A Classification of Multidimensional Open Data for Urban Morphology

ALEXANDROS ALEXIOU, ALEX SINGLETON and PAUL A. LONGLEY

Identifying socio-spatial patterns through geodemographic classification has proven utility over a range of disciplines. While most of these spatial classification systems include a plethora of socioeconomic attributes, there is arguably little to no input regarding attributes of the built environment or physical space, and their relationship to socioeconomic profiles within this context has not been evaluated in any systematic way. This research explores the generation of neighbourhood characteristics and other attributes using a geographic data science approach, taking advantage of the increasing availability of such spatial data from open data sources. We adopt a SOM (Self-Organizing Maps) methodology to create a classification of Multidimensional Open Data Urban Morphology (MODUM) and test the extent to which this output systematically follows conventional socioeconomic profiles. Such an analysis can also provide a simplified structure of the physical properties of geographic space that can be further used as input to more complex socioeconomic models.

Geodemographics is a field of quantitative geography that engages in the analysis and classification of populations into discrete classes based on socioeconomic and built environment characteristics of small-area geography. Simply put, geodemographics is the 'analysis of people by where they live' (Sleight, 1997, p. 16). Such classifications have demonstrated utility over a range of public and private sector applications (Longley, 2005; Longley and Goodchild, 2008; Reibel, 2011; Singleton and Spielman, 2013). A geodemographic analysis is essentially a data reduction methodology that aggregates populations, so that correlations between sub-populations can be drawn on with ease. It involves the process of producing key statistics of a particular area, on the basis of the characteristics of its residents and their contexts.

Geodemographic applications were initially developed as a strategy to analyse and systematically document socio-spatial segregation. The associated data reduction methods were established in the 1970s (Webber, 1978), although a wider review and interpretation would extend right back to the 'human ecology' studies from the Chicago School of Sociology in the 1920s (Burgess, 1925), social area analysis in the 1950s (Shevky and Bell, 1955) and the factorial ecologies of the 1970s (Janson, 1980). Although that geodemographics has evolved considerably over the years (Singleton and Spielman, 2013), its conceptual background is still wedded to the principle that people tend to align themselves with the behaviour and aspirations of the local communities in which they live. The inferential nature of the aggregations rely on the notion of societal homophily, or in other words, that 'birds of a feather flock together' (Harris *et al*., 2005). As such, people who live close by (e.g. in the same neighbourhood) are more likely to have commonalities in attributes and behaviours than a randomly selected group of people.

Although geodemographic frameworks can

capture a wide set of input attributes, current classification systems typically include little to no input of explicitly spatial attributes regarding the built and physical attributes of neighbourhoods. There is, however, an abundance of variables that might be collected on the built forms and relative locations that underpin neighbourhood differentiation. For instance, proximity to certain amenities is important to residential decisions such as transport nodes, parks, retail and healthcare-facilities. There has, for example, been extensive research into the topic of analysing relationships between accessibility and urban development patterns, (e.g. land use-transportation interaction (LUTI) models); and connectivity has been advanced as a key feature in shaping urban residential dynamics and socio-spatial segregation (Dear, 2002). Research on residential decisions has also attracted a lot of attention over the years, particularly through hedonic modelling. While most of the relevant research focuses on the importance of work location (Van Ommeren *et al*., 1999; Renkow and Hoover, 2000), there is strong evidence that certain demographic groups favour some relative locations over others, and that the nature and configuration of the local built environment and land-use characteristics are also relevant (Hui *et al*., 2007). For instance, individuals with children often favour green space and recreational opportunities nearby, while those without children prefer smaller residences that offer closer proximity to central services (Colwell *et al*., 2002). Other characteristics may impact the area as unfavourable due to negative externalities, such as high-speed roads or railway tracks within the vicinity of the neighbourhood (Parkes *et al*., 2002). It is unclear exactly how such characteristics impact upon residential decisions as there are many synergies involved across lifecycles (Kim *et al*., 2005). For instance, moderate proximity (200 m to 300 m) to a green space may mitigate negative effects of noise pollution (Gidlof-Gunnarsson and Ohrstrom, 2007).

Some census variables reflect limited built environment characteristics, for instance housing type and population densities. For classification systems that have been developed entirely from census variables, such as the publicly open ONS (Office of National Statistics) Output Area Classification (OAC) for 2011, attributes such as density can, however, be misleading; the arbitrary nature of the geographic extents of the administrative areas for which population measurements are offered renders comparisons between the physical features ineffective. Other proprietary geodemographic classifications, such as Mosaic by Experian (Nottingham, UK) and Acorn by CACI (London, UK) include some measures of relative location (CACI, 2013; Experian, 2014). However, to what precisely these attributes pertain, how they are used in the clustering process and the weight they are assigned in the final classification remains obscure, because of the commercial sensitivities that are inherent in 'black box' commercial solutions (Singleton and Longley, 2009).

In this paper, we test whether specific and multidimensional urban morphologies systematically correspond with socioeconomic characteristics at the neighbourhood level. In order to identify and analyse such attribute patterns, we adopt a geodemographic approach, which involves the creation of a classification for a national extent, based on clustering at the small area level. In essence, we try to identify the physical and built environment characteristics that might be used to supplement neighbourhood typologies.

