Slingsby, A., Tate, N. & Fisher, P. (2014). Visualisation of Uncertainty in a Geodemographic Classifier. Paper presented at the GIScience 2014: Uncertainty Workshop, 23 September 2014, Vienna.

Visualisation of Uncertainty in a Geodemographic Classifier

Aidan Slingsby[1], Nicholas J. Tate[2], and Peter F. Fisher[2]

[1] giCentre, Department of Computer Science, City University London, Northampton Square, London, EC1V 0HB; a.slingsby@city.ac.uk
[2] Department of Geography, University of Leicester, Leicester LE1 7RH, UK; njt9@le.ac.uk

Abstract. We explore some ideas around quantifying and visualising classification uncertainty within a geodemographic classifier. We demonstrate spatially-constrained small-multiples to show geographical variation, their combination with a Gastner population cartogram projection to normalise with respect to population, explore a fuzziness parameter when producing fuzzy-sets, and look at implications of taking into account this uncertainty when profiling population, finding that this can have significant effects that are worth investigating further.

1 Introduction

Geodemographic classifiers characterise geographical *areas* based on characteristics of those who live there. A set of a *geodemographic categories* based on a set of census-derived population data is defined – often with short descriptive labels such as 'Multicultural' and 'Blue collar' – and then one is assigned each geographical area. Thus, each small area is allocated a category that reflects the characteristics of the population living there (Figure 1, left). Geodemographics are in widespread use, helping target campaigns and advertising, assessing the viability of products and services, doing stratified sampling and enriching existing geographical data [7].

2 Classification uncertainty

Inevitably, characterising population into one of seven categories results in places whose population is characterised well and places where it is not.

The 2001 "Output Area Classification (OAC)" is a geodemographic classifier [11] which classifies Output Areas (OAs; the smallest reporting spatial units from the 2001 UK census (average population of 297 [8] for England and Wales) into seven main geodemographic categories ('super-groups') indicated in Figure 1 (left). We use it because unlike its commercial 'black-box' rivals, it is freely available and full details of how it was built, population data variables used and uncertainty information are provided. Uncertainty information for each OA is

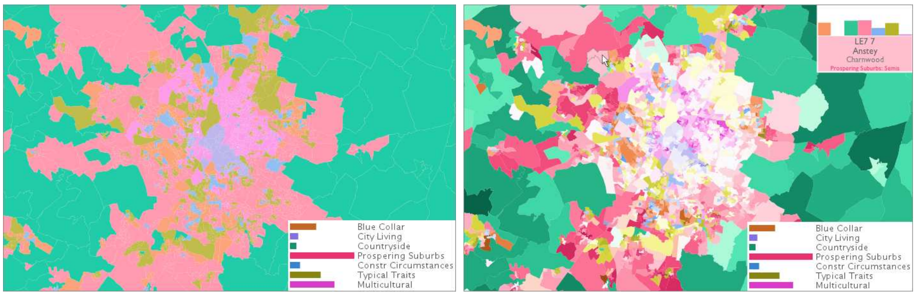

Fig. 1. *Left:* Map of OAC's geodemographic categories assigned to areas (Output Areas; OAs) in Leicester (UK). The bottom right barchart indicates population in each geodemographic category. *Right:* As left, but lightness corresponds to classification uncertainty where dark is more certain. The top right barchart shows membership of each geodemographic category for the OA indicated by the mouse pointer. *Data: 2001 Census, Output Area Boundaries. Crown copyright © 2003. Crown copyright material is reproduced with the permission of the Controller of HMSO.*

provided as a set of seven 'distances' indicating similarity to the typical population profiles of each geodemographic category. The larger the distance, the less well the category characterises the population. In normal use, the closest geodemographic category is used, but this is not always a good characterisation of the population; hence the reason for this work.

Slingsby *et al* [9] uses a measure of how well the allocated (closest) geodemographic category characterises the population (see paper for details). This is shown as colour lightness in Figure 1 (right). Hue indicates category and lightness indicates this 'typicality' measure. The figure shows that in the City centre (centre of the map), geodemographic categories poorly characterise population (pale) and characterise population better in more peripheral areas. The OA indicated with the mouse pointer is pale red which means 'Prospering suburbs' but a poor characterisation of the OA's population. The bar chart at the top right shows it is also close to 'Countryside' (green), 'Blue collar' (orange) and 'Typical traits' (yellow).

