

Developing a national geodemographic classification of Workplace Zones

Cockings Samantha¹, Martin David¹, Harfoot Andrew¹

¹Geography and Environmental Science, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Corresponding author: Samantha Cockings, Geography and Environmental Science, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: s.cockings@soton.ac.uk. Tel: +44 (0)23 8043 3140

ORCIDs: Samantha Cockings: 0000-0003-3333-4376, David Martin: 0000-0003-0397-0769, Andrew Harfoot: 0000-0001-9923-5018

Funding: This study was funded by the Quality Improvement Fund (QIF) administered by the Office for National Statistics (Grant numbers PU-10-0141 and PU-16-0031-6.009).

Conflict of Interest: The authors declare that they have no conflict of interest.

Acknowledgements

Individual level microdata were obtained from NRS and NISRA and accessed by University of Southampton researchers under secure conditions at ONS Titchfield via ONS Data Access Request 36. The authors are grateful to ONS staff (particularly Ian Coady, Andy Tait, Bruce Mitchell and Mark Baines) for their assistance and to members of the Workplace Zone User Group for providing valuable input during the early stages of the development of COWZ-EW. COWZ-UK is © Crown copyright and database right 2018. It is based on National Statistics, Ordnance Survey and Ordnance Survey Northern Ireland data © Crown copyright and database right (2018) and public sector information licensed under the terms of the Open Government Licence v3.0. Ordnance Survey POI data are © Crown copyright and database right 2019 Ordnance Survey (100025252; downloaded using EDINA Digimap Ordnance Survey Service, <http://digimap.edina.ac.uk>, 2018-05-29 13:48:43.089; items: 4258081) and include data licensed from PointX© Database Right/Copyright 2019. The views expressed in this paper are the authors' own.

Developing a national geodemographic classification of Workplace Zones

Abstract

Geodemographics conventionally refers to the classification of geographical areas based on the socioeconomic characteristics of their residents. In this paper, we develop the novel concept of a classification based on the characteristics of workers and workplaces. The paper describes the implementation of this concept at the small area level for the whole of the UK, which has involved reconciliation of three slightly different national censuses. It presents a summary of the resulting classification (a Classification of Workplace Zones for the UK (COWZ-UK)) and an innovative validation exercise based on comparison with a very large digital mapping dataset containing specific workplace locations. The openly available classification provides important new insights into the characteristics of workers and workplaces at the small area level across the UK, which will be useful for analysts in a range of sectors, including health, local government, transport and commerce. The generic concept of a classification based on the characteristics of workers and workplaces within a set of workplace zones is transferable to other countries, with refinement to reflect context- and country-specific phenomena. The concept can be readily implemented by census agencies or other data providers where individual level worker and workplace data are available.

Keywords

Geodemographics, classification, workplace, workplace zones, census, UK

1. Introduction

The conventional definition of geodemographics is “the analysis of people by where they live” (Leventhal 2016 p6), usually by the classification of small areas sharing common socioeconomic characteristics. Geodemographic classification is a powerful and internationally applicable (Singleton and Spielman 2014) data reduction tool which summarises complex multivariate data and is used to assist business location decisions and the delivery of public services. In this paper we develop – and implement on a national scale for the UK – a substantive enhancement to the conventional geodemographics concept by building a classification based on the characteristics of workers and their places of work.

The majority of openly available geodemographic classifications have for decades relied on small area data from censuses of population, which enumerate people primarily referenced to their places of residence. Various commercial organisations have combined census and non-census data to produce classifications targeted at different markets and sectors of the population, but their methods and data products are rarely published openly. More recently, new forms of data from business records and administrative sources have come to be incorporated into open classifications (Singleton and Spielman 2014), but until very recently these have similarly focused on characteristics of people based on where they live, such as household and dwelling characteristics. These residence-based classifications effectively create neighbourhood types reflecting the situation when residents are at home – mostly, but not exclusively, at night. During working hours (which encompass both day and night, but predominantly day), the working population is extensively redistributed over geographical space, meaning that the characteristics of these areas can change

greatly. These very different distributions also mean that residentially-based census reporting zones and associated data tend to be most detailed in areas where few people work, and least detailed in areas such as city centres and business districts with high workplace populations (Martin et al. 2013).

Many sectors require an understanding of the distribution and characteristics of the population at small area level, but for some important purposes it is the distribution and characteristics of population during the working day, rather than the (predominantly night-time) residential pattern, which is most relevant. For example, analysts and planners in the emergency, healthcare, transport and retail sectors require information about the spatial distribution of workers and workplaces in order to assess time-specific hazard scenarios, market opportunities and service delivery.

Geodemographic analysis based on place of residence can provide, at best, only partial insights for these purposes.

In this paper, we propose the novel concept of a workplace-based geodemographic classification, based on the characteristics of the working population at their places of work. We describe a programme of research leading to the creation of a UK-wide Classification of Workplace Zones (COWZ-UK), building on a recently created statistical geography of Workplace Zones (WZs), and an interim classification for England and Wales (COWZ-EW). COWZ-UK was published in 2018 by the Office for National Statistics (ONS) and is available as an open data product from

<https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011workplacebasedareaclassification>. UK-wide implementation has required generation of entirely new geographical units and datasets for Scotland and Northern Ireland. The classification is, nevertheless, methodologically consistent with previous UK classifications and is internationally applicable.

The remainder of this paper is structured as follows. In the next section, we review geodemographic classification, focusing on developments and data most relevant to classification of workers and workplaces. The third section describes the data and methods developed for the creation of COWZ-UK. The fourth section presents a summary of our new classification, including examples of the insights made possible. We then present a validation exercise based on spatial intersection of the new classification with a large map database indicative of different types of workplace locations. Finally, we present our conclusions and reflect on future opportunities.

2. Review

Geodemographic methods (Harris et al. 2005; Singleton and Spielman 2014; Leventhal 2016) comprise a variety of approaches to the identification of geographical patterns in multivariate socioeconomic data, usually by data reduction and classification. There is variation in the terms used to describe these classification methods in the academic and business literatures. For the sake of consistency, we here follow the convention adopted by the UK national statistical agencies and the most relevant academic commentators (ONS 2015; Gale et al. 2016; Leventhal 2016). Thus, geodemographic data products are here termed “classifications” and individual groupings of area types “clusters”, regardless of the specific algorithm employed in their construction. In geodemographic classification, a wide range of input variables describing demographic and social characteristics are combined through methods such as principal components analysis (PCA), k-means clustering or hierarchical agglomerative clustering (Harris et al. 2005; Leventhal 2016). The outputs of these methods are typically small numbers of distinctive clusters, with each area being allocated to the cluster to which it is closest in multidimensional variable space. These classifications often comprise two or three levels of nested clusters. Area cluster membership may be added as a contextual indicator to records relating to individual addresses, households and people. Textual and

statistical descriptors are usually provided for each cluster, to provide users with summaries of its distinctive characteristics.

There is a wide range of both general-purpose and market-specific geodemographic classification systems worldwide. Leventhal (2016) provides a useful summary of these, illustrating the diversity of geographical units, numbers of clusters and information sources employed, together with application examples spanning the marketing, retail, financial services, media and market research industries as well as the public and academic sectors. Diverse examples include airport marketing in Australia (Leung et al. 2016) and assessment of exposure to natural hazards in Italy (Willis et al. 2014). Many classifications have been developed and made available as commercial products (see for example Experian™'s Mosaic (<https://www.experian.co.uk/business/marketing/segmentation-targeting/mosaic/>) or TransUnion's CAMEO (<https://www.transunion.co.uk/products-and-services/consumer-marketing-data/segmentation-analysis>)), but these are not openly available. By contrast, Vickers and Rees (2007) and Gale et al. (2016), respectively, describe the development of the 2001 and 2011 Census-based Output Area Classifications (OAC) of the UK, both of which were entirely conceived of, and produced as, open data products.

Various current research trends are discernible in geodemographic classification, of which we note increasing variety of input data, creation of open classifications (Singleton and Spielman 2014; Singleton et al. 2016) and an increasing interest in aspects of temporality. Longley and Adnan (2016), for example, explore the creation of a geodemographic classification based on Twitter usage, making the important observation that these data are no longer tied only to residential locations. However, due to the importance of residential census data to many classifications (Burns et al. 2018), it has still been usual for residential census geographical units to form the spatial building blocks for such classifications.

Singleton et al. (2016) demonstrate one aspect of the temporal specificity of geodemographic classifications by investigating classification stability through time, presenting a new classification based on data from two consecutive censuses. This is an important innovation, but does not address the much shorter cyclical timescale with which we are interested here, namely the daily redistribution of population between residential and workplace locations. Singleton et al.'s (2016) classification is based entirely on residential Output Areas (OAs, mean population 309) and limited to England, thereby avoiding the challenge of reconciling censuses held in different countries. They provide an approach to validation by exploiting ancillary secondary data, which provide alternative characterisations of residential areas.

Rather than building a classification from area characteristics, Burns et al. (2018) develop a classification of individual (person-level) characteristics based on census microdata and then geographically distribute these using spatial microsimulation to match observed area characteristics. This approach addresses the important issue that small areas are never in reality comprised of individuals and households who all share the characteristics of the area-based geodemographic cluster to which they have been assigned and more often encompass very heterogeneous populations. In theory, the classified microdata employed in Burns et al. (2018) could be allocated to appropriate workplace locations, although this aspect is not explored.

Clearly, the patterns revealed by residential geodemographics do not exploit the full richness of census data, which also include questions about place of work. Workplaces represent both a different geographical distribution and a different temporal domain for the 45.2% of the UK population recorded by the 2011 Census as engaged in employment (source: 2011 Census tables QS104UK and QS601UK, <https://www.nomisweb.co.uk/>), and yet, to date, there has been no

geodemographic classification of UK workers and workplaces within bespoke small area units, primarily due to the lack of suitable data and boundaries.