Open Data Inputs

This research captures a variety of physical attributes collected for a small-area geography, and in order to enhance reproducibility, replication and extension these inputs are assembled from Open Data sources (Singleton *et al*., 2016). We produce a classification at the 2011 UK Census Output Area level for the 181,408 Output Areas (OAs) that make up England and Wales. One of the main providers of geographical data for England and Wales is the

national mapping agency Ordnance Survey (OS), and there are many datasets available within their repository, with varying degrees of granularity, depending on whether they are publicly accessible or available for purchase. As this paper focuses on Open Data sources, we use OS Open Map – Local, the most recent and detailed open OS vector data product currently available (Ordnance Survey, 2015). However, within different contexts, such data might also be supplemented by other national mapping agency data, or alternative sources such as OpenStreetMap (www.openstreetmap.org). The OS vector data product provides a variety of information including outlines of buildings, street network with hierarchy, railways, woodland areas, surface water and important functional sites.

While the OS Open Map – Local provides the main source of this data, there were a few other sources within England and Wales deemed of utility. These included data about listed buildings and historic parks and gardens supplied by the *Historic England Archive* (https://services.historicengland.org.uk/NMRDataDownload/) which is regularly updated (November 2015 update used here) and also under Open Data License. For Wales, the corresponding provider is the Cadw heritage organization (available through the UK data Service, https://data.gov.uk/dataset/listed-buildings-in-wales-gis-point-dataset), although the data are slightly outdated (September 2011). Commercial buildings for local retail centres were identified using data from the Local Data Company, an Open version of which is available through the ESRC Consumer Data Retail Centre. Finally, we included aggregated data on housing type from the 2011 Census supplied by the Office for National Statistics (ONS). Unfortunately, there are currently no Open Data available on building age or height.

Table 1 summarizes the range of inputs

Table 1. Description of the spatial dataset compiled for England and Wales.

Variable Name	*Variable Description*
D1: OA Boundaries	181,408 Output Area boundaries, as defined by the 2011 Census. All other data were spatially joined with the respective OAs that they fall into (data features were split when falling into more than one OA).
D1: Buildings	12,878,666 Building objects represented as polygons. Note that these areas do not represent individual households.
D2: Road Network	Road network is represented as line segments, approximate to the road centre. The categories include 'Motorway', 'Primary Road', 'A Road', 'B Road', 'Minor Road', 'Pedestrianized Street', 'Local Street' and 'Private Road Publicly Accessible', as well as their 'Collapsed Dual Carriageway' counterparts.
D3: Woodland	Areas of trees represented as polygons, described as coniferous and non-coniferous.
D4: Functional Sites/ Important Buildings	120,677 Building polygons that can be found within functional sites. They are categorized into themes such as Air Transport, Education, Medical Care, Road Transport and Water Transport, which are further classified into numerous more discrete classes.
D5: Railway Stations and Tracks	Railway tracks and tunnels represented as lines (in this instance we used tracks only in the analysis) and Railway Stations defined as points.
D6: Surface water	Polygons of surface water. Small rivers and streams are represented as lines and were not included in the dataset. The dataset was also supplemented with 'seawater', derived from the country's coastline.
D7: Registered Historic Buildings	406,496 listed historic buildings defined as points, which were geolocated.
D8: Registered Parks and Gardens	2,007 Polygon features with extents of the parks / gardens, classified as I, II*, or II, from most to least important. For Wales, the 372 sites were identified from points from a 'Named Places' dataset and given an approximate 200 m radius.
D9: Retail Centres	1,312 Retail Centres across England and Wales. There is no recent update for this dataset which dates back to 2004. The centres are only depicted as points and have no typology attached. We assumed an average radius of 200 m to convert them to areas.
D10: Housing Type	Percentage of households that are classified by the Census as Detached, Semi-detached, Terraced or Flat.
D11: Population	Population of total persons per OA.

used to derive measures featured in this analysis.

The classification presented later was created for Output Areas (LSOAs), and as such the input measures were assembled for this geography. These zones offer advantage over other administrative units in England and Wales since many other socioeconomic classifications are offered at the OA level, such as the 2011 ONS Output Area Classification, thus making comparisons possible. Additionally, such geography also allows the incorporation of Census data which is distributed for these units. However, for the range of the derived measures that are described in the remainder of this section, there are problems with this approach. OA borders were designed to maximize within zone homogeneity in population characteristics (population normalization), without regard to the geographical features of the area (Martin *et al*., 2001; see figure 1). As such, for proximity based inputs there were challenges about how such measures might be calculated, and to which area they should be attributed.

A similar attempt to create such a dataset was made by the Department for Communities and Local Government in 2005, within the framework of the ONS Neighbourhood Statistics, described as Land Use Statistics. The dataset was described as a generalized land-use database aggregated into OAs. The dataset contained estimates of built environment attributes, such as roads, paths, domestic and non-domestic buildings, domestic gardens, water, rail etc. Despite the fact that the proprietary OS Enhanced Basemap was used to create this resource, ONS classified it as experimental, as there were issues of accuracy, mainly arising because only the centroids of features were taken into account in class assignments of aggregations.

To facilitate these methodological shortcomings, we adopted three different types of attribute measures for each OA that related to either two types of proximity measures including *adjacency effects* or *intermediate effects*; and additionally *direct measures.* The last of these are simply attributes captured at the OA level, while the first two assume buildings as the initial unit of analysis which are then later assigned to OAs. Building polygon

Figure 1. Maps looking at the un-generalized Output Area borders (black lines) around Sefton Park, Liverpool. *Left*: Notice how the area of the park is divided arbitrarily between proximal OAs (crosshatched pattern). *Right*: Output Area borders usually coincide with the street network, making simple street network-to-area assignments impracticable.

features serve as observations in this input dataset, and represent homogenous built-up areas which can include one or more households. A graphical representation of the model is described in figure 2. All the attributes collated as input across all domains are summarized in table 2.

