Computers, Environment and Urban Systems 82 (2020) 101486

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Mapping the geodemographics of digital inequality in Great Britain: An integration of machine learning into small area estimation

Alex Singleton[a,*], Alexandros Alexiou[b], Rahul Savani[c]

[a] Department of Geography and Planning, University of Liverpool, Liverpool, UK
[b] Public Health and Policy, University of Liverpool, Liverpool, UK
[c] Department of Computer Science, University of Liverpool, Liverpool, UK

ABSTRACT

Geographic variation in digital inequality manifests as a result of a range of demographic, attitudinal, behavioural and locational factors. To better understand this multidimensional geography, our paper develops a new geodemographic classification for the spatial extent of Great Britain. In this model, we integrate a range of new small area measures that are drawn from multiple new forms of data including consumer purchasing data, survey and open data sources. Our analytical approach innovatively provides an integration of machine learning into a small-area estimation technique to obtain Lower Super Output Area / Data Zone estimates of Internet use, alongside a range of online engagement and consumption measures. Following the collation of a range of input measures, we implemented a more standard geodemographic framework that utilises the unsupervised clustering algorithm k-means to produce a map of the multidimensional characteristics of digital inequality for Great Britain; creating the Internet User Classification (IUC). Our outputs provide a new and nuanced understanding of the contemporary salient characteristics of digital inequality in Great Britain, which we evaluate both internally and externally within the context of preparations for the 2021 UK Census of the Population, exploring the geodemographic patterns of Census test response rates and the prevalence to complete the survey online. Our innovative work illustrates the strength of a geodemographic approach in mapping spatial patterns of digital inequality, and through the presented application concerning Census response rates and characteristics we demonstrate how the IUC can be operationalised within such settings for local intervention or benchmarking.

1. Introduction

Digital inequality is observable where access to online resources and those opportunities that these create are non-egalitarian. As a result of variable rates of access and use of the Internet between social and spatial groups (Inkinen, Merisalo, & Makkonen, 2018), this leads to digital differentiation, which entrenches difference and reciprocates digital inequality over time. Digital inequality has been shown to be enacted across multiple divides (Büchi, Just, & Latzer, 2015), including those of individual attributes such as age, gender, and ethnicity; and additionally, can be further impacted through interactions between themselves, or across different social or spatial contexts (Blank, Graham, & Calvino, 2018; Chang, McAllister, & McCaslin, 2015; Friemel, 2016; Hunsaker & Hargittai, 2018; Longley & Singleton, 2009); and include impacts of physical infrastructure such as connectivity (Gonzales, 2016; Grubesic, Helderop, & Alizadeh, 2018; Marler, 2018; Tsetsi & Rains, 2017).

Beyond variability in access, for those who do connect, digital inequality manifests across a range of engagement activities, for instance in terms of social network use (Yu, Ellison, McCammon, & Langa, 2016) or online shopping (Lissitsa & Kol, 2016). Within many contexts, such engagement patterns have been shown to have potential impacts upon life chances, with evidence including health (Borg, Boulet, Smith, & Bragge, 2018; Mesch, 2015) and employment (Peng, 2017). Given such social and economic imperatives of mitigating digital inequality, there is significant need for better data that delivers necessary information to enable investments to be targeted or evaluated. Within such contexts, measures are required for national extents (Szeles, 2018; Thomas et al., 2017) and, in recognition of the dynamic nature of digital exclusion (Van Dijk & Hacker, 2003), should be in a form that might be updated over time.

The interrelated factors that produce digital inequalities are highly dimensional and lend themselves to study through compound and multidimensional indicators (Vehovar, Sicherl, Hüsing, & Dolnicar, 2006), rather than monotopical measures (Barzilai-Nahon, 2006; Borg & Smith, 2018). Such an approach enables the capture of both the breadth of influences, but also instances where such relationships are complex: for example, age is generally assumed to influence rates of Internet use, however, this relationship may vary across space as aligned to the geography of infrastructure provision or other compounding geographical factors. Within such contexts, geodemographics has a legacy of successful application in both the public and private sector (Harris, Sleight, & Webber, 2005; Singleton & Spielman, 2014; Webber & Burrows, 2018). In general terms, a geodemographic

* Corresponding author at: Roxby Building, 74 Bedford St S., Liverpool L69 7ZT, UK.
E-mail address: alex.singleton@liverpool.ac.uk (A. Singleton).

https://doi.org/10.1016/j.compenvurbsys.2020.101486
Received 20 December 2019; Received in revised form 24 March 2020; Accepted 27 March 2020
Available online 08 April 2020

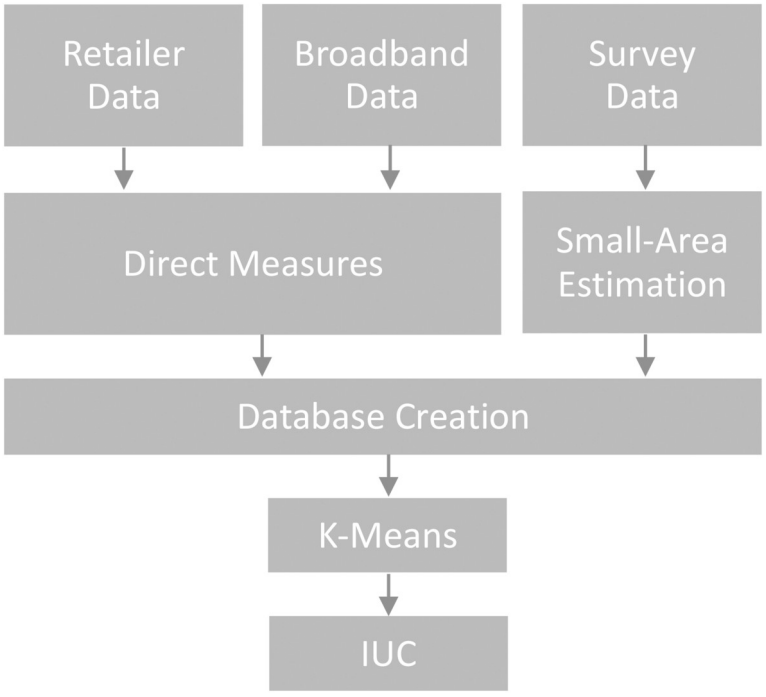

Fig. 1. Methodology flow chart.

Table 1
Domains, dimensions and measures of digital inequality.

Domains	Dimensions	Measures
Context	Durables	Own desktop PC
		Own laptop PC
	Infrastructure	Median download speed
		Broadband access
		Cable broadband by TV provider
Access	Internet history	No access
		Less than 3 years
		More than 6 years
		Never but I have access
	Internet frequency	Once a week or less
		Daily
	Access method	PC at home
		PC at work, school etc.
		Public place
		Mobile device
Behaviour	Information seeking and services	Information on hobbies or interests
		Information on products or services
		Financial services
	Communication and entertainment	Social networks
		VOIP
		Download or stream media
		Gaming
	Online shopping	Buy groceries
		Buy non-groceries
		Clothing on credit
		White goods

classification is created by assembling a wide range of measures that describe the characteristics of areas and/or those people living within them, and then, through the implementation of unsupervised learning (clustering), identifies groups of areas that share common characteristics. Emerging clusters may be divided or aggregated to create a hierarchy, and it is typical that these be accompanied by labels, descriptions, photographs, diagrams and graphs. Classifications are created to fulfil a general purpose or can be tailored to bespoke applications including substantive topics or constrained geographical extents.

The objective of this paper is therefore to develop a nationwide geodemographic classification that provides insight into the geography of digital inequality. As discussed earlier in this section, such inequality arises through digital differentiation, which is a multidimensional construct that can manifest in complex ways and across multiple different divides. For more wholistic insight, this necessitates the development of a comprehensive set of measures that aim to describe these various dimensions. As such, in Section 2, we outline the creation of a database of small area measures that have coverage for the national extent, with the objective of characterising multiple aspects of digital differentiation; either in terms of its manifestation, or to account for known exogenous drivers of these patterns. The multifaceted nature of digital differentiation is not well captured by attributes contained within traditional sources of geographically referenced data such as the Census. As such, alternative data were required to develop measures that provide insight into the different dimensions of this construct.

Both direct measures (available for a nationally extensive geography) and small area estimates created from large nationally representative surveys were identified for this purpose. We provide a theoretical framework that guided how the measures were assembled in Section 2; but broadly concerns context (constraints to getting online), access (frequency and prevalence) and behaviour (activities online). Within this framework, measures were derived from a range of different data sources. These included a number of datasets with national coverage: two capturing aspects of online consumer behaviour as recorded by two large retailers (Section 2.1), and further data that measured Internet connectivity through broadband speed (Section 2.2). To garner further insights into online consumption and capture a wider range of associated behavioural measures, a large nationally representative survey was examined. However, this did not directly have coverage or sample size that was sufficient to provide estimates for all small areas. As such, and as outlined in Section 2.3, we applied a small-area estimation technique to provide estimates for a wide variety of attributes related to Internet use and engagement. Innovatively, this applied a machine learning gradient boosting framework to obtain robust small-area estimates.

Once input measures were assembled, we implemented the unsupervised learning technique of k-means to generate a set of clusters

Table 2
Recoded variables, common between the UK 2011 Census and the BPS.

Domain	Variables
Age structure	Age 15–17; 18–24; 25–34; 35–44; 45–54; 55–64; 65 Plus.
Marital status	Single; Married; Separated; Divorced; Widowed.
Car ownership	No Car; 1 Car; 2 Cars; 3 Cars or more.
Number of people in household	Number of persons 1; 2; 3; 4; 5 or more.
Highest qualification	Qualification None & Other; Qualification 1 & 2 or Apprenticeships; Qualification 3 & 4 or above (based on 2011 Census categories).
Social grade - NS-SeC	Social Grade A & B (NS-Sec 1 & 2); Social Grade C1 (NS-Sec 3 & 4); Social Grade C2 (NS-Sec 5 & 6); Social Grade D & E (NS-Sec 7, 8 & Student).
Employment / working status	Part Time; Full Time; Self Employed; Unemployed; Retired; Housewife; Long Disabled; Student (Working and Not working); Not Working Other.
Household tenure	Owned; Mortgage; Rented (LA); Rented (Social Other); Rented (Private); Rent Free.
Ethnicity	White British; White Irish; White Other; Indian; Pakistani; Bangladeshi; Chinese; Other Asian; Black; Mixed; Other.

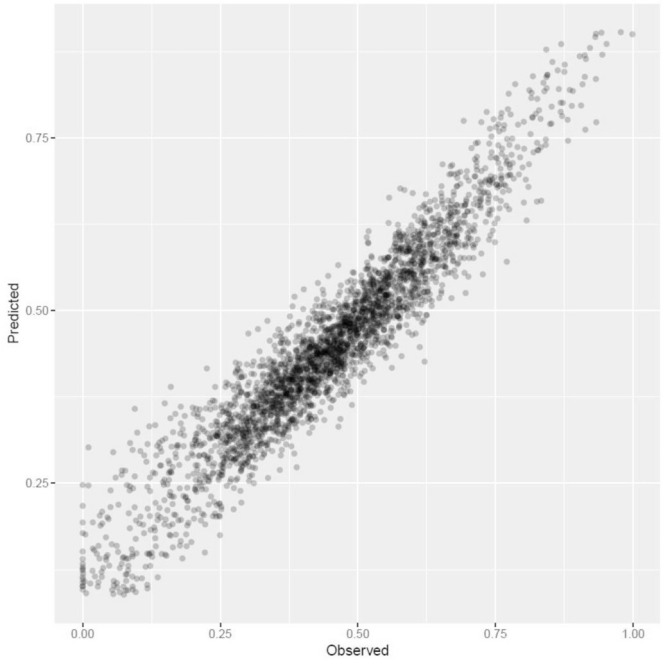

Fig. 2. Observed versus predicted values for the model describing the percentage of LSOA population that recently used the internet to visit social networking sites, blogs and forums (variable Social Networks).

that described groupings of areas with shared salient characteristics (Section 3). This created a simple indicator that highlighted patterns of digital differentiation and reflected its compound, complex and multi-dimensional nature. An alternative approach might have been to produce rankings for different input measures and to compile these into an overall index. However, such an approach does not account well for non-linearity as a result of complex interactions that might emerge between areas; for example, where infrastructure's variable provision may constrain or enable behaviour. A geodemographic or cluster-based approach mitigates such issues. In Section 3.1, to maximise the utility of the created typology, we present a range of measures that describe those salient characteristics of the clusters and map these for the national extent: producing names and "pen portrait" descriptions. These are a helpful to guide to end users in the operationalisation of the classification for applications looking at issues related to digital differentiation.

The different stages in the methodology implemented to create the classification are summarized in Fig. 1.

Although we evaluate the classification performance internally in Section 3 through examining a variety of cluster fit statistics; it is also good practice to provide an external evaluation (see Section 4). Here we present an example policy application; exploring how digital differentiation might be planned for in the implementation of the next UK Census that is due in 2021. In preparation, the Office for National Statistics (ONS) has been conducting various Census tests within England and Wales that are aimed at optimising their delivery methodology. A major change from the previous Census delivered in 2011 is that in 2021 this will predominantly be delivered online. Data from one of these tests are examined by the created classification to explore how

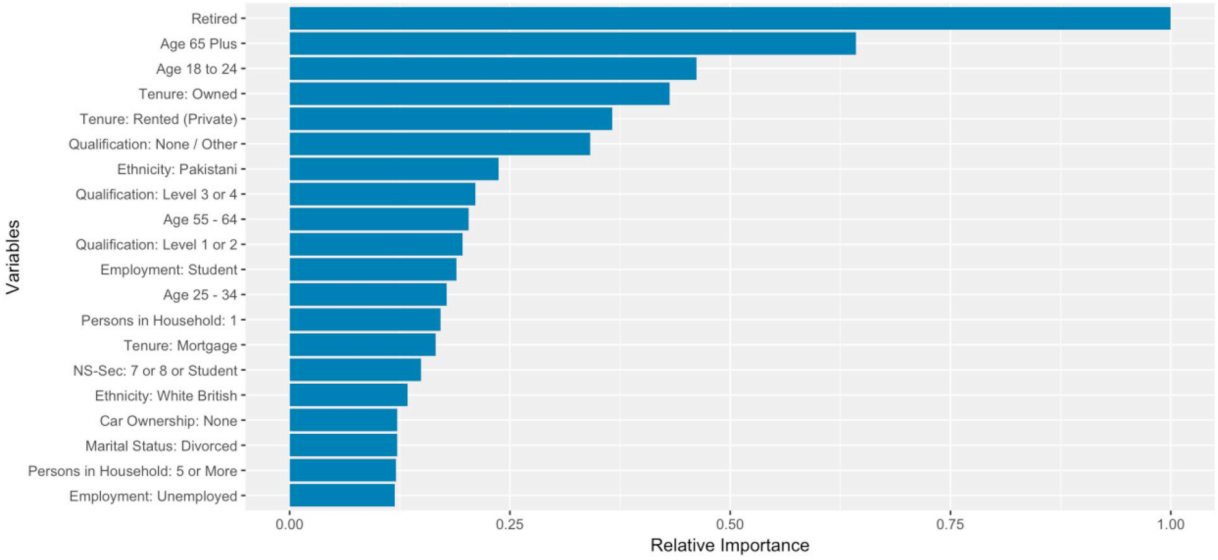

Fig. 3. Independent variable importance (1.0 being the most influential) for the final model predicting "Buy Groceries Online: Yes", top 20 variables.