Very few published studies have attempted to produce classifications of small areas based solely on workplace characteristics; where this has been attempted it has been in support of specific research objectives or without workplace-based small area data or boundaries. Hincks et al. (2018) produce a classification of commuting flows using census travel to work data, characterising the types of commuter travelling from origin and into destination areas. However, their analysis is based on Middle Layer Super Output Areas (MSOAs, mean population 7,806) and thus does not make use of the highest spatial resolution residence or workplace data available. Debenham et al. (2003) present a novel analysis, which integrates census and non-census data about the supply- and demand-side of residential and labour markets into a single classification. They include administrative data about employment in different industries but the data integration necessitates aggregation to postcode sectors (mean population in their study area 33,691) which are not specific to either residential or workplace distributions. Manaugh et al. (2010) build separate residential and workplace classifications for analysis of commuting in Montreal, to characterise the origin and destination neighbourhoods in a travel survey. The two classifications share the same spatial grid and pool of input variables, albeit refined by a factor analysis, but there are relatively few employment-related variables and no attempt to extend this to a larger region or more general application.

The closest existing attempt to a country-wide classification of workers and workplaces based on place of work at the small area level is a UK commercial product called Workforce Acorn (CACI, 2015). Workforce Acorn reweights CACI's established classification of residential areas onto areas of work, based on travel to work flows. It is therefore not a direct classification of workers and workplaces in areas of work and is unfortunately not openly available. Other commercial products which integrate census and non-census data to produce partial classifications of workers, workplaces or areas of work, include AFD Software's Censation (which provides a simplistic classification of commercial areas in the UK based on census and non-census data: <https://www.afd.co.uk/data-sets/censation/>), and Scan/US's Daytime demographics classification (classifies data on business establishments, employees and land-use at a range of geographical scales but not for specific workplace-based areas: <http://www.scanus.com/daytime-demographics-subscription.htm>).

The four countries of the UK are the setting for the empirical work described in this paper. Three separate censuses are conducted 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. Although comparable, there are important differences such as in questions asked, output categories, confidentiality thresholds and OA design. Census OAs (known as Small Areas in Northern Ireland, and with slightly different design criteria in each of the UK censuses) are designed to represent approximately equal numbers of residents and households and thus contain widely varying workplace populations, which in the 2001 Census ranged from 0 to 80,145. This immense variation and associated disclosure control concerns meant that much of the data about workers and workplaces collected by the 2001 Censuses could not be published.

A major innovation of the 2011 Census in England and Wales was the creation of an entirely new set of geographical units (WZs) designed specifically for the release of census workplace data (ONS 2014a). Using automated zone design methods developed by Martin et al. (2013) and previously applied to OAs (Cockings et al. 2011), ONS generated 53,578 WZs (mean worker count 493) by splitting, merging or retaining the 2011 OAs. Design of WZs respected confidentiality thresholds of 200 workers and 3 workplace postcodes, permitting 21 tables of 2011 Census aggregate outputs about the workplace population and workplaces to be published, based on geographical aggregation

of each person in employment to the WZ of their place of work. WZs are quite different in scale and purpose from the various types of functional labour market, travel to work or transportation zones used in the UK and internationally (Adams et al. 1999; Coombes and Bond 2008; FHA 2010). The value of these new workplace-based data and bespoke geographical units to a range of sectors has been demonstrated by various analysts. Waddington et al. (2017, 2019) highlight their value to retailers, particularly in the convenience sector, who are able to take account of the characteristics of the population working, as well as living, in the vicinity of existing or potential new stores, and to develop more time-specific demand and revenue prediction models. Woods (2017) uses the WZ demographic data as the framework for an analysis of workplace accidents. In local government, Hampshire County Council (2014) demonstrates the new insights to be gained into the characteristics of the workforce in local employment centres, while Hackney Borough Council (2015) employs WZs and associated data to inform their Transport Strategy for 2015-2025.

An interim classification of WZs for England and Wales (EW), called COWZ-EW, has been developed by Cockings et al. (2015; <http://cowz.geodata.soton.ac.uk/cowz-ew/>). This has already been utilised by Berry et al. (2016) in retail, by Greater London Authority (2016) in local government, and by Martin et al. (2018) in travel-to-work analysis. COWZ-EW has also since been refined by Singleton and Longley (2019) to produce a bespoke classification for London (UK). At the time of development of COWZ-EW, while there was strong user interest in development of a UK-wide classification, this was not possible due to the lack of WZs and corresponding census outputs for Scotland and Northern Ireland. The rest of this paper describes the development of a new, UK-wide open geodemographic classification of workers and workplaces, including creation of the necessary geographical units and aggregated census data for Scotland and Northern Ireland.

3. Creating a classification of Workplace Zones for the UK (COWZ-UK)

3.1 Data

To enable a UK-wide version of COWZ, WZ boundaries for Scotland and Northern Ireland were created by ONS on behalf of NRS and NISRA using the same software and, as far as possible, the same methods as for England and Wales (Martin et al. 2013; ONS/NRS/NISRA 2016). Capitalising on knowledge gained from the creation of COWZ-EW, and addressing specific differences between the UK censuses, the authors specified custom extracts of Scottish and Northern Irish 2011 census workplace microdata. These microdata, supplied by NRS and NISRA to ONS and accessed by the authors under secure conditions, were then aggregated to the newly created WZs for Scotland and Northern Ireland. Published aggregate WZ-level data for England and Wales were downloaded from NOMIS (https://www.nomisweb.co.uk/census/2011/workplace_population), thus providing the basis for a UK-wide geodemographic classification of 60,709 WZs.

As a by-product of this process, NRS and NISRA subsequently released openly available WZ boundaries for Scotland and Northern Ireland (available separately from NRS: <https://www.nrscotland.gov.uk/statistics-and-data/geography/our-products/census-datasets/2011-census/2011-boundaries>, or NISRA: <https://www.nisra.gov.uk/support/geography/northern-ireland-workplace-zones>, or as a complete UK set from the ONS Open Geography portal: <http://geoportal.statistics.gov.uk/>). NRS also released limited WZ-level data for Scotland (<https://www.scotlandscensus.gov.uk/ods-web/data-warehouse.html#additionaltab>), but NISRA were unable to publish WZ-level data due to disclosure concerns related to their previous release of data at the small area level.

The population base for COWZ-UK is the 2011 Census workplace population, defined as 'All usual residents aged 16 to 74 in employment in the area the week before the census', which includes any paid work (including casual or temporary work) as employees, self-employed (with or without employees), people on a government sponsored training scheme, people working for their own or family's business, people on sick leave, maternity leave, holiday or temporarily laid off, and full-time students who are working. Those usually resident in one country (e.g. England) but working in another (e.g. Scotland), having workplaces outside the UK (including offshore installations), not usually resident in the UK and full-time students who are not working are excluded from the workplace population, in part due to the separate processing of the censuses by the different national agencies. Census respondents answer questions related to their main job (most hours), which means that secondary employment is not included. Those who work mainly at or from home, or with no fixed place of work are georeferenced to their area of usual residence, while workers who report to a depot are georeferenced to the depot. In the UK, a census 'workplace' is thus defined as the place of work recorded by a census respondent and may not correspond to a workplace found in other lists of businesses, enterprises or companies. The location of a workplace was determined by matching its postcode to the georeferenced postcode directory held by the relevant statistical agency. Workplaces recorded in the individual census records for Scotland and Northern Ireland were georeferenced by postcode using the May 2012 version of the ONS Postcode Directory (ONSPD). However, some postcodes spanning local authority boundaries had previously been split by NRS and the fragments assigned a modified postcode: for the purposes of COWZ-UK, these postcode parts were georeferenced using a lookup file provided by NRS and then allocated to a WZ using a point-in-polygon operation, with the microdata being aggregated to WZs.

3.2 Methods

The generic concept of a workplace-based geodemographic classification can be implemented using a variety of methods and datasets. The rest of this section describes its implementation in COWZ-UK. The aim is to construct a new classification which is based entirely on 2011 Census outputs, is consistent with the design of the 2001 and 2011 residence-based OACs, addresses the interests of research users and statistical agencies, and is publishable as an open data product. Greater detail than can be reproduced here is documented in ONS (2018a), particularly the definitions of census variables.

3.2.1 Identification of relevant domains

As with any geodemographic classification, careful consideration of real-world phenomena and detailed exploration of candidate variables is required in order to define domains and select variables for inclusion. Both 2001 and 2011 versions of the residence-based OAC include census variables organised into five domains, representing key characteristics of residential areas: demographic structure, household composition, housing, socio-economic group and employment. These are clearly not all appropriate for the characterisation of workers and workplaces, but it is possible to identify four comparable domains: composition of the workplace population, composition of the built environment, socio-economic characteristics of the workplace population and employment characteristics of the workplace population. There is no direct workplace equivalent of household composition because, unlike households, individual workplaces are not units of observation in the UK censuses so there is no data about workforce size or industry sector for individual workplaces, nor is there a direct count of workplaces. The composition of the built environment domain in COWZ-UK is approximately equivalent to the housing domain in OAC, but is intended specifically to capture the relative workplace/residential mix of a WZ. As in OAC, distance

travelled and mode of travel to work are included within the socio-economic characteristics of workplace population domain, based on the behaviour of individual workers.