This *classification uncertainty* and how it varies across space and by category may have implications for its application underpinning resource targeting, analytical work and decision-making. Slingsby *et al* [9] explored this with some expert users who found this a thought-provoking exercise and were particularly surprised at the degree of classification uncertainty in certain areas. It was unclear how this would affect their use of geodemographics in future, but the work indicated that this issue is worth exploring.

3 Spatially-varying graphs

Thus far, we have considered uncertainty information per OA and only mapped one uncertainty value per OA. If we aggregate space into grid cells and then

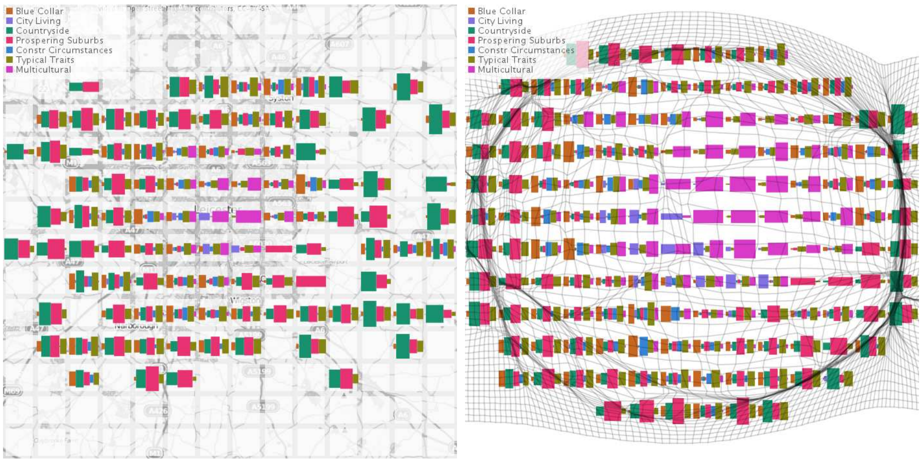

Fig. 2. *Left:* Graphs of membership of each geodemographic category of places within grid squares. x-axis indicates proportional membership; y-axis indicates absolute degree of membership. *Right:* As left, but first projecting the map as a population cartogram and then gridding that space. The overlain grid indicates geographical distortion. *Base map from OpenStreetMap.*

embed a chart that characterises the OAs within that grid cell, we can potentially provide more uncertainty information, though arguably, the process of averaging values into grid cells introduces another kind of uncertainty.

We will consider average distance to each geodemographic category for each grid cell (rather than OA). We will also consider two scalings of this: *absolute membership* which uses the inverse distances directly and *proportional membership* which scales this between the minimum and maximum average distance. These two measure are depicted in Figure 2 (left) along the y-axis and x-axis, respectively. Around the periphery, 'Countryside' (green) and 'Prospering suburbs' (red) tend to dominate in both proportional and absolute terms: i.e. places in these grid-cells are mainly characterised by these two categories. In central areas, 'Mulicultural' dominates yet it is not such a good characterisation of the population there.

To take into account the denser population in central Leicester, in Figure 2 (right) we have experimented with projecting the map as *first* projecting the map as a Gastner-type population cartogram [2] and *then* use the regular grid-based partitioning. Each grid square now contains a similar size of population. Although geographical space is distorted, more details of the dense central area are visible; in particular, 'City Living' (indigo) in the SW portion.

4 Possibilistic Fuzzy Sets

There are other ways to quantify geodemographic category membership. Possibilistic c-Means (PCM) [4] does this using fuzzy sets and the m parameter

