GEO-SPATIAL INFORMATION SCIENCE
2019, VOL. 22, NO. 4, 251–264
https://doi.org/10.1080/10095020.2019.1621549

Taylor & Francis
Taylor & Francis Group

A Principal Component Analysis (PCA)-based framework for automated variable selection in geodemographic classification

Yunzhe Liu, Alex Singleton and Daniel Arribas-Bel

Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool, Liverpool, UK

ABSTRACT

A geodemographic classification aims to describe the most salient characteristics of a small area zonal geography. However, such representations are influenced by the methodological choices made during their construction. Of particular debate are the choice and specification of input variables, with the objective of identifying inputs that add value but also aim for model parsimony. Within this context, our paper introduces a principal component analysis (PCA)-based automated variable selection methodology that has the objective of identifying candidate inputs to a geodemographic classification from a collection of variables. The proposed methodology is exemplified in the context of variables from the UK 2011 Census, and its output compared to the Office for National Statistics 2011 Output Area Classification (2011 OAC). Through the implementation of the proposed methodology, the quality of the cluster assignment was improved relative to 2011 OAC, manifested by a lower total within-cluster sum of square score. Across the UK, more than 70.2% of the Output Areas (OAs) occupied by the newly created classification (i.e. AVS-OAC) outperform the 2011 OAC, with particularly strong performance within Scotland and Wales.

ARTICLE HISTORY
Received 22 August 2018
Accepted 14 May 2019

KEYWORDS
Geodemographics; variable selection; UK census; spatial data mining; principal component analysis

1. Introduction

A geodemographic classification aims to summarise the multidimensional socio-economic and built characteristics of small area zonal geography, and are often referred as "neighbourhood" classification (Harris, Sleight, and Webber 2005). Geodemographic analysis relates to the application of such classifications and is positioned within a history of analytical frameworks that have aimed to explore the comparative context of urban areas (Bassett and Short 1980; Timms 1971). The theoretical tenet of geodemographic classification relates to the principle of homophily, which in geographic terms is the tendency for individuals to be attracted to areas that contain others with similar characteristics to themselves (Sleight 1993; Webber and Craig 1978). As such, the methodological objective when creating a geodemographic classification is, therefore, to sort a set of small areas into clusters that share similar characteristics, with the output of such groupings providing a simplified and categorical representation of the overarching multidimensional geography (Spielman and Singleton 2015).

In general terms geodemographic classifications are created in a series of stages that include the gathering of input variables that describe various characteristics of a given set of small areas; potentially normalizing these inputs and then standardizing the measures onto the same scale. Due to the high dimensionality of contemporary geodemographics, computational methods are implemented to examine the similarity between areas. This is most commonly achieved through an implementation of cluster analysis which refers to a family of computational methods that will typically have the general goal to maximize within-group similarity and between-group difference through various optimization strategies (Adnan 2011; Everitt et al. 2011). Outputs may typically be presented as a hierarchy, with larger and coarser groupings being split into smaller more specific nested groups; with such structure again created through various clustering or partitioning strategies. After this process is complete, it is typical that the characteristics of the assembled clusters are described by looking at which input variables are over- or under-represented within them; and these are then used to build written "pen portraits" and illustrative graphics.

As a methodological approach, geodemographic analysis has a lineage of application across the public and private sectors, spanning multiple decades and geographic contexts (Bassett and Short 1980; Longley 2005; Longley and Goodchild 2008; Singleton and Spielman 2013). However, the utility of a geodemographic classification for a given application is substantially determined by those methodological choices made during construction (Openshaw, Blake, and Wymer 1995). For example, it may be pertinent to align geodemographic

CONTACT Yunzhe Liu psyliu7@liverpool.ac.uk
The supplemental data for this article can be accessed here: https://doi.org/10.1080/10095020.2019.1621549.

classification inputs to those drivers of a small area differentiation within the context of a particular application (Singleton and Longley 2009) or, for analysis of specific localities, a classification may be enhanced by considering inputs derived for a focused rather than national extent (Singleton and Longley 2015). Furthermore, standardization algorithms (e.g. z-scores, range, and inter-decile range) can have various impacts upon classification shape and performance (Gale et al. 2016). Given the impact of methodological choice, it is typical that great care is taken into the testing and evaluation of different approaches along with their outputs, and this is acute within the context of those national classifications released by official statistical bodies where extensive stakeholder consultation and ratification are typically implemented as part of the construction process (Gale et al. 2016; Vickers and Rees 2007).

A primary task when building any geodemographic classification is to develop a framework for the selection of specific variables that will produce meaningful and application-relevant clusters (Murphy and Smith 2014). Those debates about the brevity of geodemographic classification inputs have been rehearsed for a long time. Openshaw and Wymer (Openshaw and Wymer 1995) advocated that "the fewer the variables the better", whereas Harris, Sleight, and Webber (2005) state that a more meaningful classification is likely to be constructed through inputting more variables, unless these variables are not "reliable, robust, and adding new information". The dimensionality of inputs (i.e. the number of zones multiplied by the number of variables) also has an interaction with the effectiveness of clustering methods to find salient structure from the data. Clustering performance can be hugely improved through the reduction of the number of variables due to this "curse of dimensionality" (Alelyani, Tang, and Liu 2014; Guyon and Elisseeff 2003; Pacheco 2015; Rojas 2015). Taking such perspectives into consideration, a typical objective of variable selection is therefore to achieve input parsimony, that is, the identification of the smallest subset of input variables that capture the most variation within the original dataset (Debenham 2002; Gale et al. 2016; Harris, Sleight, and Webber 2005). This will typically be achieved by balancing both the theoretical and empirical rationale for variable inclusion (Spielman and Singleton 2015). For example, it is common that initial inputs are presented within a framework that draws upon wider literature, guiding the type and balance between different potential influences upon or outcomes of area differentiation. More empirically, this will usually ensue a process of initial candidate input variable evaluation, typically considering a range of factors about the individual candidate variables including their correlation, distribution or spatial coverage.

The remaining sections of this paper are presented as follows. In Section 2, we introduce the Office for National Statistics 2011 Output Area Classification (2011 OAC) as an example geodemographic that has an open and reproducible methodology; and focus particularly on the variable selection method adopted to create it. This is followed by a consideration of alternative methods that have been used to select geodemographic inputs in some other past national classifications built for either the UK or Great Britain. We then consider the use of Principal Component Analysis (PCA) as an alternative methodology for automating variable selection within Section 3, alongside results of the developed methodology in Section 4. In Section 5 we compare and contrast the results of a cluster analysis using the variable selection method with 2011 OAC. The limitations of this research are discussed in Section 6, alongside some plans for further work and extension.

2. Selecting variables in national classifications

The 2011 OAC is a UK census-only geodemographic, which was released in 2014 by the Office for National Statistics (ONS) (Gale et al. 2016). This followed a similar classification created for the UK from the 2001 Census (Vickers and Rees 2007). Both the 2001 and 2011 OAC have an open methodology and data inputs, which enable reproducibility, and furthermore provide a useful framework upon which comparative studies can be designed (Gale et al. 2016). The 2011 OAC presents a three-tiered hierarchy, comprising eight supergroups, 26 groups, and 76 subgroups. Each output cluster presents a shorthand name and "pen portrait" (description) depicting the most salient multidimensional characteristics (Gale et al. 2016; Office of National Statistics 2015).

The initial variable selection for 2011 OAC only considered those non-redundant census variables that were consistently provided by the three different UK census agencies (England and Wales, Scotland, Northern Ireland); and as a result of public consultation, was also guided by the 2001 OAC inputs. In 2011 OAC, 166 prospective variables (including 94 variables that were referenced by the 2001 OAC) and a derived variable of the standardised illness ratio (SIR) were tested. Moreover, the suitability of these initial variables was also scrutinized by the ONS (Gale et al. 2016). The initial variables were rationalized with two main objectives. The first was to obtain a variable mix that represented the general characteristics of the UK's neighborhoods, meanwhile, also distinguishing salient characteristics that varied geographically. A second requirement was to minimize the number of strongly correlated census variables,

thus limiting any potential weighting effect that may be caused by collinearity. According to Gale et al. (2016), these requirements were achieved through two empirical approaches. The first was to examine the correlations of candidate variables, and specifically identifying those variable pairs for further consideration where the correlation was greater than ± 0.6. A second technique implemented cluster-based sensitivity analysis, which aimed to identify those variables that had the greatest impact, either positive or negative, on cluster formation. This method assessed the total within-cluster sum of squares and the total between-cluster sum of squares statistics after including or excluding different variables from a clustering run. After further evaluation including examination of statistical distributions and mapping, 60 variables were eventually retained to build 2011 OAC, which were broadly organized into three domains: demographic, housing and socioeconomic (Gale et al. 2016).

