Smart water metering as a non-invasive tool to infer dwelling type and occupancy – Implications for the collection of neighbourhood-level housing and tourism statistics

A. Newing [a,b,*], O. Hibbert [a,b], J. Van-Alwon [a,b], S. Ellaway [c,d], A. Smith [c]

[a] *School of Geography, University of Leeds, Leeds LS2 9JT, UK*
[b] *Leeds Institute for Data Analytics (LIDA) University of Leeds, Leeds LS2 9JT, UK*
[c] *School of Geography, Earth and Environmental Sciences, University of Plymouth, Plymouth, PL4 8AA, UK*
[d] *South West Water, Peninsula House, Exeter, EX2 7HR, UK*

A B S T R A C T

The international rollout of advanced metering infrastructure (AMI) in the residential water supply sector affords tremendous benefits in driving water-use efficiencies, accurate billing and network management (e.g. leak detection). AMI, using 'smart meters' fitted at a dwelling level, record water consumption at high temporal resolution. Since water is typically only consumed when householders are present, these data could offer a non-intrusive means of inferring dwelling occupancy patterns. These insights could have a range of benefits dependent upon the spatiotemporal scale and the intended application – our interest is in the potential of these data to identify dwelling type, specifically to identify dwellings that have occupancy patterns associated with tourism, such as second homes or short-term holiday rentals. We focus on these data in a UK context and draw on data rarely available for academic research. Our data relate to a sample of dwellings in Devon and Cornwall, South West England. They capture high-temporal resolution water consumption during Covid-19 'lockdown' and 'staycation' periods, providing a unique opportunity to demonstrate that these data can reveal the unusually pronounced property-level occupancy trends evident during this period. We apply Non-Intrusive Occupancy Monitoring (NIOM) to extract dwelling-level occupancy status (occupied/unoccupied) on a day-by-day basis. We group properties according to their occupancy trends, inferring a set of properties that exhibit occupancy characteristics associated with tourism. We demonstrate that these show correspondence with underlying indicators of tourism activity, drawn from AirDNA records of short-term tourist rental properties in this area. Ongoing global rollout of AMI means that these data will be routinely available at the dwelling level and we reflect on the benefits they could provide in generating near real time insights into dwelling occupancy. Drawing on our collaboration with the Office for National Statistics (the UKs national statistical institute) we outline the considerable potential that these data and approaches could offer in the collation of small area housing and tourism statistics.

1. Introduction and context

The introduction of Advanced Metering Infrastructure (AMI) in the residential water supply sector is part of the UK's commitment to reduce water consumption, as outlined in the DEFRA Environmental Improvement Plan (DEFRA, 2023a) and the UK Water Efficiency Strategy (Waterwise, 2022). Comparable roll-out of AMI in an international context means that globally, in excess of 6.2 M AMI-enabled meters have been installed in the residential water supply sector (Jacobs, 2023).

Typically involving instillation of a 'smart meter' on the water supply pipe to a dwelling, AMI equips consumers with near real-time information on the volume and cost of their water consumption. Meters can be read remotely in near real time, offering wider benefits including more accurate consumer billing, reduced cost of metering (no need for meters to be read manually) and improved network management (including detection of leakage). Monks, Stewart, Sahin, and Keller (2019) and Frontier Economics, Artesia, and Arqiva (2021) nicely highlight some of these benefits. Whist the overall number of dwellings

in the UK with a smart water meter fitted is unknown, some water companies have made rapid progress in installing smart meters and Arqiva announced in 2022 that the millionth water smart meter had been connected to their network, used to link meters to water company systems (Baker, 2022).

In most regions, all residential dwellings are supplied with water (and in many cases sewerage services too) by a single water company. Most water companies therefore have a regional monopoly within their supply jurisdiction, though there are some smaller sub-regional local water companies, especially in the South East (OFWAT, 2023). As penetration of AMI increases this could mean that within many regions, a single water company will collate dwelling-level smart water meter data with near-complete household coverage. This could afford tremendous potential extending far beyond water use efficiencies and network management. In this paper we highlight the potential re-use value these data could offer in inferring dwelling type, specifically the identification of properties associated with tourism, and their potential value in the generation of small area housing and tourism statistics.

Tourists staying in self-catering accommodation have long been poorly captured within official statistics, especially at the local, sub-district level (Johns & Lynch, 2007; White, 2010). Whilst some isolated reports have sought to identify the distribution of tourist accommodation and its economic impact (Newing, 2013; Scanlon, Sagor, & Whitehead, 2014), there remains a lack of robust small area data capturing the stock of tourist accommodation, especially those drawn from the housing stock. Census enumeration does not capture seasonal occupancy of dwellings (aside from student properties), and the Office for National Statistics (ONS) report difficulties in enumerating dwellings such as holiday-lets and second homes in the production of official statistics (Abbott, 2018). An ONS review of tourism statistics (ONS, 2022b), and EU-wide work by national statistical organisations to improve tourism statistics (European Commission, 2023) has recognised the potential role that novel data sources (including those from booking portals, mobile phones, payment cards and travel agencies) could offer in the production of tourism statistics. In this paper we contribute to that work, presenting novel work which assesses the potential that water metering data could offer in identifying dwellings associated with tourism and inferring their occupancy patterns.

We draw on data supplied by South West Water (SWW), the regional water company for the counties of Devon and Cornwall in South West England. These counties experience considerable seasonal population uplift driven by coastal tourism in the summer months and therefore present an excellent opportunity to explore the potential value of these data in capturing property level occupancy fluctuations associated with tourism. These data also overlap with the Covid-19 pandemic. The pronounced dwelling occupancy patterns evident during that period, including those associated with 'lockdowns' (stay at home) and 'stay-cations' (increased rates of domestic tourism, especially in coastal localities as a result of restrictions on international travel) present an additional opportunity to demonstrate the value of these data in uncovering dwelling-level occupancy trends. Furthermore, growing rates of domestic tourism during the Covid-19 pandemic has increased the number of residential dwellings being converted into short term tourist lets (Halliday & Morris, 2022).

We demonstrate that Non-Intrusive Occupancy Monitoring (NIOM) techniques can be used to infer property occupancy status on a day-by-day basis. We highlight that this could have considerable benefit as a near real-time indicator of dwelling usage profiles associated with tourism. We have worked closely with the ONS, the UKs national statistical institute. The ONS have responsibility for collecting and publishing population and neighbourhood statistics, including the decadal census in England and Wales and have an international reputation for innovation in data and methods for the production of official statistics. As outlined in section 5, the ONS have broad interests in understanding the role that novel data sources, such as those derived from smart metering, could play in the production of official statistics. ONS

methodological specialists have contributed to methodological discussion and feedback throughout the project. This work also builds upon a previous study which assessed the potential of smart-meter data from the electricity sector as a tool to support the provision of official statistics (Anderson, Lin, Newing, Bahaj, & James, 2016; Anderson & Newing, 2015; Newing, Anderson, Bahaj, & James, 2016).

The analysis and findings presented in this paper highlight the potential reuse value of these data in inferring near-real time dwelling characteristics. Specifically, we address the following research questions:

1. Can we infer property occupancy from dwelling-level water consumption data in order to infer dwelling type/usage characteristics, including identification of dwellings associated with tourism?
2. What are the potential value of these insights, and the underlying data, in supporting the provision of household and neighbourhood-level population and tourism statistics?

This work is novel as it represents the first study to explicitly consider the role of dwelling-level water supply data as an indicator of property type and local tourism activity. Whilst the work reported here is UK-centric (given the involvement of SWW and ONS), the provision of water via a supply authority is common to all developed countries and the approaches are internationally transferable. In the following section we review recent literature in this domain, considering existing studies that have used comparable data from the electricity sector to consider dwelling occupancy. In section 3 we introduce our data and methods, including pilot analysis (using data from pre-Covid) and our full analysis, based on the 2020–2022 period. Section 4 presents our results, before a more detailed discussion of their implications and wider value in section 5. Section 6 presents a summary and conclusion.

2. Literature review

Smart water meters enable water companies to obtain meter reads remotely and in near real time using Advanced Metering Infrastructure (AMI). They detect the volume of flow within the supply pipe at high temporal resolution (up to every second). Traditional 'dumb' meters record consumption on a monthly, quarterly or biannual schedule. Smart metering enables water companies to realise benefits including leak detection (Beckel, Sadamori, & Santini, 2013; Koech, Cardell-oliver, & Syme, 2021), understanding the water supply-demand balance (March, Morote, Rico, & Saurí, 2017) and informing water saving policies (Sadr, That, Ingram, & Memon, 2021). Consumers typically receive more accurate billing and are able to monitor their own usage behaviour, resulting in financial savings and reduced environmental impact (Clifford, Mulligan, Comer, & Hannon, 2018; DEFRA, 2023b; Sønderlund, Smith, Hutton, & Kapelan, 2014). AMI deployment in the water sector is not as advanced as the electricity sector (which benefited from the Smart Metering Implementation Programme), lacking central coordination and representing a 'tangled patchwork of various interest groups' (Gill, 2022). However, the rollout of AMI features heavily in water companies Water Resources Management Plans and the UK Government Policy Paper 'Plan for Water' (DEFRA, 2023b) provides further Government encouragement to water companies to make rapid progress on installing smart meters. Severn Trent Water report plans to install over 150,000 smart water meters in Coventry and Warwickshire by 2025, creating a 'smart water region' (STWater, 2023). Thames Water intends to fit smart water meters to all suitable homes in the Thames Valley by 2035, with cited benefits in relation to day-to-day network management alongside wider potential to integrate these data into 'Digital Twins', enabling greater insight and data-driven decision making in this sector (Coates, 2023).