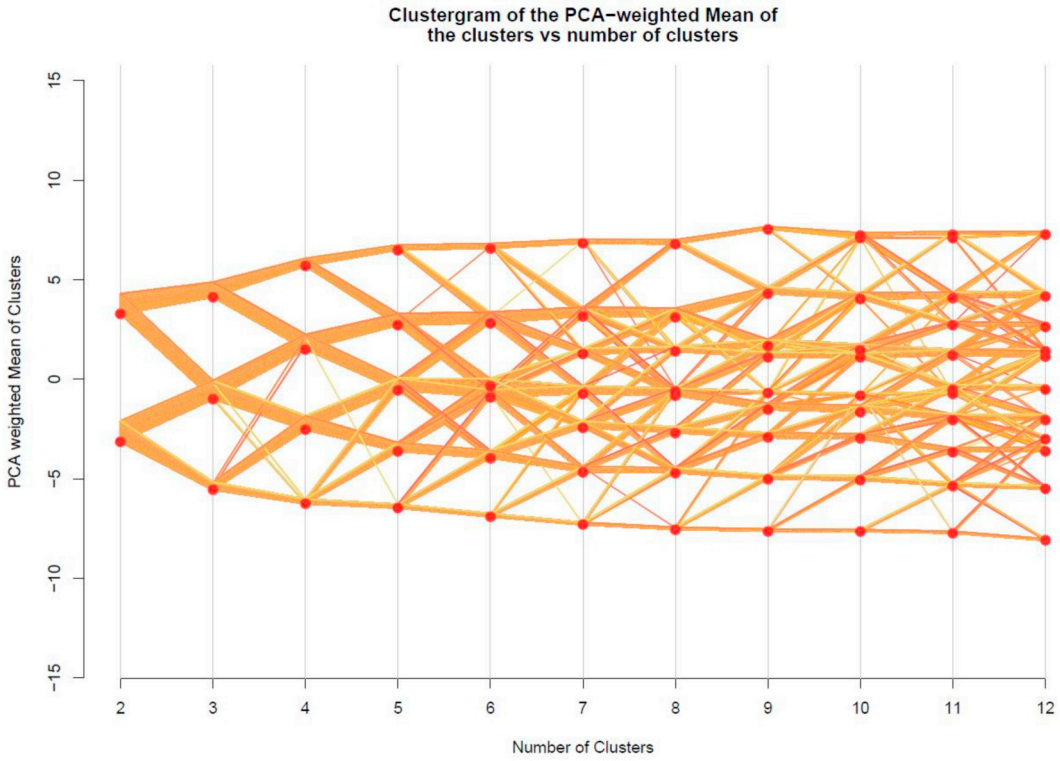

Fig. 4. Clustergram showing potential k values for the initial partitioning of the input data.

this might impact response rates and frame potential mitigation strategies. We conclude in Section 5 by drawing conclusions from this work, and discuss some areas that we would like to explore in future research.

2. Creating a database of measures describing digital inequality

Digital inequality is the produced outcome of differential consumer uptake of the Internet, alongside variations in online user behaviour. As identified in the literature in Section 1, a range of socio-economic and demographic factors influence the spatial variegation of such patterns which can be enhanced or constrained by available infrastructure. This section presents work that captured measures for small-area geography describing the multiplex of ways in which the areas or inhabiting local populations are digitally differentiated.

Data sourced for this study included a number extracts supplied by the ESRC Consumer Data Research Centre (CDRC),[1] which is a national provider of consumer data within the UK, and included: two sets of major Internet retailer transactional data pertaining to online shopping at the customer level (see Section 2.1); alongside the British Population Survey (BPS),[2] which is a commercial survey that provides individual-level behavioural characteristics regarding various aspects of Internet use (See Section 2.3). These data were supplied by the CDRC and accessed through their secure data lab. Other characteristics, such as broadband speed were supplied by Ofcom (see Section 2.2) as open data, and were publicly available.[3]

Measures were required for the extent of Great Britain (GB) and at the small area level. For England and Wales, the Census geography of Lower Super Output Area (LSOA) was selected (34,753 zones), alongside their equivalent of Data Zones (DZ) in Scotland (6976 zones). LSOAs are designed to have a population between 1000 and 3000 and DZs between 500 and 1000; with their combined geography giving coverage of the full extent of GB (41,729 zones). These geographic units offered a good balance between a granularity that provided useful local differentiation yet were aggregate enough that estimates could be calculated robustly.

In the remainder of this section, we describe the compilation of inputs to the geodemographic classification. Table 1 presents a summary of measures, which are ordered across a range of dimensions and three domains of Context, Access and Behaviour. The domain of "Context" refers to rates of durables or aspects of infrastructure that are known to enhance or constrain access. The "Access" domain considers the longevity, frequency and predominant method used to access the Internet. "Behaviour" considers those rates of different types of activity that are being conducted online. The specificity of the measures selected were shown to be either an important influence or outcome of digital inequality in the literature (see Section 1), have demonstrated utility in previous work (Dolega, Reynolds, Singleton, & Pavlis, 2019; Singleton, Dolega, Riddlesden, & Longley, 2016), or could be estimated effectively from the available data. Unlike many traditional geodemographic classifications, we did not include aspects of demography or socio-economic status directly as inputs. As we will discuss in Section 2.3, these were instead utilised as donor variables within the small-area estimation framework.

2.1. Direct measures of online consumption

Data from two major national digital retailers were supplied by the CDRC and provided insight into the online provision of goods. The first retailer specialised in white goods and home appliances, and comprised ~5 million consumer records covering sales during the period January 2013 to February 2016; providing transactional information about online orders of white goods at the product level, including attributes related to product category, value and order location. The second data

[1] The CDRC provide access to a range of consumer data for research applications in the public good – the catalogue of data holdings alongside metadata are available here - http://data.cdrc.ac.uk/

[2] http://www.thebps.co.uk

[3] https://www.ofcom.org.uk/research-and-data/data/opendata

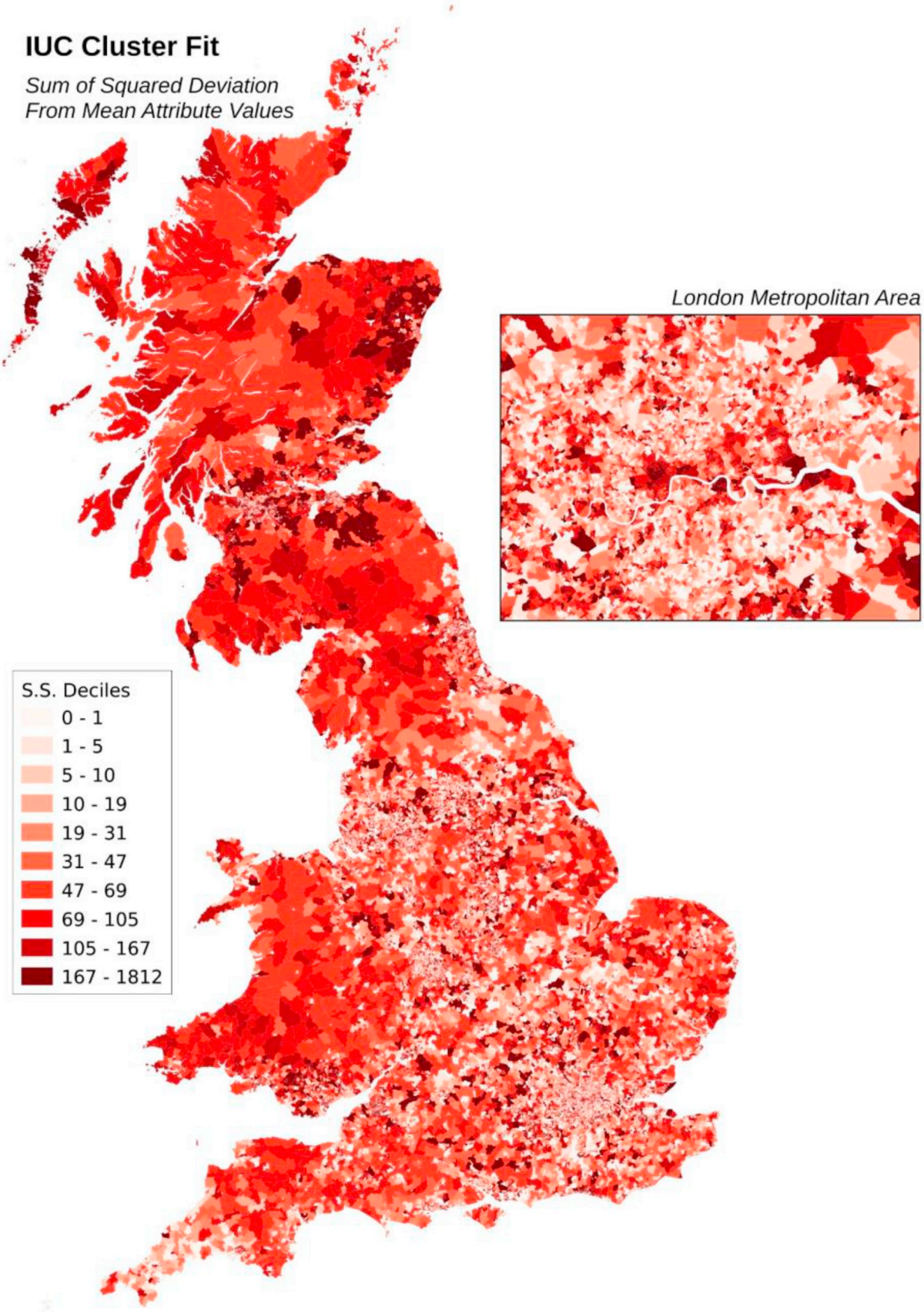

Fig. 5. Cluster fit scores presented for the GB national extent by LSOA/DZ.

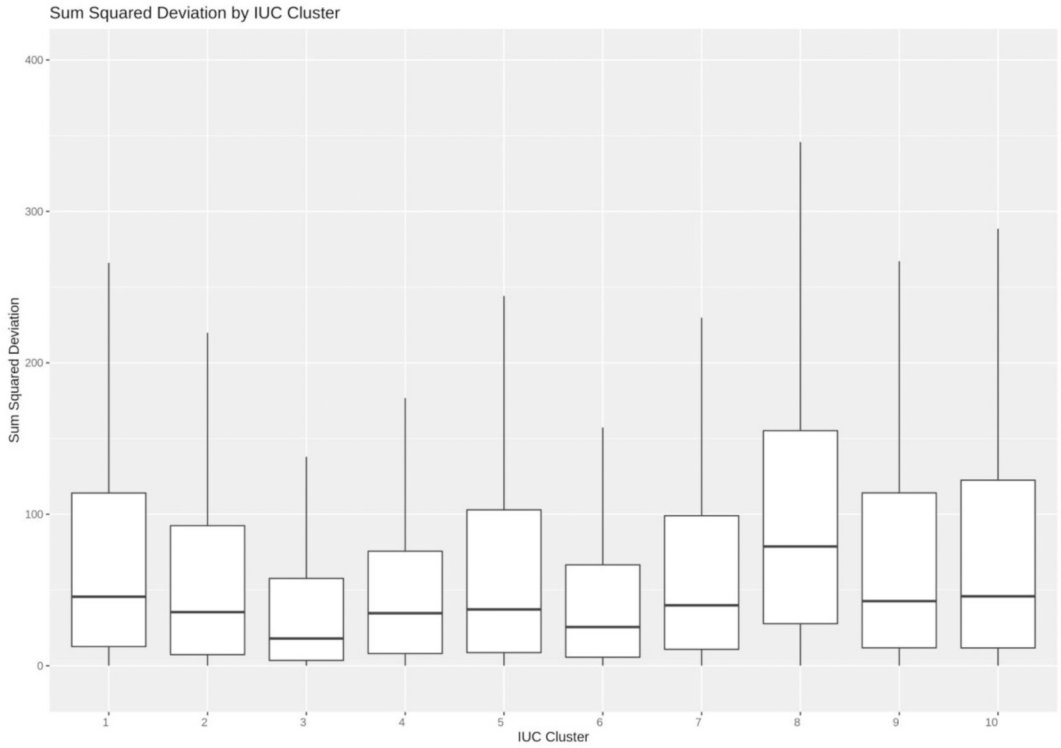

Fig. 6. Variability in cluster fit scores within the created classification.

provider was an online retailer specializing mainly in credit purchases of clothing, electronics, furniture, homeware and jewellery. These data pertained to ~1.6 m consumers making purchases within the period of January–December 2015. Full metadata for this provider are publicly redacted by the CDRC, but for the supplier of white goods and home appliances, these can be found here: https://data.cdrc.ac.uk/dataset/appliances-online-consumer-delivery-and-product-level-data.
Evaluation of these data gave assurance that these were both of a size and geographic distribution effective for the application. From these data, two measures of online shopping activity were constructed, each calculated as the total number of transactions per LSOA/DZ, normalised by the total population of that LSOA/DZ. A very small number of LSOA/DZ (0.071%) had no transactions recorded, and were given a value of 0. Other transactional measures, such as the average transaction value, were highly correlated (> 0.99) and were not used in the analysis.