However, the OAC (both 2001 and 2011 OAC) approach to variable selection deviates from those methods implemented by academics building geodemographic classifications for pre-2001 censuses in the UK. The very different computational contexts of the past made more sophisticated multidimensional processing much slower or impossible (Adnan 2011; Singleton 2016). Prior to the 2001 OAC, dimensionality reducing methods such as principal component analysis (PCA) were commonly (although not universally) integrated into some of the classification products that corresponded to the decennially released Census (e.g. Webber (1975); Webber and Craig (1978); Charlton, Openshaw, and Wymer (1985); Robinson (1998)). When a PCA is calculated for a dataset, a set of new orthogonal variables (i.e. principal components) are created which are the linear combination of the original variables. The principal component that accounts for the largest variance is called the first principal component, the second principal component that accounts for the second largest variance as the second, and so forth (Jolliffe 2002; Pacheco 2015). Clustering a set of principal components reduces the overall number of inputs to a geodemographics, making the clustering process either possible or much faster to complete; which in the past had been a key constraint given more limited computational power/availability. However, as the data handling and processing capacity of computers have increased, the necessity for PCA in this context has been reduced. Furthermore, some scholars have also argued that use of PCA to create inputs may erase interesting patterns, and particularly those which are spatially heterogeneous (Alelyani, Tang, and Liu 2014; Harris, Sleight, and Webber 2005; Leventhal 2016).

However, there are some contemporary implementations of PCA when building geodemographic classifications. For instance, Ismail, Nayan, and Ibrahim (2016) employed PCA as an inspection tool that determines whether a linear relationship exists between candidate variables; Adnan (2011) adopts PCA as a standardization technique in the progress of producing real-time geodemographics. Although not a necessity in terms of computation, and as illustrated by Debenham (2002), PCA can have a useful role as a tool that guides variable selection. Although, what is under-researched is how such a process could be automated, taking account of both the overall importance of input variables to cluster formation, but also those sensitivities of the extent to which such relationships may hold between different localities. One of the overarching objectives of this paper is therefore to re-examine the potential for PCA within a computationally intensive setting, where the benefits of PCA for the identification of variables that explain the main variance within a dataset can be integral to an automated variable selection process. We present this new methodology in the context of a UK census-based geodemographic, contrasting the output against the 2011 OAC.

3. Automated variable selection using PCA

As discussed in the previous section, an overarching objective of the variable selection stage of building a geodemographic classification is to identify the smallest possible subset of variables that can represent the main variance within a universe of potential inputs being considered, which may also be informed by theoretical or practical rationale. Although accepting of arguments that PCA can have an adverse effect when used to create inputs to a geodemographic classification (Alelyani, Tang, and Liu 2014; Harris, Sleight, and Webber 2005; Leventhal 2016), we would argue that PCA can still have utility as a tool in the identification of appropriate input variables; which is the basis of the method we introduce in the remainder of this section.

The flowchart presented in Figure 1 illustrates an approach that is comprised of five main stages. The first stage generates a set of principal components (PCs) from the input variables. Meanwhile, by summing up of the squared factor scores for the PC, the eigenvalue associated with each of the PCs can be calculated, which is utilized to define the range of iteration tests at stage 2. Additionally, the contribution of the variable to each component can be obtained by calculating the ratio between the squared factor score for a variable and the eigenvalue associated with that component. The value of a contribution is between 0 and 1. Generally, the larger the value, the more a variable contributes to the component (Abdi and Williams 2010; Pacheco 2015).

Stage 2 defines a threshold for the number of iterations to test between a "harsh" and a more "liberal" cut-

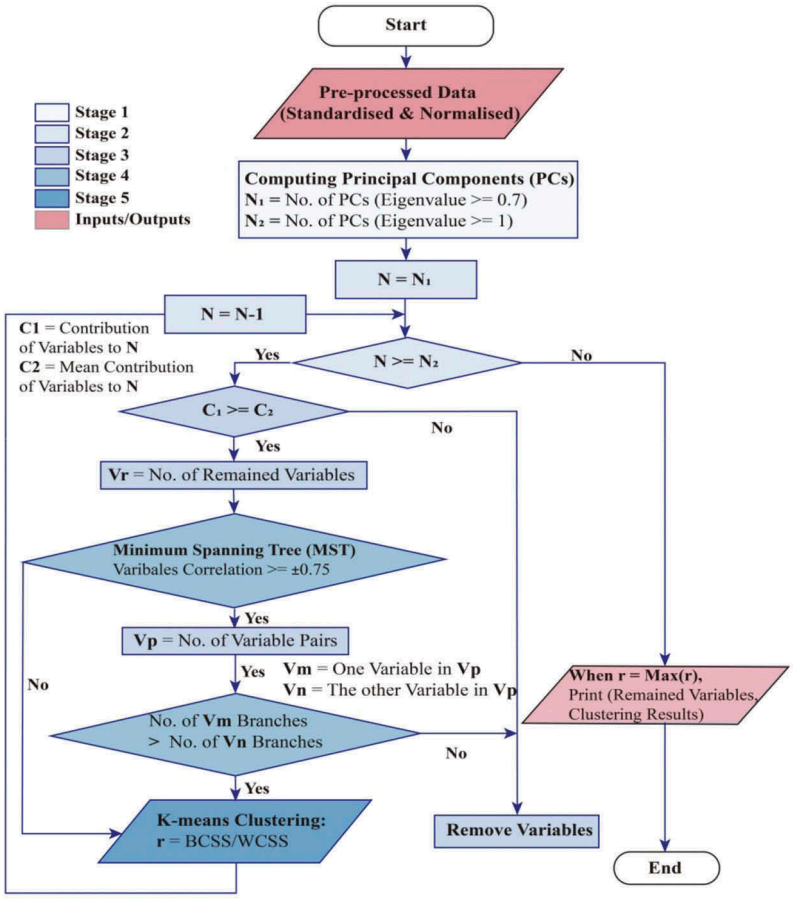

Figure 1. Proposed automated variable selection method workflow.

off point. The maximum, i.e. "harsh", threshold value is defined by a strict cut-off point that is generated from the commonly used Kaiser's rule, namely, the eigenvalue of a meaningful PC is greater than or equal to 1 (Jolliffe 2002; Kaiser 1960; Pacheco 2015). The minimum, i.e. "liberal", threshold value is determined by adopting a cut-off point that is suggested by Jolliffe (1972), namely, the eigenvalue of a meaningful PC is greater than or equal to 0.7.

Stages 3–5 are iterative, with each run successively removing a PC from the set identified at stage 2. In every iteration at stage 3, the contribution of the variables to the retained PCs is quantified by taking the sum of their individual contributions multiplied by their respective eigenvalues in each PC (Pacheco 2015). Abdi and Williams (2010, 437) suggest the use of "larger than the average contribution" as a heuristic cut-off when identifying variables with high contributions. Similarly, in this stage, where a variable contribution is greater than the average of these summed scores, it is retained for stage 4.

Stage 4 explores the correlation between the retained variables using a Minimum Spanning Tree (MST), which re-examines the level of data redundancy. Any highly correlated pairs are highlighted by the tree, defined as having a correlation coefficient of greater than or equal to ± 0.75, which is commonly cited as the "rule of thumb" indicating a high correlation (Santero, Nayan, and Ibrahim 2016; Udovičić et al. 2007). In these instances, those highly correlated variables (i.e. nodes in the MST) with the fewest branches (i.e. less connected) were removed from the candidate variable list since they are considered of lower importance (Financial Network Analytics 2012). Although automated in this instance, it follows similar methods implemented when building some commercial geodemographics (Harris, Sleight, and Webber 2005).

