Characterizing Volunteered Geographic Information using Fuzzy Clustering

S. De Sabbata[1], N.J. Tate[1], C. Jarvis[1]

[1]Department of Geography, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom
Email: s.desabbata@le.ac.uk, njt9@le.ac.uk, chj2@le.ac.uk

Abstract

This paper demonstrates the use of fuzzy clustering to characterize Volunteered Geographic Information (VGI). We argue that classifying small areas based on variables related to the amount, type, and currency of VGI can provide a more nuanced understanding of the content. We present a classification of 2011 UK Census Output Areas in Leicestershire (UK) based on content of OpenStreetMap, using a fuzzy *c*-means clustering algorithm, and we compare the resulting classification with a 'standard' socio-economic geodemographic classification.

1. Introduction

The quality of Volunteered Geographic Information (VGI) has long been a focus of research in GIScience (e.g., Haklay, 2010; Goodchild and Li, 2012; Barron et al, 2014). At the same time related questions have been raised concerning the lineage of VGI: who contributes, who is represented and who is not (e.g., Stephens, 2013; Wilson and Graham, 2013; Glasze and Perkins, 2015; Sieber and Haklay 2015). It is evident that a bias exists in VGI, as the majority of content producers seem to be composed of relatively wealthy, younger, western, tech-savvy, male users, and the interests and knowledge of this particular demographic is thus reflected in the produced content. The study of information geographies (Graham et al., 2015) focuses on how the underlying geographies of wealth and access to technology impact the geographic distribution of participation, and in turn the geographies of representation. For instance, Mashhadi et al. (2015) illustrate how population density, wealth, and centrality influence the completeness of OpenStreetMap (OSM) in London, UK. However, as high-quality authoritative benchmarks are not always available, stand-alone approaches based on data mining are necessary to further our understanding of VGI (Senaratne et al., 2016).

The aim of this paper is to characterize VGI through data mining, and explore its relationship to socio-economic variables. We use a fuzzy clustering method for the classification of 2011 UK Census Output Areas (OAs) based on OSM content, and then compare this to a 'stamdard' geodemographic classification: 2011 Output Area Classification (2011OAC) (O'Brien and Cheshire, 2016). We suggest that this process provides a) a more complete and nuanced understanding of OSM content compared to simple density maps and b) it can be the basis for further studies of information geographies, affording both qualitative and quantitative comparisons with socio-economic information, such as census data.

1.1 Numerical clustering, classification, and geodemographics

The objective of a geodemographic analysis (see e.g., Harris et al., 2005; Alexiou and Singleton, 2015) is to classify neighborhoods based on a basket of socio-economic variables, through a process of numerical clustering. Widely used both in social studies and marketing, these classifications are commonly based on census data or surveys, and are constructed through clustering methods such as hard *k*-means (e.g., Singleton and Longley, 2015), or fuzzy *c*-means (e.g., Fisher and Tate, 2015). Longley and Adnan (2016) recently applied these clustering methods to data derived from social media.

2. Methods

Our aim is to characterize VGI for further analysis and comparison, rather than conduct a 'standard' quality assessment, and thus we operate without a benchmark. Although the variables described below do not directly relate to common quality measures, the variables used in this study still aim to capture elements of the completeness, temporal accuracy, and thematic accuracy of the data. The county of Leicestershire (UK) was the focus of our case study. The Planet.osm file was downloaded on March 16th, 2016.

We first computed aggregated values per OA: counts per feature type (i.e., amenity, highway, etc.); total number of features; average number of edits (based on version numbers); the timestamp of the changeset related to the oldest and latest edited object. We then computed z-scores for the total number of features, the average number of edits, and the number of days between the two timestamps and the time of download. As most OAs contain no features of most types, the counts per feature types have been normalized simply as percentages over the total number of features, as a z-score would be skewed by the large amount of zero values.

