POPULATION, SPACE AND PLACE
Popul. Space Place **22**, 849–863 (2016)
Published online 20 July 2015 in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/psp.1972

The Role of Digital Trace Data in Supporting the Collection of Population Statistics – the Case for Smart Metered Electricity Consumption Data

Andy Newing, Ben Anderson*, AbuBakr Bahaj and Patrick James
Sustainable Energy Research Group, Energy and Climate Change Division, Faculty of Engineering and Environment, University of Southampton, Southampton, UK

ABSTRACT

Debates over the future of the UK's traditional decadal census have led to the exploration of supplementary data sources, which could support the provision of timely and enhanced statistics on population and housing in small areas. This paper reviews the potential value of a number of commercial datasets before focusing on high temporal resolution household electricity load data collected via smart metering. We suggest that such data could provide indicators of household characteristics that could then be aggregated at the census output area level to generate more frequent official small area statistics. These could directly supplement existing census indicators or even enable development of novel small area indicators. The paper explores this potential through preliminary analysis of a 'smart meter-like' dataset, and when set alongside the limited literature to date, the results suggest that aggregated household load profiles may reveal key household and householder characteristics of interest to census users and national statistical organisations. The paper concludes that complete coverage, quasi-real time reporting, and household level detail of electricity consumption data in particular could support the delivery of population statistics and area-based social indicators, and we outline a research programme to address these opportunities. © 2015 The Authors. *Population, Space and Place* published by John Wiley & Sons Ltd.

*Correspondence to: Ben Anderson, Sustainable Energy Research Group, Energy and Climate Change Division, Faculty of Engineering and Environment, University of Southampton, Southampton, UK.
E-mail: B.Anderson@soton.ac.uk

Accepted 22 May 2015

Keywords: census; energy monitoring; small area statistics; digital trace data; big data

INTRODUCTION

Provision of area-based population statistics in the UK is underpinned by the decadal census of housing and population as a crucial source of consistent baseline population estimates and robust local area statistics. Census data represent a backbone for commercial, academic and social research, widely used for in-depth analysis, policy making, and resource allocation (Eurostat, 2011), including allocation of billions of pounds of government and commercial investment at the local level. Key census outputs include estimates of usually resident population by age and sex reported using a hierarchy of output zones. However, the real value to policy and commercial analysts is derived from the publication of tables detailing the combinations of attributes of households and their usual residents at the small area level (ONS, 2014a). These provide information on indicators such as ethnic composition, education, socio-economic status, religion, and employment, and it is the combination of universal geographic coverage at the small area level coupled with this detailed attribute data that represents a major strength of the census as a tool for academic, policy, and commercial research as well as for public service resource allocation (Watson, 2009; ONS, 2013b).

Given its importance, the cost of the census (£480m in 2011) represents exceptional value,

especially when annualised over a decadal period. Nevertheless, its future has faced uncertainty given increasing costs, difficulties of enumeration, and concerns over the typically 2-year time lag between enumeration and the delivery of small area attribute estimates (Dugmore *et al.*, 2011). The Office for National Statistics (ONS) review of the strengths and weaknesses of the current census recognised that governmental administrative data, or data held by commercial organisations, could support the production of more frequent census-type statistics (Skinner *et al.*, 2013; ONS, 2014b). Further work has established that whilst linked administrative data could deliver more frequent population estimates, it would not be able to provide outputs below the Local Authority District level (ONS, 2013b; ONS, 2013a; Calder & Teague, 2013). Based on the ONS' review of options (ONS, 2014a), extensive user consultation (ONS, 2014b), and an independent review (Skinner *et al.*, 2013), the UK Statistics Authority therefore recommended that a 'traditional', albeit primarily online, decadal census should be carried out in 2021 (Dilnot, 2014).

Nevertheless, as a number of authors have noted (Dugmore *et al.*, 2011; Struijs *et al.*, 2014), large-scale transactional geocoded 'digital trace' (Lazer *et al.*, 2009) or 'digital exhaust' (Manyika *et al.*, 2011) datasets from a range of sectors could offer opportunities to supplement census or administrative data. However, the use of such data does not feature within the recommendations for UK Census taking in 2021, is only recently being explored by the ONS (Dugmore, 2009; Goodwin, 2011), and, as far as we are aware, does not currently feature within the national statistical census taking or population statistics of any nation. Yet many sources of geocoded transactional 'big data' are becoming important supplements to census-like indicators within the commercial sector, generating household or postcode level indicators of commercial value (Furness, 2008; Webber, 2009; Sleight, 2014). Subsequent aggregation of these indicators to existing small area census geographies could provide more frequent census-like small area statistics or support the development of novel indicators not currently available through census or administrative data but which would have academic, policy, or commercial research relevance.

In this paper, we briefly consider household level data held by a range of utility companies before focusing on smart meter-derived household electricity consumption data in light of the planned universal rollout of electricity smart meters (Davey, 2013; DECC, 2013; Energy UK, 2013). We present preliminary analysis of a small scale-smart meter-like dataset to demonstrate that it may be feasible to infer selected household and dwelling characteristics from household electricity demand 'profiles', which could then be aggregated to generate familiar but more frequently updated small area statistics. We conclude that the available evidence suggests that energy-monitoring data offers potential for enhancing area-based population statistics and outline a future research programme which, despite obvious challenges, could take this work forward.

Transactional 'Big Data' for Small Area Population Statistics

Commercial data from the retail, financial, leisure, utilities, and telecoms sectors could all act as a direct source of population statistics and household indicators. As Dugmore (2009) noted, service providers maintain customer or user databases such as loyalty card databases (Humby *et al.*, 2008; Burt *et al.*, 2010), geo-located positioning data from mobile phones (Pentland, 2009; McInerney *et al.*, 2013), flows inferred by usage of public transport (Batty, 2013) or behaviours, opinions, and movement patterns revealed through social media (Malleson & Birkin, 2012). Research and commercial practice acknowledges the valuable social and spatial insights that can be gained from these big data sources, many of which are geocoded at the individual or household level, thus containing vast and timely information about individual activity patterns, social and spatial interactions, and their associated characteristics (Lazer *et al.*, 2009; Pentland, 2013; Graham & Shelton, 2013), which could not only underpin 'normal' small area statistics but also novel indicators hitherto invisible to census users (Pentland, 2009).

However, unlike the official census, commercial digital trace data are not designed for statistical aggregation or data analysis. In general, they only cover a subset of the population and do not have a defined target population (Struijs *et al.*, 2014). As a result, such datasets rarely provide sufficiently robust coverage of all sectors of the population and may form a potentially biased

DOI: 10.1002/psp

self-selected sample. From a population statistics point of view, it is likely therefore that only data sourced from 'universal service' sectors, such as household utilities, might provide a suitable and robust alternative to traditional census taking and may represent a real opportunity for application to official statistics.