2.2. Direct measures of servicing fixed infrastructure

Ofcom is the UK's communications regulator and, as supplement to their statutory duties, they also provide a range of open data including information about fixed broadband characteristics, mobile and Wi-Fi coverage. Fixed broadband data for 2016 were used here, which provided attributes on broadband type and speed at the postcode level. The average median download speed per postcode was calculated for each LSOA. Mobile data were unfortunately only provided at the Local authority level and thus were unusable, whereas the Wi-Fi dataset, although offered at the postcode level, is experimental and was deemed to lack accuracy.

2.3. Small area estimation of Internet user behaviour using supervised machine learning

The British Population Survey (BPS) is a commercial survey with a monthly historical time series that dates back to January 2008. The BPS provides a representative sample of responses to a variety of questions concerning the socio-economic, demographic and consumer characteristics (including consumer durable purchasing and online behaviour) of the British population. These data are collected through face-to-face surveys with over 80,000 individuals per year, and utilises a geodemographically stratified sample with additional target group boosting. Out of the total 641,323 interviews, only those from the years 2011 to 2015 were used, amounting to 389,074 observations. In order to maximise complete cases, a small number ($< 0.85\%$) of missing values (e.g. “Refused”, “Don't know” were imputed at the individual level using a hot-deck k Nearest Neighbours (kNN) imputation technique (Kowarik & Templ, 2016). kNN replaced a missing value on a specific record with a value obtained from a set of k most similar cases in the dataset; in this instance, the median value out of 15 most similar cases. One advantage of the kNN technique is the ability to handle a variety of data, and it is particularly useful for dealing with surveys where values are not missing at random. However, one disadvantage of kNN was the large computational power that it required.

The BPS has over 30 variables regarding Internet use and engagement (as well as several other attributes related to durables, such as owning a desktop PC, laptop, or a gaming device) that appear to be asked consistently throughout the years that the survey has been run. Those input attributes selected from the survey described some aspect of online behaviour. However, we aimed to limit variables that had high correlations, which can negatively impact a geodemographic classification

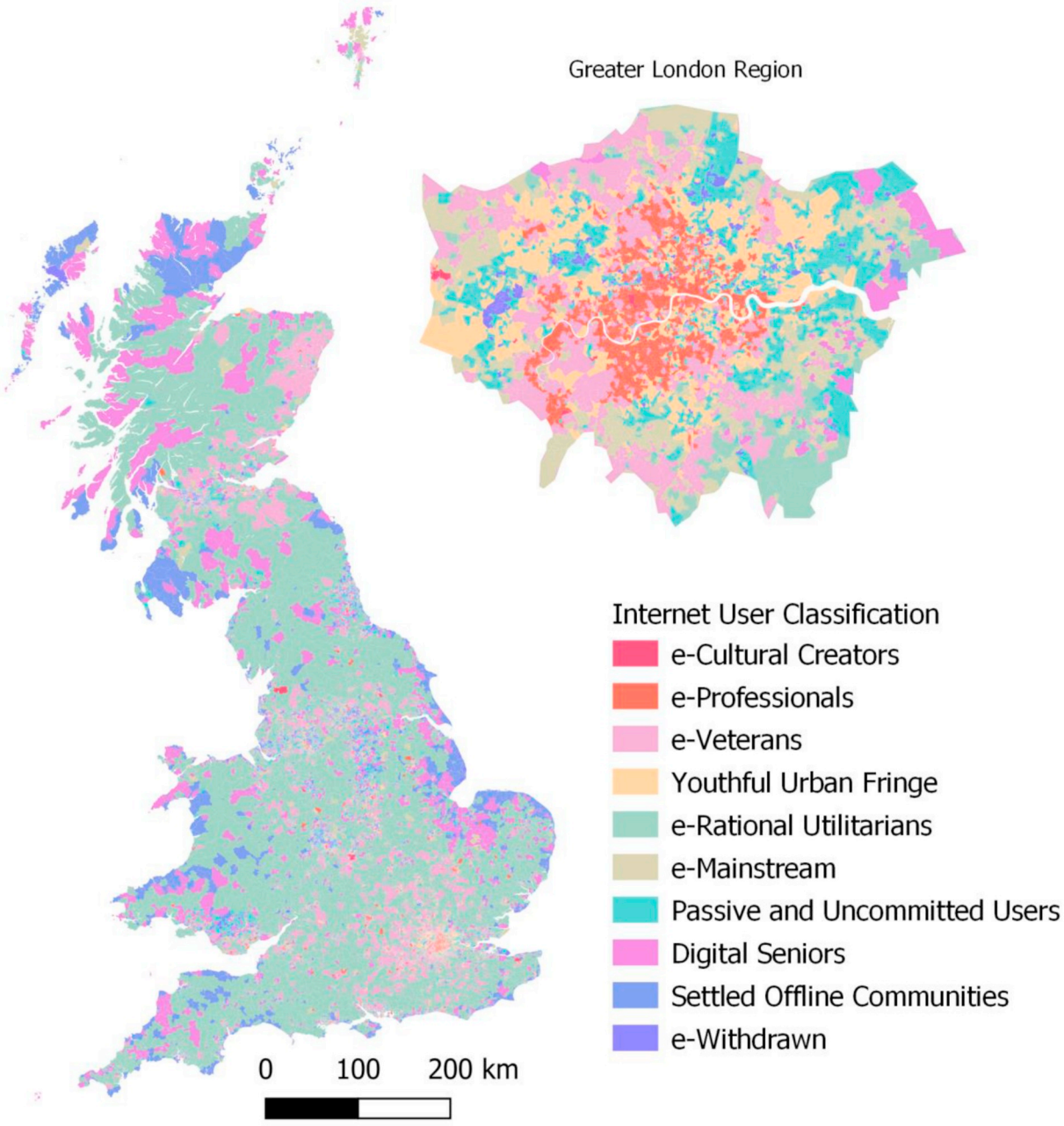

Fig. 7. A Map of the IUC for Great Britain.

(Spielman & Singleton, 2015). A secondary consideration was the extent of spatial information gained from a specific variable. For instance, "Do you engage in online gaming for money?" might offer some valuable information in the construction of a behavioural profile, but the geographic coverage of response was limited and, as such, offered limited differentiation between areas. All questions were either binary (e.g. "Do you buy groceries online?") or converted to binary (e.g. "Which method do you use to access the Internet?"). For other (ranked) questions, we merged responses to create larger aggregate groups based on response rates. For instance, in the case of the question "Which of these frequencies best describes your Internet use?", the values of "Around once a week", "2 or 3 times a month", "Around once a month" and "Less than around once a month" were recoded into "Once a week or less".

Because the aim was to gather a set of measures with coverage for all areas within GB rather than just a sample, simply pooling national survey data alone would not meet this objective. Out of the 42,729 LSOA/DZ within GB, about half of these contained no information; and several thousands more only contained a few responders. However, the volume of response was high enough that it was possible to produce estimated of rates of response for each small area using a methodological framework collectively known as Small Area Estimation (SAE). SAE techniques are diverse and well summarized elsewhere (Ghosh & Rao, 1994; Rahman, Harding, Tanton, & Liu, 2010; Rao, 2017), but these methods usually fall within two families of approach: spatial

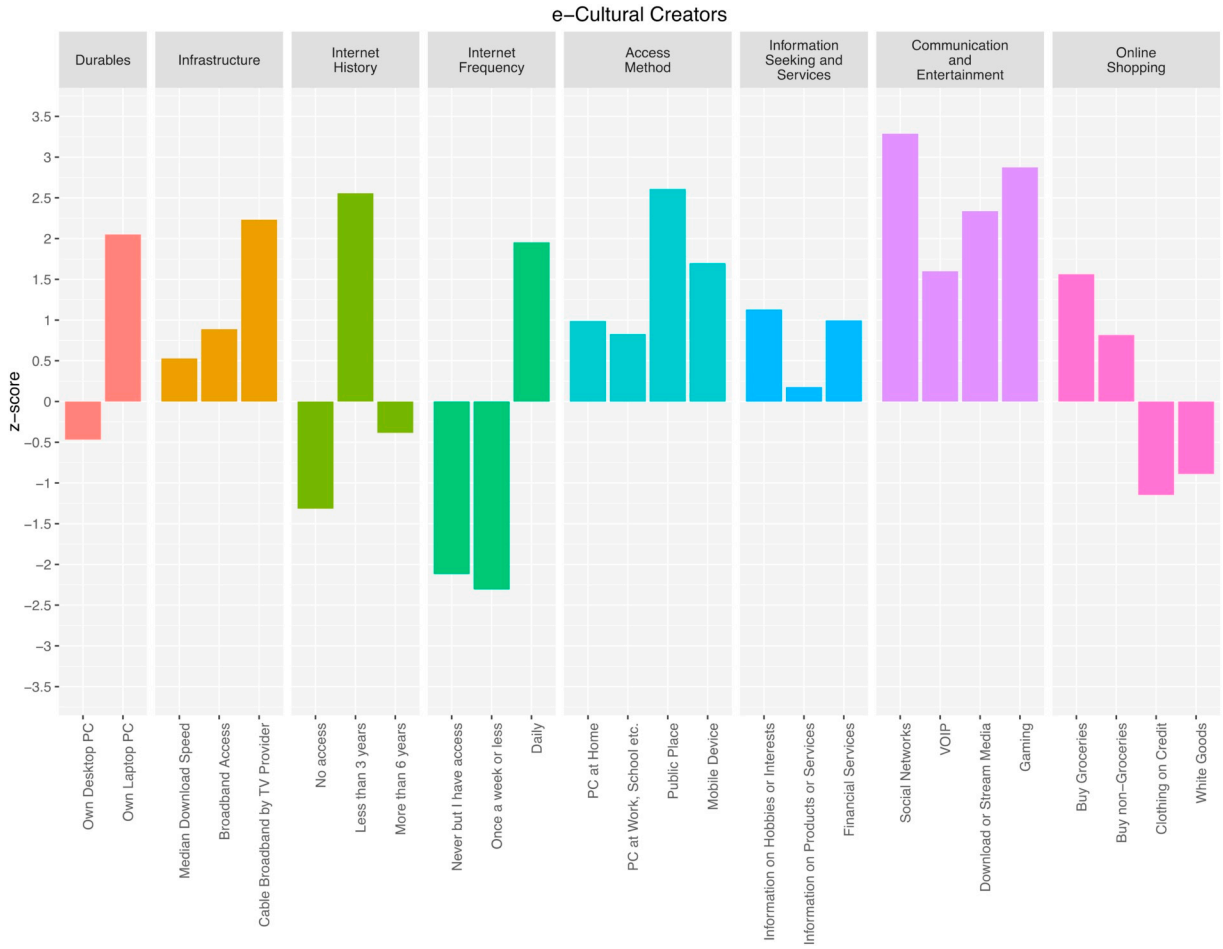

Fig. 8. Profiles of "e-Cultural Creators".

microsimulation and statistical or regression-based approaches. Spatial microsimulation approaches have the advantage of deriving synthetic microdata for each small area and not just point estimates (Lovelace & Dumont, 2016). However, they usually do not produce any measure of uncertainty regarding the estimation (Whitworth, Carter, Ballas, & Moon, 2017). Statistical or regression-based approaches to small-area estimation are trained on those areas where survey data are available. The resulting models are then applied to all small areas using a set of explanatory variables. These models do not offer distributional information and are more prone to (reverse) ecological fallacies, but have the advantage of being easier to model and to produce error estimates for the attributes under investigation.

However, in this instance we replace traditional statistical modelling with an alternative machine learning approach, namely a Gradient Boosting Regression Tree (GBRT) . GBRT is primarily a prediction model in the form of an ensemble of regression trees, where trees are employed as weak learners sequentially (i.e. every predictor "learns" from the previous one). GBRT models boast increased accuracy and flexibility; accounting for complex and non-linear relationships between variables (Ye, Chow, Chen, & Zheng, 2009). Given the recency of such methods, there are still relatively few studies on the integration of machine learning within SAE methodologies (Anderson, Guikema, Zaitchik, & Pan, 2014; Kontokosta, Hong, Johnson, & Starobin, 2018; Kriegler & Berk, 2010).

Before implementing SAE, it was necessary to adjust every LSOA/DZ BPS sample value to match target marginal distributions taken from the UK 2011 Census within GB. Features derived from the 2011 Census data were selected to be consistent with responses to socio-economic questions recorded within the BPS and included Age, Marital Status, Car Ownership, Number of Persons in Household, Children in Household, Qualification, Social Grade, Working Status, Household Tenure and Ethnicity. However, BPS categories were in some cases more or less detailed, and so required some recoding to follow those categories offered by the UK 2011 Census and vice-versa (Table 2).

