Geography and Environment

UNIVERSITY OF Southampton

A Classification of Workplace Zones for England and Wales (COWZ-EW)

User Guide

Samantha Cockings, David Martin, Andrew Harfoot

Geography and Environment, University of Southampton

August 2015

Contents

Introduction

This user guide provides an overview of the Classification of Workplace Zones for England and Wales (COWZ-EW). It describes the aims, purpose and scope of the classification, the data and methods employed to create it, and finally provides details of the outputs available.

COWZ-EW is a geodemographic classification of Workplace Zones (WZs) based on data from the 2011 Census of England and Wales. It classifies WZs according to their similarity in terms of a range of census variables. COWZ-EW was produced by the University of Southampton in collaboration with the Office for National Statistics (ONS) and is freely and openly available under an Open Government Licence from http://cowz.geodata.soton.ac.uk/. This version of COWZ-EW is not an official ONS product. In response to strong user interest, it is being released ahead of production of a UK-wide Classification of Workplace Zones (COWZ) so that users can explore and use the classification, and provide feedback to the research team.

Background, aims and scope

The UK Censuses collect data about workers and workplaces. In the 2001 Census, these 'workplace data' were released for Output Areas (OAs). OAs were designed to represent the geographical distributions of residents and residences, but these distributions are very different to those of workplaces and workers. ONS was therefore only able to publish four univariate workplace tables at the OA level in 2001 and the workplace population in these OAs varied from 0 to 80,145 workers. Following the 2011 Census, ONS employed automated zone design techniques developed by the University of Southampton to create a set of geographical areas optimised for the release of workplace data for England and Wales (Martin et al, 2013). These WZs were produced by splitting, merging or retaining the 2011 OAs, which had themselves been maintained using similar methods (Cockings et al, 2011). Boundaries for these 53,578 WZs were released in January 2013 (ONS, 2014a) and corresponding aggregate data were released in May 2014 (ONS, 2014b; http://www.ons.gov.uk/ons/rel/census/2011-census/workplace-population-statistics-for-workplace-zones-and-middle-layer-super-output-areas--msoas--in-england-and-wales/index.html). 21 univariate tables were published at WZ level for the 2011 Census (compared to four in 2001) and workplace population size was much more uniform across WZs than OAs, (range: 101 to 11,985; mean 493), reflecting the benefits of this bespoke zone design. Greater detail was provided in areas with high numbers of workers/workplaces such as city centres, retail and business parks, while the merging of OAs containing low numbers of workers, such as in suburban or rural areas, aided robust analysis. This enhanced range of data and new set of geographical areas provides real opportunities to gain new insights into the characteristics of workers and workplaces at the small area level.

One way to synthesise and extract useful information from such a rich set of data is to produce a geodemographic classification. A number of classifications of residential neighbourhoods already exist, including the 2001 and 2011 Output Area Classifications (OAC), but none of these has an explicit focus on the characteristics of workers and workplaces. The purpose of COWZ-EW is therefore to categorise groups of WZs which share similar characteristics in terms of their workers and workplaces. Its intended users are researchers, practitioners or policy makers who need to understand what makes certain areas similar or different to others in terms of their worker and workplace characteristics.

Data and definitions

COWZ-EW is based entirely on 2011 Census data. All of the input variables and the final classification are therefore open data and freely available to all users. The geographical coverage of COWZ-EW is the entire land area of England and Wales. Its population base is as per the 2011 Census definition of the workplace population: 'All usual residents aged 16 to 74 in employment in the area the week before the census'. This includes people who were working in any paid work (including casual or temporary work) within the last week. It encompasses: 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. It does not include: those usually resident in England and Wales but working in Scotland, Northern Ireland, outside the UK or on offshore installations and therefore falling outside of the land area to be classified; those with a place of work in England and Wales but who are not usually resident in England and Wales; short-term residents; or full-time students who are not working. Respondents answer questions related to their main job (or last main job) i.e. the one in which they usually work (or worked) the most hours; any secondary employment is therefore not considered. As part of the 2011 Census processing of workplace data, people who work mainly at or from home, or who do not have a fixed place of work, are georeferenced to their area of usual residence, while workers who report to a depot are asked to record the address of the depot and are georeferenced to this.

In the context of the census, a 'workplace' can be defined as a place of work recorded by a worker on their census form. It is important to note that workplaces themselves are not explicitly surveyed in the census and that a workplace in the census may be different to other entities (such as businesses, enterprises or companies) which are recorded in other datasets such as the Inter-Departmental Business Register (IDBR).

