Applied Geography 109 (2019) 102038

Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Data infrastructure requirements for new geodemographic classifications: The example of London's workplace zones

Alex D. Singleton[a,*], Paul A. Longley[b]

[a] University for Liverpool, Department of Geography and Planning, UK
[b] University College London, Department of Geography, UK

ARTICLE INFO

Keywords:
Geodemographics
Open data
Consumer data
Workplace zone
London

ABSTRACT

In recent years a mix of Open Data and commercial sources have been used to build geodemographic classifications of neighbourhoods. In this paper we argue that geodemographics are coming to embody new thinking about the emergent mixed Big Data economy. This has implications for openness and full scientific reproducibility of classifications, as well as the engagement of stakeholders in the process of building classifications. We propose and implement an operational framework for blending open and other data sources that can stimulate development of classifications that are more timely and data rich yet sufficiently open to peer scrutiny. We illustrate these ideas and challenges by describing the creation and content of the London Workplace Zone Classification.

1. Introduction

Geodemographics are small area classifications of neighbourhood conditions, conventionally used to depict the variegated residential geographies of towns and cities. Although the approach has its roots in the primary data collection of urban sociologists Park and Burgess in 1920s Chicago (Harris, Sleight, & Webber, 2005; Webber & Burrows, 2018), procedures of ascribing neighbourhoods to social, economic and demographic types came to rely upon secondary data from population censuses until the 1980s (Timms, 1971). With the advent of applications in commerce (Harris et al., 2005) and public service delivery (Longley, 2005), census data have been supplemented and partially replaced by commercial and open sources that offered greater frequency of update and depth (particularly in ascertaining income and spending preferences). Over the last ten years, improved access to censuses and the advent of the Open Data movement has led to the addition of open geodemographic classifications that present greater transparency of data and methods (Gale et al., 2016; Vickers & Rees, 2007). A final innovation has been the re-configuration and re-use of census data to provide small area classification of activities other than night-time residence, specifically workplaces (Martin, Cockings, & Harfoot, 2013) or their extension to explore varying temporal geographies (Martin et al., 2018; Singleton, Pavlis, & Longley, 2016).

Geodemographic classification has endured because of its value as an applied tool for summarising the structure and character of neighbourhoods. General purpose classifications developed using a fairly standard menu of socioeconomic and demographic variables attract wide use, whether as a means of better understanding consumption of public or of private goods and services (Grubesic, Miller, & Murray, 2014; Singleton & Longley, 2009). Full implementation of General Data Protection Regulation in Europe arguably lends the approach renewed vigour, given tightened disclosure responsibilities when identifiable individuals are profiled and targeted. Public sector applications also remain important because of the collective ways in which public services are consumed. However, the advent of many new consumer data sources is both broadening the potential range of neighbourhood activities that may be characterised, and increasing the potential depth and frequency with which such activities may be represented (Longley, Cheshire, & Singleton, 2018). Realising this potential is not, however, straightforward, as ownership and control of new data sources does not lie in the public domain. The motivation for this paper is to describe the ways in which the data landscape is changing, and to assess the implications for the creation of geodemographic classifications that are timely, data rich and sufficiently open to scrutiny by the research community. We illustrate these new developments and practices through the development of the hybrid geodemographic London Workplace Zone Classification. This is used to illustrate how, post the full implementation of EU General Data Protection Regulation, diverse data sources may be brought together in a secure setting without compromising stakeholder engagement and maintaining transparency of methodology.

* Corresponding author.
E-mail addresses: alex.singleton@liverpool.ac.uk (A.D. Singleton), p.longley@ucl.ac.uk (P.A. Longley).

https://doi.org/10.1016/j.apgeog.2019.102038
Received 14 October 2018; Received in revised form 24 March 2019; Accepted 15 June 2019
Available online 02 July 2019

2. Background

There is no open and transparent marketplace for the new forms and sources of Big Data that trace a far greater range of human actions than at any point in human history. New Big Data sources are assembled by customer-facing organisations responsible for services such as domestic energy supply, travel or general retail sales and are usually the property of the of the organisations that collected them. Other more conventional statistical sources such as market research data or surveys can belong to any of a range of organisations, and may also lie outside the public domain. Together, the availability of new data sources has the potential not only to rejuvenate the creation of geodemographic classifications, but more broadly may transform the practice of social science. However, for this to happen, vexing issues of data ownership, control and access must be addressed.

As with software, there is strong academic advocacy for data that are 'open' – that is, freely available to all with minimally restrictive licencing requirements (Singleton, Spielman, & Brunsdon, 2016). The specification, estimation and testing of 'black box' commercial geodemographic systems has provided a recurring focus of concern where they have been used to benchmark or validate research findings (Ashby & Longley, 2005). The development and dissemination of geodemographic classifications based upon principles of open Geographic Information Science (Singleton, Spielman, et al., 2016) has gone some way towards responding to these issues of transparency and scientific reproducibility, notably the 2001 and 2011 UK Output Area Classifications (OAC: Vickers & Rees, 2007; Gale et al., 2016). Yet despite advantages of transparency and reproducibility, there is some comparative evidence (Brunsdon et al., 2011) that the discriminatory power of these conventional census based systems do not match that of commercial rivals that include a wider and more contemporary range of data sources.

The best research requires the best data, and, over the last decade, the Open Data movement has gone some way towards creating a more level playing field for the creation of geodemographic classifications that utilise new data sources and enrich neighbourhood classifications. The benefits of the Open Data platforms that have been developed by government in recent years (Kitchin, 2014), and the wider recognition of the value that accrues to society when Open Data are made available unencumbered by restrictive pricing and access issues, does not take place without cost (Johnson et al., 2017). The costs of Open Data creation and maintenance are essentially ultimately borne by the taxpayer rather than specific individuals or classes of users. This is not the case where data are created and maintained by the private sector, where the immediate instincts of economic competition may override longer term or philanthropic motivations of contributing to a competitive, more socially inclusive economy.

