JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 12 (2016), pp. 1–27
doi:10.5311/JOSIS.2016.12.232

RESEARCH ARTICLE

Creating the 2011 area classification for output areas (2011 OAC)

Christopher G Gale[1], Alexander D Singleton[2], Andrew G Bates[3], and Paul A Longley[4]

[1]Administrative Data Research Centre for England, University of Southampton, UK.
[2]Department of Geography and Planning, University of Liverpool, UK.
[3]Office for National Statistics, Titchfield, Fareham, Hampshire, UK.
[4]Department of Geography, University College London, UK.

Received: May 13, 2015; returned: July 7, 2015; revised: November 26, 2015; accepted: February 21, 2016.

Abstract: This paper presents the methodology that has been used to create the 2011 Area Classification for Output Areas (2011 OAC). This extends a lineage of widely used public domain census-only geodemographic classifications in the UK. It provides an update to the successful 2001 OAC methodology, and summarizes the social and physical structure of neighborhoods using data from the 2011 UK Census. The results of a user engagement exercise that underpinned the creation of an updated methodology for the 2011 OAC are also presented. The 2011 OAC comprises 8 Supergroups, 26 Groups, and 76 Subgroups. An example of the results of the classification in Southampton is presented.

Keywords: geodemographics, cluster analysis, k-means, 2011 UK census

1 Introduction

Geodemographic classifications provide summary indicators of the social, economic, demographic, and built characteristics of small areas. Within the UK there is a lineage of freely available geodemographic classifications covering small to larger geographies that have been built from census data outputs. The earliest published work on geodemographics focused on a single city (Liverpool, UK); but later was expanded to create classifications with national coverage [62, 64, 63]. Similar classifications were also created for the 1981 [16] and 1991 UK Censuses [9, 10]. The 2001 UK Census was the first to be released with open access as opposed to licensed distribution, thus removing a constraint that had previously

restricted the creation of derivative products. The 2001 Output Area Classification (2001 OAC) [56] and 2001 Area Classifications for other geographies could therefore be used in commercial or non-commercial applications without complex or costly licensing restrictions.

The creation of the 2001 OAC in the mid-2000s was a continuation of the legacy of free and open geodemographic classifications in the UK [50]. Singleton and Spielman [50] have argued that this legacy has had a stimulus effect on the user community, opening up a wider range of applications than within other geographical jurisdictions. The 2001 OAC in particular can be seen as a catalyst for the development of open classifications in the UK, with increased prevalence in use of geodemographic in the public sector since the mid-2000s [33]. Applications that have developed over a variety of substantive areas include, but are not limited to: health [45]; education [52, 53]; law enforcement; [5] and local governance [8].

The fundamental component of all of these geodemographic classifications is the data they used, with a particular focus on each using the newest and most suitable data available. While the data used by these systems is reflective of when they were constructed, the actual process used to turn the raw data into a classification has not fundamentally changed since the 1970s. It can be argued that all geodemographic classifications are derivatives of the same process of data acquisition, data manipulation and transformation, and cluster analysis. As such, geodemographic classifications may be differentiated by: choice of input data; data transformation and standardization procedures; clustering methods; and availability of associated descriptive "pen portrait" materials. The discussion that follows pertains to a UK geodemographic classification but systems have also become established globally, with applications in South Africa [12], Nigeria [42], Japan [4], Italy [65], Spain [24], Australia [59], and the United States [50].

Building a geodemographic classification can be considered as both art and science [26], guided by changing user requirements as well as the content and coverage of available data. These influences impact upon the subjective choices and predilections of the classification builder [26, 62], guiding the methodological approaches undertaken to produce a usable classification. As such, while the creation of any new classification will follow broadly similar processes to those that have gone before, the precise methods used should be chosen in the light of user requirements and sensitivity analysis [13].

This paper describes the creation of a new open geodemographic classification of the UK, comparable to the 2001 OAC, developed using 2011 UK Census data. The work was carried out at University College London (UCL) as a knowledge exchange activity co-sponsored by the UK Office for National Statistics (ONS). The final product has been designated as a key output of the 2011 UK Census by ONS. A requirement of ONS was that the new classification should be an evolution of the previous 2001 OAC, using the methodology outlined in Vickers and Rees [56] as a guide. The strong steer of ONS at the start of the project was that the 2001 OAC had been well received by a broad user base, and thus that it would have been imprudent to jettison all of its methodological components. This is not dissimilar to the approach taken by commercial geodemographic systems, although there are exceptions—for example, recent releases of CACI Ltd.'s (London, UK) Acorn no longer uses census data as its staple data source, choosing instead to use a mixture of other open data, Freedom of Information requests, and data models [15]. Compared to this, it is correct to describe the development of the new classification as evolutionary rather than revolutionary, albeit with greater focus upon user requirements, data selection and trans-

formation issues, computational methods, and new methods of online visualization and dissemination.

Testing and evaluation of multiple methodological approaches formed a core component of this build process, with particular emphasis on exploring how interactions between different methods and techniques influenced the final cluster solutions. Developments in geodemographics over the past decade, along with advances in computer software and processing power, make it possible to address some of the methodological issues identified with the 2001 OAC [57], whilst providing methodological advances beyond this widely-used predecessor. The incorporation of the needs of those who use open geodemographic systems, as voiced through a user engagement exercise, formed a key component when developing the new classification, along with the expanded range of outputs provided by the 2011 UK Census when compared to previous UK censuses. This was all underpinned by the decision to develop and use only open source software and to release all outputs, such as code and metadata, of the classification once completed.

Alongside the methodological enhancements and adoption of open software, the overarching aim of this new classification, as with the 2001 OAC, remained: to describe the salient and multidimensional characteristics small areas across the UK, as represented by the UK Census; to provide a usable classification, with a fully transparent and reproducible methodology; and to provide an alternative to commercial geodemographic classifications, the cost of which might be prohibitive for potential end users. The methodological underpinning of this new classification subsequently formed the basis of other official geodemographic classifications created from 2011 UK Census data, such as the 2011 Area Classification for Local Authorities [40].

2 The methodology

The initial task in creating a new geodemographic classification of 2011 UK Census data was to identify appropriate methods for data manipulation and cluster analysis. These methods could then be empirically tested to determine which combination created a "best" solution. In a narrow analytical sense, an optimum solution may be defined as one in which the data points representing output areas are most tightly clustered around the seeded cluster sites. More generally it should provide a representation of the population that is useful to users of the classification. It is this combination of quantitative and qualitative measures that allows geodemographics to be considered to be as much art as it is science [26].

Data manipulation procedures are used to prepare the data prior to clustering. They are vital to ensure that all variables are measured on comparable scales to ensure equal weighting and to control for underlying population structure. In practice this requires a process of rate or ratio calculation, followed by transformation if non-normality is identified as a counter to effective cluster formation. Furthermore, as in the 2001 OAC, a standardization process may also be implemented to place the variables onto a single scale.

The 2001 OAC methodology converted 2001 UK Census counts into percentages, applied a log transformation, and then tested three standardization procedures [56]. In the analysis reported here, a greater number of transformation and rate calculation techniques were evaluated, made possible by using high performance computing and open source software to evaluate multiple combinations of options.

The two main methods of rate calculation that are commonly used in the creation of geodemographic classification are percentages and index scores. These present conceptually different approaches to classification: percentages compare areas based on rates for a particular attribute, whilst index scores compare areas based upon their difference from the national average. A third option considered was the calculation of mean differences, to identify variables with the greatest deviation away from average characteristics. These distinct rate calculation procedures create unique datasets, although the experience of 2001 OAC suggested that distributions of each dataset were unlikely to be normally distributed [56]. Highly skewed data can lead to poor assignments if clustered with algorithms that are optimized to find spherical groupings of cases with similar attributes. A critical evaluation of the different transformation techniques was therefore a vital step.

Three different transformations were investigated: $\log 10$, Box-Cox, and inverse hyperbolic sine. The Box-Cox and $\log 10$ transformations both require values to be positive and greater than 1 to effect transformation. There are different ways of managing this issue, the most common of which is to add a constant to all values prior to applying the transformation method [43]. Log transformations artificially reduce the amount of variance to that of the normal distribution by compressing the differences between the larger values and increasing those between smaller values [32]. A disadvantage of this approach is that the transformation is fixed across a dataset without sensitivity to different distributions that may appear between variables. An alternative method, which is more sensitive to these issues, is the Box-Cox transformation, which can be defined as:

$$x_i(\lambda) = \begin{cases} (x_i^\lambda - 1)/\lambda & (\lambda \neq 0) \\ \log(y) & (\lambda = 0) \end{cases} \tag{1}$$

An exponent, lambda (λ), transforms a variable (x) into a normal distribution [11]. Multiple values for lambda are tested, and the one that produces the most normal result is selected. There are numerous tests of normality that can be applied, and for the implementation here we used the Shapiro–Wilk test. A lambda value of 0 is mathematically equivalent to $\log 10$ transformation. As such, the implementation of the Box-Cox technique calculates a separate lambda value for each variable. The extent to which a variable is transformed will however depend on how skewed its distribution is, rather than a global skewness value [18].