There is a strong history of development of AMI in the electricity sector. Whilst electricity supply in the UK is a competitive market, and therefore electricity companies do not have the regional monopolies that

are present in the water sector, there has been a broader body of work exploring the use of electricity smart meter data in a social science context. This has established the link between household electricity consumption and dwelling occupancy, including the number of residents and the presence of children or older householders (Beckel et al., 2013; Mcloughlin, Duffy, & Conlon, 2012; Newing et al., 2016; Owen, 2012). Anderson et al. (2016) demonstrated that the timing of peaks in electricity demand could provide an indication of whether a given household exhibited routines associated with going to work. They were able to predict whether a householder was in paid work, reporting an accuracy of approximately 70% compared to surveyed validation data. Prior work also includes a study commissioned by the ONS which assessed the potential role that electricity data could play in the generation of official population statistics (Anderson & Newing, 2015). That work suggested that it may be feasible to predict the number of people usually resident in a household, and the likelihood that they will be at home at a given time of the day, using their electricity consumption, at a 30 min resolution.

The ability to infer the likelihood that a householder is at home at a given time of the day suggests that these data could afford insights into the occupancy status of a dwelling. McKenna et al., discussing this form of analysis note that "clearly there is a degree of speculation …… but with a little practice it is perfectly possible to identify the general movements of residents with some confidence" (2012, p808). A number of studies have linked active occupancy (periods of time when a resident is inferred to be at home and awake) to underlying smart meter electricity consumption data (López-Rodríguez, Santiago, Trillo-Montero, Torriti, & Moreno-Munoz, 2013; ONZO, 2012; Richardson, Thomson, Infield, & Clifford, 2010; Widén & Wäckelgård, 2010). These typically infer occupancy by identifying usage of appliances such as ovens, showers and televisions within high temporal resolution 'load profiles', derived from smart meter consumption records, accompanied by some form of ground-truth data for validation.

In our application we lack ground truth-data for validation – we do not know the true occupancy status of any of our study dwellings. However, four of the 93 properties used in our pilot study were thought to be associated with tourism usage (drawing on SWW in-house intelligence) and form a core part of the data used to develop our analysis routines (see section 3.2). Whilst the 'active occupancy' studies introduced above provide clear evidence that smart meter derived data can infer dwelling occupancy, they are not suitable approaches for our analysis. Given the lack of ground truth data capturing the occupancy status of the dwellings we are working with, we need to employ unsupervised approaches, with Non-Intrusive Occupancy Monitoring (NIOM) applied here given its prior usage in applications using electricity data.

Chen, Barker, Subbaswamy, Irwin, and Shenoy (2013) highlight that NIOM offers an indirect means to monitor dwelling-level occupancy solely using energy consumption, developing an algorithm to detect property-level occupancy at different time periods during the day. Their model predicts occupancy based on whether key metrics of electricity consumption - the mean, standard deviation and range – exceed a specified threshold during a given time period (typically by hour of the day). Those threshold values were based on the night-time baseload power consumption, on the assumption that the maximum night-time value for each of these metrics reflects its maximum value when the home is unoccupied. Becker and Kleiminger (2017) tested the NIOM approach on three open source datasets, demonstrating that it could capture household occupancy at 30 min intervals with reasonable accuracy.

Our study aims to detect occupancy on a day-by-day basis, rather than at smaller time intervals during the day. Eibl, Burkhart, and Engel (2018) used a modified NIOM approach to detect occupancy at a daily scale, with the specific aim of detecting holidays, again using electricity data. To do so they considered only maximum electricity consumption and added a tolerance to account for issues encountered when using

night-time threshold values on unoccupied days, which had a tendency to incorrectly infer occupancy. In common with our study, they also lacked validation data but found that their modified approach predicted unoccupied periods (in their case capturing a count of holidays) in line with their expectations. We adapt their modified approach within our methodology, outlined in section 3.

Whilst these studies highlight the link between electricity consumption and household occupancy, the fragmentation of supply in the electricity sector means that the near-complete coverage of dwellings within a given locality, via a single company, is less likely to be achievable. The regional monopolies enjoyed by water companies is thus a major potential benefit to this form of work. Furthermore, and unlike electricity (and gas) used for 'always on' appliances and space heating, water is typically not consumed when a property is unoccupied. Water using appliances which may be left on when occupants are not present, such as washing machines and dishwashers, have time limited cycles as opposed to heating/cooling systems, internet-connected devices and smart-home hubs that typically consume electricity continuously, even if occupiers are not present.

Although extensively applied in the electricity sector, there has been far less work utilising water consumption data for applications beyond consumption monitoring, demand reduction and network management. Nevertheless, studies have demonstrated the link between observable water consumption and dwelling occupancy, which is central to the analysis carried out in the following sections. Much of that work has sought to extract usage of specific water-using appliances such as showers, toilet flushes or dishwashers (see Carboni, Gluhak, Mccann, & Beach, 2016 for a review, and Cominola et al., 2019 for a good example) which could provide an indicator of household composition and routines. Applications of these approaches have included identification of seasonal variations in some activities (e.g. watering gardens) (Cardell-Oliver, Wang, & Gigney, 2016), or classification of end-users into residential or business properties (Laspidou et al., 2015). Using time-series clustering on simulated datasets, Steffelbauer, Blokker, Buchberger, Knobbe, & Abraham (2021) predicted household characteristics including household size, employment status and employment schedule from smart water meter data. However, we are unaware of any published research which has specifically used these data to infer property occupancy and identify tourist dwellings. In the following section we introduce our smart water meter dataset and the approaches we use, derived from those used with similar electricity data and introduced above, to infer dwelling occupancy characteristics.

3. Data and methods

3.1. Introduction to the data

We utilise high temporal-resolution water consumption data for a sample of households in Devon and Cornwall. These data were provided by SWW who supply approximately 800,000 properties within the region (SWW, 2023). These data are based on a sample of households subject to long term consumption monitoring by SWW, utilising high temporal resolution data loggers. They collect timestamped data recording each litre of water consumed at up to a 1 s resolution. Prior to data sharing, these data were aggregated to a 15 min resolution (count of litres consumed within each 15 min time period) and are akin to the format of data available from domestic smart meters in this sector.

We developed our data processing and analysis routines at the 15 min temporal resolution. However, these data are typically collected and processed at the hourly resolution by water companies (notably Thames water, the 'first mover' in terms of smart meter instillation at scale in the UK (Baker, 2022)). An hourly resolution has been widely applied in existing studies for a range of uses including leak detection and household consumption monitoring (Britton, Cole, Stewart, & Wiskar, 2008; Cardell-Oliver, 2013). Aggregation to a coarser temporal resolution is common when working with similar data related to

electricity consumption and has the added benefit of reducing noise, reducing the impact of missing data and smoothing spikes in consumption (Wright & Firth, 2007). Within our pilot analysis (see section 3.2) we therefore trialled our analysis on data aggregated to the 1 h resolution and found that the impact on accuracy of our occupancy monitoring technique (see section 3.2) was negligible.[1] We therefore aggregated these data to a 1 h resolution for all the data processing and analysis presented in this paper.

The set of households with data-loggers fitted is deemed by SWW to be broadly representative of the range of location and dwelling types that they supply. The household level data collected from these data loggers have been shared with us in a non-disclosive manner. Each property is identified by a unique meter ID number, but we have no access to personally identifiable customer information, meter serial numbers or linked billing information, including property address. We are therefore unable to link the observed consumption to a specific identifiable property. However, each meter ID within our data is linked to a known District Metered Area (DMA), a contiguous set of metered water supply areas introduced in the UK in the 1980s as a tool to monitor and manage leakage (Kowalska, Suchorab, & Kowalski, 2022). We are therefore able to analyse and present data at the property and area-based level without knowing or revealing the actual address of individual dwellings within our dataset. To present area-level data we use a lookup to associate each DMA with the Lower Super Output Area (LSOA) [a small area geography used for the dissemination of area based population and housing statistics, typically containing between 400 and 1200 households] that it falls within or intersects.

Our aim is to use these data to identify the occupancy status of an individual property (occupied/unoccupied) on a given day utilising NIOM techniques. The primary analysis presented in this paper is drawn from a set of 2491 properties supplied by SWW in October 2022 and covering the period March 2020 to September 2022. Data cleaning, processing and analysis routines were also initially tested on 93 properties, containing data collected by SWW between 2015 and 2019, referred to as the 'pilot analysis'. The properties used for the pilot analysis included four properties for which SWW had local intelligence that they were 'highly likely to be associated with tourism', such as a second home or holiday let. These four tourist properties were manually labelled by the research team to capture inferred occupancy status on a day-by-day basis, providing a form of ground truth data used to assess the accuracy of NIOM approaches trialled on these data. The primary analysis presented in this paper offers considerable enhancement to the pilot study. It utilises a much larger dataset and uses these data to uncover occupancy associated with specific time periods during the Covid-19 pandemic, during which pronounced dwelling occupancy patterns were observable. It also links inferred property occupancy to underlying data on occupancy rates for tourist self-catering property rentals, as outlined in the following sections.

3.2. Pilot analysis

As reported fully in van Alwon et al. (2022), pilot analysis assessed the quality and completeness of these data and trialled analytic approaches. It identified the need to account for missing data and leakage prior to analysis. Missing data is typically due to loss of wireless signal to the meter/data logger. Whilst we are aware of little work in relation to missing data in a water metering context, the body of work drawing on high-temporal resolution data from the electricity sector consistently recognises that missing or incomplete data are an inevitable challenge (Anderson & Newing, 2015; Craig, Polhill, Dent, Galan-Diaz, & Heslop,

2014; Wright & Firth, 2007). Given our interest in identifying periods of legitimate zero consumption, representing times when a property is unoccupied, it is important to be able to identify and remove missing data. van Alwon et al. (2022) noted that periods of missing data were generally longer duration, occurring less frequently than those driven by dwelling unoccupancy, enabling us to identify and exclude properties for which missing data leads to an incomplete consumption record during the period of interest.