At stage 5, the filtered variables are then clustered using the K-means algorithm with a user-specified number of clusters. This was optimized by running 10,000 times which is necessary given that the starting seeds used to initialize cluster partitioning are stochastic, and as such, there can be slight differences in outcomes between each run of the algorithm. Of the 10,000 runs that are generated for each iteration of Stage 5, the result with the lowest Total Within-cluster Sum of Squares (TWSS) statistics is extracted; representing a solution with overall more compact clusters. For this selected optimized run, two statistics were calculated as a measure of overall clustering quality: The between cluster sum of square (BCSS) and within cluster sum of square (WCSS) statistics.

At the end of each Stage 5 run, the WCSS and BCSS are then stored in association with the currently tested PC cut-off defined at Stage 3. The ratio between the WCSS and BCSS is used to monitor the impact of the specific PC selection. Generally, the larger the ratio, the better clustering results. The iteration stops when the minimum/maximum (depending on removing/adding) number of PC defined by Stage 2 is met.

4. Case study application

The automated variable selection process presented in the previous section was implemented in an example of building a UK census geodemographics that would be broadly comparable to 2011 OAC. As discussed earlier, the open methodology and data used to create this geodemographics make it a useful candidate for comparison; and in drawing parallels between the classifications we can illustrate broad comparability and the utility of the presented technique. It should be noticed that the objective of this application was, therefore, to retain broad comparability with the 2011 OAC, and to this end, the methods of standardization, normalization and clustering were mirrored. Thus, input data were normalized using an Inverse hyperbolic sine, and then range standardized onto a 1–0 scale. For the K-means implementation, only the most aggregate level of hierarchy in 2011 OAC was considered in this comparison, so k was defined as 8 for this model, although future work might consider further levels of disaggregation or a range of different k values might be tested. The rationale for the specific methodological choices in 2011 OAC can be found within Gale et al. (2016); however, the key point of departure in the presented methodology relates to how the final variables are selected for input into the clustering process.

In this application of the generally applicable methodology outlined in the previous section, we considered nearly all variables contained within the Key Statistics (KS) and Quick Statistics (QS) tables for the UK, which included the 167 initial variables considered for inclusion in 2011 OAC. Although, given that some tables contained duplicated or near identical topics, only one of these tables were included. For instance, both tables KS104 and QS108 concerned living arrangements; tables KS102 and QS102 detailed age structure. Finally, like 2011 OAC, the initial inputs also included computation of a Standardised Illness Ratio. The full initial variable specification is listed in Table S1 of the online Supplementary Materials.

After running a PCA on the input data, a total of 53 meaningful PCs were identified by examination of the eigenvalue thresholds which are plotted against the cumulative variance explained in Figure 2. If we had applied the Kaiser rule (Eigenvalue ≥1), only 30 principal components would have been selected which cumulatively accounted for about 72.4% of the variance being retained. However, it can be seen that by altering the cut-off value from 1 to 0.7, the filtering process identified 52 principal components, which cumulatively accounted for approximately 83.6% of the variance contained in the original data.

A summary of outcomes from the iteration tests is shown in Figure 3. The highest quality clustering results were identified (an objective function of maximizing the ratio between BCSS and WCSS) when the first 51 PCs were used to identify input variables to the cluster

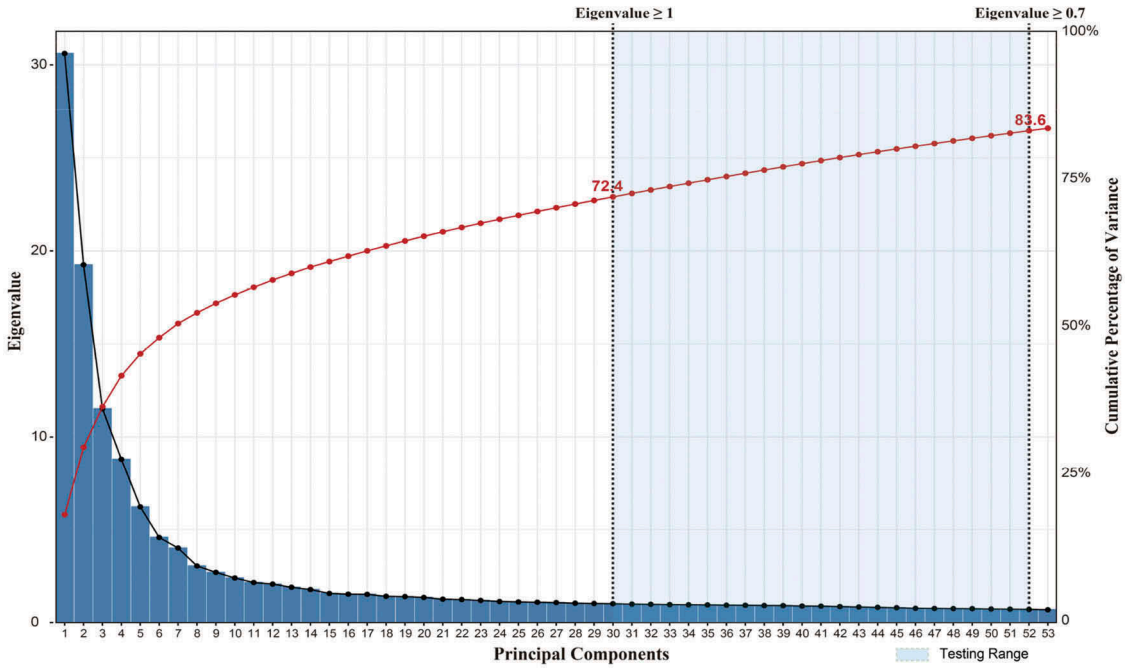

Figure 2. Scree plot: Eigenvalue vs. percentage of explained variances.

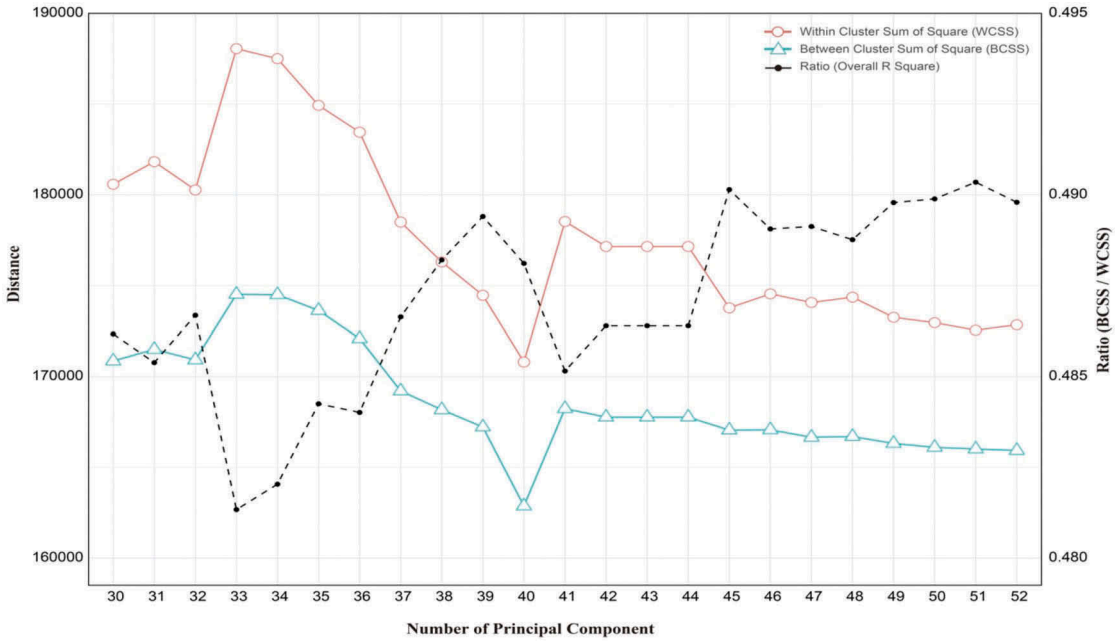

Figure 3. BCSS and WCSS result by iteration test by principal components (ratio = BCSS/WCSS).

analysis. Of the 86 variables firstly identified, 12 pairs were highly correlated (Correlation Coefficient ≥ ± 0.75); and as such, utilizing the minimum spanning tree (Figure 4), 12 variables were removed.