For both types of proximity measure, we used a series of spatial queries that identified buildings that fulfil certain criteria, for instance, which buildings are within a set distance of a major street? The buildings that met each criterion were then assigned to OA aggregations with weights determined by their attributed area. Thus, within each OA, a ratio of the area of buildings meeting the criteria relative to the total built areas was calculated for each of the attributes considered in the analysis. The necessity to differentiate between adjacency and intermediate proximity effects follows the logic that not all built environment characteristics have the same effect, and these effects may vary in scale. For example, when considering the location of a residential property, being adjacent to a very major road might be perceived as having a negative impact, given the noise/pollution associated with increased traffic volumes, whereas being near, but not adjacent to a busy road might be perceived as advantageous, given the enhanced connectivity this might facilitate.

We defined *adjacency effects* to features measured within 100 m linear distance, as commonly used in the literature on negative externality effects of built environment features, such as noise or pollution from roads (Rijnders *et al*., 2001). For *intermediate effects* a distance of 600 m was used, on the basis of various Western international definitions of 'within walking distance'. The distance figure generally varies depending on the context of analysis, but distances between 300 m and 900 m are considered appropriate for urban

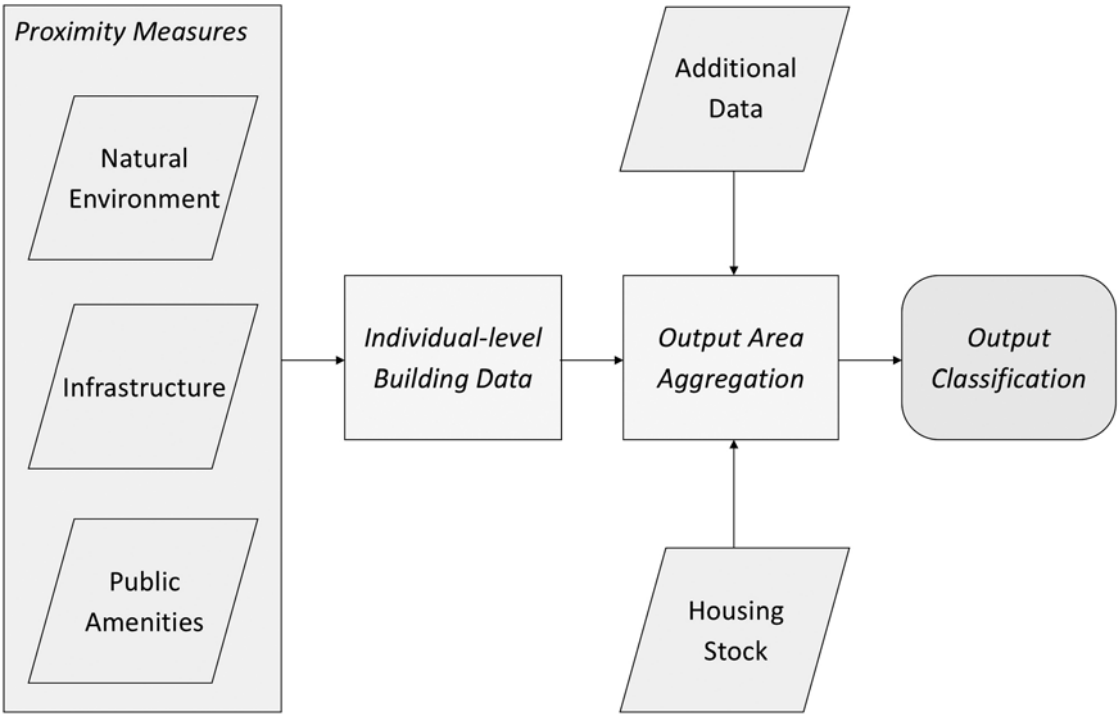

Figure 2. The spatial data model used to process data and produce Output Area inputs to the classification.

features (Hui *et al*., 2007; Barbosa *et al*., 2007; Villeneuve *et al*., 2012; Vale, 2015).

Beyond these distances we assume there are no adjacency or intermediate effects. The delineation of *adjacency effects* or *intermediate effects* brings additional practical considerations which relate to the overall density of the built environment features being considered. In common with practice when creating inputs to multidimensional classifications, preference should be for those attributes which, in addition to theoretical rationale, also provide useful differentiation between areas (Spielman and Singleton, 2015). For example, in this application, when 600 m buffers were used for major roads, this resulted in more than 50 per cent of buildings meeting this criterion, thus providing a weak differentiation. These tasks were computationally expensive, as the complete dataset contains more than 12.8 million observations (building polygons). Thus the database was pre-processed into regional datasets which were then computed separately using the R programming language.

Finally, there were two further types of *direct* measures: those which were derived from geographic features, and those which were simple inputs from secondary data. The derived *direct* measures included listed buildings and culs-de-sac (dangling segments in the road network). The latter of these was defined geocomputationally as the end of a line segment that did not intersect with any other such segment. A sensitivity of 10 m was applied to this criterion in order to avoid topological errors and intermittent street segments. The results show that such measures can capture specific urban morphologies even at the small-area level as we show in figure 3.

