City Research Online

Permanent City Research Online URL: http://openaccess.city.ac.uk/437/

Versions of research

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

Enquiries

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

Exploring Uncertainty in Geodemographics
with Interactive Graphics

Aidan Slingsby, *Member, IEEE*, Jason Dykes, and Jo Wood, *Member, IEEE*

Fig. 1. Parallel coordinate plots showing the 41 census variables used in the Output Area Classification (OAC) by super-group. Values are standardised to min/max, median values are shown with a bold line and variation is shown as shaded 1st - 9th deciles. Variables are in their original order that groups by type [51]. The five variable types are identified, along with some key distinguishing variables. *Top:* GB population as a whole. *Bottom:* GB population split by OAC super-group. See Fig. 2 for the legend.

Abstract—Geodemographic classifiers characterise populations by categorising geographical areas according to the demographic and lifestyle characteristics of those who live within them. The dimension-reducing quality of such classifiers provides a simple and effective means of characterising population through a manageable set of categories, but inevitably hides heterogeneity, which varies within and between the demographic categories and geographical areas, sometimes systematically. This may have implications for their use, which is widespread in government and commerce for planning, marketing and related activities. We use novel interactive graphics to delve into OAC – a free and open geodemographic classifier that classifies the UK population in over 200,000 small geographical areas into 7 super-groups, 21 groups and 52 sub-groups. Our graphics provide access to the original 41 demographic variables used in the classification and the uncertainty associated with the classification of each geographical area on-demand. It also supports comparison geographically and by category. This serves the dual purpose of helping understand the classifier itself leading to its more informed use and providing a more comprehensive view of population in a comprehensible manner. We assess the impact of these interactive graphics on experienced OAC users who explored the details of the classification, its uncertainty and the nature of between – and within – class variation and then reflect on their experiences. Visualization of the complexities and subtleties of the classification proved to be a thought-provoking exercise both confirming and challenging users' understanding of population, the OAC classifier and the way it is used in their organisations. Users identified three contexts for which the techniques were deemed useful in the context of local government, confirming the validity of the proposed methods.

Index Terms—Geodemographics, OAC, classification, cartography, uncertainty.

✦

1 INTRODUCTION

Large multivariate datasets can help quantify complex phenomena. These can be simplified into a manageable set of categories using dimension-reducing classifiers.

Geodemographic classifiers are types of *area classification* that characterise the populations of small areas using discrete categories, produced through the statistical clustering of demographic and lifestyle data variables [22]. Typically, a hierarchy of categories is produced, each level discriminating population to different degrees. The

• *Aidan Slingsby (a.slingsby@city.ac.uk), Jason Dykes (jad7@city.ac.uk), and Jo Wood (jwo@city.ac.uk) are based at the giCentre (http://gicentre.org/), City University London.*

resulting categories are then assigned to the geographical areas whose population characteristics best match those of the globally defined categories. This convenient means of classifying small-area populations into a manageable number of categories is widely-used in both the private and public sectors for characterising population, stratifying population samples, marketing and targeting products or services [22].

As with all processes of generalisation, there is inevitable and inherent uncertainty due to classification error [20]. Population profiles of geographical areas may share many or few characteristics with multiple categories, yet each is assigned to one category. This results in significant heterogeneity within these categories, which varies by category and geographical region. Differences in the share of population classified into each of the categories at a local level affects the discriminating potential of the classifier. This also varies by category and region. These are important considerations in the design of classifiers [34, 51] and although efforts are made to mitigate these problems, they cannot be eliminated. To a large extent, the discriminating ability of categories is context-dependent and new or tailored geodemographic

classifiers are built to meet these needs [41, 39]. Although geodemographic classifiers are designed to be used at face-value, there may be advantages in knowing and exploring the degree of variation within them, in which places and for which categories it is most significant.

The main barrier to studying this is the fact that most geodemographic classifiers are 'black-boxes' whose details are not published. A notable exception is the *Output Area Classification* (OAC) – a free and open UK geodemographic classifier [51, 36]. Details of its construction methodology, the 41 variables used in its construction and the classification uncertainty for each geodemographic category of each of the 223,060 geographic areas are in the public domain [43].

The other significant barrier is the volume and nature of the information to interpret and digest. Desktop computers are now capable of providing on-demand access to some of the large amounts of data behind classifiers. Although generalisation is an important data reduction strategy prior to visualization [2], we argue that there are benefits to augmenting information about classifiers with the data used to build them and exploring this visually [13]. There is good evidence that well-designed interactive visualization can help make large amounts of information accessible and intelligible [46]. Thomson *et al.* [47] find that analysts with effective graphical representations of uncertainty make better decisions. Our interactive visualization techniques are designed to enable analysts to "*decompose the uncertainty visualization to inspect the underlying components*" as measured levels of uncertainty – "*error*" [31] or "*derived uncertainty*" [38]. Our designs aim to address the "*lack [of] methods for depicting uncertainty [including error] simultaneously with data and interacting with those depictions in ways that are understandable, useful and usable*" [31]. Fisher [17] described the Boolean data model as "*a convenient artefact of the map production process*" that obscures important variation and yet is often used uncritically. He called for data exploration to support the critical investigation of uncertainties in geographic information characterised by innovation and dynamism framed around cartographic design principles. Other classifiers may benefit from these approaches.

This design paper has two main aims. *Firstly*, to *design* a set of interactive visualization techniques to delve into OAC and *demonstrate* their use for exploring population profiles of areas, how these relate to geodemographic categories, classification uncertainty and how these vary by category and geographically. We do this by describing and justifying our design, using this to identify and present characteristics of population and OAC. *Secondly*, to *assess the impact* of this design on a set of OAC users: their understanding of the classifier (OAC), the phenomenon (population) and potential implications for their use of OAC in the future. We ran a focussed user study with six participants with substantial experience of OAC in the workplace. Although users knew about the kind of uncertainty that exists in geodemographic classifiers and had anecdotal evidence of this, they did not know the *degree* to which this applied, where or what the implications might be and had no systematic means to find out. One participant commented: "*Although data are open, OAC has always been a black box to me. This helps unpick the classification, helping me understand how OAC works and that sometimes you've got to be a little careful*". Exploration of OAC using our techniques illustrated potential implications of classification uncertainty and the need to be able to assess this routinely. Importantly, participants found that the ability to *explore the data that drives the classifier* put the uncertainty into context in ways that could support decision-making. They showed a strong desire to use the techniques presented here to support their work.

2 RELATED WORK

The interactive and visual design presented here is based on preliminary work [57, 56], to which we have added further interactions to support comparison, improved some of the visual encodings and assessed impacts on users. Fisher [18] also explores some characteristics of uncertainty in OAC using fuzzy and probabilistic indices and presents these as static univariate small-multiple maps for a single city. Very little published work addresses uncertainty in geodemographics directly because of its commercial roots and the closed nature of most

Fig. 2. OAC's 7 super-groups and 21 groups, sized by national population. Hues are used consistently across all the figures. The number of sub-groups present in each group is shown in parentheses.

geodemographic classifiers. There is however, a body of academic work directed towards public sector use of geodemographics [41, 39].

Visual representations of population are widespread. Thematic maps depict population characteristics in discrete categories, an early example being Charles Booth's poverty map [3] (considered an early example of a geodemographic classifier). Choropleth maps show values of numerical demographic data. Innovative projections [9] and interactive analytical graphics [23, 15, 10, 1, 7, 27] help visually analyse multiple population variables through linked spatial, aspatial and semi-spatial views. Increasingly sophisticated means of selection [54, 25], interaction [58] and linking between views [52] aids comparison. Recent advances in World Wide Web standards and browser capabilities have resulted in a variety of web-based population maps, some of which have a highly innovative social and collaborative emphasis [24].