The third approach does not require the addition of a constant and takes the form of the inverse hyperbolic sine, defined as:

$$\log\bigl(x_i + (x_i^2 + 1)^{1/2}\bigr) \tag{2}$$

The inverse hyperbolic sine (IHS), proposed by Johnson [29] shares similarities with the standard $\log 10$ transformation, except that it can be defined at zero or for negative numbers [14]. As a result, the technique is often favoured when transforming wealth datasets [44]. One disadvantage relative to the Box-Cox transformation is that IHS, like $\log 10$, still maintains a single transformation applied uniformly to all attributes.

The evaluation of three transformation methods, instead of the single method used with the 2001 OAC, allowed the impact of each technique to be compared and contrasted. This was of particular importance within the context of the new classification, as it needed to be a product of the data itself, and not the quirks of a particular transformation technique. Finally, different data standardization methods were tested. Standardization places each

variable in the data set onto the same scale. There are a number of different methods through which this can be achieved, and as with the 2001 OAC, z-scores, range standardization, and inter-decile range standardization were evaluated.

Z-scores are one of the most widely used approaches of variable standardization [1]. Z-scores transform raw scores in terms of the number of standard deviations they are away from the mean, which can have the effect of assigning high leverage to outlying observations in the data. Range standardization compresses the values in a dataset into the range of 0 to 1 and was used for ONS 1991 classification of local authorities (see [60]) and for the 2001 OAC (see [56]). Finally, the inter-decile range standardization method is a variant of range standardization, in which the data is standardized over a smaller range, between the 90^{th} percentile and the 10^{th} percentile, in order to reduce the impact of outliers on the standardized data.

The three different standardization methods when combined with the three respective rate calculation and transformation techniques create a total of 27 unique data combinations which were evaluated for the new classification. This represents a substantial increase from the three different combinations that were tested for the 2001 OAC. The greater number of permutations tested can allow users of the final classification to have greater confidence that the end product provides a realistic geodemographic representation of the UK's population.

Evaluation of the 27 unique datasets used metrics such as correlation analysis and skewness, however, clustering of the different datasets was required to fully ascertain how each combination would impact upon the final classification. As such, a cluster algorithm with distance measure was required. There are numerous algorithms that can be used, which can be grouped into four types [27, 30]: partitional, hierarchical, density-based, and grid-based. The k-means algorithm [35] is a form of partitional clustering that involves an iterative process that operates on a fixed number of clusters [54]. A number of distance measures, the method by which distance between objects is measured, can be used with k-means. Most common are Euclidean distance and squared Euclidean distance (SED), although many more are available [21]. The algorithm has been used to create a number of different classifications, including the 2001 OAC [56] and a geodemographic classification of the United States [51]. Hierarchical clustering is an agglomerative, or bottom up, approach to clustering. Ward's hierarchical clustering algorithm [61] in particular has been used to construct past classifications, however it was discounted for use with the 2001 OAC because it can be unsuitable for large datasets [56]. The plethora of clustering techniques available (see Arabie et al. [3], Gordon [25], and Xu and Wunsch [66]) makes identifying the most suitable algorithm or distance measure for a particular task challenging. As such, rather than testing multiple algorithms, it was decided to select a method based on the steer from ONS and the requirements identified as part of a user consultation discussed in the next section. In addition, the testing of multiple data manipulation methods with multiple clustering algorithms would have resulted in too many potential classifications to objectively evaluate. As such, priority was given to identifying the optimum methods of preparing the data rather than how it is partitioned.

Aside from these computational and methodological enhancements, an aim of the new system was to devise a classification that could be more readily understood, evaluated, and adapted for bespoke uses than its predecessor. Although the 2001 OAC devised a commendably open methodology, it was restrictive in that it was produced in SPSS (a commercial statistical package with license requirements). Advances in the number and quality

of free open source programs over the past decade have made their use in the creation of the new classification an apt choice. It was decided to use the command line program and language R [46] which, unlike SPSS, allows universal access, free from any licensing restrictions. Fully documenting the process and publishing the associated code is beneficial for a number of reasons: the classification becomes fully reproducible; bespoke variants are easier to create; and the entire process can be examined and critiqued.

Taken together, these developments move open geodemographics further towards complete transparency. A number of key decisions were however still required, such as the extent to which non-census data sources could be included within the classification. Consultation with stakeholders was therefore vital in shaping how the methods outlined above would be used to create a new open geodemographic classification of the UK.

3 The consultation process

The creation of a new classification for the UK required a number of key decisions to be made. While the final analytical decisions would be the preserve of those directly involved with the project, it would have been inappropriate to eschew engagement with end users in order to ascertain what their needs and priorities were. As such a stakeholder consultation exercise and design evaluation, carried out at UCL in collaboration with ONS, was an integral part of creating the new classification. The results were used to guide the final choice amongst the candidate methods detailed in the previous section, although specific details were only finalized after the data had been selected and exploratory empirical analysis had been conducted.

The first stage of the user engagement entailed a pilot study with ONS at the Demographics User Group's (DUG) annual conference. DUG are a consortium of retail organizations that lobby government to open up access to data. The DUG annual conference is attended by individuals from a range of backgrounds, including business, academia, and local and central government; many of which have extensive experience of both the 2001 OAC and commercial geodemographic systems.

This pilot study led to the creation of a finalised online self-completion questionnaire. This was promoted through specific end user websites and mailing lists of potential stakeholders drawn from academia, central and local government, commercial organizations, and the health sector. Responses varied by stakeholder group, with the key findings presented in Table 1 and a detailed report published by ONS in May 2012 [38]. The results of the survey helped to refine the methodology approach for the new classification to best meet user requirements while not compromising the integrity of the final classification.

A consensus from the user engagement was that the new classification should be created at the smallest areal level with additional open data sources supplementing 2011 UK Census data. However, the dearth of open data currently available at the finest granular areal level meant that incorporating these datasets would have created too many compatibility issues with the census data, given that sources with complete UK coverage were only available at coarser levels of granularity. Therefore, the creation of a classification at the smallest areal level was prioritised, meaning only 2011 UK Census data at the output area (OA) level in England, Wales, and Scotland and small areas (SAs, introduced for the 2011 Census) in Northern Ireland were considered for use. This decision should not, however, preclude use of open data in extended models based on the same methodology, for exam-

General

- Better discrimination *vis-à-vis* commercial geodemographics; especially within London and rural areas.
- Some users are deterred from using commercial packages due to the cost.
- The open source approach is viewed as a positive attribute within certain application areas.

Methods

- Creating the new classification at the smallest areal level was preferable.
- Integrating census data with wider open data sources was considered desirable.
- No need for the new classification to be directly comparable with the 2001 OAC.
- A general purpose classification favored.
- No consensus on an appropriate geographic extent for the new classification.

Outputs

- Better promotion of the new classification to stakeholder groups.
- User friendly outputs should be included, such as pictures and pen portraits.
- Descriptive graphs and written material useful in providing greater understanding of the classification.
- Interactive maps considered a useful enhancement to the classification.
- Cluster naming very important.
- A measure of uncertainty for cluster assignment desired as an additional product.

Table 1: Summary responses from the user engagement exercise [38].

ple, in constrained geographic extents (e.g., a local authority) or bespoke domain specific applications [49].

Another important finding was that although broad comparability with the 2001 OAC was desirable, a like-for-like replacement was not. This allowed for flexibility with the variable selection and formed the basis of the rationale for using the k-means algorithm, with the SED as the distance measure, to perform the cluster analysis. Respondents had expressed a preference for the 2001 OAC structure, which was largely due to the three-tiered classification created by the k-means algorithm using SED. Using the same clustering method for the new classification maintained this familiar structure.