Open data nevertheless account for a rapidly diminishing share of all data assembled about individuals today. This is not principally because of changed social priorities or government policies – not withstanding some instances of the withdrawal of Open Data[1] or replacement of formerly open licences with more restrictive variants.[2] Rather, this is because vastly increased amounts of data are collected about citizens, year on year. Longley et al. (2018) describe many of these sources as consumer Big Data, defined as arising as a by-product of the acquisition of goods and services through business-to-consumer transactions. Examples of consumer data include traces of social media usage, evidence of customer transactions through retailers, real time smart meter readings of domestic energy consumption, and GPS traces of mobile phone use. Such data could in theory form many of the staple inputs of more detailed, pertinent and up-to-date geodemographic classifications. Data accrual today is on a vast scale and is fundamental to the operations of the behemoths on the Internet Age – Apple, Amazon, Alphabet, Facebook, Google and Microsoft – yet the inaccessibility of the enormous data silos of these and other corporations is a recurring focus of public and government concern, particularly when data breaches or inappropriate use cases periodically come to light.

If today's data are indeed the world's most valuable resource (The Economist, 2017) the concentration of ownership and control in the silos of large corporations is in some respects redolent of Galbraith's (1958) discussions of the contrast between private opulence and the relative squalor of public infrastructure in advanced societies, albeit that the world has become immeasurably more data rich in recent years. An additional issue for social scientists is that the vastly enriched depth of content of today's consumer Big Data are not entirely matched by the breadth of their coverage – for even though Internet behemoths may create and sustain near monopolies of supply, none has achieved the universality of population coverage that is sought by censuses and other government surveys.

In this paper we utilise data that have been re-purposed by the Consumer Data Research Centre (CDRC) for the social good through nascent notions of data philanthropy (Kirkpatrick, 2011). Such partnerships with more than 30 private sector data providers have been nurtured alongside the development and implementation of new access and research governance methods. From a purist perspective, the hybrid procedures that this engenders mean that the use of consumer data sources in geodemographic classification is not strictly "open" – but we illustrate that it nevertheless facilitates the creation of classifications that use rich new data resources whilst remaining sufficiently transparent and open to scrutiny.

Our approach utilises a three-tier service that facilitates access to consumer and related datasets and assemblages that have been donated or acquired from private sector organisations. In this three-tier data service, *Public* data have undergone documented pre-processing and are not disclosive or sensitive (commercial or individual) in any way, and often comprise spatially aggregate records or conflated modelled outputs. *Safeguarded* data require users to register and successfully navigate research access protocols, and concern data that will usually have some sensitivity but not have potential to be personally disclosive. Finally, *Controlled* data are the most sensitive, usually comprising individual level records and transaction histories, although most often with personal attributions removed. These are also governed by access protocols like the Safeguarded tier data, however are additionally only available through on-site access at three dedicated locations. Furthermore, outputs from the *Controlled* setting are also not immediately available and are checked for possible disclosure issues by a trained Data Scientist prior to release.

It is within the controlled setting that we developed a hybrid framework for creating a geodemographic classification of workplace zones. Our motivation was to devise a classification that was built from the best possible range of open and restricted data sources that can be fully documented and made available to any interested user. This entails a departure from the goal of truly open geodemographics, in that assent from all data providers must be gained through a gatekeeper service. Moreover, full reproducibility requires navigation of the same access protocols (albeit not unreasonably withheld for bona fide research), and the use of secure facilities (if *Controlled* tier data are used).

Fig. 1 presents a schematic diagram of the hybrid geodemographic system architecture. Data conforming to either *Public* or *Safeguarded* specifications can be ingested into the secure data laboratory and enables linkage with other *Controlled* tier data as required. All data can then be integrated, and the process of model building can begin. This will typically involve the evaluation of a set of candidate variables,

[1] The US open.whitehouse.gov website and data was removed in February 2017 and now redirects to www.whitehouse.gov/disclosures/.

[2] An example within the UK includes the Valuation Office Agency whose data concerning the ratable values of business properties was previously disseminated with an Open Government License and later replaced with a new license that has far more restrictive conditions.

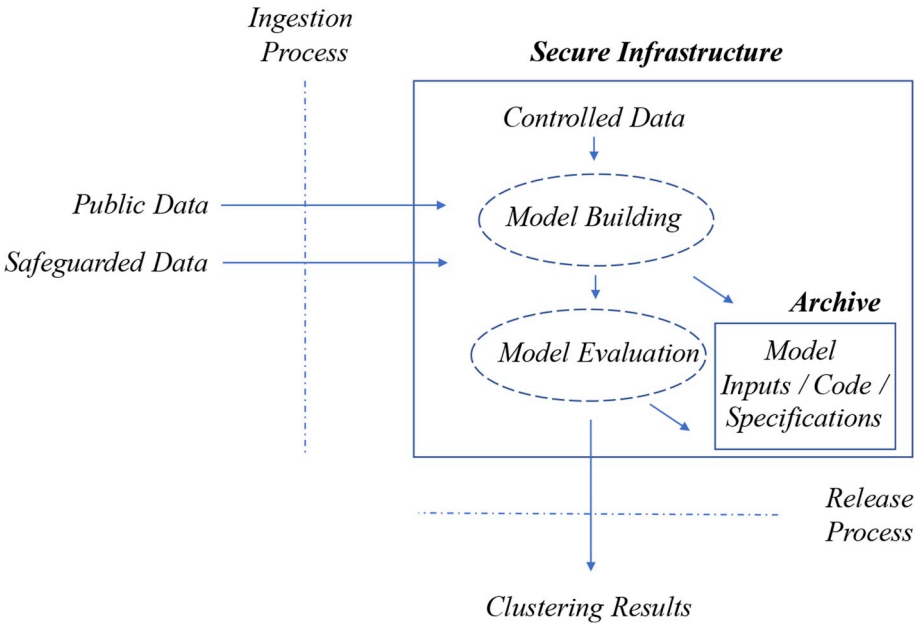

Fig. 1. A Summary of the hybrid geodemographic framework.