Details about the type and frequency of variables retained for each iteration are presented in Table 1. The variables have been divided into three different domains: demographic, socioeconomic and housing (also see Table S1 of the online Supplementary Materials). Overall the iterations, variables in the socioeconomic domain were retained less often, indicating greater redundancy. In contrast, the housing

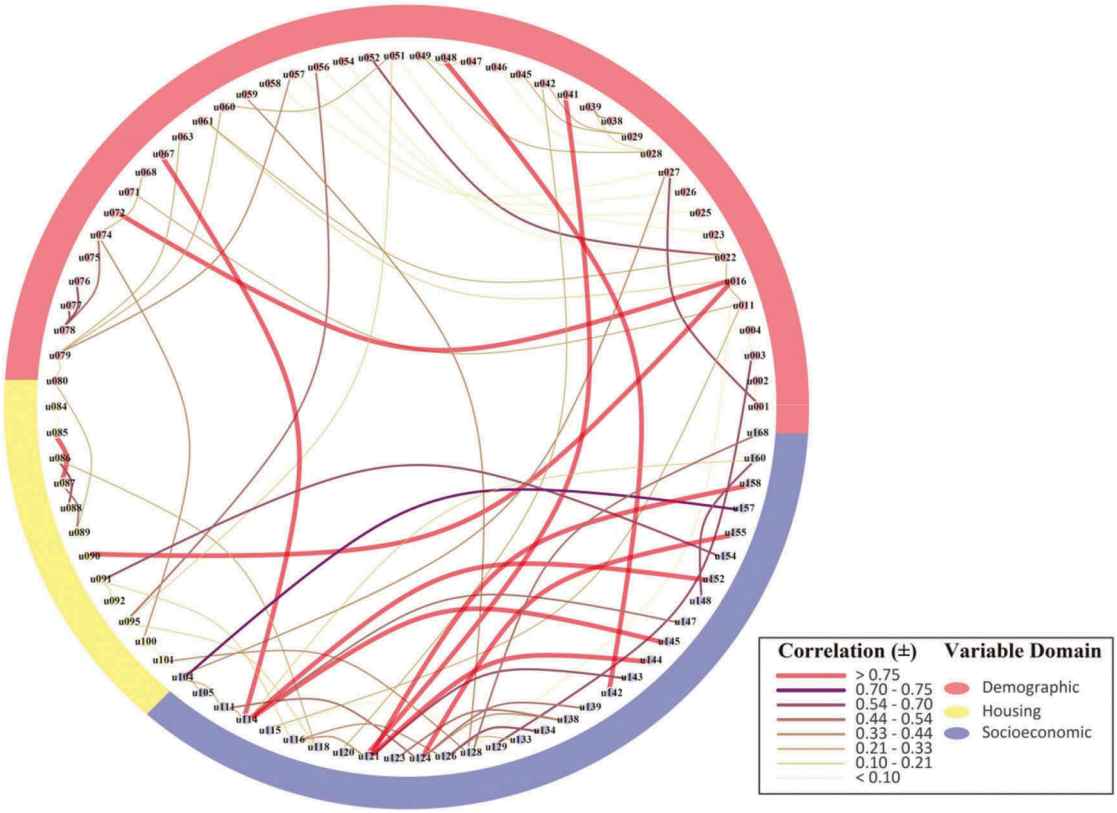

Figure 4. Minimum spanning tree (presented in circle layout) of the census variables after the PCA-based filter.

The thickness of curve indicates the absolute value of the person correlation coefficient. The value that above ±0.75 is highlighted by red thicker line, which therefore will be removed in the next phases.

Table 1. Testing results showing the number and percentage of overall retained variables and by domain.

Number of PCs	Number of retained variables	Demographic	Socioeconomic	Housing	Ratio	Demographic % by total	Socioeconomic % by total	Housing % by total
30	90	50	28	12	0.4862	62.5	41.8	60.0
31	91	50	29	12	0.4854	62.5	43.3	60.0
32	90	49	29	12	0.4867	61.3	43.3	60.0
33	88	49	27	12	0.4813	61.3	40.3	60.0
34	87	49	26	12	0.4820	61.3	38.8	60.0
35	85	48	25	12	0.4842	60.0	37.3	60.0
36	84	47	25	12	0.4840	58.8	37.3	60.0
37	81	47	24	10	0.4866	58.8	35.8	50.0
38	79	45	24	10	0.4882	56.3	35.8	50.0
39	78	43	25	10	0.4894	53.8	37.3	50.0
40	75	43	23	9	0.4881	53.8	34.3	45.0
41	75	43	22	10	0.4852	53.8	32.8	50.0
42	74	43	21	10	0.4864	53.8	31.3	50.0
43	74	43	21	10	0.4864	53.8	31.3	50.0
44	74	43	21	10	0.4864	53.8	31.3	50.0
45	74	42	21	11	0.4901	52.5	31.3	55.0
46	75	42	23	10	0.4891	52.5	34.3	50.0
47	74	41	23	10	0.4891	51.3	34.3	50.0
48	75	41	24	10	0.4888	51.3	35.8	50.0
49	75	41	23	11	0.4898	51.3	34.3	55.0
50	74	41	23	10	0.4899	51.3	34.3	50.0
51	74	40	24	10	0.4903	50.0	35.8	50.0
52	74	40	24	10	0.4898	50.0	35.8	50.0
Total	167	80	67	20		55.4	36.0	53.3

domain was reasonably stable, and for all iterations comprised between 45 and 60% of the overall variables within this domain.

In the optimized result (51 PCs), 74 variables in total were retained, distributed between 40 demographic, 24 socioeconomic, 10 housing (see Table S2 of the online Supplementary Materials). Table 2 summarizes this distribution relative to those inputs used to build 2011 OAC. Most significantly, the proportion of retained variables related to demographics was much larger, while the other domain proportions remained largely similar in size.

4.1. Describing the derived classification

In this penultimate section, we firstly present descriptions to accompany the optimized clustering result derived through automated variable selection (Automated Variable Selection OAC – AVS-OAC). The new classification created through this process had an average cluster size of approximately 29,037 Output areas (OAs), however, varied from 11,397 (E) and 41,399 (B) OAs, which, respectively, correspond to about 4.9% and 17.8% of the total number of OAs in the UK. By contrast, 2011 OAC varies from 8,589 OAs (2: Ethnicity Central) to 35,285 OAs (6: Urbanities), so the range of our presented clusters is larger.

Table 2. The number of final census variables retained by domain versus 2011 OAC.

Domain	AVS-OAC	AVS-OAC (%)	2011 OAC	2011 OAC (%)
Demographic	40	54.1%	26	43.3%
Socioeconomic	24	32.4%	26	43.3%
Housing	10	13.5%	8	13.3%
Total	74	-	60	-

Figure 5 maps the geographic distribution of the AVS-OAC clusters across the UK, and also, respectively, highlights the cluster distribution in the largest cities, namely, London, Cardiff, Edinburgh, and Belfast. The spatial distribution highlights a useful urban-rural split, and within urban areas presents a range of differentiating clusters. Additionally, and as one might expect given the methodological choices made, London is fairly poorly segmented with the majority of inner London dominated by two clusters (i.e. Cluster D and E). This effect is similar in 2011 OAC, and indeed is discussed at length elsewhere (see Singleton and Longley 2015). One potentially negative observation of the created classification was the emergence of two clusters that represented mainly rural areas (Clusters 1 and 6). In order to explore these patterns and wider interpretability of the cluster characteristics and later comparison with 2011 OAC, index scores (i.e. $x/\bar{x} * 100$) were computed for the input variables and displayed in Figure 6, with the scores ordered by domain. These scores illustrate characteristics that are over or under-represented for each of the eight clusters relative to the national average (a score of 100). An index score of 50 is, therefore, half the national average, and 200 would be double. Additionally, as is common when building a geodemographic classification, such index scores were then used to ascribe a label and brief description of each of the clusters (see S3 Section of the online Supplementary Materials).

