Inferring tourist dwelling characteristics and occupancy patterns using high temporal resolution water metering data

A technical report prepared for the Office for National Statistics as part of the ESRC funded 'WatPop: understanding seasonal population change' research project

Jacob van Alwon[1], Andy Newing[1*, 2], Alan Smith[3≠], Stuart Ellaway[3], Owen Hibbert[1] and Paul Merchant[4]

[1] Leeds Institute for Data Analytics (LIDA) University of Leeds, Leeds, UK, LS2 9JT

[2] School of Geography, University of Leeds, Leeds, UK, LS2 9JT

[3] School of Geography, Earth and Environmental Sciences, University of Plymouth, Plymouth, PL4 8AA

[4] South West Water, Peninsula House, Exeter, EX2 7HR

[*] Corresponding Author: a.newing@leeds.ac.uk, 0113 343 6720

[≠] Project Principal Investigator (PI)

October 2022

EXECUTIVE SUMMARY

This technical report has been prepared for the Office for National Statistics (ONS), reporting on findings and implications from a research project undertaken as part of a Leeds Institute for Data Analytics (LIDA) Data Science Development Programme. It forms part of a wider Economic and Social Research Council (ESRC) funded two-year research project 'WatPop: Understanding seasonal population change' undertaken by the Universities of Plymouth and Leeds in conjunction with South West Water (SWW).

The research reported here explicitly sought to address ONS' needs in relation to the development of innovative methods suitable for use in the production of official statistics. We use anonymised dwelling-level water consumption data (collected pre COVID-19, at a 15-minute temporal resolution) to infer dwelling occupancy characteristics associated with tourism. Specifically, we develop analytic approaches which can i) Infer occupancy patterns for tourist dwellings, and; ii) distinguish those tourist dwellings from residential dwellings.

We demonstrate that high temporal resolution household level water meter data could offer considerable potential as a tool to support the identification of residential dwellings used primarily as a holiday rental property or second home, and to understand their seasonal occupancy patterns. Based on our small sample of exemplar data, dwelling-level characteristics and occupancy can be inferred using computationally and conceptually straightforward approaches, including k-means clustering, drawing on key-indicators of dwelling consumption based on the magnitude, range and seasonal variations in water consumption.

These data, which are akin to those collected by smart meters being widely rolled out across the domestic water sector, could offer near-complete coverage of all properties in a given small-area, sourced from a single regional supplier. However, data pre-processing requirements should not be overlooked. Whilst leakage detection and correction is conceptually straightforward, the volume of data collected (~35,000 readings per property per year at a 15 minute resolution) and the need to identify and account for missing data (e.g. meter downtime and recording errors) present ongoing challenges. We also present some tentative findings which suggest that comparable analysis and findings can be undertaken with data at a 1-hour resolution, considerably reducing the volume of data to be processed.

Previous ONS commissioned work has recognised the potential value of commercial data sources, including those held by utility providers in the energy and water sectors (Dugmore, 2009; Anderson and Newing, 2015). This is the first study to explicitly consider the role of dwelling-level water supply data as an indicator of tourism activity. We make a series of suggestions to ONS regarding further potential work in this area around population mobility, neighbourhood characteristics, dwelling type/occupancy and use of commercial data sources.

Using household level water meter data, the automated occupancy detection method we have developed is able to determine dwelling-level occupancy on a day-by-day basis, with an accuracy of 98.8% (falling very marginally to 98.7% when using data at the 1 hour resolution). In addition, we have proven that it is possible to distinguish dwelling type (residential vs. tourist) applying k-means clustering and indicators of occupancy alongside metrics capturing magnitude and seasonality of water consumption.

A second six-month Data Science Development Programme research project commenced in October 2022 and seeks to upscale these approaches to a larger sample of dwellings supplied by SWW. We are therefore particularly interested to understand ONS' priorities in relation to this work in order to drive the objectives for that period.

CONTENTS

1 INTRODUCTION

This report describes the findings of a six month research project undertaken as part of the Leeds Institute for Data Analytics (LIDA) Data Scientist Development Programme. This internship investigates the potential to identify dwelling-level characteristics and occupancy patterns from high temporal resolution water metering data. The dwelling-level characteristics of interest reported here are driven by tourism and utilise data related to properties in South West England. These data and the approaches and methods used could be applied to identify occupancy trends for other forms of dwelling (e.g. second homes or student dwellings) or occupancy-driven changes in dwelling-level water consumption, for example those driven by a change in tenancy of a rental property.

The research forms part of a larger funded study ('WatPop'), led by Dr Alan Smith (University of Plymouth) and carried out in collaboration with the regional statutory water authority, South West Water (SWW). 'WatPop: Understanding seasonal population change', represents a two-year, £300,000 investment by the Economic and Social Research Council's (ESRC) Secondary Data Analysis Initiative. WatPop aims to assess the feasibility of inferring small-area seasonal population fluctuations from dwelling and area-level data collected by water suppliers as part of the routine supply of water to domestic properties. Alongside the dwelling-level package of work reported here, the broader project draws extensively on data at the aggregate District Metering Area (DMA) level. DMAs are a contiguous set of metered water supply areas each supplying ~1,000 households (broadly akin to a Lower Super Output Area) and could reveal additional insights into hot spots of seasonal tourist-induced population fluctuation not captured within traditional statistics.

This report has been produced to address the needs and interests of the Office for National Statistics (ONS) Methodology Division. The Methodology Division provides statistical support to all ONS business areas and develops innovative methods suitable for use in the production of official statistics. One area of interest relates to the use of administrative and commercial datasets to assist in the identification of residential dwellings that are predominantly used as second homes or tourist accommodation, alongside wider interests around neighbourhood classification and population mobility as part of their remit to produce robust and timely small area population statistics. Previous ONS commissioned work has recognised the potential value of commercial data sources, including those held by utility providers in the energy and water sectors (Dugmore, 2009; Anderson and Newing, 2015). Nevertheless, we believe this is the first study to explicitly consider the role of dwelling-level water supply data as an indicator of tourism activity.

Utilising a sample of dwelling-level water consumption data at a 15-min temporal resolution, and repeated at a 1-hour resolution, we highlight the considerable potential to infer the occupancy patterns of dwellings associated with tourism (e.g. self-catering holiday lets), and to identify dwellings with occupancy and consumption patterns consistent with tourist usage. In the following sections of this report we introduce the characteristics of dwelling-level water consumption data and highlight some of the challenges involved in cleaning these data, including the need to detect and correct for leakage. We step through a series of analytic approaches which can be used to:

1. Infer occupancy patterns for tourist dwellings, and;
2. Distinguish tourist dwellings from residential dwellings.

We make a series of recommendations which could support the ONS in understanding dwelling and neighbourhood type, area classification, population mobility and the potential

value of administrative and commercial data sources within the provision of population and neighbourhood statistics.

2 DATA

This study makes use of dwelling-level water-metering data provided by project partner South West Water (SWW). These data are collected from high temporal resolution data loggers fitted to the supply pipe of a sample of residential properties as part of an ongoing SWW household-level data collection exercise to better understand water demand and supply. Although these data loggers are different to the typical 'smart meters' being rolled out within this sector (across both residential and commercial premises), the high temporal resolution data collected is of a comparable format and temporal resolution to the smart meter consumption data captured at a dwelling-level by SWW and other water suppliers. The data provided thus enables us to assess the potential use of smart-meter data to capture dwelling-level indicators of tourist activity. SWW has one of the highest levels of metering within the UK water industry with over 80% of properties in their network metered (SWW 2018), with ongoing roll out of smart meters by all suppliers as a tool for leakage detection, more accurate billing and customer-led consumption awareness and reduction.

The high frequency data provided by smart meters can provide a number of benefits to both water suppliers and end users when compared with traditional 'dumb' meter data, which may only record water consumption on a monthly, quarterly or biannual schedule. Utilities companies can use smart water meter data for leak detection (Koech, Cardell-Oliver, and Syme, 2021), to reduce costs and better understand water demand to streamline services and improve forecasting (March et al., 2017) and to inform water saving policies (Sadr et al., 2021). End users are able to receive more accurate billing based on actual, rather than estimated, usage and are able to monitor their own usage behaviour in order to help make financial savings and reduce their environmental impact (Sønderlund et al., 2014). Whilst these benefits are regularly cited, the wider potential re-use value of these data is underreported.

Given that all domestic water is provided by a single authority with a monopoly in their supply region[1], water providers benefit from near-complete coverage of the water supply to residential properties within their geographical jurisdiction. The rollout of smart meters to domestic properties therefore provides an opportunity for data collection that offers near-complete coverage of all residential dwellings in a given locality from a single supplier. Previous ONS interest in this domain has considered data collected via electricity smart meters (see Anderson and Newing, 2015), yet the greater fragmentation of suppliers in that sector makes near-complete coverage of dwellings more difficult to achieve. Similarly, unlike other utilities (such as electricity and gas) water is generally not consumed when properties are unoccupied. Even dishwashers and washing machines have time limited cycles as opposed to a fridge, freezer, security lighting or heating/cooling systems that can operate continuously when no occupants are present. We therefore suggest that water consumption data offer a strong advantage over electricity data due to the nature of a single regional supplier for data access and limited possibilities, excluding easy to detect leaks (see below), for usage whilst properties are unoccupied.

High temporal resolution dwelling level data – akin to the data used in this report – are routinely collected by water suppliers as part of the network management and billing process. The meters themselves and the data collected from them are the property of the water supplier (rather than the individual household). With most water meters situated on the public highway or accessible via a public right of way, collection of these data are non-invasive, and can be

[1] A small number of properties are connected to a private supply

legally and legitimately collected by water suppliers or their agents as part of their statutory requirement to bill, manage and maintain water supply within their network area under the Water Act 1973. We therefore suggest that there may be considerable wider-reuse value to dwelling-level metering data collected by the water sector, and highlight one such potential use-case in this study. We aim to demonstrate the potential value of high frequency water consumption data as a tool to support the identification of residential dwellings used primarily as a holiday rental property and to understand their seasonal occupancy patterns.

For this study, SWW provided water consumption data at 15-minute intervals over an extended period (up to ~4 years, ending 2019[2]) for 93 properties. Of the 93 properties, SWW in-house intelligence[3] suggested that seven (7.5%) are very likely to represent self-catering tourist rental properties. Whilst change-of-use may mean that some properties are incorrectly categorised, SWW intelligence alongside our interrogation of these properties' water consumption gives us a high level of confidence in their status and dwellings will be referred to as either 'residential' or 'tourist' throughout this report.

Prior to analysis, pre-processing was required in order to convert dwelling-level consumption records into an appropriate format in relation to data temporal coverage and completeness. Given the seasonal nature of tourism in the study region and the methods applied, we required temporal coverage to capture a one-year period, although the start date of that yearlong period was not important. We excluded all properties for which one years' worth of continuous data could not be extracted. For those properties with greater than 12 months' worth of data we extracted only one years' continuous data. Potential future work could seek to assess the viability of these analysis methods to extract occupancy trends from data of variable temporal durations, although it is likely that a minimum of one year of data would be required given the seasonal dimension to tourism-driven occupancy patterns.

Within the dataset, 32 properties did not meet this minimum data standard due to missing data, characterised by extended periods of zero water consumption, indicating signal loss from the water meter. Figure 2.1(a) shows the total daily water consumption over a one-year period for an example property characterised by a lost signal for an extended period (8 months). Any property which had lost signal, as characterised below, so as not to provide at least 1 year of continuous data was removed from the dataset. The presence of tourist dwellings within our dataset means that some properties will have legitimate periods of zero consumption – these are the very features we seek to identify. Removal of properties with lost signal – which also manifests itself as periods of zero consumption – thus runs the risk of excluding the very properties we seek to identify. A manual process was therefore used, accounting for:

- The length of the period with zero consumption, as periods of lost signal were generally far longer than periods of dwelling non-occupancy.
- The time of year when the zero consumption occurred, as zero consumption over the summer and continuous consumption over the winter would likely indicate a lost signal rather than a tourist property.
- The number of periods of zero consumption, as periods of lost signal generally occurred only a small number of times (once in most cases) over the measurement

[2] All data used within this report are pre COVID-19 and so the methods and approaches developed are suitable for ongoing usage and are not dependent on the specific household behaviours exhibited during lockdown periods.

[3] Determined with reference to registered supply and billing addresses, consultation with local authorities and information obtained from online listings/advertisement of tourist properties.

period, whereas, for tourist properties there were generally many periods of shorter duration vacancy, between usage indicating occupancy.

Manual removal of properties with lost signal was achievable for this relatively small dataset and represented an important data-familiarisation step. One of the future aims of this research is to develop a method to automate the identification and removal of properties with lost signal.

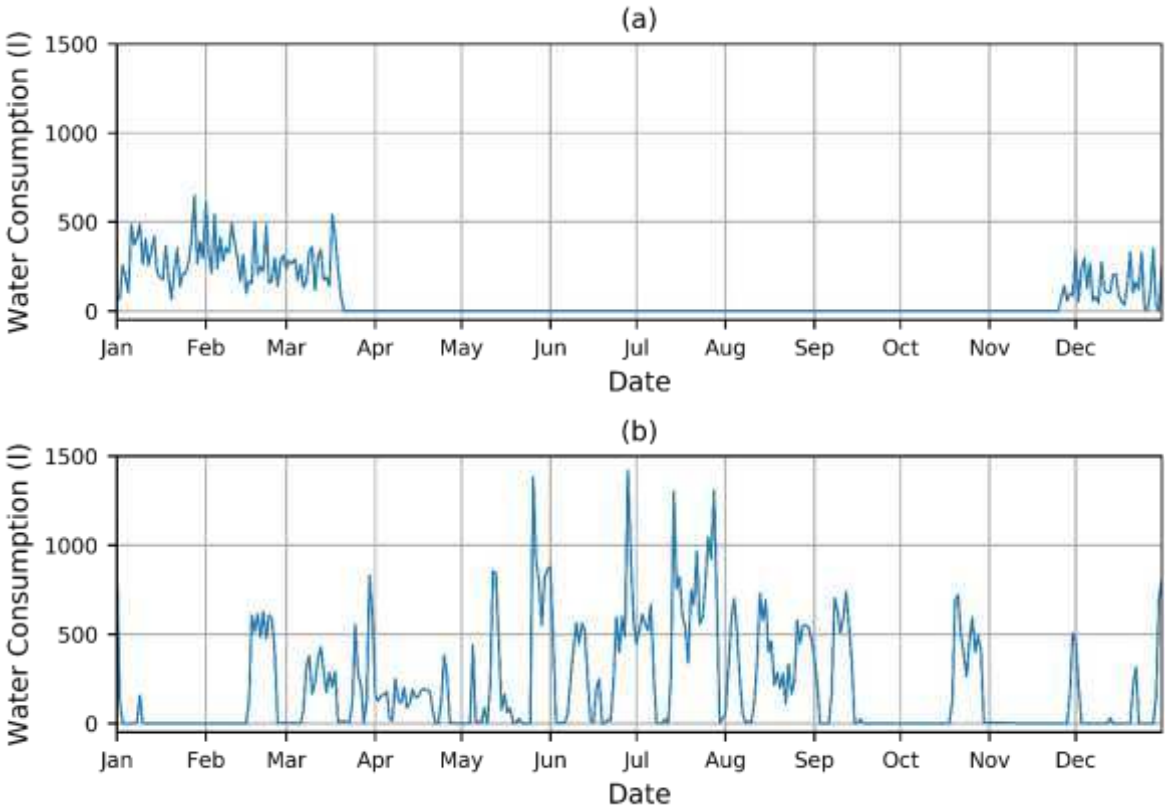


Figure 2.1 – Total daily water consumption over a year period for: (a) a property with a lost signal from the water meter for an extended period (b) a tourist property.

These processing steps produced a dataset of 15-minute water consumption records over exactly one calendar year for 61 properties, four of which SWW were confident were tourist properties. Throughout the following sections of this report we present a number of dwelling-level water consumption profiles, typically illustrated over a one-year or 24-hour period. Figure 2.2 presents two typical profiles which seek to introduce the features typically evident within these data at a 15-minute resolution. We also aggregated dwelling-level consumption records to a 1-hour temporal resolution, serving the dual purpose of reducing data volume (24 readings per property per day as opposed to 96 readings per property per day) and smoothing these data. We briefly assess whether data at the 1-hour resolution could offer comparable insights, whilst considerably reducing data processing requirements.

Figure 2.2(a) shows a typical one-year profile of total daily consumption for an exemplar residential property from within our data. Fluctuating but broadly consistent consumption is evident, with two short periods of zero consumption evident in August and November, suggesting that the property was unoccupied during these periods, potentially due to a summer holidays and weekend break. Figure 2.2(b) shows a typical daily consumption profile of 15-minute readings for an occupied property. The typical profile shows little or no water consumption during the night, followed by a peak in the morning as occupants wake up and start to use appliances such as toilets and showers. Subsequently there is less use during the

day working daytime hours, followed by another peak in the evening. Although this profile is typical, the occurrence, timing and magnitude of the peaks and troughs in consumption vary from day to day as well as from property to property.

