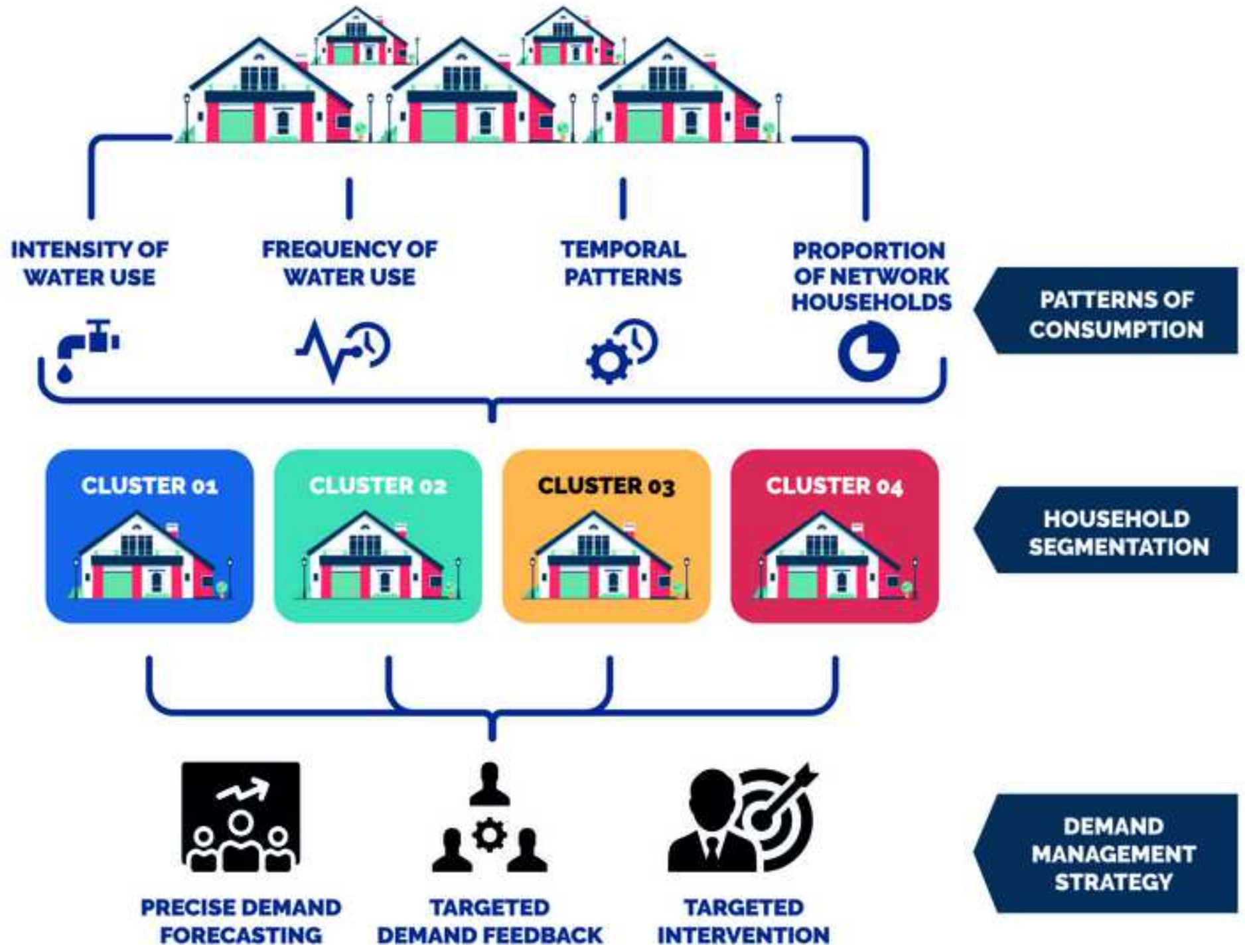
Graphical Abstract



Manuscript

Click here to view linked References

Click here to access/download;Manuscript;Final_manuscript_JRCR_PLAIN.