3.2.2 Selection and preparation of variables for classification

When producing the interim COWZ-EW classification, an initial long list of 501 candidate variables, from 13 of the 21 published WZ-level tables, was identified, covering the four domains (see Table 1 in Cockings et al. 2015). Three additional bespoke variables (the number of workplace postcodes (obtained directly from ONS), density of workplace postcodes (per hectare), and ratio of the number of OAs to WZs) were also explored as proxies for the composition of the built environment as this was not otherwise reflected in the published data. The remaining eight WZ-level tables were excluded because they did not fall within the scope of the classification, or because they were highly correlated with variables in other tables. Following in-depth analysis, this long list was reduced first to 63 and then to 48 variables, which formed the final inputs to the COWZ-EW cluster analysis.

In producing COWZ-UK, the full long list of 504 variables is not re-evaluated, as the previous COWZ-EW analysis had already shown that many of these contribute little useful information to a workplace classification and there is no reason to expect this to be different for the whole of the UK. Analysis for COWZ-UK instead focuses on the 63 variables which comprised the intermediate set of candidate variables for COWZ-EW. These were constructed by the authors for Scotland and Northern Ireland from the microdata extracts supplied by NRS and NISRA.

The definitions of the vast majority of variables are consistent across the UK, or can readily be mapped to equivalents. However, special treatment is required for four variables whose differences originate from variations in the census questions and/or coding of responses. Variable names refer to those shown in Table 1. The categorisation of White British (WP201_WhiteBrit) ethnicity is slightly different in the Scotland and Northern Ireland censuses compared to England and Wales and it is not possible to fully reconcile the Northern Ireland definition with the others. Similarly, there are slight differences in the classification of highest level of qualification (WP501_GE_L4, WP501_L3, WP501_NoQual), for which the Scottish definition cannot be completely reconciled with the others. Full-time students (WP601_FT_Stud) are identified slightly differently in all three censuses, although a comparable count for each country can be obtained from: the employment status variable for England and Wales; a separate student status variable for Northern Ireland; and socio-economic classification for Scotland. The England and Wales census asked only about travel to place of work (WP702 and WP703), but in Scotland and Northern Ireland the equivalent question asked about travel to place of work or study. For full-time students also in employment, it is not possible to determine whether the response relates to place of work or study and this cannot be inferred from either the routing of the questions, or from post-enumeration processing. In order to reduce the impact of this discrepancy on COWZ-UK, full-time students were excluded from just the calculation of travel to work variables for Scotland and Northern Ireland. Full details of the differences and associated solutions are provided in ONS (2018a). All counts are expressed as percentages of the workplace population (all usual residents aged 16 to 74 in employment in the week before the census), except for Approximated social grade (WP613) for which the denominator is all usual residents aged 16 to 64 in employment in the week before the census.

Despite the essentially data-driven nature of geodemographic classification, the analyst must make a number of key decisions concerning the selection and refinement of variables. For COWZ-UK, comparable methodological steps to those set out by Vickers and Rees (2007) in their development of the original 2001 OAC are followed. The statistical and geographical distributions and inter-relationships between the 63 intermediate variables are explored using a combination of simple

summary statistics (mean, median, range, standard deviation), histograms, Quantile-Quantile (Q-Q) (normal probability) plots, maps at various geographical scales, and a correlation matrix.

Decisions concerning whether to normalise variables prior to cluster analysis, and which variables to include/exclude are non-trivial (Spielman and Singleton 2015; Singleton et al. 2018). Geographically-aggregated census data typically exhibits many zero counts and outliers for specific variables. Such skewed distributions can be problematic for cluster analysis. Both 2001 and 2011 OACs were normalised prior to clustering. Following Gale et al (2016), the log, Box-Cox and inverse hyperbolic sine (IHS) methods of transformation are evaluated for COWZ-UK, with the Box-Cox method producing the best overall results and therefore being implemented for all variables. Pair-wise correlation between the 63 normalised variables is then assessed using Pearson's Product Moment Correlation coefficient (see Online Resource 1). Rather than employing rigid correlation coefficient or significance thresholds, final inclusion/exclusion decisions are based on statistical and mapped analysis and assessment of the variable's overall contribution to the classification.

One such example includes two highly correlated groups of variables: Approximated Social Grade (ASG) (variables WP613_AB to WP613_DE) and National Statistics Socio-economic Classification (NS-SeC) (WP607_1 to WP607_7). Both ASG and NS-SeC are derived variables, but NS-SeC is a National Statistics product and incorporates information on supervisory relationships and size of employer. The NS-SeC variables are therefore retained (with the exception of WP607_4 (Small employers and own account workers) which is very strongly correlated with other variables such as Self-employed with employees, Self-employed without employees, No fixed place of work and Work from home), while the ASG variables are excluded.

A further example includes composite and component variables: Employed (WP601_Emp) and Self-employed (WP601_SEmp) are strongly negatively correlated. Employed exhibits distinct spatial patterning, particularly concentrated in urban areas and positively correlated with intermediate occupations and travelling greater than 20km. The two component variables of Self-employed: Self-employed with employees (WP601_SEmp_w_E) and Self-employed without employees (WP601_SEmp_wo_E) demonstrate important relationships with other variables: the former positively correlated with longer and shorter working hours, Work from home, No fixed place of work, Construction and Education; the latter with industries such as Arts and entertainment. Employed, Self-employed with employees and Self-employed without employees are therefore retained whereas the composite Self-employed variable are excluded. Following these detailed evaluations, the 63 intermediate variables are refined to a final input set of 48, shown in Table 1 (grouped by domain).

<Table 1 about here>

Various researchers note the importance of standardising variables prior to cluster analysis to avoid those with a much greater range in their values carrying a disproportionate weight in the classification. As per Vickers and Rees (2007), Spielman and Singleton (2015) and Gale et al (2016), the final 48 variables for COWZ-UK are range standardised to produce values in the range 0 to 1 prior to clustering.

3.2.3 Cluster analysis

There is a range of algorithms for creating geodemographic classifications, including top-down and bottom-up approaches (Spielman and Singleton, 2015). To maintain consistency with previous open residence-based classifications, COWZ-UK employs a top-down k -means clustering method, with repeated application to create a nested hierarchy i.e. the top level of the hierarchy is created first

and each cluster is then further sub-divided to create the next level down. This is implemented in R using the *kmeans* function and the default Hartigan-Wong algorithm, with 10,000 random starts. The squared Euclidean distance is used to evaluate the degree of similarity within and between clusters.

For the top level of the COWZ-UK hierarchy (termed Supergroups in line with the 2001 and 2011 OACs), solutions containing two to twelve clusters are evaluated. Each of these solutions is then further subdivided into two to six clusters (termed Groups). Evaluation of potential cluster analysis solutions typically involves application of both objective and subjective criteria. Here, solutions are evaluated using: prior expectations based on the literature and the authors' knowledge of specific locations; statistical and graphical diagnostics (such as compactness of cluster solutions, homogeneity of cluster sizes, and stability of clusters); and mapping using geographical information systems (GIS) and visual imagery. Analysis takes place at various geographical scales (local, regional and national) to ensure a thorough understanding of the spatial and statistical patterns.

As observed by others (e.g. Singleton et al. 2018), scree plots showing the mean distance to the cluster centre relative to the number of clusters prove generally unhelpful here in determining the most suitable number of clusters, as the gradients of the curves are relatively smooth throughout. Similarly, while it can be useful to aim for homogeneity of cluster size throughout the hierarchy (expressed as the range in the number of members (WZs) per cluster, relative to the number of clusters), some clusters will inevitably have slightly higher/lower membership due to their distinctiveness and it is often desirable to retain this specificity. The most insightful diagnostics include: clustergrams (see for example Figure 1), radial plots (Figure 2), and sense-checking using tools such as Google Maps and Street View to corroborate the authors' expectations of the classification of specific areas.

Figure 1 is a bespoke clustergram (Schonlau, 2004) of the potential Supergroup solutions, which permits evaluation of the stability of potential clusters and solutions. Each row represents a solution, with each circle representing a cluster within that solution i.e. row 1 shows all 60,709 WZs in one cluster, row 2 shows the two-cluster solution, row 3 the three-cluster solution, and so on. For each cluster, the area of the circle is proportional to the number of members (WZs) in that cluster. The thickness of the arrows between clusters shows the percentage of WZs in the cluster in row n contributing to the relevant cluster in row $n + 1$. For example, of the 60,709 WZs in the one-cluster solution, 54.3% contributed to the left-hand cluster in the two-cluster solution, and 45.7% to the right-hand cluster. To aid clarity, only contributions greater than 5% are drawn.

Using Figure 1, it is possible to identify the levels at which clusters emerge and remain stable. For example, the two-cluster solution effectively splits the WZs into "Rural" (right) and "Urban" (left) clusters. The "Rural" cluster remains relatively distinct throughout the two- to seven-cluster solutions, whereas the "Urban" cluster splits off into various combinations of groupings, which gradually coalesce into stable solutions at around the six- to eight-cluster solutions. By evaluating clustergrams for both the Supergroup and Group levels of the hierarchy, it is also possible to identify clusters from the Supergroup solutions which become clusters in their own right at the Group level of the hierarchy. For example, in the eight-cluster solution in Figure 1, the "Rural" cluster splits off into two distinct new clusters: in the final COWZ-UK classification, these both become Groups. Fundamentally, beyond the seven-cluster solution, any sub-divisions of the Supergroups become Groups in the next level of the hierarchy. The seven-cluster solution is therefore selected for COWZ-UK, and Figure 1 shows the clusters at this level labelled according to the codes used in the published classification (and as per Table 2).