Leakage is also common in water supply networks. Leakage occurring within the dwelling or its supply pipe will be captured by the smart meter and therefore these data require correction for leakage prior to usage. Leaks typically present as a period of non-zero consumption (a 'prolonged continuous flow') that increases over time (as the leak worsens), prior to a sudden fix. The non-zero baseline leakage recorded by the meter will be interspersed with periods of legitimate usage by householders, and thus it is necessary to identify the magnitude of the leak and remove this from recorded consumption values prior to further analysis. Pilot analysis, reported fully in van Alwon et al. (2022) developed an automated approach to identify and remove inferred leakage on a property-by-property basis, in conjunction with advice from project partners at SWW.

Leaks were identified with reference to the recorded consumption in each time period between midnight and 6 am, termed 'baseline consumption', consistent with the in-house approach used by SWW. Baseline consumption is at a time when householders are typically asleep and least likely to be using high water consuming appliances. The leak detection algorithm identifies the lowest non-zero recorded consumption for each property during the nightline period and subtracts this from all recorded consumption records for that day. Unlike missing data, which results in properties being deemed unsuitable for inclusion within subsequent analysis, we have been able to identify and correct for leakage, allowing properties with detected leakage to form part of subsequent analysis.

Pilot analysis enabled us to develop a Non-Intrusive Occupancy Monitoring (NIOM) technique to infer whether a property was occupied on a given day. As noted in section 2, NIOM is commonly used with electricity data. van Alwon et al. (2022) built upon the modified NIOM approach used by (Eibl et al., 2018) at a daily scale. van Alwon et al. (2022) trialled a range of consumption metrics, thresholds and tolerance values, to address issues encountered by Eibl et al. (2018) in using night-time threshold values to infer occupancy on unoccupied days. These included the mean, median, standard deviation and range of consumption alongside the number of 'usage events' (any period with non-zero consumption, after accounting for leakage). The chosen approach compares water consumption on the day of interest to consumption across the entire one year study period on a property-by-property basis.

Our modified NIOM approach considers a property to be occupied on a given day if:

i. The number of usage events (non-zero readings) for a given day is greater than 25% of the mean number of usage events of all days for that property, and;

ii. The daily mean volume of water consumed is greater than 25% of the average daily mean of all days for that property.

In relation to the manually labelled test data for the four tourist properties (within our pilot dataset of 93 properties), and based on data recorded at an hourly temporal resolution, this approach correctly identified occupancy status on 98.7% of days. This means that the average tourist property from within that dataset ($n = 4$) had just 5 days incorrectly assigned as occupied/unoccupied during the 1 year period (van Alwon et al., 2022). Using this approach, we are able to calculate the occupancy ratio for each property, capturing the proportion of days on which that property was occupied during the 12-month period of interest. Occupancy ratio is the key indicator used within the primary analysis presented within this manuscript.

---

[1] Based on four manually labelled test properties, the overall accuracy of our occupancy detection technique fell marginally from 98.8% accuracy (proportion of days correctly labelled as occupied/unoccupied at 15 min resolution) to 98.7% (at 1 h resolution).

3.3. Overview of the primary analysis

Analysis reported in the following section is based on the 2491 properties supplied by SWW in October 2022 and covering September 2018 – September 2022. 516 properties were excluded from further analysis as they contained missing data (gaps in recorded consumption), with a further 21 properties removed due to recording unrealistically low consumption. Those properties failed to record a minimum 30 L consumption (approximately equivalent to a 5 min shower) during any 1-h recording period and are therefore deemed to be wholly unoccupied or to have long-term meter faults. After accounting for missing data, 1882 usable properties remained, each providing at least one consecutive years' worth of valid data during the March 2020 – September 2022 study window.

We also corrected for leakage, with our leak detection and correction method (see above) identifying that 1199 properties (64% of usable properties) exhibited some form of leakage. 205 properties (11% of usable properties) had a sufficiently large leak that their maximum recorded consumption reduced post leak-detection, in one case by over 370 L. The effect on most properties was negligible, with leakage detection primarily serving to reduce excessively extreme recorded consumption in those properties with substantial leaks. The number of unusable properties within this dataset (approx. 12% of those supplied) and the high prevalence of leakage suggests that the data preparation and cleaning requirements would be substantial should this analysis be up scaled to larger datasets, as suggested in section 5. However, the automated routines that we have developed provide a mechanism through which data cleaning and leakage correction could be undertaken.

After accounting for skewed consumption due to leakage and missing data, we work with two subsets of these properties, enabling us to capture occupancy trends in two different time periods:

1. A subset of 784 properties have a complete consumption record between 26th March 2020 and 22nd September 2021, with no missing data in any 1 h time period). We use these properties to infer occupancy trends associated with Covid-19 lockdown and staycation periods.
2. A subset of 753 properties have a complete consumption record from 1st January 2022 to 30th September 2022. We use these properties to infer occupancy trends across 9 months (Jan – Sept) of 2022, free of Covid-19 restrictions, comparing these to underlying indicators of tourism activity within the region during this time period.

Since these properties are drawn from the same set of 2491 properties supplied by SWW, it is possible for a study property to appear within both of these subsets, subject to having a continuous consumption record from March 2020 to September 2022, with no missing data. 225 properties appear in both subsets, with 1312 unique properties falling within only one subset.

3.3.1. Analysis based on March 2020 – September 2021

Analysis of occupancy trends during the Covid-19 pandemic is based on our occupancy metric, calculated using the magnitude of water consumed by each dwelling on an hour-by-hour basis for all days (24 h periods) between 26th March 2020 and 22nd September 2021, coinciding with the first Covid-19 national lockdown in 2020, through to the end of summer 2021, when most Covid-19 restrictions were lifted. As previously noted, our specific interest is not to uncover household behaviour during the pandemic. Rather, the dwelling-level occupancy patterns observed during this period due to 'stay at home' guidance (lockdowns) and unusually high rates of domestic tourism (staycation), provide a unique opportunity to evaluate the ability of our occupancy detection method to uncover the pronounced occupancy characteristics observed during these periods.

Occupancy detection, introduced above, identifies whether each property is inferred to be 'occupied' or 'vacant' on a given day based on the magnitude of consumption and the presence of specific water usage events. We calculated this metric on a daily basis and reported the occupancy ratio (proportion of days during which a given dwelling was occupied) for five time periods:

1. 26th March 2020 to 1st June 2020: First national lockdown
2. 20th June 2020 to 22nd September 2020: Summer – lockdown restrictions eased, first 'staycation' summer.
3. 5th November to 2nd December 2020: Second national lockdown
4. 26th March 2021 to 1st June 2021: Spring – gradual easing of lockdown restrictions.
5. 20th June 2021 to 22nd September 2021: Summer – 'staycation' with high rates of domestic tourism.

In section 4 we illustrate that the occupancy ratio in each of these time periods varies between properties, enabling us to draw inferences about property type. We subsequently use property occupancy ratios in each time period to segment properties into groups that share similar occupancy characteristics across the time periods. A range of clustering approaches were considered including K-means, the Gausian Mixture Model (GMM) and DBSCAN. K-means clustering was chosen due to its widespread application in studies seeking to segment properties based on the characteristics of their water or electricity consumption (Abu-Bakar, Williams, & Hallett, 2021; Anderson et al., 2016; Cominola et al., 2018), and following advice from our project stakeholders.

K-means requires the analyst to specify the number of clusters into which the data should be split. In our application this was determined with reference to the Scree Plot ('elbow method') described in detail by Singleton & Longley (2015), coupled with the Silhouette Index, Calinski-Harabasz score and Davis-Bouldin scores as validation tools. In identifying the optimum number of clusters, we applied the k-means clustering algorithm to a number of cuts of the data based on different temporal recording periods. Consistently a 4 cluster solution (based on the 2020–2021 data) and three cluster solution (based on the 2022 data) were identified as being the most appropriate for these data and were applied within our analysis. Our application of k-means produces compact and distinct clusters – the within cluster sum of squares (WCSS) ranges from 0.001 (Cluster 1) to 0.295 (Cluster 2), with a between cluster sum of squares (BCSS) of 53.4. In the three cluster solution (2022), the WCSS ranges from 0.037 (Cluster 1) to 1.122 (Cluster 3), with a BCSS of 69.98. We acknowledge that there is no objective method of choosing the number of clusters but are confident that the use of these approaches, coupled with analyst expertise has generated a set of clusters that match our expectations and can be used to identify property types which are logical given our knowledge of property occupancy within this region during the time periods of interest.

3.3.2. Analysis based on January – September 2022

We apply similar tools and techniques to the consumption records for 753 properties for which we have data covering 1st January 2022 to 30th September 2022. This period was free of Covid restrictions, allowing us to assess the extent to which these data and approaches can capture nuanced property-level occupancy trends during this period. We calculate the occupancy ratio on a property-by-property basis for each month (Jan – Sept) and use these to cluster properties (k-means) according to their occupancy ratios in each time period. We compare our calculated occupancy ratios with data on known occupancy rates for self-catering tourist properties in Devon and Cornwall, extracted from AirDNA data.

AirDNA captures registered short term rental properties from the Airbnb and Vrbo booking platforms. AirDNA data is an important tool for academic research into tourism and mobility trends (Martí, Serrano-Estrada, & Nolasco-Cirugeda, 2019). We extracted 18,226 properties from AirDNA capturing self-contained rental properties across categories typically drawn from the housing stock (Home, Cottage, Rental

Unit, Condo, Vacation Home, Bungalow, Guest Suite, House, Town-house and Apartment), excluding property types such as 'B&B' and 'Guesthouse' which typically have an owner living on site and therefore exhibit more complex water use profiles, including consumption by owners when guests are not present. We extracted the month-by-month occupancy rate for each property, which AirDNA calculates as "Total Booked Days / Active Listing Nights" (AirDNA, 2023).