The clustering procedure was based on 18 variables: z-scores based on the number of features, average edits, time since oldest edited object and latest edited object, and 14 feature types percentages. The percentage of *building* feature type was excluded, due to high negative correlation value with the *highway* feature type (Pearson's $r=-0.79$, $t(3052)=-71.1$, $p<.001$).

Fuzzy clustering was performed using the *cmeans* function of the e1071 R library. We assigned weights to variables, aiming to highlight the amount, type of content, and edits: 0.25 to both number of features and average edits; 0.125 to both time since oldest and latest edited object; 0.25 equally split among the different feature type counts. Based on the analysis of the within-cluster sum of squares, we selected the target number of clusters to be mined to be six.

3. Results and discussion

Among the six classes identified in the classification (OA hard membership based on highest membership value, see Figure 1), one clearly characterizes the city centres ("Central"), with the highest total features density and low *highway* percentages, and another characterizes the countryside ("Countryside") with the lowest total feature density and high percentages of *natural* features. Two classes group most residential areas, one seemly less up-to-date ("Residential A") than the other ("Residential B").

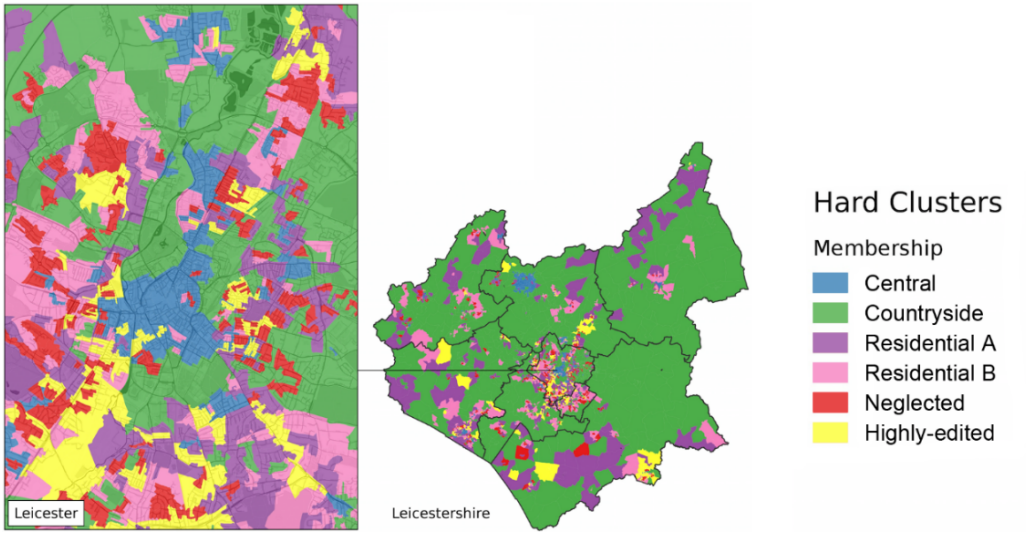

Figure 1. Classification of 2011 OAs based on OSM content. (Contains National Statistics data © Crown copyright and database right, 2016. Contains OS data © Crown copyright and database right, 2016. © OpenStreetMap contributors).

Two other classes are of particular interest in assessing the quality of OSM content. The first is "Neglected", illustrated in Figure 2 (top). The centre of this class has the lowest value of feature density beside the "Countryside" class, and the highest value for the variable related to the time since the last edit has been made. Members of this class are areas which have seen little editing and contain few points of interest. Preliminary analysis suggests that there is a significant association between being classified as "Hard-Pressed Living" in the 2011OAC and being classified as "Neglected" ($\chi^2=15.45$, $p<.001$), with the odds of an area being 'neglected' on OSM being 1.63 (1.26, 2.10) higher if classified as "Hard-Pressed Living".

The second is "Highly-edited", illustrated in Figure 2 (bottom). The centre of this class has a particularly high value for the variable related to the number of edits, which seems to be mostly due to highly edited highways features, or polygons. There seems to be a weak association between being classified as "Cosmopolitans" in the 2011OAC and being classified as "Highly-edited" ($\chi^2=5.83$, $p<.02$), with the odds of an area being 'highly-edited' on OSM being 1.84 (1.05, 3.06) higher if classified as "Cosmopolitans".