4 Initial variable selection

The decision to use only 2011 UK Census data meant the new classification, or the 2011 Area Classification for Output Areas (2011 OAC), required outputs from three different UK census agencies. Outputs of the 2011 UK Census at the OA and SA level were released in stages by ONS for England and Wales, National Records of Scotland (NRS) for Scotland, and the Northern Ireland Statistics and Research Agency (NISRA) for Northern Ireland. The outputs that were most appropriate for creating the 2011 OAC were those provided in univariate format. This data was less likely to be impacted by the perturbations linked with disclosure control methods applied to the 2011 UK Census [37], where values were modified if the raw data could conceivably be used to identify specific individuals, households or businesses. For the 2011 UK Census data these were identified by the statistical

bodies as Key and Quick univariate statistics, with Key Statistics providing summaries and Quick Statistics being more detailed. ONS released 35 tables as Key Statistics and 73 tables as Quick Statistics for England and Wales. NISRA released 45 tables as Key Statistics and 58 tables as Quick Statistics for Northern Ireland. By December 2013 NRS had released 34 tables as Key Statistics and 59 tables as Quick Statistics for Scotland. The combined dataset for England and Wales contained 2,139 variables, in Scotland there were 1,326 and in Northern Ireland 1,378. Only variables that were consistent across the whole UK were considered for use in creating the 2011 OAC, however, within this reduced dataset, there were numerous cases of duplicated variables (for example, tables KS101EW and QS101EW contained identical data regarding the usual resident population of England and Wales).

The objective of the 2011 OAC variable selection process was to obtain the smallest subset of variables that captured the main variations within the 2011 UK Census, consistent with the broad approach used in previous classifications [6, 7]. As with the 2001 OAC, candidate variables were classified as belonging to one of five domains that aimed to best represent drivers of socio-spatial differentiation in the UK: Demographic Structure, Household Composition, Housing, Socio-Economic, and Employment [56]. Keeping the domains consistent with the 2001 OAC allowed for a similar variable structure to be maintained for the 2011 OAC, without restricting the final variable selection to being merely a replication of the previous classification. This also accommodates a key finding of the 2011 OAC user engagement exercise, which highlighted that although broad comparability was desirable, a like-for-like replacement of the 2001 OAC was not.

The output of this initial variable filter was 167 prospective variables that were used as the basis for the final attribute selection. This variable set aimed to assure coverage over the pre-identified domains. The 94 variables initially considered for the 2001 OAC [56] guided the selection, whilst also accommodating the expanded number of outputs available from the 2011 UK Census, such as the increased number of ethnic group categories. A total of 166 variables were taken directly from the 2011 UK Census, with an additional variable derived from a number of census outputs. This additional variable, a Standardised Illness Ratio (SIR) measure, compares observed illness counts within an area relative to expected values accounting for underlying age structure. The use of the health variables contained within the 2011 UK Census without modification was not considered, as they do not account for the age structure of an area having a significant impact on recorded illness rates. As such, areas that contain a high concentration of older individuals are more likely to be associated with higher illness rates than areas containing a high proportion of younger people if values are not standardized.

The suitability of the 167 prospective variables to represent drivers of socio-spatial differentiation across the UK and therefore provide the basis of the new classification was assessed by ONS. Exploratory analysis was also undertaken, with histograms being used to explore the degree to which the attribute data was normally distributed, with the majority of variables exhibiting varying degrees of skew. A large number of outliers also existed towards the high end of the value scale, notably concerning population density. Retaining such variables unmodified would have been undesirable as Kumar et al. [31] note, the k-means clustering algorithm can be adversely affected by skewed data distributions, because Euclidean distances are calculated between points and cluster centroids, meaning the algorithm is optimised to find spherical groupings of cases with similar attributes [28]. A choice was made with the 2011 OAC to use the different combinations of data manipulation and transformation procedures outlined in section 2 to reduce the impact of skewed

variables. Singleton and Spielman [50] do however caution that global normalization may smooth away interesting local patterns, albeit that some such patterns may reflect the effects of outlying observations. Harris et al. [26] outline an alternative method found in commercial geodemographics systems, where skewness is accommodated through weighting. The inherent subjectivity of such an approach however led to its rejection in the creation of the 2001 OAC and with the 2011 OAC.

5 Final attribute selection and testing

The initial selection process returned 27 prospective variable datasets generated through combinations of different rate calculation, transformation, and standardization techniques. To select a final set of attributes for use with cluster analysis, further analysis was required. Each of the 27 datasets underwent additional testing, with the results from each contributing to the final selection of variables. However, for the sake of clarity only the results from the percentage rate calculation, IHS transformation, and range standardization prospective variable dataset are presented in this section to illustrate the methods used.

The process of selecting the final variables took into consideration two key requirements. The first was for the 2011 OAC to be a general-purpose geodemographic classification. This required inclusion of a collection of variables that reflected the general characteristics of the UK's population, whilst also emphasizing characteristics that varied between OAs and SAs in order for distinct clusters to form. The second requirement was inclusion of the minimum number of variables in order to limit any potential weighting effect arising from co-linearity within the final list of inputs. The process of reducing the prospective variables to a final list of inputs was conducted with the overarching aim of retaining those variables that were likely to be the most important for the 2011 OAC, and removing those that would only add limited information to the final assignment of areas into clusters. Two empirical methods were utilized to guide the selection of final inputs for cluster analysis. The first explored variable correlation, and the second implemented a clustering based sensitivity analysis technique based on total within cluster sum of squares (WCSS) values to give an indication of those variables that are important when forming output clusters.

It can be expected within any large multidimensional dataset that there will be an element of correlation between some variables. Correlated variables act in a similar way to a weight, giving added prominence to a shared dimension. Although most correlated variables were removed from the 2001 OAC, some which were considered highly correlated were retained for use. This was done in order to add predictive and descriptive power to the classification. There were three options available after identifying highly correlated variable pairs: remove the variable from the final selection; group it with other variable(s) to create a new composite variable; or disregard the correlation. For the prospective variables, a Pearson correlation matrix was created, and those variables that had correlation of greater than 0.6 or less or -0.6 were examined. The percentage calculation, IHS, and range standardization dataset had a total 104 variables with correlations in excess of these values with at least one other variable within the dataset. Table 2 is a summary of the ten variables that showed the greatest correlation with the other prospective variables. The three potential options to address these intercorrelations were carefully considered, with

the most appropriate action dependent on the importance of each variable to the classification.

Variable Name	*Number of variables with intercorrelation values greater than* 0.6 *or less than* -0.6
Households that only contain persons aged over 16 who are living in a couple and are married	17
Persons aged over 16 who are single	16
Persons aged over 16 who are married	16
Mean age	14
Persons whose country of birth is the UK	14
Households that only contain persons aged over 16 who are not living in a couple and are single (never married or never registered a same-sex civil partnership)	14
Households with two or more cars or vans	14
Median age	13
Persons who are white British and Irish	13
Households who have two or more rooms than required	13

Table 2: Intercorrelation frequency of the 2011 OAC prospective variables.

A second consideration before final attribute selection was identification of those variables that would have the greatest impact on cluster formation. Clustering of the 2011 OAC was completed using the previously discussed k-means algorithm [35] and maintains broad comparison with the 2001 OAC by using the same methodological approach. This algorithm initialises with k "seeds" randomly placed within the multidimensional attribute space of the input dataset, and the OAs and SAs are assigned to their closest seed, thereby creating an initial cluster assignment. Cluster centroids are then recalculated as the average of the attribute values for all data points assigned to each cluster. The data points are then reassigned if they become closer to new cluster centroid. This process is repeated until no further data points move between clusters, thus meeting the convergence criterion. This process aims to create clusters that are as homogenous as possible with variability within each cluster minimized, with the composition of the final n clusters being as heterogeneous as possible based on the initial seed assignment.

Given the initialization procedures, the k-means algorithm is stochastic, and as such, multiple runs are required with randomly allocated initial seeds. The WCSS and total between cluster sum of squares (BCSS) are calculated as part of the clustering process. These values indicate how tightly clustered a particular dataset is. The WCSS value signals how close objects within each cluster are to their centroids, thereby providing an indication of cluster homogeneity. The BCSS value measures the distances between the clusters, and as such, quantifies how similar they are to each other. In constructing the 2011 OAC, we were guided by the WCSS statistic, given our emphasis upon identifying homogenous groups within the UK's population rather than ensuring that clusters were as dissimilar to each

other as possible. Consequently, the WCSS statistic was used as a global goodness-of-fit criterion to identify optimised cluster solutions. This was done by selecting the cluster solution with the lowest WCSS value after running the k-means algorithm 10,000 times, using multiple random seed sites to generate different outcomes [49].