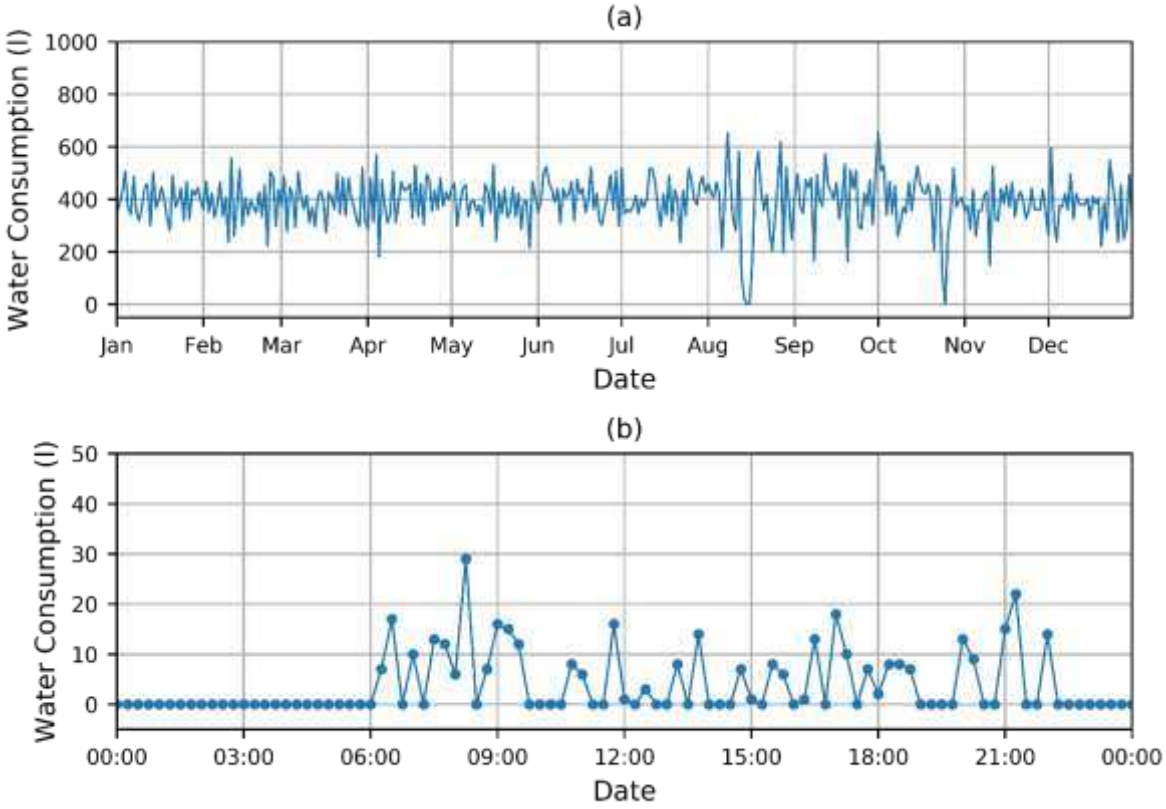


Figure 2.2 – Typical water consumption: (a) daily total over one year for a residential property (b) 15 minute consumption over 1 day for an occupied day.

As outlined in this section, almost a third of the supplied dwellings have been removed from subsequent analysis due to data coverage or completeness that falls-short of minimum requirements given the context of seasonal population fluctuation. Whilst these data thus afford tremendous potential (as highlighted in Sections 4 and 5) the degree of data pre-processing and cleaning required should not be underestimated, alongside the inevitable loss of data due to failure of the wireless recording technology. In the following section we outline a further data-preparation stage which seeks to identify and correct for leakage. In future work beyond this project we hope that it might be possible to make better use of even partial household usage profiles. However, this depends on the desired outputs. As a proof of concept and for confidence in this study, properties with incomplete data signals were removed.

3 LEAK DETECTION AND CORRECTION

It is common for water leaks to occur in residential dwellings, which will be recorded by smart water meters. Water leakage within a dwelling causes the high temporal resolution water consumption data to show a reasonably constant “baseline” water consumption at a non-zero value, with additional water consumption by residential usage also recorded, typically as short-term peaks (Figure 3.1). In order to prevent the additional recorded water consumption caused by leakage from affecting the analysis methods applied in this study, leakages observed within our dwelling-level data were identified and removed from the data. By removing the baseline water consumption caused by leakage, the water consumption magnitude and profile due to residential consumption was preserved.

Figure 3.1 shows the daily water consumption over a one year period for a property which had a water leak, as well as the 15 minute water consumption profile on a specified day for which leakage occurred (2nd July). The yearly water consumption profile shows that the volume of water consumed daily tended to increase gradually up until 21st July, at which point the water consumption reduced considerably and the daily consumption was generally lower thereafter. This is a typical water consumption profile for a property with a leak. The gradual increase in consumption occurred as the leak gradually became worse and then the sudden reduction in consumption occurred when the leak was fixed. The 15 minute water consumption profile for the same property shows that the water usage throughout the day is similar to the typical water consumption profile shown in Figure 2.2, but with a consistent non-zero baseline consumption.

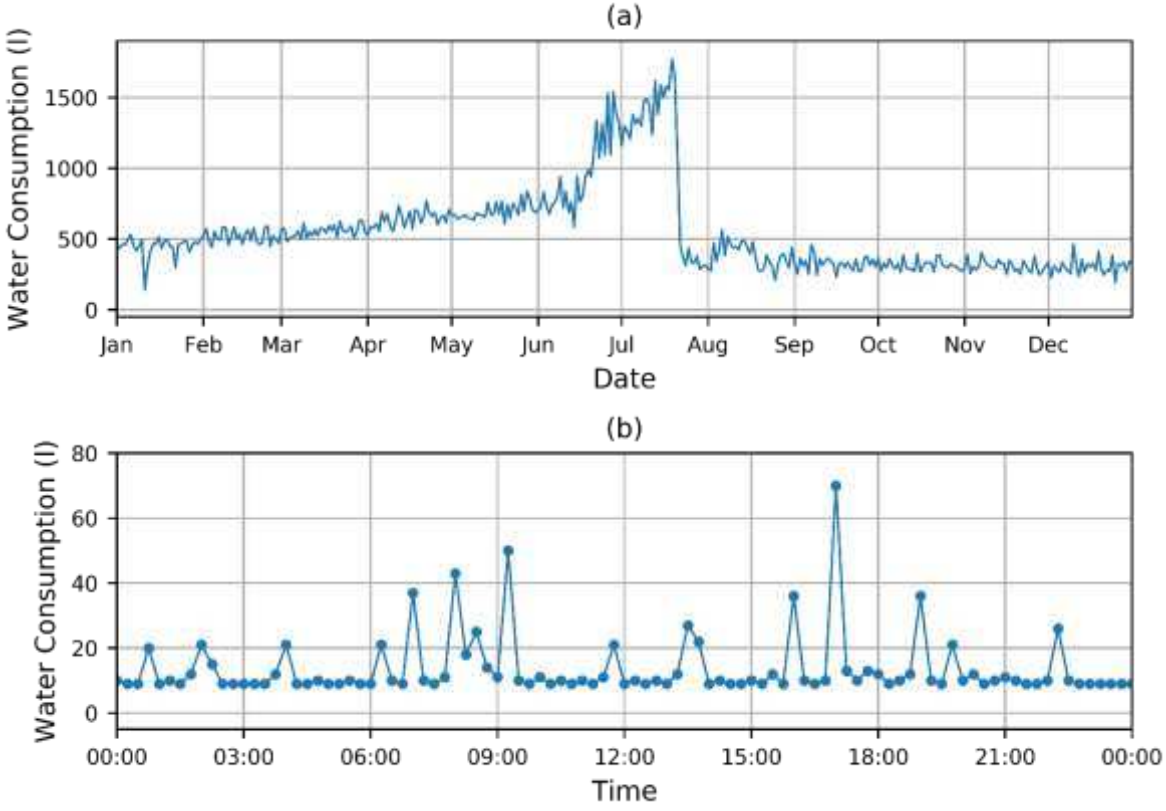


Figure 3.1 – Water consumption profiles with leakage: (a) daily water consumption over a 1 year period. (b) 15 minute water consumption on 2nd July.

Water leakage volume can vary from day to day as the hydraulic head (water pressure) within the pipework may vary depending on the water consumption in the local area. Leak

identification and removal from the data was thus conducted on a property-by-property and day-by-day basis using an automated procedure. Generally, it is expected that less water would be consumed during the night and, following discussions with representatives from SWW, it was determined that a continuous water consumption throughout the night was the most appropriate indicator of a leak. Therefore, for a particular day, a leak was identified if the minimum recorded water consumption during the night (between 00:00 and 5:45) was greater than 0 (Box 3.1). For those instances where baseline load was greater than zero, the effect of the leak was removed by subtracting the baseline load (a single value) from each water consumption reading (one per 15-min period, n = 96 for a given 24 hour period) for each day for which a leak was identified. In order to prevent negative water consumption measurements in the data after a leak is removed, the baseline load was defined as the minimum recorded water consumption for that day (Box 3.1).

IF

The minimum water consumption reading between 00:00 and 5:45 > 0

THEN

Subtract the minimum daily reading from all readings for that day

ELSE IF

The minimum water consumption reading between 00:00 and 5:45 = 0

THEN

Do nothing

Box 3.1 – Summary of procedure to identify and correct for leakage

Figure 3.2 shows the daily water consumption over a year period, as well as the 15-minute water consumption profile on 2nd July, for the same property as displayed in Figure 3.1, following the leak identification and removal procedure. Both plots show that the effect of the leak has been reduced. The baseline leakage load was reasonably constant, however, some variation was apparent, which remains evident in Figure 3.2. Small readings of 1 L occur a number of times, e.g. before 06:00 and around 15:00. This would be expected from a small leak as losses build up until triggering the meter to record a unit of usage at the meter's finest recording resolution which appears as a 1 L timestamped peak in these data. Removal of these small volume readings which remained following leakage removal would require a more complex leak removal procedure, which is deemed unnecessary for this study, since the magnitude of the water consumption events of interest (i.e. those consistent with appliance usage within an occupied property) are greater than the magnitude of the remaining minor fluctuations. Given their very low volume, there is little evidence that they have any meaningful impact on the analysis methods used in this study.

Figure 3.2(a) shows annual daily water consumption for the household with a leak in question. It shows a considerably higher peak on 21st July, the day that the leak was fixed. This is expected as water is lost whilst a leak is repaired. Figure 3.3 shows the 15 minute water consumption profile for the 21st July. It can be seen that the leak appears to have been fixed at approximately 15:00, when the baseline leakage load drops suddenly to 0 L. The leak removal procedure subtracts the minimum daily reading from all readings for a day for which

a leak has been identified. As the leak was fixed on this day, the minimum reading was 0 L. Therefore, the baseline load prior to the leak being fixed was not subtracted from any readings for that day. In order to account for this at scale using an automated procedure, a more complex leak removal procedure is required. However, as days on which leaks were fixed would only account for a small number of days over a yearlong measurement period (1 day in the case of this property), the impact on the objectives of this study, i.e. to identify tourist dwellings and to infer occupancy patterns for these properties is negligible.

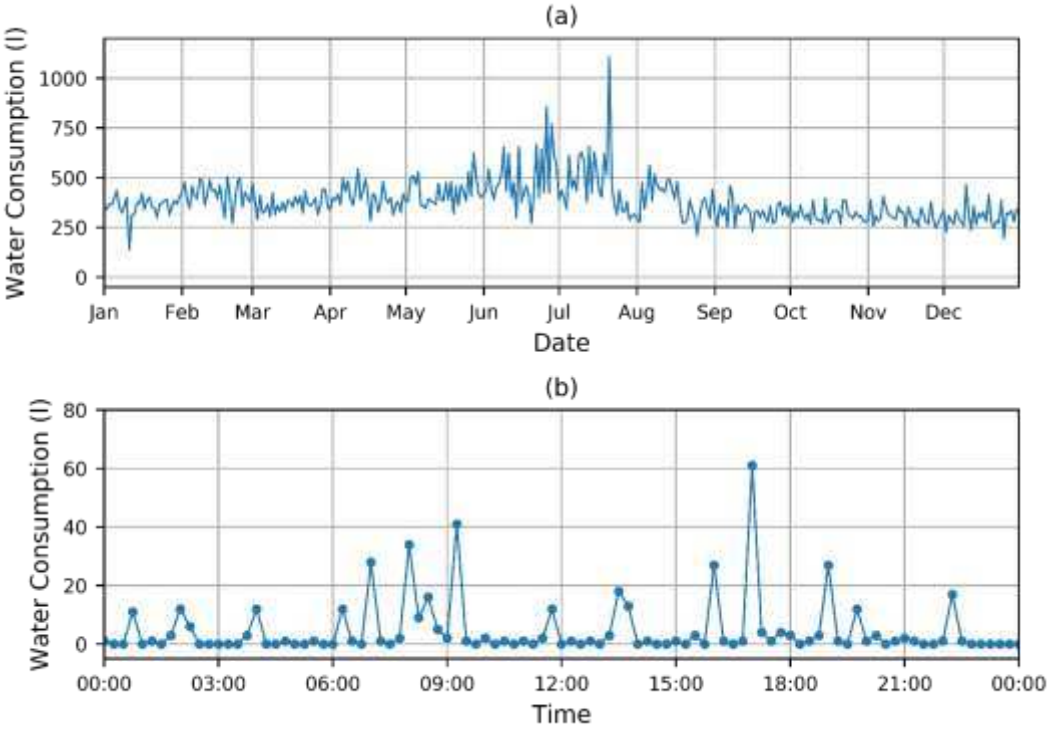


Figure 3.2 – Water consumption profiles following leak identification and removal: (a) daily water consumption over a 1 year period. (b) 15 minute water consumption on 2nd July.

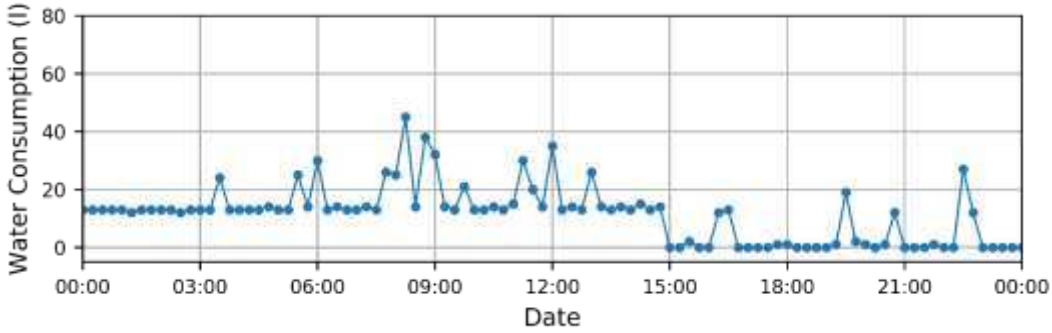


Figure 3.3 – 15 minute water consumption on 21st July, the day the leak was fixed.

Identification and correction for leakage represents an important data cleaning and preparation phase. Unlike Section 2 (where properties were removed from our dataset due to inadequate temporal coverage or completeness), no dwellings were removed from analysis since leakage detected within the sample of properties can be identified, and its effects removed, prior to further analysis. As such, a total of 61 properties (4 of which are associated with tourism) are utilised within the remainder of this report, beginning with occupancy detection as outlined in Section 4.

4 TOURIST PROPERTY OCCUPANCY DETECTION

4.1 Introduction

The first aim of this study is to develop methods to infer occupancy patterns for dwellings thought to be associated with tourism. In order to achieve this aim, a number of different methods to infer occupancy using the 15-minute water consumption data, for a one year period, were investigated. When applied to both tourist and residential properties, the occupancy detection methods investigated were also found to provide a key discriminator between residential and tourist properties, as outlined fully in Section 5.

4.2 Validation Data

In order to assess the accuracy of each occupancy detection method, it is important to validate results against ground truth data on occupancy patterns. However, no ground truth data was available for the occupancy patterns of the four tourist properties over the measurement periods provided. Therefore, in order to assess the accuracy of the occupancy detection methods investigated, for the four tourist properties, each day was labelled as either occupied or unoccupied, based on the daily water consumption profiles. This method of manual labelling would not be viable for a dataset with a large number of tourist properties, hence the requirement for an automated occupancy detection method. However, for the four tourist properties available for the current study, this method of manual labelling was achievable and provides a dataset of inferred occupancy which can be used to assess the accuracy of a range of automated occupancy detection approaches.

Manual labelling of occupancy may not provide a completely accurate validation dataset, as labelling was based on daily water consumption profiles and the true occupancy status of each day was not known. However, for the vast majority of the days investigated, inspection of the daily water consumption profile provided a clear indicator of the occupancy status. For this reason, the authors are confident that this method of manual labelling provides a very reasonable approach to assess the accuracy of the automated occupancy detection methods.

Figure 4.1 shows two examples of daily water consumption profiles for days during which a property was clearly occupied. In each case, there were a large number of non-zero measurements, with a considerable volume of water consumption recorded throughout the day. Figure 4.1(a) shows a day in which water consumption occurred reasonably frequently throughout the day, indicating that the property was occupied for all or most of the day.

As well as days which are either occupied or unoccupied, there are also check-in and check-out days, on which tourists either arrived at or left a property. Figure 4.3(a) shows a typical daily water consumption for a check-in day, which is indicated by zero water consumption during the first portion of the day followed by typical afternoon/evening consumption. Figure 4.3(b) shows a typical daily water consumption profile for a check-out day, which is indicated by a typical morning water consumption profile followed by zero water consumption for the remainder of the day. Check-in and check-out days could be considered as either occupied or unoccupied, as in each case the property is occupied for part of the day and unoccupied for part of the day. For the purposes of consistency in the manual labelling process, all check-in and check-out days were labelled as occupied.

(b) shows a day with water consumption in the morning and the evening, with no water consumption in between, indicating that the occupants were likely out of the property for a period during the day, but undertook morning and evening routines (e.g. washing and use of the toilet) consistent with a property occupied overnight.