Methods

The overarching methodological approach adopted in the creation of COWZ-EW is deliberately similar to that used in OAC. This provides consistency between the classifications and capitalises on insights gained from previous detailed analysis of small area UK census data. Rigorous assessment of the various design options was undertaken at each stage of the process. An overview of the methodology is provided below.

First, the key domains related to workers and workplaces were identified and a set of candidate variables representing each of these domains was selected. Exploratory analysis was undertaken in order to understand the statistical and geographical distributions of the variables and to evaluate the strength of correlation between them. Various methods of transformation were explored prior to the correlation analysis in order to aid inference. Based on the results of this exploratory analysis, the set of variables was then reduced to the final set employed in the classification. The variables were standardised to ensure that all contributed equally to the clustering process. A *k*-means clustering algorithm was then employed to produce specified numbers of clusters for the top level of the hierarchy. These clusters were evaluated using a range of objective and subjective statistical and graphical methods. The *k*-means algorithm was then reapplied to the top-level clusters to subdivide them, thus creating the next level of the hierarchy. Again, various values of *k* were trialled. This

hierarchical subdivision continued until meaningful results were no longer achieved. Neither the number of clusters at each level nor the number of levels of the hierarchy was predetermined. Once the number of levels of the hierarchy and the most suitable set of clusters for each level were determined, the clusters were profiled and named to enable users to interpret and use them more readily. The rest of this section describes the implementation of these methods in more detail.

Definition of domains

The 2001 and 2011 versions of OAC both identified five domains to represent the key characteristics of residential areas: demographic structure, household composition, housing, socio-economic group and employment. Equivalent pre-defined domains did not exist for areas containing workplaces, but four domains were identified with reference to literature and consultation with ONS and users: composition of the workplace population, composition of the built environment, socio-economic characteristics of the workplace population and employment characteristics of the workplace population. Note that there is no direct workplace equivalent for household composition, because characteristics of each workplace, such as size or type of employer, were not measured by the 2011 Census. The composition of the built environment domain is the equivalent of the housing domain in OAC but differs in that it is explicitly designed to capture the workplace/residential mix and workplace population density of a WZ. Aspects of transportation and travel to work are included within the socio-economic characteristics of workplace population domain (as per OAC) rather than introducing a separate, narrowly focused, transportation domain.

Selection and refinement of variables

The 21 WZ tables from the 2011 Census were downloaded from NOMIS (https://www.nomisweb.co.uk/census/2011/workplace_population). 13 of these tables (shown in Table 1) were selected for preliminary analysis because they contained variables which most closely represented the domains identified above. The other eight tables were excluded either because they contained variables which were of analytical interest in their own right but not within the scope of the classification, or because they were highly correlated with the variables in other tables.

This gave a pool of 501 candidate variables for the exploratory analysis. Three further variables (number of workplace postcodes (obtained from ONS specifically for this work), density of workplace postcodes (per hectare), and ratio of the number of OAs to WZs[1]) were included specifically to measure the composition of built environment as there were no variables in the published WZ tables which achieved this. This brought the total number of candidate variables to 504.

[1] Derived from open lookup tables from the ONS Open Geography portal (https://geoportal.statistics.gov.uk/geoportal/catalog/main/home.page) and intended as a proxy for the workplace/residential mix of a WZ. Values greater than one (where OAs have been merged to form a WZ) tend to indicate low numbers of workplaces compared to residential properties, such as in rural or suburban areas; values less than one (OA split to produce more than one WZ) indicate high numbers of workplaces relative to residential properties, such as in city centres, retail or industrial areas. Areas with a mixed composition, such as neighbourhood high streets, where the original OA has been retained to form a WZ in its own right will have a ratio of exactly 1. The relatively few 'complex' situations where WZs have been formed by parts of an OA will have fractional values.

Table 1 Workplace Zone tables included and excluded from the preliminary analysis

Table code	Table name
Included:	
WP102EW	Population density
WP1101EW	Sex by single year of age
WP201EW	Ethnic group
WP203EW	Country of birth (detailed)
WP501EW	Highest level of qualification
WP601EW	Employment status
WP604EW	Hours worked
WP605EW	Industry
WP607EW	NS-SeC
WP613EW	Approximated social grade
WP702EW	Distance travelled to work
WP703EW	Method of travel to work (2001 specification)
WP204EW	Main language (detailed)
Excluded:	
WP206WA	Welsh language skills
WP207WA	Welsh language skills (detailed)
WP210EW	Religion
WP212EW	Passports held (detailed)
WP302EW	General health
WP403EW	Tenure - People
WP606EW	Occupation
WP803EW	Length of residence in the UK

All variables besides workplace population density, workplace postcode density and the ratio of OAs to WZs (all kept as ratios) were then converted from counts to percentages. The population base for all of these conversions was the workplace population i.e. all usual residents aged 16 to 74 in employment in the area the week before the census, other than for WP613EW (Approximated social grade) which used all usual residents aged 16 to 64 in employment in the area the week before the census.