For the other non-derived *direct* measures, the variables were simply aggregated directly at the OA level, such as the housing type. Population density was calculated using a

Figure 3. *Left*: Attribute of cul-de-sac ratio per OA at Kingston-upon-Hull, Yorkshire. *Right*: The ratio of listed (registered) buildings per OA area in Liverpool.

ratio of persons per total building area, which potentially would give more accurate results regarding housing conditions. The final OA attributes along with their descriptions are provided in table 2.

A Multidimensional Classification of the Built Environment

Methodologically, our cluster analysis follows a conventional approach as detailed in Harris *et al*. (2005); however, here we use only built environment data to create the typology. A common clustering technique used in geodemographic analyses is the iterative allocation – reallocation algorithm, known as k-means. Although this algorithm has been used in a variety of geodemographic applications, our dataset is sparsely populated, and k-means is known not to respond well to the non-Gaussian distributions that characterize such datasets (Everitt *et al*., 2011).

Table 2. Built environment attributes used in the classification.

Variables	*Variable Description, Aggregated per OA Code*
Adjacent effects	
1. Major Roads	Percentage of the area of buildings that the centroid is within 100 m of a major road to the total building area. We defined major as those of type 'Motorway', 'A Road' and 'Primary Road'.
2. Arterial Roads	Percentage of the area of buildings that their centroid is within 100 m of an arterial road to the total building area. We defined Arterial roads as those with type 'B Road'.
3. Pedestrian Roads	Percentage of the area of buildings that their centroid is within 100 m of a pedestrian road or footway to the total building area.
4. Railway Tracks	Percentage of the area of building units that their centroid is within 100 m of railway tracks, excluding tunnels, to the total building area.
5. Woodland Areas	Percentage of the area of building units that their centroid is within 100 m of woodland features to the total building area.
6. Surface Water	Percentage of the area of building units that their centroid is within 100 m of surface water (inland) and seafront (calculated by the distance from the coastal line), but excluding small rivers and streams, to the total building area.
Intermediate effects	
7. Railway Stations	Percentage of the area of building units that their centroid is within 600 m from the centroid of a railway station to the total building area.
8. Parks & Gardens	Percentage of the area of building units that their centroid is within 600 m from the registered site extents to the total building area.
9. Retail Centres	Percentage of the area of building units that their centroid is within 600 m from the retail centre centroid plus 200 m to the total building area.
10. Schools	Percentage of the area of building units that their centroid is within 600 m from the sites that are identified as primary through secondary education to the total building area.
11. Higher Education	Percentage of the area of building units that their centroid is within 600 m from the sites that are identified as further and higher education to the total building area.
Direct measures	
12. Detached Ratio	Percentage of unshared households that are classified by the 2011 Census as detached housing to the total building area.
13. Semi-Detached Ratio	Percentage of unshared households that are classified by the 2011 Census as semi-detached housing to the total building area.
14. Terraced Ratio	Percentage of unshared households that are classified by the 2011 Census as terraced housing to the total building area.
15. Flat Ratio	Percentage of unshared households that are classified by the 2011 Census as Flats to the total building area.
16. Density	Ratio of persons to total building area (people/he).
17. Cul-de-sac	Ratio of culs-de-sac or dead-end road points to the total OA area (points/he).
18. Registered Buildings	Ratio of listed buildings to the total OA area (points/he)

As such, in this framework we adopt the alternative technique of a Self-Organizing Map (SOM). A SOM is an unsupervised classifier that uses artificial neural networks to classify multidimensional observations in two-dimensional space based on their similarities (Kohonen, 2001). A SOM typically organizes observations by projecting them onto a plane, and through consecutive iterations finds the best configuration of observations so that every observation is most similar to the others closest to them. Typically, the SOM mapping process employs a lattice of squares or hexagons as the output layer, and the results are therefore easily mapped as they retain their topology. SOMs have many applications in a broad range of fields, from medicine and biology to image analysis and computer science. SOMs have also been tested as an alternative classifier of census data (Spielman and Thill, 2008; Arribas-Bel and Schmidt, 2013) where they seem to perform well for socioeconomic data at the US Census tract scale. Arribas-Bel *et al*. (2011) have also demonstrated the algorithm capabilities to measure urban sprawl in Europe using a similar attribute set, specifically six variables: connectivity; decentralization; density; scattering; availability of open space; and land-use mix. The technique also has the advantage of not assuming any hypotheses regarding the nature or distribution of the data, and responds well to geographic sensitivity. A further advantage of using a SOM is the capacity to visualize the structure of data values aiding initial data exploration. This feature can be very useful when analyzing datasets such as our built environment measures, where there are little to no *a-priori* hypotheses on their underlying distribution.

As input to this analysis the dataset comprising the eighteen variables described in table 2 was transformed into z-scores in order to standardize the measures. The majority of the analysis and output production was performed in the R programming language using the 'Kohonen' library (Wehrens and Buydens, 2007). More specifically, we adopted a SOM approach to cluster our input dataset using the methodology described by Spielman and Folch (2015). A relatively unexplored built environment classification with too many clusters would be difficult to interpret, so we selected a 4-by-2 hexagonal grid, which produces eight distinct clusters. We implemented a hexagonal geodesic grid to project results. A geodesic plane forces the cells' relations to 'loop' around the edges, while the hexagonal representation is typically favoured over grids, as this configuration benefits from every cell having six immediate neighbours. The other main parameters of the SOM algorithm are the learning rate alpha, which we defined to progress linearly from 0.05 to 0.01 over fifty reconfigurations (updates), and the initial size of the neighbourhood, in this instance a distance chosen in such a way that two-thirds of all distances of the map units fall within the topological extents. The neighbourhood decreases linearly during training until the algorithm reaches equilibrium. The algorithm has achieved equilibrium at ~25 iterations, meaning that no more changes to the observations' configuration were required, with the mean distance to the closest unit in the map at 11.34. Once areas were assigned to clusters, we then implemented a radar plot to map their characteristics on the basis of the input variables as we show in figure 4. This enables classes to be labelled and the following short descriptions to be created:

High Street and Promenades. These clearly depicted areas represent the main retail centres of urban regions located along the main commercial streets. This cluster also includes areas with significant pedestrianized street networks, especially along seafronts, where a lot of recreational and leisure venues can be found.