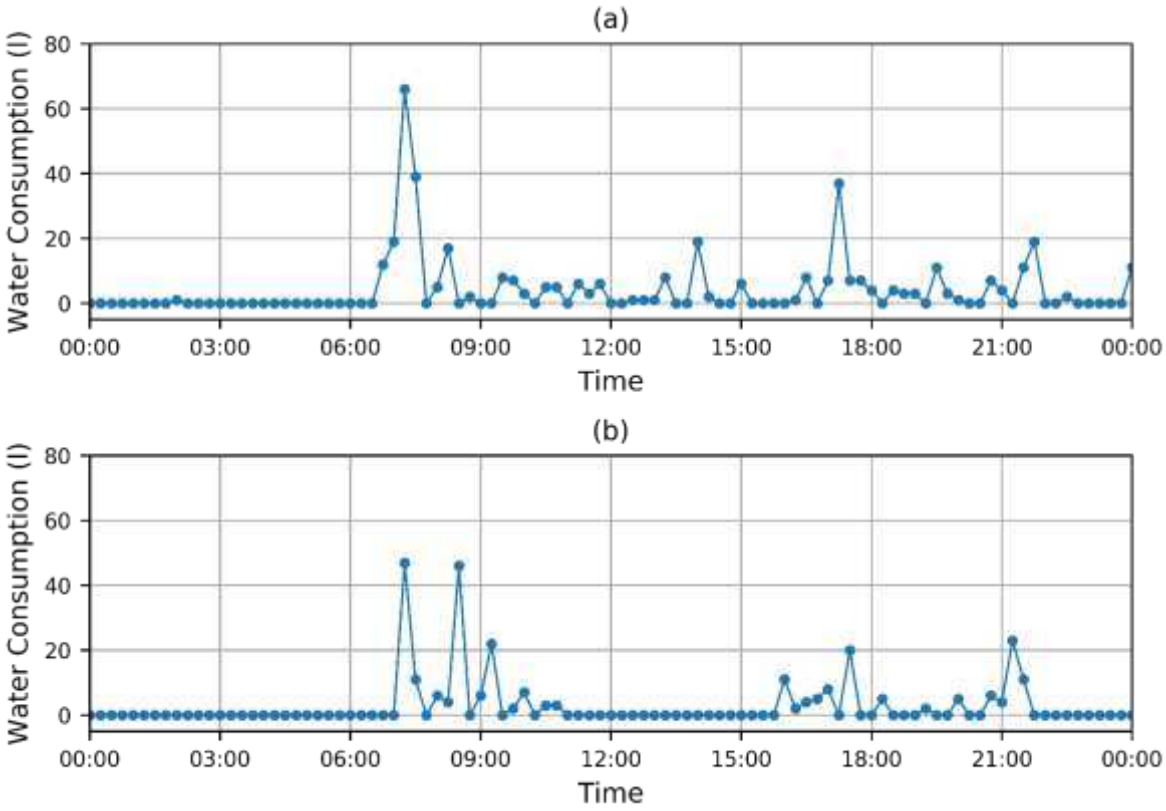


Figure 4.1 – Example daily water consumption profiles for occupied days

Figure 4.2 shows examples of typical daily water consumption profiles for unoccupied days. Figure 4.2(a) shows a day with no water consumption throughout the day, indicating that the property was unoccupied on this day. Figure 4.2(b) shows a day with only a small number of low volume water consumption measurements. These low volume measurements often occur due to low volume appliance leaks (e.g. a dripping tap) which slowly cumulatively generate a sufficient magnitude of consumption (1 L) for the recording device to register these leaks periodically (in this case approx. every six hours) and therefore, do not indicate occupancy. As no other larger volume measurements occurred during this day, it is clear that the property was unoccupied on this day.

As well as days which are either occupied or unoccupied, there are also check-in and check-out days, on which tourists either arrived at or left a property. Figure 4.3(a) shows a typical daily water consumption for a check-in day, which is indicated by zero water consumption during the first portion of the day followed by typical afternoon/evening consumption. Figure 4.3(b) shows a typical daily water consumption profile for a check-out day, which is indicated by a typical morning water consumption profile followed by zero water consumption for the remainder of the day. Check-in and check-out days could be considered as either occupied or unoccupied, as in each case the property is occupied for part of the day and unoccupied for part of the day. For the purposes of consistency in the manual labelling process, all check-in and check-out days were labelled as occupied.

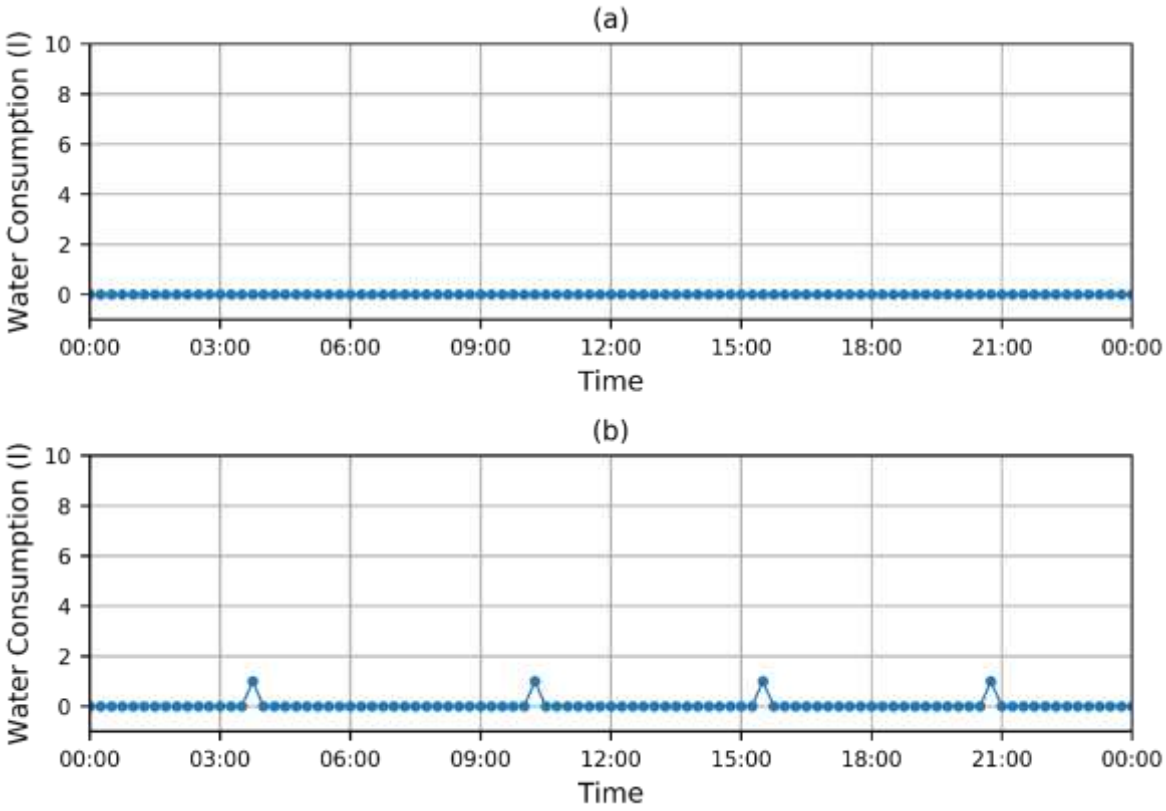


Figure 4.2 – Example daily water consumption profiles for unoccupied days

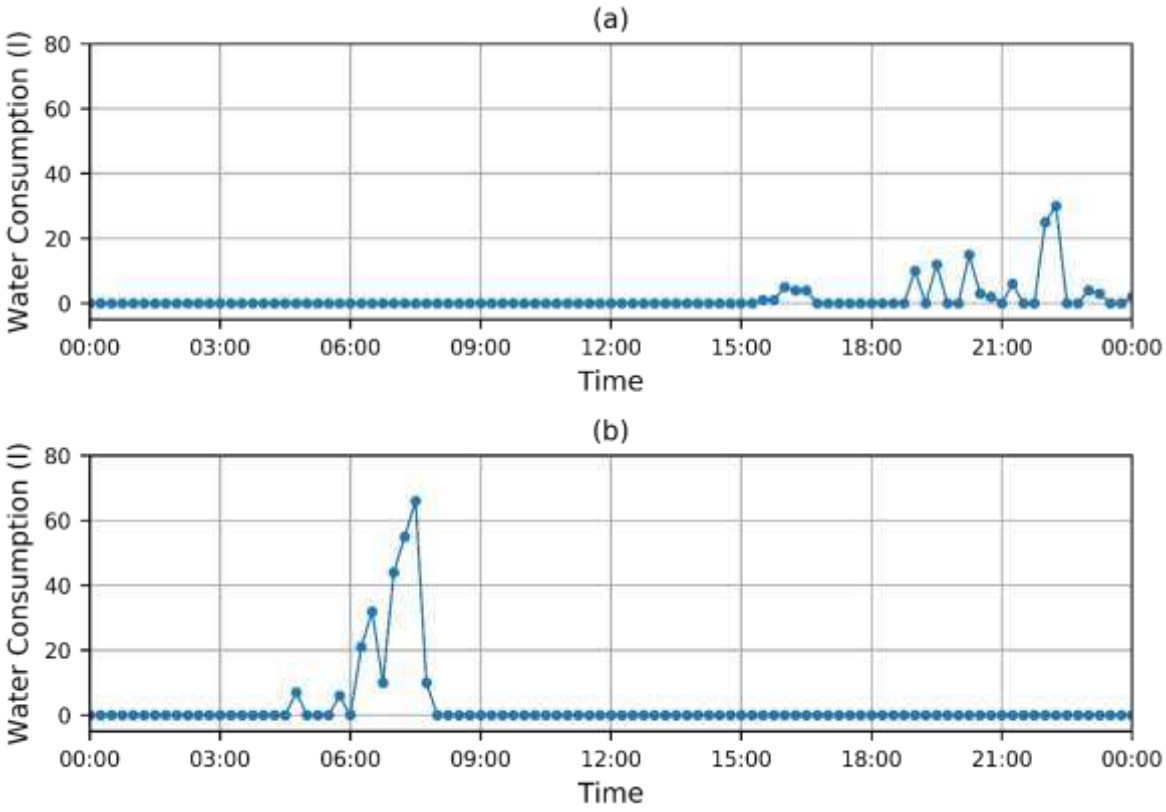


Figure 4.3 – Example daily water consumption profiles for: (a) a check-in day and (b) a check-out day

The typical daily consumption profiles shown in Figures 4.1 – Figure 4.3 represent the vast majority of water consumption profile types which were observed in the data. However, a very small number of ambiguous daily water consumption profiles were observed in the data. Figure 4.4 shows one such ambiguous profile. A small number of non-zero measurements can be observed, which are greater than the typical low volume measurements shown in Figure 4.2(b). Figure 4.4 may represent a check-out day, as the only water consumption during the day occurred in the morning. However, the water consumption in this period was lower than a typical check-out day. To determine whether an ambiguous water profile represented a check-in or check-out day, the preceding and succeeding days were investigated. For example, a check-in day is unlikely to follow an occupied day and a check-out day is unlikely to follow an unoccupied day. A possible explanation for these ambiguous profiles is that a day was unoccupied by tourists, however, a small volume of water was consumed over a short period during cleaning or maintenance. Ambiguous days such as these have been labelled as unoccupied.

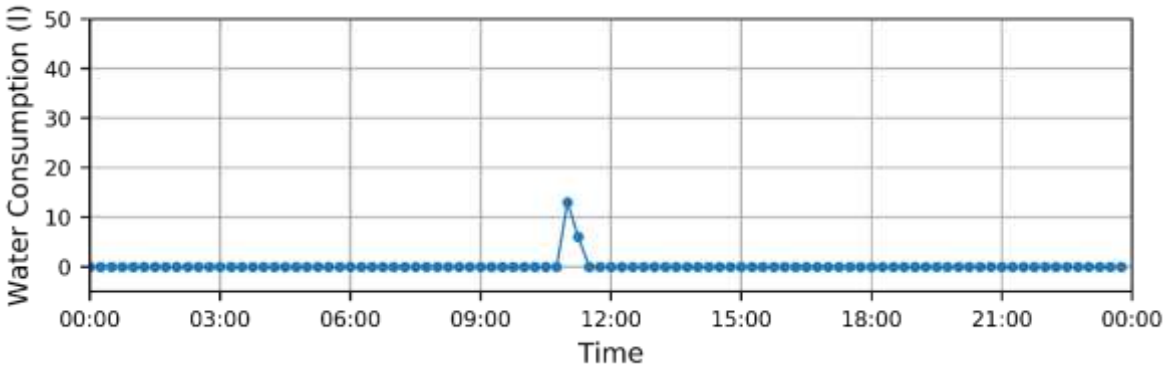


Figure 4.4 – Ambiguous daily water consumption profile

4.3 Occupancy Detection Methodology

Occupancy detection was investigated using the same concept as the Non-Intrusive Occupancy Monitoring (NIOM) method, introduced by Chen *et al.* (2013). They detected occupancy throughout the day using high-temporal resolution data from smart electricity meters. In the study by Chen *et al.* (2013), the NIOM method detects occupancy for a specified period during a day based on whether one of three metrics of power usage over a particular time period are greater than threshold values. Their measures capture the mean, the standard deviation or the range, each of which are calculated dynamically for each day, based on the night-time usage.

Our study aims to detect occupancy on a day by day basis, rather than at smaller intervals throughout the day. For an unoccupied day, the night-time water consumption would be expected to be similar to the day-time water consumption, which causes problems when defining threshold values based on the night-time water consumption. For example, Figure 4.2(b) shows zero water consumption during the night (between 00:00 and 6:00) but a small volume of water consumption during the day. Using the standard NIOM approach, this day would be defined as occupied, when it clearly was not. Eibl *et al.* (2018) used a modified version of NIOM to detect occupancy at a daily scale, in their case to predict holidays. In the modified NIOM method, only the maximum electricity consumption was considered and the threshold values were still based on the night-time usage, with a tolerance added which was intended to combat the issues encountered using night-time threshold values for unoccupied days. In our study, a range of tolerance values were investigated, however, it was found that

all values investigated produced poor accuracy (~ 50% or lower) when compared with the manually labelled validation data.

Therefore, in this work, alternative methods to define threshold values were investigated, based on the water consumption over the entire one year dataset for each property. The metrics used to define occupancy, e.g. the mean or the standard deviation, were also investigated in further detail.

The first thresholds which were investigated were based on which percentile of all days within the measurement period the mean, standard deviation or range of values for each day fell within. Table 4.1 shows the accuracy of this method, when compared with the manually labelled validation data, for a range of percentile values. The maximum average accuracy (across all four tourist properties) was 90.6% at the 50th and 60th percentiles. This means that 90.6% of our manually labelled occupancy 'events' (n = 1460[4]) are correctly labelled as 'occupied' or 'unoccupied' by this algorithm if the occupancy threshold is set as the 50th or 60th percentile of the mean, standard deviation, or range. Thus, a given day is identified as occupied if the mean, standard deviation or range of recorded water consumption that day (24-hour period) exceeds the 50th or 60th percentile of the mean, standard deviation and range for that property across the entire recording period (365-days).

However, as can be seen in Table 4.1, at these percentiles, the accuracy for each of the four properties varied considerably, indicating that threshold values based on percentiles may not be a consistently accurate method to define occupancy.

Table 4.1 – Accuracy of occupancy detection determined by whether the mean, standard deviation or range of water consumption for each day fall within various percentiles of all days

Threshold Percentile of the mean	**Accuracy %**				
	Property 1	**Property 2**	**Property 3**	**Property 4**	**Mean**
1	92.3	78.6	77.3	78.6	81.7
10	92.3	78.6	77.3	78.6	81.7
20	98.1	78.6	77.3	78.6	83.2
25	97.3	78.6	77.3	78.6	82.9
30	94.0	78.6	77.3	78.6	82.1
40	87.7	85.5	84.4	91.2	87.2
50	80.8	96.2	94.2	91.2	90.6
60	73.2	93.4	97.5	98.4	90.6

The next threshold values to be investigated were defined as a percentage of the mean value of each metric over all days within the yearlong dataset for each property. Table 4.2 shows the accuracy of this method for various percentages of the mean values. The maximum mean accuracy was found to be 98.2% with a threshold value of 40% of the average. Unlike for the percentile threshold values, the accuracy was also found to be reasonably consistent for all properties.

[4] Four tourist properties, each of which has 365 days (one year) of individually labelled occupancy 'events' (either occupied or unoccupied).

Table 4.2 – Accuracy of occupancy detection determined by whether the mean, standard deviation or range of water consumption for each day fall above various percentages of the mean of all days

Threshold % of Average	Accuracy %				
	Property 1	Property 2	Property 3	Property 4	Mean
1	92.6	78.6	77.3	78.6	81.8
10	99.2	88.8	95.1	94.8	94.5
20	99.5	96.7	95.3	94.8	96.6
25	99.2	97.0	96.4	94.8	96.8
30	98.9	97.8	97.3	96.4	97.6
40	98.1	98.1	98.9	97.5	98.2
50	97.8	97.3	98.6	98.1	97.9
60	95.6	97.3	97.8	98.1	97.2

An accuracy of 98.2% is extremely high and suggests that this occupancy detection method is suitable to infer occupancy patterns for individual dwellings. However, additional methods to define occupancy were also investigated, in an attempt to further improve accuracy. The NIOM method defines occupancy if any one of the three consumption metrics are greater than their given threshold. The next method to be investigated defined occupancy only if all three metrics were greater than their given threshold. Again, threshold values were defined as a percentage of the average values across all days within the dataset. Table 4.3 shows the accuracy of this method for a range of threshold values. The maximum mean accuracy is slightly higher than the previous method, at 98.7%, again, with a reasonably consistent accuracy for each property.

Table 4.3 – Accuracy of occupancy detection determined by whether the mean, standard deviation and range of water consumption for each day fall above various percentages of the mean of all days

Threshold % of Average	Accuracy %				
	Property 1	Property 2	Property 3	Property 4	Mean
1	98.4	84.4	88.2	78.6	87.4
10	99.7	97.5	97.0	95.9	97.5
20	98.9	98.6	98.9	98.4	98.7
25	98.1	98.4	99.7	98.4	98.6
30	97.8	98.1	98.9	98.4	98.3
40	95.9	97.3	97.3	98.9	97.3
50	92.1	96.2	96.4	98.4	95.8
60	87.9	94.5	94.8	98.1	93.8

The next occupancy detection methods to be investigated considered each of the three water consumption metrics in isolation, again with the threshold values defined as a percentage of the average values for all days within the measurement period. Table 4.4, Table 4.5 and Table 4.6 show the accuracy for various threshold values when considering only the mean, standard deviation and range, respectively. Of these methods, considering only the mean provided the maximum mean accuracy of 98.8%, with a threshold value of 25% of the average. The maximum mean accuracy for the standard deviation and the range were also extremely high, however, at 98.4% and 97.8% respectively, both for a threshold value of 40% of the average.

Table 4.4 – Accuracy of occupancy detection determined by whether the mean water consumption for each day falls above various percentages of the mean of all days

Threshold % of Average	Accuracy %				
	Property 1	Property 2	Property 3	Property 4	Mean
1	92.6	84.4	88.2	78.6	86.0
10	99.2	93.4	97.0	95.9	96.4
20	98.9	98.6	98.9	98.4	98.7
25	98.4	98.6	99.7	98.4	98.8
30	98.1	98.4	98.9	98.4	98.4
40	95.9	97.5	97.8	98.9	97.5
50	92.1	96.4	97.0	98.4	96.0
60	88.2	94.8	95.6	98.4	94.2

Table 4.5 – Accuracy of occupancy detection determined by whether the standard deviation of water consumption for each day falls above various percentages of the mean of all days

Threshold % of Average	Accuracy %				
	Property 1	Property 2	Property 3	Property 4	Mean
1	93.4	78.6	77.3	78.6	82.0
10	99.7	94.5	95.3	94.8	96.1
20	99.2	97.3	95.9	95.6	97.0
25	98.6	98.4	96.4	96.2	97.4
30	98.6	98.4	98.1	97.3	98.1
40	98.1	98.4	99.2	98.1	98.4
50	97.5	97.0	98.4	98.4	97.8
60	94.2	96.7	96.7	98.1	96.4

Table 4.6 – Accuracy of occupancy detection determined by whether the range of water consumption for each day falls above various percentages of the mean of all days

Threshold % of Average	Accuracy %				
	Property 1	Property 2	Property 3	Property 4	Mean
1	98.1	78.6	77.3	78.6	83.2
10	99.7	92.9	95.1	94.8	95.6
20	99.5	96.7	95.6	94.8	96.6
25	98.9	96.7	96.4	94.8	96.7
30	98.4	97.5	97.3	96.4	97.4
40	98.1	97.3	98.4	97.5	97.8
50	97.8	97.0	97.8	98.1	97.7
60	94.8	96.4	96.4	97.8	96.4

All of the occupancy detection methods discussed thus far have considered the three metrics utilised in the original NIOM method: mean, standard deviation and range. The occupancy detection methods described below also considered an additional metric which captured the number of water consumption events per day. The data utilised in this study consists of the total water consumed every 15 minutes, so individual water consumption events (such as a toilet flushing) cannot be identified. Therefore, the term “usage event” is used to describe any 15-minute period during which water was consumed.