We also group AirDNA properties by LSOA and find that all but 17 of the 935 LSOAs in Devon and Cornwall contain properties listed on Airbnb/Vrbo. Most LSOAs have between 1 and 25 rental properties (mean 19.8), with 3 LSOAs having in excess of 25 rental properties (max = 464). We have multiplied the number of AirDNA properties in each LSOA by the mean occupancy rate for properties within that LSOA, calculating the number of 'occupied nights' by LSOA across our study period. This enables us to understand the spatial distribution of rental properties and account for their occupancy rate. We use these data to identify spatial clusters of self-catering accommodation activity. We applied Getis Ord Gi* (Getis & Ord, 1992) to reveal statistically significant hot spots within these data – groups of neighbouring LSOAs that share higher than average counts of occupied nights. We also applied Local Moran's I (Anselin, 2010) to generate a local indicator of spatial association (LISA), capturing the extent to which the number of occupied nights in a given LSOA is similar to adjacent LSOAs, both computed using an inverse distance spatial relationship. These have been reported at the LSOA level, enabling us to compare the concentration of tourism activity to the location of properties of interest (our inferred tourist properties), as revealed in section 4.2. The following section presents our findings, beginning with the analysis of dwelling occupancy during 2020 and 2021.

4. Results

4.1. Occupancy trends and inferred property type, March 2020 – September 2021

Across the 18-month period, the mean occupancy ratio (proportion of days during which a given property was occupied) was 93.2%. 89 properties (just over 11% of the 784 study properties) were inferred to be 'fully occupied' during this period (i.e. there were no nights when the property was unoccupied) whilst four properties were empty (no evidence of occupancy).

There are considerable variations between properties and between time periods, illustrated in Fig. 1(a) by comparing Spring 2020 (first

national lockdown) with summer 2020 (lockdown restrictions eased) occupancy ratios by property. A number of properties have high occupancy in both periods (top right quadrant). A number of properties exhibit either: i) low occupancy during lockdown and high occupancy during the summer (top left), or; ii) high occupancy during lockdown and lower occupancy during the summer (bottom right). Fig. 1(b) compares summer 2020 and summer 2021 'staycation' periods. Again, many properties exhibit high occupancy in both periods (top right) and are likely to represent residential dwellings. Those with lower occupancy in one or both periods may have more complex usage patterns including empty or under-utilised dwellings (bottom left), tourist properties or/ second homes, or possibly those that have undergone a change of usage or occupier between these two time periods. Whilst we have no validation data capturing the actual status of any of these properties during the study period, those properties with non-standard occupancy patterns are of particular interest. These will include short-term tourist rental properties (which we would expect to have low occupancy during lockdown periods and higher occupancy during stay-cations) or second/holiday homes which may have more complex and individualised non-standard occupancy patterns during this period.

Based on their occupancy ratios during the five time periods of interest (Table 1), these properties cluster into four distinct groups as shown in Fig. 2 (see section 3.3.1 for discussion of the selection of the appropriate number of clusters). The largest group, cluster 1, contains 694 properties. These properties exhibit the profile typically expected of residential dwellings, with near-complete occupancy in all time periods, especially during the two national lockdowns. Some properties in cluster 1 have lower occupancy during summer 2020 and 2021, when holidays are more likely to have been taken.

Cluster group 2 comprises 42 households and is also likely to represent predominantly residential dwellings. These properties have greater variability in occupancy ratios between these five time periods relative to those properties in cluster 1. Occupancy is high during the

Table 1
Inferred occupancy ratio for our study properties at various time points during 2020 and 2021 to coincide with Covid-19 events.

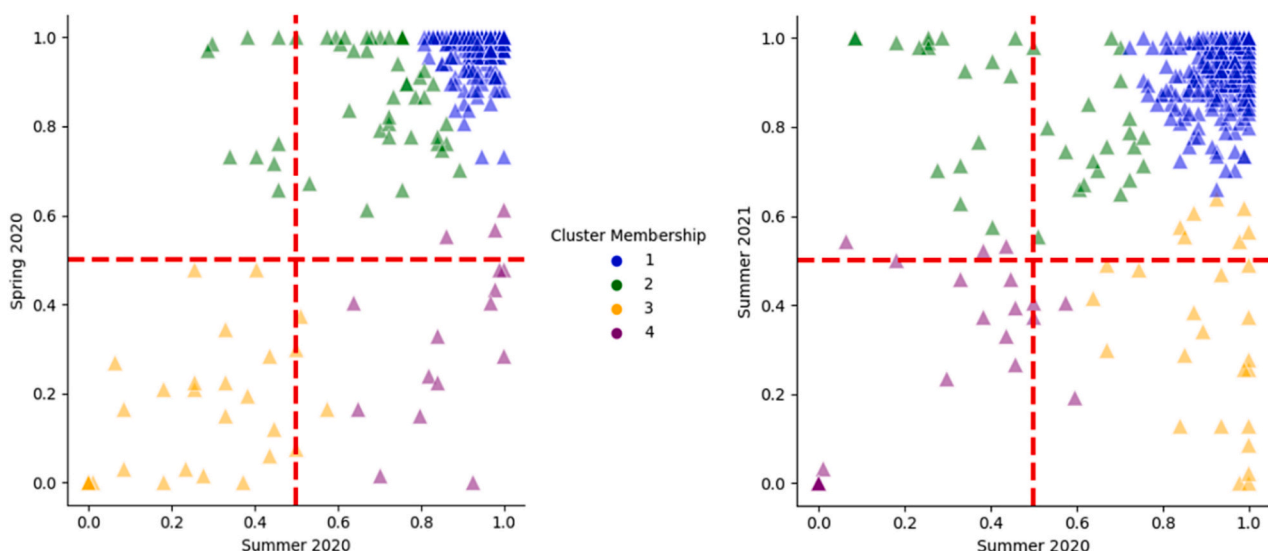| Proportion of days occupied	Spring: 26th March – 1st June	Summer: 20th June – 22nd September	Autumn: 5th November – 2nd December
2020	93.95%	92.65%	96.34%
2021	95.19%	91.77%	N/A

Fig. 1. comparison of dwelling level occupancy between a) Spring and Summer 2020, and b) Summer 2020 and Summer 2021. Cluster membership is also shown.
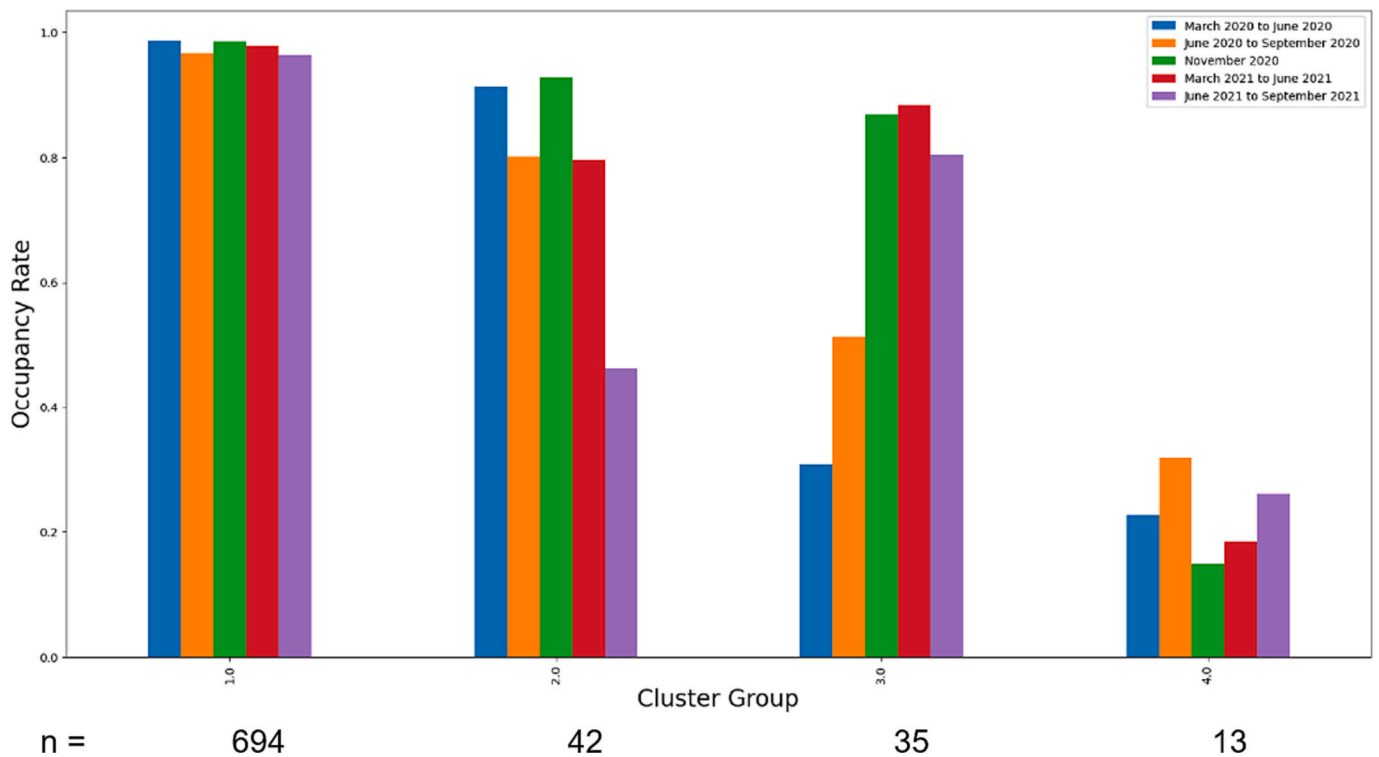
Fig. 2. Classification of dwellings according to their occupancy trends during five periods of interest between March 2020 and September 2021.

two national lockdowns, but many properties have lower occupancy in summer, especially summer 2021, which may suggest that residents were away from home for work, study or leisure. Cluster groups 3 and 4 are most likely to represent dwellings that have occupancy patterns not associated with traditional residential usage.

We infer that the 35 properties in cluster group 3 are second homes (dwellings that are primarily used as holiday homes or occupied while working away from home (House of Commons Library, 2022)) or associated with tourism (e.g. short term self-catering holiday lets). They exhibit low occupancy during the first and most stringent national lockdown and also in summer 2020, with occupancy rising in the second national lockdown (winter 2020) and peaking during the summer 2021 staycation period. Higher rates of occupancy among these properties in winter 2020 could reflect greater occupancy of second homes during this period or longer-term lets in holiday accommodation amidst a growing trend for remote working in this region (Shaw, 2021). The smaller group of 13 properties in cluster 4 may represent under-occupied dwellings including residential properties that have been empty for a longer period of time, tourist lets with low occupancy rates (including weekend-only lets) or properties with alternative occupancy trends.