4.2. Classification performance and comparison to 2011 OAC

In this final section, we first evaluate AVS-OAC performance internally to explore cluster robustness, and

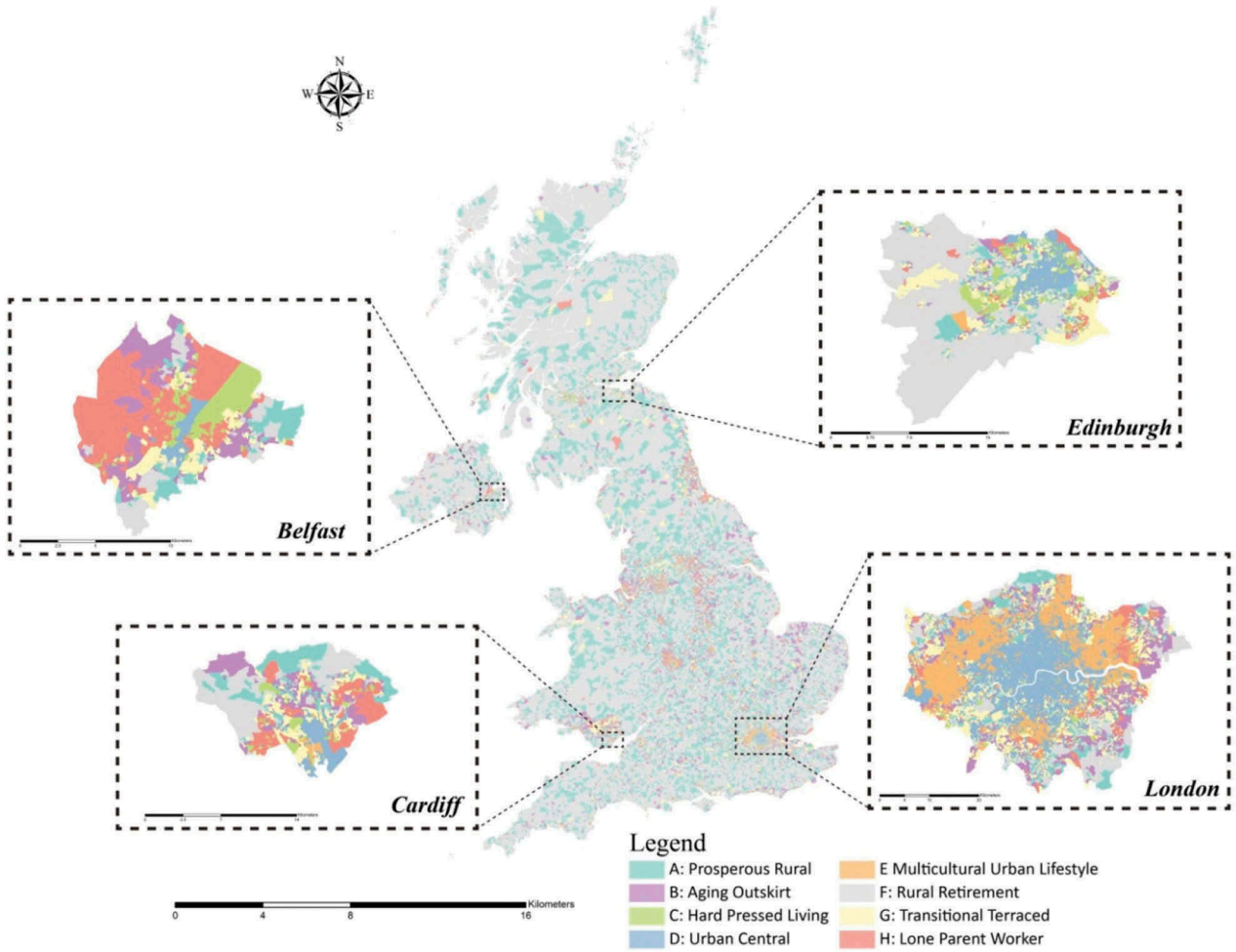

Figure 5. Geographic distribution of AVS-OACs with highlighted major cities.

then make some external comparisons with 2011 OAC; to establish those broad similarities or differences that emerge through the application of this alternative methodology, and examine the impact this has on the overall discriminatory power.

An objective when building this classification was to provide an output that would make a suitable benchmark against 2011 OAC; achieved through maintaining both a broadly similar potential attribute input pool and output cluster frequency. However, a disadvantage of constraining the number of clusters to match 2011 OAC was that two very similar rural clusters emerged: Cluster A: Prosperous Rural and Cluster F: Rural Retirement; which represented considerable redundancy. When building a geodemographic classification for operational rather than methodological evaluation purposes, there is typically a stage that will test multiple potential cluster frequencies with the objective of mitigating such issues. However, conversely, the post-analysis merging or splitting of clusters is also prevalent when building many geodemographic classifications (Harris, Sleight, and Webber 2005). For the purposes of this illustration we decided to keep this artifact, although in an operational model such as 2011 OAC, we would expect that such issues would be resolved pre or post clustering through manual intervention after stakeholder consultation.

Correspondence between 2011 OAC supergroups and AVS-OAC clusters is highlighted in Figure 7 which presents the percentage by of OAs that overlap between the two classifications for the UK extent. As might be expected given the differing inputs, the correspondence between the two classifications varies; and highlights the importance of stakeholder engagement when selecting appropriate cluster representations in operational models. For example, we can see that the AVS-OAC Cluster: "F: Rural Retirement" is composed predominantly by OA identified by 2011 OAC as within Supergroups "1. Rural Residents" and "6. Suburbanites", thus representing a blend of both rural and the connecting hinterland at the periphery of urban areas. The AVS-OAC Cluster "D. Urban Central" combines many OA that are identified by the 2011 OAC Supergroups "2. Cosmopolitans", "3. Ethnicity Central", but not some other predominantly urban clusters such as "7. Constrained City Dwellers", which emerged with greater correspondence to AVS OAC Cluster "C. Hard Pressed Living". Or, the AVS-OAC cluster "B: Ageing Outskirt" can be seen to correspond to a diffuse number of 2011 OAC Supergroups located

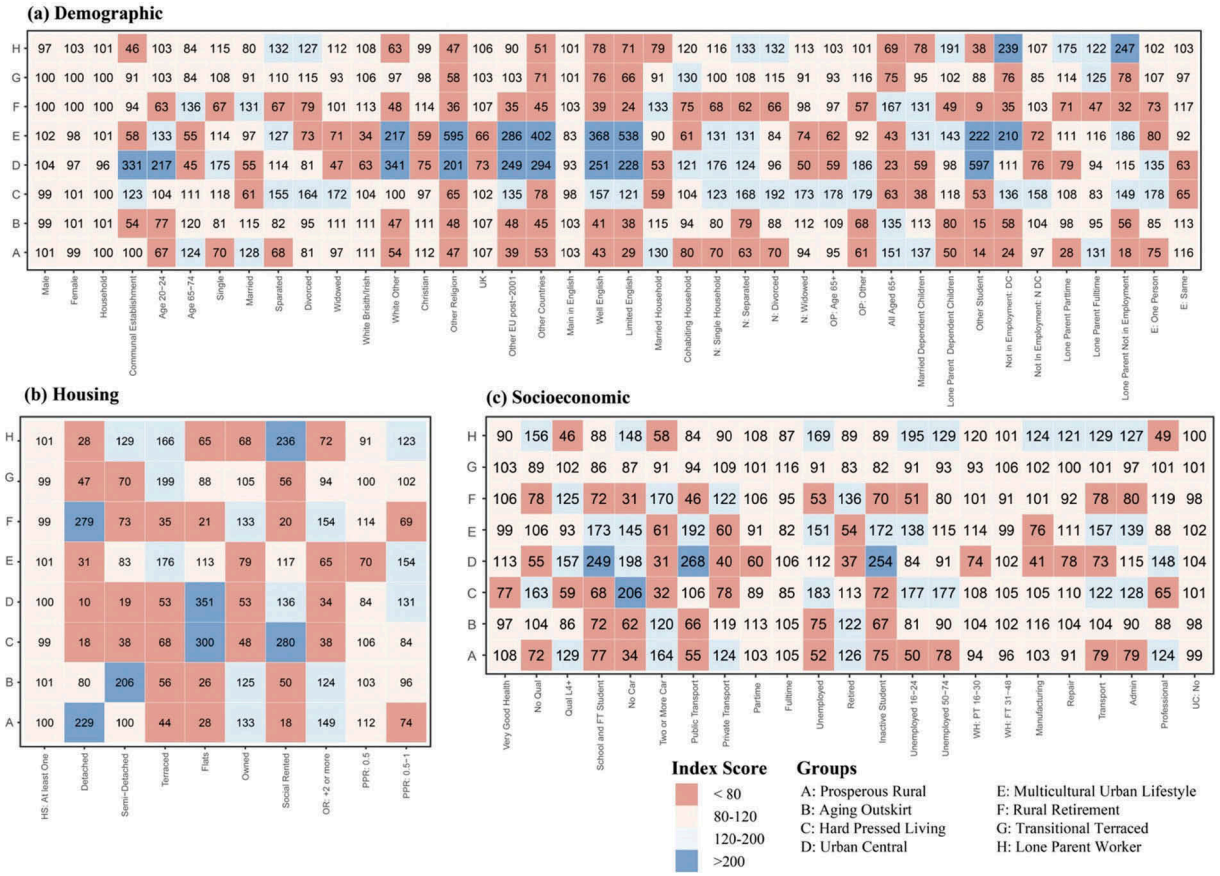

Figure 6. AVS-OAC results (Index scores) grouped by variable domains.

