Towards Automated Variable Selection in Geodemographic Analysis: A Case Study of New York City

Yunzhe Liu*[1], Dani Arribas-Bel[1], Guanpeng Dong[1], Alex Singleton[1]

[1]Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool, Roxby Building, Liverpool, L69 7ZT

November 30, 2016

KEYWORDS: Open Data, Geodemographic Classification, American Community Survey

1. Introduction

Sleight (1997, p16) defines geodemographics as "an analysis of people by where they live", which is configured through a clustering process that organises each area (often specified as small geographic scale, shorthanded as 'neighbourhood') into group based upon the overall similarities concealing within those multivariate attributes which they share (Singleton and Longley, 2015; Leventhal, 2016). The conceptual and theoretical origins of geodemographics can be traced back to the intellectual heritage of urban ecology study in the 1920s (Harris et al., 2005), however, methodologically are aligned to work in the late 1970s by Richard Webber (Singleton and Spielman, 2014). An important data source in the construction of geodemographic classification has been the decennial census of the population which offers an extensive range of contextual variables describing the multidimensional characteristics of each neighbourhood. However, such data are of low temporal resolution, and as such can draw criticism that any classification created from them will become dated during the inter-censal period. For this reason, and an additional aim to achieve greater differentiation between zones, commercial geodemographic classifications will often supplement or replace these data with other non-census sources. For non-commercial classifications, the availability of Open Data are providing a wider pool of attributes from which we can potentially add value to geodemographic classification (Leventhal, 2016). However, as Openshaw, Cullingford, and Gillard (1980) point out, variable selection is a subjective procedure and must reflect the major requirement of the classification. In the case of building a general-purpose classification, this would be to provide area-level profiles that are useful for a wide variety of purposes, however, to balance this range of potential applications versus attribute parsimony is a challenge.

Within this context, the aim of this paper is to produce a methodological framework that identifies suitable types of input data for the construction of a general-purpose geodemographic classification. This will concentrate on the study area of New York City, which is selected given their well-documented and abundant open data.

2. An Overview of American Community Survey (ACS) and Open Data

The American Community Survey (ACS) has replaced the long-form 2010 Decennial Census, US. Although the ACS offers similar demographic and economic data when compared to the long-form survey, the form of the survey has changed completely, and is now based on a large sample rolling survey of around 3.5 million addresses per year (Spielman and Singleton, 2015; Glenn, 2016). Although the census has a relatively high resolution, for some dynamic regions, the data can be soon out-of-date, indicating a coarser temporal resolution. The ACS, on the other hand are collected more frequently with small area estimates produced for 1-year and 5-year intervals (Alexander, 2002). The main benefit of this approach is the timeliness, however, there is uncertainty around the estimates and especially when estimating for small areas (e.g. block group level) (Spielman and Singleton, 2015).

Open data is summarily defined as data that can be freely accessed, used and re-used, modified and redistributed by anyone for any purpose (Open Definition, n.d.). Many government reformers believe that opening data is one of the key stages towards altering how people engage with governments at all levels (Pew Research Centre, 2015). A number of 'open data' and 'open government' initiatives have been developed

* psyliu7@liverpool.ac.uk

across the US, which not only attempt to utilise data as a lever to spur economic growth, but also aim at establishing a system of transparency, accountability, public participation, and collaboration so as to improve government performance, and encourage warmer citizen's attitude toward government (Gurin, 2014; Lathrop and Ruma, 2010).

Within the context of New York, over 1500 machine-readable datasets generated by various government agencies and organisations have been initiated by the New York City Council for public use since the year 2009 (Roest, 2016). Moreover, the City Council passed seven pieces of legislation to reinforce and bolster the first Open Data Law of 2012, fostering a more user-centric approach to open data. Additionally, according to the Open Data Plan 2016, more than 200 datasets are listed in the Plan and prepared to be released to the public. Most of the open datasets are available online (at https://data.cityofnewyork.us/), which is the main platform for the City of New York Open Data. The content of these data vary among eleven main categories, ranging from business, education, environment, health, public safety, recreation, and so forth. According to the Open Data Plan (2016), most of the open datasets are updated and maintained periodically, ranging from daily, monthly, quarterly, and annually.

3. Methodology

Given the plethora of data that might be used to build a geodemographic classification, we are interested here in how attributes can be selected in an automated fashion, with the objective of selecting those that will offer the greatest descrimination between small areas. For the case study area of New York, and utilising both the ACS and Open Data, we consider a range of attributes including: geographic coverage (distribution and granularity), uncertainty (margins of error), temporal resolution (refresh rate), provenance, redundancy with other variables such as sensitivity and variance. There are four main goals of this variable evaluation. First and foremost, to identify those that varied most between the chosen zonal geography, and accordingly might offer the most discrimination potential. Secondly, to minimise the potentiality for attribute exaggeration in a certain dimensions of the classification, which is influenced by highly correlated pairs. Thirdly, to ensure a proper level of quality of the selected variables, and therefore to improve the reliability of the output classification. Finally, to make sure there is a good diversity of variables that not only differentiate places but also fulfil the aim of a classification within some given theoretical framework.