Whilst the impact of Covid-19 on dwelling level occupancy rates is not our primary focus, it is clear that these data enable us to identify high rates of property occupancy during lockdown periods, especially the most severe first and second national lockdown (March to June and November 2020). Properties exhibited distinct occupancy patterns in this period, with an identifiable group of properties (cluster 3) likely to represent tourist properties and second homes. In the following section, we present findings from application of the same approaches to dwelling-level data for the year 2022, which was not subject to the same extreme occupancy trends associated with Covid-19 lockdowns and staycations.

4.2. Occupancy trends and inferred property type, 2022

During our 9-month period of interest in 2022, the mean occupancy ratio for the 753 study properties was 94%, with 117 properties (almost

16% of our sample) fully occupied. Most properties show near-complete occupancy in January (mean occupancy ratio of 96%), with lower occupancy evident in April (Easter – mean occupancy 93%) and August (summer – mean occupancy 91%), consistent with residential households taking holidays away from home. Properties have been clustered based on their monthly occupancy ratio during this 9 month period, with 3 clusters representing the optimal cluster solution (see Fig. 3).

In common with the 2020/21 data, clusters 1 and 2 likely represent residential dwellings with near complete occupancy (cluster 1) and short periods away from home (cluster 2) as a result of holidays, leisure and work travel, especially in July and August (Fig. 3). Cluster 3 is likely to represent second homes, tourist lets and under-occupied dwellings, with higher occupancy rates evident during the key summer tourist period (Fig. 3). Once again, the lack of intelligence on the actual status of these properties limits our ability to validate these findings. However, and as outlined fully in section 3.3.2, we are able to compare the occupancy rates calculated via our analysis with reported occupancy rates for tourist properties in this region derived from AirDNA data. Fig. 4 illustrates the mean monthly occupancy rate for properties from each of our three clusters, alongside observed occupancy rates derived from the AirDNA data. Cluster 3 (representing inferred second home, tourist let and under-occupied dwellings) exhibits occupancy trends most associated with tourism. Calculated occupancy ratios closely follow the month-by-month occupancy rates drawn from the AirDNA data for tourist rental properties in Devon and Cornwall.

We have also linked each of our study properties to their respective neighbourhood (LSOA) and compare the distribution of LSOAs containing properties in Cluster 3 (predominantly tourism) with underlying indicators of tourism activity. Fig. 5 illustrates the output of our Local Moran's I, capturing statistically significant localised clusters of LSOAs which share high or low counts of 'occupied AirDNA nights' (see section 3.3.2). Most statistically significant clusters or pockets of tourist activity
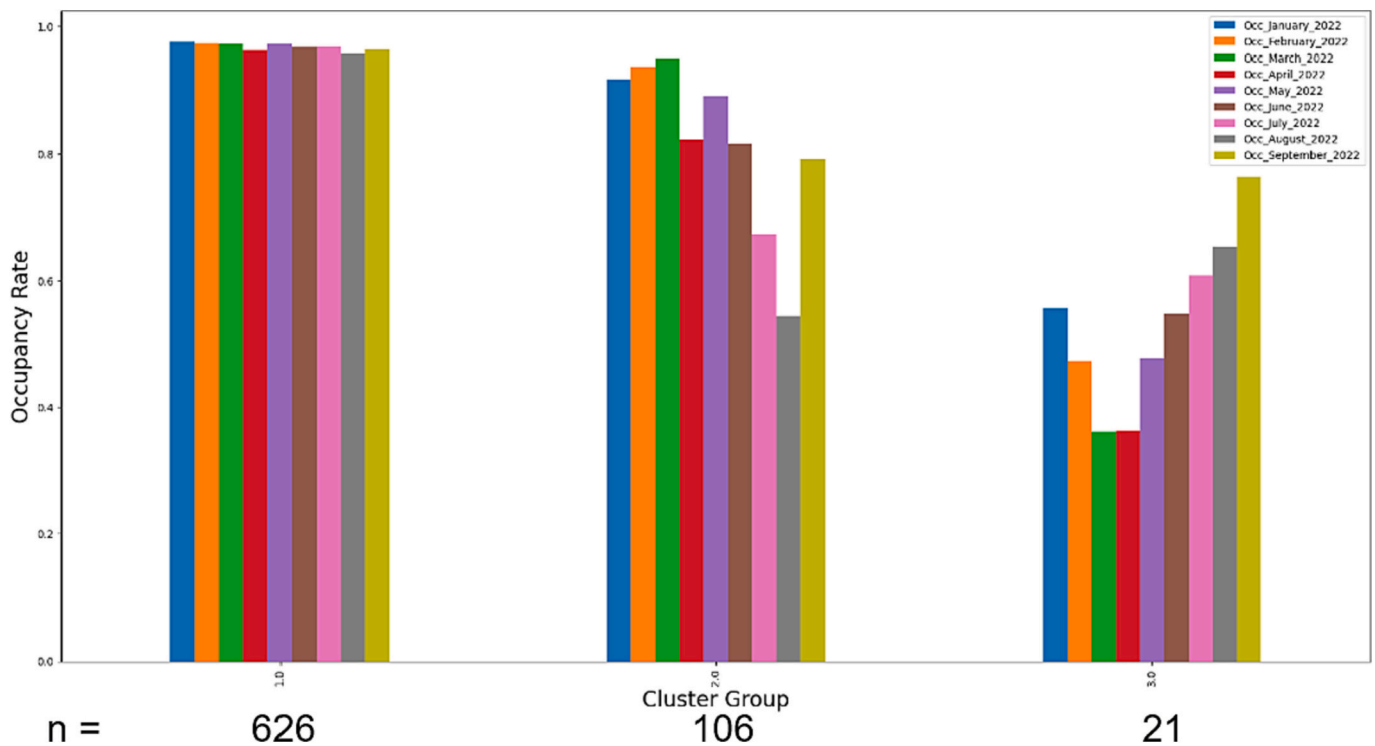
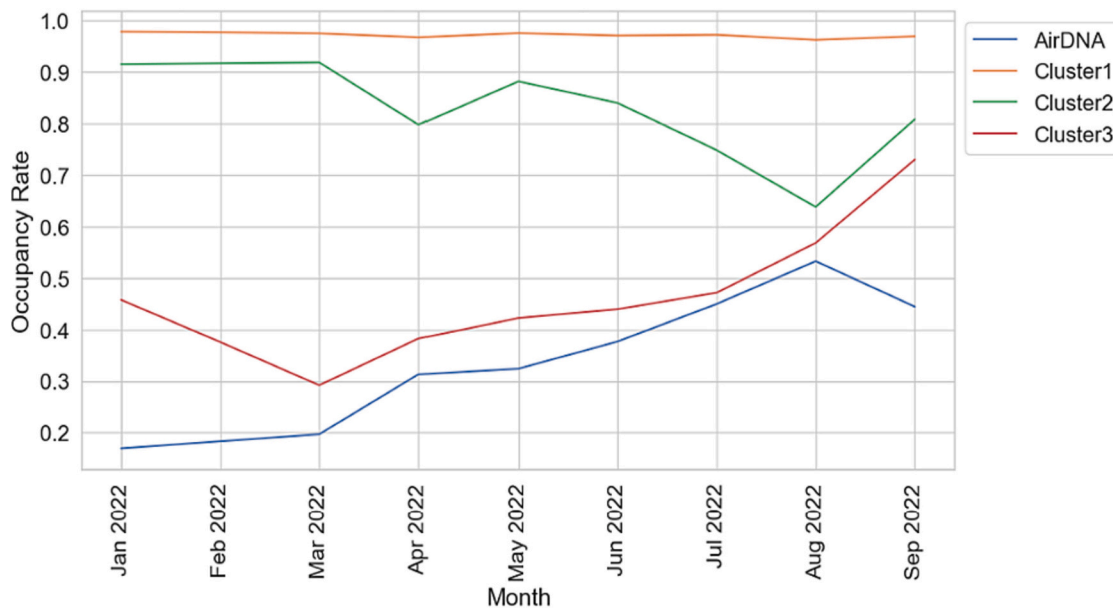**Fig. 3.** Classification of dwellings according to their occupancy trends during nine months of 2022.

Fig. 4. Mean occupancy rates by cluster and from AirDNA data for rental properties in Devon and Cornwall by month (Jan – Sept 2022).

(shown as High-High on Fig. 5) are on the north and south Devon coasts. There are also smaller clusters of occupied tourist properties on the north Cornwall coast.