Using visualization in the context of the classification of multi-band imagery is common practice and novel approaches have been developed and demonstrated to be effective in research [30] and education [32]. Little is published on how exploring classifiers might allow them to be used more effectively, yet we know that sophisticated users of classifiers and models do their own investigations to help optimise their use of them. For example, our work with colleagues in the insurance industry confirms that sensitivity analysis of commercial 'blackbox' catastrophe models is an important portfolio optimisation tool.

3 OUTPUT AREA CLASSIFICATION (OAC)

A significant barrier to opening up classifiers in this way is the 'blackbox' nature of commercial geodemographic classifiers such as Experian's Mosaic [16] and CACI's Acorn [5]. The Output Area Classification (OAC) [51, 36] is a notable exception. It is a free geodemographic classifier that openly publishes details of the data variables used in its construction and similarity measures of each geographical area to each demographic category [43]. The most recent OAC was built using data acquired from the most recently published census (2001). A new version is planned using the 2011 census [37].

OAC classifies 223,060 Output Areas (OAs; the smallest geographical reporting units of the UK census [33]) into a hierarchy of 7 super-groups, 21 groups and 52 sub-groups (Fig. 2). Here, we consider those in Great Britain (GB; this excludes Northern Ireland). Each OA has a characteristic profile based on the 41 census variables used in its construction. Super-groups' and groups' labels are designed to evoke the essence of these classes. We do not consider sub-groups here as these are primarily used for modelling and re-aggregating to make tailored versions of OAC (though some of the participants from the evaluation wanted sub-groups be available for this purpose; section 5.5.3).

4 DESIGN AND OAC EXPLORATION

Our interactive visualization techniques are designed to explore classification uncertainty in OAC, geographically and by OAC category. These are described and demonstrated in this section and are available through our software (Fig. 3) which produced most of the figures in this paper and is demonstrated in the accompanying video (http://vimeo.com/25460216).

The software was designed to help us demonstrate and evaluate these techniques through a 'chauffeuring' approach with users. The uncluttered interface has functionality that can be hidden, yet remains
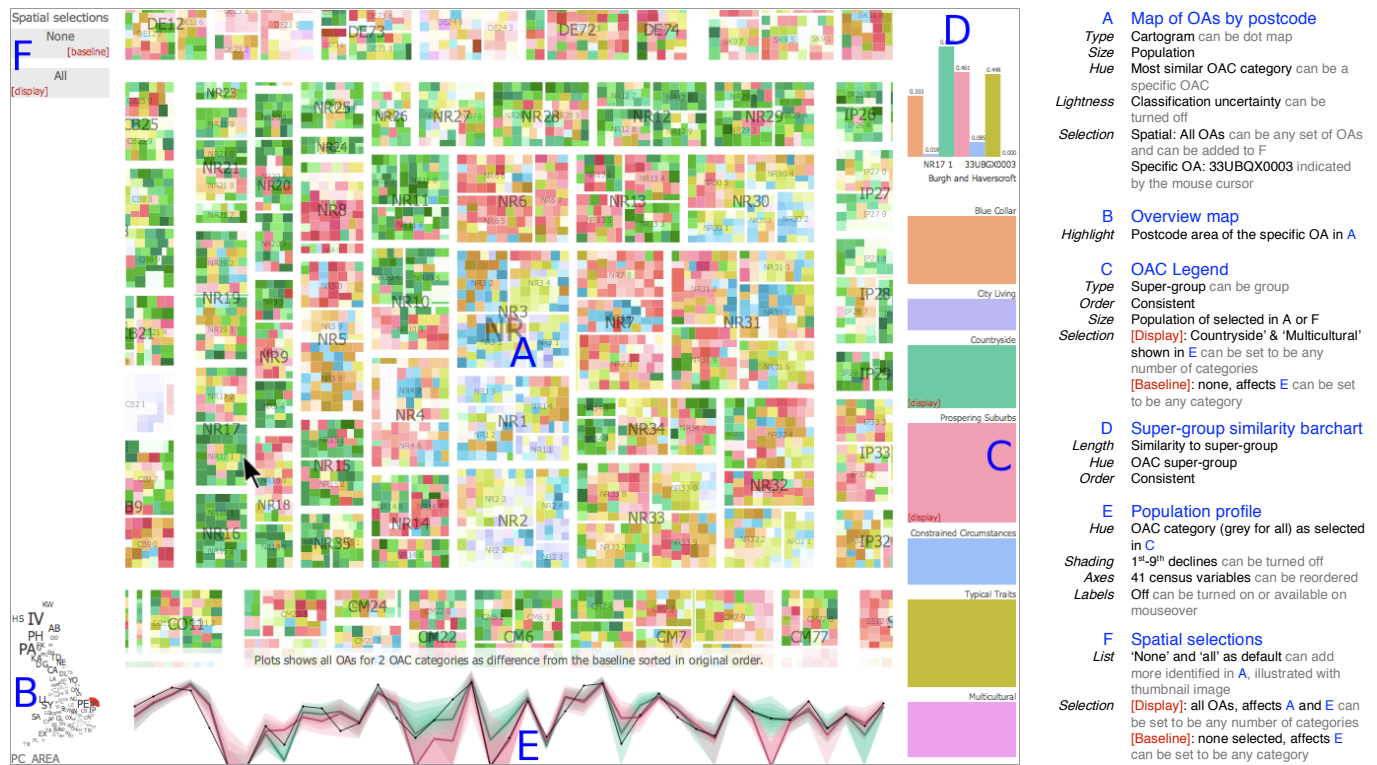
Fig. 3. Screenshot of the software that implements our design as described in section 4. Details listed on the right. Here, *all* OAs are selected (see [display] in F). A is zoomed to the NR postcode (overview in B). Mouse cursor indicates a 'Countryside' OA that is also quite similar to two other super-groups (D). Its population profile (41 census variables) is the black line in E shown alongside two national profiles indicated by [display] in C.

accessible. This suits demonstration and chauffeuring well because functionality is only revealed when required; a fact positively commented upon by one of the participants in our evaluation. There is evidence that cluttered user interfaces can detract from the data being shown [40] and that aesthetics can have a significant effect on user experience [6]. These considerations have strongly influenced our design.

4.1 Colour

Perceptually evenly-spaced hues of equal lightness [56, 53] depict super-groups (×7), shown in Fig. 2 and used consistently across views. Groups' (×21) colours are derived from their parents' hues and are also perceptually equally spaced. Although too indistinguishable for lookup tasks, they allow heterogeneity in adjacent areas to be detected.

Since lightness is held constant, it is available to encode other information. Perceptually-uniform variations in lightness represent the relative similarity to allocated category (section 4.4), such that lightness can be directly compared across hues [56]. As the similarity decreases, colours converge to white indicating that the allocated category is a poor characterisation of the area. Lightness is considered an appropriate visual variable for this kind of information [31]. The result is a categorical sequential bivariate colour scheme [4] showing category and uncertainty as is appropriate for encoding attribute accuracy for categorical data in areal coverages [47].

4.2 Geographical distribution

Each Output Area (OA) is assigned an OAC category. The dot map in Fig. 4 reveals that most of the land area contains OAs classified as 'Countryside', yet Fig. 2 shows this is a relatively small proportion of the population. The zoomed-in portion in Fig. 4 (right) reveals huge variations in OA density, illustrating the difficulty of producing national population maps [42].