Central Business District. The area often called city centre. Typically high-rise buildings with a lot of commercial and office spaces, hence the relatively low net population density. These areas have proximity to the majority

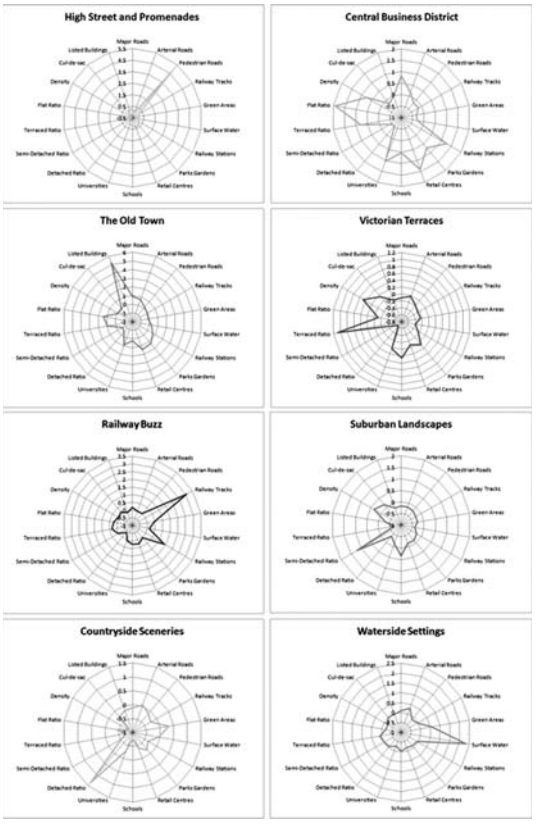

Figure 4. Final cluster results produced by the SOM, with mean attribute centres per cluster.

of public amenities, and have plenty of access via major roads and railways. For moderate-size cities the title holds true, but in areas such as London they tend to be too expansive to be labelled as central (figure 6).

The Old Town. The traditional town centre, usually close by the main high street. It is strongly defined by the amount of registered buildings. Typically a lot of recreational facilities can be found there, like pubs and restaurants, along with many administrative buildings and some historical major roads. Although it does have a considerable amount of flats, densities remain low, potentially due to refurbishments and change of usage.

Railway Buzz. These areas are dominated by railway tracks and railway stations. They have no other major distinguishing attributes which may suggest that they are actually rather heterogeneous in physical structure.

Suburban Landscapes. These areas are typically of semi-detached houses, with good access to parks. They tend to be quite distant from town centres. They are primarily residential areas, and close to schools. Culs-de-sac are relatively common, probably because of organized developments and gated communities.

Countryside Sceneries. These areas are dotted with detached houses, and are located either near or within open countryside. Most rural villages fall into this category, along with some city fringe developments that lie beyond the classic suburbs.

Waterside Settings. The principal defining attribute of these neighbourhoods is their proximity to surface water such as rivers, canals or sea. Some of these areas are ports, industrial or post-industrial sites. Distinctive infrastructure is arterial roads, i.e. roads wide enough to be used by lorries for the distribution of goods.

A Comparison of MODUM and OAC

In order to test whether the Multidimensional Open Data Urban Morphology (MODUM) classification systematically follows the conventional OAC geodemographic classification, we correlate the two sets of output classes via a contingency table. Table 3 shows the frequency distribution of MODUM within OAC 2011. *Supergroup 6. Rural residents* seems to be identified fairly well by the morphological features, with a correlation of more than 82 per cent, followed by a small percentage of *Waterside Settings* and *Suburban Landscapes*. About half the areas categorized as suburban also fall into this category, which is to be expected taking into account that typologies tend to blend out at the urban

Table 3. Contingency tables showing frequencies of OAC 2011 classes within MODUM.

	Output Area Classification 2011 – Supergroup Level								
MODUM Cluster Description	1. Rural residents	2. Cosmo-politans	3. Ethnicity central	4. Multi-cultural metro-politans	5. Urbanites	6. Suburban-ites	7. Constrained city dwellers	8. Hard-pressed living	OA Amounts
	%	%	%	%	%	%	%	%	
1. Suburban Landscapes	5.53	2.83	3.38	24.82	23.77	38.97	22.12	43.33	46,788
2. Railway Buzz	0.99	10.61	13.50	10.09	8.31	3.08	7.31	5.33	12,186
3. The Old Town	0.25	17.87	5.35	0.58	4.05	0.05	4.76	0.30	2,812
4. Victorian Terraces	1.20	14.43	16.56	43.93	24.59	1.79	39.38	34.98	49,860
5. Waterside Settings	8.43	5.03	3.56	6.98	12.08	6.73	8.04	8.82	12,468
6. Countryside Sceneries	82.45	2.05	0.43	2.91	18.89	47.79	2.14	3.90	3,172
7. High Street and Promenades	1.07	6.20	4.28	3.00	4.03	1.50	4.98	2.47	1,299
8. Central Business District	0.08	40.99	52.94	7.68	4.26	0.09	11.27	0.88	52,823
Sum (%)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	181,408