selected and analysed using essentially the same procedures as conventional open geodemographics. The hybrid geodemographic is therefore specified, estimated and tested within the secure environment; yet non-disclosive intermediate outputs and software may also be exported from the controlled setting for subsequent refinement or use by other researchers. Any outputs from the process are governed by the data export procedures of the secure lab, and in this context would typically consist of the cluster assignments and descriptive profiles of the groups, alongside code used in the build process. Copies of the code and all input data can remain within the controlled setting as an archive should anybody wish to reproduce these results. The only constraint upon this is the additional steps of having to register a project with the data custodian (in this case the CDRC) to enable access to the secure lab in order to complete this work. Thus, although not fully open, a hybrid approach does enable the creation of geodemographics with wider data inputs, while maintaining the essence of reproducibility that has been championed by open geodemographics.

3. A hybrid geodemographic: the London Workplace Zone Classification

3.1. The requirement

This section sets out how the schema set out in Fig. 1 was implemented when creating a geodemographic classification of workplaces in London. Workplace zone classifications have emerged in recent years as a novel re-use of census data to provide information for economic planning of local diversification or regeneration, alongside evidence for transport planning of improved accessibility across transport networks. The core methodology entails reassignment of census data related to employment to the work destination (Martin et al., 2013). There are uncertainties inherent in this assignment – for example, many individuals do not have a single regular place of work – but the result is useful for planning purposes as it provides a guide to the functional characteristics of areas during the working day.

The requirement for a London-specific workplace zone classification arises in a significant part from the functional differences between the world city of London and the rest of the United Kingdom. The notion that London's labour market is structured in a fundamentally different way to the rest of the UK has echoes in both open and closed geodemographic classifications of residence. For example, the open 2011 Output Area Classification of residential areas (Gale et al., 2016) spawned a London specific variant using essentially the same open methodology (Singleton & Longley, 2015) in order to recognise a number of distinctive characteristics of the Capital, notably its intricate and variegated residential structure. This was possible because both the software and the data were open. With respect to Workplace Zones, the majority of Greater London is assigned to just two of the seven Super Groups in the UK national Classification of Workplace Zones (COWZ) (see Fig. 2). Additionally, within areas that have a very significant presence of retail such as central London (see the cut out map in Fig. 2), the Retail Supergroup is almost entirely absent. Such issues motivated the decision to devise a new classification that better represented the diverse functions of London's workplaces.

The core motivations for creating a specific workplace zone classification for London were to incorporate a wider range of up-to-date data that bore testimony to London's unique employment structure. With specific end uses in mind, we also convened a stakeholder group comprising local authority end users of the classification as well as representatives of the Greater London Authority. This made us aware of additional requirements viz: (a) updating 2011 Census data to more accurately reflect London's dynamic economy; (b) incorporating broader occupational data consistent with the breadth of economic activities taking place in London; (c) incorporating indicators of activities arising from employment, since these might have important implications for planning; and (d) devising a readily intelligible classification that could be used to understand the interactions between different employment sectors, such as retailing and head or back office functions.

The schema set out in Fig. 1 was implemented in dialogue with the stakeholder group that was periodically updated with interim outputs. Candidate inputs to the typology were identified from the literature, and assembled around five domains deemed relevant by the stakeholder group. These were:

1. ***Employment Structure***: to capture the mix and type of industry and occupations (Gordon, Champion, & Coombes, 2015; O'Donoghue, 2016; Youn et al., 2016; Faggio and Silva et al., 2017; Frey & Osborne, 2017)
2. ***Dynamism/Attractiveness***: to capture both long and short term indicators of change (Meerow, Newell, & Stults, 2016)
3. ***Employee characteristics***: the skills and demographic characteristics