The raw BPS survey data were reshaped so that categorical responses were converted to binary, and then aggregated into their associated LSOA/DZs. Since the survey sample's population characteristics did not necessarily correspond to the population characteristics of their associated zone, an Iterative Proportional Fitting Procedure (IPFP) was implemented to adjust the values. IPFP is an iterative procedure commonly used with the SAE microsimulation framework in order to reweight a micro-data file for each small area in accordance to a set of small area benchmarks derived from another source (Lomax & Norman, 2016). For most applications, several n benchmark characteristics are considered, and IPFP is applied to fill in cell values of a contingency table of n dimensions with known marginals. In this case, the IPFP was applied in order to obtain weights that "fit" the survey marginal

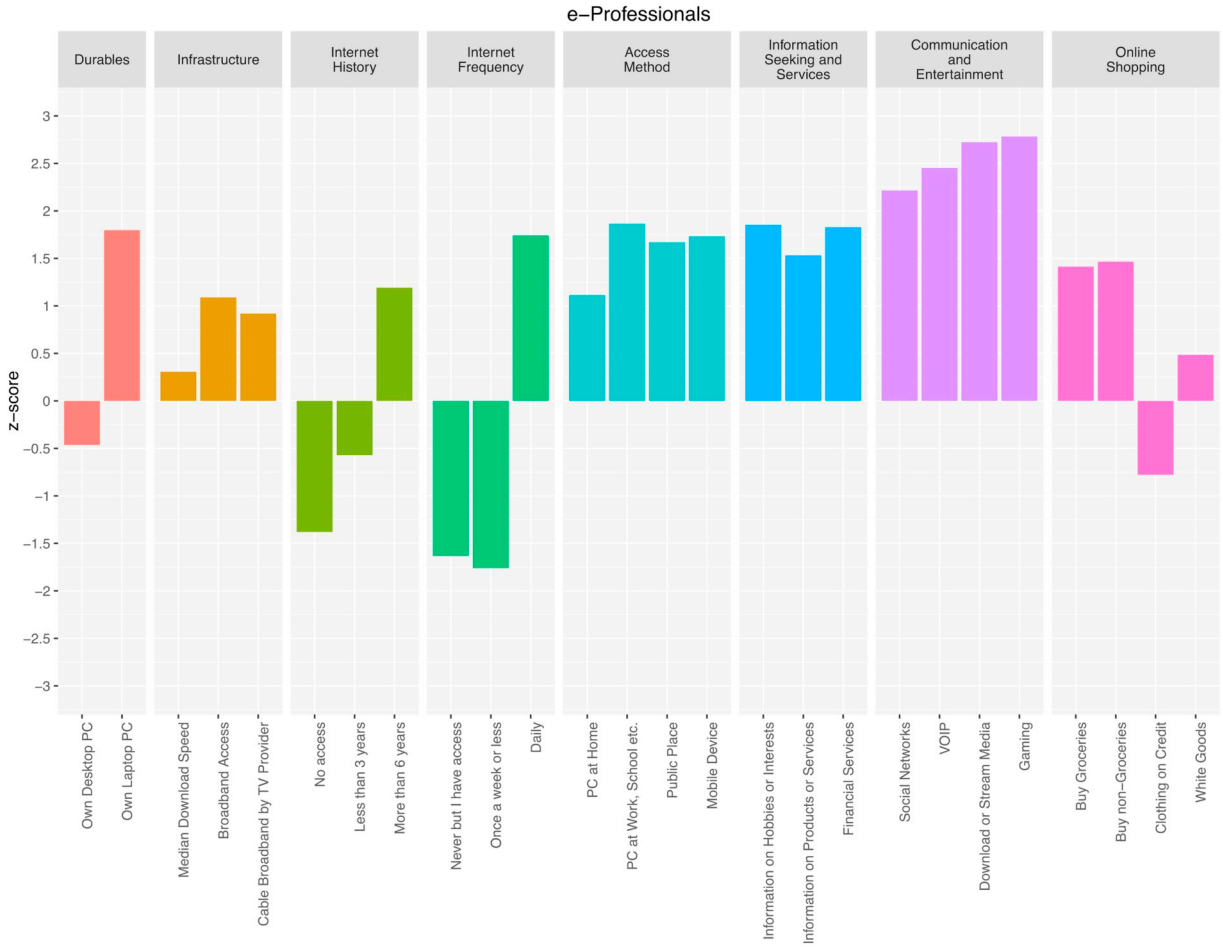

Fig. 9. Profiles of "e-professionals".

distributions to the UK 2011 Census marginal distributions within GB. Benchmark characteristics included Age Structure, Employment and Highest Qualifications categories as described in Table 2, which corresponded to a 7x9x3 contingency table. Adjusted LSOA/DZ ratios were then computed for each variable of interest, and provided base data that could be used to train supervised learning models that aimed to estimate rates of response for all areas.

2.3.1. *Predicting missing values with supervised machine learning*

Once the LSOA/DZ BPS survey sample data were adjusted by those weights created in the previous section, these data were then used to train a series of predictive models that produced feature estimates for areas with unknown values. All input data described ratios (with values from 0 to 1) for every binary value regarding an attribute of online behaviour. For this application, model training was carried out using the stochastic GBRT implementation "XGBoost" within the R programming language (Chen & Guestrin, 2016). The stochastic framework assumes that trees are trained on a randomly selected subset of the training data and thus are less prone to over-fitting. In order to make the estimations as robust as possible, only LSOA/DZs with 30 or more observations were used in the analysis, which amounted to 3222 LSOA/DZs. Models were trained using a tree booster for a maximum of 8000 rounds (i.e., the maximum number of trees) assigning a regression as the learning task and Root Mean Square Error (RMSE) as the evaluation metric. Since model overfitting (bias) is a known issue within gradient boosting models, the dataset was split randomly into a training and test dataset, using an 80% to 20% ratio as internal validation. One of the features of XGBoost is the capacity to follow the progress of the learning after each round and stop when necessary, as having too many rounds can lead to overfitting; the test dataset is utilised in this regard. Based on exploration, specific tuning of hyper-parameters for each model were used. In general, fine-tuning was performed with the learning rate (eta) ranging between 0.0005 and 0.001, the maximum depth of the tree subject between 6 and 11, and the minimum number of observations per terminal node (minimum child weight) ranged between 1 and 2. Observations were also weighted individually by log(N), where N was the number of observations within each LSOA/DZ (sample size). In this scenario LSOA/DZs that were more heavily sampled were given more weight in the prediction model. All models were also run with a subsample per tree subject of 0.8 and a column sample of 0.4 (i.e. only a random 80% of observations and 40% of variables are used in the construction of every tree, to reduce bias).

To illustrate results, a plot of predicted versus observed values reporting the ratio of response of people within LSOA/DZ who were identified within the BPS as using "Social Networks" (Fig. 2). Final R^2 values differed depending on the model, and ranged between 0.72 and

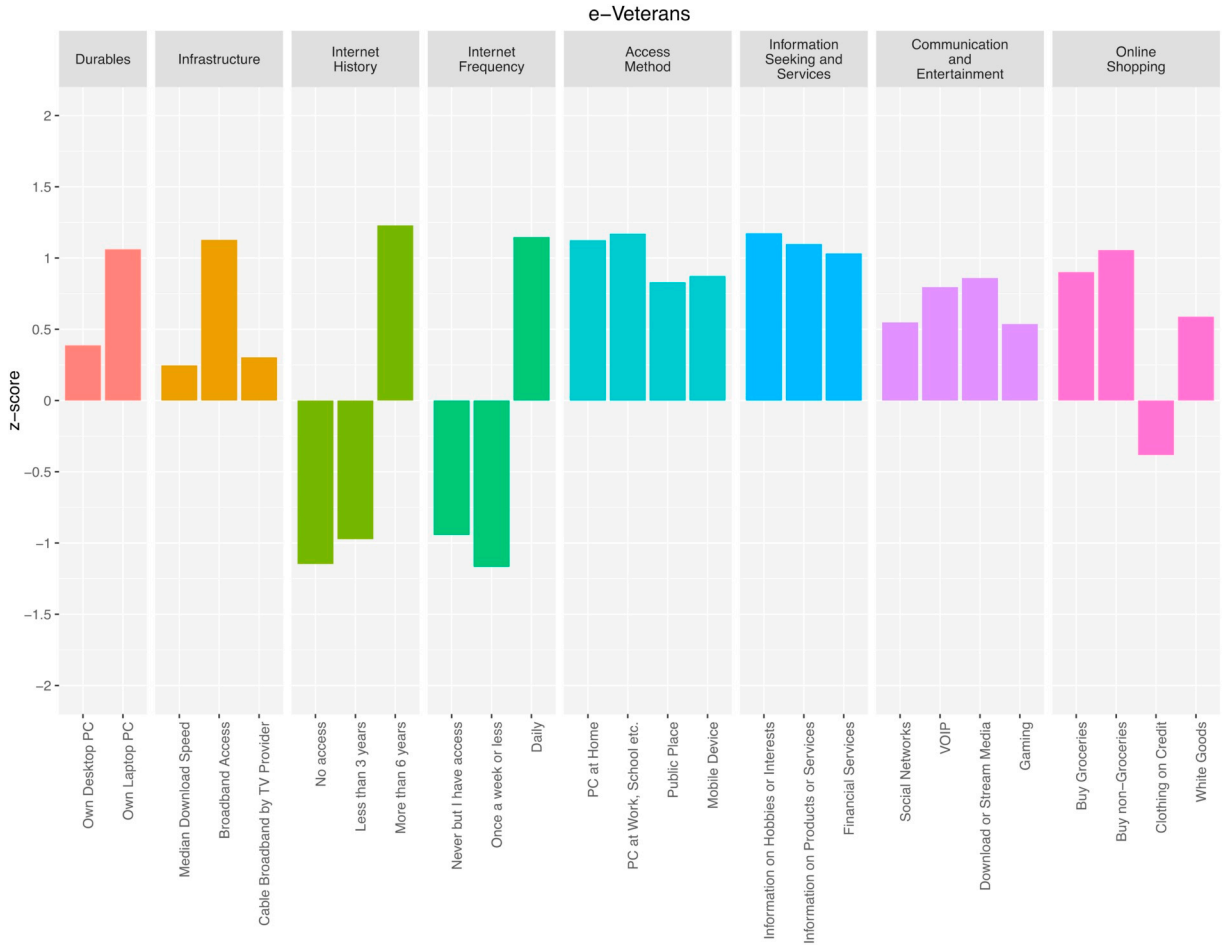

Fig. 10. Profiles of "e-Veterans".

0.89, while RMSE values ranged from 0.181 to 0.043, with feature "PC at Home" scoring the highest and "Cable Broadband by TV Provider" scoring the lowest. Instead of the coefficients associated with traditional regression, gradient boosting features relative importance; the more a variable is used to make key decisions in predicting behaviour, the higher its relative importance. Fig. 3 shows the relative importance of the question "Buy Groceries Online: Yes", scaled as such that the most important attribute equals 1. In this case, the variable that provides the most information on whether an individual buys groceries online is the *Retired* variable. Such outputs were evaluated during the parameter tuning process in order to identify potential model performance issues.

3. Using unsupervised machine learning to map digital differentiation

The outcome of the analysis presented in Section 2 was a set of small-area measures that described many of those different ways in which digital differentiation is manifest. Maps of these attributes might derive insight, but consolidating information from so many different variables and their geographic representationwould challenge human interpretability. Geodemographic classification is a computational technique that was introduced in the 1970s in both the US and UK with the explicit aim of rendering saliency from diverse sets of small-area measures (Singleton & Spielman, 2014). This technique draws lineage most directly from Social Area Analysis which was a quantitative method applied to urban areas in the 1950–60s (Timms, 1971); and also arguably, to much earlier qualitative work including Booth's poverty maps of London (Webber & Burrows, 2018). Although some geodemographic classification might be described as "black-box" (Singleton & Longley, 2009), and there are well founded critiques surrounding aspects of surveillance and the ascription of labels to populations (Dalton & Thatcher, 2015), there has also been much work to create more transparent classifications that are open to scrutiny (Gale, Singleton, Bates, & Longley, 2016; Singleton & Longley, 2019; Vickers & Rees, 2007). Geodemographic classification remain a widely applied technique in both the public and private sectors (Longley, 2005); with recent notable applications in health (Moon, Twigg, Jones, Aitken, & Taylor, 2019; Wami et al., 2019), education (Xiang, Stillwell, Burns, Heppenstall, & Norman, 2018) and the built environment (Alexiou, Singleton, & Longley, 2016).

The process of creating a Geodemographic classification tends to follow a reasonably standard process of applying a clustering technique (typically k-means or wards clustering) to group small areas based on feature similarity. For this application, inputs included all measures created in Section 2 and the values were standardised using *z*-scores

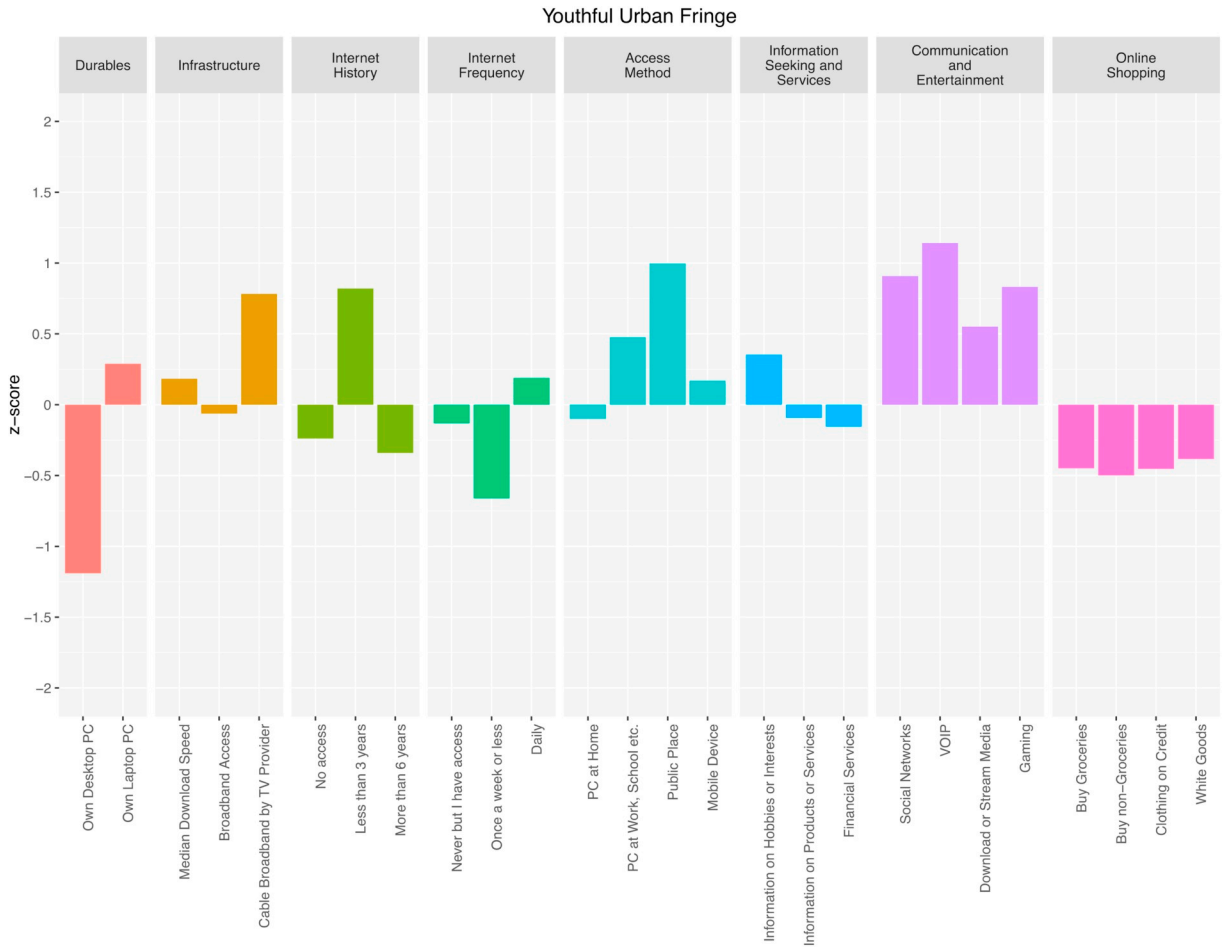

Fig. 11. Profiles of "Youthful Urban Fringe".

prior to clustering, with no normalization applied given that the distributions were already very close to normality.