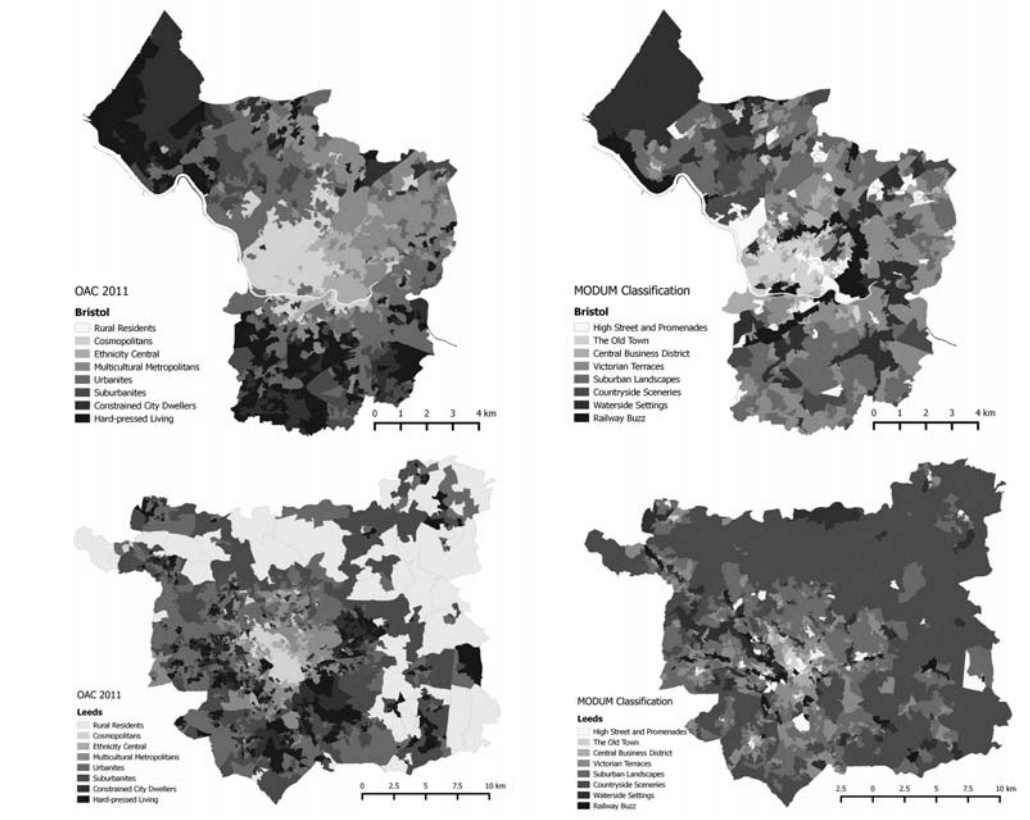

Figure 5. Built environment and socio-spatial patterns for the cities of Bristol (*top*) and Leeds (*below*). The two classifications, MODUM and OAC 2011, share many common locations, especially towards the city centre. In general, axial zones exhibit much more strongly in the morphological classification, while OAC seems to have a more 'regionalized' patterning, at least within local extents.

edges. The expansive central areas seem to be mainly populated by *Supergroup 2. Cosmopolitans* and *Supergroup 3. Ethnicity Central*. Moving out of the centre, Victorian Terraces seem to be scattered across three classes, *Supergroup 4. Multicultural Metropolitans*, *Supergroup 7. Constrained City Dwellers* and *Supergroup 8. Hard-Pressed Living*. The suburban class is most interesting, as 43 per cent of the areas classified as suburban is populated by areas identified as hard-pressed living. Generally speaking, unique classes in the MODUM classification such as the old city centre and railway-heavy areas seem to be equally dispersed among classes. Some further analysis could provide better insight as to why, and even reveal interesting patterns. Figure 5 provides two different sets of maps of the area of Bristol and Leeds, in order to demon-strate the overall pattern relationships between MODUM and OAC.

A chi-square test of the two categorical values shows that the two classifications have a significant relationship between them. We can measure the strength of the association by calculating the Cramer's V value $\varphi_c = 0.328$, which indicates an important level of association, given that φ_c can take values between 0 (no association) and 1 (complete association).

Discussion and Further Research

The development of MODUM illustrates that the production and analysis of a classification of the built environment using Big and Open Data can offer unique insights into some aspects of geodemographic structure of urban areas. The results capture, through the multi-