The simplest method to define a usage event would be any non-zero reading. However, as described in Section 3, following leak detection and elimination some very low volume water consumption readings remain, which represent variation in the original leak volume rather than usage events. Therefore, for a particular day, a usage event has been defined as any water consumption reading greater than the mode reading for that day. For cases in which the most common water consumption reading is non-zero following leak detection, these non-zero values will not be considered usage events. However, for cases in which the mode reading is 0 L, but small non-zero readings remain following leak detection which do not represent usage events, these small non-zero readings would be considered usage events using this method. This, therefore, is not a perfect method to define a usage event. As described previously, removal of these small non-zero readings following leak detection and elimination would require a more complex leak removal procedure, which was beyond the scope of this work due to time constraints. Nevertheless, the usage event metric was able to produce accurate results, as described below.

Initially, the usage event metric was considered in isolation, again with the threshold value defined as a percentage of the average value. Table 4.7 shows the accuracy of this method with a range of threshold values. The maximum accuracy was achieved with a threshold value of 60% and was found to be 97.1%. This is a very high level of accuracy, however, is lower than any of the three previously investigated metrics when considered in isolation.

Table 4.7 – Accuracy of occupancy detection determined by whether the number of usage events for each day falls above various percentages of the mean of all days

Threshold % of Average	Accuracy %				
	Property 1	Property 2	Property 3	Property 4	Mean
1	97.3	78.6	77.3	78.6	82.9
10	98.6	84.1	88.5	93.7	91.2
20	98.4	88.5	91.8	96.4	93.8
25	98.4	91.5	95.1	96.4	95.3
30	98.1	91.5	96.7	98.4	96.2
40	98.1	92.1	98.1	99.5	96.9
50	96.7	93.2	98.1	99.5	96.8
60	97.0	93.4	98.6	99.2	97.1

The final occupancy detection methods which were investigated combined the number of usage events with the mean, as the mean had provided the highest accuracy when considered in isolation. Two methods to combine the mean and the number of usage events were investigated:

i) a day was defined as occupied when either the mean or the number of usage events were greater than their specified thresholds, and;
ii) a day was defined as occupied when both the mean and the number of usage events were greater than their specified thresholds.

Again, the threshold values were based on the average of the values across all days within the dataset for each property. Table 4.8 and 4.9 show the accuracy of these two methods for various threshold values. Of the two methods, the highest accuracy was found to be 98.8 %, at a threshold value of 25% of the average, when occupancy was defined when both metrics were greater than their specified threshold values. This was the equal highest accuracy found for any occupancy detection method investigated, along with the mean considered in isolation.

We have, therefore, identified two occupancy detection methods which give the equal highest accuracy when compared with the manually labelled validation data:

Method A: when the mean on a given day of interest is greater than 25% of the average mean for all days

Method B: when both the mean and the number of usage events are greater than 25% of the average for each criteria.

Method A utilises only one metric so is computationally less expensive. Therefore, this method may be preferable, especially for larger sample sizes. However, in order to determine which method is preferable, the differences between the two methods were investigated further.

Table 4.8 – Method A: accuracy of occupancy detection determined by whether the mean water consumption or the number of usage events for each day fall above various percentages of the mean of all days

Threshold % of Average	Accuracy %				
	Property 1	Property 2	Property 3	Property 4	Mean
1	92.3	78.6	77.3	78.6	81.7
10	98.6	84.1	88.5	93.7	91.2
20	98.6	88.5	91.8	96.2	93.8
25	98.6	91.5	95.1	96.2	95.3
30	98.1	91.5	96.7	97.8	96.0
40	98.1	92.1	98.1	98.6	96.7
50	96.7	93.4	98.4	98.6	96.8
60	97.0	94.2	98.9	98.9	97.3

Table 4.9 – Method B: accuracy of occupancy detection determined by whether the mean water consumption and the number of usage events for each day fall above various percentages of the mean of all days

Threshold % of Average	Accuracy %				
	Property 1	Property 2	Property 3	Property 4	Mean
1	97.5	84.4	88.2	78.6	87.2
10	99.2	93.4	97.0	95.9	96.4
20	98.6	98.6	98.9	98.6	98.7
25	98.1	98.6	99.7	98.6	98.8
30	98.1	98.4	98.9	98.9	98.6
40	95.9	97.5	97.8	99.7	97.7
50	92.1	96.2	96.7	99.2	96.0
60	88.2	94.0	95.3	98.6	94.0

As both Methods A and B require the mean consumption to be greater than 25% of the average, whereas Method B requires the number of usage events to be greater than 25% of the average as well, the only differences between the two methods will occur when Method A defines a day as occupied but Method B does not. The opposite case does not exist, as all days which are defined as occupied by Method B would also be defined as occupied by Method A. Across our four tourist properties, there were 93 days (6.4% of all days), which were defined as occupied by Method A but not by Method B.

Figure 4. shows examples of days which Method A defined as occupied but Method B defined as unoccupied. *Figure 4.*(a) shows a day for which there were only two non-zero readings,

one of which was reasonably large. The large reading caused Method A to define the day as occupied, whereas, the low number of usage events caused Method B to define the day as unoccupied. It is not possible to confirm whether this day was occupied or not, however, the daily profile does not show a trend which would usually be associated with occupancy. Therefore it is likely that this day was unoccupied. *Figure 4.*(b) shows a day for which all readings are either zero or one litre, with more readings of one litre than zero litres. This day clearly represents an unoccupied day and the variation in water consumption may be a result of a relatively low volume leak, as described in Section 3. The large number of 1 litre readings caused the mean consumption for the day to be greater than 25% of the average for all days, however, as 1 litre was the mode reading, no usage events were recorded. Similar cases also occur, for which the mode reading was zero. These cases would result in a large number of usage events, however, as there would be fewer non-zero readings, the likelihood that the mean would fall above 25% of the average would be reduced.

Figure 4.(a) and Figure 4.(b) show days which are likely unoccupied and would be defined as occupied by Method A but unoccupied by Method B. Figure 4.(c) shows a check-out day. In this case, the mean water consumption was sufficiently high to be defined as occupied by Method A, but the number of usage events was not high enough to be defined as occupied by Method B. The manual occupancy labelling procedure defined check-in and check-out days as occupied, however, as previously described, this is a subjective determination, as these days could be described as either occupied or unoccupied. Due to check-in and check-out days being occupied for only a proportion of the time, they are more likely to fall close to the limits of the thresholds used to define occupancy than clearly occupied or unoccupied days. For this reason, it is unsurprising that a check-in or check-out day may be defined differently by Methods A and B.

Overall, the majority of the 93 days which were defined differently by methods A and B had consumption profiles which resembled those displayed in Figure 4.5, with a smaller proportion of ambiguous days for which it was not clear whether the day was occupied or unoccupied. As Figures 4.5(a) and (b) both show days for which Method B would produce a ‘correct’ result, and there is subjectivity in how to define the occupancy status for check-in or check-out days, such as in Figure 4.5(c), Method B was chosen as the most appropriate occupancy detection method. This occupancy detection method is defined as in Box 4.1.

A day is classified as occupied:

IF

The number of usage events for a given day is greater than 25% of the mean number of usage events of all days for that property.

AND

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

Box 4.1 – Summary of final occupancy detection method

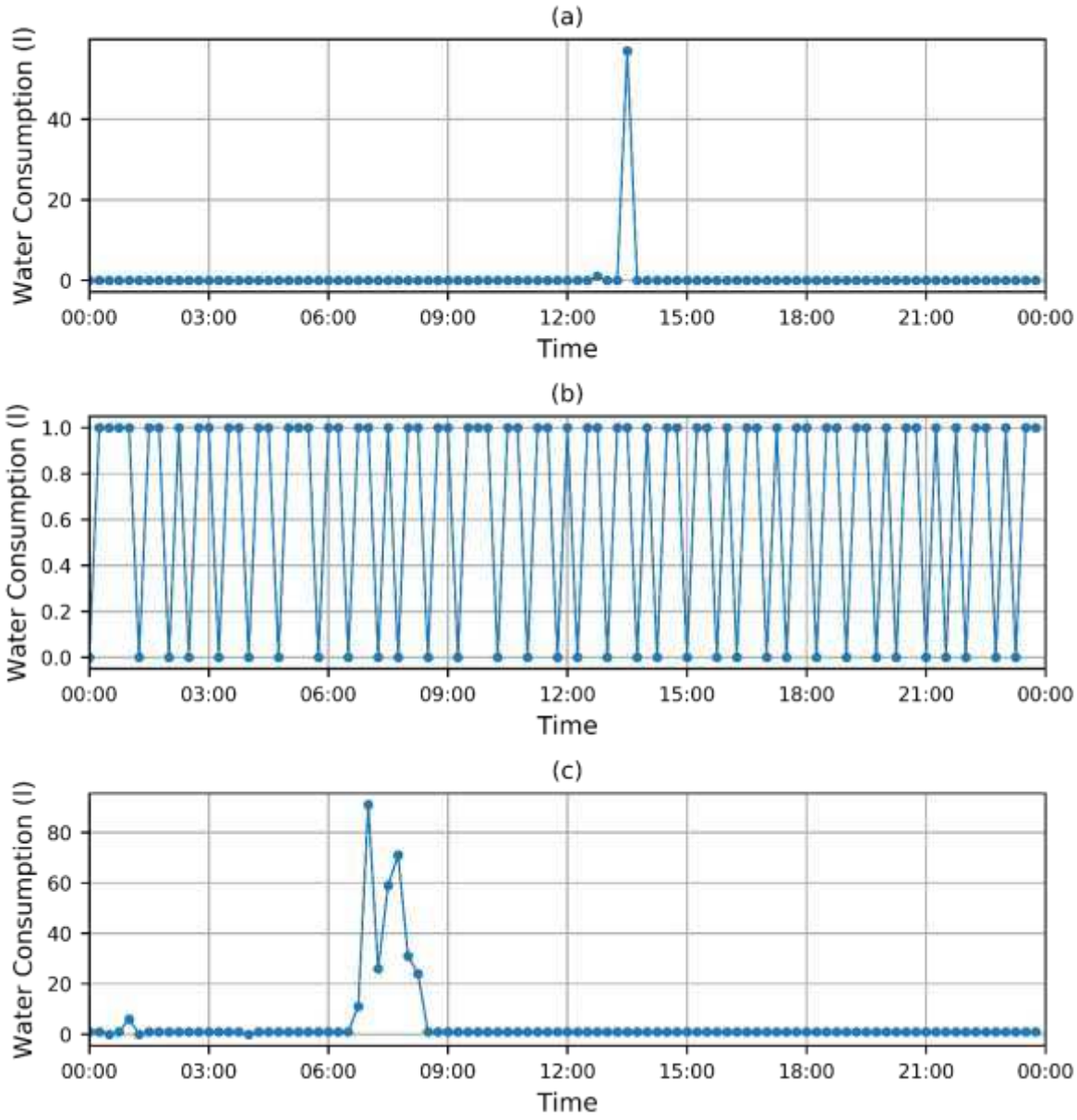


Figure 4.5 – Example differences in occupancy detection between Method A and Method B.

Figure 4.6 shows the daily total water consumption over the yearlong measurement period for each of the tourist properties, with the occupancy status of each day, as determined by the chosen occupancy detection method, indicated by blue bars. Figure 4.6(a) – (d) show the data for Properties 1 – 4, respectively. For each property, the occupancy detection method has determined the property to be unoccupied when there is little or no water consumption and occupied when there is noticeable water consumption. Properties 2 – 4 show similar trends, in that there are a reasonably large number of transitions from occupied to unoccupied days and vice versa. Property 1, however, shows a slightly different trend as there are fewer transitions from occupied to unoccupied days and the periods of occupancy are generally longer than those for properties 2 – 4. This difference in occupancy trends between Property 1 and Properties 2 – 4 will also be apparent in subsequent analysis.

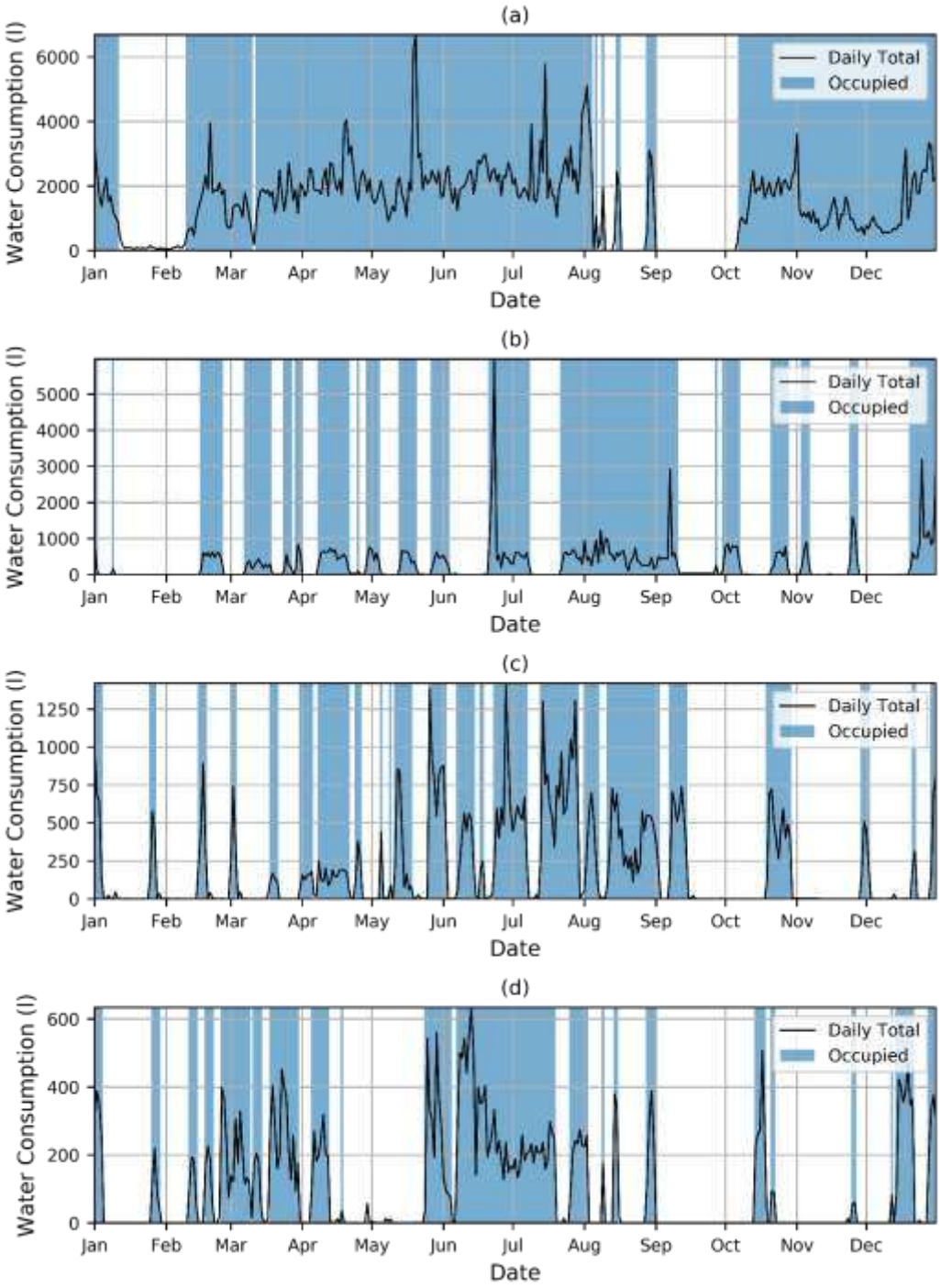


Figure 4.6 – Daily total water consumption for each tourist property with the occupancy status, as determined by the occupancy detection procedure (15-min resolution), indicated.

We are confident that our occupancy detection method is able to infer, on a day-by-day basis, periods of occupancy and vacant periods within our four tourist study properties. Using our manually labelled validation data, this approach was able to correctly identify the property occupancy status 98.8% of the time. We re-ran the occupancy detection algorithm utilising the dwelling-level water consumption data aggregated to a 1-hour resolution. This resulted in a negligible decrease in the accuracy of this approach, with an overall accuracy of 98.7% against the same labelled validation data. Across these four properties, this equates to just two additional days that have been incorrectly assigned an occupancy status, relative to the 15-

min resolution data. This strongly suggests that this approach could be applied successfully across data of varying temporal resolutions and that the trade-off between data volume and occupancy detection is negligible in this context.

The occupancy detection method developed here can identify transition points between occupancy statuses and could be applied to other property types (e.g. second homes, student households or other dwellings) to infer changes in occupancy status and periods of utilisation. Given that these data enable us to infer occupancy status for tourist dwellings, we now seek to develop approaches to identify dwelling with characteristics associated with tourist utilisation (seasonal occupancy and frequent transitions between occupancy statuses), as explored in Section 5.

5 IDENTIFICATION OF DWELLING TYPE

5.1 Introduction

The second aim of this study was to assess potential approaches to distinguish tourist dwellings from residential dwellings based on their recorded 15-min water consumption profiles. Whilst we have a good indication as to which of our study dwellings represent tourist accommodation (see Section 3), the intention is to develop methods that can categorise properties according to dwelling type (at this stage tourist vs. residential) with no additional information. Therefore, it was decided that an unsupervised machine learning approach should be used, as unsupervised approaches do not require validation data to train the algorithm, so these methods would be applicable for samples of properties without validation data. Since we seek to categorise properties according to their status, we utilise a clustering approach, seeking to group dwellings (n = 61) into two clusters, 'tourism' and 'residential'. It would be entirely feasible to adapt this approach to capture a broader range of dwelling types (e.g. residential property permanently occupied, second/holiday home, self-catering holiday let).