Resources, Conservation and Recycling, Volume 174, November 2021, Article number 105792
DOI:10.1016/j.resconrec.2021.105792

An empirical water consumer segmentation and the characterisation of consumption patterns underpinning demand peaks

Halidu (Halid) Abu-Bakar [a], Leon Williams [a*], Stephen H Hallett [b]

[a] *Cranfield Centre for Competitive Creative Design; School of Water, Energy and Environment Cranfield University, College Rd, Cranfield, Bedford MK43 0AL*

[b] *Centre for Environment and Agricultural Informatics, Cranfield University, College Rd, Cranfield, Bedford MK43 0AL*

*Professor Leon Williams; Cranfield Centre for Competitive Creative Design, Cranfield University, College Rd, Cranfield, Bedford MK43 0AL; l.williams@cranfield.ac.uk

Abstract

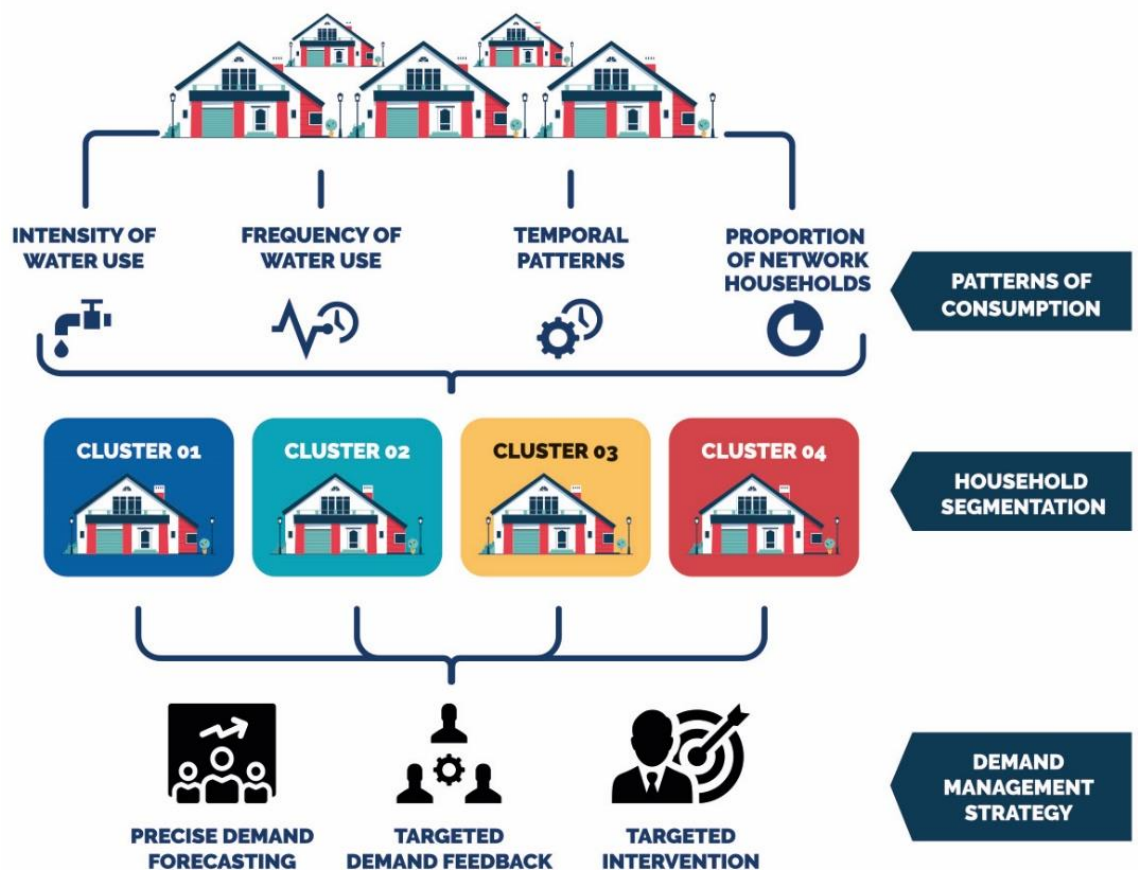
Water utility companies are increasingly searching for ways to understand the complete patterns of household water use and how to improve demand forecasting to achieve effective water demand-side management (DSM). This paper presents an empirical characterisation of household water consumption patterns, based on consumer segmentation, to improve the accuracy of demand forecasting and to develop both proactive and responsive water conservation strategies. Medium resolution smart meter data for 2019 for 10,000households were analysed, revealing four clusters (EP, LM, EM and MP) whose significant differences are underpinned by a variety of indicators in their consumption quartiles. For example, on average households in the EM cluster only record one peak event in 24 hours, compared with the MP clusters' four peak events. Moreover, 2 in every 5 households in MP have a confirmed internal leak compared with 1 in every 5 for the other three clusters. A total of 31,788 Cubic metres (m^3) was consumed, constituting a monthly mean of 2,649m^3, equating to a per household consumption (PHC) of ~270 litres per household per day (l/h/d). Results also revealed the clusters' distributed dominance of hourly demand and the most active clusters in different seasons. The paper concludes that identifying the significant differences characterising consumption patterns and their concomitant impact on network demand will not only serve to enhance demand forecasting and the prediction of geographical consumption hotspots but will also allow the delivery of targeted intervention measures according to households' shared characteristics.