4 of the 20[2] LSOAs containing inferred tourist properties (cluster 3) fall within these clusters of high tourism activity, yet 9 of the LSOAs containing our inferred tourist properties fall within clusters of low tourism activity (typically a result of fewer tourist properties and lower occupancy rates in those LSOAs). These areas of lower tourism activity incorporate some of the major cities in these counties and industrial inland areas less-traditionally associated with tourism. Whilst these areas are dominated by high density residential properties, it is not unreasonable to assume that some properties in these neighbourhoods could represent short term rental properties for tourism, leisure and work-related visitors. Our outputs (not shown on Fig. 5) from running Getis-Ord Gi* on these data present a similar pattern, identifying coastal 'hot spots' of tourism activity, with only three of the LSOAs containing our inferred tourist properties falling within those statistically

---

[2] Although there are 21 properties in cluster 3, only 20 could be linked to a known geographic location due to incomplete geo-location information in the data supplied.
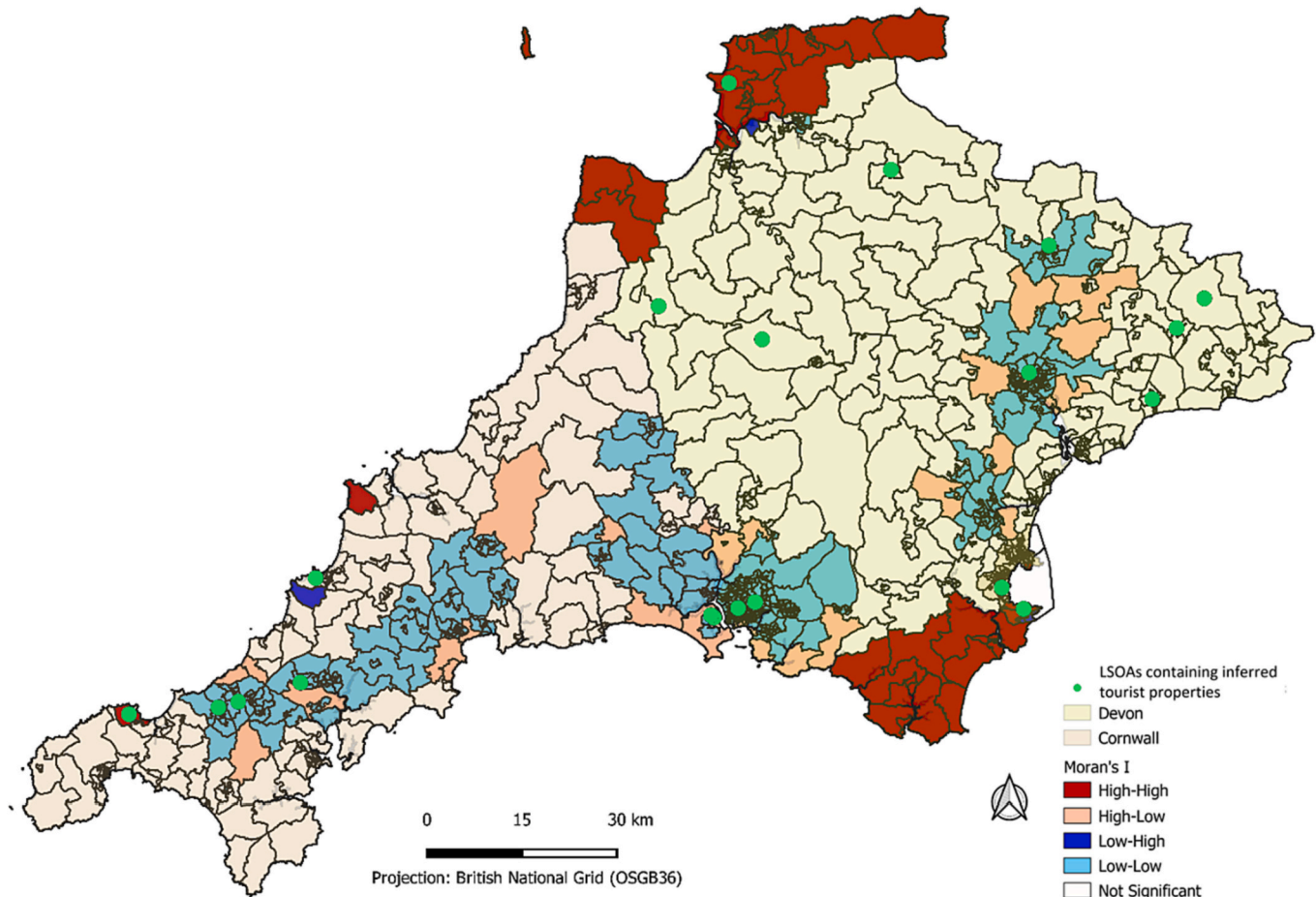
Fig. 5. Local Moran's I capturing statistically significant clusters of high and lot tourism activity, with LSOAs containing inferred tourist properties also shown.

significant hot spots.

The outputs from Moran's I (Fig. 5) also allow us to identify statistically significant outliers, including neighbourhoods with low evidence of tourism activity which are surrounded by neighbourhoods with higher tourism activity. We could hypothesise that these outliers may represent neighbourhoods in which there is a propensity for residential dwellings to be converted into tourist lets (Halliday & Morris, 2022), but none of our inferred tourist properties fall within these neighbouhoods. Moreover, 9 LSOAs containing our inferred tourist properties fall within areas deemed to be 'not significant' and therefore not representing any form of localised cluster of tourism activity. Whilst properties inferred to be tourist properties do not show a clear propensity to be located within spatial clusters of tourism activity, we must acknowledge that are working with a relatively small subset of properties which may be too small to draw robust conclusions at the LSOA level.

Although the spatial correlation with clusters and hot spots of tourism activity is not as pronounced as we may have hoped, our analysis suggests that 21 of the properties within our our analysis (2.8%) have occupancy patterns consistent with non-residential usage. Data from the UK Government 'Council Taxbase' (DLUHC, 2023) suggests that, across Devon and Cornwall, the proportion of dwellings that are classed as second homes, including those available as tourist rental properties, is around 3.5%. Whilst our analysis has revealed fewer potential tourist properties in our data, this may be an artefact of both the coverage of our data and the difficulties in identifying the number of non-residential properties in Devon and Cornwall, discussed in more detail in the following section.

In spite of occupancy trends being less pronounced (due to the reduced impact of Covid-19), the analysis of the 2022 data has

demonstrated that it is possible to extract dwelling-level occupancy trends. These have been used to group properties by occupancy and identify a subset of properties with occupancy trends associated with tourism. As reflected on further in the following section, these findings are incredibly encouraging, especially the close correspondence between AirDNA occupancy rates and the occupancy rates for our inferred tourism properties.

5. Discussion and wider value

The analysis presented above provides very encouraging evidence that these data can reveal occupancy characteristics which enable us to infer dwelling type. Our contacts at SWW and their subsidiary Bristol Water have shown considerable interest in our analysis. They have an important role in balancing the demand for and supply of water, which includes a need to forecast periods of increased seasonal demand due to tourism, as laid out in Water Resources Planning Guidance (House of Commons Treasury Committee, 2008). Our interest in this section is on the limitations of these data and our analysis, alongside a detailed consideration of the wider re-use value of these data beyond the water sector.

Our experience suggests that water companies can deliver these data in near-real time, with data for the period ending 30th Sept 2022 delivered to the project team on 5th October 2022, with little time lag required for internal data processing prior to release. Given the geographical monopolies present in the residential water supply sector, these data also offer considerable advantages over other commercial data sources (including similar data from energy companies) where typically households are distributed across a range of companies.

Alongside these advantages (and the sector-specific benefits presented in section 2), we must acknowledge that the deployment of AMI within the domestic water supply sector is not as advanced as in the electricity sector (Gill, 2022), where much of the previous work on occupancy monitoring has taken place. However, as noted in section 2, strong government encouragement is driving more rapid roll out of smart metering among many water companies. Whilst these data thus present many potential advantages, our analysis is subject to a number of limitations, most notably the lack of ground truth data available for validating the true occupancy status of any of the study properties. Nevertheless, the close correspondence between our inferred tourist property occupancy rates and the corresponding occupancy rates for tourist properties in Devon and Cornwall (see section 4.2) is very encouraging. The collation of validation data would be crucial for any future study utilising the datasets outlined below. Additionally, the small number of inferred tourist properties within our dataset has limited our ability to undertake more detailed spatial analysis in relation to localised occupancy rates (as reported by AirDNA) for comparable tourist properties. We consider this to represent a priority for further analysis, utilising one of the larger datasets introduced below.

Our analysis assumes that a property retains its identified status (residential or tourist) throughout the period of analysis. Given the propensity for residential dwellings to be converted to short term tourist lets (Halliday & Morris, 2022) and the relative ease of entry to the self-catering sector (using platforms such as 'Airbnb'), it is reasonable to assume that some properties may have changed status during our analysis period. Whilst not reported here, this is an area of investigation that we considered during our analysis and with a larger sample of properties we believe it would be feasible to identify properties that show occupancy characteristics that are consistent with different clusters at different time points in order to identify change of usage. Similarly, it is reasonable to assume that some residential properties will have undergone a change in occupier (e.g. as a result of a property sale/purchase) during the study period. We have not been able to capture this in our analysis but the detection of a change in occupier could represent an interesting application of these data, which we briefly reflect on below.

We have carefully considered how ONS could unlock additional value from these data, sharing our approaches and findings with the ONS' Methodology Hub throughout this work. The Methodology Hub provides statistical support to all ONS business areas and develops innovative methods suitable for use in the production of official statistics. They have a long-standing interest in the potential that administrative and commercial data sources could offer to supplement ONS' traditional (predominantly census and survey-based) sources of population, housing and tourism statistics (ONS, 2003, ONS, 2014, ONS, n.d). Previous ONS commissioned work has highlighted the potential value of commercial data sources held by utility providers in the energy and water sectors (Anderson & Newing, 2015; Dugmore, 2009) but this is the first study to explicitly make recommendations on the use of smart-meter derived water consumption data in this context.

As noted in section 1, tourists staying in self-catering accommodation are poorly captured within official statistics. Whilst relatively novel data sources, including the AirDNA data used within our analysis, provide information on the location and recorded occupancy rates for self-catering accommodation registered on those platforms, this remains a difficult to capture subset of visitor accommodation when predominantly drawn from the housing stock. An ONS review of their travel and tourism statistics recognised that many of the existing surveys of tourism activity lacked the timeliness, accuracy, coverage or level of disaggregation required by end users (ONS, 2022b). ONS work in this area has focused on international tourism statistics (volume and value of inbound and outbound visits) and the demand side (spend, duration of stay, number of visits). That work has explored the potential to supplement survey-based sources of tourism data with mobile phone data (capturing volume of visits) and financial transactions data (tourist spend) (ONS,

2022b). We suggest that further consideration could be given to the tourism supply side, with our analysis demonstrating that water metering data could provide new indicators of the tourist dwelling stock and their occupancy rates.