4. Preliminary Results and Conclusion

Table 1 presents the preliminary conceptual model which is organised by three layers hierarchically, namely, concepts, domains, and measures. The framework guides those attributes that will be considered by the evaluation, with preliminary outputs shown in Table 2 and mapped in Figure 1.

This paper present has presented some initial consideration of variable evaluation in relation to data extracted from ACS and NYC OpenData repository, with the ultimate aim that these will be used to create a new general geodemographic classification for New York City.

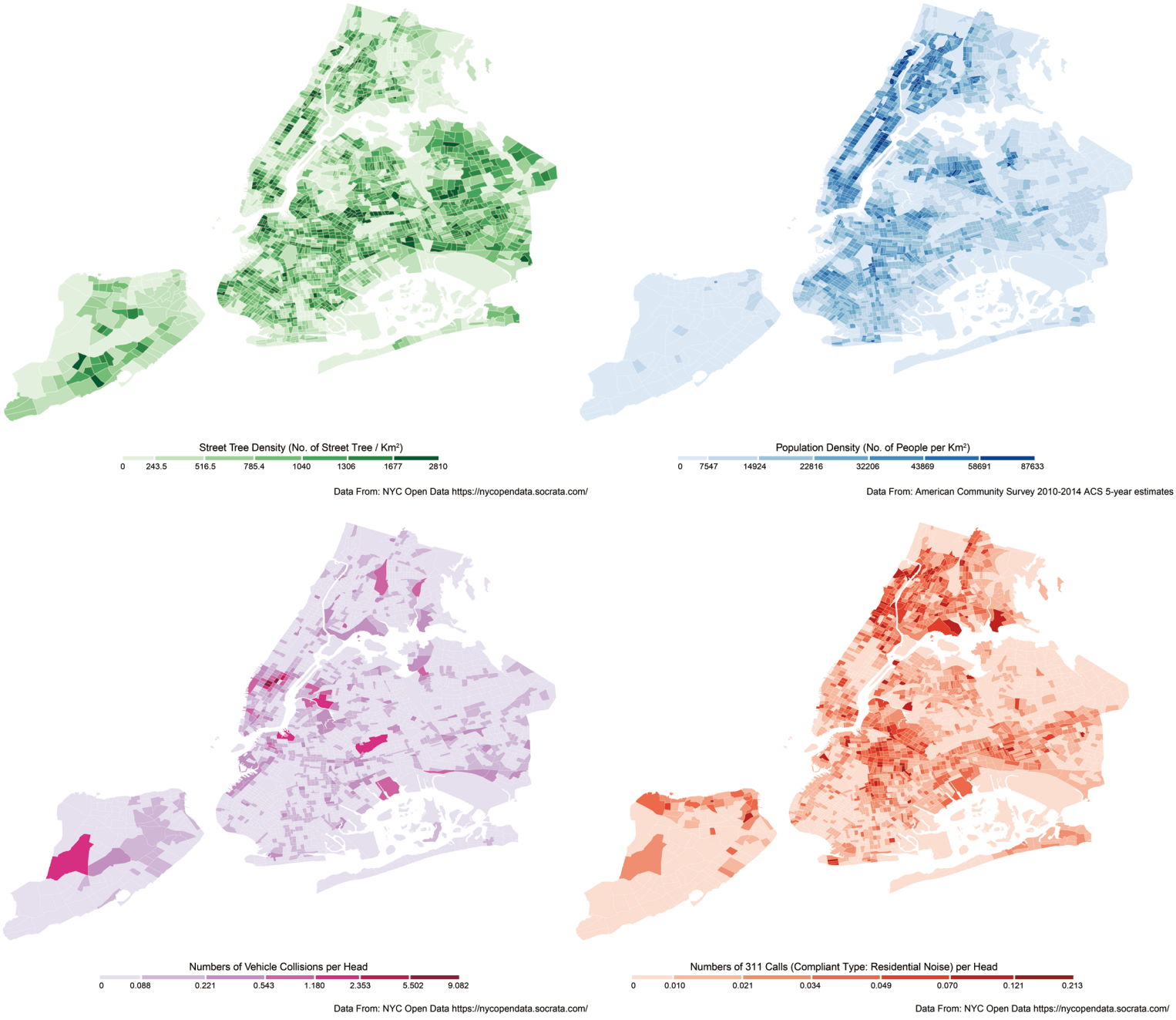
Table 1 Examples of Conceptual Model

Concepts	Domains	Measures (*Not Inclusive)
Demography	Age	Age bands
	Ethnicity	Ethnic groups
	Language	Language speak at home
	Household Structure	Martial type; with dependent children
Socioeconomic	Occupation	Types of employment
	Income	Median income; Vehicle ownership
	Education	Degree of qualification
	Mobility	Means of travel to work; Travel time to work
Living Environment	Housing	Year of built; Property value; Tenure
	Density	Population density; Building density
	Quality of Life	Trees; Proximity to amenities; 311
	Public Safety	Vehicle collisions