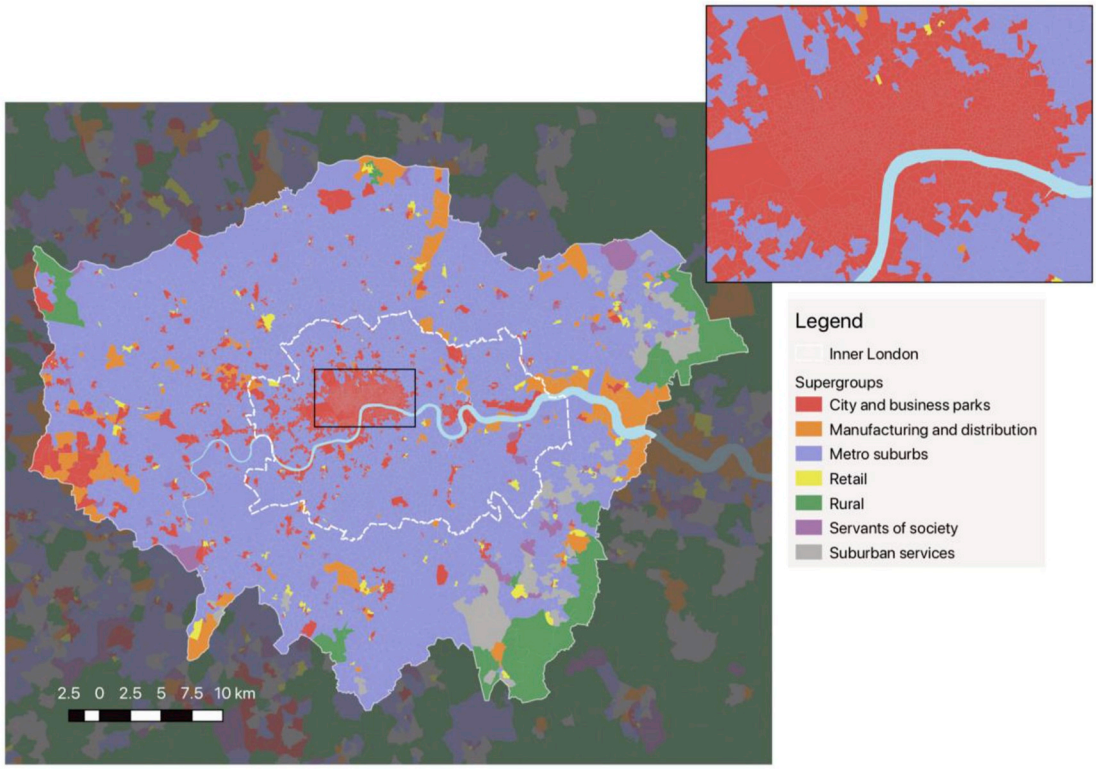

Fig. 2. Greater London Workplaces represented by the national COWZ.

of workers (Flynn, Schröder, & Chan, 2016; Salvatori, 2018)

4. ***Employment characteristics***: the nature of work undertaken, including hours worked and full/part time mix (Clayton, Williams, & Howell, 2014; Dawson and Henley et al., 2014; Green et al., 2016)
5. **Commuting/connections**: location accessibility and travel-to-work patterns (Jahanshahi, Jin, & Williams, 2015; Martin et al., 2018)

Within each domain, a series of sub domains and candidate measures were identified and evaluated using similar procedures to those described in Gale et al. (2016). An overview of the classification framework is provided in Table 1, which additionally details for each measure the relevant three-tier data access control (See Fig. 1), and the spatial scale at which the measures are available. As with COWZ, many variables were sourced from the 2011 Census of Population,[3] but this open source was supplemented with other recent data, sourced through the CDRC, the Office for National Statistics (ONS) and Transport for London (TfL): and used to create indicators pertaining to dynamism and attractiveness of workplace settings, retail structure, and accessibility. CDRC data that were used to create measures within the Retail Density and Night Time Economy Sub Domains were derived from the Local Data Company (LDC). LDC are a retail data and intelligence company who have a team of their own surveyors visiting UK retail centres on a rolling basis throughout each year, and record the location of retail premises alongside details of the specific retailer (or vacancy if an empty unit) which are classified into their own functional categories; for example, enabling differentiation between food versus clothing retailers. The latest extract that was made available for this study pertained to 2016. Non census data supplied by the ONS was used within the Dynamism/Attractiveness Domain, and included data taken from the Inter-Departmental Business Register (IDBR) which is a comprehensive list of UK businesses, and is used by government for statistical purposes. Within the Distance/Accessibility Sub Domain, the Public Transport Accessibility Levels (PTALS) were considered, and are a pre-calculated measure supplied by Transport for London (TfL), and provide an accurate measure of the accessibility of a work place zone (and other geography) to the public transport network, considering walk access time and service availability.

The availability of public domain data for classifications of this nature is potentially problematic where rights of use do not extend to the creation of derivative products. Strictly speaking, reuse of UK open data should always be under the terms of an Open Government License (OGL), although in practice, other datasets are available under unrestrictive licensing terms. All publicly available sources used in this study brought no restrictions on circulation of the resulting classification.

3.2. Variable evaluation, final selection and standardisation

Consistent with the data analysis procedures used in other geodemographic classifications (Gale et al., 2016; Vickers & Rees, 2007), candidate input variables were examined and problematic variables removed. However, given the sensitive nature of some data sources, this procedure was carried out within a CDRC secure data laboratory. Considerations for exclusion included: very low variability with limited discriminatory power; high positive or negative correlation with resultant undue impact upon cluster formation; and similar distributions or low counts. Exploratory statistical analysis and mapping, in conjunction with consultation with the stakeholder group led to removal of several measures. The excluded variables are listed in Table 2 alongside the Domain and Sub Domain from which they were drawn. The remaining variables were then range standardized onto a 1-0 scale in order to limit the impact of outliers; and following Spielman and Singleton (2015), no other normalisation was implemented.

[3] Census data were obtained from Nomis: https://www.nomisweb.co.uk/census/2011; and for those Tables selected, all variables were considered for evaluation as presented, with the exception of age and health, where bands were created.

Table 1
Classification framework, security restrictions and spatial scale.

Domain	Sub Domain	Measure*	Data Access Type	Spatial Scale
Employment Type	Employment	Worker density	Public	Workplace Zone
		Worker industry %		
	Occupation Types	Worker occupation %	Public	Workplace Zone
	Retail Density	Density of retailers[+]	Controlled	Address co-ordinate
		Density of retailers by category[+]		
Dynamism/Attractiveness	Change	Workplace % change 2009–2015[&]	Public	Workplace Zone
	Night-time Economy	Night-time economy businesses %[+a]	Controlled	Address co-ordinate
Employee Characteristics	Demographic	Age All/Male/Female % 16-24	Public	Workplace Zone
		Age All/Male/Female % 25-39		
		Age All/Male/Female % 40-64		
		Age All/Male/Female % 65 +		
	Diversity	Ethnic group %	Public	Workplace Zone
		Country of birth categories %		
		Length of residence in the UK categories %		
	Socio-economic	General health categories %	Public	Workplace Zone
		Tenure categories %		
Job Characteristics	Qualifications	Qualification categories %	Public	Workplace Zone
	Working day	Employment status categories %	Public	Workplace Zone
		Hours worked categories%		
	NS-SEC	NS-SeC top level categories %	Public	Workplace Zone
Commuting/Connections	Distance/Accessibility	Distance travelled to work categories %	Public	Workplace Zone
		Average distance travelled to work		
		Public Transport Accessibility Levels[^]		
		Workers from outside of London %		
	Mode	Transport mode categories %	Public	Workplace Zone

Notes: * = Where not otherwise specified, data are sourced from the 2011 Census; + = Supplied by ESRC Consumer Data Research Centre (CDRC); & = Supplied by the Office for National Statistics (http://bit.ly/2qF0KMl); ∧ = PTAL data were created by TfL and are available: http://bit.ly/2raLR8b.