In this study, the k-means, Gaussian Mixture Model and DBSCAN clustering approaches were all investigated. k-means clustering was found to be the simplest approach to implement and provided the most accurate results, relative to our known property status. Therefore, only the k-means clustering results are described in this report. k-means clustering is one of the simplest, easy to implement and most well-known unsupervised machine learning algorithms. The algorithm groups objects into clusters by minimising the distances between points within a cluster and the number of clusters, k, must be manually input.

In order to apply k-means clustering to the sample of 61 properties, a series of features of water consumption behaviour for each property were extracted from the data. This section of this report describes the methods used to extract these features, as well as the results of the k-means clustering analysis. It should be noted that, although an unsupervised machine learning approach has been investigated in this study, which does not require validation data on dwelling type, validation data were available, which was used to determine the most appropriate features for clustering. The trends identified using this validation data may vary somewhat for different properties in the South West of England and also for different geographical regions. This should be considered when applying the methods described to different samples of properties.

5.2 Feature Investigation

5.2.1 Occupancy Ratio

It would be expected that a tourist property would generally be occupied for a considerably shorter proportion of the time than a residential property. Therefore, the ratio of occupied days

to total number of days (in this case one year – 365 days – per property) was calculated as a feature to discriminate between tourist and residential properties. Figure 5.1 shows the occupancy ratio for the tourist and residential properties, with the tourist properties numbered 1 – 4 and displayed in blue. Properties 2 – 4 have a noticeably lower occupancy ratio than all of the residential properties, as might be expected. Property 1 has a lower occupancy ratio than all but one of the residential properties, but a considerably higher occupancy ratio than the other tourist properties. This is another indication of the difference in occupancy trends between Property 1 and the other tourist properties, which was discussed in Section 4.3. Although Property 1 has a higher occupancy ratio than one residential property (Property 17), when combined with other features the occupancy ratio may still provide a useful discriminator between residential and tourist properties.

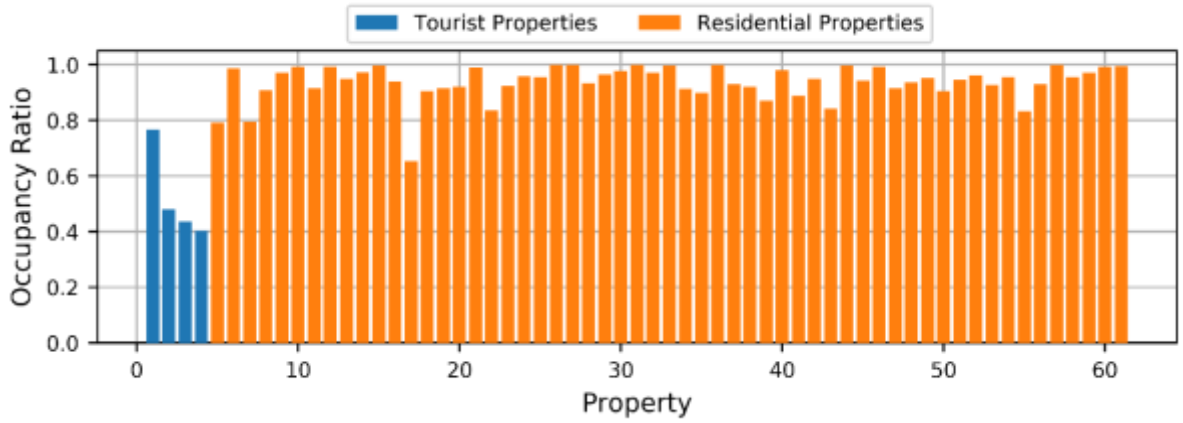


Figure 5.1 – Occupancy ratio for tourist and residential properties.

5.2.2 Statistical Methods

As tourist properties are not generally occupied all of the time, it may be expected that the average water consumption of these properties would be lower than that of a comparable residential property. It may also be expected that the variation in the daily water consumption would be higher for a tourist property, as different size parties of guests, with different routines and habits, may occupy the property at different times. A large number of unoccupied days may also increase the daily variation in water consumption as there would be a larger number of days on which little or no water was consumed.

The average water consumption was calculated using both the mean and the median, with the average water consumption calculated for the daily total consumption over the yearlong measurement period. The variation in water consumption was quantified using standard deviation, again calculated using the daily total over the measurement period. Not all properties are comparable, however. For example, a large ten-bedroom holiday let occupied for only four months of the year may have a higher average water consumption than a small residential dwelling with a single occupant. Similarly, a proportionally small variation in water consumption for a large residential dwelling with many occupants may equate to a large magnitude variation in water consumption when compared with a small, one bedroom tourist property. For these reasons, the daily water consumption totals were normalised, so that water consumption data with diverse magnitude ranges could be compared. Min-max normalisation was utilised. For each property, the minimum daily total was assigned the value 0, the maximum daily total was assigned the value 1 and all intermediate daily totals were transformed into the corresponding value between 0 and 1.

For certain properties, a small number of days can be observed during which an extremely large volume of water was consumed, when compared with the usual daily water consumption

for each property. Figure 5.2(a) shows the daily total water consumption for one such example in a residential property. Five days with exceptionally high water consumption volumes are evident between May and September. Figure 5.2(b) shows the 15-minute water consumption readings for one of these days. For a period of approximately 4 hours, the water consumption on this day was unusually high. It is unclear what the causes of these high water consumption readings are. Possible explanations include the filling of outdoor appliances such as hot tubs and paddling pools, domestic cleaning such as the use of a pressure washer or taps being accidentally left on for extended periods. Extremely high daily totals such as these may skew the mean, median and standard deviation, possibly affecting their suitability as discriminators between tourist and residential dwellings. In order to reduce the effect of extremely high daily water consumption readings on the mean, median and standard deviation, smoothing of the data was conducted as outlined in Box 5.1.

IF

The daily total water consumption > the mean daily total water consumption + 2 x the standard deviation

THEN

Replace the 15-minute water consumption profile for that day with the mean 15-minute water consumption profile for all days.

Box 5.1 – Overview of approach to smooth abnormally high consumption at a dwelling level

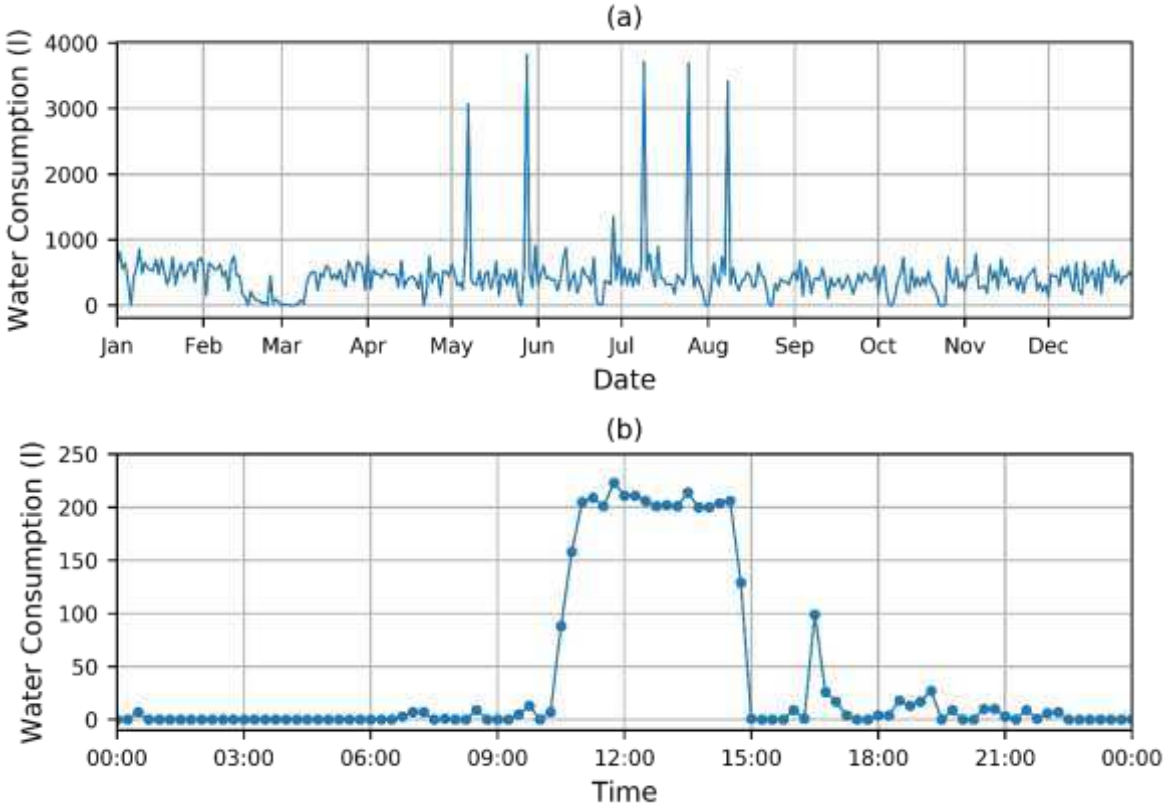
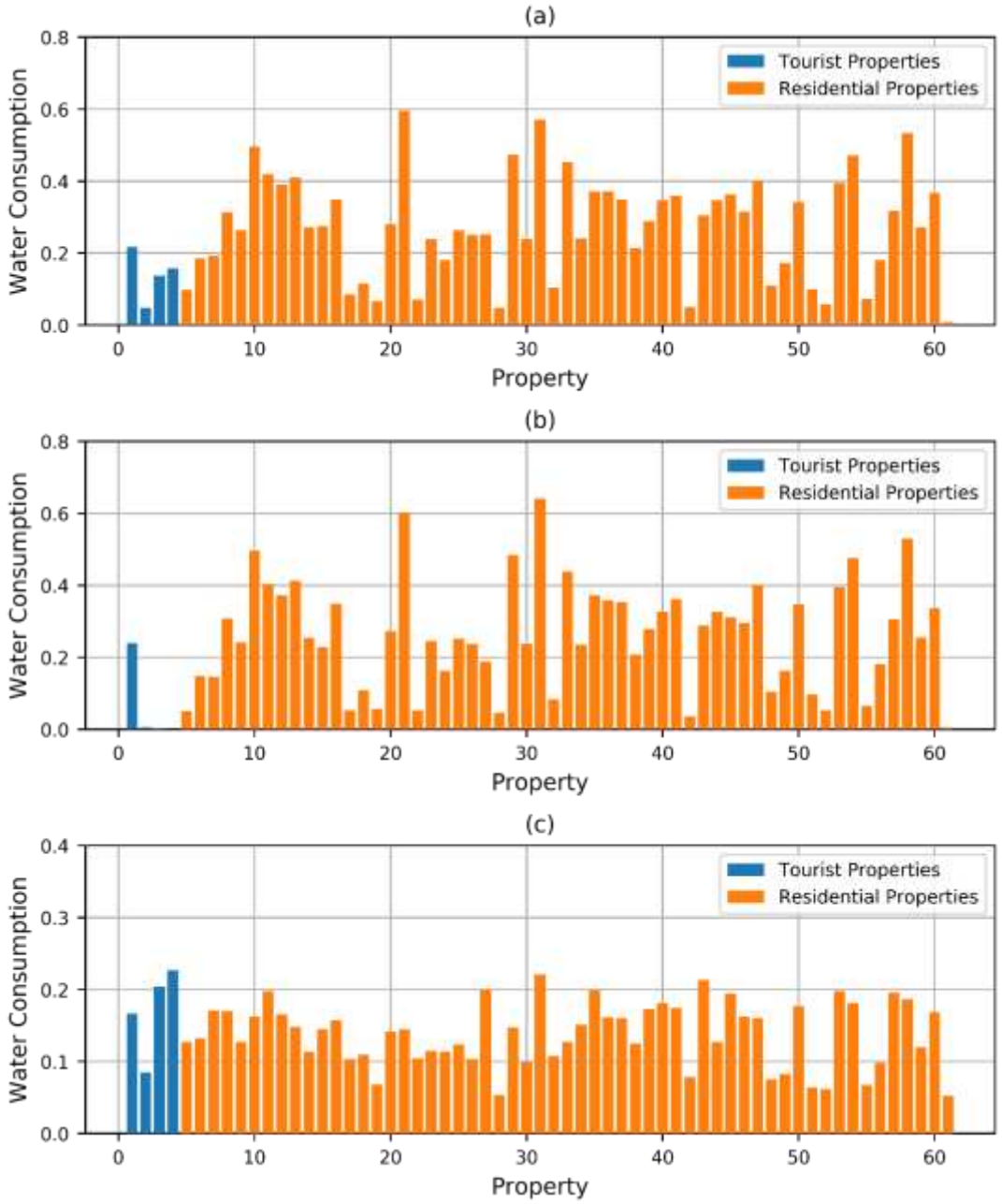


Figure 5.2 – Examples of extremely high water consumption readings: (a) daily totals over the yearlong measurement period. (b) 15 minute readings for an example high consumption day.

Figure 5.3 and Figure 5.4 show the mean, median and standard deviation of the normalised daily total water consumption for each property, calculated with both the unsmoothed and smoothed data respectively. It can be seen that the smoothed data provides a greater distinction between the tourist and residential properties. Therefore, only the smoothed data



was considered for further analysis conducted using these statistical methods. Note that, for all other analysis in this study, the unsmoothed data was utilised. The occupancy ratio was defined using the occupancy detection method, described previously, for which data smoothing would not be appropriate as this would introduce artificial data into the dataset which may produce inaccurate determinations of occupancy status. For other analysis methods, both the smoothed and unsmoothed data was investigated and it was found that both datasets produced similar results. Therefore, it was decided that the unsmoothed data should be utilised, so as not to introduce the artificial data contained within the smoothed data into these analysis methods.

Figure 5.3 – Statistical methods calculated for the normalised, unsmoothed daily totals for tourist and residential properties: (a) mean. (b) median. (c) standard deviation.

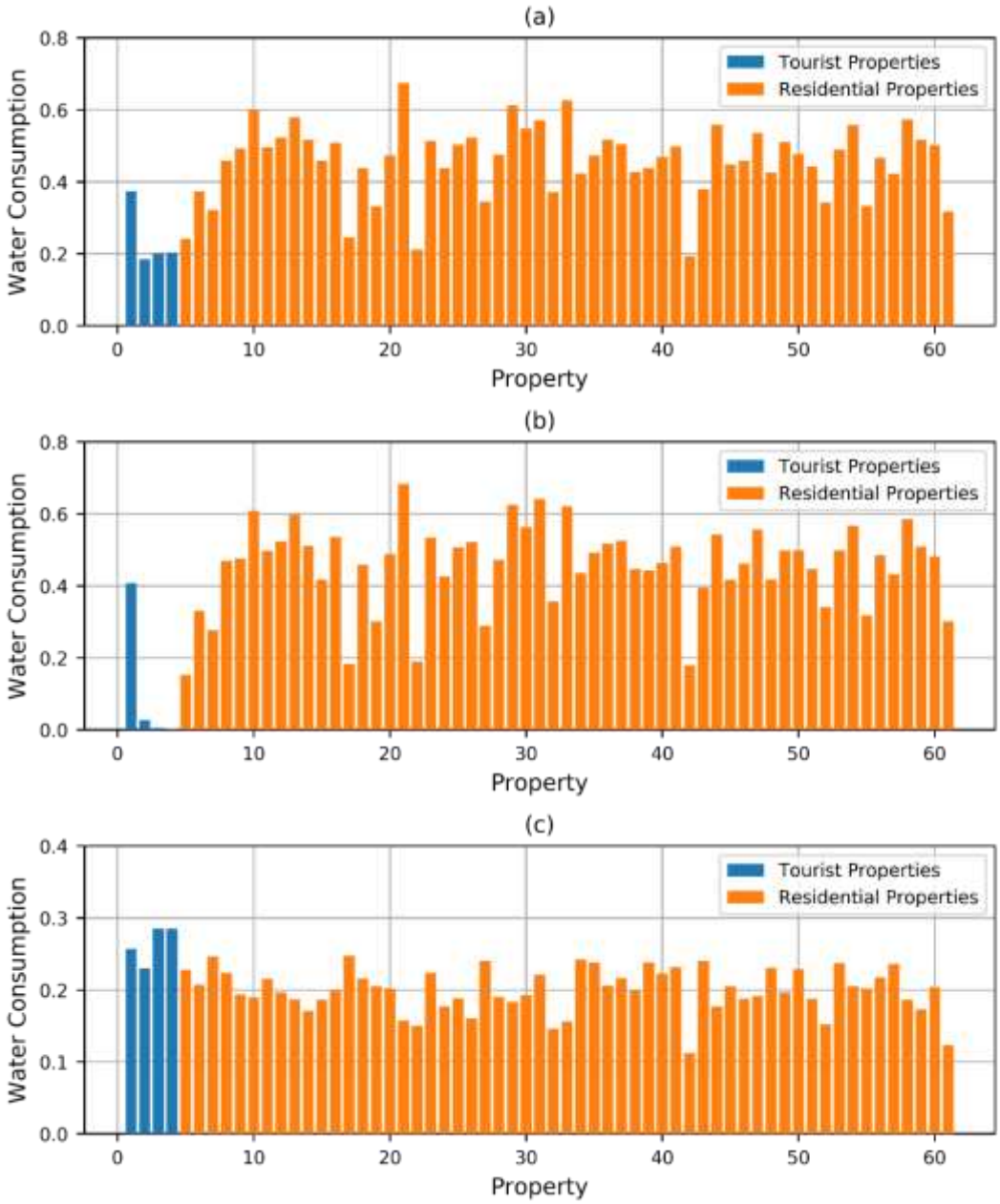


Figure 5.4 – Statistical methods calculated for the normalised, smoothed daily totals for tourist and residential properties: (a) mean. (b) median. (c) standard deviation.