At this stage, an optimal number of k clusters for the classification was not known, and $k = 6$ to $k = 8$ were selected for use in the sensitivity analysis as these corresponded to the most aggregate level of other UK geodemographic classifications, although these numbers were otherwise arbitrary. For each value of k a different variable was held back each time. This enabled identification of those individual attributes that impacted cluster performance by examining the WCSS and BCSS values. Although the tests were run for $k = 6$ to $k = 8$, the results identified similar variables having impact on the clustering performance. As an example, for $k = 8$, the WCSS results returned as a result of removing each of the variables in the percentage calculation, IHS, and range standardization dataset are shown in Figure 1, and a summary of the variables that influence cluster homogeneity the most and least are shown in Table 3. Where the removal of a variable resulted in a marked decrease in the WCSS value and a larger BCSS value, this suggested that the omission of that variable would result in more homogenous clusters and greater heterogeneity between clusters. While this indicated that the variable should be discarded, the decision whether or not to do so was dependent on the importance of the variable in capturing key characteristics of the 2011 UK Census; for example, it was important a number of housing variables were retained as they detailed the built-up environment of areas. Comparatively small decreases in the WCSS value and increases in the BCSS value for a variable suggested that the impact on the homogeneity of the clusters and heterogeneity between clusters was minimal, increasing the likelihood that the variable would be retained. Gale [23] provides a more detailed account of how this method was applied and the results analyzed across the different datasets.

The results from the correlation and sensitivity analysis provided a greater understanding of how individual variables would affect the final classification. However, these results were only a secondary consideration for the final variable selection, with the primary focus being upon ensuring the relevant population and household structure was captured. This meant a variable that could be empirically shown to have a negative impact on the global homogeneity of clusters was still considered if it represented a key facet of the five census data domains identified by Vickers and Rees [56]. For example, Table 3 highlights the negative impact that certain housing variables can have on the clustering process. It was however essential that a selection of these variables were included in the final classification to represent the built environment of the UK.

The application of these methods across the 27 datasets showed consistent results, with housing variables in each dataset having a greater propensity to negatively influence cluster homogeneity. The use of these quantitative techniques to aid the final selection of variables, alongside the use of more data manipulation and transformation methods, can be seen as an enhancement when compared to the 2001 OAC. The consistency shown across the different datasets is an indication that the results gathered are a true reflection of the patterns within the data and not an artificial artifact of a particular set of methods used, a possibility when only a limited number of different techniques are tested. This in turn makes the 2011 OAC variable selection more robust when compared to its predecessor.

In total, from the 167 prospective variables, 86 were removed; 41 were retained without being combined with other variables and 40 were merged with at last one other variable to

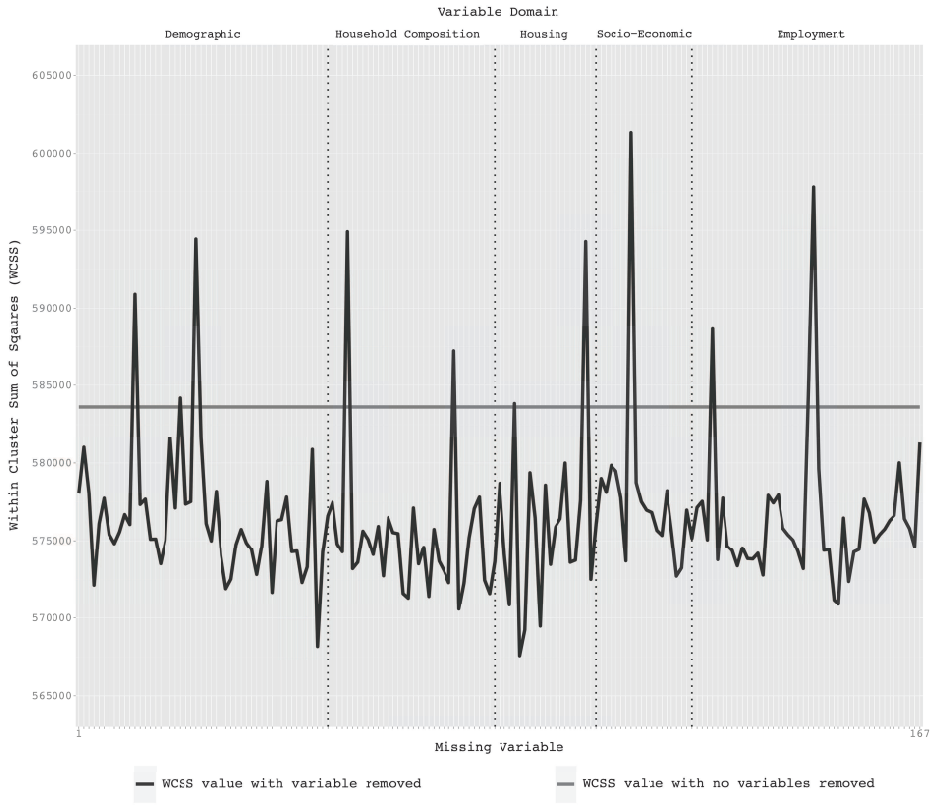

Figure 1: Total within cluster sum of squares values for the 167 variables initially selected for the 2011 OAC for $k = 8$ with the percentage calculation, inverse hyperbolic sine and range standardization dataset (See Gale [23] for variable names).

create 19 composite variables. This created a final set of 60 variables. Composite variables were created by combining variables which shared the same denominator. This was used most frequently to reduce inter-correlation within the dataset. Gale [23] provides a full explanation of why variables were removed, merged, or retained.

The total of 60 variables, compared to the 41 variables used by the 2001 OAC, offers additional dimensions to differentiate the UK's population. Arguably, this is a less parsimonious solution than that created in the 2001 OAC but it was decided that the benefit of increased differentiation merited this [21]. For example, an additional age variable is included to split those aged over 65 into two categories, to reflect the UK's ageing demographic [41]. Furthermore, the inclusion of a communal establishment indicator with the included age variables aims to make a distinction between areas where older members of society live independently *vis-à-vis* those who live within communal facilities such as residential care homes.

A greater range of demographic variables have also been included. These provide a more in-depth perspective on individuals' ethnic background, country of birth, and ability to speak the English language. Finally, more education variables have been incorporated,

Variables that have the most negative influence on cluster homogeneity when included in cluster analysis

- Households who live in a terrace or end-terrace house,
- Persons whose main language is not English but can speak English well,
- Households who live in a flat,
- Households who are private renting,
- Households with lone parent in part-time employment,
- Households who live in a detached house or bungalow,
- Employed persons aged between 16 and 74 who work in the financial and insurance activities industries,
- Employed persons aged between 16 and who work in the information and communication industry,
- One family households containing cohabiting couples with dependent children,
- Households with dependent children.

Variables that have the least negative influence on cluster homogeneity when included in cluster analysis

- Persons providing unpaid care,
- Employed persons aged between 16 and who work in the construction industry,
- Households that only contain persons aged over 16 who are not living in a couple and classed as single (never married or never registered a same-sex civil partnership),
- Persons aged over 16 who are in a registered same-sex civil partnership,
- Households with over 0.5 and up to 1.0 persons per room,
- Persons aged 25 to 29,
- Persons aged between 16 and 74 who unemployed and economically active,
- Households with no adults in employment and no dependent children,
- Employed persons aged between 16 and who work in the water supply; sewerage, waste management, and remediation activities industries,
- Median age.

Table 3: Variables that influence cluster homogeneity the most and least based on WCSS values in the percentage calculation, inverse hyperbolic sine, and range standardization dataset.

along with an expanded number of employment industry types. A full list of the 60 variables are provided by ONS [39], with additional accuracy of the finalized dataset assured through rigorous testing and validation by ONS prior to clustering.

6 Dataset selection and assessing similarity between areas

The previous section outlined how the final set of 60 variables used to create the 2011 OAC was formulated. However, a further decision on what set of rate calculation, data transformation, and standardization methods to be used to create the final classification was required. This decision was also co-dependent on the number of clusters to be included in the new classification and what its structure would be. As with the 2001 OAC, the intention was to build a hierarchical classification to provide greater flexibility for potential applications. A top down approach was adopted, whereby clusters are created for the most aggregate level of the classification, and then used to subdivide the input data, which are then successively clustered separately, forming the hierarchical levels of the classification.

It was argued with the 2001 OAC that this method was favorable because the objects of study (OAs and SAs) are always clustered, rather than cluster centroids, as might be the case with "bottom up" methods implemented using alternate hierarchical algorithms such as Wards clustering. This is important because cluster centroids will only rarely be representative of an entire cluster. Going up a hierarchical level and clustering using these centroids can create clusters containing objects with little in common, and thus result in clusters with low homogeneity.

The 2001 OAC adopted a three-tiered hierarchical approach, and the cluster frequency at each level was deemed satisfactory by the 2011 OAC user engagement. However, there was also an expectation that the new classification would not necessarily have identical numbers of clusters at each level. The 2011 OAC therefore aimed to have similar, but not necessarily identical numbers in a three-tiered structure with the Supergroup, Group, and Subgroup terminology being retained.