Density-normalising population cartograms size geographic areas by population. By distorting geographical space, the visual prominence of areas becomes more proportional to population. This has
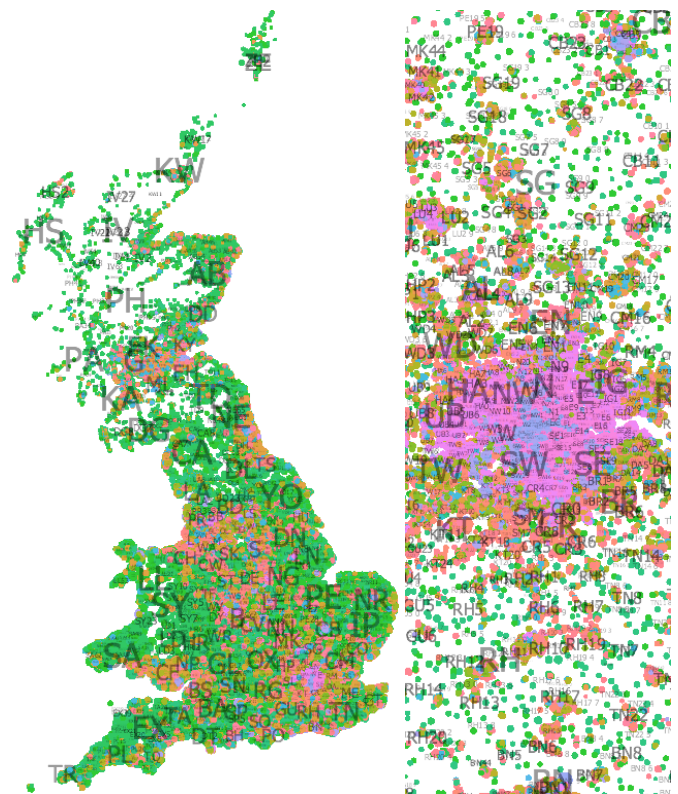
Fig. 4. Centroids of OAs coloured by super-group. Labels indicate postcodes. *Left:* National map shows most of the land *area* is 'Countryside'. *Right:* The zoomed-in portion centred on London shows marked geographical differences in OA density and classification.
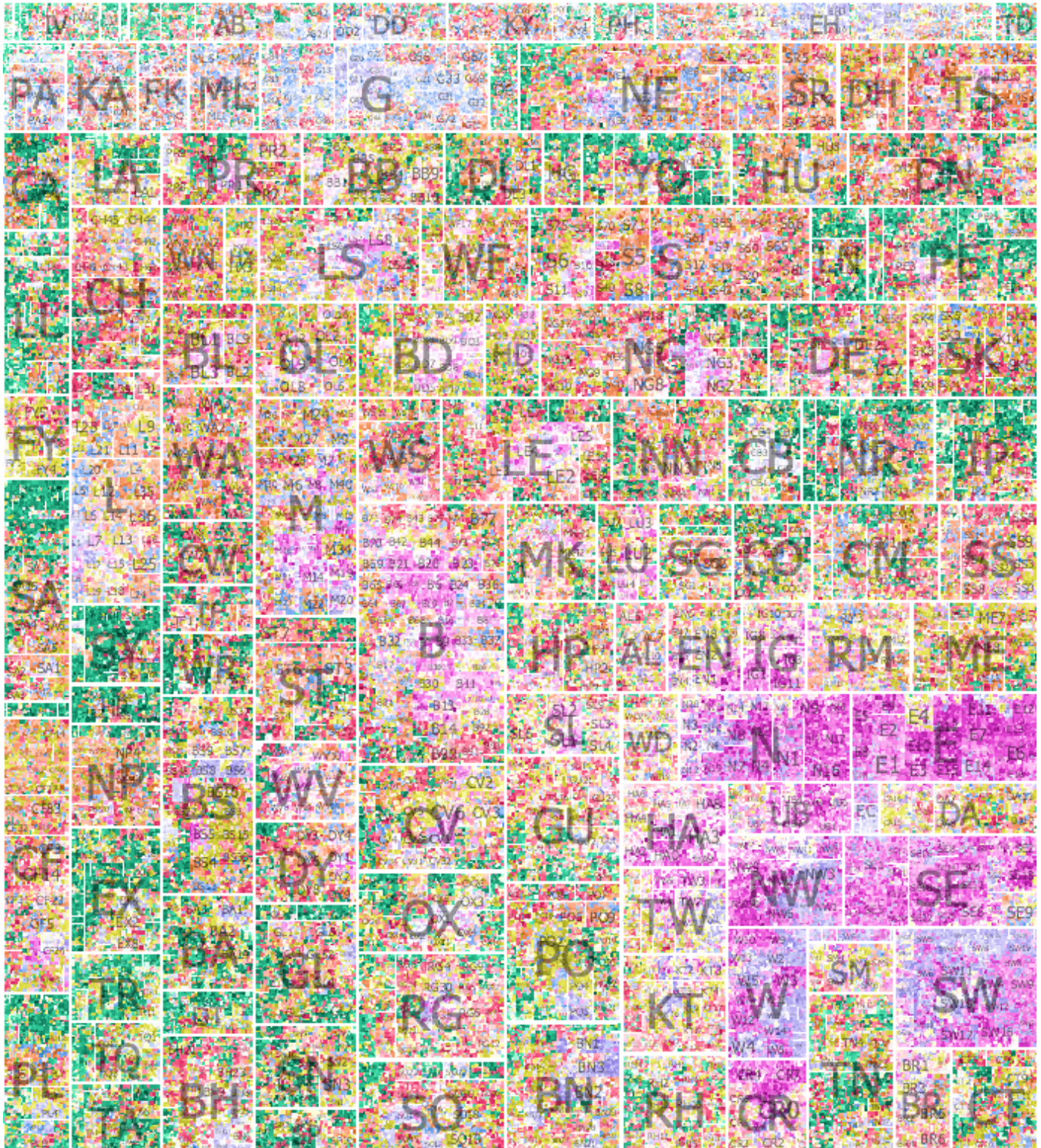
Fig. 5. Hierarchical rectangular cartogram of >200,000 OAs, organised into the postcode hierarchy. Hue denotes their OAC super-groups (Fig. 2) and lightness indicates the classification uncertainty of each. The screenshot of the software in (Fig. 3A) contains this graphic, zoomed in on the NR postcode area. The N, NW, E, W, SE and SW postcode areas correspond to area of London, showing the unique characteristics of the capital city and its dominant effect on OAC. The lighter colours of the Scottish OAs (top) indicates that these OAs tend to be less typical of their allocated super-groups than for those in other parts of the country.
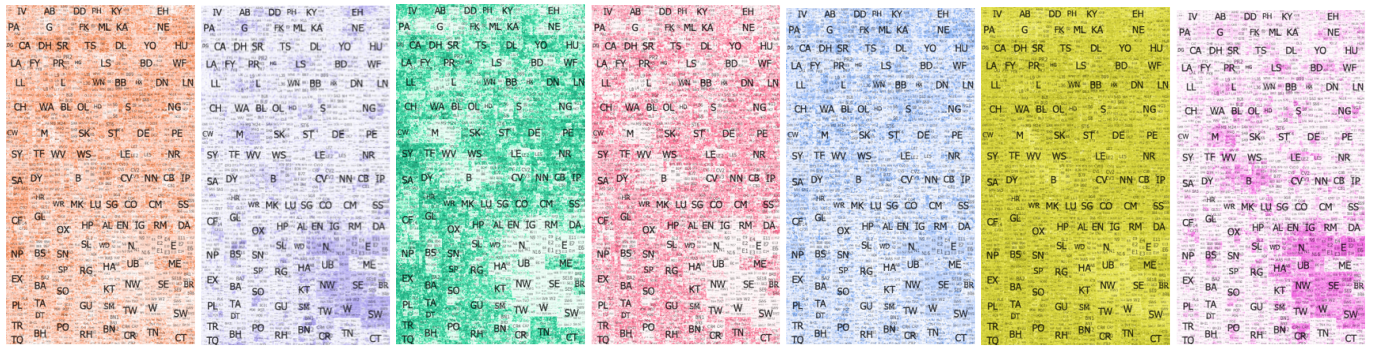
Fig. 6. Hierarchical rectangular cartogram of each super-group, where lightness indicates the relative similarity of each OA to each super-group.