The geodemographic was created as an initial set of larger clusters and were then considered in terms of whether further splits would provide additional meaningful groups. An exploration of the initial value of k (cluster frequency) was carried out using a Clustergram (Schonlau, 2002) which plots a series of potential k values alongside the weighted mean of their first principal component (Fig. 4). This visual method shows where clusters are well separated across the y axis and might represent a sensible number of clusters given the data input. While K = 5 and K = 7 both seemed plausible, K = 7 yielded a more reasonable result in terms of the nature of the cluster centres and the spatial patterns of cluster membership when mapped. This process was repeated for each of the data partitions identified by the initial clustering; three of the clusters were divided into two distinct groups, while the other four clusters remained, creating a total of ten clusters.

To better understand the robustness of the created cluster assignments, a set of statistics were calculated to contextualise the fit for each area. The input variables associated with every LSOA/DZ were compared with their assigned cluster mean values; and the squared difference of these scores returned. A higher score indicated an area with worse fit, as the area input values are positioned further from the cluster mean. These scores were mapped to examine if the error had systematic geographic bias; and were also plotted by cluster, to highlight potential reliability issues between the clusters. The results from these analysis are presented in Figs. 5 and 6. There is no particular systematic geographic bias to the error, with values reasonably random across both urban and rural areas. One exception is that the error is marginally higher within inner London (see Fig. 5 inset map), and in particular those areas lying north and along the River Thames . For a national geodemographic classification this is not unexpected given London's unique geography relative to other parts of GB; and a reason why bespoke geodemographic classifications have been created for this particular geographic extent (Singleton & Longley, 2015, 2019). Between cluster error was also reasonably evenly distributed (Fig. 6) with marginally higher error in cluster 8.

3.1. A National map of digital differentiation

To give saliency to the emergent clusters, these were provided with labels, written "pen portraits" and graphics. Such contextual details are essential to maximise the utility of a classification and aim to give an end user a clear indication of the main features of each cluster. As such, the cluster centres for each of the input variables were plotted to illustrate those attributes that were over or underrepresented within each cluster. These were considered alongside general socio-economic

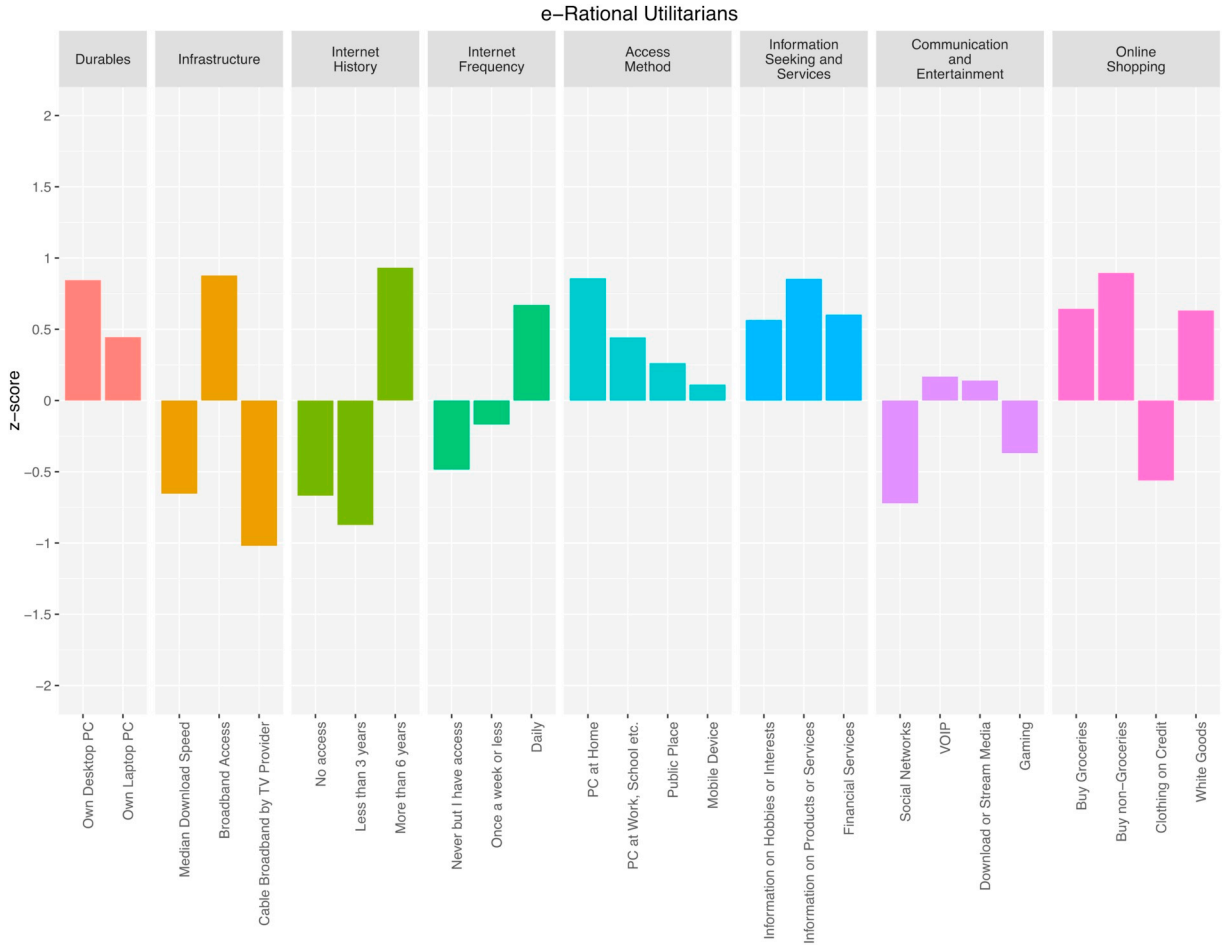

Fig. 12. Profiles of "e-Rational Utilitarians".

characteristics and maps and then used to assign labels and descriptions to each cluster, which are described in the remainder of this section but mapped for the national extent in Fig. 7. The clusters are presented broadly in the order of their relative rates of Internet engagement (high-low) and, going forward, we refer to the created typology as the Internet User Classification (IUC).

3.1.1. *e-Cultural Creators*

This Group was labelled as "e-Cultural Creators" on the basis of those attributes highlighted in Fig. 8. This Group has high levels of Internet engagement, particularly regarding social networks, communication, streaming and gaming, but relatively low levels of online shopping, besides groceries. They are new but very active users, with a very high proportion of the constituent population engaging on a daily basis. Their online behaviour can be explained by a demographic base that suggests a transitionary nature; the age structure of the Group is young, typically aged between 18 and 24, and with a strong presence of multicultural and student populations. They have a well-above average ownership of laptop devices, and an above average Internet access via mobile and in public places. Geographically, this Group is mainly located close to city centers or within the proximity of Higher Education Institutes, where infrastructure accessibility such as cable broadband is sufficient.

3.1.2. *e-Professionals*

The *e-Professionals* Group (Fig. 9) have high levels of Internet engagement, and comprise fairly young populations of urban professionals, typically aged between 25 and 34. They are experienced users and engage daily with the Internet and in a variety of settings. While communication and entertainment activities are very common, they tend to favour entertainment such as gaming more than social networks. They also carry out a significant portion of shopping activities online, particularly for non-groceries, and they use a variety of devices and methods to access the Internet. This Group is ethnically diverse, with a very strong representation of white, non-British populations. They are well-qualified and have very high availability of Internet at work. This Group tends to be found in residential areas abutting city centres or within affluent suburbs.

3.1.3. *e-Veterans*

The e-Veterans Group (Fig. 10) has a profile distribution that represents affluent families, is usually located within low-density suburbs, with populations of mainly middle-aged and highly qualified professionals. They are more likely to be frequent and experienced users of the Internet, having the second highest levels of Internet access at work after *e-Professionals* users. They engage with the Internet using multiple devices and in a variety of ways. They are populations of fairly mature users and, as such, they have higher levels of engagement for

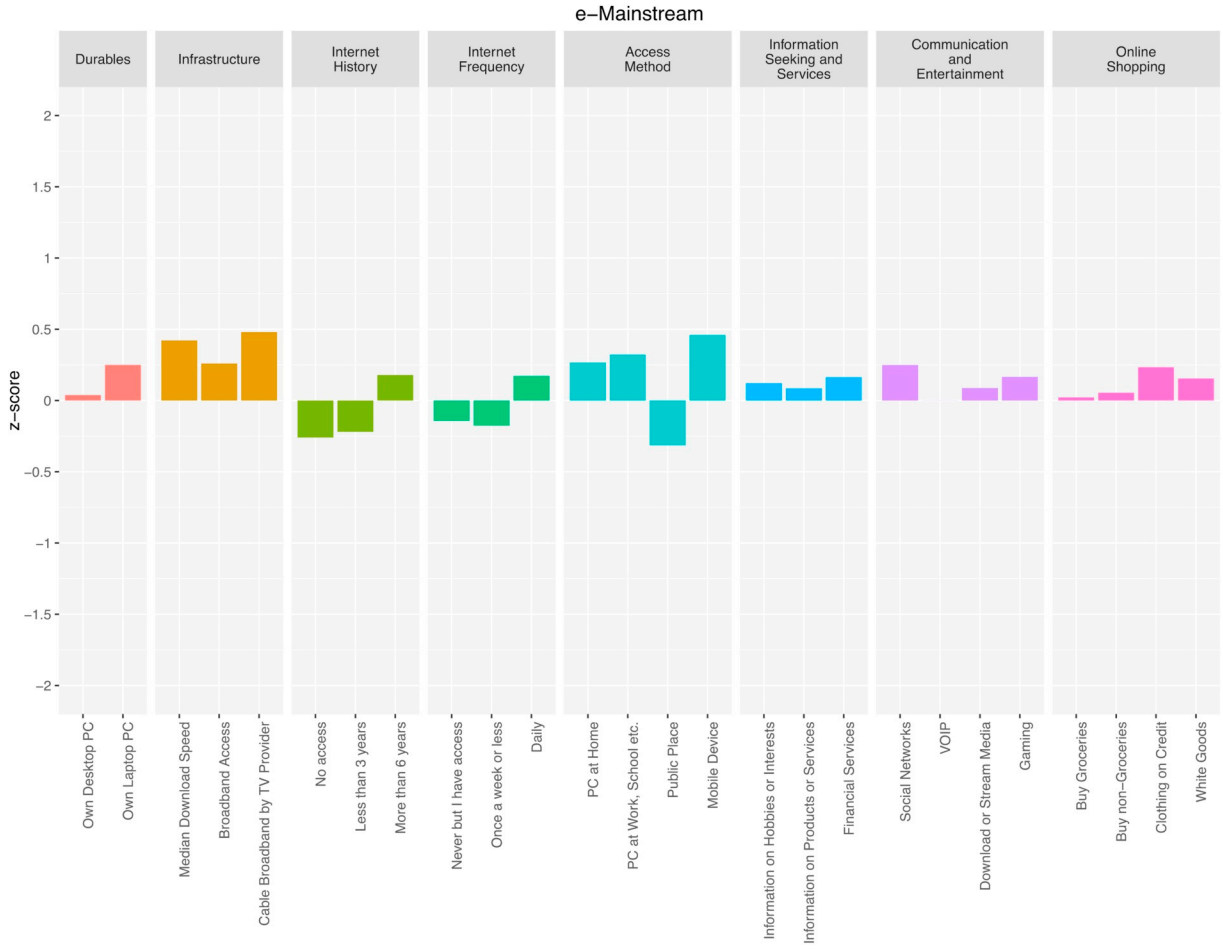

Fig. 13. Profiles of "e-Mainstream".

information seeking, online services and shopping, but relatively less so for communication and entertainment, particularly social networks or gaming.

3.1.4. *Youthful Urban Fringe*

This Group often resides at the edge of city centres, are young and are drawn from ethnic minorities. The populations typically include a mixture of students and other young urbanites living in informal households, often at the edges of materially deprived communities. Access through desktop devices is particularly low (Fig. 11), suggesting a young and mobile profile of individuals. Access to broadband is average, possibly due to other modes of access taking priority, such as Internet usage in public places. The levels of Internet engagement are fairly average, with high levels of social media use but low patronage of online retailing.

3.1.5. *e-ational utilitarians*

This Group mainly consists of rural and semi-rural areas at the city fringe. High demand for Internet services by members of this Group is often constrained by poor infrastructure (Fig. 12). Users undertake online shopping, particularly for groceries, perhaps because of the limited offer from "bricks and mortar" retailers. Users tend to be late middle-aged or elderly, and as might be expected, include a high percentage of retired home owners. The preferred method of engagement with the Internet is through personal computers located at home, with low levels of mobile access. In addition to shopping, users search for information or access online banking rather than engage with social networks or gaming: the Internet is used as a utility rather than a conduit for entertainment.