There is no over-all consensus as to how to structure a geodemographic classification. Singleton and Spielman [50] point out that three tier hierarchies tend to predominate in UK national classifications, albeit with different number of clusters per level. With no common methodological approach, the number of clusters required to best represent a population is unique to every classification; and deciding upon the optimum number of clusters is an inherently subjective process [49]. For the 2001 OAC, cluster numbers were decided upon following personal consultations with experts [55], cf. [58]. Decisions made with the 2001 OAC were therefore a balance between the ideal number of clusters identified by the experts, aligned with maximizing cluster integrity and ensuring reasonably even cluster size.

Formalising the structure of the 2011 OAC meant identifying which combination of data manipulation and transformation methods to apply to the final set of 60 variables, and, simultaneously, the number of clusters in each level of the hierarchy. Identifying the best solution for the 2011 OAC thus entailed exploratory analysis. Using the final set of 60 variables, 27 datasets were created using the same combination of data manipulation methods used to aid variable reduction. Different k seeds were used on each dataset to identify cluster solutions that created distinct clusters, provided a scattered geographic distribution across the UK, and also gave an even assignment of OAs and SAs across all levels of the hierarchy. To ensure a similar hierarchical structure to the 2001 OAC, the top level of the hierarchy was examined for 5 to 9 cluster options, while the middle and bottom levels had 2 to 4. Each cluster analysis was repeated 10,000 times to identify the optimum solution, in line with the method outlined in the previous section and other optimised geodemographics [49].

This analysis led to a dataset constructed using percentages for rate calculation, IHS for data transformation and range standardization with eight clusters forming the top of the hierarchy being selected to form the 2011 OAC. While this decision can be considered arbitrary, it can be justified within the context of the project. The selected dataset produced the most robust classification in terms of cluster homogeneity and differentiation between clusters, but also in terms of fulfilling user requirements as expressed during the engagement exercise reported in Table 1. In particular, respondents desire for a classification that offered good discrimination across the UK and offered improved discrimination within London when compared to the 2001 OAC, without having a detrimental impact on the functionality of the classification in other areas of the UK. The same processes were repeated for the second and third levels of the hierarchy, resulting in the final classification

having 8 Supergroups, 26 Groups, and 76 Subgroups. For a more detailed explanation regarding the formalization of the final dataset and selected cluster numbers see Gale [23].

The incorporation of user requirements, and analysis of multiple combinations of data manipulation methods and cluster solutions, demonstrates a notable advancement with the 2011 OAC methodology compared to its predecessor. However, while the combination of selected methods proved to be the best for the 2011 OAC, each technique individually had desirable traits in terms of building a geodemographic classification. Percentages compare similarities between places, rather than measure how places deviate from a global average, which is important where many attributes are known to be regionalised. The IHS method and log 10 methods act similarly by compressing differences between larger values relative to those between smaller values. This reduces the variability in a skewed dataset and provides a closer approximation to a normal distribution. However, unlike log 10, IHS can be defined at zero or for negative numbers [14]. Rather than adding a constant to all values of the 2011 OAC dataset, the IHS method accommodates the 14% of values that were zero without the need to modify them. Finally, range standardization has been shown to be less susceptible to outliers and skewed variables when compared to using z-scores and the inter-decile range [58], making it advantageous for use with the 2011 OAC dataset.

7 Describing UK geodemographic patterns

Cluster names and descriptions are an important aspect of the user interface of geodemographic classifications, and are intended to represent the underlying complexity of the cluster compositions. Names and short "pen portrait" descriptions were therefore developed for the final clusters making up each level of the 2011 OAC hierarchy. Vickers et al. [58] noted that names and descriptions of clusters may be contentious, especially if they reinforce negative stereotypes. Additionally, the procedure of conflating individual variables with one another opens up the risk of ecological fallacy within the analysis [47]. However, past experience in both the private and public sector confirms that this procedure does help end users to identify with the names and descriptions given to local areas.

Given the predominant public sector usage that was anticipated in creating the 2011 OAC, words and phrases that might be construed as pejorative were not used, and all descriptors had strong and literal links to the underlying distributions revealed by the data. Words that were overtly negative or positive were avoided where possible, and the abiding sense of the descriptors was to present the characteristics of each cluster as consequences of factors that have happened to areas rather than the consequences of human agency. This was consistent with avoiding value judgments when assigning cluster names and descriptions.

Names and descriptions of clusters are, of course, based on the characteristics of their centroids, and the workings of the cluster assignment procedure means that specific areas differ in the degrees to which they conform to these average characteristics. Attribute range statistics were thus used to avoid descriptors that pertained only to the OAs and SAs that were closest to their assigned cluster's centroid. This consideration was balanced by avoiding names that are too broad and too vague to offer any insight into an area's attributes. Additional checks were undertaken to ensure that names where possible did not duplicate those used by previous commercial and non-commercial geodemographic classifications,

and final approval of the cluster names was sought from ONS who conducted their own internal review and consultation.

The final names for the 2011 OAC are shown in Table 4, and Figure 2 presents an illustration of the 2011 OAC for the city of Southampton and the surrounding area in southern England at the Supergroup level. The mapping of clusters allowed for internal validation, where names were checked against local knowledge of areas. For example, the share of Southampton's total population classified as being white and being born outside the UK was 7.4% in 2011. The clusters in the 2011 OAC that were the most likely to contain persons with these characteristics matched areas of Southampton known for containing higher concentrations of such populations. Another example is clusters whose characteristic profiles reflect large proportions of students, and named accordingly, matched areas in Southampton known for their large residential student population. Finally, the clusters most likely to contain populations with higher levels of deprivation correspond to areas in Southampton identified as being the most deprived by the English Indices of Deprivation in 2010 and 2015 [19, 20]. The English Indices of Deprivation are composite measures primarily based on administrative data [20] alongside 2011 Census data.

Supergroups	*Groups*	*Subgroups*
1 - Rural Residents	1a - Farming Communities	1a1 - Rural Workers and Families
		1a2 - Established Farming Communities
		1a3 - Agricultural Communities
		1a4 - Older Farming Communities
	1b - Rural Tenants	1b1 - Rural Life
		1b2 - Rural White-Collar Workers
		1b3 - Ageing Rural Flat Tenants
	1c - Ageing Rural Dwellers	1c1 - Rural Employment and Retirees
		1c2 - Renting Rural Retirement
		1c3 - Detached Rural Retirement
2 - Cosmopolitans	2a - Students Around Campus	2a1 - Student Communal Living
		2a2 - Student Digs
		2a3 - Students and Professionals
	2b - Inner-City Students	2b1 - Students and Commuters
		2b2 - Multicultural Student Neighbourhoods
	2c - Comfortable Cosmopolitans	2c1 - Migrant Families
		2c2 - Migrant Commuters
		2c3 - Professional Service Cosmopolitans
	2d - Aspiring and Affluent	2d1 - Urban Cultural Mix
		2d2 - Highly-Qualified Quaternary Workers
		2d3 - EU White-Collar Workers

Supergroups	*Groups*	*Subgroups*
3 - Ethnicity Central	3a - Ethnic Family Life	3a1 - Established Renting Families
		3a2 - Young Families and Students
	3b - Endeavouring Ethnic Mix	3b1 - Striving Service Workers
		3b2 - Bangladeshi Mixed Employment
		3b3 - Multi-Ethnic Professional Service Workers
	3c - Ethnic Dynamics	3c1 - Constrained Neighbourhoods
		3c2 - Constrained Commuters
	3d - Aspirational Techies	3d1 - New EU Tech Workers
		3d2 - Established Tech Workers
		3d3 - Old EU Tech Workers
4 - Multicultural Metropolitans	4a - Rented Family Living	4a1 - Social Renting Young Families
		4a2 - Private Renting New Arrivals
		4a3 - Commuters with Young Families
	4b - Challenged Asian Terraces	4b1 - Asian Terraces and Flats
		4b2 - Pakistani Communities
	4c - Asian Traits	4c1 - Achieving Minorities
		4c2 - Multicultural New Arrivals
		4c3 - Inner City Ethnic Mix
5 - Urbanites	5a - Urban Professionals and Families	5a1 - White Professionals
		5a2 - Multi-Ethnic Professionals with Families
		5a3 - Families in Terraces and Flats
	5b - Ageing Urban Living	5b1 - Delayed Retirement
		5b2 - Communal Retirement
		5b3 - Self-Sufficient Retirement
6 - Suburbanites	6a - Suburban Achievers	6a1 - Indian Tech Achievers
		6a2 - Comfortable Suburbia
		6a3 - Detached Retirement Living
		6a4 - Ageing in Suburbia
	6b - Semi-Detached Suburbia	6b1 - Multi-Ethnic Suburbia
		6b2 - White Suburban Communities
		6b3 - Semi-Detached Ageing
		6b4 - Older Workers and Retirement