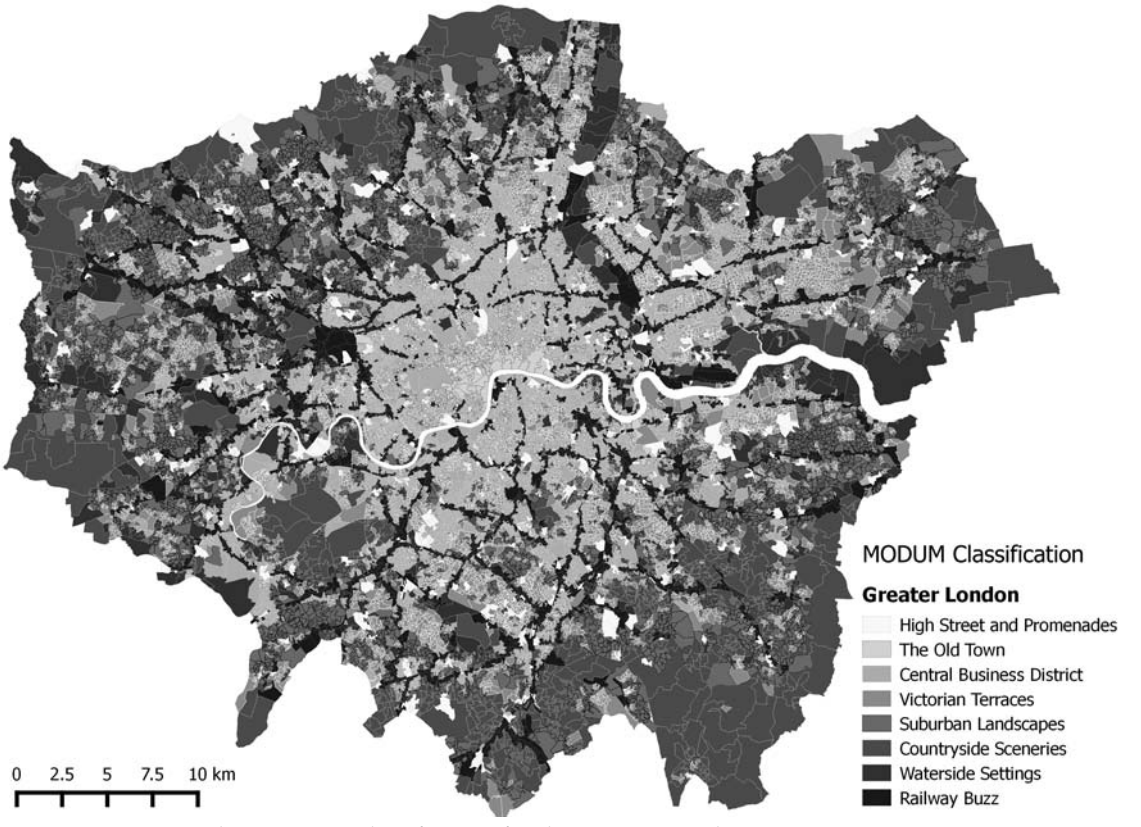

Figure 6. Mapping the MODUM classification for the Greater London Area.

dimensionality of the data, both microscopic and macroscopic identifiers of urban morphology. The classification can be used as input to more complex socioeconomic models, increasing robustness. There is strong evidence that residential preference is in significant part related to form of the built environment, suggesting that there is an important dimension to residential decisions beyond homophily. This raises some logical discrepancies in current socioeconomic geodemographic classifications; the conceptual 'control by aggregation' does not account for these unobserved variables. For instance, one would expect house prices to drop significantly very close to railway tracks. However, these localized phenomena are aggregated in the general context of the area, and thus patterns get 'smoothed away', raising some issues about the success of geo-classifications (Voas and Williamson, 2001). While gathering this type of behavioural data would be next to impossible, their outcomes can be observed through peoples' residential decisions on local morphology.

Furthermore, the MODUM classification can not only enhance socioeconomic classifications, and take into account microscopic variation, but also prove useful in itself; it can provide a simplified structure of the physical properties of geographic space that can be used to explore correlations with other spatial phenomena, potentially in a variety of applications, from real estate and house prices to health and wellbeing. In a dynamic sense, it can be used by urban planners and investors in the built environment to identify the areas in which the physical preconditions exist for neighbourhood renewal or upscaling.

On the other hand, the classification process described here is very specific to the underlying data and methodology. An inherent disadvantage of all geodemographic classifications is that lack of a single global optimization function during the classification procedure, making them highly susceptible to the operational decisions during the creation process (Openshaw and Gillard, 1978). However, geodemographics are nevertheless still valuable in many circumstances, mainly because they are practicable. Our own classification is easy to use, and offers the ability to append and update data as it becomes available, while keeping the same model infrastructure intact. In general, it meets the growing need for geodemographic systems that are open and versatile enough to handle the abundance of big data that is currently available.

REFERENCES

Arribas-Bel, D., Nijkamp, P. and Schoelten, H. (2011) Multidimensional urban sprawl in Europe: a self-organizing map approach. *Computers, Environment and Urban Systems*, **35**(4), pp. 265–275.

Arribas-Bel, D. and Schmidt, C.R. (2013) Self-organizing maps and the US urban spatial structure. *Environment and Planning B*, **40**(2), pp. 362–371.

Barbosa, O., Tratalosa, J.A., Armsworth, P.R., Davies, R.G, Fuller, R.A., Johnson, P. and Gaston, K.J. (2007) Who benefits from access to green space? A case study from Sheffield, UK. *Landscape and Urban Planning*, **83**, pp. 187–195.

Burgess, E.W. (1925) The growth of city: an introduction to a research project, in Park, R.E., Burgess, W. and McKenzie, R.D. (eds.) *The City*. Chicago, IL: University of Chicago Press.

CACI (2013) *The ACORN User Guide: The Consumer Classification.* London: CACI. Available at: http://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf.

Colwell, P., Dehring, C. and Turnbull, G. (2002) Recreation demand and residential location: the influence of sensitivity for road traffic noise on residential location: does it trigger a process of spatial selection? *Journal of Urban Economics*, **51**, pp. 418–428.

Dear, M. (2002) Los Angeles and Chicago School: invitation to debate. *City and Community*, **1**(1), pp. 5–32.