In the case of telephony, around 84% of UK households have a fixed line ('landline') telephone service (Ofcom 2013), and this increased during 2013 driven largely by the need to have a fixed line for most broadband services. However, those households without a fixed line tend to be younger and from lower socio-economic groups (Ofcom 2013) re-introducing problems of potential bias.

Similarly, whilst almost 85% of households receive mains gas, non-gas households tend to have a well-known rural and low-income distribution (Baker 2011) and include some 5% of households who could receive mains gas but do not do so. In addition, gas usage, especially in winter, is predominantly used for space heating under thermostat and often timer control (Building Research Establishment 2013) so that distinguishing between occupant and dwelling characteristics from such data may prove difficult even with the introduction of smart meters.

In contrast, almost 100% of households are connected to mains water (Communities and Local Government 2010), but only 43% of residential dwellings are metered in England and Wales and less than 1% in Scotland and Northern Ireland.[1] Thus, whilst water companies' address bases may form a useful resource to support development of robust address listing for census or survey enumeration (Dugmore, 2009), little is reliably known about the occupants nor their water consumption profiles at the dwelling or aggregate level.

As with water, almost 100% of households receive mains electricity with only an estimated 200,000 households 'off-grid' for electricity supply (Communities and Local Government 2010). The electricity industry 'meter postcode address file' records all billable supply points (including all 'on-grid' dwellings), and as with the water service address databases, this could support development of robust address listings. Until recently, however, the standard service delivery and billing process relied on six monthly (or less) readings of traditional 'dumb' meters and therefore provided little in the way of additionally useful information other than aggregated and often estimated small area consumption statistics (DECC, 2013).

However, a UK Department of Energy and Climate Change (DECC) mandated Smart Metering Implementation Programme (DECC, 2011) aims to install smart meters into all domestic dwellings and many small businesses by 2020. The installed meters will be capable of recording electricity consumption at 5- to 10-second intervals and storing half-hour aggregates, representing the standard 'settlement period' used for load analysis and billing in the UK, for 13 months (DECC, 2013). The meters will feed near real-time energy consumption data to (optional) in-home display units and/or third party energy management services whilst energy suppliers will be able to extract half-hourly consumption data from the meter for customer billing, fraud prevention, and fault detection via the newly formed Data Communications Company (DCC). Critically, the DCC will also have the capability to extract half-hourly consumption data for all domestic meter points for authorised purposes irrespective of supplier. The presence of this single data gateway together with the near 100% population coverage of electricity metering suggests a focus on the potential value of smart meter electricity consumption data.

Smart Meter Data for a 'Smart' Census

Smart meter data is recognised as a potential resource by the 'UK strategy for Data Resources for Social and Economic Research' (UK Data Forum, 2013) with a particular focus on research on energy demand or applications and services (Firth *et al.*, 2008; Kohnstamm, 2011; Roberts, 2013; Naus *et al.*, 2014). It has also been recognised as a potential resource for official statistics (Caroll *et al.*, 2013; UNECR, 2013; Struijs *et al.*, 2014), but with few exceptions (e.g. Caroll *et al.*, 2013), we are unaware of any studies that have explicitly explored the value of analysing such data in support of the production of small area statistics.

Nevertheless, a range of studies including the Department for Environment, Food and Rural Affairs (DEFRA)/Energy Saving Trust 'Powering the Nation' project (Zimmermann *et al.*, 2012), academic research (Craig *et al.*, 2014), national statistical office work (Caroll *et al.*, 2013), and the

DOI: 10.1002/psp

large-scale Energy Demand Research Programme (AECOM, 2011) and Low Carbon Network Fund trials (Haben *et al.*, 2014) have all illustrated the role of household characteristics as drivers of energy use. In the commercial context, energy management services such as Onzo (2012) claim to have accumulated extensive databases of household energy smart meter data, and their development of commercial services based on the data suggests that household behaviours and routines can be usefully inferred through high temporal resolution energy monitoring data.

In essence, the approach proposed in this paper builds on these insights to invert the current commercial practice of combining observed average power demand profiles with census and other small area data to forecast electricity demand (Kleiminger *et al.*, 2013). Instead, we seek to use observed household electricity demand to infer household and dwelling characteristics before aggregating these to established small area boundaries to generate population statistics. The identification of attributes of particular households is therefore *not* the end goal of this work. Rather, the end goal for official statistical purposes is the estimation of the characteristics of *groups of households or dwellings at small area level*s such as the Census output area (OA). Having estimated the probability that a household, dwelling, or person has a particular characteristic, these estimates would then be appropriately aggregated to produce and publish estimated household attribute counts or proportions within a Census small area subject to the usual statistical disclosure controls and existing codes of practice (UK Statistics Authority 2009). Thus, generating small area estimates requires a two-stage process: firstly, inferring household characteristics from transactional datasets and secondly, aggregating these characteristics to relevant output levels for publication of area-based statistics. The analysis presented in this paper is concerned solely with the first stage and does not seek to generate area-based indicators although we return to a discussion of the development and validation of such small area indicators in our conclusions.

Clearly, success in stage one (and subsequently) is dependent on there being an identifiable link between household characteristics and observed electricity consumption, and one potential approach is to make use of temporal load profiles constructed from high temporal resolution electricity demand data. In contrast to summary (e.g. annualised) consumption figures or six monthly meter reads, load profiles typically consider mean load during each 30-minute period of the day providing insights into the timing of different levels of demand. Clearly, these can represent an individual household on a given day, or a summary over multiple days or weeks, and households are commonly grouped together to produce aggregate load profiles. This approach smooths individual household level variation and helps to identify general trends by type of household for use in the targeting of interventions aimed at improving network efficiency (Haben *et al.*, 2014) or reducing energy use (Bardsley *et al.*, 2013). In the following sections, we explore the links between load profiles and underlying household composition and characteristics.

Load Profiles and Household Composition and Characteristics

The magnitude and timing of demand emerge from complex household behaviours, driven largely by household occupancy levels and the energy-using activities undertaken by household residents. The literature suggests that whilst consistent routines tend to produce fairly consistent load profiles for a given household on a day-to-day basis (Ning & Kirschen, 2010), load profiles vary considerably *between* households, in terms of the timing and magnitude of demand. There is evidence that household load profiles vary by the number of household residents, the presence of children, the size of the dwelling, and householder employment status (Firth *et al.*, 2008; Wright, 2008; Beckel *et al.*, 2013; McLoughlin *et al.*, 2013; Hughes & Moreno, 2013) suggesting considerable potential as a tool to differentiate between households. Given that we would expect these household characteristics to exhibit spatial variation across small area geographies, yet some degree of homogeneity within a given small area, there is considerable scope to use household load profiles to infer these characteristics at the small area level.