Supergroups	*Groups*	*Subgroups*
7 - Constrained City Dwellers	7a - Challenged Diversity	7a1 - Transitional Eastern European Neighbourhoods
		7a2 - Hampered Aspiration
		7a3 - Multi-Ethnic Hardship
	7b - Constrained Flat Dwellers	7b1 - Eastern European Communities
		7b2 - Deprived Neighbourhoods
		7b3 - Endeavouring Flat Dwellers
	7c - White Communities	7c1 - Challenged Transitionaries
		7c2 - Constrained Young Families
		7c3 - Outer City Hardship
	7d - Ageing City Dwellers	7d1 - Ageing Communities and Families
		7d2 - Retired Independent City Dwellers
		7d3 - Retired Communal City Dwellers
		7d4 - Retired City Hardship
8 - Hard-Pressed Living	8a - Industrious Communities	8a1 - Industrious Transitions
		8a2 - Industrious Hardship
	8b - Challenged Terraced Workers	8b1 - Deprived Blue-Collar Terraces
		8b2 - Hard-Pressed Rented Terraces
	8c - Hard-Pressed Ageing Workers	8c1 - Ageing Industrious Workers
		8c2 - Ageing Rural Industry Workers
		8c3 - Renting Hard-Pressed Workers
	8d - Migration and Churn	8d1 - Young Hard-Pressed Families
		8d2 - Hard-Pressed Ethnic Mix
		8d3 - Hard-Pressed European Settlers

Table 4: 2011 OAC cluster names and hierarchy.

8 Discussion

This paper has outlined the process that underpinned the creation of the 2011 OAC classification. As with all geodemographic classifications, it is inevitable that subjective decisions are made based upon our own accumulated experience, and in this context "best" should be thought of in terms of creating a general purpose classification that fulfills utilitarian objectives—of use to the greatest number of people and the largest number of applications.

Creating geodemographic classifications can lean heavily on past methodologies or be created with new techniques utilizing methods and procedures never previously applied

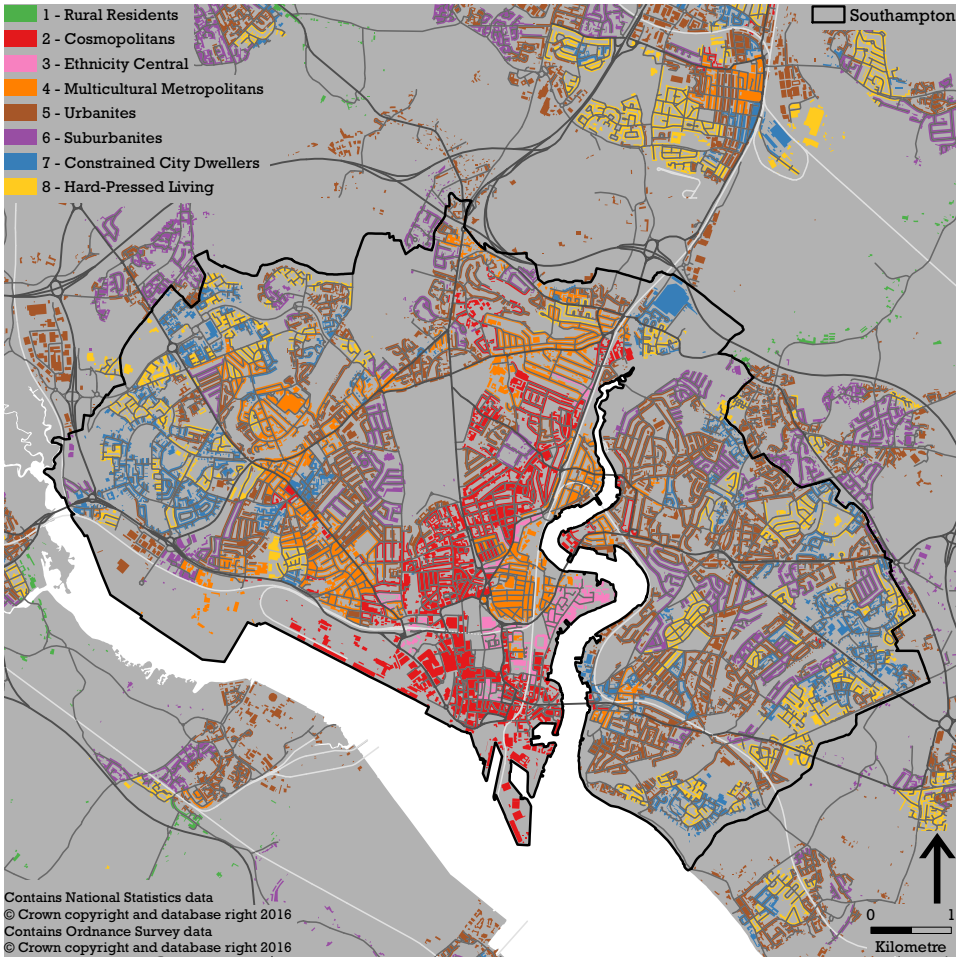

Figure 2: The 2011 OAC Supergroups in the city of Southampton and surrounding area.

in the field. Ultimately, however, whatever the techniques used, the aim is the same: to create a geodemographic classification that summarizes the varying characteristics of the study population and built environment. The methodology for the 2011 OAC is an evolution of that used for the 2001 OAC, and not a radical departure. This takes on board the requirements set out by ONS and views expressed by current and past users of the 2001 OAC in terms of the structure, outputs, and general characteristics of the classification. As such, improving the 2001 OAC methodology, rather than creating something entirely new has been the focus of this work.

A comparison of the two classifications is shown in Table 5. It is notable that the 2011 OAC has, on average, clusters that are less homogeneous across the UK than the 2001 OAC, suggesting greater variation in the characteristics of the residents within each 2011 OAC cluster when compared to their 2001 OAC counterparts. Cluster homogeneity was calculated using the distance measure used with the k-means clustering algorithm, SED. The SED is a dissimilarity measure; the larger the SED value for each OA or SA, the more dis-

similar it is to the cluster centroid (the average characteristics of that cluster's population). SED values for each OA and SA were calculated, thereby providing a proxy measure of cluster assignment uncertainty; a requested output from the 2011 OAC user engagement. The calculation of SED values does allow for comparisons between the cluster homogeneity of the 2011 OAC and the 2001 OAC. However any direct comparisons are likely to be misleading because of changes in the census questionnaire and broader secular changes in society over the inter-censal period. The increased number of input variables almost certainly accounts for some of the observed decrease in cluster homogeneity, yet the revised design also fulfills key user requirements shown in Table 1. London and other urban centers, for example, are better described by the widened variable specification, not least because the ethnic diversity of their populations are quite fully represented. In London for instance, the "Ethnicity Central" and "Multicultural Metropolitans" Supergroups contain distinct populations, whereas with the 2001 OAC they were grouped together into a single cluster.

	2001 OAC	*2011 OAC*
Number of input variables	41	60
Rate calculation method	Percentages	Percentages
Data transformation method	log 10	Inverse Hyperbolic Sine
Standardization method	Range	Range
Clustering algorithm	k-means	k-means
Cluster numbers	7 Supergroups 24 Groups 52 Subgroups	8 Supergroups 26 Groups 76 Subgroups
Range of cluster assignments (across the UK)	21.2% to 7.5% for Supergroups 8.2% to 2.3% for Groups 3.3% to 0.8% for Subgroups	20.2% to 5.1% for Supergroups 11.6% to 0.7% for Groups 4.2% to 0.1% for Subgroups
Mean cluster homogeneity (squared Euclidean distance)	0.82 for Supergroups 0.75 for Groups 0.70 for Subgroups	0.87 for Supergroups 0.81 for Groups 0.77 for Subgroups

Table 5: Comparison of the 2001 OAC and the 2011 OAC.

Ultimately the relative merit and robustness of the 2011 OAC can only be assessed by how well it differentiates the UK's population based on 2011 UK Census data. In addition to the statistics shown in Table 5, we have also evaluated the 2011 OAC based upon our own knowledge of development and other changes over the 2001 to 2011 inter-censal period. The validation of the 2011 OAC using Southampton in the previous section, while

limited, indicates the classification does appear to offer a realistic representation of population groups, at least in areas known to the authors.