[a] In consultation with the stakeholder group, "Night-time economy businesses" were defined as: LDC designations of "Bars, Pubs & Clubs", "Off Licences", "Restaurants"; LDC sub category designations of "Cafes & Fast Food" defined as "Fast Food Takeaway", "Take Away Food Shops", "Fish & Chip Shops", "Pizza Takeaway", "Chinese Fast Food Takeaway", "Indian Takeaway"," Fast Food Delivery"; and LDC sub category designations of "Entertainment" defined as "Amusement Parks & Arcades", "Theatres & Concert Halls", "Cinemas", "Snooker, Billiards & Pool Halls", "Bowling Alleys".

Table 2
Variables removed by domain and sub domain.

Domain	Sub Domain	Removed Variables
Employment Type	Employment	None
	Occupation Types	A Agriculture, forestry and fishing; D Electricity, gas, steam and air conditioning supply; U Activities of extraterritorial organisations and bodies; T Activities of households as employers; E Water supply, sewerage, waste management and remediation activities; B Mining and quarrying
	Retail Density	Convenience retail density; Comparison retail density; Leisure retail density; Service retail density
Dynamism/Attractiveness	Change	None
	Night-time Economy	None
Employee Characteristics	Demographic	Female 65 +
	Diversity	Africa: North Africa; Mixed/multiple ethnic group: Other Mixed; Mixed/multiple ethnic group: White and Asia; Mixed/multiple ethnic group: White and Black African; Mixed/multiple ethnic group: White and Black Caribbean; Black/African/Caribbean/Black British: Other Black; Other ethnic group: Any other ethnic group; White: Gypsy or Irish Traveller
	Socio-economic	Bad & Very Bad Health
		Good & Very Good Health
Job Characteristics	Qualifications	None
	Working day	Self-employed with employees: Part-time
	NS-SEC[1]	None
Commuting/Connections	Distance/Accessibility	None
	Mode	Motorcycle, scooter or moped; Taxi

1 – NS-SEC – National Statistics Socio-Economic Classification (https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010).

3.3. Estimating cluster frequency and clustering

We utilised k-means clustering, which has a long history of application when building geodemographics. This method begins by assigning an initial set of "seeds", typically at random locations within the attribute space of the data inputs. The distance between all records (workplace zones) and their nearest seeds were assigned, and the mean value of the clusters initially identified were calculated; these new mean locations were then used to re-assign records to their nearest centroid. This process continued iteratively until no further reassignments occurred. Given that the initial seed locations are random, the optimised outcomes were stochastic, and as such require multiple re-runs to estimate a globally optimal solution relative to the ascribed starting seed locations. Automated comparison between solutions often uses the ratio between the within and between cluster sum of squares, and was implemented here. The process of cluster analysis was essentially a statistical procedure, but the results were summarised for deliberation by the stakeholder group. The first stage in building the geodemographic was to select an appropriate number of clusters that both effectively represented salient groupings within the data, and would be of utility to the stakeholder group. A clustergram (Schonlau, 2002) suggested that five or six clusters presented a stable solution.

Table 3
Group level clustering results.

Clusters	Population		Workplace Zones	
	N.	%.	N.	%.
A	830552	18.5	1774	21.8
B	1464405	32.5	1668	20.5
C	814008	18.1	1443	17.7
D	724861	16.1	1766	21.7
E	666655	14.8	1503	18.4

Following discussion with the stakeholder group, it was decided to proceed with five clusters, with the greater ethnic differentiation of the workforce afforded by the six cluster solution not seen as a priority in end uses of the classification.

A summary of the cluster distribution is shown in Table 3 and forms the "Group" level of the typology; listing each of the five clusters alongside their constituent population and workplace zone frequencies and proportions. Both the total population and workplace zones had a reasonably even distribution, except for one cluster where the population was marginally higher. These are also mapped for the Greater London Extent in Fig. 3, and show differentiation within and between central and suburban locations.

A second tier was then built to provide greater detail within each Group. This involved splitting the input data up by each WZ identified Group cluster, and separately running further k-means on the disaggregated data. Again, clustergrams were used to explore structure within the data, and the final allocation of k within each Group cluster analysis agreed with the stakeholder group. After exploration of a range of results, those Sub Groups that showed the most effective partitioning in terms of within and between Group differentiation are presented in Table 4, and were agreed with the stakeholder group. All Groups were partitioned into two further clusters, with the exception of D, which was split into three. This created the Sub Group level of hierarchy and consisted of 11 clusters.

After building the Sub Group tier of the classification it was possible to examine the cluster fit of each WZ by comparing the relative difference between the input attributes for the zone and their assigned Sub Group cluster mean. This creates a score for each input variable and can be summed for each area, thus creating an overall measure. A higher score indicates a poorer fit, as the WZ attributes are further from their assigned cluster mean. These are mapped in Fig. 4 and there is a reasonably even fit, with no particular spatial pattern emerging.

Table 4
Sub group distribution.