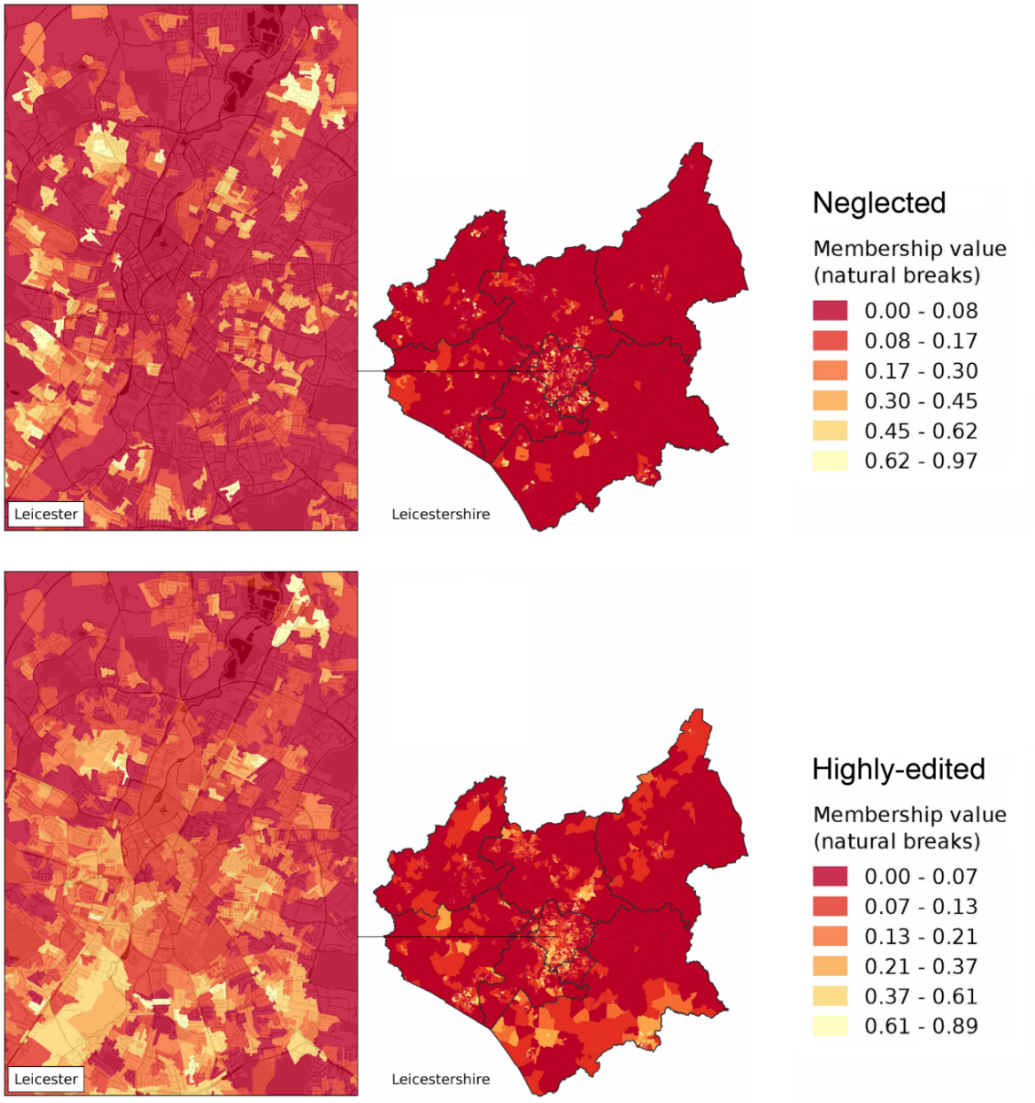

Figure 2. OA membership values for the classes "Neglected" (top) and "Highly-edited"(bottom). (Contains National Statistics data © Crown copyright and database right, 2016. Contains OS data © Crown copyright and database right, 2016. © OpenStreetMap contributors).