The ability to examine in any detail the validity of the 2011 OAC at the small area level across the whole of the UK is beyond the scope of this methodological paper. However, the commitment of the 2011 OAC to be open and transparent and meet the needs of users means a range of tools are now available that allow anyone to integrate the classification. The website http://http://www.opengeodemographics.com has been created to provide access to all outputs from the 2011 OAC. It hosts a number of outputs as requested by participants of the 2011 OAC user engagement exercise, including pen portraits, full data downloads, and documentation. In addition, the website provides a search facility and structured method of leaving feedback, for which the benefits have been argued elsewhere [34]. As of October 2015, the classification has been searched 29,500 times by 12,000 unique users.

The availability of this data has led to third party websites creating their own resources. An example of this is the DataShine website, which offers an interactive web-based map of the 2011 OAC that allows the social geography of local neighborhoods to be explored. As such, it is regarded as a key output of the new classification by users. Stakeholders also expressed a desire for the new classification to be more widely promoted. To this end, the Consumer Data Research Centre (CDRC), a major UK Big Data research initiative funded by the Economic and Social Research Council, has included 2011 OAC data as part of their 27.6 gigabytes (as of October 2015) of consumer focused datasets. The platform the CDRC provides allows the 2011 OAC to be more widely promoted and accessed by users whose interests are broader than just geodemographics.

The aspirations of openness with the 2011 OAC far exceed that of its predecessor. However, the dissemination of data and materials only formed one part of this. The other was with the use of open source software such as R. To this end all the code used in the creation of the 2011 OAC has been put in the public domain using GitHub. It is hoped that this resource will allow others to build their own classifications, such as that already done so for London [48] and with workplace data [17]; along with those seeking to substitute open data for data from the 2011 UK Census and those seeking to create geodemographic classifications for niche applications, through creating, for example, a topic specific classification.

The ultimate test of the classification will not however be the extent to which user requirements were met or how open it is, but rather how well it represents the UK's population. The internal validation undertaken would indicate the 2011 OAC meets the expectations of the authors, however it will be for users to ultimately decide if the classification meets their needs. This is not to say improvements cannot be made to future classifications of this type. The choice of data in the future should expand as more sources are made available, with less reliance on the decennial census opening up the possibility of including a wider range of topics, such as income, and allowing for more frequent updates. The continued dearth of such data at the smallest area levels may however require creating a classification at less granular spatial scales to accomplish this, with initial research having been undertaken to understand how feasible this is [2]. Additionally, the focus of the 2011 OAC on how data is handled should not preclude future classification creators from exploring different cluster algorithms. Although k-means using SED was deemed the most appropriate method of fulfilling the requirements of ONS and users for this project, other algorithms or distance measures not traditionally used with geodemographics may be more suitable for future applications. An example of this is the potential of using Ma-

halanobis distances [36], rather than SED, with the k-means algorithm to better handle the complexities of census data.

The practices of geodemographic classification continue to evolve. The primary motivation for the 2001 OAC was to demonstrate the feasibility of creating a free open geodemographic classification at the highest available level of granularity. The 2011 OAC extends this methodology, while for the first time in open geodemographics in the UK having no restrictions to the data, methodology, or software used. It is hoped it will also establish a benchmark for monitoring change in subsequent years using open data sources (see [22]), although in-depth analysis of this topic lies beyond the scope of this paper.

The 2011 OAC follows in the same linage as previous free geodemographic systems in the UK. However, the methodological improvements made and the fulfilled desire to be as open and transparent as possible mean the new classification can be considered an advance in open geodemographics in the UK.

Acknowledgments

This research was sponsored by the Office for National Statistics under a UCL "Impact" PhD studentship, with further analysis undertaken under EPSRC grants EP/J004197/1 (Crime, policing and citizenship (CPC)—space-time interactions of dynamic networks) and EP/J005266/1 (The uncertainty of identity: linking spatiotemporal information between virtual and real worlds) and ESRC grants ES/K004719/1 (Using secondary data to measure, monitor and visualise spatio-temporal uncertainties in geodemographics) and ES/L011840/1 (Retail Business Datasafe).

References

[1] ADNAN, M., LONGLEY, P., SINGLETON, A., AND BRUNSON, C. Towards real-time geodemographics: Clustering algorithm performance for large multidimensional spatial databases. *Transactions in GIS 14*, 3 (2010), 283–297. doi:10.1111/j.1467-9671.2010.01197.x.

[2] AJEBON, M. O., AND NORMAN, P. Can administrative data be used to create a geodemographic classification? In *GISRUK 2015 Proceedings* (Leeds, UK, 2015), pp. 20–32.

[3] ARABIE, P., HUBERT, L. J., AND DE SOETE, G., Eds. *Clustering and Classification*. World Scientific, River Edge, USA, 1996.

[4] ASAI, Y., AND YANO, K. Construction of geodemographics based on the 1995 population census of Japan. *Papers and proceedings of the Geographic Information Systems Association 10* (2001), 279–284.

[5] ASHBY, D. I., AND LONGLEY, P. Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS 9*, 1 (2005), 53–72. doi:10.1111/j.1467-9671.2005.00205.x.

[6] BAILEY, S., CHARLTON, J., DOLLAMORE, G., AND FITZPATRICK, J. Which authorities are alike? *Population Trends 98* (1999), 29–41.

[7] BAILEY, S., CHARLTON, J., DOLLAMORE, G., AND FITZPATRICK, J. Families, groups and clusters of local and health authorities of Great Britain: Revised for authorities in 1999. *Population Trends 99* (2000), 37–52.

[8] BATEY, P., BROWN, P., AND PEMBERTON, S. Methods for the spatial targeting of urban policy in the UK: A comparative analysis. *Applied Spatial Analysis and Policy 1*, 2 (July 2008), 117–132. doi:10.1007/s12061-008-9007-3.

[9] BLAKE, M., AND OPENSHAW, S. GB profiles: A user guide. Tech. rep., School of Geography, University of Leeds, Leeds, UK, 1994.

[10] BLAKE, M., AND OPENSHAW, S. Selecting variables for small area classifications of 1991 UK census data. Tech. rep., School of Geography, University of Leeds, Leeds, UK, 1995.

[11] BOX, G. E. P., AND COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 26*, 2 (Jan. 1964), 211–252.

[12] BREETZKE, G. D., AND HORN, A. C. A geodemographic profiler for high offender propensity areas in the city of Tshwane, South Africa. *Environment and Planning A 41*, 1 (2009), 112 – 127. doi:10.1068/a40159.

[13] BRUNSDON, C., LONGLEY, P., SINGLETON, A., AND ASHBY, D. Predicting participation in higher education: A comparative evaluation of the performance of geodemographic classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174*, 1 (Jan. 2011), 17–30. doi:10.1111/j.1467-985X.2010.00641.x.

[14] BURBIDGE, J. B., MAGEE, L., AND ROBB, A. L. Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association 83*, 401 (Mar. 1988), 123–127. doi:10.2307/2288929.

[15] CACI. Acorn technical guide, 15th March 2013.

[16] CHARLTON, M., OPENSHAW, S., AND WYMER, C. Some new classifications of census enumeration districts in Britain: A poor man's ACORN. *Journal of Economic and Social Measurement 13*, 1 (Apr. 1985), 69–96.

[17] COCKINGS, S., MARTIN, D., AND HARFOOT, A. A classification of workplace zones for England and Wales (COWZ-EW). Tech. rep., University of Southampton, Southampton, UK, 2015.

[18] DAG, O., ASAR, O., AND ILK, O. A methodology to implement box-cox transformation when no covariate is available. *Communications in Statistics - Simulation and Computation 43*, 7 (Jan. 2014), 1740–1759. doi:10.1080/03610918.2012.744042.

[19] DEPARTMENT FOR COMMUNITIES AND LOCAL GOVERNMENT. The English indices of deprivation 2010. Tech. rep., Mar. 2011.

[20] DEPARTMENT FOR COMMUNITIES AND LOCAL GOVERNMENT. The English indices of deprivation 2015. Tech. rep., Sept. 2015.

[21] EVERITT, B. S., LANDAU, S., LEESE, M., AND STAHL, D. *Cluster Analysis*, 5th ed. Wiley Series in Probability and Statistics. Wiley, Chichester, UK, 2011.

[22] GALE, C., AND LONGLEY, P. Temporal uncertainty in a small area open geodemographic classification. *Transactions in GIS 17*, 4 (2013), 563–588. doi:10.1111/tgis.12035.