Figure 5.3(a) shows that, for three of the four tourist properties, the mean consumption is lower than that of the majority of the residential properties. The remaining tourist property, however, has a higher mean than the other tourist properties. This is Property 1, which has a dissimilar occupancy trend to the other tourist properties, as described in Section 4.3. The median, shown in Figure 5.3(b), shows the same trend. However, the median values for Properties 2 – 4 are extremely low when compared with the other properties. Figure 5.3(c) shows that Properties 1, 3 and 4 provide the three highest values of standard deviation. Property 2 provides a standard deviation value lower than that of a small number of residential properties, but still generally high. Despite the fact that Property 1 produced significantly higher mean and

median values than the other tourist properties, these statistical methods do provide a distinction between the tourist and residential properties and, therefore, the smoothed, normalised mean, median and standard deviation were defined as features to be used in clustering analysis.

5.2.3 Seasonality

The peak tourist season in the South West of England, as in much of the UK, is the summer. As such, it would be expected that tourist dwelling occupancy, and therefore water consumption in these properties, would be greater during the summer months and may provide a useful factor to distinguish between tourist and residential dwellings. Therefore, an indicator of the variation between summer and winter consumption could afford discriminative power between properties in the clustering process. This indicator first calculates the monthly mean of the daily total consumption on a property-by-property basis. Seasonality was defined as the ratio of the mean monthly consumption of the summer months to the mean monthly consumption of the winter months.

Figure 5.5(a) shows a bar chart of the seasonality of each property, with summer defined as April – September and winter as October – March, so that all months of the year were included in the seasonality calculation. The ratio of summer to winter consumption was generally higher for the tourist properties, but not in all cases and the summer/winter ratio for three of the four tourist properties was reasonably similar to that of many of the residential properties. However, defining summer and winter in this way also includes the spring and autumn periods, which may not provide as clear a discriminator between tourist and residential properties. Figure 5.5(b) shows the summer/winter ratio for each property with summer defined as May – August and winter as November – February. In this case, there is a more clear distinction between the summer/winter ratios for the tourist and residential properties, with only a small number of residential properties having similar ratios to the tourist properties. Therefore, the summer/winter ratio, with summer and winter defined as in

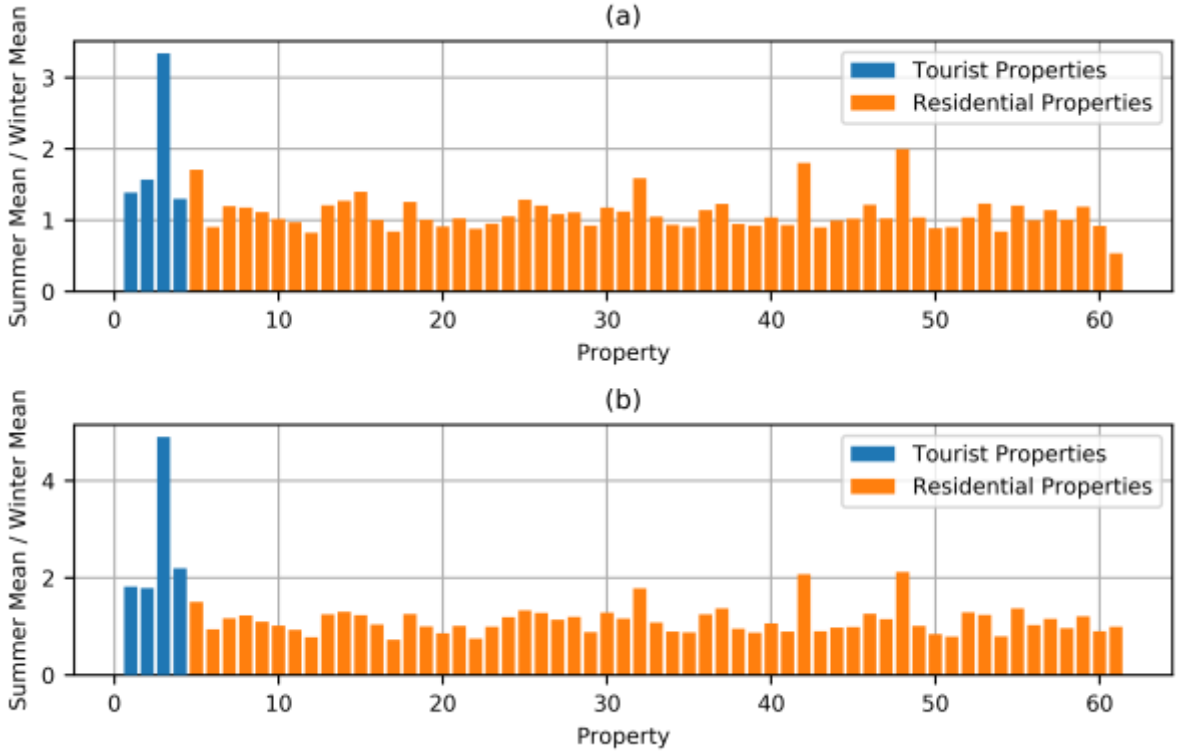


Figure 5.5(b), was selected as a feature for the clustering algorithms.

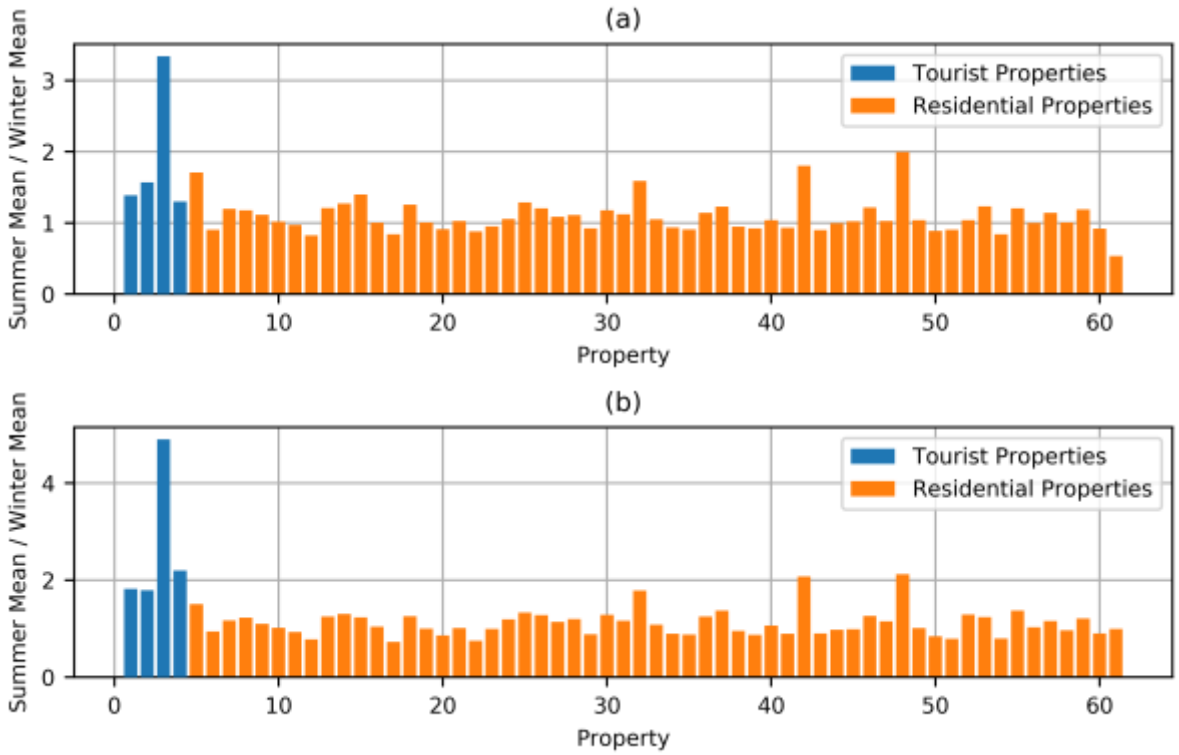


Figure 5.5 – Seasonality in water consumption with: (a) summer defined as April – September and winter defined as October – March. (b) summer defined as May – August and winter defined as November – February.

5.2.4 Weekday and Weekend Behaviour

For many residential properties, it may be expected that the routine of occupants would vary between weekdays and the weekend, as many people (pre coronavirus pandemic in our data) have a 9 – 5, Monday – Friday working pattern, and this variation in routine may be reflected by a variation in water consumption behaviour. In contrast, for occupants of tourist lets, it might be expected that there would not be a large variation in the weekday and weekend routines, as tourists would not usually be restricted in their behaviour by a work schedule, so the weekend routine may not vary from the weekday routine. Of course, not all residential occupants follow a 9 – 5, Monday – Friday working pattern, for example shift workers or retirees. However, as many workers do follow this working pattern, the potential for the variation in weekday and weekend water consumption behaviour to distinguish between residential and tourist properties was investigated.

Figure 5.6 demonstrates this idea. Figure 5.6(a) shows that, for an example tourist property from our data, the average weekday and weekend daily water consumption profiles are extremely similar. Conversely, Figure 5.6(b) shows that, for a residential property, there is notable variation between the weekday and weekend profiles. Figure 5.6(b) demonstrates the idea of variation between weekday and weekend consumption in residential properties and is indicative of many residential properties. However, there were also a large number of properties for which this trend was not apparent, demonstrating a wide variation in the routines of local residents. Figure5.6(a) is reasonably indicative of the other tourist properties, on the other hand.

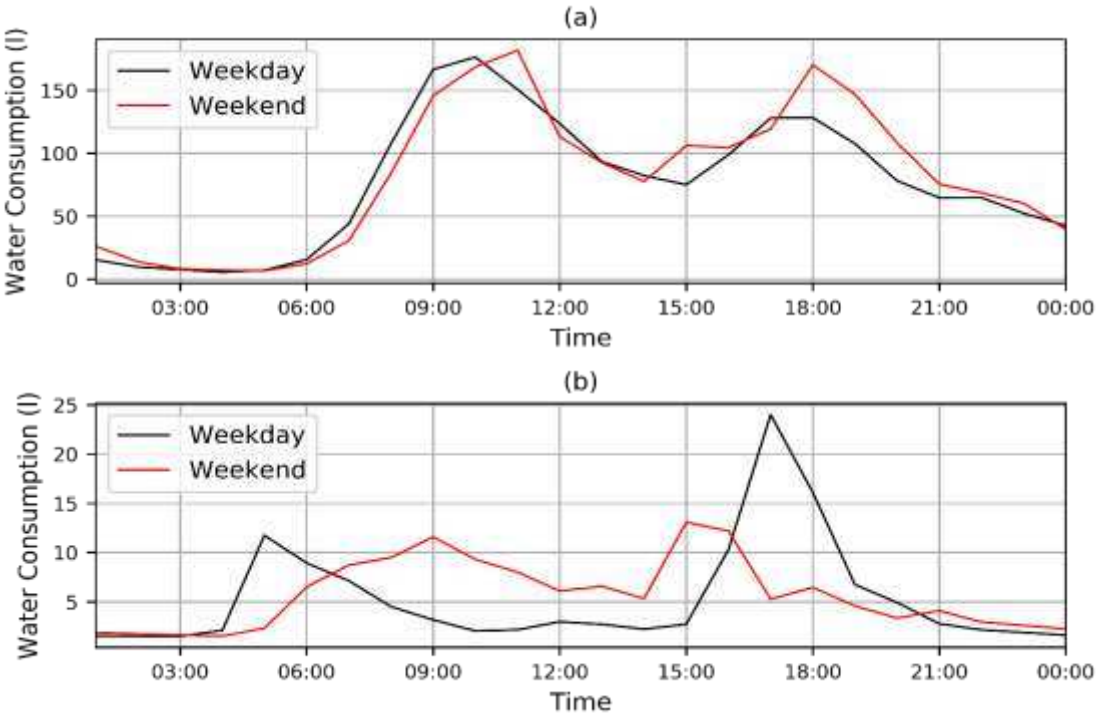


Figure 5.6 – Average weekday and weekend consumption profiles for: (a) a tourist property. (b) a residential property.

A number of metrics to quantify the variation between weekday and weekend consumption were investigated, such as the difference in the volume of water consumed in specific time periods of the day. The metric which provided the greatest discriminator between tourist and residential properties was found to be the time difference between the weekday and weekend morning and evening peak consumption, with the morning defined as between 05:00 and 12:00 and the evening defined as between 16:00 and 23:00. This is demonstrated by Figure 5.6, where the residential property has a weekend morning peak over three hours later than the weekday morning peak and a weekend evening peak approximately two hours earlier than the weekday evening peak. The tourist property, on the other hand, has much smaller time differences between the weekday and weekend morning and evening peaks. This is an imperfect metric, however, as there is much variation in the general residential consumption profiles and many properties did not exhibit either a clear morning or clear evening peak. It should be noted that, for all properties, the average weekend and weekday profiles were calculated using only the days which were defined as occupied using the occupancy detection algorithm (see Section 4), in order to prevent unoccupied days with little or no water consumption from affecting the average consumption profiles.

Figure 5.7 and Figure 5.8 show the time differences between the weekday and weekend morning and evening peaks for each property, calculated using the mean and median weekday and weekend consumption profiles, respectively. It can be seen that there is variation in the peak consumption time differences for all properties. The most clear distinction between tourist and residential properties occurs for the PM peaks, when the average weekday and weekend profiles were calculated using the mean (Figure 5.8(b)). In this case, all of the tourist properties produced positive values, whereas, the majority of the residential properties produced negative values. Although this case produced the greatest distinction between tourist and residential properties, in general, the difference in peak consumption times between weekdays and weekends did not provide as clear a distinction between tourist and residential properties as other methods investigated in this study. For this reason, the differences between weekday and weekend behaviour was not selected as a feature for clustering analysis.

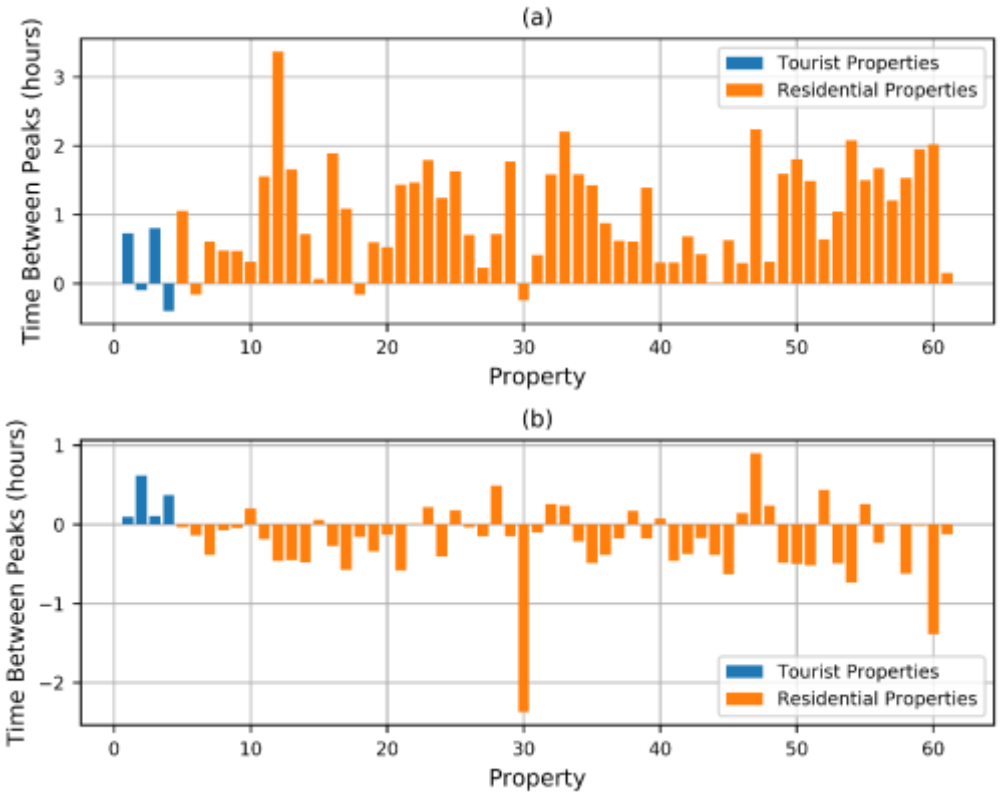


Figure 5.7 – Time difference between the weekend mean and weekday mean peak water consumption for each property: (a) during the morning. (b) during the evening.

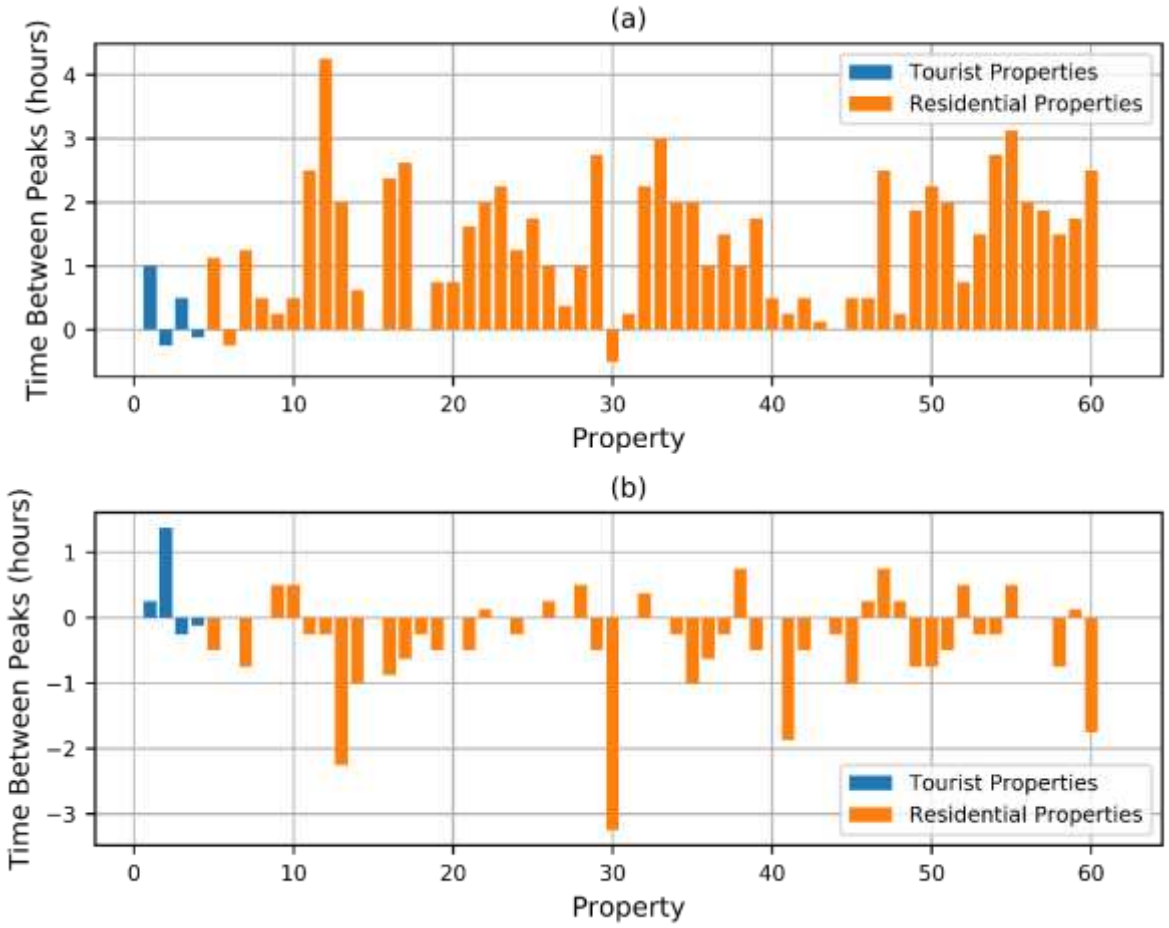


Figure 5.8 – Time difference between the weekend median and weekday median peak water consumption for each property: (a) during the morning. (b) during the evening.