< Figure 1 about here >

Figure 2 is an illustrative radial plot. Each radial axis of the plot relates to one of the 48 variables used in COWZ-UK, labelled according to the codes in Table 1 but ordered as per the final ONS published products. The value of the cluster centre for each variable is plotted on these axes and these points are then joined to form an irregular polygon (blue) which reveals the distinctive characteristics of the cluster, relative to the UK mean values for the variables (indicated by the (red) circle). If the point lies outside the circle then that variable has a higher than average value, and vice versa. Thus, Figure 2 (for Supergroup E: Manufacturing and distribution, taken from the final COWZ-UK classification) shows that this cluster is characterised by higher than average percentages in Manufacturing (WP605_C), Transport and storage (WP605_H), Employee (WP601_Emp), Lower supervisory and Routine occupations (WP607_5 and WP607_7) and Travelling over 20km to work (WP702_GE20), together with lower than average percentages of Females (WP1101_F_16_74), High qualifications (WP501_GE_L4) and Work from home (WP702_WkFromHome).

< Figure 2 about here >

Once the number of levels in the hierarchy and clusters per level has been determined, a pen portrait for each cluster is produced and the clusters are named. This naming process is challenging and often contentious (Vickers and Rees, 2007). Names are proposed by the authors and then jointly reviewed by themselves and ONS, using experience gained from the production and naming of other official area classifications. The names are intended to be: accurate in terms of the input variables; neutral rather than complimentary or derogatory; consistent throughout the hierarchy; consistent with, but not duplicative of, their use in other classifications; and independent of placenames to avoid undue specificity.

3.3 Results: a classification of Workplace Zones for the UK

COWZ-UK is a two-tiered geodemographic classification of WZs, comprising 7 Supergroups and 29 Groups. Table 2 shows the hierarchy of Supergroups and Groups, together with the codes and names assigned to each cluster and various descriptive statistics. The full open datasets, including cluster memberships, squared Euclidean distance from centroids and pen portraits for the Supergroups and Groups are available from

<https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011workplacebasedareaclassification>. Interactive maps facilitating exploration of COWZ-UK are provided by Consumer Data Research Centre (CDRC) Maps:

<https://maps.cdrc.ac.uk/#/geodemographics/cowzuk11/default/BTTTFFT/10/-0.1500/51.5200/>.

<Table 2 about here>

As expected, Table 2 shows that the rural Groups cover the greatest land areas, with Rural with Mining and Quarrying at 42.55% and Traditional Countryside at 28.39%, although in combination they account for only 12.8% of WZs and 8.6% of workers. The Suburban Services Supergroup contains the two Groups with the largest percentage of workers, at 14.79% in combination, in just 3.62% of the land area. The highest spatial concentration of workers is achieved in the Global Business group, with 3.77% of workers in just 0.01% of land area (a density of 599 workers per hectare).

Figure 3 maps the national distribution of the Supergroups. As highlighted by Table 2, it is clear that the land area is dominated by the Rural Supergroup, with all others being concentrated in urban areas. In Figure 4, the spatial distributions of Groups for three coastal cities in different parts of the

UK are illustrated, masked with a buildings layer to aid interpretation. Belfast and Aberdeen are characterised by extensive industrial and docks-related areas categorised as Groups of the Manufacturing and Distribution Supergroup. Both have city centres reflecting a mix of Groups primarily within the Retail, City and Business Parks and Servants of Society Supergroups, but in Belfast this is much more closely ringed by the Suburban Services Groups, while Aberdeen demonstrates a more complex patchwork of Groups across the whole urban area. The Brighton map contrasts greatly, having only a small area of port-related activity but a centre of intense Retail and City and Business Park Groups, surrounded by a wide band of Metro Suburbs, which tend to have higher status occupations than those of the extensive Suburban Services seen in Belfast. The complex internal structure of each city is reflected in the classification, while the substantial differences between them are also captured, reflecting their unique industrial, social and settlement histories.

< Figure 3 about here >

< Figure 4 about here >

4. Validation of the new classification

Spielman and Singleton (2015) and Openshaw et al. (1980) note that the accuracy of classifications cannot really be measured; rather, they can only be evaluated relative to the purpose for which they were designed. Vickers and Rees (2011) describe 'ground-truthing' a geodemographic classification by recruiting a panel of 61 peer reviewers able to comment on results in areas with which they are familiar. As a preliminary validation step, the COWZ-UK categories of approximately 100 specific places (such as large shopping centres, industrial areas, educational and health sites, suburbs, high streets, different types of rural areas) are sense-checked against the authors' prior expectations using maps and visual imagery. This proves generally reassuring in terms of anticipated classifications but is clearly limited in scope, coverage and objectivity.

Leventhal (2016, p55) suggests that, as well as mapping and 'sense checking' the locations of the clusters, validation of geodemographic classifications should include comparison against other datasets. There are many datasets which could potentially be intersected with COWZ-UK, but none which offers a definitive way to assess 'accuracy' per se, given that there is no 'true' classification against which to compare the results. One approach is to compare the COWZ-UK outputs with independent location data for identifiable workplace types such as hospitals, business parks, farms etc. While COWZ-UK is a classification of small areas rather than individual workplaces, large workplaces or groups of similar workplaces will strongly influence the classification: comparing patterns of association between the location of such features and the COWZ-UK categories therefore provides an independent source of verification. Open source datasets such as OpenStreetMap cover the entire UK, but lack consistency of feature coverage and definitions. Ordnance Survey GB publishes a Points of Interest® (POI) database (Ordnance Survey, 2014) containing 4,258,081 geocoded features (March 2018 release), using a detailed 3-level classification. This database permits identification of relevant features for comparison with the COWZ-UK classification. Northern Ireland is covered by a separate national mapping agency with no directly comparable dataset available: the following analysis is therefore necessarily restricted to England, Wales and Scotland. POI features are date-stamped with the date of data supply but it is not possible to extract the feature set current at a historical date (such as 2011 for the census), so any comparison is inevitably impacted by changes in POI features over time.

The POI database is general purpose and therefore contains many natural and built physical features, such as ponds and wind turbines, which are not associated with a workplace type and will rarely be places of usual employment: these are excluded from the analysis. Similarly, many POI features which probably do represent places of work but which are likely too small (in terms of number of workers) to influence the classification of a WZ are not included. For example, the high level Education and Health code 05 includes a range of facilities ranging from code 05280368 'Dental surgeries' to 05280370 'Hospitals': while an individual hospital's workforce will often be large enough to influence the classification of a WZ, that of a dental surgery will not.

Table 3 identifies 17 POI feature codes, or combinations of codes, likely to be associated with distinctive, large, workplaces. These account for 1,054,697 POIs, 24.9% of the total. These point features are extracted and intersected with WZ boundaries to identify the COWZ-UK Group to which they are allocated. Less than 0.01% of POIs cannot be matched, mostly because they are offshore and fall outside the WZ geography (e.g. POI code 07390512 'Fish and shellfish', within 0739 'Farming', comprises mostly offshore mussel beds). This results in 1,053,645 POI records with a matched COWZ-UK Group code (shown broken down by POI code group in the final column of Table 3).

<Table 3 about here>

The 17 POI code groups are then cross-tabulated with the 29 COWZ-UK Groups. In order to account for the very different number of features in each comparison group, expected values ((row total*column total)/grand total) are calculated for each intersection, assuming a totally uniform distribution of POI code groups across the COWZ-UK Groups. The number of features observed in each intersection is then compared to the expected value, with differences expressed as a ratio [(observed-expected)/(observed+expected)], thus allowing groups with much greater (or lower) than expected prevalence to be identified.

The results are summarised as a heatmap in Figure 5, which provides a clearer visualisation of the relationships than a simple numerical table. Higher than expected values are shown in red; lower than expected in blue. The POI code groups are re-ordered from Table 3 to aid interpretation. An exact relationship between POI code groups and COWZ-UK Groups is not to be expected, but there are clear structural relationships, with a broad diagonal of strongest association through retail and business POI code groups in the relevant COWZ-UK Groups, through industrial and rural activities to education health and public administration. The extractive industry and farming POI code groups are strongly present in the Rural COWZ-UK Groups and strongly absent from all other types. There are also isolated hotspots such as the strong presence of government and hospitals in the COWZ-UK Administrative centres Group. Clothing and accessories retail is under-represented in almost all COWZ-UK groups except those associated with retail and business (and highest in 'Shop until you drop'). As might be expected, suburban areas offer some of the least clear patterns, but show some overrepresentation of food and multi-item retail, sport, entertainment and education, reflecting local community services and businesses. Taken as a whole, this analysis provides further confidence in the COWZ-UK classification.

<Figure 5 about here>

5. Discussion and conclusions

This paper has introduced the novel concept of a workplace-based geodemographic classification and demonstrated its implementation via COWZ-UK, a national small area level classification of WZs for the UK. While validation of such classifications is difficult, the analysis presented here suggests

that COWZ-UK performs as per *a priori* expectations and that users can have confidence in its outputs. The concept of a workplace-based geodemographic classification is generic and could readily be implemented by census or statistical agencies in other countries where the characteristics of workers and their workplaces are recorded at the individual level. Refinements to the domains and variables are likely to be required, depending on context- and country-specific phenomena. The classification could be produced for any geographical units to which the individual level data can be aggregated. A link between place of work and place of residence also permits the inclusion of variables related to travel to work.

COWZ-UK is an openly available data product, consistent with other existing census-based classifications. As described in Section 3.1, not all of the source data employed in the creation of COWZ-UK are publicly available, primarily due to its retrospective creation following the processing and release of 2011 Census data in the UK. While the methods are transparent, users are therefore unable to fully replicate its production.