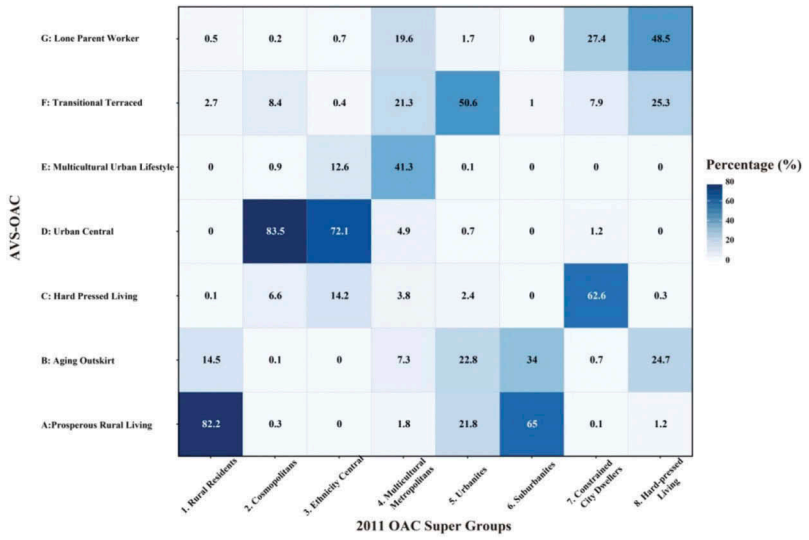

Figure 7. Cross-tabulation: OA percentage by AVS-OAC and 2011 OAC.

in suburban areas. A similarly defuse pattern can also be identified in "G: Transitional Terraced", although just over half of those areas identified by the 2011 OAC Supergroup "5. Urbanites" also correspond with this cluster.

As a measure of comparative clustering quality, a Total Within-cluster Sum of Squares (TWSS) statistic was calculated for each classification (i.e. AVS-OAC and 2011 OAC) by taking the sum of the squared difference between every classification input attribute within an area and the mean of the assigned cluster centroid. A higher score indicates an area where the attribute values for the OA are further from their assigned cluster mean (the centroid generated via k-means clustering), in other words, the quality of cluster assignment is poorer. Box plots in Figures 8 and 9, respectively delineate the

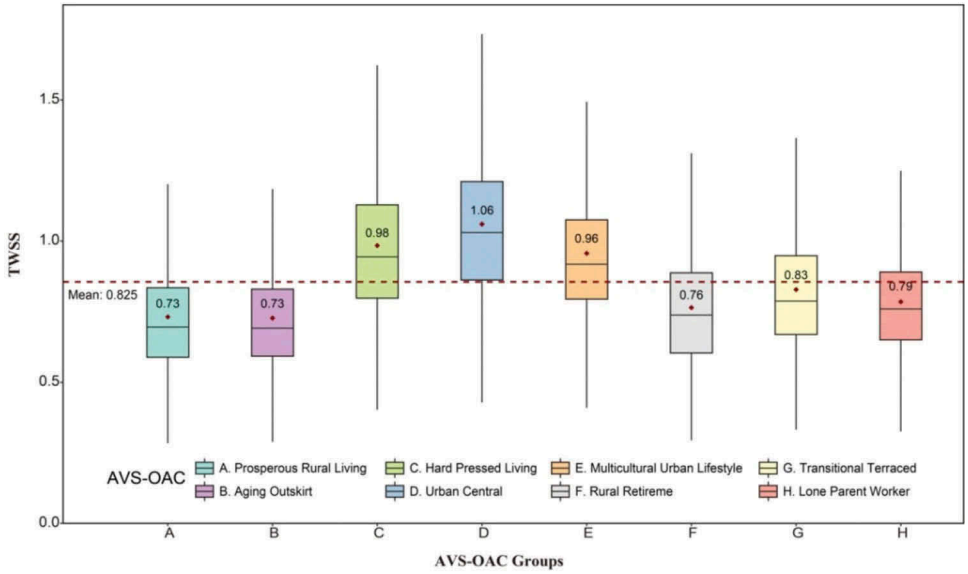

Figure 8. Total within-cluster sum of squares (TWSS) by the AVS-OACs. Mean value for each of the cluster is calculated and illustrated by the red point within the boxplot. The total mean value is presented by the dashed line.

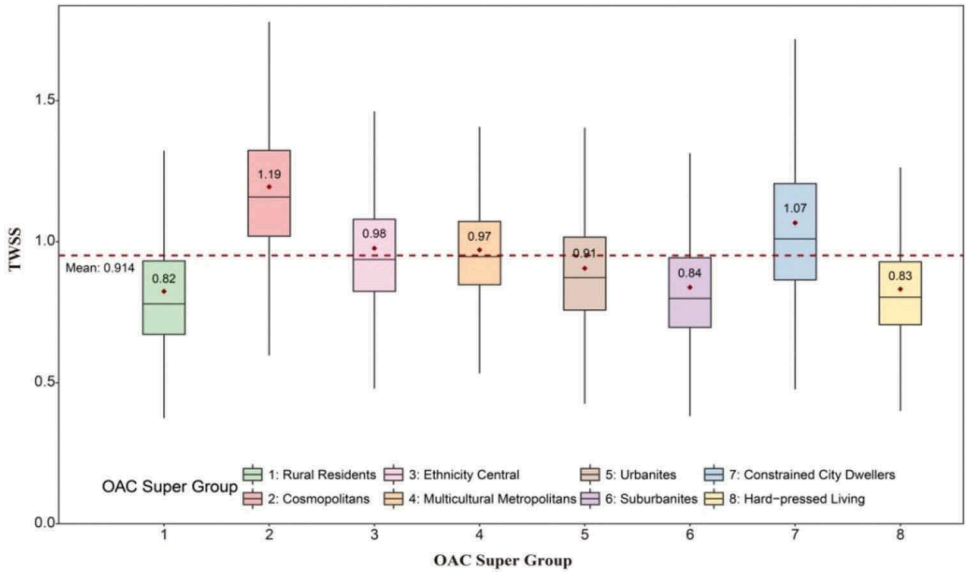

Figure 9. Total within-cluster sum of squares (TWSS) by the 2011 OAC. Mean value for each of the cluster is calculated and illustrated by the red point within the boxplot. The total mean value is presented by the dashed line.

TWSS by the AVS-OAC Clusters and the 2011 OAC Supergroups.

Overall, the AVS-OAC clusters have lower TWSS than 2011 OAC, statistically indicating a better fit, which is manifested by the average value (i.e. 0.825 and 0.914). Within AVS-OAC, we can see that the clusters "D: Urban Central", "C: Hard Pressed Living" and "E: Multicultural Urban Lifestyle" contain the highest TWSS value (average value, which are 1.06, 0.98, and 0.96) and the greatest variability (standard deviations, which are 0.286, 0.257, and 0.241, respectively), which might be considered the three least successful AVS-OAC clusters. These clusters are concentrated in both densely populated urban centers and transitional areas on the periphery of urban cores. In some sense, this is to be expected given the heterogeneous nature of urban centers and is an issue acute between Greater London and other parts of the UK which leads to larger variability. In particular, residents of AVS-OAC cluster "E: Multicultural Urban Lifestyle" are mainly concentrated within Greater London, which is a region known to be not well represented by 2011 OAC (Singleton and Longley 2015).