Whilst a number of large scale energy monitoring datasets are becoming available for analysis, unless additional information is known about the characteristics of households, it is very difficult to provide further explanations of the patterns observed (see also Craig *et al.*, 2014) and

DOI: 10.1002/psp

thus impossible to develop techniques to infer those household characteristics from such data. As far as we are aware, the only attempt to carry out this kind of analysis to date used data from the Irish Commission for Energy Regulation's Smart Metering Electricity Consumer Behaviour Trials (CER, 2012). A pre-trial and post-trial survey collected data on household characteristics including occupant counts by age and sex, social class, and employment status of the chief income earner, indicators of daytime occupancy, dwelling type, tenure, income, and appliance use, and this can be linked to 30-minute electricity demand data for over 5,000 households for 18 months.

McLoughlin *et al.* (2012) used this dataset to characterise households based on their temporal load profiles, identifying key features such as the timing and relative magnitude of peak demand. They recognised that the magnitude of power import (demand) was strongly influenced by household composition (such as the number of residents and the presence of children), whilst household and dwelling characteristics appeared to influence the timing of use. Beckel *et al.* (2013) used the same data to develop a classification system, which they claim is able to estimate household characteristics, including income, number of residents, and floor area with accuracy levels of up to 80%. However, whilst these studies are both valuable examples of the potential use of these datasets, neither seeks to derive population statistics at the small area level. Instead, McLoughlin *et al.* (2012) seek to describe variations in electricity demand by household type in order to target energy savings and reduce peak demand, whereas the end goal for Beckel *et al.* (2013) is to support the targeting of value added 'energy efficiency consulting' services.

In contrast, Caroll *et al.* (2013) made use of the same dataset to assess the feasibility of determining household composition from smart meter data with the explicit aim of supporting national statistical organisations in the production of population statistics. They recognised that new techniques for handling, storing, and analysing the volumes of household level data produced would be needed, and they attempted to reduce the volume of data by deriving a series of summary indicators from load profiles. They concluded that a number of dimensions of electricity demand could prove to be a usable predictor of household composition and thus form the basis for area level estimates of household types, although they noted considerable challenges encountered during their analysis associated with data storage, processing, and manipulation. Notwithstanding these challenges, their exploratory analysis suggests that load profiles or associated profile indicators could offer considerable potential as a predictor of household characteristics. In the remainder of this paper, we illustrate the potential value using preliminary analysis of a smart meter-like dataset.

UoS-E Smart Meter-like Dataset

Building on this discussion, the remainder of this paper uses a smart meter-like dataset to explore the feasibility of deriving census-like indicators at the household level. The data were collected as part of a University of Southampton Household Energy Monitoring Study (UoS-E). Study households were instrumented to collect instantaneous power demand ('power import') every second, and this data was continuously uploaded to a secure database at the University of Southampton The power import data can be linked to periodic surveys of dwelling characteristics as well as household composition, householder behaviours, and attitudes, and these characteristics are summarised in Tables 1–3 for the 95 study households used in this research. Households were recruited from two areas in the south of England, which comprised predominantly suburban neighbourhoods of relatively affluent owner-occupiers. Almost three quarters of the study households were drawn from output area classification Supergroup 4, 'Prospering Suburbs'. Households in this supergroup tend to show a high propensity to represent detached homes, with residents tending to be well-qualified professionals, typically aged 45–64 years and with no dependent children, thus representing some of the least deprived neighbourhoods (Williams & Botterill, 2006; Vickers & Rees, 2006).

We draw on data collected over a 3-week period running from 00.00 on Saturday, 1 October 2011, until 23.59 on Friday, 21 October 2011, which correspond closely with survey data collection to avoid household-type misclassification. The early October time period avoids holidays associated with the education sector, public holidays, and major religious festivals, all of which could considerably alter the temporal characteristics of

DOI: 10.1002/psp

Table 1. Self-reported household composition for 95 study households from the UoS-E dataset.

Counts of households by composition					
		Number of usually resident adults			
Number of usually resident children		1	2	3+	Total
	None	6	30	10	48
	1	3	12	1	15
	2	3	19	1	21
	3+		10		11
	Total	12	71	12	95

UoS-E, University of Southampton Household Energy Monitoring Study.

Table 2. Self-reported dwelling type by OAC supergroup (allocated using household postcode and based on 2001 Census Data) for 95 study households from the UoS-E dataset.

Counts of dwellings by and OAC supergroup						
		OAC supergroup				
Dwelling type		3	4	5	6	Total
	Detached	4	61	1	10	76
	Semi-detached house	3	4	2	3	12
	Terrace	1	4	1		6
	Unknown		1			1
	Total	8	70	4	13	95

OAC, output area classification; UoS-E, University of Southampton Household Energy Monitoring Study.
3 = countryside; 4 = prospering suburbs; 5 = constrained by circumstances; 6 = typical traits.

power demand although, as discussed in the succeeding paragraphs, such features may turn out to be useful indicators in themselves. Likewise, we also avoided a mid-summer or mid-winter observation period in order to negate the effects of electrically powered space heating or cooling, which may be driven by devices using timers or thermostatic control and thus not represent behaviours and routines associated with household occupancy.

We have cleaned and aggregated this 1-second level data to 30-minute settlement periods on a household-by-household basis, calculating the mean of instantaneous power import for each half-hour of each day. Not only does summarising over a 30-minute period have the effect of smoothing short-term fluctuations and enables general trends to be observed but it also mimics the half-hourly data that will be available in the future from smart meters. Figure 1 represents an aggregated temporal load profile for all 95 study households across the mid-week period of Tuesday, Wednesday, and Thursday during our observation period. Each boxplot corresponds to a half-hour period and shows the statistical dispersion of mean household power import within that period with the 50% of households falling within the interquartile range (IQR) represented by the shaded box.

Table 3. Self-reported HRP employment status for 95 study households from the UoS-E dataset.

Counts of households by employment Status of HRP	
Employed	
Work full-time (employed, including maternity leave, etc)	37
Work part-time (employed, including maternity leave, etc)	22
Freelance/self employed	13
Not in active employment	
Full time student	1
House wife/husband	2
Retired	18
Unemployed/between jobs	2
Total	95

HRP, household reference person; UoS-E, University of Southampton Household Energy Monitoring Study.