COWZ-UK shares some limitations with all census-derived datasets. Despite the overall high quality of census data, questions relating to topics such as place of work and industry of employer are among the least accurately completed (ONS 2014b). Inaccuracies are also introduced both by the assignment of workers with no fixed place of work to their residential OA, and incorrect allocations of some workers to business headquarters rather than actual places of work. A specific direct consequence of the separate production of 2011 workplace census statistics by ONS, NRS and NISRA is that cross-border flows are entirely missing from the source datasets and therefore from COWZ-UK. Only people living and working within a specific country are included – people living in England and working in Scotland, for example, are not included. Nevertheless, the analysis presented here suggests that COWZ-UK provides an original and meaningful characterisation of workers and workplaces at the small area level.

COWZ-UK is a classification of both workers and workplaces. 31 out of the 48 input variables relate to workers, with variables which are more geographically concentrated notably driving the classification. For example, ethnic group/country of birth and the presence of public transport systems result in clear differentiation between metropolitan and non-metropolitan areas. London is so distinct in a range of variables compared to the rest of the country that, relatively, it lacks internal differentiation. Singleton and Longley's (2019) recent London-specific classification emphasizes this local variation compared to the national COWZ-UK version.

While 17 of the input variables relate to workplaces, 15 of these are a typology of industry. The influence of industry type on COWZ-UK is therefore quite profound. A broader range of workplace-related variables would be useful but, as noted previously, the UK census is not a register of businesses and only indirectly records information about employers and businesses via workers' reporting of them. Industry type also arguably exerts a strong influence on the validation of COWZ-UK presented in this paper. Much subjective *a priori* expectation of the classification of small areas is based on the type of industry present, probably because it is the most easily observed visually, compared to the less obvious characteristics of the workforce in an area. Similarly, the POI feature-type categorisations are largely based on industry type rather than any workforce characteristics. Independent observation, ground-truthing and measurement of the characteristics of the workforce are therefore more challenging in this respect.

Current UK statistical agency plans are to update OAs and WZs for the 2021 Censuses, thus maintaining stable geographies (HM Government, 2018). This presents a real opportunity to harmonise definitions and processing across the three UK statistical agencies, and to address some

of the limitations of the current WZ and COWZ outputs such as definitional and geocoding differences and the treatment of cross-border flows. Unlike in 2011, WZs and associated data should be able to be released with the rest of the census outputs, increasing user awareness and uptake. It is reasonable to expect that new versions of COWZ will be produced following the 2021 Censuses, thus providing users with a valuable time series of workplace-based classifications upon which to base analyses. WZs have also recently been produced by the Central Statistical Office for Ireland (CSO, 2018), therefore it would now also be possible to extend the COWZ suite of products to Ireland.

As with any census-derived data product, the accuracy of COWZ-UK will degrade over time. The rate of change in the distribution and characteristics of workers and workplaces is likely to be more rapid than that of residents and households. It is highly likely that such changes will be geographically concentrated, with certain COWZ-UK categories being affected more than others, due to their varying relative concentrations in different sectors of the economy. The maintenance required for WZs may be higher than that predicted for OAs and will therefore need to be factored into Census 2021 plans.

Looking further ahead, beyond the 2021 Census, it can be anticipated that workplace, business and employment data will form an important aspect of the broader transformation of traditional census products to make much greater use of administrative and survey data (ONS, 2018b), which in turn has the potential to address some of the challenges identified here and offer more frequent updating of derived classifications such as COWZ-UK.

References

Adams, J. S., VanDrasek, B. J., & Phillips, E. G. (1999). Metropolitan area definition in the United States. *Urban Geography*, 20(8), 695-726.

Berry, T., Newing, A., Davies, D., & Branch, K. (2016). Using workplace population statistics to understand retail store performance. *The International Review of Retail, Distribution and Consumer Research*, 26(4), 375–395.

Burns, L., See, L., Heppenstall, A. & Birkin, M. (2018). Developing an individual-level geodemographic classification. *Applied Spatial Analysis and Policy*, 11, 417-437.

CACI (2015). *Workforce Acorn Product Sheet*. CACI Ltd.
https://www.caci.co.uk/sites/default/files/resources/Workforce_Acorn_product_sheet.pdf. Accessed 24 July 2019.

Cockings, S., Harfoot, A., Martin, D., & Hornby, D. (2011). Maintaining existing zoning systems using automated zone design techniques: Methods for creating the 2011 Census output geographies for England and Wales. *Environment and Planning A*, 43(10), 2399–2418.

Cockings, S., Martin, D., & Harfoot, A. (2015). *A Classification of Workplace Zones for England and Wales (COWZ-EW): User Guide*. University of Southampton.
http://cowz.geodata.soton.ac.uk/download/files/COWZ-EW_UserGuide.pdf. Accessed 24 July 2019.

Coombes, M. G., & Bond, S. (2008). *Travel-to-Work Areas: the 2007 review*. London, UK: Office for National Statistics.

CSO (2018). *Workplace Zones and 1km² population grids – Ireland*. Central Statistics Office.
<https://www.cso.ie/en/census/census2016reports/workplacezonesand1kmpopulationgrids/>. Accessed 24 July 2019.

Debenham, J., Clarke, G., & Stillwell, J. (2003). Extending geodemographic classification: A new regional prototype. *Environment and Planning A*, 35(6), 1025–1050.

FHA (2010). *TAZ Delineation Business Rules (March 2010)*. Federal Highway Administration.
http://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/tazddbrules.cfm. Accessed 24 July 2019.

Gale, C. G., Singleton, A. D., Bates, A. G., & Longley, P. A. (2016). Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science*, 12, 1-27.

Greater London Authority (2016) *Economic Evidence Base for London 2016*. Greater London Authority.
https://www.london.gov.uk/sites/default/files/economic_evidence_base_2016.compressed.pdf. Accessed 19 December 2019.

Hackney London Borough Council (2015) *Hackney Transport Strategy 2015-2025: Evidence Base Paper 1: Census 2011 Travel to Work Data-Transport Analysis*. Hackney London Borough Council.
https://drive.google.com/file/d/1tu4p_CGFB29e-nBExceDR72Br-RsTLqE/view. Accessed 19 December 2019.

Hampshire County Council (2014) *Census 2011 Workplace Zones: Examples of Workplace Population Data*. Hampshire County Council.

<https://documents.hants.gov.uk/Economy/ExamplesofWorkplaceZones.pdf>. Accessed 19 December 2019.

Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS and neighbourhood targeting*. Chichester, UK: Wiley.

Hincks, S., Kingston, R., Webb, B. & Wong, C. (2018). A new geodemographic classification of commuting flows for England and Wales. *International Journal of Geographical Science*, 32, 663-684.

HM Government (2018). *Help Shape Our Future: The 2021 Census of Population and Housing in England and Wales*. HM Government.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/765089/Census2021WhitePaper.pdf. Accessed 24 July 2019.

Leung, A., Yen, B.T.H., & Lohmann, G. (2016). Why passengers' geo-demographic characteristics matter to airport marketing. *Journal of Travel and Tourism Marketing*, 34, 833-850.

Leventhal, B. (2016). *Geodemographics for marketers*. London, UK: Kogan.

Longley, P.A., & Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, 30 (2), 369–389.

Manaugh, K., Miranda-Moreno, L. F., & El-Geneidy, A. M. (2010). The effect of neighbourhood characteristics, accessibility, home–work location, and demographics on commuting distances. *Transportation*, 37(4), 627–646.

Martin, D., Cockings, S., & Harfoot, A. (2013). Development of a geographical framework for census workplace data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2), 585–602.

Martin, D., Gale, C., Cockings, S., & Harfoot, A. (2018). Origin-destination geodemographics for analysis of travel to work flows. *Computers, Environment and Urban Systems*, 67, 68-79.

ONS (2014a). *Workplace Zones: A new geography for workplace statistics*. Office for National Statistics. <https://geoportal.statistics.gov.uk/datasets/workplace-zones-a-new-geography-for-workplace-statistics>. Accessed 24 July 2019.

ONS (2014b). *2011 Census Quality Survey*. Office for National Statistics.

<https://www.ons.gov.uk/file?uri=/census/censustransformationprogramme/consultations/the2021censusinitialviewoncontentforenglandandwales/2011censusqualitysurveyreport.pdf>. Accessed 24 July 2019.

ONS (2015). *Methodology note for the 2011 area classification for output areas*. Office for National Statistics. <http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/methodology-and-variables/methodology-oa.pdf>. Accessed 24 July 2019.

ONS (2018a). *Classification of workplace zones for the UK: methodology and variables*. Office for National Statistics.

<https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011workplacebasedareaclassification/classificationofworkplacezonesfortheukmethodologyandvariables>. Accessed 24 July 2019.

ONS (2018b). *Annual assessment of ONS's progress on the Administrative Data Census: July 2018*. Office for National Statistics.

<https://www.ons.gov.uk/file?uri=/census/censustransformationprogramme/administrativecensususproject/administrativecensusannualassessments/annualassessmentofonssprogressontheadministrativecensusjuly2018/annualassessment2018.pdf>. Accessed 24 July 2019.

ONS/NRS/NISRA (2016). *Workplace zones 2011 supporting information*. Office for National Statistics, National Records of Scotland, Northern Ireland Statistics and Research Agency.

<https://www.nrscotland.gov.uk/files/geography/products/workplace-zones2011-supporting-information.pdf>. Accessed 24 July 2019.

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

Ordnance Survey (2014). *Points of Interest database: user guide and technical specification*. Ordnance Survey. <https://www.ordnancesurvey.co.uk/docs/user-guides/points-of-interest-user-guide.pdf>. Accessed 24 July 2019.

Schonlau, M. (2004). Visualizing Hierarchical and Non-Hierarchical Cluster Analyses with Clustergrams. *Computational Statistics*, 19(1), 95-111.

