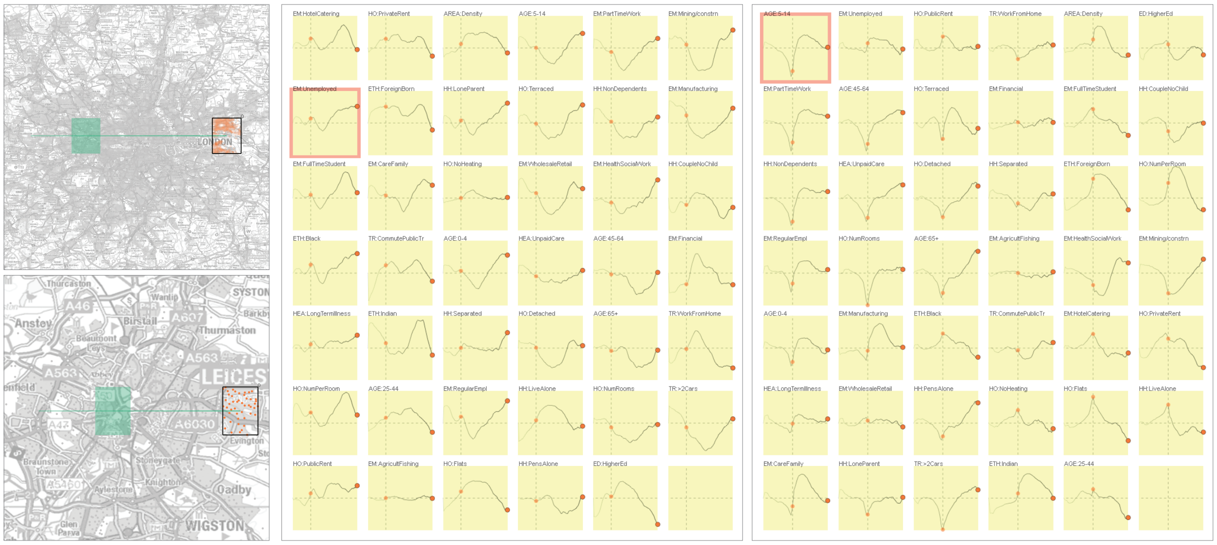

Fig. 1. Attribute signatures are dynamically created in response to interactive geographic selection sequences. A transect through the centers of London (polycentric city, left, top) and Leicester (monocentric city, left bottom). Attribute signatures for London (centre) and Leicester (right) are ordered by similarity to that at the top-left of each series of small multiples.

Abstract— The visual analysis of geographically referenced datasets with a large number of attributes is challenging due to the fact that the characteristics of the attributes are highly dependent upon the locations at which they are focussed and the scale at which they are measured. Here, we develop *attribute signatures* – interactively crafted graphics that show the geographic variability of statistics of attributes through which the extent of dependency between the attributes and geography can be visually explored. We compute a number of statistical measures and use them as a basis for our visualizations. Our methods allow variation in multiple statistical summaries of multiple attributes to be considered concurrently and geographically, as evidenced by an example in which the census geography of two cities in UK are explored.

Index Terms—Visual analytics, multivariate data, geographic information

1 Introduction

In some domains, multivariate data have a strong geographical component which dominates variation. Examples include population demographics, multivariate spatial interaction models, species distribution models and land-use models. Knowing how multiple attributes vary over space is critical in interpreting the phenomena that these data and models represent.

Designing mechanisms to support the exploration of the geographical variation in multiple attributes simultaneously is challenging due to the specific characteristics that geographical data have [1]. Geographical distributions tend to be heterogeneous and are often strongly related and influenced by topographic features. Some phenomena vary greatly, such as population density – a phenomenon that is highly dependent upon the extent of the spatial units used to measure it as well as the location at which it is measured.

Maps are often appropriate means for depicting geographical variation in data, graphically. However, this is only really effective where there are few attributes. Since maps already use position- and size-related visual variables, visual variables for depicting other attributes are limited. Choropleth maps [9] and geographical heatmaps [15] convey data using different aspects of colour, including lightness, saturation and hue. Additional attributes can be added by combining more visual variables or by using glyphs or other embedded matrices, but often at the expense of the geographical resolution at which the data are displayed. Additionally, *interactive techniques* are widely used to help make sense of many variables. These often avoid the problem of depicting spatial variation directly by facilitating *geographical filtering*. This usually results in non-geographical graphical depictions of the multiple attributes for one location only (e.g. [8]).