3.1.6. *e-Mainstream*

This Group exhibits average Internet user characteristics (Fig. 13) and comprises populations drawn from a wide range of social echelons as defined using conventional socioeconomic data, and mostly represent heterogeneous neighbourhoods. Geographically, the Group is usually located at the periphery of urban areas.

3.1.7. *Passive and Uncommitted Users*

Fig. 14 shows that within this Group many individuals have limited or no interaction with the Internet. They tend to reside outside city

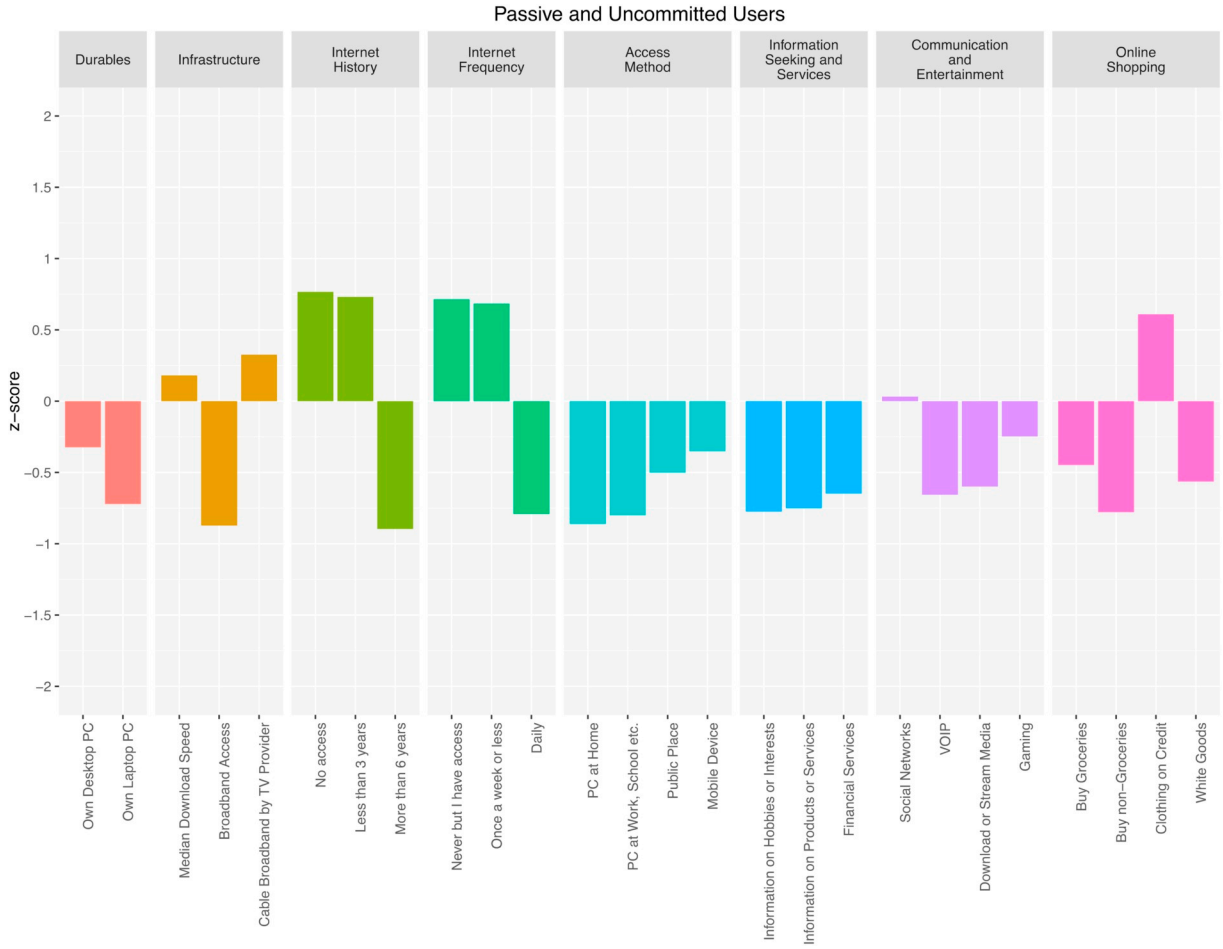

Fig. 14. Profiles of "Passive and Uncommitted Users".

centres and close to the suburbs or semi-rural areas. Members of this Group have few distinctive characteristics in conventional socio-economic terms, albeit higher levels of employment in semi-skilled and blue-collar occupations. Individuals are rarely online, and most commonly report use once a week or less. Access to broadband is well below average, and for those online, there is mild preference for access via smartphones. The Internet is typically used for social networks, gaming and some limited online shopping.

3.1.8. *Digital Seniors*

Members of this Group are ageing and predominantly White British, retired and relatively affluent. They make average use of the Internet (Fig. 15), typically using a personal computer at home. Despite being infrequent users, they are adept enough to use the Internet for information seeking, financial services and online shopping, but less so for social networks, streaming or gaming. Members of this Group typically reside in semi-rural or coastal regions, where infrastructure provision is often limited.

3.1.9. *Settled Offline Communities*

Most members of this Group are elderly, White British and retired, and tend to reside in semi-rural areas. They undertake only limited engagement with the Internet (Fig. 16), they may only have rare access or, indeed, no access to it at all. Any online behaviour tends to be through home computers rather than mobile devices, and is focused upon information seeking and limited online shopping (particularly for more bulky items such as white goods) rather than social networking, gaming or media streaming.

3.1.10. *e-Withdrawn*

This Group is mainly characterised by individuals who are the least engaged with the Internet (Fig. 17). Their geography is expressed by areas that are associated with those more deprived neighbourhoods of urban regions. The socio-economic profile of the population is characterised by less affluent white British individuals or areas of high ethnic diversity; and it has the highest rate of unemployment and social housing among all other Groups. The *e-Withdrawn* Group appears to have the highest ratio of people that don't have access, or have access but never engage with the Internet. It also expresses the lowest rates of engagement in terms of information seeking and financial services, as well as the lowest rate in terms of online access via a mobile device. Rates of online shopping are also particularly low, with the exception of Clothing on Credit, suggesting an opportunistic dimension to Internet usage. This is further reinforced by the higher than average access to

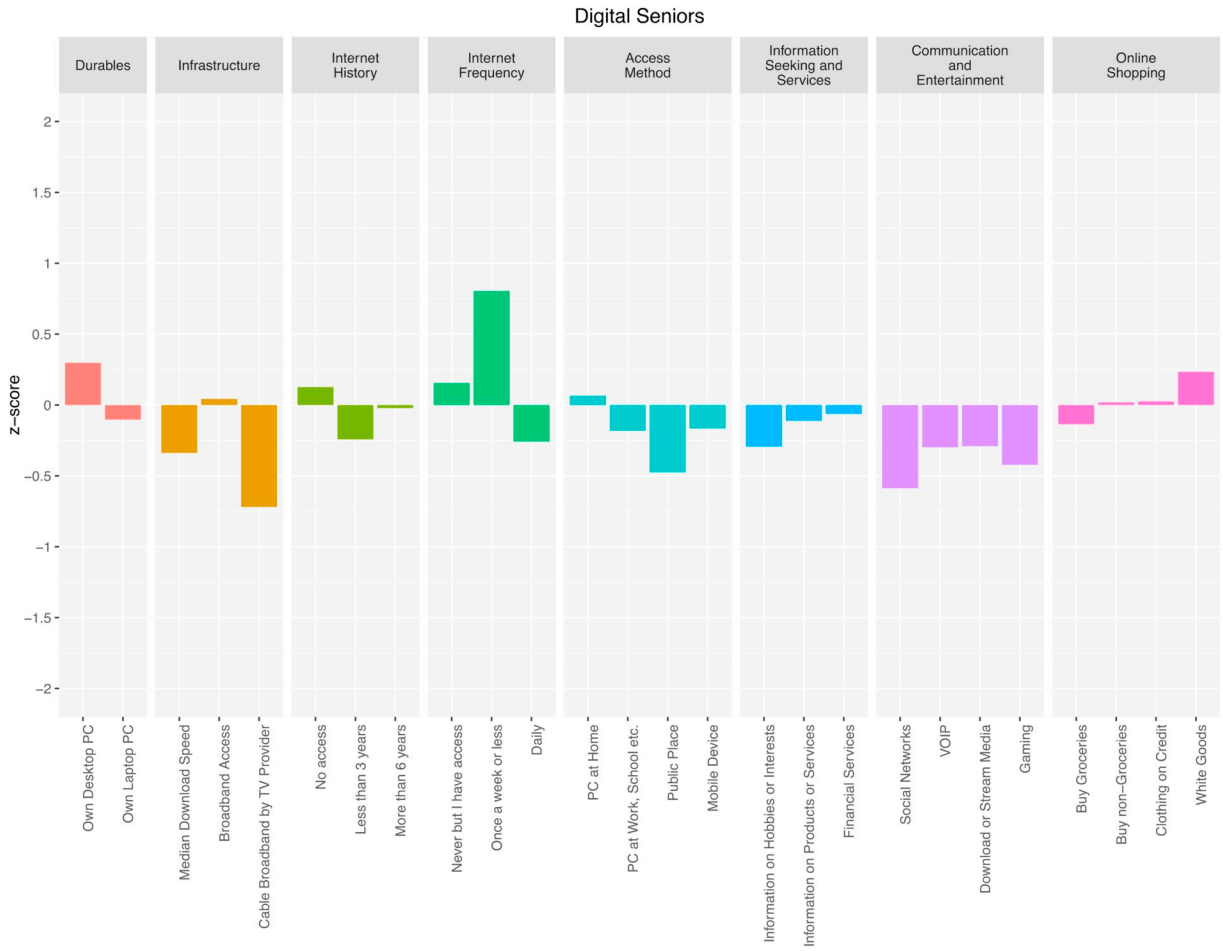

Fig. 15. Profiles of "Digital Seniors".

Cable broadband by TV Provider, which may suggest that some individuals have opted into broadband mainly for the TV-associated benefits. It is possible that many people within this Group have opted out of online engagement, either because it is considered unnecessary or because of economic reasons.

4. External evaluation: digital differentiation and 2021 UK census online response rates

A useful method of providing external validation to a new geodemographic classification is to examine its utility within the context of a focused case study. In this illustrative example, we explore how those insights into the geography of digital differentiation presented by the IUC might be applied to a real-world case study exploring potential patterns of response to a national Census. Within the UK, the ONS have been conducting tests to refine their delivery methodology that will be implementing for the 2021 UK Census; which will be conducted predominantly online for the first time. Digital differentiation is therefore an obvious policy concern within this context, as a lack of response from those who are less engaged could have a significant impact upon the quality of those estimates that are produced.

In 2017, an extensive Census test was conducted, involving a total of 208,000 households across England and Wales. This comprised a series of components, with invitations being sent to 100,000 households located within seven local authority areas which are shown in Fig. 18, alongside a further 108,000 randomly selected households across the rest of England and Wales. Full details of the sampling strategy can be found within the ONS documentation,[4] alongside LSOA level aggregated data pertaining to the response rates within these locations, and splits between the proportion of people completing online and offline versions of the Census form. There were numerous objectives of the test, but one relevant to this study/research was to develop insight into rates of online take-up, and those characteristics of areas and respondents who will or can only respond on paper.

We first mapped the IUC within the survey location areas (Fig. 18), which illustrated a variable Internet user geography. These areas also represented a variety of different urban-rural settings, with Powys, South Somerset and West Dorset being large and predominant rural areas, Barnsley and Sheffield being more mixed, and Southwark being densely populated and within Central London.

If effective, the IUC should differentiate between areas where online response rates are likely to be lower (i.e. lower Internet engagement). After appending response rates to the IUC, the relationship broadly follows what might be expected; with the more Internet engaged groups (left side of the chart in Fig. 19) showing higher rates of online survey completion.

Further insights are revealed when we examine the overall rates of completion (online and paper).[5](Fig. 20). The lowest rates of response

[4] https://www.ons.gov.uk/census/censustransformationprogramme/2017censustest/2017censustestreport

[5] The ONS derive response rate as a percentage of valid Census Test household returns made, divided by the valid household sample for each LSOA; A valid household return is one where there is enough information submitted to denote at least one person is normally resident at the address.