DOI: 10.1002/psp

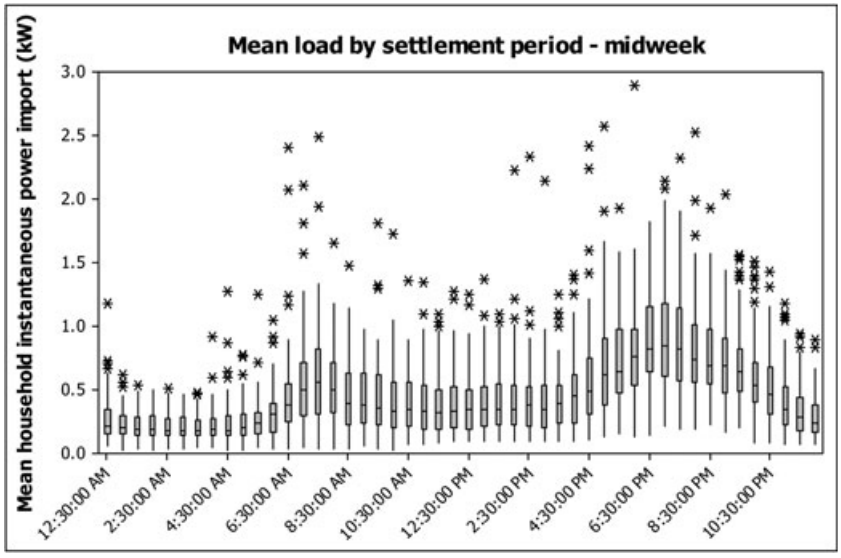

Figure 1. Temporal load profile for 95 study households ($n = 95$) for the midweek period (Tuesdays to Thursdays) during 1–21 October 2011. Load profiles are presented as boxplots, representing the statistical dispersion of household mean instantaneous power import by settlement period. Shaded box represents inter-quartile range; outliers are denoted with *.

Figure 1 exhibits a familiar temporal load profile, with pronounced morning and evening peaks in demand (Elexon, 2013). The broader IQR and whiskers during those peak periods, plus the presence of outliers, suggest that considerable variation is evident between households, especially during the evening peak period. It is this variation between households that offers potential for determining household characteristics provided these differences are driven, at least in part, by the characteristics that we seek to infer.

In the following section, we use this relatively small dataset to consider the link between selected load profiles and the census-like household characteristics available within the linked survey data. Table 4 shows that a number of census household characteristics are present in the UoS-E-linked household survey data, and the literature suggests that several of them may be associated with distinctive household load profiles. Employment status, for example, is collected via the census as part of questions related to economic activity and forms an important tool at the local, regional, and national level. It is an important indicator for the commercial sector, enabling household classification and acting as a predictor for purchasing behaviours. The literature identifies a strong link between the magnitude and timing of electricity demand and householder employment status (Yohanis *et al.,* 2008) and is one of the available characteristics from the UoS-E study that we explore in the succeeding paragraphs.

We also consider household income, which does not form part of census data collection, yet represents an indicator of considerable value to policy makers and the commercial sector. Despite frequent calls for the inclusion of an income or earnings question within the census, plans to collect this information within the 2011 census were dropped amidst concerns of under-response driven by the perceived intrusion of this question (ONS, 2008). Attempts to incorporate an income question in 2001 and previous censuses also failed, and the lack of information on income collected via the census is frequently cited as a weakness especially by commercial users who use income extensively as a predictor of consumption (Dugmore *et al.,* 2011). Commercial sources such as household energy monitoring data may thus represent the best solution for obtaining small area insight into household attributes such as incomes, which may be more difficult to collect via the Census and other major surveys, enhancing small area statistics and addressing census-users' needs. The literature also recognises household income as a driver of household load profiles, and income is

DOI: 10.1002/psp

Table 4. Census 2011 household variables – availability in UoS-E-linked survey data and potential to predict these from load profiles.

Census 2011 household level variables	UoS-E linked survey variables	Existing evidence for links to load profiles
Household composition		
Number of persons	Y	Beckel *et al.* (2013)
Age of HRP	Y	McLoughlin (2013)
Number of children	Y	Yohanis *et al.* (2008)
Marital Status	Y	
Dwelling Characteristics		
Household dwelling type	Y	McLoughlin (2013); Wright (2008)
Household tenure	Y	Druckman and Jackson (2008)
Number of (bed)rooms		See dwelling floor area as a proxy
Number of cars/vans	Y	
Presence of and fuel used for heating	Y	McLoughlin (2013)
Economic Activity		
NS-SEC of HRP		Hughes and Moreno (2013); Druckman and Jackson (2008); McLoughlin (2013)
Economic activity of HRP/hours worked	Y	Yohanis *et al.* (2008); McLoughlin (2013)
Householder characteristics		
Ethnic group/country of birth of HRP/ main language		
Presence of person with limiting long term illness		
Additional characteristics within UoS-E Data		
	Income	McLoughlin *et al.* (2012); Beckel *et al.* (2013); Craig *et al.* (2014)
	Dwelling floor area	McLoughlin *et al.* (2012); Beckel *et al.* (2013); Craig *et al.* (2014)

UoS-E, University of Southampton Household Energy Monitoring Study; HRP, household reference person.

considered further in the following section not only for its policy relevance but also because the literature suggests a relationship with load profiles.

Exploring Variations in Load Profiles by Household Composition and Characteristics

The literature clearly identifies that both the magnitude and timing of electricity demand are a function of the number of household residents and household composition, the latter referring to the age structure and presence of children. Energy monitoring studies strongly suggest that households with more residents generally exhibit a higher magnitude power demand than smaller households, whilst the presence of children impacts upon the timing of peak loads, with pronounced morning and late afternoon peaks associated with preparation for and return from school or other education (e.g. Druckman & Jackson, 2008; Firth *et al.*, 2008; Wright, 2008).

Figure 2 presents temporal load profiles for 87 of the study households grouped by the number of residents and their broad age group. Eight study households are not incorporated as their combination of residents (e.g. one adult and one child) generated categories where $n < 5$. The midweek period (Tuesday–Thursday) is representative of weekday routines, and whilst the night-time base load appears consistent across all households, it is clear that larger households exhibit a higher magnitude daytime load, particularly during the evening peak period, with households comprising two adults and three children exhibiting a higher maximum mean load (1.31 kW) than smaller one adult households (0.57 kW). Thus, electricity loads during the peak period appear to represent an indicator of household size. Furthermore, households where

DOI: 10.1002/psp

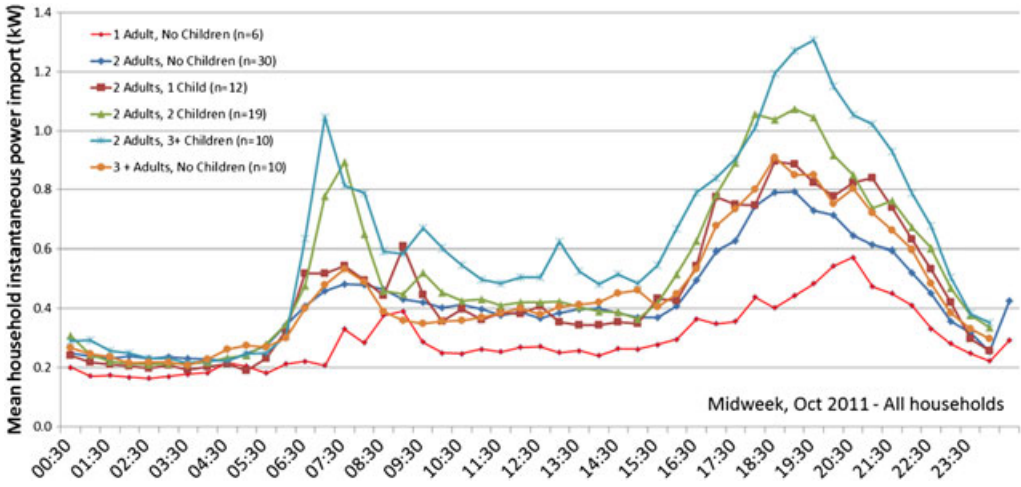

Figure 2. Household load profiles by household composition.

children are present show more pronounced morning peak demand, potentially driven by routines associated with preparation for school. Even if the underlying cause of these observed behaviours are uncertain, Figure 2 suggests that key household compositional indicators such as the number of household residents or the presence of children could be inferred from temporal load profiles.