5.2.5 Fourier Analysis

Fourier analysis is a well-established mathematical approach for analysing spatial and temporal signals and is based on the idea that any signal can be broken down into a number of simpler trigonometric functions which, when summed together, produce the original signal. Fourier transformation is used to transform a signal from the spatial or temporal domain to the frequency domain, providing a useful tool for analysing dominant frequencies in a noisy signal. The Fast Fourier Transform (a computationally efficient method to compute the Fourier transform) was applied during this analysis.

Figure 5.9 shows the Fourier transformation of the water consumption signal for a residential property. The x axis represents the period of the signal (converted from frequency for simplicity of analysis) and the y axis represents the Energy Spectral Density (ESD). The ESD provides an indication of how dominant each frequency is in the overall signal, so a high ESD value at a particular frequency implies that that frequency has a high prevalence in the original signal. In Figure 5.9, it can be seen that the highest ESD values occur at periods of 15 minutes and 24 hours. It should be noted that Figure 5.9 shows the ESD values only between 0 and 25 hours, where the highest ESD values were observed. It is unsurprising that the highest ESD values were observed at periods of 15 minutes and 24 hours as the sampling rate of the data is 15 minutes and a 24 hour water usage pattern would be expected due to the daily routine of occupants.

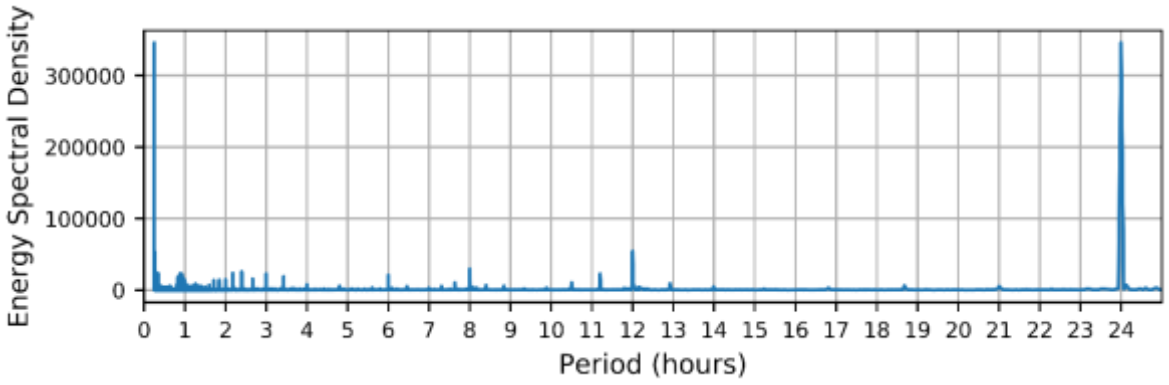


Figure 5.9 – Energy Spectral Density vs period for a Fourier transformed water consumption signal.

Given that tourist properties are unoccupied for periods of time, during which little or no water consumption occurs, it may be expected that the 24 hour period would be less dominant in the overall water consumption signal than that of a residential property. In order to quantify this idea, the ESD value at a period of 24 hours was calculated for each property. These values are displayed in *Figure 5.10*. It can be seen that the ESD value at 24 hours varies considerably over the 61 properties and there appears to be no trend in the values between the tourist or residential properties. Note that the y axis of Figure 5.10 was limited to 100,000 so that the lower ESD values could be clearly observed. A number of properties produced ESD values far higher than 100000. The ESD values at periods of 15 minutes, 12 hours and 1 week were also investigated and similar variations between properties were observed. For these reasons, Fourier analysis was not used to create a feature for clustering of properties.

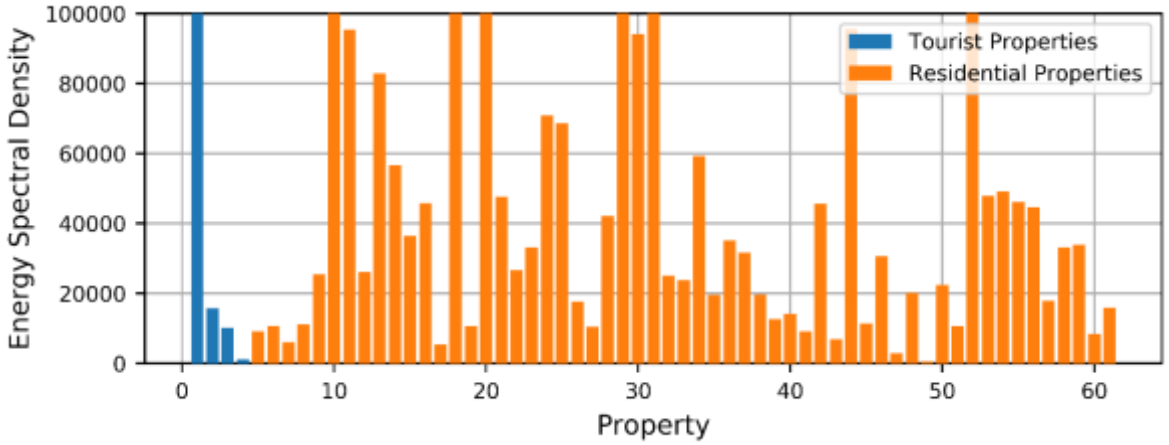


Figure 5.10 – Energy Spectral Density values at a period of 24 hours for each property investigated.

5.2.6 Time Series Clustering

k-means clustering of time series data can also be conducted, so that time series with similar profiles are grouped together. Time series clustering of the daily water consumption profiles was trialled for each property in order to investigate whether this may provide a useful discriminator between tourist and residential properties. No useful metrics to distinguish between the property types were identified, although we acknowledge that time limitations and a lack of expertise within the research team prevented this idea from being explored further. This may be an analysis method which could be investigated further in future work.

5.2.7 Features Summary

A number of different features to distinguish between tourist and residential properties have been investigated. Table 5.1 summarises these features and details whether they have been selected for clustering or discounted.

Table 5.1 – Investigated features which have been either selected for clustering or discounted.

Selected Features	Discounted Features
Occupancy Ratio	Weekday vs Weekend Usage Behaviour
Normalised and Smoothed Mean	Fourier Analysis
Normalised and Smoothed Median	Time Series Clustering
Normalised and Smoothed Standard Deviation	
Seasonality	

5.3 k-means Clustering

5.3.1 Binary two cluster solution

k-means clustering was conducted using the five selected features identified in Table 5.1. The number of clusters was set to two in order to produce a binary distinction between tourist and residential properties. Figure 5.11 shows the results of the k-means clustering. All four of the tourist properties were grouped together, however, five of the residential properties (including Property 17, discussed below) were also grouped into the same cluster as the tourist properties. These results were positive, as only 5 of 61 properties were not grouped into the intended cluster (92% of properties accurately classified). However, in an attempt to improve the clustering results, different combinations of the selected features were utilised.

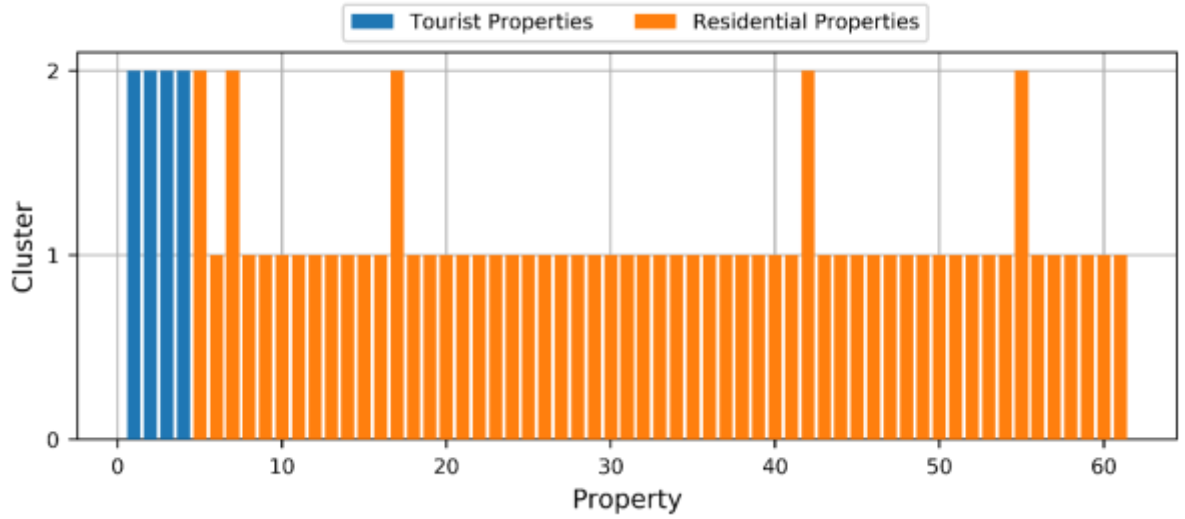


Figure 5.11 – k-means clustering results for two clusters using all selected features.

Following clustering using all of the selected features, clustering was conducted using only the occupancy ratio, as this feature showed a strong discrepancy between tourist and residential properties. Clustering would usually be conducted using more than one feature.

However, in order to better understand how each feature affected the clustering results, a single feature was investigated and then features were added iteratively. Figure 5.12 shows the results of this clustering procedure. Three of the four tourist properties were grouped together, along with Property 17. The remaining tourist property, Property 1, was grouped with the other residential properties. This is unsurprising as the occupancy ratio for Property 1 was greater than that of the other tourist properties and that of Property 17. Although the occupancy ratio of Property 1 was lower than that of the remaining residential properties, it was more similar to that of many of the residential properties than it was to that of the other tourist properties.

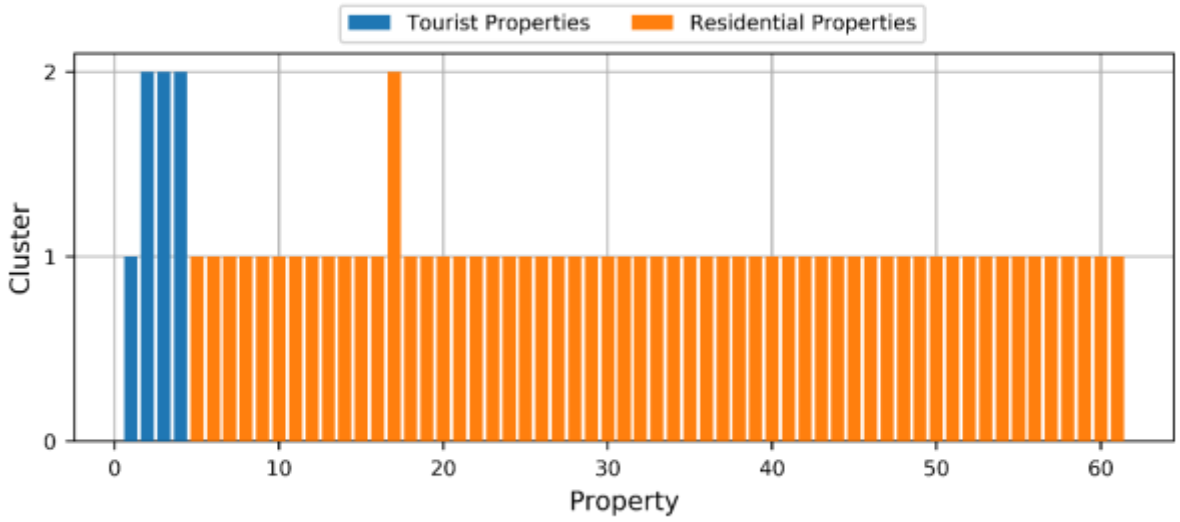


Figure 5.12– k-means clustering results for two clusters using only the occupancy ratio feature.

Following clustering using only the occupancy ratio, clustering was conducted using both the occupancy ratio and the normalised, smoothed standard deviation. The results are shown in Figure 5.13. In this case, all four of the tourist properties were clustered together, along with Property 17 again. This result is an improvement on the previously described clustering results, however, in all cases thus far, Property 17 has been grouped with the tourist properties.

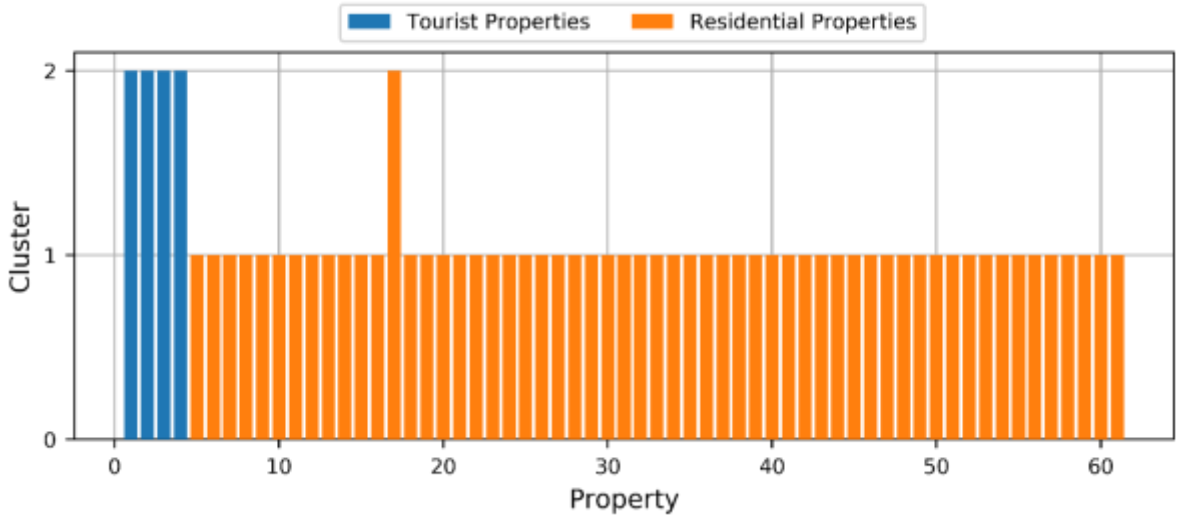


Figure 5.13 – k-means clustering results for two clusters using the occupancy ratio and the normalised, smoothed standard deviation.

Figure 5.14 shows the daily total water consumption over the yearlong measurement period for Property 17, with the occupancy status of each day, as determined by the chosen occupancy detection method, indicated by blue bars. On the surface, the pattern of occupancy appears similar to those of Properties 2 – 4 in that the total period that the property is unoccupied is reasonably high and there are a large number of transitions from occupied to unoccupied days, and vice versa. However, further analysis of the occupancy pattern reveals a dissimilar trend to that of Properties 2 – 4 (which is again dissimilar to the trend for Property 1). Property 17 was occupied throughout most of January and February, but unoccupied over Christmas. There were a large number of unoccupied periods during late spring and early summer and the longest unoccupied period occurred in August, during the school summer holidays. This is the opposite seasonal trend of what might be expected for a holiday let in the South West of England, for which it would be expected that the property was occupied more often during the peak tourist season summer months and, in fact, Property 17 produced the lowest ratio of summer to winter water consumption (our seasonality feature) of all properties (Figure 5.5(b)). It can also be observed that the unoccupied periods were generally shorter for Property 17 than they were for Properties 2 – 4. A possible explanation for the different occupancy trends may be that Property 17 represents a dwelling for which the occupants own a second home which they use for holidays.

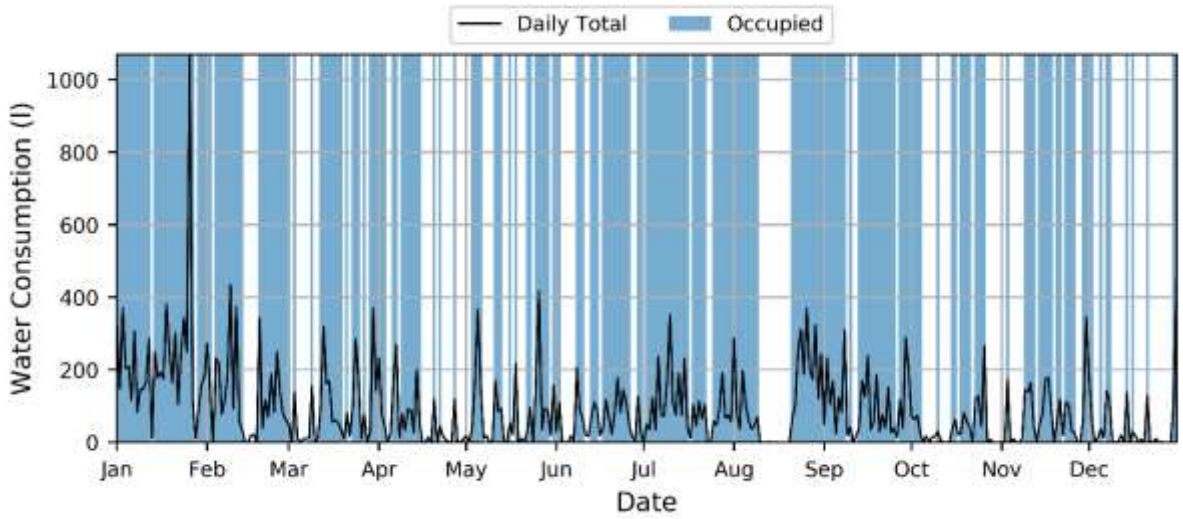


Figure 5.14 – Daily total water consumption for Property 17 with the occupancy status, as determined by the occupancy detection procedure, indicated.