Published by Elsevier. This is the Author Accepted Manuscript issued with: Creative Commons Attribution Non-Commercial No Derivatives License (CC:BY:NC:ND 4.0). The final published version (version of record) is available online at DOI:10.1016/j.resconrec.2021.105792. Please refer to any applicable publisher terms of use.

Keywords:

Peak demand; household water consumption; demand-side management; consumption patterns; segmentation; demand forecasting.

Graphical abstract:



1. Introduction

Peak demand pattern management forms the core of water Demand-Side Management (DSM) strategies and the accurate detection and characterisation of consumption patterns that underpin peak temporal demand can provide a robust framework for improving demand forecasting (Beal and Stewart, 2014; Beal et al., 2016). Studies have established how both exogenous (e.g.seasonal) and endogenous (e.g.household characteristics) variables (Abu-Bakar et al., 2021; Beal and Stewart, 2014) do contribute to peak demand, thereby enabling demand forecasting.

Conventionally, water utilities rely on existing socioeconomic (SE) and sociodemographic (SD) household attributes, e.g. ACORN (Parker and Wilby, 2013), and self-reported behaviours gathered through surveys and water use diaries (Giurco et al., 2008; Willis et al., 2011) used to characterise

household water consumption patterns. However, companies continue to face difficulties in their ability to track effectively individual households' patterns of consumption, establishing the extent to which these patterns change, how they underpin peak demand and their impact on aggregate demand.

Household water consumption patterns are temporally varied and are characterised by peaks and troughs, with diurnal demand being the most diverse (McDonald et al., 2011). While most households exhibit many similarities in these patterns, especially in their sub-diurnal consumption patterns, typically between 00:00 and 06:00 hours, there exist significant differences (Lavee et al., 2013) in peak demand characteristics in terms of their values, temporal and spatial dynamics, average frequency and proportion of household (relative) consumption and network (absolute) demand.

Globally, smart metering in the household water sector has immensely facilitated data collection, providing real-time or near real-time data, and offering powerful insights into household water consumption patterns as well as delivering behaviour-changing feedback to consumers (Cardell-Oliver et al., 2016; Nguyen et al., 2018; Sønderlund et al., 2016). To improve the effectiveness of metering, the UKWIR's (2015) guidance manual on household consumption forecasting emphasised the importance of consumer segmentation, which involves clustering households into categories of water consumption characteristics, in improving the accuracy of demand forecasting and targeted feedback.

The disaggregation of smart meter data remains an effective means of revealing consumption patterns at micro-component levels. However, the process relies on high-resolution data (5 – 15 second interval) and specialised software in combination, making it difficult to incorporate into a utility's routine DSM framework (Abu-Bakar et al., 2021a; Cominola et al., 2018a).