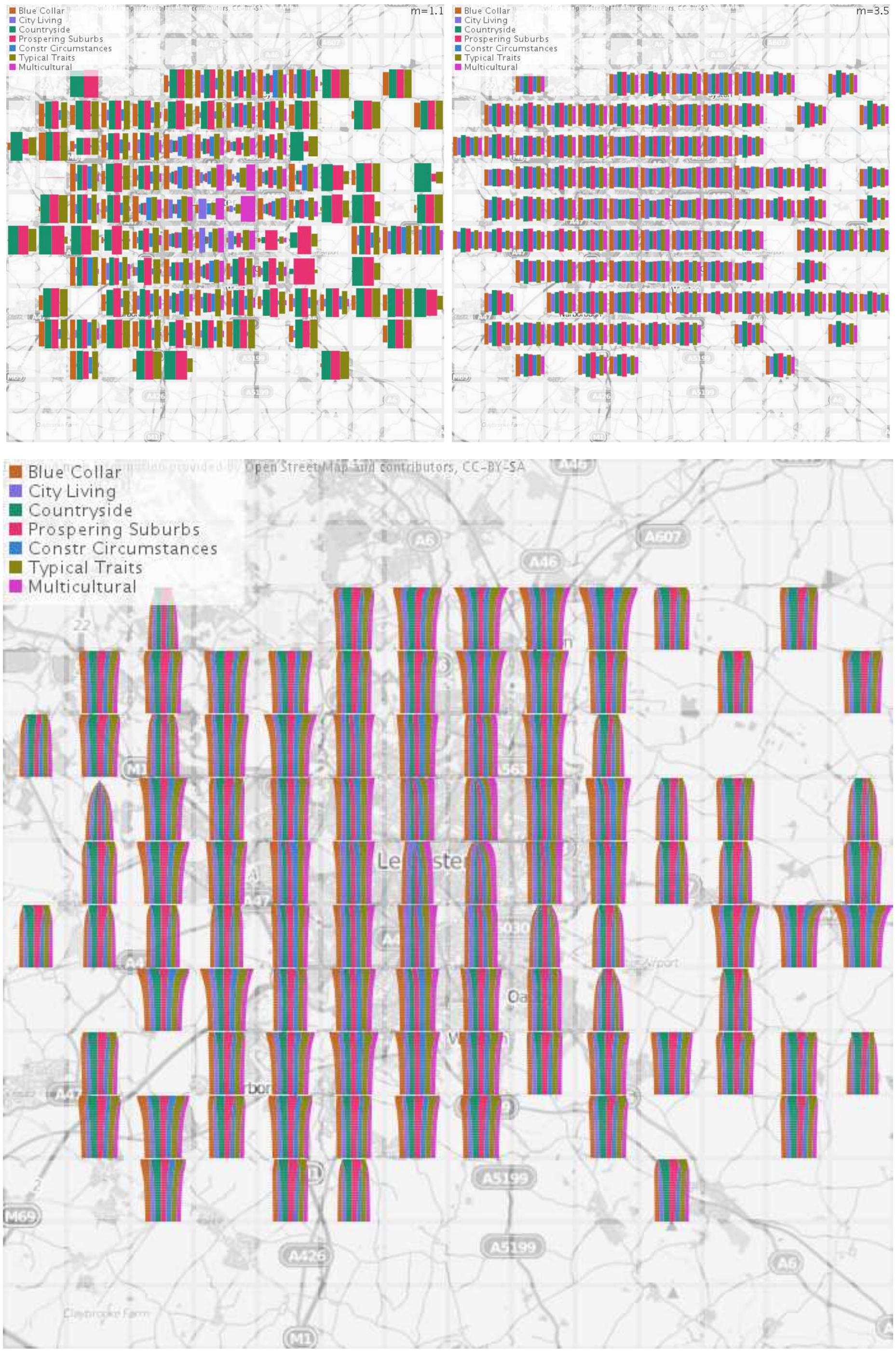

Fig. 3. As the graphs in Figure 2 but using Possibilistic-Means (PCM) [6]. *Top left:* $m = 1.1$; *Top right:* $m = 3.5$. *Bottom:* Graphs of category membership (x axis) that shows the effect of continuously varying m (y-axis) from $m = 1.1$ at the top to $m = 3.5$ at the bottom. *Base map from OpenStreetMap.*

that adjusts the fuzziness applied to the membership set. There has been debate about what the best value to use for m [5] and Okeke and Karnieli's [10] tried multiple values of m. We investigate the effect of this parameter using the graphs from Figure 2. Figure 3 (top) shows the effect of low m and high m, with the latter almost completely smoothing out category memberships. In Figure 3 (bottom) we continuously vary m from 1.1 to 3.5 along the y axis from top to bottom with absolute membership on the x axis. Low m-values give lower memberships in some areas (narrower at the top) and high memberships in other areas (wider at the top). As m approaches 2, memberships differences are smoothed out.