Table 2 Examples of Variable Evaluation (Not Inclusive)

Data	Geographic Resolution	Coverage	Variance σ^2	Domains	Skewness (Distribution)	Provenance	Temporal Resolution
311 Calls	Points	NYC	10131.16	Quality of Life	1.8908	NYC OpenData	Daily
Population Density	Census Tracts	Nation	185168917	Density	1.2547	ACS2010-14	Annually
Tree Density	Points	NYC	243384.6	Quality of Life	0.4723	NYC OpenData	Decennially
Vehicle Collisions	Points	NYC	86589.88	Public Safety	3.5876	NYC OpenData	Weekly

Figure 1. Examples of Variable Visualisation (Resolution: Census Tracts)



5. References

Alexander, C. (2002) *A discussion of the quality of estimates from the American Community Survey for small population groups*. Available: https://www.census.gov/content/dam/Census/library/working-papers/2002/acs/2002_Alexander_01.pdf

Glenn, E. (2016) *Working with the American Community Survey in R: A guide to using the acs Package.* Gewerbestrasse: Springer.

Gurin, J. (2014) *Open Data Now: The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation*. McGraw-Hill Education.

Harris, R Sleight, P and Webber, R (2005). *Geodemographic, GIS and Neighbourhood Targeting.* Chichester: John Wiley & Sons Ltd.

Lathrop, D. and Ruma, L. (2010) *Open Government: Collaboration, Transparency, and Participation in Practice.* Sebastopol: O'Reilly.

Leventhal, B (2016). *Geodemographics for Marketers: Using Location Analysis for Research and Marketing.* London: Kogan Page

MacDonald, H (2006) *The American Community Survey: warmer (more current), but fuzzier (less precise) than the decennial census.* Journal of the American Planning Association, 72(4), 491-503

Open Definition (n.d.) *The Open Definition.* Available at: http://opendefinition.org/

Openshaw, S., Cullingford, D., and Gillard, A. (1980), A critique of the national classifications of OPCS/PRAG. *Town Planning Review.* 51(4):421

Pew Research Centre (2015) *Americans' view on open government data.* Available at: http://www.pewinternet.org/2015/04/21/open-government-data/

Roest, A. (2016) *NYC Open Data Plan 2016.* Available: http://www1.nyc.gov/assets/doitt/downloads/pdf/open-data-update-2016-final.pdf

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

Singleton, A., Spielman, S. (2014) The past, present, and future of geodemographic research in the United States and United Kingdom, *The Professional Geographer,* 66:4, 558-567

Sleight, P. (1997). *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*, Henley-on-Thames, NTC Publications.

Spielman, S. and Singleton, A. (2015) Studying neighbourhoods using uncertain data from the American Community Survey: A contextual approach. *Annals of the Association of American Geographer*, 105:5, 1003-1025

Biography

Alex Singleton is a Professor of Geographic Information Science at the University of Liverpool and Deputy Director of the ESRC Consumer Data Research Centre (CDRC). In a general sense his research is concerned with how the complexities of individual behaviours manifest spatially and can be represented and understood though a framework of geographic data science. In particular, this research has extended a tradition of area classification and He has developed a broad critique of the ways in which geodemographic methods can be refined through modern scientific approaches to data mining, geographic information science and quantitative human geography.

Yunzhe Liu is currently a first year PhD student in the in the Geographic Data Science laboratory at University of Liverpool mainly under Prof. Singleton's supervision. Prior to this he awarded distinct graduate from the MSc Geographic Information Sciences at UCL with Prof. Tao Cheng's supervision. He is interested in researching about Big Data mining, Open Geodemographics, and urban geography/planning.

Dani Arribas-Bel is a Lecturer in Geographic Data Science at the Department of Geography and Planning at the University of Liverpool (UK), where he directs the MSc in Geographic Data Science. He is also part of the development team of the open source library PySAL for spatial analysis in Python. His main research interests are: Urban Economics and Regional Science, Spatial Analysis and Spatial Econometrics, and Open source scientific computing

Guanpeng Dong is a Lecture in Geographic Data Science at Department of Geography and Planning, University of Liverpool. His research interests are in spatial statistics and multilevel models to explore urban housing market dynamics, inequality and segregation, and in economic valuation of environmental amenities and neighbourhood socioeconomic characteristics using economics sorting models.