In this paper, we consider key issues associated with the geographic variation of multivariate data and develop interactive visual approaches to support the analysis of such datasets. We design, discuss and demonstrate how this can be done using map brushing in multiple coordinated views that show how multiple attributes vary in

- *Cagatay Turkay, Aidan Slingsby, Helwig Hauser, Jo Wood, and Jason Dykes are with the Dep. of Computer Science, City University London, UK. E-mail: {Cagatay.Turkay.1, Aidan.Slingsby.1, J.D.Wood, J.Dykes}@city.ac.uk*
- *Helwig Hauser is with the Department of Informatics, University of Bergen, Norway. E-mail: Helwig Hauser@ii.uib.no.*

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

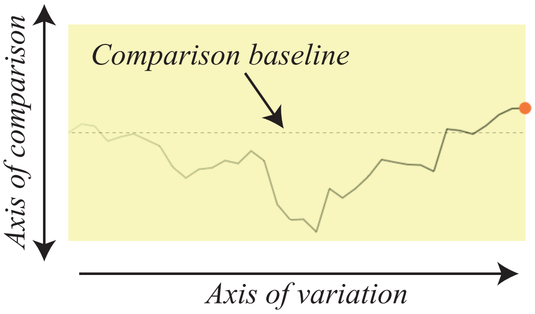

Fig. 2. An *attribute signature* represents changes in a single (or more) attribute along the axis of variation. The x-axis is the axis of variation, corresponding to the geographical aspect (location, extent or resolution) interactively defined by the user. The y-axis represents change in the computed statistics in response to this, comparing dynamically computed values to an appropriate baseline.

geographical space, extent and resolution. We introduce dynamically generated visual summaries, called *attribute signatures*, of how data varies along geography.

When these visual summaries are generated, we consider different aspects that may vary independently at various *geographical scales*. Within the context of this paper, we distinguish three aspects of space that are the basis of our analysis: location, scale extent and scale resolution [6]. *Location* is the geographical point at which a measurement is made. *Scale extent* (or *domain*) is the geographical extent around a location that is under consideration that defines an area [6]. *Scale resolution* is the amount of detail that is considered in characterising a location. It may be related to sampling strategy or data availability. The nature of the summaries will change as they are computed at these different spatial resolutions – an understanding of which reveals the scales at which homogeneity or heterogeneity exist in different aspects of population.

We use a single data set dataset through this paper to demonstrate the methods developed for analyzing multivariate geographic data. It consists of records taken from the UK Census of Population for the 181,000 output areas (OA) of England and Wales [7]. Each OA has 41 attributes associated with it, those deemed discriminating in developing the Output Area Classifier (OAC) [14] The result is a 41 x 181,000 multivariate table of values containing geographic characteristics.

2 Attribute Signatures

An *attribute signature* depicts a user-defined geographical variation of an attribute using one or more summary statistics as a sparkline [10]. Figure 2 illustrates how these aspects are represented in an instance of an attribute signature. The *axis of variation* (x-axis) represents either SL, SE or SR. When plotted against the *axis of comparison* (y-axis), variation in the attribute along the axis of variation is depicted using one or more summary statistics (section 2) compared to an appropriate baseline.

For each attribute, we construct a single attribute signature and arrange these using ordered juxtaposition [5] as a series of small multiples [11], one for each attribute (see Figure 1, right). These can be ordered in various configurations according to their similarity. The small multiples view is a component of a multiple-coordinated views environment in which interactive selections can be performed on location, extent and resolution on a map view to undertake location, extent and resolution style analysis. Linking signatures to a map view can be seen in Figure 1. Here, the user performs a sequence of selections on the map and the attribute signatures are generated dynamically in response to support SLc analysis. Since we are varying *location* (by moving the selection on the map) *location* becomes the *variation axis* on the signatures. For each point on x-axis, a *comparative statistic* (e.g. normalized difference between means) is computed between the selection and the baseline.

Statistical summaries for comparison : In response to the interactive selections on a map, we dynamically compute statistics to help investigate how attributes vary along the axis of variation. We employ a multiple-coordinated views approach, in which brushing on a map geographically conditions the data. Each attribute is summarised with a summary statistic relating to this area using Turkay *et al.*'s methods [12] whereby the statistic λ, e.g., a descriptive statistic such as mean μ or standard deviation σ, is computed using only the data points that are selected S_i at a particular location i on the variation axis. We then compare these "locally" computed results λ^{S_i} to a *baseline* value λ^{B_i} to calculate the difference at location i with: $\Delta_i = \lambda^{S_i} - \lambda^{B_i}$ similar to difference plots by Turkay *et al.* [13]. These computations are undertaken in real time for all the attributes, so Δ_i and λ are vectors of size p – the number of attributes in the data.