Analysis of the geographic variability in classification performance can be expanded by mapping how well the input attributes of each OA fit their assigned cluster from both AVS-OAC and 2011 OAC, again using the TWSS statistics. The frequency of OAs that performed better by AVS-OAC relative to 2011 OAC (attributes values that

are closer to their assigned cluster mean) was counted within each UK Local Authority District (LAD), and are presented in the choropleth map in Figure 10. Overall, 390 out of 404 local authority districts in the UK have greater than 50% of their constituent OAs statistically better represented by AVS-OAC relative to 2011 OAC. There are however some clear regional patterns that emerge; with particularly strong performance in Scotland and Wales where, respectively, all unitary authorities had more than 70% and 65% of OAs with better fit by AVS-OAC relative to 2011 OAC statistically. More negatively, there are some LADs that experience relatively poor performance, which are indicated by dark red in the choropleth (Figure 10): and include a cluster of boroughs alongside the River Thames within Greater London alongside some other London Boroughs. Additionally, some of the LADs located within Northern Ireland also exhibit poorer clustering performance. These instances support an argument for more consideration within an automated variable selection process of those characteristics specific to regional geographies. The need for greater regional consideration when building geodemographics is a well-established argument (Alexiou 2016), which also points to future work outside of the scope of this paper when selecting variables automatically.

Despite these regional differences, the overall performance of AVS-OAC relative to 2011 OAC is strong, with 70.2% of OA having a better fit. This is highlighted in Table 3 which presents the frequency and percentage of OA within each 2011 OAC

Table 3. Checklist: number of OAs that AVS-OAC outperforms than 2011 OAC.

2011 OAC	Frequency	Percentage (%)
1. Rural Residents	20,279	74.3
2. Cosmopolitans	7520	57.3
3. Ethnicity Central	6010	50.7
4. Multicultural Metropolitans	16,633	70.8
5. Urbanites	28,241	73.0
6. Suburbanites	33,678	71.9
7. Constrained City Dweller	20,780	76.6
8. Hard-pressed Living	29,982	68.4
Total	163,123	70.2

Frequency: Number of OAs better performed by AVS-OAC.

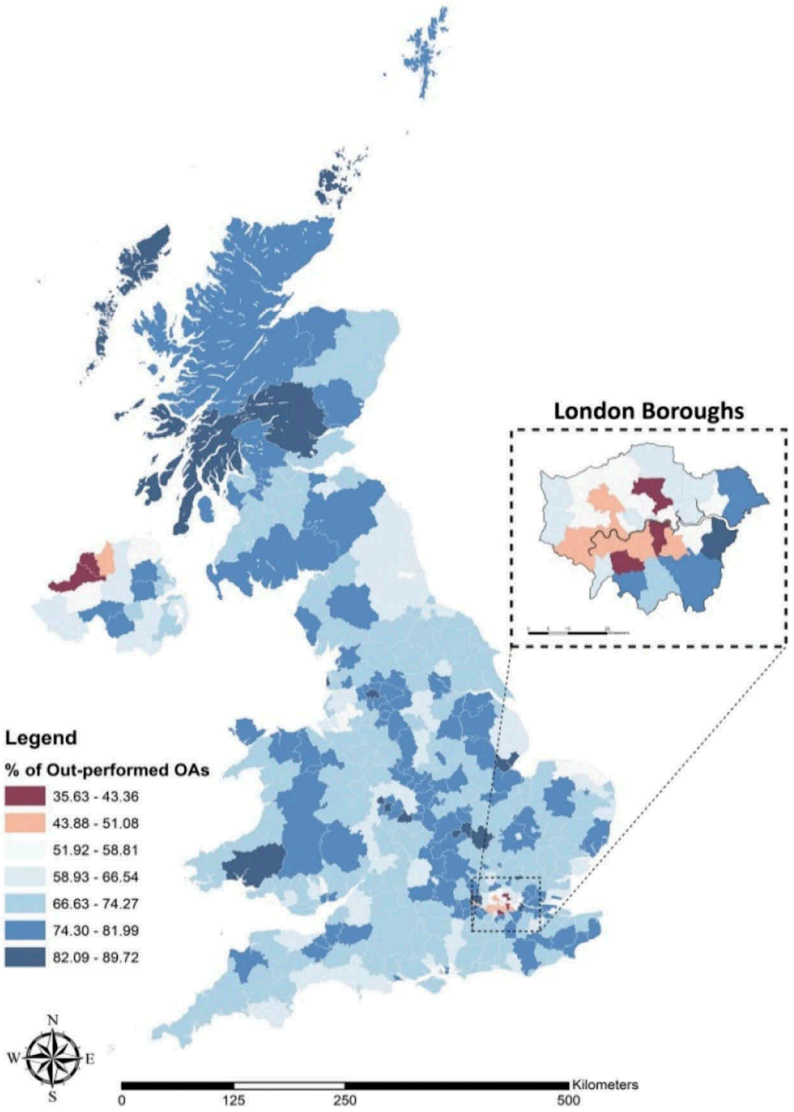

Figure 10. Percentage of out-performed OAs by AVS-OAC by local authorities and London boroughs.

Supergroup that were outperformed by their AVS-OAC cluster assignment.

5. Conclusions

The consideration of which variables are input into a cluster analysis is a common preliminary stage when building a geodemographic classification. The overarching objective is typically to achieve input parsimony, but there are multiple views on how this is best achieved, balancing theoretical considerations, practicalities of available data or attribute statistical properties and those past experiences or embedded knowledge of the classification builder(s). The primary objective of this paper was to extend such considerations by developing and testing an automated method of variable selection, and then benchmarking the presented technique within the context of building a UK national geodemographic from 2011 Census data.

The objective was to illustrate how automated variable selection could be implemented to identify inputs that produce a plausible and comparable classification. In doing so, we are not claiming that this be of equivalence to an operational model, specifically as the methodology presented here lacks user consultation; but rather provides an innovative tool that might be useful to inform variable choices. It is not difficult to envisage a build process within an operation setting where differing variable selection sets might be specified and evaluated in consultation with stakeholders.

Our heuristic process was built around Principal Component analysis that automated input variable selection, feeding these into a classification model that in our example broadly followed the 2011 OAC methodology. The application presented here was primarily data-driven for the purposes of methodological illustration; however, the technique itself is flexible and generic, and lends itself to other applications with any set of variables, thus also transferring well as a component of more theoretical expositions of geodemographic structure.

The method as implemented here was within the context of consistently available variables from the 2011 Census for the UK geographical extent. Through a five-stage variable selection procedure, 74 census variables were retained from 171 initial candidates. The clustering was constrained to mirror 2011 OAC cluster frequency, creating a final typology of eight clusters. This output was subsequently evaluated through comparison with 2011 OAC to examine both cluster similarity and relative performance. Overall, the quality of the cluster assignment is statistically better than 2011 OAC in more than 70.2% of the OAs across the UK; with particularly strong performance within Scotland and Wales.

The application of our method illustrated good comparative performance relative to 2011 OAC; however, there are several limitations that could be alleviated in future work. First, there may be potential for integrating regional and subregional evaluation when selecting variables, which might evolve into a set of heuristics that would potentially identify a more effective variable input mix. A counter view would be that this would be at the expense of computation time; and indeed, may be not resolve an inherent constraint in regional variability when building geodemographics from data pertaining to a national extent. Secondly, this automated process decouples stakeholder user input from the classification process; and as Openshaw, Blake, and Wymer (1995) state that "there is no simple relationship between optimizing a statistical measure of classification performance such as the within-cluster sum of squares and the end-users' perception of classification performance in a particular context". Such considerations could be integrated into a fuller process of classification building, which may be particularly important within the context of an operational classification, such as those built for a national statistical agency. Finally, it is also worthy of recalling that the presented method utilizes PCA and there is also potential to integrate alternate and more explicitly spatial techniques, which may also enhance regionally variable performance, for example through Geographically Weighted PCA (Harris, Brunsdon, and Charlton 2011). Furthermore, there is also potential that additional steps could be implemented that assess an appropriate cluster frequency for a given problem, although there would be significant challenges when balancing such considerations with computational efficiency when input variable combinations were also being assessed in parallel.