been demonstrated to stark effect generally [9] and for OAC specifically [50]. There are a variety of techniques for creating cartograms [48, 8] with different properties. In Fig. 5, we use a hierarchical rectangular cartogram [42] (a treemap with spatial ordering [55]), where lightness indicates similarity to the OA's geodemographic category (section 4.4). OAs are organised within three levels of the UK postcode hierarchy. Although the census geography is more appropriate for population studies [33], we use postcodes because of their widespread familiarly and their compact and familiar labelling. Postcodes provide 125 postcode areas (e.g. B), 3064 postcode districts (e.g. B12) and 11598 postcode sectors (e.g. B12 5). OAs themselves do not have names, but we provide the names of their LSOAs (a census geography coarser than OAs). Each rectangle represents an OA and these are geographically arranged within the postcode hierarchy. The resulting non-occluding, space-filling, population-normalised cartogram has some advantages over the dot map. Coloured area is now proportional to the population of OAs classified as each super-group – notice how much less 'Countryside' (green) area there is in the cartogram than the dot map. However, significant geographical distortion is evident when compared with the dot map and some of the spatial patterns are due to artefacts of the layout algorithm – as with all cartograms there are trade-offs between shape, position and adjacency to different degrees. For example, in Vickers *et al.*'s Gastner Cartogram [50], adjacency is preserved at the expense of convexity of shape, making fill colour more difficult to discern and populations sizes more difficult to compare. In the rectangular cartogram case, screen-space is used more efficiently (less white space) and the regular shape makes size more comparable [28]. Geographical coherence is better at deeper levels of the hierarchy and worst between postcode areas. In most cases, the core-periphery structure is preserved in postcode areas (e.g. B), many (but not all) of which correspond to city centres. Whilst Fig. 2 shows overall proportions of each category, the cartogram reveals that these are subject to strong geographical variation. 'Multicultural' and 'Constrained by Circumstances' dominate the centres of some postcode areas, with the former particularly evident towards the southeast and in the north. 'Countryside' and 'Prospering Suburbs' tend to co-occur around postcode area peripheries, particularly in the west and north. 'Typical Traits' is more scattered, but cohesive blocks are evident. The nature of this variation will be explored later in this section.

4.3 Classification distribution

For geographical selections, we provide a legend whose elements are sized by population (Fig. 3C), following recommendations for legend design [12]. Fig. 7 highlights some of the problems with using a national classifier such as OAC for regions of a country. Leicestershire County Council manage services in a rural county which *excludes* the urban centre governed by Leicester City Council). This split in local government between rural and urban is now a relatively common arrangement in the UK. Subsequently, super-groups do not stratify the population very effectively – about half of Leicestershire is 'Prospering Suburbs' (Fig. 7) – limiting its effectiveness when used within the county (section 5.5). Such geographic heterogeneities affect
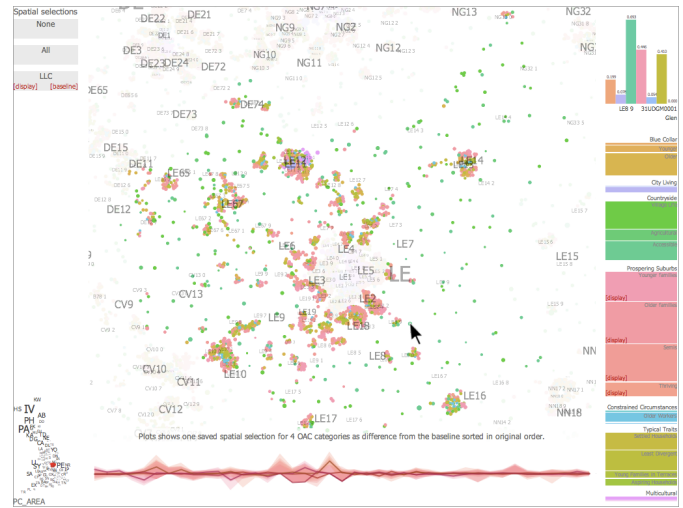

Fig. 7. A dot-map of Leicestershire County Council's OAs. Leicestershire is a rural county with clusters of OAs that represent towns and villages. Almost half the population is in the 'Prospering Suburbs' supergroup with three super-groups barely represented (right). The profiles of the four constituent groups making up 'Prospering Suburbs' are shown with respect to their parent super-group for Leicestershire (baseline).

the discriminating quality of national geodemographic classes when used locally (another example is 'Multicultural' in North London; Fig. 8). This can be addressed for local areas by tailoring existing geodemographic classifiers (e.g. [41]) or constructing specialised classifiers (e.g. [39]), but makes comparison outside this area more difficult.

4.4 Classification uncertainty

In common with other geodemographic classifiers, OAC allocates to each OA the category that best characterises it. Unlike other classifiers, the statistical distances between each OA to each super-group cluster centre are published [51, 43]. These can be considered as measures of similarity [51]. We normalise these to the most dissimilar super-group [57, 56] to give a *relative* measure of similarity and show this for each OA as a bar chart (Fig. 3D) where bar heights represent relative similarity. Where one bar dominates, the OA is more similar to that super-group than any other; where bars are the same heights, the OA is equally similar to all the super-groups. The OA in Fig. 8 is on the brink of being allocated 'City Living' or 'Multicultural', a situation that is widespread across large parts of London. This is one of the reasons Peterson *et al* [39] constructed a classifier for London using values standardised to the population of London.

In Fig. 5, we show the similarity of OAs to their *most similar* category. Significant similarities and differences between areas and categories are apparent. For example, OAs classified as 'Multicultural'
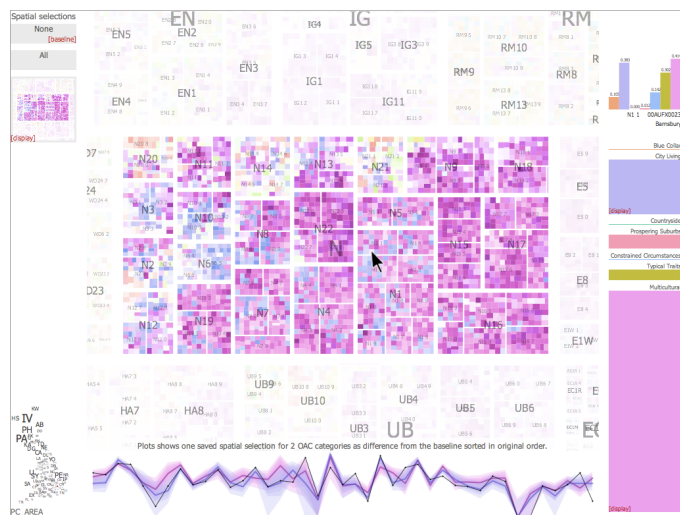
Fig. 8. The N postcode area in North London has been selected (other OAs are faded out) and this saved selection (see top left) is being displayed (see left-aligned [display] label). The legend indicates over half the people in N are in OAs classified as 'Multicultural'. The barchart shows that the OA indicated by the mouse cursor is strongly similar to both 'City Living' and 'Multicultural', a common situation London. These two population profiles are shown (see left-aligned [display] labels in the legend) for this spatial selection (N) along with the OA identified with the mouse (thin black line), showing similar variables values except the rightmost variable (wholesale/retail trade employment), identifiable with a tooltip in the software. *See Fig. 3 for help with interpretation.*

tend to be strongly 'Multicultural' in the south east, but less so elsewhere. But it is also interesting to show maps of the similarity of each OA to each super-group *separately* as shown in Fig. 6. OAs similar to 'Countryside' have a similar geographical distribution to 'Prospering Suburbs', with very strong or very weak levels of similarity. Those with strong similarity have an inverse geographical distribution to 'Multicultural' and 'City Living' OA. The similarity of OAs to 'Typical Traits' is strong for all OAs. This consistent and systematic pattern of similarity and dissimilarity is likely to have an impact for the use of OAC, particularly within constrained geographical areas – for example in local government. There is also scope for exploring these with other measures of classification uncertainty [18].