[23] GALE, C. G. *Creating an open geodemographic classification using the UK census of the population*. Doctoral thesis, University College London, Department of Geography, London, UK, Sept. 2014.

[24] GONZÁLEZ-BENITO, Ó., BUSTOS-REYES, C. A., AND MUÑOZ-GALLEGO, P. A. Isolating the geodemographic characterisation of retail format choice from the effects of spatial convenience. *Marketing Letters 18*, 1-2 (June 2007), 45–59. doi:10.1007/s11002-006-9000-z.

[25] GORDON, A. *Classification*, 2nd ed. No. 82 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC Press, London, UK, 1999.

[26] HARRIS, R., SLEIGHT, P., AND WEBBER, R. *Geodemographics, GIS and Neighbourhood Targeting*. Wiley, London, UK, 2005.

[27] JAIN, A., MURTY, M. N., AND FLYNN, P. J. Data clustering: A review. *ACM Computing Survey 31*, 3 (Sept. 1999), 264–323. doi:10.1145/331499.331504.

[28] JAIN, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters 31*, 8 (June 2010), 651–666. doi:10.1016/j.patrec.2009.09.011.

[29] JOHNSON, N. L. Systems of frequency curves generated by methods of translation. *Biometrika 36*, 1-2 (June 1949), 149–176. doi:10.2307/2332539.

[30] KOVÁCS, F., LEGÁNY, C., AND BABOS, A. Cluster validity measurement techniques. In *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence* (Nov. 2005), pp. 18–19.

[31] KUMAR, N. S., RAO, K. N., GOVARDHAN, A., REDDY, K. S., AND MAHMOOD, A. M. Undersampled K-means approach for handling imbalanced distributed data. *Progress in Artificial Intelligence 3*, 1 (Aug. 2014), 29–38. doi:10.1007/s13748-014-0045-6.

[32] LEYDESDORFF, L., AND BENSMAN, S. Classification and powerlaws: The logarithmic transformation. *Journal of the American Society for Information Science and Technology 57*, 11 (Sept. 2006), 1470–1486. doi:10.1002/asi.20467.

[33] LONGLEY, P. A. Geographical information systems: A renaissance of geodemographics for public service delivery. *Progress in Human Geography 29*, 1 (Feb. 2005), 57–63. doi:10.1191/0309132505ph528pr.

[34] LONGLEY, P. A., AND SINGLETON, A. Classification through consultation: Public views of the geography of the e-society. *International Journal of Geographical Information Science 23*, 6 (2009), 737–763.

[35] MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (1967), University of California Press, pp. 281–297.

[36] MELNYKOV, I., AND MELNYKOV, V. On K-means algorithm with the use of Mahalanobis distances. *Statistics & Probability Letters 84* (Jan. 2014), 88–95. doi:10.1016/j.spl.2013.09.026.

[37] OFFICE FOR NATIONAL STATISTICS, ONS. Statistical disclosure control for 2011 census, 2012.

[38] OFFICE FOR NATIONAL STATISTICS, ONS. User engagement on a new United Kingdom output area classification - summary of responses, May 2012.

[39] OFFICE FOR NATIONAL STATISTICS, ONS. Methodology note for the 2011 area classification for output areas, 2014.

[40] OFFICE FOR NATIONAL STATISTICS, ONS. Methodology note for the 2011 area classification for local authorities, July 2015.

[41] OFFICE FOR NATIONAL STATISTICS, ONS. Population ageing in the United Kingdom, its constituent countries and the European Union, 2nd March 2012.

[42] OJO, A. A., VICKERS, D., AND BALLAS, D. The segmentation of local government areas: Creating a new geography of nigeria. *Applied Spatial Analysis and Policy 5*, 1 (Mar. 2012), 25–49. doi:10.1007/s12061-010-9058-0.

[43] OSBOURNE, J. Notes on the use of data transformation. *Practical Assessment, Research & Evaluation 8*, 6 (2002).

[44] PENCE, K. The role of wealth transformations: An application to estimating the effect of tax incentives on saving. *Contributions to Economic Analysis & Policy 5*, 1 (2006), 1430–1430. doi:10.2202/1538-0645.1430.

[45] PETERSEN, J., GIBIN, M., LONGLEY, P., MATEOS, P., ATKINSON, P., AND ASHBY, D. Geodemographics as a tool for targeting neighbourhoods in public health campaigns. *Journal of Geographical Systems 13*, 2 (June 2011), 173–192. doi:10.1007/s10109-010-0113-9.

[46] R DEVELOPMENT CORE TEAM. *R: A Language and Environment for Statistical Computing*. The R Foundation for Statistical Computing, Vienna, Austria, 2011.

[47] ROBINSON, W. S. Ecological correlations and the behavior of individuals. *American Sociological Review 15*, 3 (June 1950), 351–357. doi:10.1093/ije/dyn357.

[48] SINGLETON, A., AND LONGLEY, P. The internal structure of Greater London: a comparison of national and regional geodemographic models. *Geo: Geography and Environment 2*, 1 (June 2015), 69–87. doi:10.1002/geo2.7.

[49] SINGLETON, A., AND LONGLEY, P. A. Creating open source geodemographics - refining a national classification of census output areas for applications in higher education. *Papers in Regional Science 88*, 3 (2009), 643–666. doi:10.1111/j.1435-5957.2008.00197.x.

[50] SINGLETON, A., AND SPIELMAN, S. The past, present and future of geodemographic research in the United States and United Kingdom. *Professional Geographer 66*, 4 (2014), 558–567. doi:10.1080/00330124.2013.848764.

[51] SINGLETON, A., AND SPIELMAN, S. An open geodemographic classification of the United States. *Annals of the Association of American Geographers* (In Press).

[52] SINGLETON, A. D. The geodemographics of educational progression and their implications for widening participation in higher education. *Environment and Planning A 42*, 11 (2010), 2560 – 2580. doi:10.1068/a42394.

[53] SINGLETON, A. D., WILSON, A. G., AND O'BRIEN, O. Geodemographics and spatial interaction: An integrated model for higher education. *Journal of Geographical Systems 14*, 2 (Apr. 2012), 223–241. doi:10.1007/s10109-010-0141-5.

[54] VAN LAERHOVEN, K. Combining the self-organizing map and K-means clustering for on-line classification of sensor data. In *Artificial Neural Networks—ICANN* (2001), G. Dorffner, H. Bischof, and K. Hornik, Eds., vol. 2130 of *Lecture Notes in Computer Science*, pp. 464–469.

[55] VICKERS, D. *Multi-level Integrated Classifications Based on the 2001 Census*. Doctoral thesis, University of Leeds, Department of Geography, Leeds, UK, 2006.

[56] VICKERS, D., AND REES, P. Creating the UK national statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 170*, 2 (2007), 379–403. doi:10.1111/j.1467-985X.2007.00466.x.

[57] VICKERS, D., AND REES, P. Ground-truthing geodemographics. *Applied Spatial Analysis and Policy 4*, 1 (Feb. 2011), 3–21. doi:10.1007/s12061-009-9037-5.

[58] VICKERS, D., REES, P., AND BIRKIN, M. Creating the national classification of census output areas: Data, methods and results. Tech. rep., School of Geography, University of Leeds, Leeds, UK, 2005.

[59] VOLKOV, M., HARKER, D., AND HARKER, M. Who's complaining? Using MOSAIC to identify the profile of complainants. *Marketing Intelligence & Planning 23*, 3 (May 2005), 296–312. doi:10.1108/02634500510597328.

[60] WALLACE, M., AND DENHAM, C. *The ONS Classification of Local and Health Authorities of Great Britain*. No. 59 in Studies on Medical and Population Subjects. HMSO, London, UK, 1996.

[61] WARD, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*, 301 (Mar. 1963), 236–244. doi:10.1080/01621459.1963.10500845.

[62] WEBBER, R. J. An introduction to the national classification of wards and parishes. Tech. Rep. 23, Centre for Environmental Studies, London, UK, 1977.

[63] WEBBER, R. J., AND CRAIG, J. Which local authorities are alike? *Population Trends 5* (1976), 13–19.

[64] WEBBER, R. J., AND CRAIG, J. *Socio-Economic Classification of Local Authority Areas*. No. 35 in Studies in Medical and Population Subjects. Office of Population Censuses and Surveys, 1978.

[65] WILLIS, I., GIBIN, M., BARROS, J., AND WEBBER, R. J. Applying neighbourhood classification systems to natural hazards: a case study of Mt Vesuvius. *Natural Hazards 70*, 1 (2014), 1–22.

[66] XU, R., AND WUNSCH, D. C. *Clustering*. IEEE series on computational intelligence. Wiley, Oxford, UK, 2009.