This paper has presented a new methodology that optimizes the selection of an initial list of candidate variables that are input into a cluster analysis used to build a geodemographic classification. The performance of this methodology is implemented within the context of the 2011 UK Census, and comparison is made with 2011 OAC. Performance was comparable to 2011 OAC over the evaluated metrics, although the shape of the classification varied, and there were also some regional differences in performance. The methodology presented here provides a generally applicable tool that integrates well with both theoretical and user embedded classification building programs over multiple international contexts, and, will likely have particular relevance for the creation of future geodemographics for the UK 2021 Census and beyond.

Notes on contributors

Yunzhe Liu is currently a PhD student the Geographic Data Science lab at University of Liverpool. He was graduated (with Distinction) from the MSc Geographic Information Sciences at University College London with

Professor Tao Cheng's supervision. He completed a BA (Hons) in Environment and Planning at the University of Liverpool in 2015.

Alex Singleton is a professor of geographic information science and Deputy Director of the ESRC Consumer Data Research Centre (CDRC) at the University of Liverpool, where he was appointed as a Lecturer in 2010. Previously he held research positions at University College London, where he was also awarded a PhD in 2007. He completed a BSc in Geography at the University of Manchester, graduating with a First-class honours degree in 2003.

Daniel Arribas-Bel is a lecturer in geographic data science at the Department of Geography and Planning, and member of the Geographic Data Science Lab, at the University of Liverpool (UK), where he directs the MSc in Geographic Data Science.

ORCID

Yunzhe Liu http://orcid.org/0000-0002-7189-3323
Alex Singleton http://orcid.org/0000-0002-2338-2334
Daniel Arribas-Bel http://orcid.org/0000-0002-6274-1619

References

Abdi, H., and L. Williams. 2010. "Principal Component Analysis." *Wiley Interdisciplinary Reviews: Computational Statistics* 2 (4): 433–459. doi:10.1002/wics.101.

Adnan, M. 2011. "Towards Real-Time Geodemographic Information Systems: Design, Analysis and Evaluation." PhD thesis, University College London.

Alexiou, A. 2016. "Putting "Geo" into Geodemographics: Evaluating the Performance of National Classification Systems within Regional Contexts." PhD thesis, University of Liverpool.

Bassett, K., and J. Short. 1980. *Housing and Residential Structure: Alternative Approaches*. London: Routledge.

Charlton, M., S. Openshaw, and C. Wymer. 1985. "Some New Classifications of Census Enumeration Districts in Britain: A Poor Man's ACORN." *Journal of Economic and Social Measurement* 13 (1): 69–96.

Debenham, J. 2002. "Understanding Geodemographic Classification: Creating the Building Blocks for an Extension." Accessed May 26 2017. http://eprints.whiterose.ac.uk/5014/1/02-1.pdf

Everitt, B., S. Landau, M. Leese, and D. Stahl. 2011. *Cluster Analysis*. 5th ed. London: Wiley.

Financial Network Analytics. 2012. "Correlation Networks." Accessed February 16 2018. https://blog.fna.fi/knowledge_center/page/2/

Gale, C., A. Singleton, A. Bates, and P. Longley. 2016. "Creating the 2011 Area Classification for Output Areas (2011 OAC)." *Journal of Spatial Information Science* 12: 1–27.

Guyon, I., and A. Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning* 7 (8): 1157–1182.

Harris, P., C. Brunsdon, and M. Charlton. 2011. "Geographically Weighted Principal Components Analysis." *International Journal of Geographical Information Science* 25 (10): 1717–1736. doi:10.1080/13658816.2011.554838.

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

Ismail, K., N. Nayan, and S. N. Ibrahim. 2016. "Improving The Tool for Analyzing Malaysia's Demographic Change: Data Standardization Analysis to Form Geo-demographics Classification Profiles Using K-means Algorithms." *Geografia Malaysian Journal of Society and Space* 12 (6): 34–42.

Jolliffe, I. 1972. "Discarding Variables in a Principal Component Analysis. I: Artificial Data." *Applied Statistics* 21 (2): 160–173.

Jolliffe, I. 2002. *Principal Component Analysis*, 1–10. 2nd ed. New York: Springer.

Kaiser, H. 1960. "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement* 20 (1): 141–151. doi:10.1177/001316446002000116.

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

Longley, P. 2005. "Geographical Information Systems: A Renaissance of Geodemographics for Public Service Delivery." *Progress in Human Geography* 29 (1): 57–63. doi:10.1191/0309132505ph528pr.

Longley, P., and M. Goodchild. 2008. "The Use of Geodemographics to Improve Public Service Delivery." In *Managing to Improve Public Services*, edited by J. Hartley, C. Donaldson, C. Skelcher, and M. Wallace, 176–194. Cambridge: Cambridge University Press.

Murphy, S., and M. Smith. 2014. "Geodemographic Model Variable Selection Spatial Data Mining of the 2011 Irish Census." 2014 IEEE International Advance Computing Conference (IACC). IEEE Xplore Digital Library, Gurgaon, India, February 21–22, 613–622.

Office for National Statistics. 2015. "Pen Portraits and Radial Plots." Accessed July 20 2018. https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications/penportraitsandradialplots

Openshaw, S., and C. Wymer. 1995. "Classifying and Regionalising Census Data." In *Census User's Handbook*, edited by S. Openshaw, 353–361. Cambridge: Geoinformation International.

Openshaw, S., M. Blake, and C. Wymer. 1995. "Using Neurocomputing Methods to Classify Britain's Residential Areas." *Innovations in GIS* 23: 97–111.

Pacheco, E. 2015. *Unsupervised Learning with R*, 111–146. Birmingham: Packt Publishing.

Robinson, G. 1998. *Methods and Techniques in Human Geography*. Chichester: John Wiley & Sons, Ltd.

Rojas, R. 2015. "The Curse of Dimensionality." Accessed May 30 2019. https://www.inf.fu-berlin.de/inst/ag-ki/rojas_home/documents/tutorials/dimensionality.pdf

Santero, K., N. Nayan, and S. Ibrahim. 2016. "Improving the Tool for Analyzing Malaysia's Demographic Change: Data Standardization Analysis to Form Geo-Demographics Classification Profiles Using K-Means Algorithms." *Geografia - Malaysian Journal of Society and Space* 12 (6): 34–42.

Singleton, A. 2016. "Cities and Context: The Codification of Small Areas through Geodemographic Classification." In *Code and the City*, edited by B. Kitchin and S. Perng, 215–235. New York: Routledge.

Singleton, A., and P. Longley. 2009. "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. doi:10.1111/pirs.2009.88.issue-3.

Singleton, A., and P. Longley. 2015. "The Internal Structure of Greater London: A Comparison of National and Regional Geodemographic Models." *Geography and Environment* 2 (1): 69–87. doi:10.1002/geo2.7.

Singleton, A., and S. Spielman. 2013. "The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom." *The Professional Geographer* 66 (4): 558–567. doi:10.1080/00330124.2013.848764.

Sleight, P. 1993. *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*. Henley-on-Thames: NTC Publications.

Spielman, S., and A. Singleton. 2015. "Studying Neighborhoods using Uncertain Data from the American Community Survey: A Contextual Approach." *Annals of the Association of American Geographers* 105 (5): 1003–1025. doi:10.1080/00045608.2015.1052335.

Alelyani S., J. Tang, and H. Liu. 2014. "Feature selection for clustering: A review." In *Data Clustering: Algorithms and Applications*, edited by C. C. Aggarwal and C. K. Reddy, Boca Raton, FL: CRC Press.

Timms, D. 1971. *The Urban Mosaic: Towards a Theory of Residential Differentiation*. Cambridge: Cambridge University Press.

Udovičić, M., K. Baždarić, L. Bilić-Zulle, and M. Petrovečki. 2007. "What We Need to Know When Calculating the Coefficient of Correlation?" *Biochemia Medica* 17 (1): 10–15. doi:10.11613/BM.2007.002.

Vickers, D., and P. Rees. 2007. "Creating the UK National Statistics 2001 Output Area Classification." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170 (2): 379–403. doi:10.1111/rssa.2007.170.issue-2.

Webber, R. 1975. "Liverpool Social Area Study 1971 Data." *PRAG Technical Paper 14*. London.

Webber, R., and J. Craig. 1978. *Socio-Economic Classifications of Local Authority Areas (Studies on Medical and Population Subjects)*. London: Office of Population, Censuses and Surveys.

Webber, R., and R. Burrows. 2018. *The Predictive Postcode: The Geodemographic Classification of British Society*. London: SAGE Publications Ltd.