4.5 Profiles and demographic variables

Studying the values of the demographic variables that drive geodemographic classification can help evaluate the use of geodemographics in particular contexts.

The values of the 41 census variables used in the original classification can be used to profile individual OAs or groups of OAs. We use the transformed and standardised versions employed in the construction of OAC [51] and present these as an interactive parallel coordinate plot [26], a widely-used technique for showing multivariate socio-economic geographic information [13, 10, 1, 11, 44, 14, 45, 7] with increasingly sophisticated symbolism, interaction and functionality. We summarise the distributions of variables in Fig. 1, using thick coloured lines to show the median for OAs and shading with linearly decreasing lightness around the median for the 1st to 9th deciles (similar to Fua *et al.*'s [19] technique). Variables/axes in all figures are ordered using the original order (unless stated otherwise) where census variables are grouped by type as in Fig. 1. These graphs reveal that profiles are remarkably similar. Variation for many of the variables is relatively low, suggesting that OAC is driven by the few key differentiating variables labelled in Fig. 1. Note that median lines do not show real profiles, rather the set of medians for each variable for a selected set of OAs. An individual OA can be identified by moving the mouse over it and its profile superimposed onto the parallel coordinates, fa-
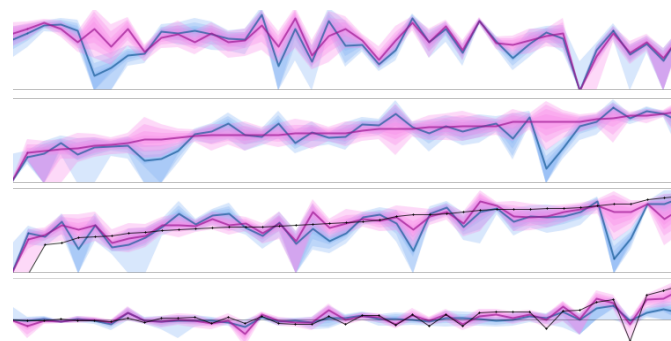

Fig. 9. Alternative variable/axis orders for the profile in Fig. 10 (right). *From top to bottom: (a)* In original order (as in all other figures; see Fig. 1). *(b)* In order of median variable values for 'Multicultural'. Left-most (lowest) variable is fishing/agriculture employment, right-most (highest) is terraced housing. *(c)* In order of the values of one particular 'Multicultural' OA. *(d)* With a base set to 'Constrained by circumstances' for the G (Glasgow) postcode area, ordered by absolute deviation from the baseline. Variables with the highest difference from G relate to ethnicity.

cilitating comparison. This is shown as a thin black line in Fig. 3E, where the characteristics of the identified OA are close to both of the displayed super-groups for different sections along the profile. Labels are not displayed to reduce clutter, but can be turned on in the software and through mouseover interactions.

Axes can be sorted, helping to rank the significance of variables in terms of median values, variable values for the chosen OA or the differences in either of these from the baseline (Fig. 9).

4.6 Comparison

Comparison is key to visual analysis and underlies many of the software requirements for exploratory data analysis [49]. Our software facilitates geographical selection and enables such selections to be saved and re-used, offering four comparison options.

Switching back and forth. In Fig. 10 our saved selections of the Glasgow (G) and Birmingham (B) postcode areas enable users to switch back and forth for visual comparison. This is cumbersome, becoming more so as the number of selections for comparison increases.

Using the same coordinate space. Plotting the elements for comparison in the same coordinate space can be effective. We employ this for showing multiple population profiles concurrently (e.g. Fig. 1, bottom). Occlusion is an obvious problem for many concurrently displayed elements, but there are situations where it works well.

Side-by-side comparison. Conventional maps can be problematic for comparing geographically distant areas. In Fig. 10 we take the areas of interest and position those of interest spatially [56], hiding all others. This allows us to compare areas of interest in detail.

Showing differences. More direct comparison is afforded by plotting differences. In Fig. 10 (right), we *set a baseline* and plot the difference from this to all the data we wish to compare this with. The baseline can be set for both a spatial selection and an OAC category. The software makes it easy to do this, using right-aligned [baseline] labels to indicate this. In Figure 10 (right), the baseline has been set to the 'Constrained by Circumstances' OAs in the G (Glasgow) postcode area. The difference between this and the displayed selections (indicated by left-aligned [display] labels in the figure; Birmingham's 'Multicultural' and 'Constrained by Circumstances' populations in this case) are shown in the profile. This allows some quite complex geographical and attribute comparisons to be made. The first large peak indicates higher ethnic minority populations in B than G. The second large peak indicates differences in housing rent and housing type. Other variables with higher values for 'Multicultural' relate to education. These graphs help identify key population differences between areas and geodemographic categories. In this example, we may begin to wonder how useful and appropriate it is to differentiate between groups of people on the basis of ethnicity and housing type.

Fig. 10. Comparing populations of the Glasgow (G; left) and Birmingham (B; right) postcode areas. Other postcode areas have been hidden. Category similarity is not shown here. *Left:* The G postcode area is selected and other OAs faded out (see left-aligned [display] label in spatial selections C). Its centre is dominated by 'Constrained by Circumstances' OAs. The population profile shows only this category for G (see left-aligned [display] label in legend C). *Right:* The B postcode area's centre is dominated by 'Multicultural' OAs. Notice the *right*-aligned [baseline] label in the spatial selections (F) and legend (C), that *sets the baseline* to G's 'Constrained by Circumstances'. The profile now shows the *difference* between the baseline and B's 'Multicultural' and 'Constrained by Circumstances' population. The plot shows some similarities and differences as the deviation from the central line. *The blue letters refer to the elements of the display labelled in Fig. 3, which help with interpretation.*

In certain cases this may be appropriate, in others it may not be. Understanding the composition of the classifier and what is lost through generalisation that results (Fisher's "artefact of the map production process" [17]), underlies any such consideration.

5 IMPACT ON USERS

Classification uncertainty, categorical variation and geographical variation in OAC are likely to have implications for its use. We wanted to find the impact of providing expert OAC users with the means to explore variation and uncertainty in OAC. Specifically, we wanted to see whether it improves their understanding of OAC and the nature of the population, whether this would influence their use of OAC in future and whether they would routinely use similar techniques to support their use of OAC in future.