Singleton, A. D., & Longley, P. A. (2019). Data infrastructure requirements for new geodemographic classifications: The example of London's workplace zones. *Applied Geography*, 109, <https://doi.org/10.1016/j.apgeog.2019.102038> Accessed 24 July 2019.

Singleton, A., Pavlis, M., & Longley, P. A. (2016). The stability of geodemographic cluster assignments over an intercensal period. *Journal of Geographical Systems*, 18, 97-123.

Singleton, A. D., & Spielman, S. E. (2014). The Past, Present, and Future of Geodemographic Research in the United States and United Kingdom. *Professional Geographer*, 66(4), 558-567.

Spielman, S., & Singleton, A. (2015). Studying Neighborhoods Using Uncertain Data from the American Community Survey: A Contextual Approach. *Annals of the Association of American Geographers*, 105(5), 1003-1025.

Vickers, D., & Rees, P. (2007). Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2), 379–403.

Vickers, D., & Rees, P. (2011). Ground-truthing geodemographics. *Applied Spatial Analysis*, 4, 3-21.

Waddington, T. B. P., Clarke, G. P., Clarke, M., & Newing, A. (2017). Open all hours: spatiotemporal fluctuations in U.K. grocery store scale and catchment area demand. *The International Review of Retail, Distribution and Consumer Research*, 28, 1-26.

Waddington T., Clarke G., Clarke M.C., Hood N., & Newing, A. (2019). Accounting for Temporal Demand Variations in Retail Location Models, *Geographical Analysis*, 51(4), 426-447.

Willis, I., Gibin, M., Barros, J., & Webber, R. (2014). Applying neighbourhood classification systems to natural hazards: a case study of Mt Vesuvius. *Natural Hazards*, 70, 1-22.

Woods, L. (2017). *A statistical and geographical analysis of workplace accidents in England and Wales*. PhD thesis. Liverpool, UK: University of Liverpool.

Table 1 Final set of 48 variables input to COWZ-UK classification, grouped by domain

Variable code ^a	Variable name
Composition of workplace population	
WP102_WPPOPDENS	Workplace population density (number of persons per hectare)
WP1101_F_16_74	All categories: Age 16 to 74, Females
WP1101_F_25_39	All categories: Age 25 to 39, Females
WP1101_MF_16_24	All categories: Age 16 to 24, All categories: Sex
WP1101_MF_GE60	All categories: Age 60 to 74 , All categories: Sex
WP201_Black	Black: African/Caribbean/Other Black
WP201_In_Pk_Bng	Asian/Asian British: Indian, Pakistani, Bangladeshi
WP201_WhiteBrit	White: English/Welsh/Scottish/Northern Irish/British
WP203_EU_2001_11	Europe: Other Europe: EU countries: Accession countries April 2001 to March 2011: Total
WP203_EU_pre2001	Europe: Other Europe: EU countries: Member countries in March 2001: Total
Composition of built environment	
OAWZRATIO	Ratio of Output Areas to Workplace Zones
Socio-economic characteristics of workplace population	
WP501_GE_L4	Level 4 qualifications and above
WP501_L3	Level 3 qualifications
WP501_NoQual	No qualifications
WP607_1	1. Higher managerial, administrative and professional occupations
WP607_2	2. Lower managerial, administrative and professional occupations
WP607_3	3. Intermediate occupations
WP607_5	5. Lower supervisory and technical occupations
WP607_6	6. Semi-routine occupations
WP607_7	7. Routine occupations
WP702_GE20	20km and over
WP702_LT5	Less than 5km
WP702_NoFixedPl	No fixed place
WP702_WkFromHome	Work mainly at or from home
WP703_PubTrans	Underground, metro, light rail or tram, Train, Bus, minibus or coach
WP703_Walk_Bike	Bicycle, On foot
Employment characteristics of workplace population	
WP601_Emp	Employee: Full or Part-time
WP601_FT_Stud	Full-time student
WP601_SEmp_w_E	Self-employed with employees: Full or Part-time
WP601_SEmp_wo_E	Self-employed without employees: Full or Part-time
WP604_FT_GE49	Full-time: 49 or more hours worked
WP604_PT_16_30	Part-time: 16 to 30 hours worked
WP604_PT_LE15	Part-time: 15 hours or less worked
WP605_A	A Agriculture, forestry and fishing
WP605_B	B Mining and quarrying

WP605_C	C Manufacturing
WP605_D_E	D Electricity, gas, steam and air conditioning supply / E Water supply; sewerage, waste management and remediation activities
WP605_F	F Construction
WP605_G_L	G Wholesale and retail trade; repair of motor vehicles and motor cycles / L Real estate activities
WP605_H	H Transport and storage
WP605_I	I Accommodation and food service activities
WP605_J	J Information and communication
WP605_K	K Financial and insurance activities
WP605_M	M Professional, scientific and technical activities
WP605_O	O Public administration and defence; compulsory social security
WP605_P	P Education
WP605_Q	Q Human health and social work activities
WP605_R_S	R,S Arts, entertainment and recreation; other service activities

^a Variable codes are derived from the published Workplace Zone table names for England and Wales, but relate to data for all of the UK

Table 2 COWZ-UK Supergroups and Groups: codes, names and descriptive statistics

Supergroup	Supergroup name	Group	Group name	Number of WZs	% of WZs	% of workers ^a	% of land area
A	Retail	A1	Low density wholesale and retail	1,070	1.76	2.07	0.20
		A2	Market squares	1,336	2.20	1.67	0.06
		A3	Multicultural urban high streets	986	1.62	1.58	0.05
		A4	Traditional high streets	1,092	1.80	1.48	0.02
		A5	Shop until you drop	778	1.28	1.76	0.02
		A6	Eat, drink and be merry	1,039	1.71	1.45	0.05
B	City and business parks	B1	Global business	1,258	2.07	3.77	0.01
		B2	Administrative centres	1,023	1.69	3.27	0.04
		B3	Big city life	1,324	2.18	2.24	0.02
		B4	Regional business centres	1,032	1.70	2.70	0.04
		B5	Science and business parks	1,047	1.72	2.81	0.17
C	Metro suburbs	C1	Teachers and carers in metro suburbs	1,971	3.25	2.84	0.29
		C2	Independent professional metro services	1,985	3.27	2.63	0.14
		C3	Metro suburban distribution industries	1,939	3.19	2.86	0.26
		C4	Self-employed tradespeople in multicultural metro suburbs	1,764	2.91	2.48	0.33
		C5	Professional home-workers in outer suburbs	1,873	3.09	2.39	0.58
D	Suburban services	D1	Non-metropolitan suburban areas	6,458	10.64	8.23	1.81
		D2	Primarily residential suburbs	5,335	8.79	6.56	1.81
E	Manufacturing and distribution	E1	Mining and quarrying facilities	1,552	2.56	3.37	0.91
		E2	Industrial units	2,119	3.49	4.50	0.39
		E3	Business parks	1,851	3.05	4.08	0.57
		E4	Manufacturing, energy and utilities	2,523	4.16	4.88	1.01
F	Rural	F1	Town fringe countryside	2,503	4.12	3.11	7.87
		F2	Accessible countryside	2,442	4.02	3.39	11.09
		F3	Rural with mining or quarrying	4,321	7.12	4.88	42.55
		F4	Traditional countryside	3,449	5.68	3.73	28.39
G	Servants of society	G1	Public administration and security	2,859	4.71	6.06	0.53
		G2	Healthcare and public administration	1,746	2.88	5.66	0.27

		G3	Large scale education		2,034	3.35	3.55	0.53
				Total	60,709	100	100	100

^a Percentage of workers for Great Britain (England, Wales, Scotland) only as Northern Ireland data are not publicly available

Table 3 Ordnance Survey GB Points of Interest (POI) code groups selected for COWZ-UK validation

POI code groups selected for study		Number of matched features
Level 1 POI Groups	Level 2 POI Categories	
01 Accommodation, eating and drinking		202,602
02 Commercial services	0209 Legal and financial	170,768
	0214 Research and design	26,052
	0215 Transport, storage and delivery	63,677
03 Attractions	0320 Tourism	14,971
04 Sport and entertainment		131,508
05 Education and health	05280780 Accident and emergency hospitals	1,860
	05280370 Hospices	
	05280371 Hospitals	
	0531 Primary, secondary and tertiary education	30,393
06 Public infrastructure	0633 Central and local Government	19,718
07 Manufacturing and production	0738 Extractive industries	43,112
	0739 Farming	69,763
	07410531 Business parks and industrial estates	8,373
	0742 Industrial products	91,588
09 Retail	0946 Clothing and accessories	53,132
	0947 Food, drink and multi item retail	98,560
	0949 Motoring	21,510
10 Transport	1057 Public transport, stations and infrastructure	6,058
	Total	1,053,645