Sub Group	Population		Workplace Zones	
	N.	%.	N.	%.
A1	554110	12.3	1064	13.0
A2	276442	6.1	710	8.7
B1	406049	9.0	294	3.6
B2	1058356	23.5	1374	16.9
C1	211154	4.7	407	5.0
C2	602854	13.4	1036	12.7
D1	233217	5.2	571	7.0
D2	194466	4.3	535	6.6
D3	297178	6.6	660	8.1
E1	431987	9.6	945	11.6
E2	234668	5.2	558	6.8

3.4. Cluster description

A common practice in geodemographics is to create verbal 'pen portraits' to describe the melange of numerical scores that characterise each Group or Sub Group. The variables are transformed to index scores, where 100 is the all zone average, 50 one half, 200 double, and so forth. Using the scores, both labels and descriptions were created for the two-tier hierarchy and were ratified with the stakeholder group. This was the final analysis to be completed in the secure lab, with the following labels and a lookup between WZ, Group and Sub Group then output.

Group – A: Residential Services: These workplace zones are characterised by services offered to local communities by local community members. Occupations include classroom assistants, domestic

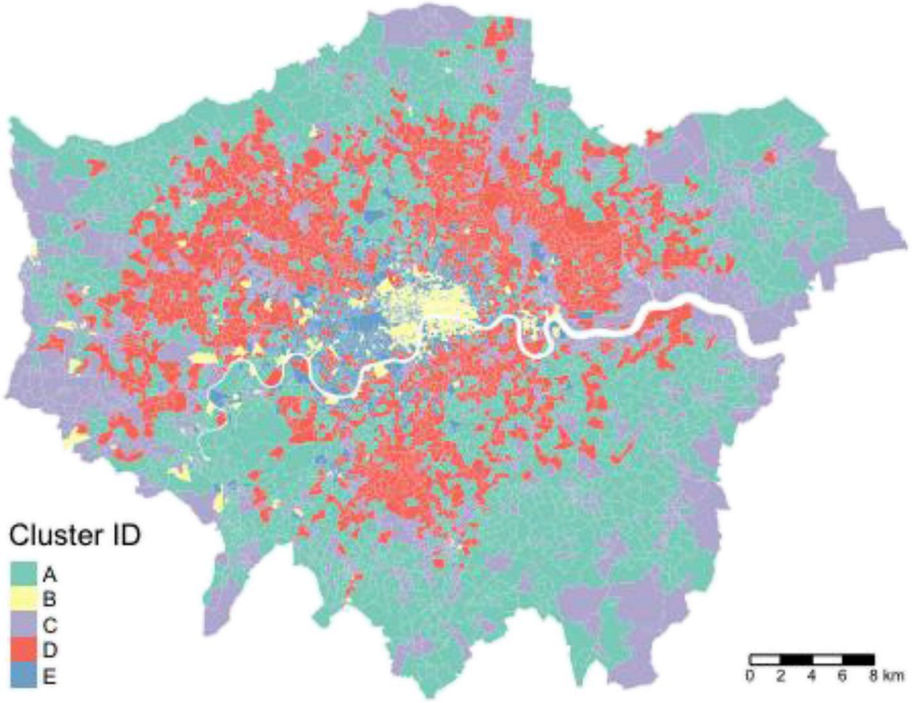

Fig. 3. Group level clustering results mapped.

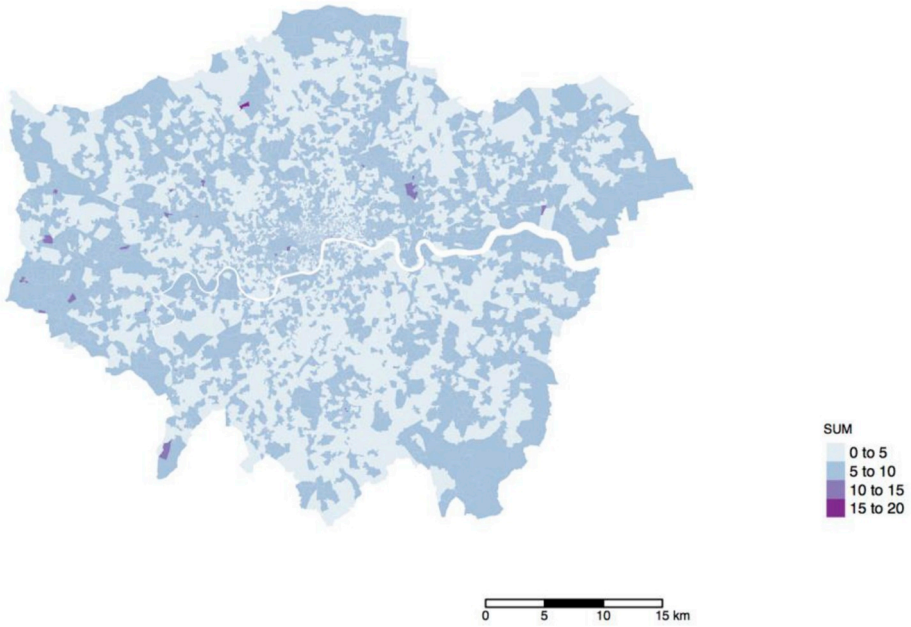

Fig. 4. Sub group cluster fit.

assistants and self-employed cleaners. Workers, particularly women, are typically older than average and some above normal retirement age.

A1: Predominantly older, local education and health workers: These workers are frequently sourced locally, and predominantly work in the health and education sectors. Caring and leisure services are well represented, and some individuals work in professional occupations.

A2: Low qualified workers in construction and allied local trades: Work in construction and related skilled and unskilled trades predominates, although there is also some representation of health, social work and education-related activities. Many workers are self-employed or work for small companies. Workers tend to be drawn from the older age cohorts, although some are employed as apprentices.