We ran an evaluation session to find this out, targeting six sophisticated users of OAC with practical experience of using it in the workplace (mainly local government). One participant from the Greater London Authority (GLA) is a statistical analyst, a member of the OAC User Group (which promotes the use of OAC across the public and private sectors) and manages census analysis software [29] which includes OAC in some of its outputs. We had hoped to involve other users from the GLA, but found that their use of geodemographics was quite low. Another participant is an academic with OAC experience in the private and academic sectors and is a member of the OAC User Group. As he is no longer actively involved in information provision to support policy-making in local government, the scope of the evaluation with him was more limited. The four remaining users were analysts from Leicestershire County Council (LCC), who use OAC to code survey data that they collect so that they can be correlated across datasets and in the reporting statistics by geodemographic group. As we saw in Fig. 7, there is an imbalance in the way in which the county's population is allocated, with about half the population allocated to 'Prospering Suburbs'. For this reason, LCC is in the process of producing a tailored version of OAC for the county by aggregating OAC's sub-groups into new categories whose populations are more balanced. This is different from Peterson *et al.*'s [39] approach of reclustering from reweighted variables, but has the advantage that it still corresponds to the national OAC, making inter-region comparison more straightforward.

As sophisticated users, participants knew OAC well and already had an understanding and anecdotal evidence of the type of variation in OAC. We used our software to provide them with the means to explore OAC using open questions to prompt discussion which was recorded. We found that capturing the richness of the open discussion was appropriate given the small number of users and the complexity and subtlety of the issues involved.

Key to the validity of the evaluation is the participants' clear understanding of the visual encodings used and how to interpret these. We ensured that this was the case by demonstrating the techniques and subsequently 'chauffeuring' participants though their exploration and investigation. This significantly reduced the barrier that unfamiliar user-interfaces impose on software use. When users articulated an analyst task (such as the complex one described in section 4.6), it was relatively straightforward to chauffeur them through the use of the software to study this. We also asked for feedback on the specific interactions and visual encodings offered by the software.

The evaluation had two parts. In the first part, we presented the techniques and demonstrated how they could be used for characterising variation and uncertainty in OAC. We used a similar structure as section 4, starting with simpler types of comparison and hiding aspects of the display yet to be introduced, and then introducing the visual encodings and interactions to support more complex and subtle types of exploration. We expected that some of the variation presented would be well-known to participants. This was deliberate, as confirming 'knowns' is an important step in gaining confidence in using visualization [40]. In the case of Leicestershire County Council, where four analysts took part, this was run as a group seminar where participants were encouraged to ask questions, offer their perspectives and discuss the data and methods. This enabled us to gauge whether the techniques being presented were understood and the extent to which the characteristics of OAC seen in the demonstration matched their experience.

In the second part, participants were evaluated individually whilst being 'chauffeured' through their own lines of enquiry. This is appropriate for testing the interaction and visualization methods themselves rather than the software usability [21]. Free exploration was encouraged, but participants were asked to focus on areas they knew well, characterise these and then compare them with other known areas. The intention was to make the tasks as unconstrained, relevant and grounded in areas of participant knowledge as possible, so that partic-

ipants could objectively reflect on how the techniques might help them in future. The chauffeuring process ensured that the participants had to think-aloud, articulating what they wanted to find out and why. Since the evaluation was carried out by one individual, audio dialogue was recorded and notes made subsequently. This was time-consuming, but the rich dialogue captured in the recording was a valuable resource for subsequent reflection and reporting. Responses were not coded due to the small number of participants and their individually themed exploration. Once users had finished exploring, questions were used to ensure that users reflected on the aspects they were evaluating. Since questions were answered verbally, answers were richer than if participants had been asked to write down the answers as a list. Responses are reported later in this section.

1. *What did you learn that was useful about OAC?*
2. *What did you learn that was useful about the population?*
3. *Which of the functionality (graphics and interactions) did you find most useful, and why?*
4. *Might any of these findings influence the way you or you organisation use, promote or describe OAC?*
5. *Would you use these techniques routinely to support your job?*

5.1 Understanding of OAC

As expected, participants knew OAC well, were aware of the inherent within-class variation (*"we always knew that OAC had variation in it"*) and knew about the imbalance in classified population for their local areas (*"some people get the idea that OAC is no good for LCC"*). Users felt reassured that patterns they knew or suspected existed were confirmed. However, the *degree* of classification uncertainty and the heterogeneity of adjacent OAs of the same category surprised participants (*"you begin to question things you took for granted"*). The large number of OAs that showed strong similarities with other categories raised questions concerning the validity of decision-making based upon closest groups alone (*"some categories appear to be just as valid as others"*) and there was surprise at the extent to which OAC appeared to be driven by just a few key variables.

5.2 Understanding of population

All participants explored their home areas and found that the OAC classification and demographic data matched their expectations and understanding of their local area (*"fits in very nicely with how I imagined it"*). A participant who grew up in Cambridge (CM), found that small collections of OAs with low similarity to their classified groups corresponded to the student accommodation areas. Comparison between these and similar areas of student population in their university town of Loughborough proved interesting. One individual was struck by how the dot-maps Fg. 4 revealed apparent local population segregation, with pockets of differentiated population types. Reasons for these patterns were speculated upon (e.g. corresponding to historically industrial parts of a city or pockets of student population) and the demographic profiles of these areas were studied.

5.3 Visual techniques

We did not use a standard set of colours [35] as we wanted to make better use of colour-space. This did mean that users were initially unfamiliar with the coding used. They were able to adapt, but the importance of visual familiarity in encoding and the ability for users to learn were highlighted.

The rectangular cartogram was not immediately intuitive to participants (*"takes time to get your head round this projection"*), but all were able to interpret after explanation and with the help of the dynamically linked overview map with highlighting (Fig. 3B), the postcode labels and area names. The lack of guaranteed adjacency between zones was considered to be problematic by some, but it was considered more effective for studying trends over large geographical areas. Where absolute geographical positioning of OAs was required, participants were able to use the dot map (e.g. Fig. 7), but as one participant later commented, most meaningful questions of the data were at a coarser geographical scale that that of individual OAs.

Most participants had not seen the similarity information presented before. Since we presented the published uncertainty data directly without further transformation (except by normalising against the least similar), users quickly understood what the data showed and were confident in their interpretation of it. They found the barchart effective (*"it demonstrates there's quite a lot of variation – key for me is the barchart"*), particularly for OAs that were strongly similar to multiple categories (e.g. Fig. 8). A number of participants wanted to see this information summarised for coarser geographies – one in particular was critical of the emphasis on individual OAs, saying that *in vivo* most questions would start with a profile or region rather than a specific OA. This could be addressed by providing spatial selections that correspond to the coarser geographies relevant to the users. Participants liked the use of lightness to indicate similarity (*"lightness is a clever way to show similarity"*). They particularly liked the single super-group maps (Fig. 6) because they had not looked at the data in this way and could see its utility. One participant suggested it might be useful to produce maps of the second and third most typical supergroups in any area. Interest was expressed for seeing composite versions of these maps, e.g. the degree to which a region was generally typical of both 'City Living' and 'Multicultural'.

The parallel coordinates plots that summarised variation in the demographic variables for selected areas by super-group and group as in Fig. 10, were universally liked. Participants found depicting variation as shading around the median to be intuitive. Those working at LCC wanted access to this kind of view more regularly, but do not have the means to do so in a systematic way. The use of a baseline took "*some getting one's head around*", but such direct comparison was found to be useful once this had been understood.

One participant was concerned about the fact that no indication of sample sizes or statistical significance was given, something they regard as essential for interpreting their data and reporting their findings. For example, it was perfectly possible to generate decile representations of variation on ten OAs representing a tiny proportion of the population with no indication of this being the case. It was suggested that the statistical significance of sample sizes and differences should be encoded in these views, something we plan to implement in future.

5.4 Impact

All participants were positive about their experience of exploring OAC and the demographic variables, reporting that they found it interesting and useful (*"need to shatter the illusion that categories are the same"*; *"really useful to get a sense of how much variation there is"*; *"we've had strange results using OAC is some areas, and this would help see why"*; *"lays bare the frailties of the technique – highlights that such issues probably exist in other schemes"*).