Of the household response persons in the UoS-E study, over 75% reported being in employment, almost 20% were retired (Table 3), and the remaining 5% were unemployed, studying or full-time housewife/househusband. The latter categories were combined with retired households for subsequent analysis, giving two groups: *Employed* ($n = 72$) and *Not in active employment* ($n = 23$). Householder employment status should, however, be treated with some caution. Self-reported employment status must be treated as an indicator only as response categories provided by the UoS-E study do not account for the full range of nuanced employment patterns that may exist, such as homeworking and flexible working arrangements, which would impact considerably on behaviours, routines, and domestic electricity loads.

As aforementioned, Figure 3 presents midweek (Tuesday to Thursday) load profiles for these households, and there is some evidence that households where the HRP is in employment exhibit a higher early evening peak load, generally associated with a return from daytime employment and commencement of domestic activities such as cooking (Yohanis *et al.*, 2008). Figure 3 supports this to an extent, yet daytime loads among 'inactive' households are not noticeably higher than those in employment although we should remember that we are representing an entire household's employment status via that of the HRP. Nevertheless, inactive households, which in this sample tend to be those with retired HRPs, also tend to exhibit a slightly higher ratio of mean load to maximum load (known as the 'load factor'), driven by the lower peak load relative to households in employment. Thus whilst the actual timing and magnitude of loads could potentially be used to identify household composition, household employment status (or at least the presence of retired householders) could potentially be inferred from an indicator such as the load factor.

The literature suggests that higher income households tend to occupy larger dwellings than low income households, with a commensurate increase in consumption (Yohanis *et al.*, 2008). In addition, even after accounting for dwelling size, higher income households have been found to exhibit a higher overall electricity consumption and tend to exhibit more pronounced morning and evening peak loads possibly driven by routines associated with employment (Yohanis *et al.*, 2008; Craig *et al.*, 2014). The UoS-E dataset records each household occupant's self-reported net annual income within bands. Taking the midpoint of each band as an exact proxy, net household income after tax, national insurance, and pension contributions range from £4,000 to

DOI: 10.1002/psp

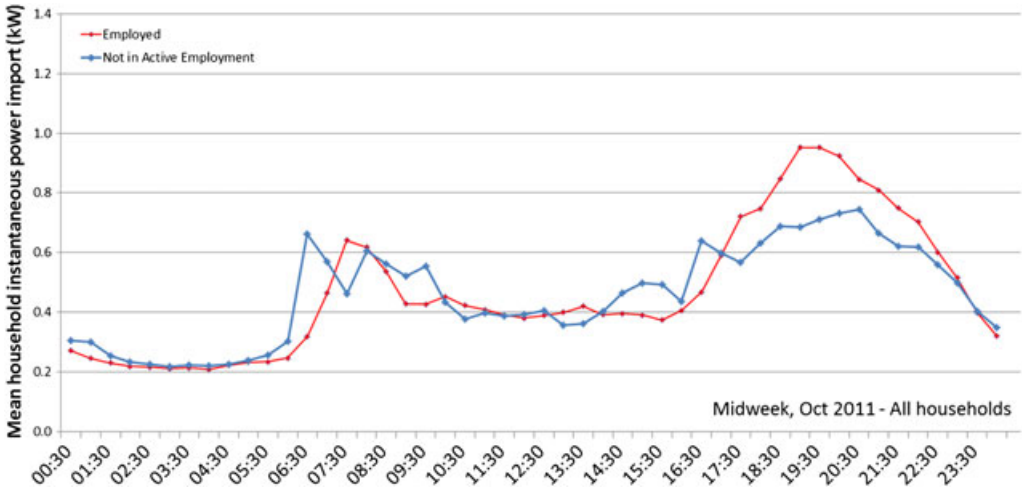

Figure 3. Household load profiles by employment status.

£156,000, with a mean of £46,149. Given that we have already identified a clear link between household composition and load profiles, our exploration of income as a factor influencing electricity loads takes account of household composition and considers only those households comprising two adults and no children ($n = 30$) and those comprising two adults and one or two children ($n = 31$), representing two of our largest groups of study households.

Figure 4 presents load profiles for those households, grouped by relative household income (above/below median). After accounting for household composition, it is apparent that there is a general trend towards higher mean loads by higher income households especially during the early evening peak period. Notwithstanding the fact that our analysis is based on self-reported income for a small sample of households, Figure 4 provides further evidence for the link between household income and load profiles identified within the literature. Thus, there is potential that smart metered electricity consumption data of this nature could afford value as a potential indicator of household income, especially if underlying confounding characteristics, such as household composition, are known, as explored in the following section.

DISCUSSION AND NEXT STEPS

Our initial exploration of the UoS-E smart meter-like dataset suggests that household power demand profiles exhibit variations between households in terms of both the magnitude and timing of demand (*c.f.* Figure 1). The literature review and exploratory analysis presented here suggest that these differences may be linked to household characteristics such as the number of residents and presence of children. We also find that load profiles appear to be driven by HRP employment status and household income, although as yet the analyses do not fully account for interactions between these factors and for confounding effects. Whilst we acknowledge that these observations are based on a small sample of households, they are all the more encouraging given that the 95 households used from the UoS-E dataset are broadly similar in terms of their characteristics. The literature strongly suggests that greater variation in the timing and magnitude of electricity loads would be expected where more pronounced underlying variation exists between household and dwelling characteristics including, for example, housing tenure and householder socio-economic status (Firth *et al.*, 2008; Wright, 2008; Beckel *et al.*, 2013; McLoughlin *et al.*, 2013; Hughes & Moreno, 2013). This should, in turn, increase the capacity to distinguish between groups of households with similar load profiles.