Interpreted in the context of the literature discussed above (e.g., Glasze and Perkins, 2015; Mashhadi et al., 2015; Sieber and Haklay 2015), the associations between the classification presented here and the 2011OAC suggest that they could be both related to underlying socio-economic geographies. Future work will focus on a more detailed analysis of the relationships between VGI and socio-economic factors (Senaratne et al., 2016).

The main disadvantages of the methods demonstrated above are the partial arbitrariness of a) the selected attributes, b) the areal units used (and related Modifiable Areal Unit Problem), and c) the interpretation of the resulting clusters. The main advantage is the flexibility of the clustering approach, which allows to create a coherent perspective on the data from a large number of diverse attributes. A broader, open classification has the potential to create a powerful tool for researcher, producers, and users for the analysis, development, and critique of single datasets (as in the present case study), as well as combining multiple VGI sources.

Acknowledgements

The 2011 UK Census Output Area (OA) boundaries and attributes were obtained via the UK Data Service, retrieved from SN:5819. http://discover.ukdataservice.ac.uk/catalogue/?sn=5819
Figure 1 and 2 use map tiles by Stamen Design, under CC BY 3.0. http://maps.stamen.com

References

Alexiou A and Singleton A, 2015, Geodemographic analysis. In: Brunsdon C, and Singleton A, (eds) *Geocomputation: A Practical Primer*, SAGE, London, 137–151.

Barron C, Neis P and Zipf A, 2014, A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS,* 18: 877–895.

Fisher P and Tate NJ, 2015, Modelling class uncertainty in the geodemographic Output Area Classification. *Environment and Planning B*, 42(3): 541–563.

Glasze G, and Perkins C, 2015, Social and political dimensions of the OpenStreetMap project: Towards a critical geographical research agenda. In: Arsanjani JJ, Zipf A, Mooney P, and Helbich M, (eds), *OpenStreetMap in GIScience: Experiences, Research, and Applications*, Springer: Heidelberg, 143–166.

Goodchild MF, and Li L, 2012, Assuring the quality of volunteered geographic information. *Spatial Statistics,* 1: 110–120.

Graham M, De Sabbata S, Zook, MA, 2015, Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information. *Geo:Geography and Environment,* 2(1): 88–105.

Haklay M, 2010, How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B* 37: 682–703.

Harris R, Sleight P and Webber R, 2005, *Geodemographics, GIS and Neighbourhood Targeting*. John Wiley and Sons, Chichester, UK.

Longley PA and Adnan PA, 2016, Geotemporal Twitter demographics. *International Journal of Geographical Information Science* 30(2): 368–389.

Mashhadi A, Quattrone G, and Capra L, 2015, The impact of society on volunteered geographic information: the case of OpenStreetMap. In: Arsanjani JJ, Zipf A, Mooney P, and Helbich M, (eds), *OpenStreetMap in GIScience: Experiences, Research, and Applications,* Springer: Heidelberg, 125–141.

O'Brien O, Cheshire J, 2016, Interactive mapping for large, open demographic data sets using familiar geographical features. *Journal of Maps* 12, 676–683.

Senaratne H, Mobasheri A, Ali AL, Capineri C, Haklay M, 2016, A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*. Forthcoming. DOI: 10.1080/13658816.2016.1189556

Sieber,RE, and Haklay M, 2015, The epistemology(s) of volunteered geographic information: a critique. *Geo: Geography and Environment*, 2(1): 122–136.

Singleton AD and Longley PA, 2015, The internal structure of Greater London: a comparison of national and regional geodemographic models. *Geo:Geography and Environment*, 2(1): 69–87.

Stephens M, 2013, Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal*, 78(6): 981–996.

Wilson MW, and Graham M, 2013, Situating Neogeography. *Environment and Planning A*, 45: 3–9.