The members of LCC felt strongly about using an open classifier and believed that it was important to make the most of this (*"so we can be more sure that decisions are more transparent"*). They recognised the importance of not taking such classifiers at face value and believed that doing regular "sense-checking" was important (*"we should never rely completely on OAC"*), a conviction reinforced during this evaluation exercise.

There was widespread concern about allowing inexperienced users access to these techniques (*"we don't want to undermine people's confidence in OAC"* and *"people at 'end of chain' need it to be simple and understandable."*). We were surprised at how strongly-held this view was, but it resonates with the experience that "policy-makers typically want issues presented with no ambiguity" [31]. However, participants were unanimous in their view that these techniques could be enormously valuable to experienced users of OAC, such as themselves. As one participant noted *"[these techniques] might either undermine the whole basis of geodemographics. . . or put it into a proper context"*.

5.5 Potential future use in the workplace

This question was only asked of the LCC group, as they were the only participants who used OAC regularly in the workplace. The free exploration we encouraged was directed towards enabling participants to identify the kind of tasks they wanted to do and evaluate whether

the interactions and graphics adequately supported these tasks. Three tasks emerged: *sense-checking*; *comparison with other data sets and alternative classifiers*; *building a tailored OAC*.

5.5.1 "Sense-checking"

The degree of geographic and categorical variation in OAC highlighted the importance of understanding what impact this might have in various contexts. The lack of methods that allowed this to be achieved easily was reported as a barrier to this kind of activity. Participants were impressed with the speed and ease with which comparisons could be made using the methods shown in our application and wanted to use them on a regular basis to check the classifications that they were using. They understood that the software in this participatory activity was a prototype developed to propose and evaluate generic methods in a research context, but the methods were so successful that participants were keen to have a copy of the software to try out.

5.5.2 Comparison with other data and alternative classifiers

Several participants expressed an interest in *"show[ing] other data as well as OAC variables"* and comparing with other classifiers. Although non-open classifiers preclude the study of the variables that drive them, comparing the population discriminating potential was considered important. This related to a recent exercise undertaken at LCC where the discriminating potential of OAC and a commercial alternative were compared for reporting the results of a crime survey, taking into consideration the key variables for each. The study led to a decision about which was used in a published report. This 'one-off' piece of work, was time-consuming and used relatively *ad hoc* methods but was something they *"wanted to do more of"*.

5.5.3 Building a tailored OAC

The participants are in the process of designing a tailored version of OAC that better stratifies the LCC population. They are doing this by aggregating OAC's sub-groups to form new categories. They see these techniques as *"useful to help with the classification of those low levels [sub-groups] and hopefully build a more meaningful OAC because of it"*. 'Prospering Suburbs' is a super-group of particular interest. One participant wanted to be able to find the *"next most similar sub-group"* to help identify how sub-groups should be best aggregated into new tailored groups. Fig. 7 compares groups against the super-group baseline. We had not anticipated the need to include sub-groups but sub-groups could also be compared in this way.

6 CONCLUSION

OAC is a free and open geodemographic classifier designed to help characterise population, stratify population samples and target products or services. It is typical of a number of alternatives. As with any classification, whilst it is valuable in terms of the generalisation that is involved, it contains an inevitable degree of uncertainty. Importantly, this varies geographically and by population type, often in a systematic manner. We have found that users of OAC find it useful to study this variability and that it has the potential to help them assess its impact in their particular use-cases. The interactive graphical designs developed here were found to be useful means of supporting this activity, however, this finding may not apply to inexperienced users.

The first aim of this work was to design and demonstrate interactive visualization techniques to help identify the variables that drive the OAC classifier, to characterise variability and uncertainty in OAC and to show how this varies geographically and by OAC category. Our design combines a variety of techniques including dot maps, rectangular hierarchical cartograms, barcharts, parallel coordinates plots, perceptual colour schemes and means to select and compare geographical regions, OAC categories and multiple demographic variable values. It does so in the context of a national data set described through over 200,000 small areas. The parallel coordinate plots summarise geographical and categorical selections of population, through the median and decile values of 41 socio-economic variables. These can be compared to population profiles of particular OAs or classes by reconfiguring the graphics on demand. The examples and user reactions sug-

gest that these techniques fulfil this aim to a large degree. They also confirm the feasibility of using interactive visualization on standard desktop computers to delve into multivariate classification schemes as applied to large numbers of areas and explore the underlying data.

The second aim of this work was to assess the impact of these techniques on analysts. We found strong evidence that providing experienced users of OAC with these exploratory techniques increased their understanding of how OAC works, helped them connect OAC more strongly to the underlying population and had the potential to support their ongoing use of OAC in the workplace. Although the analysts anticipated many of the issues we explored, they were surprised at the degree to which heterogeneity existed in data and areas that they knew well. The main impact of this exercise was that participants in our study felt that they had a better understanding of OAC and its uncertainties. They also had a desire to study these effects regularly using interactive graphics. They enjoyed exploring the data using our methods, regarded them as interesting and thought that systems providing this kind of functionality would help them and other sophisticated users make more informed use of OAC. They were cautious, however, about providing access to this kind of information to individuals inexperienced with OAC or geodemographics, believing that exposing the uncertainty associated with the classification had the potential to undermine confidence in its use. Communicating the uncertainties inherent in abstractions derived from multivariate geographic information to a less specialist audience is a challenge to which visualization may usefully be applied in the future.

The evaluation design was appropriate for the small number of expert users and it fulfilled our aim. The *chauffering* approach enabled users to evaluate the techniques and their potential impacts without being hindered by an unfamiliar user-interface. This, along with a *clean interface* uncluttered with controls/options allowed users to concentrate on the data and tasks they wanted to perform. Capturing *verbal* answers through discussion and thinking-aloud, provided rich and detailed responses appropriate for the complexity and subtlety of the issues involved. *Free exploration* helped the OAC experts apply these issues to their own work. They were able to identify three types of task they would like to do regularly to support their work and suggested further useful functionality to support these. The evaluation results help validate our initial assertion that sophisticated users of OAC benefit from being able to explore the classifier itself, the variation and uncertainty that exists within, how these vary geographically and by category and the variables that drive the classifier. This will certainly apply to other open geodemographic classifiers around the world and is very likely to apply more generally in other contexts, such as remote sensing, land-use classification and biodiversity research.

REFERENCES

[1] G. Andrienko and N. Andrienko. Interactive maps for visual data exploration. *Int. J. Geog. Inf. Science (IJGIS)*, 13(4):355–374, 1999.

[2] G. Andrienko, N. Andrienko, J. Dykes, S. Fabrikant, and M. Wachowicz. Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research. *Information Visualization*, 7(3-4):173–180, 2008.

[3] C. Booth. *Labour and Life of the People: Descriptive Map of East End Poverty*. Macmillan, 1889.

[4] C. Brewer. Color use guidelines for mapping and visualization. In A. MacEachren and D. Taylor, editors, *Visualization in modern cartography*, volume 2, chapter 7, pages 123–148. Elsevier Science, Tarrytown, NY, USA, 1994.

[5] CACI. ACORN classification website. http://www.caci.co.uk/acorn-classification.aspx, 2011.

[6] N. Cawthon and A. V. Moere. The effect of aesthetic on the usability of data visualization. In *Proceedings of the 11th International Conference*

Information Visualization, page 637–648, Washington, DC, USA, 2007. IEEE Computer Society. ACM ID: 1271539.

[7] R. Chang, G. Wessel, R. Kosara, E. Sauda, and W. Ribarsky. Legible cities: Focus-Dependent Multi-Resolution visualization of urban relationships, 2007.