Geodemographic classification is predominantly data-driven, but the analyst designing the classification must always make key decisions concerning the selection and refinement of variables. In common with Vickers and Rees (2007), objective and subjective criteria were employed in the design of COWZ-EW, including: the exclusion of variables representing small percentages of the population; the reduction of redundancy by exclusion of one of any pair of very highly correlated variables; the removal of variables with inappropriate distributions e.g. highly skewed; the calculation of composite variables e.g. when variables were strongly related but represented only small percentages of the population; the avoidance of variables for which there were significant data quality concerns (ONS, 2015a); and the exclusion of variables with relatively uniform geographical distributions which added little to the classification. A range of analytical methods was employed to assess the statistical and geographical distributions of the 504 candidate variables and their inter-relationships, including summary statistics (mean, median, range, standard deviation), histograms, Q-Q plots (to assess normality) and maps.

Some variables were excluded at this early stage because they represented a very small percentage of the population, for example many of the country of birth and ethnic group categories. While contributing to the distinct characteristics of a small number of areas, these variables have values of zero across most WZs and therefore offer very little differentiation between areas. Only the top level of the NS-SeC hierarchy was retained for the same reason. In cases where there were small numbers but good evidence (from the literature or from initial analysis of the data) that producing a composite variable may be more appropriate, such as for age or ethnic groups, these variables were composited and then explored further. Following this initial exploration of the data, a subset of 63 variables was retained for further investigation.

The 63 variables were then re-analysed. In addition to the previous analytical methods, a correlation matrix (using Pearson's Product Moment Correlation coefficient) was employed to assess the degree of correlation between each pair of variables. This required normalisation of the data prior to its calculation. Three transformation methods were evaluated for this purpose: log, Box-Cox and inverse hyperbolic sine (IHS). Theoretically, the log transform deals with extreme outliers well, the Box-Cox method should deal with a range of different distributions better than the log, and the IHS is particularly suited to distributions with a large number of zero values (ONS, 2015b). For COWZ-EW, the Box-Cox transformation was found to perform best, probably because it is more adaptive to the specific distribution of each variable being transformed. Following this detailed exploratory analysis, the subset of 63 variables was further reduced down to 48. This final set of variables formed the inputs to the clustering analysis and is shown in Table 2 (grouped according to domain).

Table 2 Final set of 48 variables employed in COWZ-EW, grouped by domain

Variable code	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

Standardisation of variables

It was important to ensure that all variables were measured on the same scale and magnitude otherwise variables with a greater range in their potential values may carry a disproportionate weight in the classification. The 48 variables were therefore standardised using the range standardisation technique shown in Equation 1 (after Vickers and Rees, 2007):

$$R_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \qquad \text{Equation 1}$$

where R_i is the standardised value of the variable which has a value of x_i in area i and where x_{max} is the maximum value taken by x in the data and x_{min} the minimum. This technique produces output values in the range 0-1.

Workplace population density had a few very extreme outliers (artefacts arising from the WZ design process where large numbers of workers work in workplaces which are all stacked on the same geographical location, such as large office tower blocks). These outliers prevented the range standardisation technique from working effectively. The standardisation process was therefore undertaken with the top 0.01% of values excluded. These outlier values were assigned a value of 1 and added back to the distribution after standardisation of the rest of the values.

Cluster analysis

A hierarchical *k*-means clustering method was employed to group the WZs into clusters based on their similarity in terms of the 48 variables. This was implemented in R using the *kmeans* function and the default Hartigan-Wong algorithm. The squared Euclidean distance was used to evaluate the degree of similarity within and between clusters, as shown in Equation 2 (after Vickers and Rees, 2007):

$$E_c = \sum_{i=1}^{n_c} \sum_{j=1}^{m} \left(z_{ij} - \bar{z}_{cj}\right)^2 \qquad \text{Equation 2}$$

where m is the number of variables, n_c is the number of members within a cluster and c is the index of the cluster, z_{ij} is the value for variable j of member i, and z_{cj} is the mean value for the cluster. 10,000 random starts were employed.

Solutions were generated for 2 to 12 clusters for the top level of the COWZ-EW hierarchy. These solutions were evaluated and solutions 5 to 8 were then further subdivided into 2 to 6 clusters. All of the potential solutions were systematically evaluated using a combination of: prior assumptions based on theory and practice, statistical and graphical techniques (such as compactness of cluster solution, homogeneity of cluster size, and stability and robustness throughout the hierarchy), together with mapping to confirm whether the outputs made sense on the ground. Analysis took place at various geographical scales to ensure a thorough understanding of the patterns was obtained.