Whilst the preliminary results presented earlier are encouraging, as we have noted there remain very few studies in the literature that are based on high temporal resolution smart meter data for any more than 200–300 households in a UK context. There is therefore currently a lack of suitable household level datasets which could be used to further explore the link between

DOI: 10.1002/psp

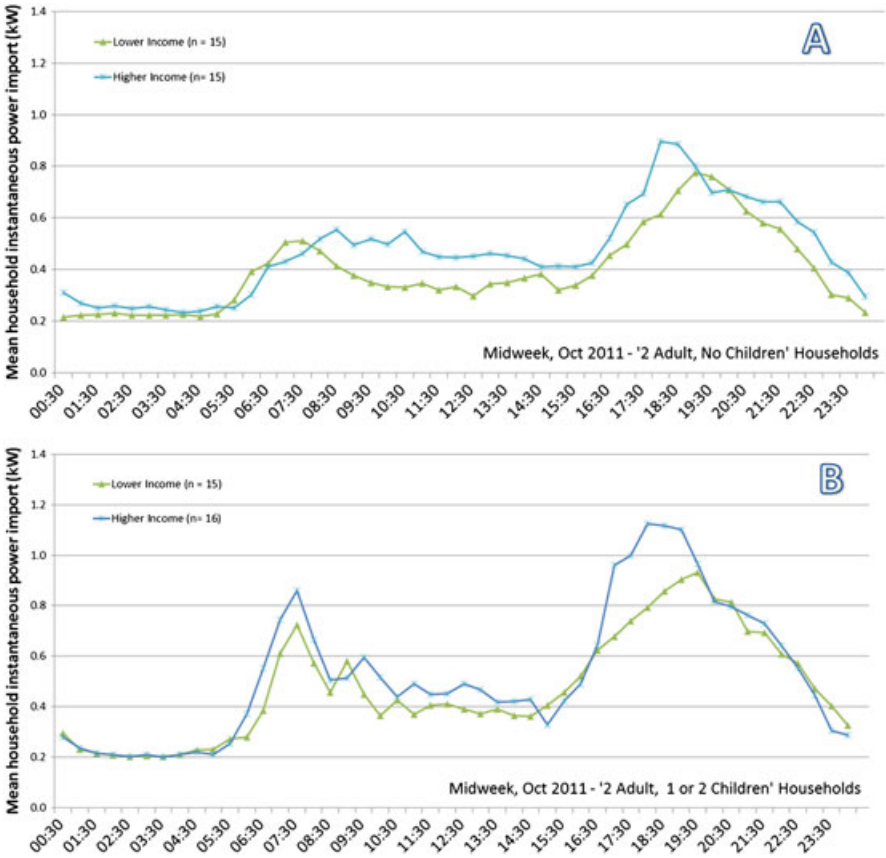

Figure 4. Household load profiles by broad income group (above/below median) for (A) households with two adults and no children and (B) households with two adults and one or two children.

household load profiles and their underlying characteristics for the purposes proposed within this paper. With the notable exception of the Irish Commission for Energy Regulation's smart meter trial data, this is largely because larger-scale smart meter and energy demand reduction trials rarely include sufficient socio-demographic data for the purposes outlined here. Furthermore, because these datasets are generally collected for energy policy studies and are usually self-selecting samples, they tend to focus on a limited range of household types, such as those eligible for government grants to improve energy efficiency or those who are prepared to take part in an experimental trial. As such, there are very few datasets that consider temporal patterns of domestic energy loads (a point previously noted by Wright & Firth, 2007), and even fewer that link fine grained temporal load profiles with household characteristics for a large-scale representative population sample.

It is nevertheless the case that smaller datasets such as those used within this paper provide the opportunity to explore the feasibility of generating population estimates from household energy loads and to experiment with different approaches without encountering storage, processing, or analytic limits. Even for just 95 households over a 3-week period, our time series data at a 30-minute resolution produced almost 100,000 individual readings, although through aggregation, this was reduced to a much smaller dataset for the analysis presented here. Nevertheless, manipulation of 1-second level data to form cleaned half-hourly aggregates necessitated the use of High Performance Computing infrastructure, and the literature recognises these challenges when working with larger datasets of the order of thousands of households and with sub-half-hourly data (Caroll *et al.*, 2013; Emerson & Kane, 2012; Thumim & Wilcox, 2013).

Whilst analytic processing power will inevitably increase, our ongoing analysis seeks to reduce

DOI: 10.1002/psp

data storage and processing requirements by utilising summative indicators (e.g. mean load, base load, and peak load) and temporal properties (e.g. timing of peak load and ratio of peak to off-peak loads) of individual household load profiles The literature suggests that indicators which summarise key features of household load profiles can smooth the variable nature of electricity loads and reduce data volumes whilst enabling identification of key consumption 'events' (such as the timing and magnitude of peak load), which may be driven by underlying household characteristics and routines (Yohanis *et al.*, 2008; Beckel *et al.*, 2012; McLoughlin, 2013; Ning & Kirschen, 2010). If these indicators could be derived prior to transfer from the smart meter to the DCC, then overall data storage and processing requirements for this form of work may be reduced.

The next step (stage one) for the approach outlined in this paper must therefore be to develop a range of appropriate predictive multivariate statistical models to derive household characteristics of interest to national statistical bodies from electricity load profiles. We have argued that stage one can only be achieved using 'labelled' consumption data from relatively large-scale representative samples where household characteristics are known. Our exploratory analysis suggests that inferring household composition, such as the number of residents or presence of children, would make an effective starting point. However, it is realistic to assume that future forms of the Census in the UK will almost certainly make use of administrative data to provide population counts and basic household composition for small area geographies (ONS, 2013b; ONS, 2014a). If these were available at the household level then it may be possible to use known household composition alongside household load profiles in order to infer household attribute information such as employment status or income, for example. The approach would then seek to derive valuable novel indicators that supplement area-based attribute information, rather than generate indicators of household composition which could be provided by census or administrative data.

The second step (stage two) must then be the application of these models to a larger scale representative random sample or complete 'census' of anonymised smart meter data extracted from known geographical areas to produce estimates of inferred household characteristics at small area levels. Ideally such estimates would be produced at the OA level with validation taking place against other sources of small area population statistics for those areas such as from the 2011 Census. In particular, validation would seek to assess whether aggregation over zones as small as Census OAs could generate estimates of household characteristics with acceptable accuracy for use by national statistical organisations. Unfortunately, datasets suitable for stage two are currently rare, and those that do exist tend to be small scale or have restricted access even though they would offer considerable potential to understand the relationship between household characteristics and the temporal profile of electricity loads.

Future work could make use of greater time-series information within the underlying dataset from which our sample is drawn. As previously noted, we have made use of a 3-week period in October to coincide with linked survey data collection and to avoid institutional holidays linked to the education system or public holidays. Nevertheless, because households with certain characteristics may exhibit different routines and energy using behaviours during school holidays (where children are present) or on bank holidays (where householders are employed full-time), differences in an individual households load profile at different times of the year could potentially afford some insight into characteristics and behaviours. Consideration of the seasonal dimension may also reveal behaviours associated with use of primary or secondary electric space-heating. As we have already noted, use of timers or thermostatic controls to operate these devices may mask some of the energy-using behaviours that can be used to identify household characteristics, but may also provide a valuable insight into energy inequality, enabling novel policy-relevant indicators to be developed.