Group – B: City Focus: These areas bring focus to a range of specialised professional activities. They also host more general support services and retail activities. The portfolio of over-all activities may also be catalytic to the vitality of night-time economies. Workers in all of these activities are predominantly drawn from a core (age 25–39) labour force.

B1: Dynamic financial centres with extended operating hours: These areas form the close knit financial heart of the City. Much of the younger, and predominantly full-time workforce commute by rail over significant distances. Many workers fulfil managerial roles within their organisations. The areas also host significant retail and leisure functions that contribute to a vibrant night time economy.

B2: Professional, retail and leisure Services in dynamic central locations: This predominantly full-time, well qualified labour force often commutes long distances to work in Central and West London locations. These tight knit employment zones host a range of professional and scientific and technical activities. There is also strong representation of supporting retail and leisure services, and a night-time economy.

Group – C: Infrastructure Support**:** Workers in these areas provide direct or indirect support for the physical infrastructure of the economy – in transport, utilities and the retail trade. Workers are drawn from the traditional workforce and there is strong labour force participation from Asian ethnic minorities.

C1: Younger customer service workers in wholesale or retail occupations: This young, locally based and studentified labour force are employed at locations scattered widely across Outer London. Commuting is typically by car or bus. Employment includes retail and customer service with workers drawn disproportionately from Asian backgrounds. Workers have relatively low-level qualifications and part-time working is common.

C2: Blue collar, manufacturing and transport services: These workers find employment at locations scattered throughout London, with some concentration on the Capital's outermost fringes. Employment is found in a wide range of occupations and workers tend to have low or intermediate level educational qualifications. Travel to work is often by car.

Group – D: Integrating and Independent Service Providers: These areas are characterised by high levels of self-employment, and significant numbers work part-time. Workers may be based at home, or travel to deliver services to local communities. The areas attest to the dynamism of London's economy in recent years, providing employment for recent migrants and longer settled members of ethnic minorities. These zones predominantly make up an annular tract of land surrounding the inner core of London.

D1: Health care support staff and routine service occupations: This heavily multicultural workforce is very locally based and employed in a wide range of occupations. Although some workers are skilled, many have low levels of educational qualifications and work in unskilled or semi-routine occupations. Levels of self-employment are high and residential context is characterised by higher than average unemployment.

D2 Locally sourced, home helps and domestic or manual workers: Domestic employers requiring caring, recreational and other services are an employment mainstay of these areas. Other trades and activities are present. Levels of self-employment and work for small employers is higher than average.

D3: Travelling or home-based general service providers: This generally low-skilled labour force has changed in markedly in recent years, in significant part as the result of immigration. Employment in low-skilled manual and administrative occupations predominates.

Group – E: Metropolitan Destinations: These areas are overwhelmingly located in Inner London, especially its West End, and many serve as retail destinations. A very international range of workers provide a wide range of high value services as well as retailing. Many of these workers also reside in Central London.

E1: High street destinations and domestic employers: Employment in these areas has a strong international service

orientation, although households also provide an important source of employment. Real estate and entertainment activities are in evidence, and various forms of retailing also underpin local economies. Journeys to work are typically short distances, mainly by public transport.

E2: Accessible retail, leisure and tourist services: These densely occupied destinations offer services in retailing, leisure and accommodation. They have important night time economies. Public transport predominates in the journey to work over short to medium length commutes. There is high turnover in the workforce and routine occupations predominate are common.

The pen portraits were designed with the objective of giving oversight to salient characteristics where there is clear coincidence of a distinctive labour market profile with employment location, and additionally, differentiate the unique characteristics of London employment from those found within national CoWZ. Their utility and appropriateness of such descriptions and labels was assured through stakeholder consultation in their design.

4. Discussion and conclusion

The abiding message of this paper is that geodemographic classification remains a tried and tested approach to area profiling through shorthand descriptors of work place as well as residential locations. Fully open classifications require that the requisite data be made available without encumbrance. But in the age of Big Data this requirement is increasingly likely to mean that the data are not the most detailed, up-to-date or relevant to the purpose of the classification. Further progress thus becomes more contingent upon successfully navigating issues of data access and control, while retaining the confidence of stakeholders that their requirements remain paramount. The increasing real share of consumer data and the role of customer facing organisations inevitably means that a large proportion of the rich data that are assembled about citizens will have private sector custodians. For such sources to be made available for the public good, access protocols will need to be negotiated and data licencing agreements that respect commercial interests are likely to replace fully open licencing.

This paper has illustrated that these issues are thrown into sharp focus when the remit of geodemographics is extended from geographies of night-time residence to geographies of workplace location. In methodological terms we have demonstrated how a hybrid data access framework may be developed to blend potentially sensitive data sources alongside those that lie entirely within the public domain. There are a number of promising avenues for further development in this regard. The London Workplace Zone Classification successfully reuses a core of 2011 Census data (like the national COWZ classification), but blends it with other sources that are not in the public domain, while retaining engagement with end users of the product. By extension, our future goals are to use other consumer data to bring these classifications closer to real time updating and to introduce new data pertaining to social and workplace interactions. One objective is to use footfall data on retail centre activity both as an external descriptor of the existing classification and, prospectively, as an input variable to further classifications.

Taken together, these developments suggest that issues of data resourcing and custodianship need to be rethought if the best available data are to find their way into the best classifications. The advent of commercial geodemographic systems in the 1980s crystallised data as a commodity and strategic resource, with the consequence that some academics felt increasingly estranged from the best data required to develop policy tools. The advent of Big Data has brought new challenges in terms of metadata creation and establishing the provenance of detailed yet partial representations of socioeconomic and demographic systems, and has also created new barriers to academic access to new forms and sources of data. However, the advent of distributed and secure methods of data access creates new opportunities for implementation of derived measures within geodemographic classifications where appropriate consents have been obtained and data licences granted. We have illustrated how this mixed data economy can facilitate the creation of products that are data rich, salient and up-to-date. Such products are not, in the strictest sense, entirely scientifically reproducible in the spirit of open data, but they are nonetheless transparent and, we argue, can be sufficiently open to scrutiny.