[8] D. Dorling. *Area Cartograms: Their Use and Creation*. Concepts and Techniques in Modern Geography. University of East Anglia, 1996.

[9] D. Dorling, A. Barford, and M. Newman. Worldmapper: The world as you've never seen it before. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):757–764, 2006.

[10] J. Dykes. Cartographic visualization: Exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv. *The Statistician*, 47(3):485–497, 1998.

[11] J. Dykes, D. Dorling, and J. Carter. Visualizing census data. In P. Rees, D. Martin, and P. Williamson, editors, *The Census Data System*, pages 97–110. Wiley, Chichester, 2002.

[12] J. Dykes, J. Wood, and A. Slingsby. Rethinking map legends with visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16:890–899, Nov. 2010. ACM ID: 1907962.

[13] J. A. Dykes. Exploring spatial data representation with dynamic graphics. *Computers & Geosciences*, 23(4):345–370, May 1997.

[14] R. Edsall. The parallel coordinate plot in action: design and use for geographic visualization. *Computational Statistics & Data Analysis*, 43(4):605–619, 2003.

[15] S. Egbert and T. A. Slocum. EXPLOREMAP: an exploration system for choropleth maps. *Annals of the Association of American Geographers*, 82(2):275–288, 1992.

[16] Experian. Mosaic UK website. http://www.experian.co.uk/business-strategies/mosaic-uk-2009.html, 2011.

[17] P. Fisher. Is GIS hidebound by the legacy of cartography? *Cartographic Journal, The*, 35(1):5–9, 1998.

[18] P. Fisher. Uncertainty in the 2001 Output Area Classification for the census of England and Wales. In *Proceedings of the 17th Annual GIS Research UK Conference*, pages 161–166, 2010.

[19] Y. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *Proceedings of the conference on Visualization '99*, VIS '99, page 43–50, Los Alamitos, CA, USA, 1999. IEEE Computer Society Press. ACM ID: 319355.

[20] M. Gahegan and M. Ehlers. A framework for the modelling of uncertainty between remote sensing and geographic information systems. *IS-PRS Journal of Photogrammetry and Remote Sensing*, 55(3):176–188, 2000.

[21] S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 111–120, New York, 2008. ACM.

[22] R. Harris, P. Sleight, and R. Webber. *Geodemographics, GIS and neighbourhood targeting*. John Wiley and Sons, Feb. 2005.

[23] J. Haslett, G. Wills, and A. Unwin. SPIDER – an interactive statistical tool for the analysis of spatially distributed data. *International Journal of Geographical Information Science*, 4(3):285–296, 1990.

[24] J. Heer. Voyagers and voyeurs: Supporting asynchronous collaborative information visualization. *Proceedings of CHI'07*, pages 1029—1038, 2007.

[25] H. Hofmann and M. Theus. Selection sequences in MANET. *Computational Statistics*, 13(1):77–88, 1998.

[26] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[27] M. Jern. Collaborative web-enabled geoanalytics applied to OECD regional data. In *Proceedings of the 6th international conference on Cooperative design, visualization, and engineering*, CDVE'09, page 32–43, Berlin, Heidelberg, 2009. Springer-Verlag. ACM ID: 1812988.

[28] N. Kong, J. Heer, and M. Agrawala. Perceptual Guidelines for Creating Rectangular Treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):990–998, 2010.

[29] Local Government Association. SASPAC: About. http://saspac.wordpress.com/about-2/, 2011.

[30] A. Lucieer and M. Kraak. Interactive and visual fuzzy classification of remotely sensed imagery for exploration of uncertainty. *International Journal of Geographical Information Science*, 18(5):491–512, 2004.

[31] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32:139–160, July 2005.

[32] S. Marsh and J. Dykes. Using usability to evaluate geovisualization for learning and teaching. In *Proceedings of the 13th Annual GIS Research UK Conference*, 2005.

[33] D. Martin. Geography for the 2001 census in England and Wales. *Population Trends*, 108:7–15, 2002.

[34] G. Milligan. Clustering validation: results and implications for applied analyses. In P. Arabie, L. J. Hubert, and G. de Soete, editors, *Clustering and classification*. World Scientific, 1996.

[35] OAC User Group. Standard colours for OAC. http://areaclassification.org.uk/2010/09/07/standard-colours-for-oac/, 2010.

[36] Office of National Statistics. National Statistics 2001 Area Classification website. http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/, 2008.

[37] Office of National Statistics. National Statistics 2011 Area Classification. http://ur.ly/HHa0, 2011.

[38] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.

[39] J. Petersen, M. Gibin, P. Longley, P. Mateos, P. Atkinson, and D. Ashby. Geodemographics as a tool for targeting neighbourhoods in public health campaigns. *Journal of Geographical Systems*, 2010.

[40] A. C. Robinson. A design framework exploratory geovisualization in epidemiology. *Information Visualization*, 6:197–214, Dec. 2007. ACM ID: 1375942.

[41] A. D. Singleton and P. Longley. Creating open source geodemographics: Refining a national classification of census output areas for applications in higher education. *Papers in Regional Science*, 88(3):643–666, 2009.

[42] A. Slingsby, J. Dykes, and J. Wood. Rectangular hierarchical cartograms for Socio-Economic data. *Journal of Maps*, pages 330–345, 2010.

[43] Social and Spatial Inequalities Group, Sheffield University. The National Classification of Census Output Areas website. http://www.sasi.group.shef.ac.uk/area_classification, 2011.

[44] M. Takatsuka and M. Gahegan. GeoVISTA Studio: a codeless visual programming environment for geoscientific data analysis and visualization* 1. *Computers & Geosciences*, 28(10):1131–1144, 2002.

[45] M. Theus. Interactive data visualization using Mondrian. *Journal of Statistical Software*, 7(11):1–9, 2003.

[46] J. Thomas and K. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.

[47] J. Thomson, B. Hetzler, A. M. MacEachren, M. Gahegan, and M. Pavel. A typology for visualizing uncertainty. In *Conference on Visualization and Data Analysis*, 2005.

[48] W. Tobler. Thirty five years of computer cartograms ground-truthing geodemographics. *Annals of the Association of American Geographers*, 94(1):58–73, 2004.

[49] A. Unwin. Requirements for interactive graphics software for exploratory data analysis. *Computational Statistics*, 14(1):7–22, 1999.

[50] D. Vickers and J. Pritchard. Visualising the output area classification. *Journal of Maps*, pages 410–416, 2010.

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

[52] C. Weaver. Building highly-coordinated visualizations in improvise. In *10th IEEE Symposium on Information Visualization (InfoVis 2004)*, pages 159–166. IEEE Computer Society, 2004.

[53] M. Wijffelaars, R. Vliegen, J. J. van Wijk, and E. van der Linden. Generating color palettes using intuitive parameters. *Computer Graphics Forum*, 27(3):743–750, 2008.

[54] G. Wills. Selection: 524,288 ways to say ïhis is interesting. In *2nd IEEE Symposium on Information Visualization (InfoVis 1996)*, pages 54–61. IEEE Computer Society, 1996.

[55] J. Wood and J. Dykes. Spatially ordered treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1348–1355, 2008.

[56] J. Wood, A. Slingsby, and J. Dykes. Layout and colour transformations for visualising OAC data. In *Proceedings of the 17th Annual GIS Research UK Conference*, pages 455–462, 2010.

[57] J. Wood, A. Slingsby, and J. Dykes. OAC Explorer: Interactive exploration and comparison of multivariate socioeconomic population characteristics. In *Proceedings of the 17th Annual GIS Research UK Conference*, pages 167–174, 2010.

[58] J. Yi, Y. ah Kang, J. Stasko, and J. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007.