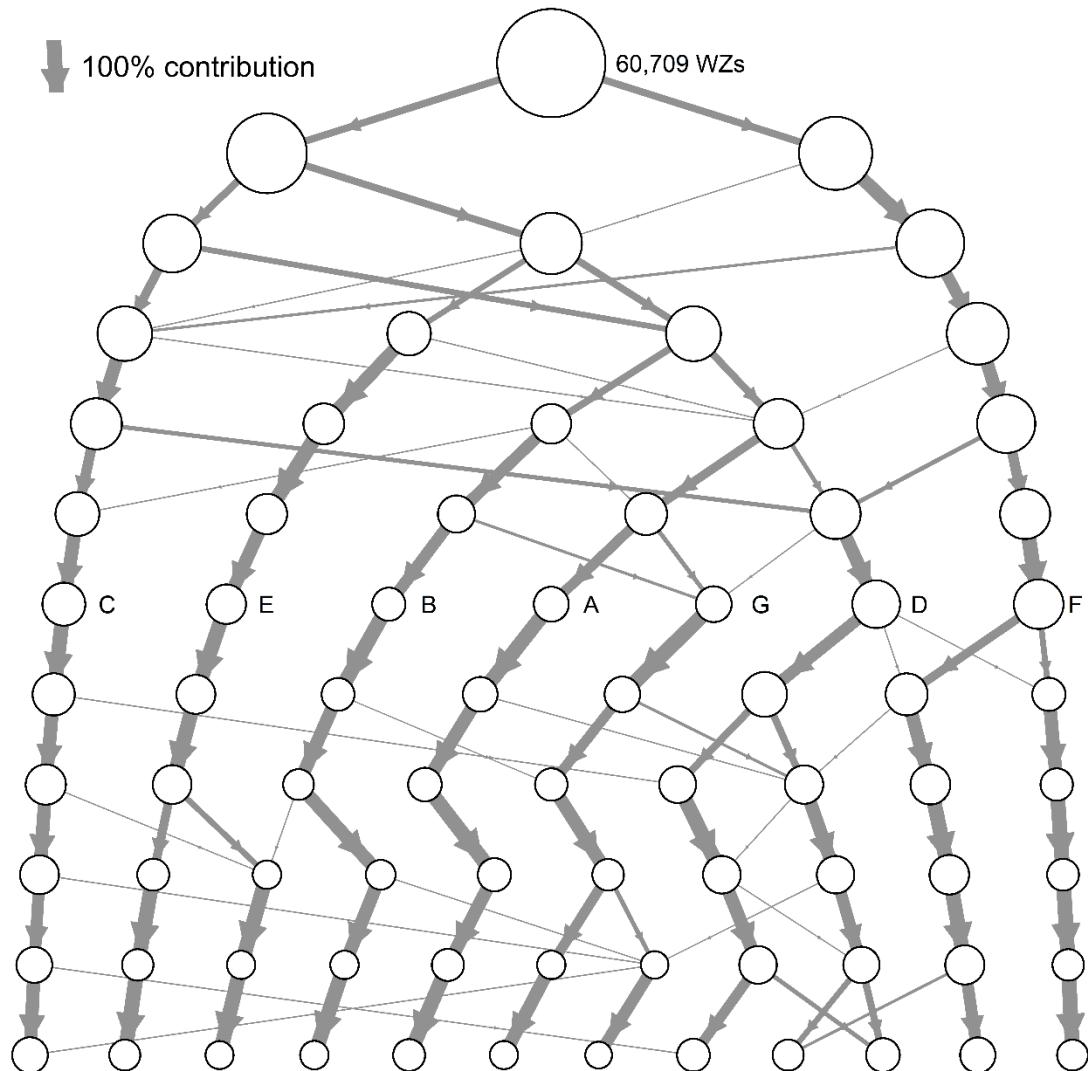


Fig. 1 Clustergram showing potential clusters in one- to twelve-cluster solutions for the Supergroup level of the COWZ-UK hierarchy. Rows represent solutions. Circles represent clusters. Area of circle proportional to number of WZs in the cluster. Thickness of arrows proportional to percentage of WZs contributed from relevant cluster in solution above. Labels for seven-cluster solution indicate named Supergroups in final COWZ-UK classification: A = Retail; B = City and business parks; C = Metro suburbs; D = Suburban services; E = Manufacturing and distribution; F = Rural; G = Servants of society

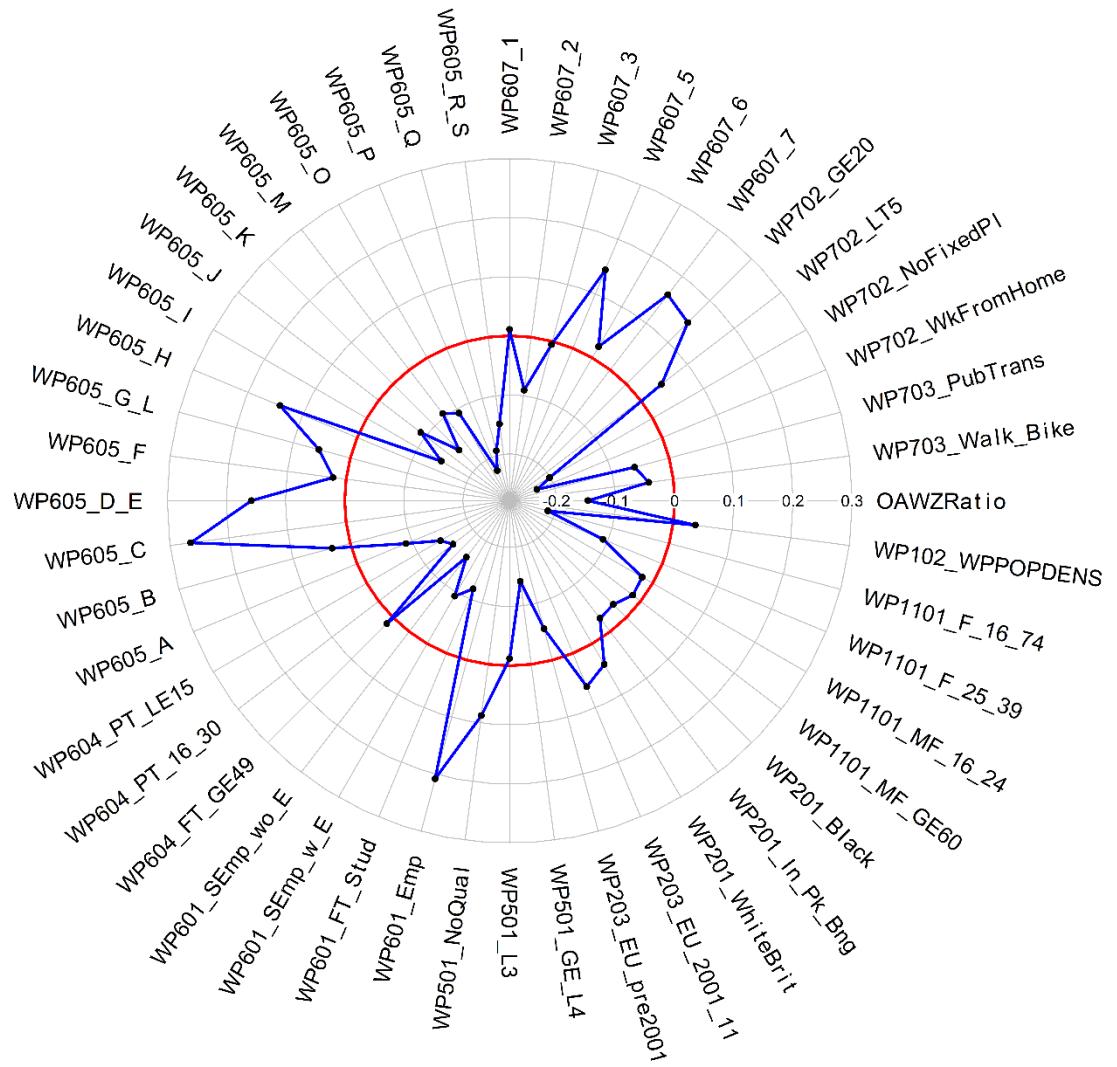


Fig. 2 Illustrative radial plot: Supergroup E - Manufacturing and distribution. Values of cluster centres for final 48 variables plotted on radial axes and joined to form (blue) irregular polygon. UK mean values shown as (red) circle. Points falling outside the circle indicate higher than average values for that variable and vice versa. Codes for variables labelled as per Table 1, but reordered as per final ONS published products