Although our interest is in seasonal occupancy patterns driven by tourism, it is entirely feasible to use our approach to identify other dwelling characteristics and events which are related to occupancy. These may include a change in residents within a residential dwelling (e. g. a house sale or change in tenancy), identifiable via a period of non-occupancy and/or a change in occupancy patterns, affording new and timely insights into small-area population change. Subject to further consideration of privacy and ethical concerns, alongside the challenges of mining these data at scale, they could offer near real time insights into dwelling-level occupancy, addressing questions such as 'was this property/properties in this area occupied on a particular date?' or, 'are properties in this area typically occupied on a given day of the year'?. We should note, however, that we are aware of no current plans by ONS to attempt to monitor dwelling-level occupancy in that level of detail.

These data and approaches could support ONS' 'Future of Population and Social Statistics work package' (ONS, n.d.), which seeks to transform ONS' population, migration and social statistics via the provision of more frequent and detailed statistics using administrative data soruces. One specfic administrative data source of interest to the ONS is held by the Valuation Office Agency (VOA) (responsible for banding properties for tax purposes) and captures property address alongside attributes such as size (e.g. number of bedrooms and floor area) and age (ONS, 2023). The ONS report that these data could provide more frequent census-like housing statistics, via an experimental Admin Based Housing Stock (ABHS) dataset (ONS, 2022a). However, they note challenges in distinguishing between occupied and vacant residential properties using the VOA data (ONS, 2022a). Use of smart-meter derived water consumption data, coupled with our NIOM approach could offer considerable value here in identifying property occupancy status if these datasets could be linked, for example using Unique Property Reference Numbers (UPRN). Discussion with out contacts at ONS suggests that this could be a feasible next step in these analyses.

Wider uses of our approach could support the UK Government 'Levelling-up' campaign to maximise employment opportunities and living standards across the UK. Levelling-up has recognised that second homes (especially when used as a holiday let) benefit tourism in many localities, but that they can also price others out of the housing market, especially in major tourist areas where wages are typically lower (House of Commons Library, 2022; ONS, 2021). The 'Levelling up and Regeneration Bill' proposes doubling council tax (a dwelling-level tax collected by the local authority and shared among organisations providing local and regional services) paid by owners of second homes in order to discourage their use as tourist accommodation and promote affordable homes for local residents (DLUHC, 2022a). Whilst empty dwellings are captured by the decadal census (i.e. households with no usual residents), the status of a dwelling as a second/holiday home is not recorded. Some information on the second home dwelling stock is held by local authorities (primarily collated from council tax records) yet this is not usually available for analysis at small-area geographies. Furthermore, many second homes are also part of the rental accommodation stock and are therefore registered as business premises and excluded from council tax records. We strongly suggest that our analysis – if appropriately up-scaled - could identify the number and location of short term tourist rental properties in order to enact new housing policy in this area (DLUHC, 2022b).

Our ongoing work seeks to upscale these analysis to a larger collection of dwellings. Potential datasets include the 35,000+ smart meters that SWW will install in North Devon (coinciding with one of our hot spots of tourism activity) as part of their Green Recovery Initiative (SWW, 2022). We are also keen to extend our analysis to other water companies which could include Thames Water or Severn Trent Water. Thames were the first UK water company to install smart meters at scale

and currently have at least 620,000 instillations, collecting data at an hourly resolution (Baker, 2022). Their supply area includes London and could offer an opportunity to unpick the more complex range of dwelling types and occupancy patterns that may be evident. Severn Trent Water have an ambitious plan to install over 150,000 smart meters in the cities of Coventry and across the county of Warwickshire, creating a 'smart water region' (STWater, 2023). The higher density of meters in these areas could allow us to capture a higher proportion of dwellings and could offer scope to assess the potential to use our approaches to generate experimental area-based statistics capturing dwelling occupancy.

6. Conclusions

The analysis reported in this paper sought to assess the feasibility of using high temporal resolution water consumption data to identify dwelling-level occupancy, as a proxy for tourism activity. Specifically we used these data to infer the presence of dwellings with occupancy patterns consistent with usage as second homes or short-term rental tourist accommodation. The continued rollout of AMI in the residential water supply sector means that these data will become more routinely available. They are collected by water companies using non-intrusive means, and could be delivered to end-users in near-real time, offering tremendous re-use potential beyond their intended purpose in driving water-use efficiencies, accurate billing and network management. Our engagement with the ONS throughout these analysis highlight wider interest in these forms of data. We strongly assert that these data could afford potential as indicators of area-based housing status and tourism activity.

Our analysis has been carried out in relation to dwellings in Devon and Cornwall, using data supplied by SWW. Whilst our analysis has used these data in a non-personally identifiable and ethically compliant manner, their linkage to specific identifiable dwellings, in order to facilitate the type of analysis suggested in section 5, would require further assessment of privacy and ethical concerns. Our analysis reveals that these data can be used to identify property-level occupancy trends, highlighted by our ability to pull out unique dwelling-level usage characteristics exhibited during the Covid-19 period. We have been able to infer a set of properties that exhibit occupancy characteristics which may be associated with tourism and demonstrate that these show some correspondence with underlying indicators of tourism activity. Our sample of properties is likely too small to effectively assess the location of these properties in relation to hot spots of tourist activity, and this is an area where we recommend additional work with a larger sample of properties.

Whist our data has not enabled us to validate these findings (we do not know the true status of any of these properties), we strongly assert that the findings of this study add considerable novelty and value to a range of stakeholders. It provides further evidence of the potential of high temporal resolution household-level data collected by utilities providers, extending the work of Anderson et al. (2016) to the water sector. It highlights the wider potential re-use value these data to water suppliers including SWW, adding further weight to the notion that these data could be used to generate 'smart water regions' (STWater, 2023) and 'digital twins' (Coates, 2023) with applications extending beyond the management of water supply networks. Our primary impact, however, is in the potential application of these data in the generation of area-based housing and tourism statistics, as highlighted by the involvement of the ONS at every stage of this project.

We recommend further work in a UK and international context, bringing together water companies, academia and organisations such as the ONS to explore the potential these data could offer in the generation of area-based housing and tourism statistics. The value of near real time dwelling-level occupancy insights could extend beyond tourism to include other forms of non-standard dwelling including second homes or student residences. If up-scaled to a larger sample of properties

(specifically with some form of geo-reference such as postcode or UPRN in a UK context), it would be entirely feasible for this approach to provide additional indicators of neighbourhood characteristics related to dwelling occupancy and utilisation. Our analysis is UK-centric, benefitting from data from SWW and the interest and methodological insight provided by the ONS. However, wider roll out of AMI internationally provides considerable scope for comparable analysis.

CRediT authorship contribution statement

A. Newing: Conceptualization, Methodology, Formal analysis, Investigation, Validation, Writing – original draft, Resources, Project administration, Funding acquisition. **O. Hibbert:** Methodology, Formal analysis, Investigation, Writing – review & editing, Data curation. **J. Van-Alwon:** Methodology, Formal analysis, Investigation, Writing – review & editing. **S. Ellaway:** Methodology, Validation, Writing – review & editing. **A. Smith:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

At the time of analysis, Stuart Ellaway was employed as a Research Fellow at the University of Plymouth. He has subsequently taken up employment with South West Water, a data supplier for this project. The remaining authors have no competing interests to declare.

References

Abbott, O. (2018). *Beyond the traditional - Data Science in Official Statistics. Seminar delivered at Leeds Institute for Data Analytics (LIDA), 22nd February 2018*.

Abu-Bakar, H., Williams, L., & Hallett, S. H. (2021). An empirical water consumer segmentation and the characterisation of consumption patterns underpinning demand peaks. *Resources, Conservation and Recycling, 174*, Article 105792.

AirDNA. (2023). Glossary of Terms [Online]. Available: https://www.airdna.co/airdna-glossary-of-metric-definitions [Accessed 2nd June 2023 2023].

van Alwon, J., Newing, A., Smith, A., Ellaway, S., Hibbert, O., & Merchant, P. (2022). *Inferring tourist dwelling characteristics and occupancy patterns using high temporal resolution water metering data. Technical report prepared for the ONS*. Leeds: University of Leeds.

Anderson, B., Lin, S., Newing, A., Bahaj, A., & James, P. (2016). Electricity consumption and household characteristics: Implications for census-taking in a smart metered future. *Computers, Environment and Urban Systems, 63*, 58–67.

Anderson, B., & Newing, A. (2015). *Using Energy Metering Data to Support Official Statistics: A Feasibility Study*. Southampton: Sustainable Energy Research Group, University of Southampton.

Anselin, L. (2010). Local Indicators of Spatial Association-LISA. *Geographical Analysis, 27*, 93–115.

Baker, P. (2022). *The evidence base is burgeoning for investment in AMI to make the cut at PR24 [Online]*. Chalfont: Arqiva. Available: https://www.arqiva.com/news-views/news/smart-money-investment-in-ami-to-make-the-cut-at-pr24 [Accessed 9th May 2023 2023].

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*.

Becker, V., & Kleiminger, W. (2017). Exploring zero-training algorithms for occupancy detection based on smart meter measurements. *Computer Science - Research and Development, 33*, 25–36.

Britton, T., Cole, G., Stewart, R., & Wiskar, D. (2008). Remote diagnosis of leakage in residential households. *Water, 35*, 56–60.

Carboni, D., Gluhak, A., Mccann, J. A., & Beach, T. H. (2016). Contextualising water use in residential settings: A survey of non-intrusive techniques and approaches. *Sensors (Basel), 16*.

Cardell-Oliver, R. (2013). Water use signature patterns for analyzing household consumption using medium resolution meter data. *Water Resources Research, 49*, 8589–8599.

Cardell-Oliver, R., Wang, J., & Gigney, H. (2016). Smart meter analytics to pinpoint opportunities for reducing household water use. *Journal of Water Resources Planning and Management, 142*, 04016007.

Chen, D., Barker, S. K., Subbaswamy, A., Irwin, D. E., & Shenoy, P. J. (2013). *Non-Intrusive Occupancy Monitoring using Smart Meters*.