5 'Monte Carlo' type Simulation

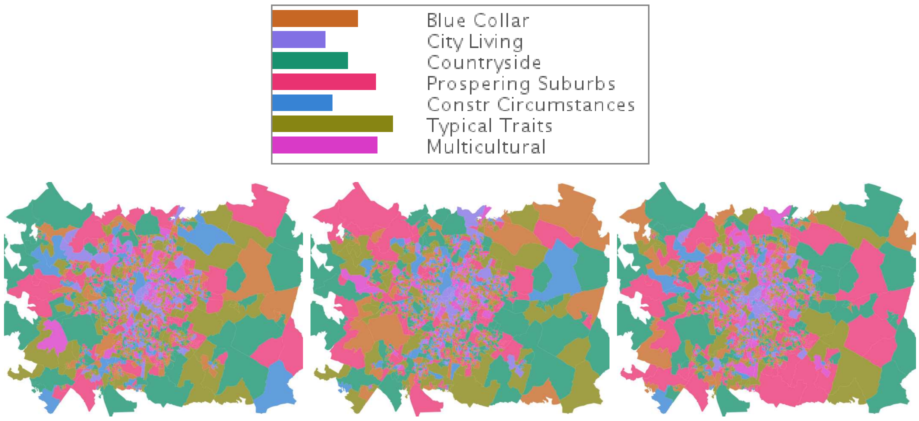

Fig. 4. *Top:* Amount of population in each geodemographic category after 1000 'Monte-Carlo' type runs. *Bottom:* Three alternative maps [3]. Notice how some of the largest OAs switch between 'Countryside', 'Prospering suburbs' and 'Typical Traits'. *Data: 2001 Census, Output Area Boundaries. Crown copyright © 2003. Crown copyright material is reproduced with the permission of the Controller of HMSO.*

Finally, we turn our attention to possible *implications* of classification uncertainty. In Figure 4 we do a 'Monte-Carlo' type simulation where we randomly assign a geodemographic category to each OA weighted by the category membership. This means that if a geodemographic category has double the membership as another, it will be twice as likely to be allocated. The population barchart in Figure 4 shows the median population allocated to each category after 1000 runs. Significantly, although Figure 1 shows that 'Prospering suburbs' has the largest population share, here the *greatest share of the population is 'Typical traits'.* This is because 'Typical traits' is close to most OAs but is rarely the closest. Although a very simple experiment, it indicates that *taking the degree* *of classification uncertainty into account may affect geodemographics-supported analysis and decision-making.*

6 Conclusion

We have explored some ideas around quantifying and graphically depicting geographical classification uncertainty within the OAC geodemographic classifier and consider possible implications of this. We have suggested gridding space to produce regular geographically-constrained small-multiples and have suggested using a Gastner Cartogram projection to give a population-weighted depiction of the results. We have quantified classification uncertainty as relative (proportion), absolute and fuzzy sets; in the latter case, we used graphics to depict the effect of changing fuzziness (m) parameter. Finally, using a 'Monte Carlo' style approach, we look at some of the implications of taking into account this uncertainty when profiling population and we believe that finding ways to take account of this uncertainty will help make more informed use geodemographics.

Acknowledgements

OAC data were obtained from http://www.sasi.group.shef.ac.uk/area_classification and with help from Dan Vickers.

References

1. Bezdek, J.C. Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York (1981).
2. Dorling, D., Barford, A. and Newman, M. Worldmapper: The world as you've never seen it before. IEEE Transactions on Visualization and Computer Graphics, 12 (5), 757–764 (2006).
3. Fisher, P. Visualizing Uncertainty in Soil Maps by Animation. Cartographica: The International Journal for Geographic Information and Geovisualization, 30 (2), 20–27 (1993)
4. Fisher, P.F. and Tate, N.J. Modelling Class Uncertainty in the Geodemographic Output Area Classification. Environment and Planning B(in press).
5. Fisher, P. Remote sensing of land cover classes as type 2 fuzzy sets. Remote Sensing of Environment 114 (2), 309–321 (2010).
6. Fisher, P., Tate,N. and Slingsby, A. Type-2 Fuzzy Sets Applied to Geodemographic Classification. Proceedings of GIScience (extended abstracts), Vienna, (2014).
7. Harris, R., Sleight, P. and Webber, R. Geodemographics, GIS and neighbourhood targeting. John Wiley and Sons. (2005).
8. ONS. Changes to Output Areas and Super Output Areas in England and Wales, 2001 to 2011 http://www.ons.gov.uk/ons/guide-method/geography/products/census/report--changes-to-output-areas-and-super-output-areas-in-england-and-wales--2001-to-2011.pdf (2012)

9. Slingsby, A., Dykes, J. and Wood, J. Exploring Uncertainty in Geodemographics with Interactive Graphics. IEEE Transactions on Visualization and Computer Graphics 17 (12), 2545–2554 (2011).
10. Okeke, F., and Karnieli, A. Linear mixture model approach for selecting fuzzy exponent value in fuzzy c-means algorithm. Ecological Informatics 1, 117–124 (2006).
11. Vickers, D. and Rees, P. Introducing the national classification of census output areas. Population Trends, 125:380403, (2007).