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

Experian (2014) *Mosaic: The Consumer Classification Solution for Consistent Cross-Channel Marketing.* Nottingham: Experian Ltd. Available at: http://www.experian.co.uk/assets/marketing-services/brochures/mosaic_uk_brochure.pdf.

Gidlöf-Gunnarsson, A. and Öhrström, E. (2007) Noise and well-being in urban residential environments: the potential role of perceived availability to nearby green areas. *Landscape and Urban Planning*, **83**, pp. 115–126.

Harris, R., Sleight, P. and Webber, R. (2005) *Geodemographics, GIS, and Neighbourhood Targeting*. Chichester: Wiley.

Hui, E., Chau, C., Pun, L. and Law, M. (2007) Measuring the neighboring and environmental effects on residential property value: using spatial weighting matrix. *Building and Environment*, **42**(6), pp. 2333–2343.

Janson, C.G. (1980) Factorial social ecology – an attempt at summary and evaluation. *Annual Review of Sociology*, **6**, pp. 433–456.

Kim T., Horner, M.W. and Marans, R.W. (2005) Life cycle and environmental factors in selecting residential and job locations. *Housing Studies*, **20**(3), pp. 457–473.

Kohonen, T. (2001) *Self-organizing Maps*. Berlin: Springer.

Longley, P.A. (2005) Geographical information systems: a renaissance of geodemographics for public service delivery. *Progress in Human Geography*, **29**(1), pp. 57–63.

Longley, P.A. and Goodchild, M.F. (2008) The use of geodemographics to improve public service delivery, in Hartley, J., Donaldson, C., Skelcher, C. and Wallace, M. (eds.) *Managing to Improve Public Services*. Cambridge: Cambridge University Press, pp. 176–194.

Martin, D., Nolan, A. and Tranmer, M. (2001) The application of zone-design methodology in the 2001 UK Census. *Environment and Planning A*, **33**, pp. 1949–1962.

Openshaw, S. and Gillard, A.A. (1978) On the stability of a spatial classification of census enumeration district data, in Batey, P.W.S. (ed.) *Theory and Methods in Urban and Regional Analysis*. London: Pion, pp. 101–119.

Ordnance Survey (2015) *Open Map – User Guide and Technical Specification* v1.4. Crown Copyright, London: HMSO.

Parkes, A., Kearns, A. and Atkinson, R. (2002) What makes people dissatisfied with their neighborhoods? *Urban Studies*, **39**, pp. 2413–2438.

Reibel, M. (2011) Classification approaches in neighborhood research: introduction and review. *Urban Geography*, **2**(3), pp. 305–316.

Renkow, M. and Hoover, D. (2000) Commuting, migration, and rural-urban population dynamics. *Journal of Regional Science*, *40* (2), pp. 261–287.

Rijnders, E., Janssen, N.A., van Vliet, P.H. and Brunekreef, B. (2001) Personal and outdoor nitrogen dioxide concentrations in relation to degree of urbanization and traffic density. *Environmental Health Perspectives*, **109**(3), pp. 411–441.

Shevky, E. and Bell, W. (1955) *Social Area Analysis*. Stanford, CA: Stanford University Press.

Singleton, A.D. and Longley, P.A. (2009) Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*, **29**(3), pp. 289–298.

Singleton, A.D. and Spielman, S.E. (2013) The past, present and future of geodemographic research in the United States and United Kingdom. *The Professional Geographer*, **66**(4), pp. 558–567.

Singleton, A.D, Spielman, S.E. and Brunsdon, C. (2016) Establishing a framework for open geographic information science. *International Journal of Geographical Information Science*, **30**(8), pp. 1507–1521.

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

Spielman, S.E. and Folch, D.C. (2015) Social area analysis with self-organizing maps, in Singleton, A. and Brundson, C. (eds.) *Geocomputation: A Practical Primer*. London: Sage.

Spielman, S.E. and Thill, J.C. (2008) Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, **32**(2), pp. 110–122.

Spielman, S.E. and Singleton, A.D. (2015) Studying neighborhoods using uncertain data from the American Community Survey: a contextual approach. *Annals of the Association of American Geographers*, **105**(5), pp. 1003–1025.

Vale, D.S. (2015) Transit-oriented development, integration of land use and transport, and pedestrian accessibility: combining node-place model with pedestrian shed ratio to evaluate and classify station areas in Lisbon. *Journal of Transport Geography*, **45**, pp. 70–80.

Van Ommeren, J., Rietveld, P. and Nijkamp, P. (1999) Job moving, residential moving, and commuting: a search perspective. *Journal of Urban Economics*, **46**, pp. 230–253.

Villeneuve, P.J., Jerrett, M., Su, J.G., Burnett, R.T., Chen, H., Wheeler, A.J. and Goldberg M.S. (2012) A cohort study relating urban green space with mortality in Ontario, Canada. *Environmental Research*, **115**, pp. 51–58.

Voas, D. and Williamson, P. (2001) The diversity of diversity: a critique of geodemographic classification. *Area*, **33**(1), pp. 63–76.

Webber, R. J. (1978) Making the most of the census for strategic analysis. *Town Planning Review*, **49**(3), pp. 274–284.

Wehrens, R. and Buydens, L.M.C. (2007) Self- and super-organising maps in R: the kohonen package. *Journal of Statistical Software*, **21**(5), 23–29.

ACKNOWLEDGEMENTS

This research was funded by Economic and Social Research Council grant ES/L011840/1 (Retail Business Datasafe).