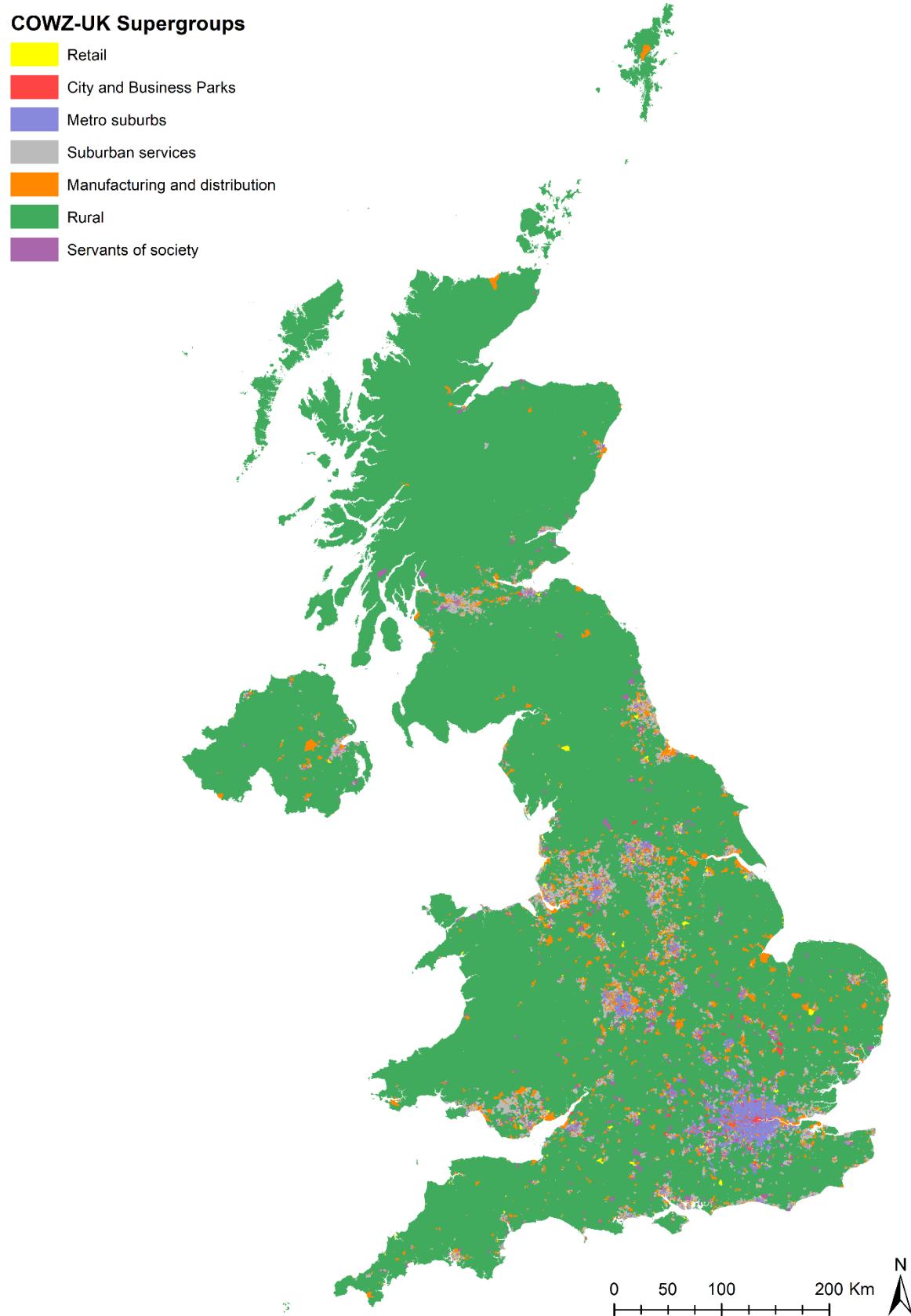


Fig. 3 COWZ-UK Supergroups for the UK

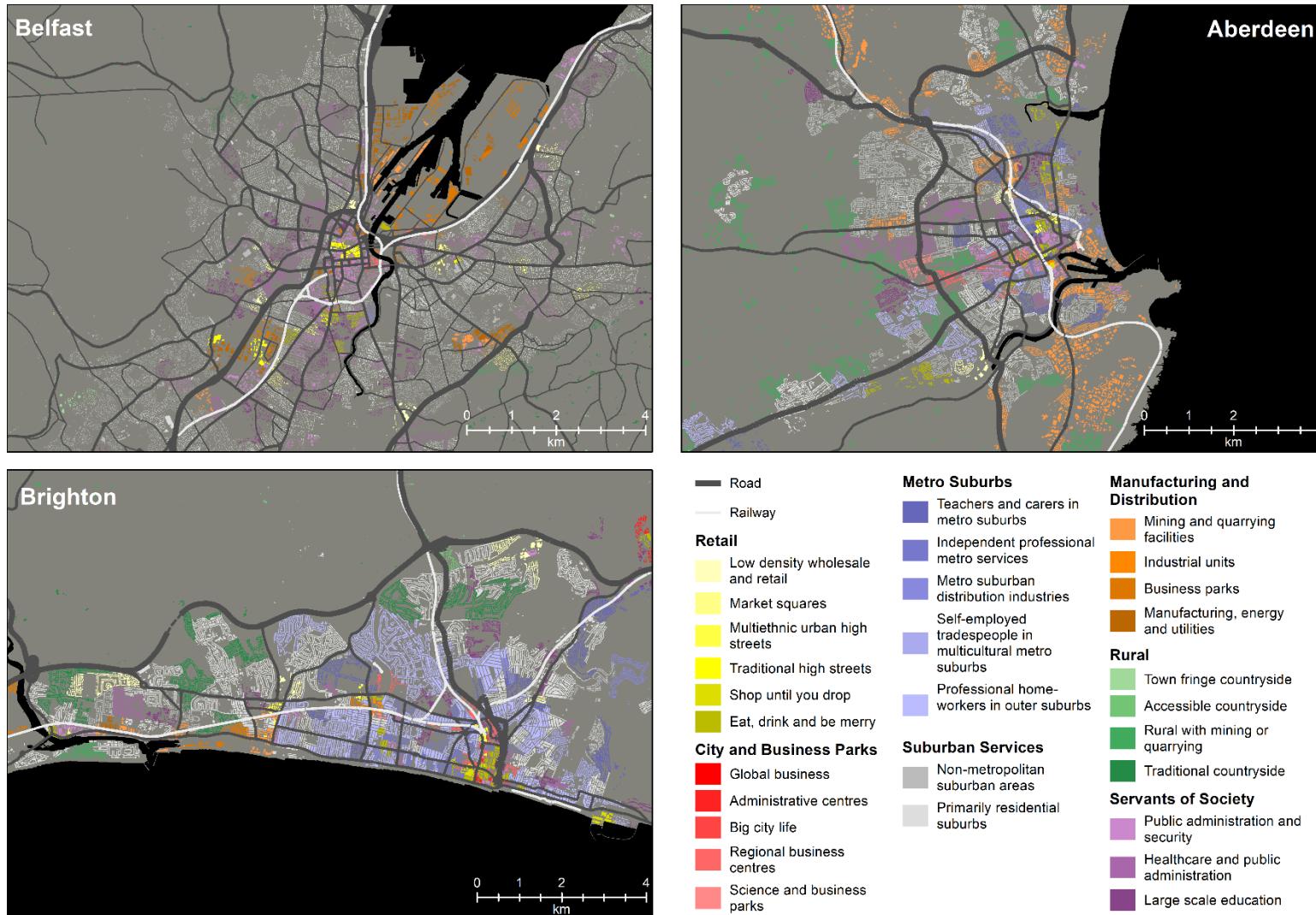


Fig. 4 COWZ-UK Groups for three coastal cities in the UK (a) Belfast, (b) Aberdeen, (c) Brighton, masked by a buildings layer to aid interpretation

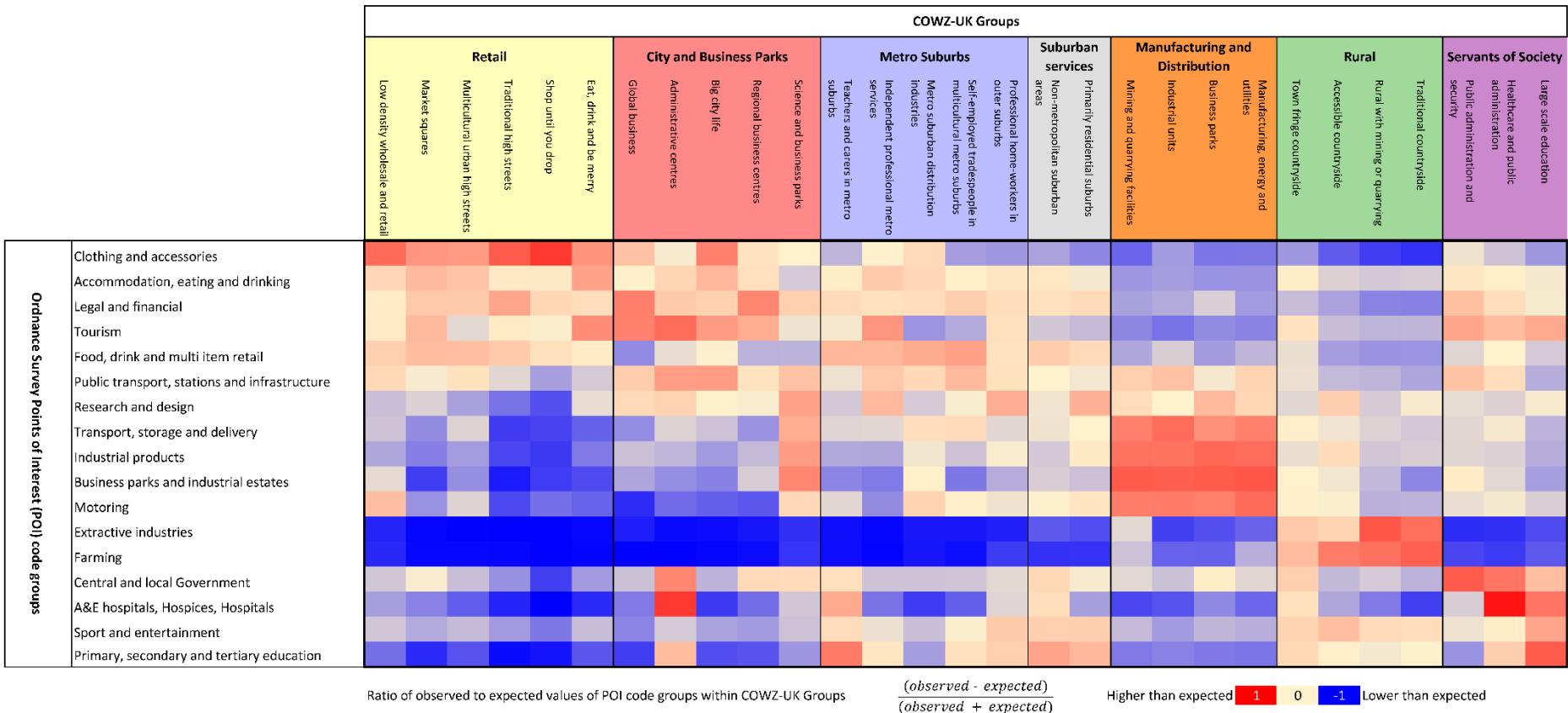


Fig. 5 Heatmap of Points of Interest (POI) code groups cross-tabulated against COWZ-UK Group