Acknowledgement

This work was supported by ESRC Consumer Data Research Centre (ES/L011840/1).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apgeog.2019.102038.

References

Ashby, D. I., & Longley, P. A. (2005). Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS*. https://doi.org/10.1111/j.1467-9671.2005.00205.x.

Brunsdon, C., et al. (2011). Predicting participation in higher education: A comparative evaluation of the performance of geodemographic classifications. *Journal of the Royal Statistical Society - Series A: Statistics in Society*. https://doi.org/10.1111/j.1467-985X.2010.00641.x.

Clayton, N., Williams, M., & Howell, A. (2014). *Unequal opportunity: How jobs are changing in cities. London*.

Dawson, C., Henley, A., & Latreille, P. (2014). Individual motives for choosing self-employment in the UK: Does region matter? *Regional Studies*. https://doi.org/10.1080/00343404.2012.697140.

Faggio, G., Silva, O., & Strange, W. C. (2017). 'Heterogeneous agglomeration'. *Review of Economics and Statistics*. https://doi.org/10.1162/REST_a_00604.

Flynn, M., Schröder, H., & Chan, A. C. (2016). The impact of national context on age diversity and age management: The case of the UK and Hong Kong. In E. Parry, & J. McCarthy (Eds.). *The palgrave handbook of age diversity and work* (pp. 499–519). London: Palgrave.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2016.08.019.

Galbraith, J. K. (1958). *The affluent society*. Boston: Houghton Mifflin.

Gale, C. G., et al. (2016). Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science*(12), https://doi.org/10.5311/JOSIS.2016.12.232.

Gordon, I., Champion, T., & Coombes, M. (2015). Urban escalators and interregional elevators: The difference that location, mobility, and sectoral specialisation make to occupational progression. *Environment and Planning A*. https://doi.org/10.1068/a130125p.

Green, A., et al. (2016). *Improving progression from low-paid jobs at city-region level. York*.

Grubesic, T. H., Miller, J. A., & Murray, A. T. (2014). Geospatial and geodemographic insights for diabetes in the United States. *Applied Geography*. https://doi.org/10.1016/j.apgeog.2014.08.017.

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

Jahanshahi, K., Jin, Y., & Williams, I. (2015). 'Direct and indirect influences on employed adults' travel in the UK: New insights from the National Travel Survey data 2002-2010'. *Transportation Research Part A: Policy and Practice*. https://doi.org/10.1016/j.tra.2015.08.007.

Johnson, P. A., et al. (2017). The cost(s) of geospatial open data. *Transactions in GIS, 21*(3), 434–445. https://doi.org/10.1111/tgis.12283.

Kirkpatrick, R. (2011). *Data philanthropy: Public & private sector data sharing for global resilience*. United Nations Global Pulse. Available at: https://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience Accessed: 18 September 2018 .

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

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

Longley, P. A., Cheshire, J., & Singleton, A. D. (2018). *Consumer data research*. London: UCL Press.

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*. https://doi.org/10.1111/j.1467-985X.2012.01054.x.

Martin, D., et al. (2018). Origin-destination geodemographics for analysis of travel to work flows'. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2017.09.002.

Meerow, S., Newell, J. P., & Stults, M. (2016). Defining urban resilience: A review. *Landscape and Urban Planning*. https://doi.org/10.1016/j.landurbplan.2015.11.011.

O'Donoghue, D. (2016). 'Exploring the links between employment clusters and economic diversity in the British urban system'. *Modern Economy*. https://doi.org/10.4236/me.

2016.77079.
Salvatori, A. (2018). The anatomy of job polarisation in the UK. *Journal for Labour Market Research*. https://doi.org/10.1186/s12651-018-0242-z.
Schonlau, M. (2002). 'The Clustergram : A graph for visualizing hierarchical and non-hierarchical cluster analyses'. *STATA Journal, 3*, 316–327.
Singleton, A. D., & Longley, P. A. (2009). Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*. https://doi.org/10.1016/j.apgeog.2008.10.006.
Singleton, A. D., & Longley, P. (2015). The internal structure of greater London: A comparison of national and regional geodemographic models. *Geo: Geography and Environment, 2*(1), 69–87.
Singleton, A., Pavlis, M., & Longley, P. A. (2016). The stability of geodemographic cluster assignments over an intercensal period. *Journal of Geographical Systems, 18*(2), 97–123. https://doi.org/10.1007/s10109-016-0226-x.
Singleton, A. D., Spielman, S., & Brunsdon, C. (2016). Establishing a framework for open geographic information science. *International Journal of Geographical Information Science, 30*(8), 1507–1521. https://doi.org/10.1080/13658816.2015.1137579.
Spielman, S. E., & 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.
The Economist. The world's most valuable resource is no longer oil, but data - regulating the internet giants. (2017). Available at: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data Accessed: 19 September 2018 .
Timms, D. (1971). *The urban mosaic: Towards a theory of residential differentiation.* Cambridge: Cambridge University Press.
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. https://doi.org/10.1111/j.1467-985X.2007.00466.x.
Webber, R., & Burrows, R. (2018). *The predictive postcode: The geodemographic classification of British society.* London: SAGE.
Youn, H., et al. (2016). Scaling and universality in urban economic diversification. *Journal of The Royal Society Interface*. https://doi.org/10.1098/rsif.2015.0937.