Linking signatures : To more effectively study how the attributes vary over space, we display the path along which the map was brushed or display the set of discrete locations selected (see Figure 1). This has the effect of leaving trails on the map. This allows us to see how attributes vary as we move along the trail on the map. Highlighting the interaction location along the x-axes of all signatures ensures that signatures are interactively linked to each other (via small dots displayed on the sparklines) and to the location and extent on the map at which the summary statistics are computed (via a path and a rectangle showing the selection). Moreover, bidirectional linking between the map and attribute signatures enable the identification of locations, extents or resolutions at which variations in the statistics occur. This type of linking between the map and abstract visual representations is shown to be effective in understanding the urban structures [3] and supporting multi-focus analysis in the paper by Butkiewicz et al. [2] where the authors developed probes to aggregate data on several locations on the map.

Key-framed brushing : In order to support users in developing their selection sequences we also introduce a semi-automated interaction mechanism called *keyframed brushing*. This method aids the user in quickly defining selection sequences that are precisely structured, by making equally placed selections that follow a straight line. This provides a regular spatial sample across any linear transect. In this mechanism, the user defines two or more brushes (according to their analytical goal) and using these *key brushes*, a sequence of *intervening* brushes are generated automatically over a linear path that connects these key brushes. After the brush sequence is computed, the system starts traversing through this without the need for further input by the user.

3 Analysis Example – Transects through cities

Here, we present an analysis example where attribute signatures are utilized. Inspired by Duany's concept of the 'urban transect' [4], we explore transects through London (a polycentric city) and Leicester (a monocentric city) in this example. We employ our key-framed brushing mechanism to create a linear west-east *transect* that starts at the westernmost outskirts of a city, passes through the center and continues to the eastern outskirts (Figure 1). We report values in attribute signatures using *effect size* and local baselines so we can compare local variation in cities.

Attribute signatures across London (Fig. 1, center, left-top) are variously shaped as *m*, *v*, *u* or *n* – highlighting differences between inner and outer London, with significant differences in central London for the *m-shaped* signatures, such as for commuting using public transport.

The lack of symmetry as we move across London reveals interesting structure, such as the low proportion of home workers, high proportion of infants and proportion of adults separated or divorced at the eastern fringes of the city compared to the west. These figures vary significantly despite other similarities between the east and west ends of the city relating to population density, tenure and the data on commuting.

In Leicester (Fig. 1, right, left-bottom) the very sharp dips or peaks at a single location at the city centre for attributes such as % of detached houses, % of children, or % people living alone reflects the concentration of students and young professionals living in apartments in this part of town, which is distinct in character from the other locations along the selected path. This is typical of a monocentric city and the variation is captured with the scale of the extent used here.

4 Conclusions

Our stated aim is to develop techniques to help understand how multiple attributes vary over space as a means of gaining knowledge of the phenomena represented by geographic data. Being able to access these different representations of the data and perform comparative visual analysis on them simultaneously is important in dealing with the characteristics of geographic data that make them interesting. It enables us find and present the stability, or otherwise, of the numbers that we compute to describe geography and use broad visual channels to show how they vary using visualization methods that are applicable to a broad range of multivariate geographic data.

References

[1] L. Anselin. *What is special about spatial data?: alternative perspectives on spatial data analysis*. National Center for Geographic Information and Analysis Santa Barbara, CA, 1989.

[2] T. Butkiewicz, W. Dou, Z. Wartell, W. Ribarsky, and R. Chang. Multi-focused geospatial analysis using probes. *IEEE TVCG*, 14(6):1165–1172, 2008.

[3] R. Chang, G. Wessel, R. Kosara, E. Sauda, and W. Ribarsky. Legible cities: Focus-dependent multi-resolution visualization of urban relationships. *IEEE TVCG*, 13(6):1169–1175, 2007.

[4] A. Duany. Introduction to the special issue: The transect. *Journal of Urban Design*, 7(3):251–260, 2002.

[5] 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, 2011.

[6] 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.

[7] Office for National Statistics. Census Dataset Finder - Data Explorer (Beta) - http://j.mp/onsDX, 2014.

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

[9] T. A. Slocum. *Thematic cartography and visualization*. Prentice hall Upper Saddle River, NJ, 1999.

[10] E. Tufte. Sparklines: Intense, simple, word-sized graphics. *Beautiful Evidence*, 1:46–63, 2004.

[11] E. R. Tufte. *The visual display of quantitative information*. Graphics Press, Cheshire, CT, 1983.

[12] C. Turkay, P. Filzmoser, and H. Hauser. Brushing dimensions – a dual visual analysis model for high-dimensional data. *IEEE TVCG*, 17(12):2591 –2599, dec. 2011.

[13] C. Turkay, A. Lex, M. Streit, H. Pfister, and H. Hauser. Characterizing cancer subtypes using dual analysis in caleydo stratomex. *IEEE CG&A*, 34(2):38–47, Mar 2014.

[14] D. Vickers and P. Rees. Introducing the national classification of census output areas. *Population Trends*, 125:380–403, 2007.

[15] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2), 2009.