Outputs: COWZ-EW

COWZ-EW is a two-tiered hierarchical classification. Subdivision below this level leads to excessively small cluster size membership and fragmentation of the clusters. To ensure comparability with OAC and other classifications, the levels of the COWZ-EW hierarchy are termed 'Supergroups' (top level), and 'Groups' (second level).

COWZ-EW consists of 7 Supergroups and 29 Groups, as shown in Tables 3 and 4. These lookups can be downloaded from http://cowz.geodata.soton.ac.uk/ in .csv or ESRI Shapefile format. Annexes A and B (available from the same website) provide a full set of profiles for each of the Supergroups and Groups, including pen portraits, radial plots, example locations and images.

Codes and names have been allocated to each of the Supergroups and Groups by the University of Southampton research team, based on detailed analysis of the data and results. Users should note that the codes and names employed here are specific to the COWZ-EW version of the classification. As it is anticipated that further versions of COWZ will be released in the future, such as a UK-wide version (COWZ-UK), users are advised to prefix any labelling in maps, reports and so on with COWZ-EW. This will avoid potential confusion with any subsequently-released versions which may contain the same or different codes and names depending on the characteristics of the clusters in that version.

Table 3 Codes and names for COWZ-EW Supergroups

Supergroup	Supergroup name
1	Retail
2	Top jobs
3	Metro suburbs
4	Suburban services
5	Manufacturing and distribution
6	Rural
7	Servants of society

Table 4 Codes and names for COWZ-EW Groups

Group	Supergroup name	Group name
1a	Retail	Low density retail and wholesale
1b	Retail	Market squares
1c	Retail	Multicultural urban high streets
1d	Retail	Traditional high streets
1e	Retail	Shop until you drop
1f	Retail	Eat, drink, shop and be merry
2a	Top jobs	Global business
2b	Top jobs	Administrative centres
2c	Top jobs	Big city life
2d	Top jobs	Regional business centres
2e	Top jobs	Science and business parks
3a	Metro suburbs	Metro suburban distribution
3b	Metro suburbs	Cosmopolitan metro suburban mix
3c	Metro suburbs	Independent professional metro services
3d	Metro suburbs	Suburban metro infrastructure
4a	Suburban services	Non-metropolitan suburban areas
4b	Suburban services	Primarily residential suburbs
5a	Manufacturing and distribution	Mining and quarrying facilities
5b	Manufacturing and distribution	Industrial units
5c	Manufacturing and distribution	Business parks
5d	Manufacturing and distribution	Manufacturing, energy and utilities
6a	Rural	Rural with core services
6b	Rural	Rural with non-local workers
6c	Rural	Rural with mining or quarrying
6d	Rural	Traditional countryside
7a	Servants of society	Large scale education
7b	Servants of society	Public administration
7c	Servants of society	Major hospitals
7d	Servants of society	Highly qualified workforces and professional services

Acknowledgements

This project was funded by the Quality Improvement Fund (QIF). The authors are grateful to ONS staff for their assistance in the project, particularly Ian Coady, Andy Tait, Bruce Mitchell and Mark Baines. They also thank Chris Gale and Alex Singleton for sharing their OAC/geodemographic classification code and members of the Workplace Zone User Group for providing valuable input during the early stages of the project. COWZ-EW is based on National Statistics data © Crown copyright and database right 2015 and Ordnance Survey data © Crown copyright and database right 2015. The views expressed in this report are the authors' own.

References

Cockings S, Harfoot D, 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.

Martin D, Cockings S, Harfoot A (2013) Development of a Geographical Framework for Census Workplace Data, *Journal of the Royal Statistical Society, Series A*, 176(2), 585-602

ONS (2014a) *Workplace Zones: A new geography for workplace statistics*. ONS. Available from http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/workplace-zones--wzs-/index.html [Last accessed 26/08/15]

ONS (2014b) *Workplace population analysis, 2011 Census*. ONS. Available from http://www.ons.gov.uk/ons/rel/census/2011-census/workplace-population-statistics-for-workplace-zones-and-middle-layer-super-output-areas--msoas--in-england-and-wales/workplace-population-analysis--2011-census.html [Last accessed 26/08/15]

ONS (2015a) *2011 Census: General report for England and Wales*. London: Office for National Statistics. Available from http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-did-we-do-in-2011-/2011-census-general-report/index.html [Last accessed 26/08/15]

ONS (2015b) Methodology note for the 2011 Area Classification for Output Areas. ONS. Available from 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 [Last accessed 26/08/15]

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