As Property 17 showed a different seasonal trend to the tourist properties, the next feature to be included in the clustering procedure was the seasonality. The results of this clustering analysis are shown in Figure 5.15. Including seasonality in the clustering algorithm distinguished Property 17 from the tourist properties so that all tourist properties were grouped correctly into one cluster and all residential properties were grouped into the other cluster.

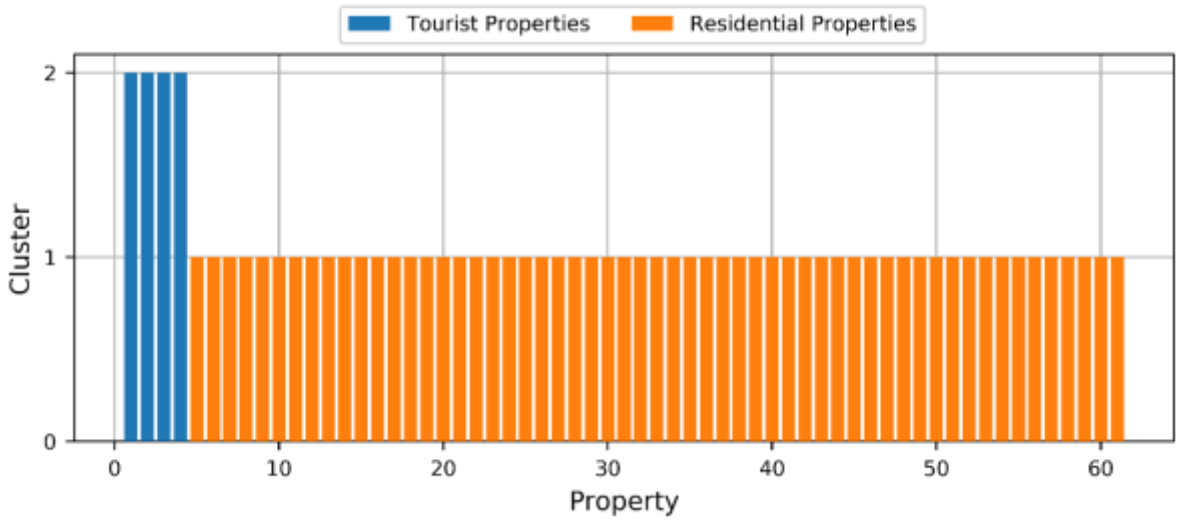


Figure 5.15 – k-means clustering results for two clusters using the occupancy ratio, the normalised, smoothed standard deviation and the seasonality.

For the reasonably small sample of properties available, k-means clustering was able to distinguish between tourist and residential properties with 100% accuracy, when the occupancy ratio, the normalised, smoothed standard deviation and the seasonality were applied to the algorithm. The normalised, smoothed median provided a strong distinction between tourist Properties 2 – 4 and the residential properties and the normalised, smoothed mean also provided a reasonably good distinction between tourist Properties 2 – 4 and the residential properties, as was shown in Figure 5.4. Property 1 showed a dissimilar occupancy trend than the other tourist properties (as shown in Figure 4.6) and produced a higher normalised, smoothed mean and median than the other tourist properties. It is possible that Property 1 represents an outlier and the majority of tourist properties show similar occupancy trends to those of Properties 2 – 4. In this case, the normalised, smoothed mean and median may provide useful discriminators between tourist and residential dwellings. However, Property 1 may not be an outlier and there may be a large variation in the occupancy patterns for tourist dwellings. It is not possible to determine this with a small sample size of four tourist properties. Future work, with a larger sample size of tourist properties, may investigate this further.

5.3.2 Testing the stability of the clustering approach and its applicability at a 1-hour resolution

In order to investigate how stable the grouping of the four tourist properties was, the number of clusters was increased. If the number of clusters were increased and the four tourist properties remained grouped together, with no residential properties, then this would suggest the features selected for k-means clustering provide a strong discriminator between tourist and residential properties.

Figure 5.16 shows the clustering results for three clusters using the occupancy ratio, the normalised, smoothed standard deviation and the seasonality. The four tourist properties were again grouped into the same cluster, with the additional cluster enabling sub-division of the residential properties. This indicates that the three features selected for clustering provide a useful distinction between tourist and residential properties.

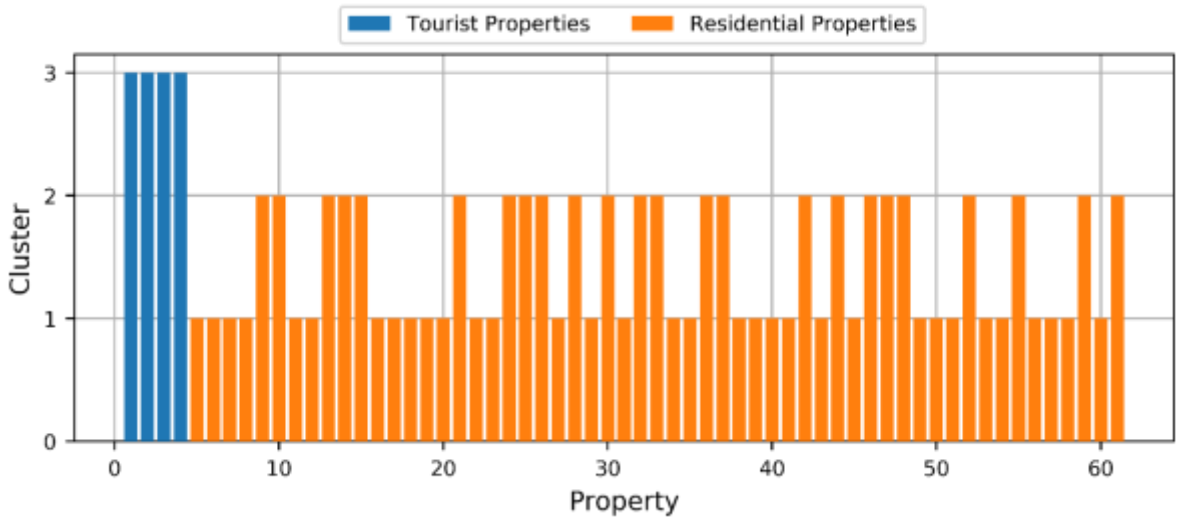


Figure 5.16 – k-means clustering results for three clusters using the occupancy ratio, the normalised, smoothed standard deviation and the seasonality.

When the number of clusters was increased above three, Properties 2 – 4 remained grouped together with no residential properties in the same cluster, however, Property 1 was grouped into a different cluster, with some residential properties. This demonstrates the different water consumption trends shown by Property 1 relative to the other tourist properties in our small sample and may indicate that Property 1 is an outlier, as previously discussed.

Properties 2 – 4 remained in the same cluster as the number of clusters was increased, up to 11 clusters (as shown in Figure 5.17). When the number of clusters was increased above 11, Properties 2 – 4 were grouped into different clusters. Figure 5.17 also shows that Property 17, which had a low occupancy ratio when compared with other residential properties and a low ratio of summer to winter consumption, was grouped into its own cluster.

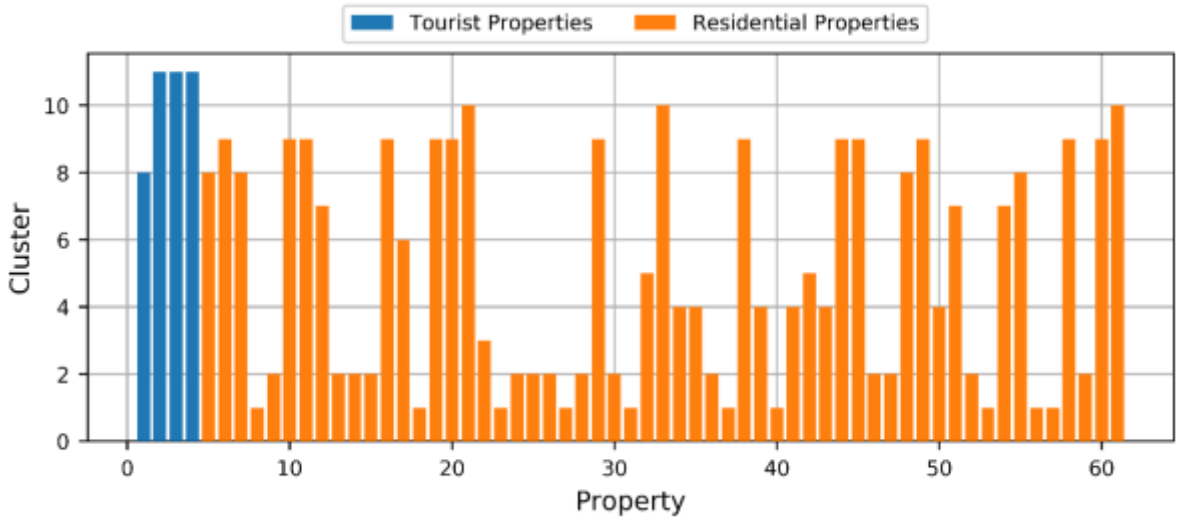


Figure 5.17 – k-means clustering results for three clusters using the occupancy ratio, the normalised, smoothed standard deviation and the seasonality

We thus find that analysis of dwelling level water consumption behaviour may provide a useful mechanism through which to categorise dwellings according to key occupancy trends and characteristics, as explored further in section 6.

We also re-ran the clustering using the property level data aggregated to a 1-hour resolution in order to identify the potential application and stability of these clusters at that resolution. We find that aggregation to the 1-hour resolution naturally smooths the input variables, reducing the variation between tourist and residential properties evident within the means, medians, and standard deviation features. The clustering process outlined above was repeated across all combinations of input features (Table 5.1) and with varying numbers of output clusters.

Using these data at a 1 hour resolution, the 'best' clustering solution – when seeking to generate only two clusters (residential and tourism) - was achieved when using only the features capturing the occupancy ratio and seasonality. Whilst those features did correctly group all residential properties together into a single cluster, only three tourist properties (properties 2-4) were grouped together as tourist properties. The aforementioned property 1, which has many water-use characteristics associated with a residential property, was incorrectly assigned to the residential cluster.

Utilising this small sample of properties, we have been unable to identify a clustering solution which groups all four tourist properties together when using the 1-hour resolution data. Nevertheless, the ability of a two-cluster solution to correctly group three of the four tourist properties utilising only the occupancy ratio and seasonality is encouraging and suggests that a more parsimonious solution, utilising fewer features may be possible. We are, however, limited by the small number of properties and vagaries around the true status of property 1, which has an occupancy pattern and consumption habits that are dissimilar to our other tourist properties, which will be addressed in subsequent work, utilising a larger sample of properties as discussed in the following section.

6 SUMMARY, IMPLICATIONS AND RECOMMENDATIONS

6.1 Summary of findings and implications for ONS

This technical report assessed methodological approaches to infer dwelling characteristics and occupancy trends from high temporal resolution dwelling-level water consumption data, drawing on a small sample of properties in South West England.

We strongly assert that dwelling-level water consumption data can reveal sufficient indicators of dwelling occupancy in order to:

i) **Identify tourist dwellings as a subset of the dwelling stock.** Unsupervised clustering (k-means) was used to distinguish tourist properties from residential properties, allowing future upscaling to a broader sample of dwellings without the need for validation data to train the model. Features of water consumption behaviour for each property were extracted from the data and utilised in the clustering algorithm. A range of features were investigated and the occupancy ratio, the normalised and smoothed mean, median and standard deviation and the seasonality of water consumption are recommended. Variation between weekday and weekend behaviour, Fourier analysis and time series analysis were also investigated, however, no features from these distinguished sufficiently between tourist and residential properties for inclusion in the final model.

ii) **Infer occupancy status of a tourist dwelling (occupied or unoccupied) on a given day.** Our occupancy detection method is able to determine dwelling-level occupancy on a day-by-day basis, with an accuracy of 98.8% (15-min resolution), and 98.7% (1-hour resolution), within our sample data, validated against manual

labelling of the same dataset. A number of occupancy detection methods were investigated, and the applied method defined a property as occupied when both the mean and the number of usage events are greater than 25% of the average for each criteria, for a given property.

As noted in Section 2, collection of these data are a routine and non-invasive activity carried out as part of routine billing, network management and leak detection by water suppliers. Given the geographical monopolies present in the residential water supply sector, a single water supplier can offer near-complete coverage of all domestic properties within their supply region, this may offer considerable advantages over other administrative and commercial data sources (including similar data from energy suppliers) where typically households are distributed across a range of suppliers. The ability to derive indicators of dwelling type and occupancy from these data could afford tremendous re-use value for these data, allowing water suppliers and the ONS to unlock additional value from these commercially held transactional data sources.

Based on the small sample of exemplar properties utilised in this analysis, we strongly suggest that high-temporal resolution water metering data at the dwelling level can be used to identify dwellings with seasonal occupancy patterns. We developed our approaches using data at the 15-min resolution, but suggest that use of these data at a 1-hour resolution (considerably reducing data volumes) offers comparable results. We have demonstrated that occupancy can be reported on a day-by-day basis, with wider implications for understanding usage of other forms of non-standard dwelling which could include 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), it would be entirely feasible for this approach to provide additional indicators of neighbourhood characteristics related to dwelling occupancy and utilisation.

Although our interest is in seasonal occupancy patterns driven by tourism, we suggest that it would be entirely feasible to use our occupancy detection method and some of our key-features of consumption behaviour to identify other dwelling-level 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) which may be identifiable via a period of non-occupancy and/or a change in consumption habits, affording new and timely insights into small-area population mobility. Subject to further consideration of privacy and ethical concerns, alongside the challenges of mining these data at scale, these data could in theory offer near real time insights into dwelling-level occupancy, addressing questions such as ‘was this property occupied on a particular date?’ or, ‘are properties in this area typically occupied on a given day of the year’?.

Whilst these data afford tremendous potential, challenges around data quality assessment, cleaning and pre-processing should not be underestimated. Approximately one-third of the available properties for this study were excluded from analysis due to extended periods of lost signal from the data-logger, which meant that a sufficient temporal coverage (a one year period) could not be extracted. Prior to analysis, these data also required detection of and correction for leakage. The custom-built leakage detection algorithm applied here eliminated the majority of the influence of leaks on recorded water consumption. Some remnants of the leak remained within the data but it was determined that they would not have a meaningful effect on the analysis methods utilised in this study.

6.2 Recommendations for further analysis

The analysis presented in this report drew on a small sample of properties (n=61) to assess feasibility and develop methods. Upscaling this to a larger sample of properties will require

development of additional automated procedures to assess data coverage and completeness (to account for loss of data due to signal loss from data loggers). Additionally, subsequent analysis may address some/all of the following points:

- Further development of the analysis methods to use more than one years' worth of data for each property, and to integrate data from properties for which a different temporal span is available. All tools applied within the analysis reported here utilise exactly one years' worth of data (to tie in with the annual cycle of tourism) potentially wasting valuable data for properties for which in excess of one years' worth of data are available.
- Minor improvements to the leak detection and elimination procedure to remove the remaining effects of leaks on the water consumption data.
- Improve the method used to define usage events for occupancy detection. However, this may not be required if the residual effects of leaks are removed from the data.
- Further investigation into features for clustering of properties could be conducted. For example, time series clustering may provide useful insights, which were not identified in this study.
- The clustering algorithm could seek to distinguish a broader range of dwelling types, for example attempting to separately distinguish second homes based on their occupancy patterns and water consumption profiles.
- Greater insight into verified occupancy patterns for known tourist dwellings in our sample of properties in order to assist model validation. It may be possible to obtain these insights using AirDNA data.
- Investigation of changes in tourist dwelling occupancy characteristics as a result of the COVID-19 pandemic including usage during lockdown periods and evidence of staycation trends.
- Integration of higher temporal resolution data (1-second) which could provide a more granular resolution (and therefore scope to extract specific practices such as toilet flushes) and which could provide additional insights into occupancy (e.g. number of people present within a property).

We are particularly keen to tailor this work towards the needs and interests of the ONS and would be very glad for suggestions as to which of the suggested points above are of greatest value

ACKNOWLEDGEMENTS

This work was supported by the Economic and Social Research Council [grant number ES/T005904/1]and the Leeds Institute for Data Analytics. We are very grateful for the support and data provided by South West Water, especially Paul Merchant (co-author of this report), James Mercer and Josh Pocock.

In developing this work we have had a number of useful discussions with representatives from ONS and would particularly like to acknowledge the input of Susan Williams, Charlie Wroth-Smith, Dr Chris Gale, Donna Clarke and Michael Hawkes.

BIBLIOGRAPHY

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

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

Chen, D., Barker, S., Subbaswamy, A., Irwin, D. and Shenoy, P., 2013, November. Non-intrusive occupancy monitoring using smart meters. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings* (pp. 1-8).

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

Eibl, G., Burkhart, S. and Engel, D., 2018. Unsupervised Holiday Detection from Low-resolution Smart Metering Data. In *ICISSP* (pp. 477-486).

Gössling, S., Peeters, P., Hall, C.M., Ceron, J.P., Dubois, G. and Scott, D., 2012. Tourism and water use: Supply, demand, and security. An international review. *Tourism management*, *33*(1), pp.1-15.

Koech, R., Cardell-Oliver, R. and Syme, G., 2021. Smart water metering: adoption, regulatory and social considerations. *Australasian Journal of Water Resources*, *25*(2), pp.173-182.

March, H., Morote, Á.F., Rico, A.M. and Saurí, D., 2017. Household smart water metering in Spain: Insights from the experience of remote meter reading in Alicante. *Sustainability*, *9*(4), p.582.

Martin, D., Cockings, S. and Leung, S., 2015. Developing a flexible framework for spatiotemporal population modeling. *Annals of the Association of American Geographers*, *105*(4), pp.754-772.

Sadr, S.M., That, L.T., Ingram, W. and Memon, F.A., 2021. Simulating the impact of water demand management options on water consumption and wastewater generation profiles. *Urban Water Journal*, *18*(5), pp.320-333.

Sønderlund, A.L., Smith, J.R., Hutton, C. and Kapelan, Z., 2014. Using smart meters for household water consumption feedback: Knowns and unknowns. *Procedia Engineering*, *89*, pp.990-997.

SWW, 2018. South West Water annual performance report and regulatory reporting 2018. Exeter: South West Water.