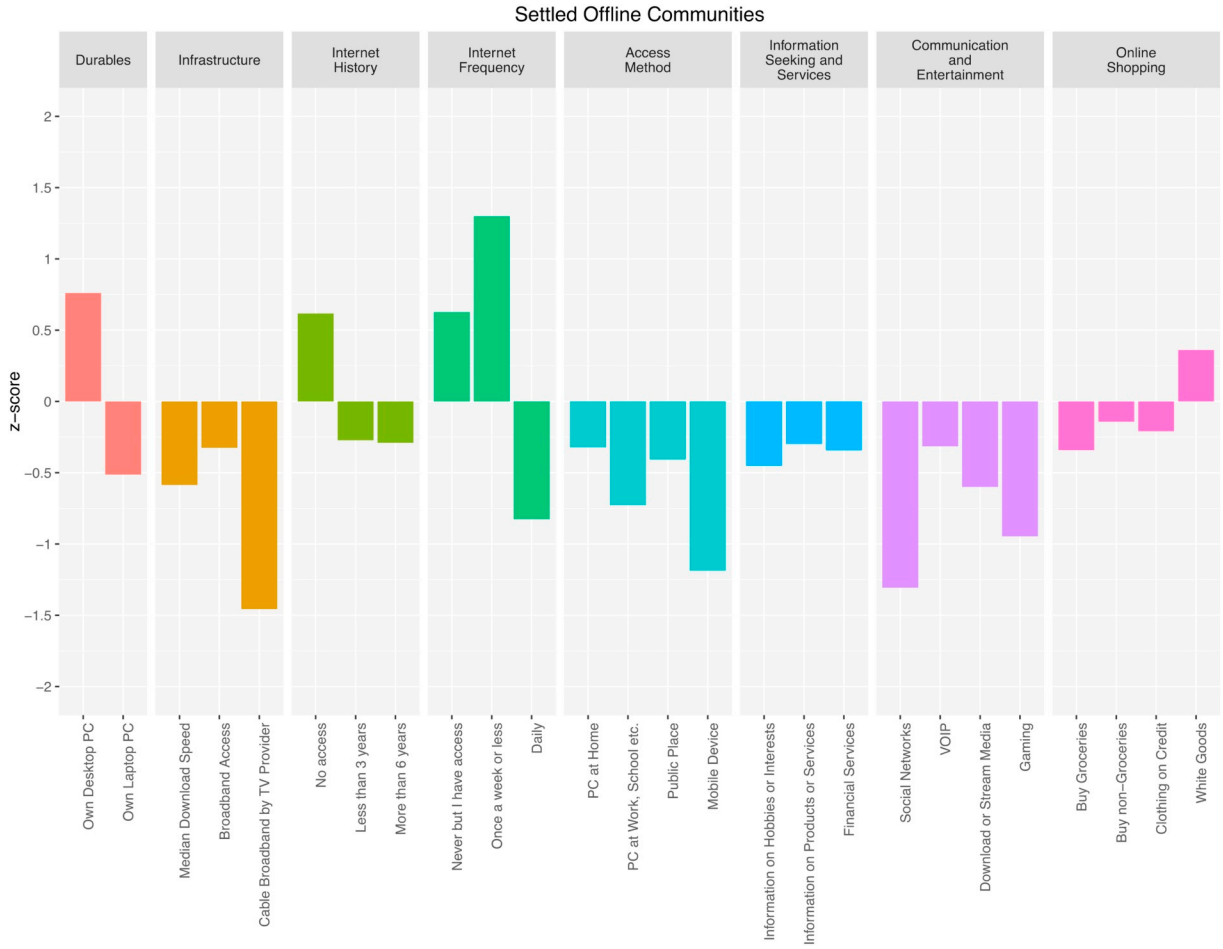

Fig. 16. Profiles of "Settled Offline Communities".

fall within the "e-Cultural Creators" group and also the "Youthful Urban Fringe"; which although were both likely to complete surveys online, were less likely to respond in general. In both these clusters, there is a greater prevalence for residents aged 18–24 and areas being of higher ethnic diversity. For those living within areas identified by the former Group, there is a very high propensity for residents to be studying full time; whereas in "Youthful Urban Fringe" these areas have lower employment, but also correspondingly lower qualifications. Such insights might be used prospectively to develop appropriate strategies to improve response rates from people living within areas identified by these groups.

More generally, the approach adopted here of building a multi-dimensional classification, rather than ranking areas by monotopical factors (e.g. access or use of the Internet) is emphasised; in that richer and discursive profiles can be generated that guide further insights, explorations or response. For example, given that the "e-Cultural Creators" and "Youthful Urban Fringe" Groups both have higher Internet engagement, they may be more responsive to online strategies that improve completion rates; although given their different economic circumstances, such messaging may warrant further differentiation. We also see one of the communities who are least engaged online "Settled Offline Communities" producing one of the higher rates of overall response. Such differences indicate areas where online completion would generate worse returns over paper-based forms, and it is apparent that the IUC might provide a useful tool through which such necessary differentiation could be understood.

5. Discussion and conclusion

The objective of this paper was to map those ways in which digital inequality is differentiated between small areas across GB. There are a range of consequences of digital inequalities for both individuals and households; including, but not limited to, access to goods and service, civic engagement / participation, and development of skills, alongside social and economic mobility. For the public sector, where services are increasingly delivered online, such differences can result in variable take-up of opportunities and negative social impact. Many services offered by Local Councils are predominantly carried out online, and online services offered by the UK National Health Service (NHS) are becoming increasingly important for the everyday life of individuals

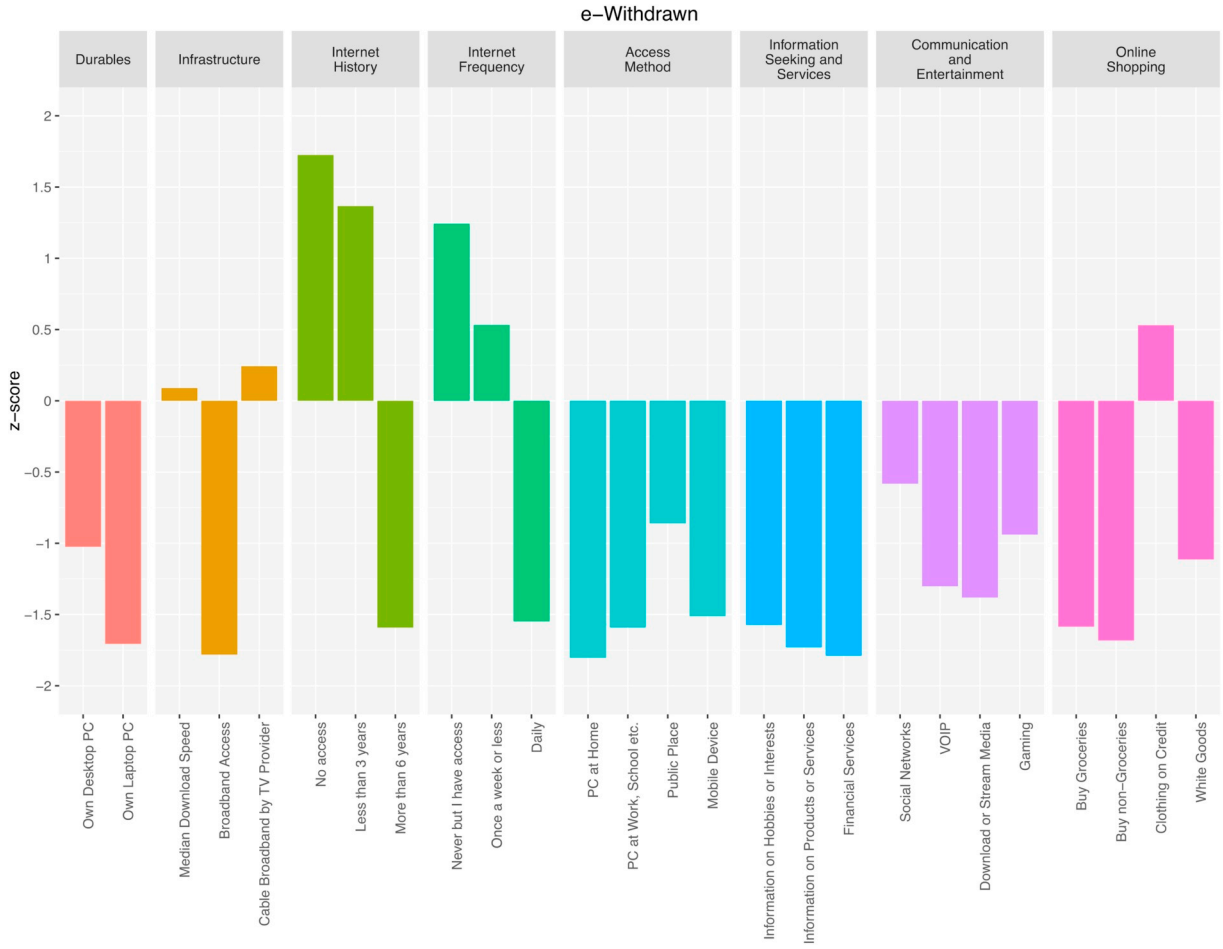

Fig. 17. Profiles of "e-Withdrawn".

(Honeyman, Dunn, & McKenna, 2016). The practicalities associated with the Internet may contribute to the tendency for some groups of people and areas to "lag behind", leading to new forms of not only digital but also social exclusion.

Following previous work developing compound and multi-dimensional indicators of digital inequality (Barzilai-Nahon, 2006; Borg & Smith, 2018; Vehovar et al., 2006), our approach takes a similarly holistic view through the development of a new national geodemographic classification that benefits from the integration of a range of topical consumer data supplied by the ESRC Consumer Data Research Centre, alongside demographic measures and other open data. We extract a range of measures from these data that were guided by the literature, and described the geography of access to the Internet, how this use is differentiated through various types of online activities, and how the local context of access vary spatially. Survey data were utilised extensively in the creation of behavioural measures around Internet use and required the application of small area estimation to derive local response estimates. A methodological contribution of this work beyond the substantive findings was in supplement of a standard inferential model within a small-area estimation framework for gradient boosting, which is a supervised machine learning algorithm. Through this integration, we were able to produce a series of estimates for LSOA/DZs concerning Internet user behaviours from a nationally representative survey. The assembly of behavioural and contextual measures describing those different ways in which digital inequality manifests were integrated within a geodemographic framework to create a composite but multidimensional categorical indicator. This typology was composed of ten distinctive groups representing salient patterns of digital differentiation that were mapped at the LSOA/DZ spatial scale for the extent of GB.

Internal and external evaluation revealed the robustness and utility of the Internet User Classification . Through consideration of the ONS 2017 Census test results we provided a set of profiles that considered both the geodemographics of survey completion online, but also overall response rate. Through such profiling the advantages of having a multidimensional indicator were highlighted as, by consideration of the similarities and differences between clusters, hypothesis could be generated that may warrant further investigation, or could be used in a practical sense to help formulate differentiated methodological interventions.

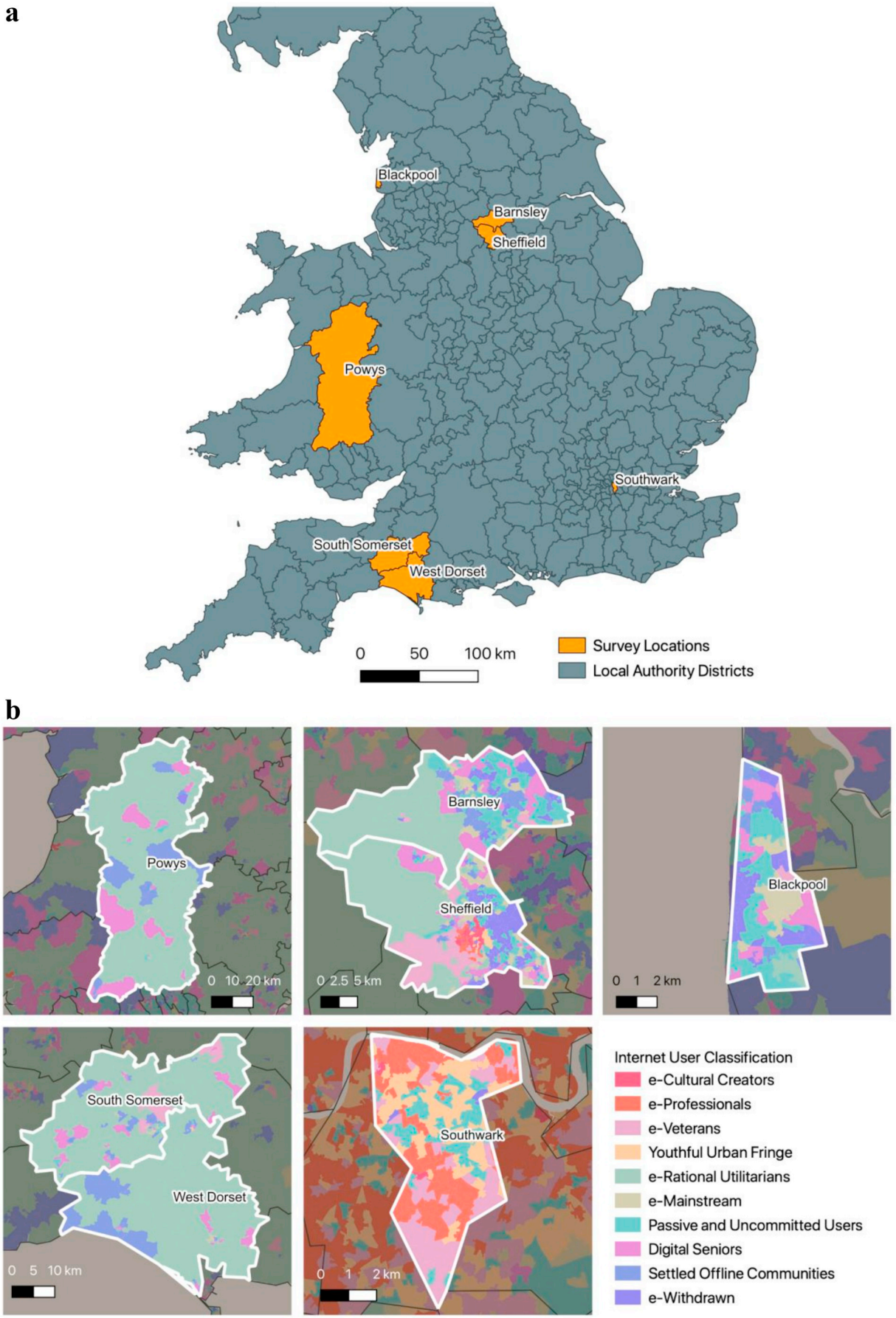

Fig. 18. 2017 Test area locations and IUC profiles.