CONCLUSIONS

Overall, our review of evidence to date, combined with preliminary descriptive analysis, has suggested that there is value in exploring the use of smart meter electricity consumption data for the purposes of deriving population statistics and indicators of household attributes at small area levels of geography. With electricity smart-meter roll out underway or anticipated in a number of the largest domestic markets including the US, China, Brazil, India, and Japan, alongside

DOI: 10.1002/psp

France, Germany, and Spain (Deloitte 2011), we suggest that there is considerable potential for this form of analysis in a variety of national contexts. In the UK, this form of data could supplement official census-type statistics, addressing users' requirements for timely reporting of small area statistics. In particular, regular data extraction via the DCC could be used to infer households or dwelling characteristics at any given time and to monitor changing small area characteristics over intercensal periods (Claxton *et al.*, 2013), yet to date, there has been little empirical research in this area.

Using a smart meter-like dataset, we observe that household power demand profiles exhibit variations between households, likely to be driven by the complex inter-relationship between factors such as the number of occupants, the presence of children, the type and size of dwelling, and household socio-economic and geodemographic characteristics. As discussed earlier, the next stage requires development of predictive multivariate statistical models to derive those characteristics of interest from household electricity load profiles. We argue that the research programme outlined in this paper would go some way to addressing this gap although there are clear and currently unmet data requirements. These include large-scale representative population samples where a wide range of household attributes ('labels') can be linked to the households' smart meter data and anonymised large-scale smart meter data extracts from known small area geographies, which could be used to validate the approach.

At the same time, national statistical organisations may be able to derive additional novel local area statistical indicators based on the energy consumption data itself. These might include novel indicators of local energy inequality and indicators of potential for demand reduction as well as indicators of local demand for other kinds of public or retail services, a use to which many 'added value' data processors currently put the existing census. The potential for additional value extraction from smart meter data through an aggregated data analytics market, whilst maintaining household level non-disclosure and privacy, is clear.

ACKNOWLEDGEMENTS

The development of this paper was supported by the ESRC funded 'Census 2022: Transforming Small Area Socio-Economic Indicators through "Big Data"' project (ES/L00318X/1). We would like to thank the members of the Census 2022 Advisory Group for comments on presentations of early versions of this material.

The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

NOTES

(1) Authors' calculation using ONS Living Costs and Food Survey 2012

REFERENCES

AECOM. 2011. *Energy Demand Research Project: Final Analysis*. AECOM: St Albans.

Baker W. 2011. *Off-gas Consumers*. Consumer Focus: London.

Bardsley N, Buchs M, James PA, Papafragrou A, Rushby T, Saunders C, Smith G, Wallbridge R, Woodman N. 2013. Initial effects of a community-based initiative for energy saving: an experimental analysis. *Working Paper*. Universities of Westminster, Reading, Southampton and Exeter.

Batty M. 2013. Big data, smart cities and city planning. *Dialogues in Human Geography* **3**: 274–279.

Beckel C, Sadamori L, Santini S. 2012. Towards automatic classification of private households using electricity consumption data. *In: BuildSys Conference, 6th November 2012, Toronto, Canada*.

Beckel C, Sadamori L, Santini S. 2013. Automatic socio-economic classification of households using electricity consumption data. *In: e-Energy Conference, 21st-24th May 2013, Berkeley, California*.

Building Research Establishment. 2013. *Energy Follow-Up Survey 2011 Report 4: Main Heating Systems*. Prepared by BRE on behalf of Department of Energy and Climate Change: London.

Burt S, Sparks L, Teller C. 2010. Retailing in the United Kingdom – a synopsis. *European Retail Research* **21**: 173–174.

Calder A, Teague A. 2013. *The Census and Future Provision of Population Statistics in England and Wales: Presentation Delivered at RGS 'The Future of Small Area Population Statistics' 21st October 2013*. Office for National Statistics: Newport.

Caroll P, Dunne J, Hanley M, Murphy T. 2013. Exploration of electricity usage data from smart meters to investigate household composition. *In: Conference of European Statisticians, 25-27 September 2013, Geneva, Switzerland*.

DOI: 10.1002/psp

CER. 2012. *Smart Meter Electricity Consumer Behaviour Trial Data*. Irish Social Science Data Archive: Dublin.

Claxton R, Reades J, Anderson B. 2013. On the value of digital traces for commercial strategy and public policy: telecommunications data as a case study. In *The Global Information Technology Report 2012*, Dutta S, Bilbao-Osorio B (eds). World Economic Forum: Geneva.

Communities and Local Government. 2010. *English Housing Survey: Housing Stock Report 2008*. Department for Communities and Local Government: London.

Craig T, Polhill JG, Dent I, Galan-Diaz C, Heslop S. 2014. The North East Scotland energy monitoring project: exploring relationships between household occupants and energy usage. *Energy and Buildings* **75**: 493–503.

Davey E. 2013. Written ministerial statement by Edward Davey: Smart Metering: 10th May 2013 [online]. [Accessed 4th October 2013]. Available from: https://www.gov.uk/government/speeches/written-ministerial-statement-by-edward-davey-smart-metering.

DECC. 2011. *Smart Metering Implementation Programme: Response to Prospectus Consultation*. Department of Energy and Climate Change: London.

DECC. 2013. *Smart Metering Equipment Technical Specifications Version 2*. Department of Energy and Climate Change: London.

Dilnot A. 2014. *The Census and Future Provision of Population Statistics in England and Wales. Letter to Rt. Hon. Francis Maude MP from the Chair of the UK Statistics Authority, Sir Andrew Dilnot CBE, 27th March 2014*. UK Statistics Authority: London.

Druckman A, Jackson T. 2008. Household energy consumption in the UK: a highly geographically and socio-economically disaggregated model. *Energy Policy* **36**: 3177-3192.

Dugmore K. 2009. *Information Collected by Commercial Companies: What Might Be of Value to ONS?* Demographic Decisions Ltd.: London.

Dugmore K, Furness P, Leventhal B, Moy C. 2011. Beyond the 2011 census in the United Kingdom: with an international perspective. *International Journal of Market Research*, **53**: 619.

Elexon. 2013. *Load Profiles and Their Use in Electricity Settlement*. Elexon: London.

Emerson J, Kane M. 2012. Don't drown in the data. *Significance* **9**: 38–39.

Energy UK. 2013. About smart meters [online]. [Accessed 05 November 2013]. Available from: http://www.energy-uk.org.uk/customers/about-smart-meters/how-much-data-is-collected-with-smart-metering-and-is-it-secure.html.