Clifford, E., Mulligan, S., Comer, J., & Hannon, L. (2018). Flow-signature analysis of water consumption in nonresidential building water networks using high-resolution and medium-resolution smart meter data: Two case studies. *Water Resources Research, 54*, 88–106.

Coates, M. (2023). Smart metering: a step towards water's digital twin? [Online]. Available: https://www.bimplus.co.uk/smart-metering-a-step-towards-waters-digital-twin/ [Accessed 23rd May 2023 2023].

Cominola, A., Nguyen, K., Giuliani, M., Stewart, R. A., Maier, H. R., & Castelletti, A. (2019). Data mining to uncover heterogeneous water use behaviors from smart meter data. *Water Resources Research, 55*, 9315–9333.

Cominola, A., Spang, E. S., Giuliani, M., Castelletti, A., Lund, J. R., & Loge, F. J. (2018). Segmentation analysis of residential water-electricity demand for customized demand-side management programs. *Journal of Cleaner Production, 172*, 1607–1619.

Craig, T., Polhill, J. G., 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.

DEFRA. (2023a). *Environmental Improvement Plan 2023*. London: Department for Environment, Food and Rural Affairs.

DEFRA. (2023b). *Plan for Water: our integrated plan for delivering clean and plentiful water*. London Department for Environment Food & Rural Affairs.

DLUHC. (2022a). *Gove closes tax loophole on second homes [Online]*. London: Department for Levelling Up, Housing and Communities. Available: https://www.gov.uk/government/news/gove-closes-tax-loophole-on-second-homes [Accessed 26th May 2023 2023].

DLUHC. (2022b). *New holiday let rules to protect local people and support tourism [Online]*. London: Department for Levelling Up, Housing and Communities. Available: https://www.gov.uk/government/news/new-holiday-let-rules-to-protect-local-people-and-support-tourism [Accessed 26th May 2023 2023].

DLUHC. (2023). *Local authority Council Taxbase in England: 2022 (revised) [Online]*. London: Department for Levelling Up, Housing and Communities. Available: https://www.gov.uk/government/statistics/council-taxbase-2022-in-england/local-authority-council-taxbase-in-england-2022 [Accessed 26th May 2023 2023].

Dugmore. (2009). *Information collected by commercial companies: What might be of value to ONS?* London: Demographic Decisions Ltd.

Eibl, G., Burkhart, S., & Engel, D. (2018). Unsupervised holiday detection from low-resolution smart metering data. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy - ICISSP*. SciTePress.

European Commission. (2023). Web Intelligence Network: Work Package 3 - New use-cases [Online]. Available: https://cros-legacy.ec.europa.eu/content/work-package-3-%E2%80%93-new-use-cases_en [Accessed 25th May 2023 2023].

Frontier Economics, Artesia & Arqiva. (2021). *Cost benefit analysis, assessing the social and environmental case for a smart water meter rollout*. London: Frontier Economics, Artesia and Arqiva.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*, 189–206.

Gill, J. (2022). *Smart water meter innovation key to taking control of consumption [Online]*. Utility Week. Available: https://utilityweek.co.uk/smart-water-meter-innovation-key-to-taking-control-of-consumption/ [Accessed 24th April 2023 2026].

Halliday, J., & Morris, S. (2022). Holiday homes are 'hollowing out' coastal areas, says MP. *The Guardian*, 18th June 2022 https://www.theguardian.com/society/2022/jun/18/holiday-homes-are-hollowing-out-coastal-areas-says-mp accessed on 1st June 2023.

House of Commons Library. (2022). *Reasearch Briefing: Second homes and holiday-lets in rural communities*. London: UK Parliament.

House of Commons Treasury Committee. (2008). *Counting the population: Eleventh Report of Session 2007-08. Volume 1, Report, together with formal minutes, oral and written evidence*. London: The Stationery Office Limited.

Jacobs. (2023). *Smart Water Metering [Online]*. London: Jacobs. Available https://www.jacobs.com/solutions/markets/infrastructure/water/smart-water-metering [Accessed 1st June 2023 2023].

Johns, N., & Lynch, P. (2007). The self-catering accommodation market: A review of electronic and other sources. *Hospitality Management, 26*, 293–309.

Koech, R., Cardell-oliver, R., & Syme, G. (2021). Smart water metering: adoption, regulatory and social considerations. *Australasian Journal of Water Resources, 25*, 173–182.

Kowalska, B., Suchorab, P., & Kowalski, D. (2022). Division of district metered areas (DMAs) in a part of water supply network using WaterGEMS (Bentley) software: a case study. *Applied Water Science, 12*.

Laspidou, C., Papageorgiou, E., Kokkinos, K., Sahu, S., Gupta, A., & Tassiulas, L. (2015). Exploring patterns in water consumption by clustering. *Procedia Engineering, 119*, 1439–1446.

López-Rodríguez, M. A., Santiago, I., Trillo-Montero, D., Torriti, J., & Moreno-Munoz, A. (2013). Analysis and modeling of active occupancy of the residential sector in Spain: An indicator of residential electricity consumption. *Energy Policy, 62*, 742–751.

March, H., Morote, Á.-F., Rico, A.-M., & Saurí, D. (2017). Household smart water metering in Spain: Insights from the experience of remote meter reading in alicante. *Sustainability, 9*.

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2019). Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems, 74*, 161–174.

Mckenna, E., Richardson, I., & Thomson, M. (2012). Smart meter data: Balancing consumer privacy concerns with legitimate applications. *Energy Policy, 41*, 807–814.

Mcloughlin, F., Duffy, A., & Conlon, M. (2012). Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. *Energy and Buildings, 48*, 240–248.

Monks, Stewart, Sahin, & Keller. (2019). Revealing unreported benefits of digital water metering: Literature review and expert opinions. *Water, 11*.

Newing, A. (2013). *Incorporating seasonal visitor demand in retail location modelling* (Doctor of Philosophy). University of Leeds.

Newing, A., Anderson, B., Bahaj, A., & James, P. (2016). The role of digital trace data in supporting the collection of population statistics – the case for smart metered electricity consumption data. *Population, Space and Place, 22*, 849–869.

OFWAT. (2023). *Your Water Company [Online]*. London: Ofwat. Available: https://www.ofwat.gov.uk/households/your-water-company/ [Accessed 9th May 2023 2023].

ONS. (2003). *Census strategic development review - Alternatives to a Census: Linkage of existing data sources*. Newport: Office for National Statistics.

ONS. (2014). *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*. Newport: Office for National Statistics.

ONS. (2021). *House prices in tourist hotspots increasingly out of reach for young and low paid [Online]*. Newport: Office for National Statistics. Available: https://www.ons.gov.uk/economy/inflationandpriceindices/articles/housepricesintouristhotspotsincreasinglyoutofreachforyoungandlowpaid/2021-09-28 [Accessed 26th May 2023 2023].

ONS. (2022a). *Developing admin-based housing stock statistics for England and Wales: 2020 [Online]*. Newport: Office for National Statistics. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/developingadminbasedhousingstockstatisticsforenglandandwales/2020 [Accessed 26th May 2023 2023].

ONS. (2022b). *Travel and tourism review: final report*. Newport: Office for National Statistics.

ONS. (2023). *Valuation Office Agency data [Online]*. Newport: Office for National Statistics. Available: https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/censusanddatacollectiontransformationprogramme/futureofpopulationandsocialstatistics/datasourceoverviews/valuationofficeagencydata [Accessed 26th May 2023 2023].

ONS. n.d. Future of population and social statistics: Our progress towards more regular and responsive statistics, building the richest picture of our population. [Online]. Newport: Office for National Statistics. Available: https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/censusanddatacollectiontransformationprogramme/futureofpopulationandsocialstatistics [Accessed 23rd May 2023 2023].

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

Owen, P. (2012). *Powering the nation: Household electricity habits revealed*. London: Energy Saving Trust.

Richardson, I., Thomson, M., Infield, D., & Clifford, C. (2010). Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings, 42*, 1878–1887.

Sadr, S. M. K., That, L. T., Ingram, W., & Memon, F. A. (2021). Simulating the impact of water demand management options on water consumption and wastewater generation profiles. *Urban Water Journal, 18*, 320–333.

Scanlon, K., Sagor, E., & Whitehead, C. (2014). *The economic impact of holiday rentals in the UK: Final Report*. LSE: London School of Economics.

Shaw, D. (2021). Airbnb boss: 'Cornwall's more popular than London'. *BBC News, 25th May 2021 [Online]*. London: BBC. Available: https://www.bbc.co.uk/news/technology-57240212 [Accessed 9th May 2023 2023].

Singleton, A. D., & Longley, P. (2015). The internal structure of Greater London: a comparison of national and regional geodemographic models. *Geo: Geography and Environment, 2*(1), 69–87.

Sønderlund, A. L., Smith, J. R., Hutton, C., & Kapelan, Z. (2014). Using smart meters for household water consumption feedback: Knowns and unknows. *Procedia Engineering, 89*, 990–997.

Steffelbauer, D. B., Blokker, E. J. M., Buchberger, S. G., Knobbe, A., & Abraham, E. (2021). *Dynamic Time Warping Clustering to Discover Socioeconomic Characteristics in Smart Water Meter Data., 147*, 04021026.

STWater. (2023). *Coventry and Warwickshire gets smart with water!* Coventry: Severn Trent Water.

SWW. (2022). *Green Recovery Initiative*. Exeter: South West Water.

SWW. (2023). What we do [Online]. Available: https://www.southwestwater.co.uk/wholesale/what-we-do/south-west-water/ [Accessed 25th May 2023 2023]..

Waterwise. (2022). *UK Water Efficiency Strategy to 2030*. London: Waterwise.

White, S. (2010). *Measuring tourism locally: guidance note five: Measuring the supply side of tourism*. Newport: Office for National Statistics.

Widén, J., & Wäckelgård, E. (2010). A high-resolution stochastic model of domestic activity patterns and electricity demand. *Applied Energy, 87*, 1880–1892.

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.