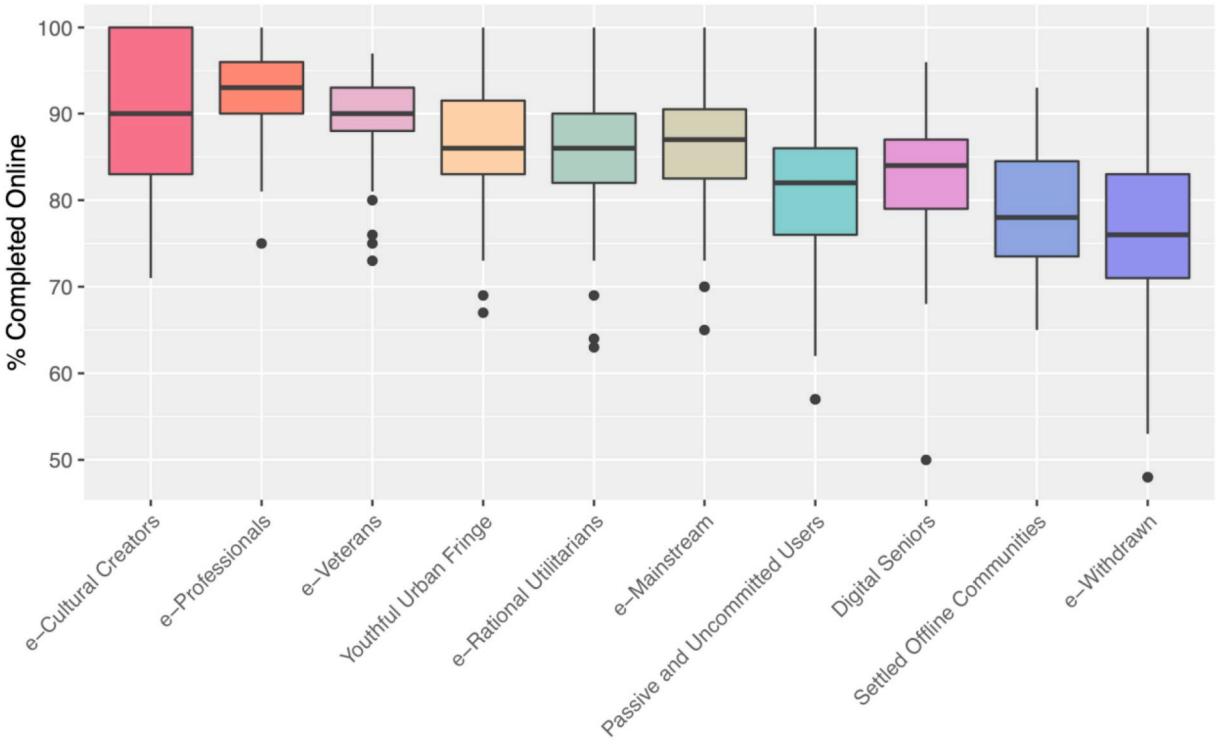

Fig. 19. Box and Whisker plot of the online completion rates by IUC groups.

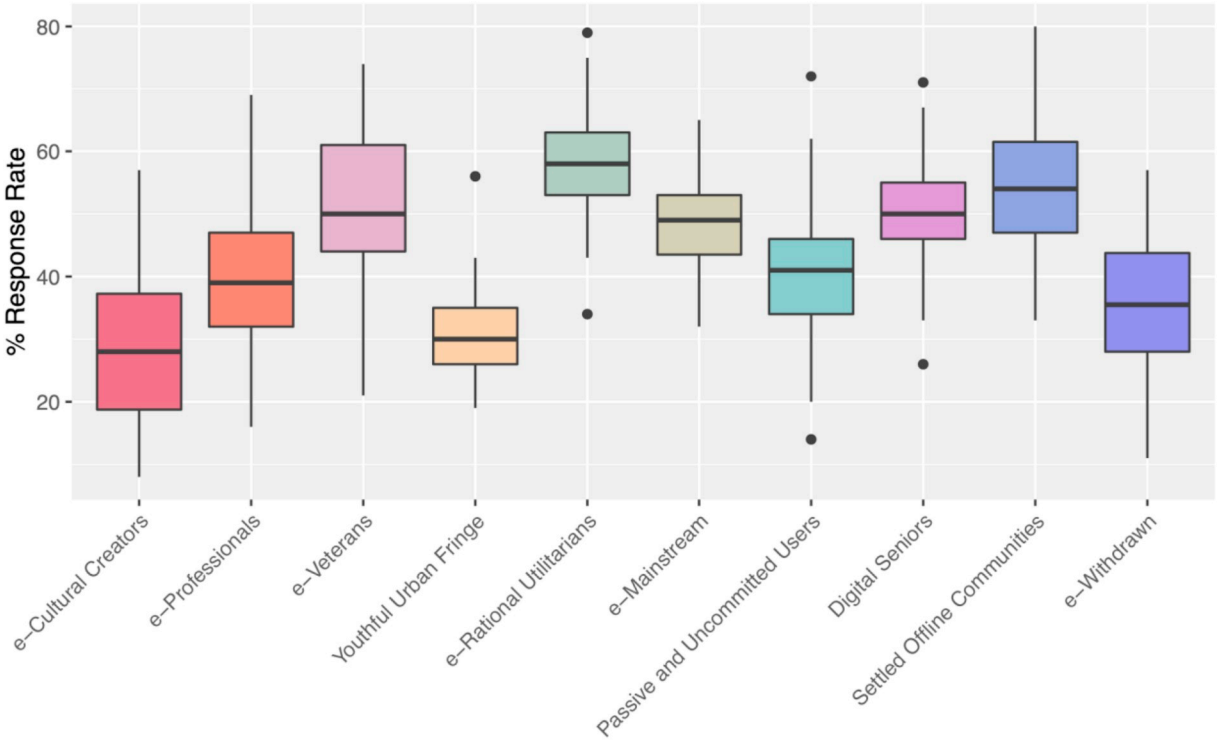

Fig. 20. Box and Whisker plot of completion rates by IUC groups.

Acknowledgments

This work was funded by the UK Economic and Social Research Council, grant number: ES/L011840/1.

References

Alexiou, A., Singleton, A. D., & Longley, P. A. (2016). A classification of multidimensional open data for urban morphology. *Built Environment*. https://doi.org/10.2148/benv.42.3.382.

Anderson, W., Guikema, S., Zaitchik, B., & Pan, W. (2014). Methods for estimating population density in data-limited areas: Evaluating regression and tree-based models in Peru. *PLoS One*. https://doi.org/10.1371/journal.pone.0100037.

Barzilai-Nahon, K. (2006). Gaps and bits: Conceptualizing measurements for digital divide/s. *The Information Society*. https://doi.org/10.1080/01972240600903953.

Blank, G., Graham, M., & Calvino, C. (2018). Local geographies of digital inequality. *Social Science Computer Review*. https://doi.org/10.1177/0894439317693332.

Borg, K., Boulet, M., Smith, L., & Bragge, P. (2018). Digital inclusion & health communication: A rapid review of literature. *Health Communication, 0*(0), 1–9. https://doi.org/10.1080/10410236.2018.1485077.

Borg, K., & Smith, L. (2018). Digital inclusion and online behaviour: Five typologies of Australian internet users. *Behaviour & Information Technology*. https://doi.org/10.1080/0144929X.2018.1436593.

Büchi, M., Just, N., & Latzer, M. (2015). Modeling the second-level digital divide: A five-country study of social differences in internet use. *New Media & Society, 18*(11), 2703–2722. https://doi.org/10.1177/1461444815604154.

Chang, J., McAllister, C., & McCaslin, R. (2015). Correlates of, and barriers to, Internet use among older adults. *Journal of Gerontological Social Work*. https://doi.org/10.1080/01634372.2014.913754.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*https://doi.org/10.1145/2939672.2939785.

Dalton, C. M., & Thatcher, J. (2015). Inflated granularity: Spatial "big data" and geodemographics. *Big Data & Society, 2*(2), https://doi.org/10.1177/2053951715601144.

Dolega, L., Reynolds, J., Singleton, A. D., & Pavlis, M. (2019). Beyond retail: New ways of classifying UK shopping and consumption spaces. *Environment and Planning B: Urban Analytics and City Science.*. https://doi.org/10.1177/2399808319840666.

Friemel, T. N. (2016). The digital divide has grown old: Determinants of a digital divide among seniors. *New Media & Society*. https://doi.org/10.1177/1461444814538648.

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*. https://doi.org/10.5311/JOSIS.2016.12.232.

Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science, 9*(1), 55–76. https://doi.org/10.1214/ss/1177010647.

Gonzales, A. (2016). The contemporary US digital divide: From initial access to technology maintenance. *Information, Communication & Society*. https://doi.org/10.1080/1369118X.2015.1050438.

Grubesic, T. H., Helderop, E., & Alizadeh, T. (2018). Closing information asymmetries: A scale agnostic approach for exploring equity implications of broadband provision. *Telecommunications Policy*. https://doi.org/10.1016/j.telpol.2018.04.002.

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

Honeyman, M., Dunn, D., & McKenna, H. (2016). A digital NHS?: An introduction to the digital agenda and plans for implementation. *The Kings Fund*.

Hunsaker, A., & Hargittai, E. (2018). A review of internet use among older adults. *New Media & Society*. https://doi.org/10.1177/1461444818787348.

Inkinen, T., Merisalo, M., & Makkonen, T. (2018). Variations in the adoption and willingness to use e-services in three differentiated urban areas. *European Planning Studies, 26*(5), 950–968. https://doi.org/10.1080/09654313.2018.1448756.

Kontokosta, C. E., Hong, B., Johnson, N. E., & Starobin, D. (2018). Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. *Computers, Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2018.03.004.

Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*. Doi:10.18637/jss.v074.i07.

Kriegler, B., & Berk, R. (2010). Small area estimation of the homeless in los Angeles: An application of cost-sensitive stochastic gradient boosting. *Annals of Applied Statistics*. https://doi.org/10.1214/10-AOAS328.

Lissitsa, S., & Kol, O. (2016). Generation X vs. Generation Y - A decade of online shopping. *Journal of Retailing and Consumer Services*. https://doi.org/10.1016/j.jretconser.2016.04.015.

Lomax, N., & Norman, P. (2016). Estimating population attribute values in a table: "Get me started in" iterative proportional fitting. *The Professional Geographer*. https://doi.org/10.1080/00330124.2015.1099449.

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

Longley, P. A., & Singleton, A. D. (2009). Linking social deprivation and digital exclusion in England. *Urban Studies, 46*(7), 1275–1298. https://doi.org/10.1177/0042098009104566.

Lovelace, R., & Dumont, M. (2016). *Spatial microsimulation with R*. CRC Press.

Marler, W. (2018). Mobile phones and inequality: Findings, trends, and future directions. *New Media & Society, 0*(0), https://doi.org/10.1177/1461444818765154.

Mesch, G. S. (2015). Ethnic origin and access to electronic health services. *Health Informatics Journal, 22*(4), 791–803. https://doi.org/10.1177/1460458215590863.

Moon, G., Twigg, L., Jones, K., Aitken, G., & Taylor, J. (2019). The utility of geodemographic indicators in small area estimates of limiting long-term illness. *Social Science and Medicine*. https://doi.org/10.1016/j.socscimed.2018.06.029.

Peng, G. (2017). Do computer skills affect worker employment? An empirical study from CPS surveys. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2017.04.013.

Rahman, A., Harding, A., Tanton, R., & Liu, S. (2010). Methodological issues in spatial microsimulation modelling for small area estimation. *International Journal of Microsimulation, 3*(2), 3–22.

Rao, J. N. K. (2017). Small-area estimation. *Wiley StatsRef: Statistics Reference Online* (pp. 1–8). Doi:https://doi.org/10.1002/9781118445112.stat03310.pub2.

Schonlau, M. (2002). The Clustergram: A graph for visualizing hierarchical and non-hierarchical cluster analyses. *The Stata Journal: Promoting Communications on Statistics and Stata*. https://doi.org/10.1177/1536867x0200200405.

Singleton, A. D., Dolega, L., Riddlesden, D., & Longley, P. A. (2016). Measuring the spatial vulnerability of retail centres to online consumption through a framework of e-resilience. *Geoforum, 69*, 5–18.

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. A. (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. D., & Longley, P. A. (2019). Data infrastructure requirements for new geodemographic classifications: The example of London's workplace zones. *Applied Geography, 109*, 102038. https://doi.org/10.1016/j.apgeog.2019.102038.

Singleton, A. D., & Spielman, S. E. (2014). The past, present, and future of geodemographic research in the United States and United Kingdom. *The Professional Geographer, 66*(4), 558–567.

Spielman, S. E., & Singleton, A. D. (2015). Studying Neighborhoods using uncertain data from the American community survey: A contextual approach. *Annals of the Association of American Geographers, 105*(5), 1003–1025. https://doi.org/10.1080/00045608.2015.1052335.

Szeles, M. R. (2018). New insights from a multilevel approach to the regional digital divide in the European Union. *Telecommunications Policy, 42*(6), 452–463. https://doi.org/10.1016/j.telpol.2018.03.007.

Thomas, J., Barraket, J., Wilson, C., Ewing, S., MacDonald, T., Tucker, J., & Rennie, E. (2017). *Measuring Australia's digital divide: The Australian digital inclusion index,* 2017. https://doi.org/10.4225/50/59b762ab75714.

Timms, D. (1971). *The urban mosaic: Towards a theory of residential differentiation*. Cambridge University Press.

Tsetsi, E., & Rains, S. A. (2017). Smartphone internet access and use: Extending the digital divide and usage gap. *Mobile Media and Communication.*. https://doi.org/10.1177/2050157917708329.

Van Dijk, J., & Hacker, K. (2003). The digital divide as a complex and dynamic phenomenon. *In Information Society.*. https://doi.org/10.1080/01972240309487.

Vehovar, V., Sicherl, P., Hüsing, T., & Dolnicar, V. (2006). Methodological challenges of digital divide measurements. *The Information Society.*. https://doi.org/10.1080/01972240600904076.

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.

Wami, W. M., Dundas, R., Molaodi, O. R., Tranter, M., Leyland, A. H., & Katikireddi, S. V. (2019). Assessing the potential utility of commercial "big data" for health research: Enhancing small-area deprivation measures with Experian™ mosaic groups. *Health and Place*. https://doi.org/10.1016/j.healthplace.2019.05.005.

Webber, R., & Burrows, R. (2018). *The predictive postcode: The Geodemographic classification of British society*. SAGE.

Whitworth, A., Carter, E., Ballas, D., & Moon, G. (2017). Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an old problem. *Computers, Environment and Urban Systems, 63*, 50–57. https://doi.org/10.1016/j.compenvurbsys.2016.06.004.

Xiang, L., Stillwell, J., Burns, L., Heppenstall, A., & Norman, P. (2018). A geodemographic classification of sub-districts to identify education inequality in Central Beijing. *Computers, Environment and Urban Systems, 70*, 59–70. https://doi.org/10.1016/j.compenvurbsys.2018.02.002.

Ye, J., Chow, J. H., Chen, J., & Zheng, Z. (2009). Stochastic gradient boosted distributed decision trees. *International Conference on Information and Knowledge Management, Proceedings.*. https://doi.org/10.1145/1645953.1646301.

Yu, R. P., Ellison, N. B., McCammon, R. J., & Langa, K. M. (2016). Mapping the two levels of digital divide: Internet access and social network site adoption among older adults in the USA. *Information, Communication & Society*. https://doi.org/10.1080/1369118X.2015.1109695.