Eurostat. 2011. *EU Legislation on the 2011 Population and Housing Censuses: Explanatory Notes*. European Commission (Eurostat): Luxembourg.

Firth S, Lomas K, Wright A, Wall R. 2008. Identifying trends in the use of domestic appliances from household electricity consumption measurements. *Energy and Buildings* **40**: 926–936.

Furness P. 2008. Real time geodemographics: new services and business opportunities (and risks) from analysing people in space and time. *Journal of Direct, Data and Digital Marketing Practice* **10**: 104–115.

Goodwin G. 2011. *Counting New Information About the UK's Population "Beyond 2011": Presentation Delivered at the DUG Conference, 12th October 2011*. Office for National Statistics: Newport.

Graham M, Shelton T. 2013. Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography* **3**: 255–261.

Haben S, Ward J, Vukadinovic Greetham D, Singleton C, Grindrod P. 2014. A new error measure for forecasts of household-level, high resolution electrical energy consumption. *International Journal of Forecasting* **30**: 246–256.

Hughes M, Moreno J. 2013. *Further Analysis of Data from the Household Electricity Usage Study: Consumer Archetypes – Report for DECC and DEFRA*. Element Energy Limited: Cambridge.

Humby C, Hunt T, Phillips T. 2008. *Scoring Points: How Tesco Continues to Win Customer Loyalty.* Kogan Page: London.

Kleiminger W, Beckel C, Staake T, Santini S. 2013. Occupancy detection from electricity consumption data. *Paper presented at BuildSys'13, November 14 - 15 2013, Rome, Italy.*

Kohnstamm J. 2011. *Article 29 Data Protection Working Party: Opinion 12/2011 on Smart Metering (00671/11/EN).* European Commission: Brussels.

Lazer D, Pentland A, Adamic L, Aral S, Barabási A-L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M. 2009. Computational social science. *Science*, **323**: 721-723.

Malleson N, Birkin M. 2012. New insights into individual activity spaces using crowd-sourced big data. In: *2014 Big Data Conference, 27-31 May 2014, Stanford, CA, USA*.

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung Byers A. 2011. *Big Data: The Next Frontier for Innovation, Competition and Productivity.* McKinsey Global Institute: San Francisco.

McInerney J, Rogers A, Jennings N. 2013. Bus, bike and random journeys: crowdsourcing aid distribution in Ivory Coast. *Significance*, **10**: 4–9.

McLoughlin F. 2013. Characterising domestic electricity demand for customer load profile segmentation [THESIS]. Thesis, Dublin Institute of Technology.

McLoughlin F, Duffy A, Conlon M. 2012. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic

DOI: 10.1002/psp

variables: an Irish case study. *Energy and Buildings* **48**: 240–248.

McLoughlin F, Duffy A, Conlon M. 2013. Evaluation of time series techniques to characterise domestic electricity demand. *Energy* **50**: 120–130.

Naus J, Spaargaren G, van Vliet BJM, van der Horst HM. 2014. Smart grids, information flows and emerging domestic energy practices. *Energy Policy* **68**: 436–446.

Ning Z, Kirschen D. 2010. *Preliminary Analysis of High Resolution Domestic Load Data*. School of Electrical & Electronic Engineering, University of Manchester: Manchester.

Ofcom. 2013. *Communications Market Report 2013*. Ofcom.

ONS. 2008. *2007 Census Test: The Effects of Including Questions on Income and Implications for the 2011 Census*. Office for National Statistics: Newport.

ONS. 2013a. *Beyond 2011: Newsletter – July 2013*. Office for National Statistics: Newport.

ONS. 2013b. *Beyond 2011: Options Explained 2*. Office for National Statistics: Newport

ONS. 2014a. *The Census and Future Provision of Population Statistics in England and Wales: Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority*. Office for National Statistics: Newport.

ONS. 2014b. *The Census and Future Provision of Population Statistics in England and Wales: Report on the Public Consultation*. Office for National Statistics: Newport.

Onzo. 2012. *ONZO Application Detection Technology*. Onzo Ltd.: London.

Pentland A. 2009. Reality mining of mobile communications: toward a new deal on data. *In:* Dutta S, Mia I (eds). *The Global Information Technology Report* 2008–2009. World Economic Forum: Geneva.

Pentland A. 2013. *The Data Driven Society Scientific American* **309**: 64–69.

Roberts S. 2013. Energy gap claptrap [online]. [Accessed 17th June 2014]. Available from: http://www.cse.org.uk/news/.

Skinner C, Hollis J, Murphy M. 2013. *Beyond 2011: Independent Review of Methodology*. Independent review for the UK Statistics Authority: London.

Sleight P. 2014. Geodemographic Classification Systems – The New Breed [online]. [Accessed 17th July 2014]. Available from: http://www.geodemographics.org.uk/blog/gkb/peter_sleight_geodemographic_classification_systems_-andnbsp;the_new_breed/id/796.

Struijs P, Braaksma B, Daas PJ. 2014. Official statistics and big data. *Big Data & Society* **1**: 1–6.

Thumim J, Wilcox T. 2013. Managing and 'mining' smart meter data at scale. *Energise (Autumn 2013): News from the Centre for Sustainable Energy.*

UK Data Forum. 2013. *UK Strategy for Data Resources for Social and Economic Research*. UK Data Forum: Swindon.

Uk Statistics Authority. 2009. *Code of Practice for Official Statistics*. UK Statistics Authority: London.

UNECR. 2013. *What Does "Big Data" Mean for Official Statistics*. United Nations Economic Commission for Europe - Conference of European Statisticians: Geneva.

Vickers D, Rees P. 2006. Introducing the area classification of output areas. *Population Trends* **125**: 15–29.

Watson G. 2009. Making the case for the 2011 census. *Presentation Delivered by at the DUG Annual Conference, 8th October 2009*. Office for National Statistics: Newport.

Webber R. 2009. Response to 'the coming crisis of empirical sociology': an outline of the research potential of administrative and transactional data. *Sociology* **43**: 169–178.

Williams S, Botterill A. 2006. Profiling areas using the output area classification. *Regional Trends* **39**: 11–18.

Wright A. 2008. What is the relationship between built form and energy use in dwellings? *Energy Policy* **36**: 4544–4547.

Wright A, Firth S. 2007. The nature of domestic electricity-loads and effects of time averaging on statistics and on-site generation calculations. *Applied Energy* **84**: 389–403.

Yohanis YG, Mondol JD, Wright A, Norton B. 2008. Real-life energy use in the UK: how occupancy and dwelling characteristics affect domestic electricity use. *Energy and Buildings* **40**: 1053–1059.

Zimmermann J-P, Evans M, Griggs J, King N, Harding L, Roberts P, Evans C. 2012. *Household Electricity Survey: A Study of Domestic Electrical Product Usage*. Intertek: Didcot.

DOI: 10.1002/psp