Consumer segmentation, a common practice in the water and energy sectors, supports demand forecasting, forming the basis for targeted feedback and helping to account for the variations in consumption patterns between different types of households (Cominola et al., 2018b; UKWIR, 2015). Segmentation in the water sector may be based on socioeconomic, sociodemographic characteristics, attitudinal or behavioural typologies (Orr et al., 2018; Owen et al., 2009; Rahim et al., 2020), consumption patterns, and geographical location (Avni et al., 2015; Cominola et al., 2018b; McKenna et al., 2014).

The use of Machine Learning (ML), a branch of Artificial Intelligence (AI), has supported numerous water studies by offering tools for high-level exploratory and statistical analysis of big water consumption data

(Cominola et al., 2019b; Dsuerr et al., 2018; Garcia et al., 2017; Rahim et al., 2020). As a sector, the water industry has traditionally tended to be conservative in the uptake of such techniques (Ponce Romero et al., 2017).

Cardell-Oliver (2013) presented a segmentation approach where households were clustered according to their water use signature patterns using medium resolution data from smart meters. Laspidou et al. (2015) performed clustering through the construction of Kohonen Self-Organized Maps to build consumer profiles according to their water consumption, drawing on low resolution (quarterly) data and exploring the potential of differentiation among different consumer types, both for business and household use. In Cardell-Oliver et al's (2016) five-stage water consumer segmentation, smart meter hourly data was used to identify consumers' "regular high-magnitude behaviours", that is, water use behaviours of a consistent temporal occurrence. In their study, the integration of smart meter data with a consumer survey on household characteristics revealed the time and purpose of consumption as well as the principal water use behaviours. The results revealed that 12–15% of households from two areas of Western Australia were responsible for over 80% of the peak-hour demand. Cominola et al. (2018) developed a three-phase segmentation analysis of over 1,000 household accounts to examine the heterogeneity of their water-electricity demand characteristics. Hourly consumption data was combined with household survey information to provide insights for coordinated water-energy DSM. Candelieri (2017) presented a two-stage data-driven approach to clustering for the identification of a variety of anomalies in metered data, such as faults, frauds, cyber-physical attacks, or meaningful changes in water consumption habits, from individual household hourly water consumption data. Padulano and Del Giudice (2018) performed a two-phased cluster analysis of hourly smart meter data on households in Naples, Italy, to classify households according to annual demand patterns. Cominola et al. (2019) developed a data-driven approach for the extraction of information on heterogeneous water end use events, main end use components, and temporal patterns from 1,000 smart-metered households. The model was tested on high-resolution (5-second interval) smart meter data from 327 Australian households, leading to the creation of strategic DSM policies and forming the basis for tailored, non-intrusive feedback intervention. More recently, in a study to quantify the effect of the COVID-19 pandemic on household water consumption patterns in England, Abu-Bakar et al., (2021b) performed an ML-based consumer segmentation of over 11,000 households over 20 weeks, highlighting household

clusters' with the highest and lowest impact on network water supply and revealing a substantial change in water consumption patterns attributable to the COVID-19 lockdown.

This paper presents a data-driven segmentation approach as a means of characterising households' distinctive patterns of consumption, using medium resolution smart meter data logged at an hourly interval. The approach employs Machine Learning (ML) and data analytic (DA) techniques to mine the data for common temporal patterns. The paper aims to provide a basis for identifying the significant differences in the temporal patterns of household water consumption and their concomitant impact on network demand to enhance demand forecasting, enable the prediction of geographical consumption hotspots and facilitate the delivery of targeted intervention according to households' shared characteristics.

2. Materials and methods

All households in the study are from a single water provider, spanning two geographical areas 50 miles apart consisting of 1,034 unique postcodes across 24 District Metering Areas (DMAs). The

data, produced by "Sensus Flexnet" AMI meters, came in a "comma-separated-values" (CSV) format and featured water consumption in litres (l) at a time resolution of one-hour interval.

In this study, 12 months (January to December 2019) of anonymised hourly smart meter data was obtained from about 18,000 household accounts. No endogenous variables about the participants were available, hence no inferences were made as to the socio-demographic (SD) or socio-economic (SE) make-up of the participating households.

ML processes (Fig.1) were employed to prepare and explore the datasets for common patterns before building a predictive model to classify new datasets.

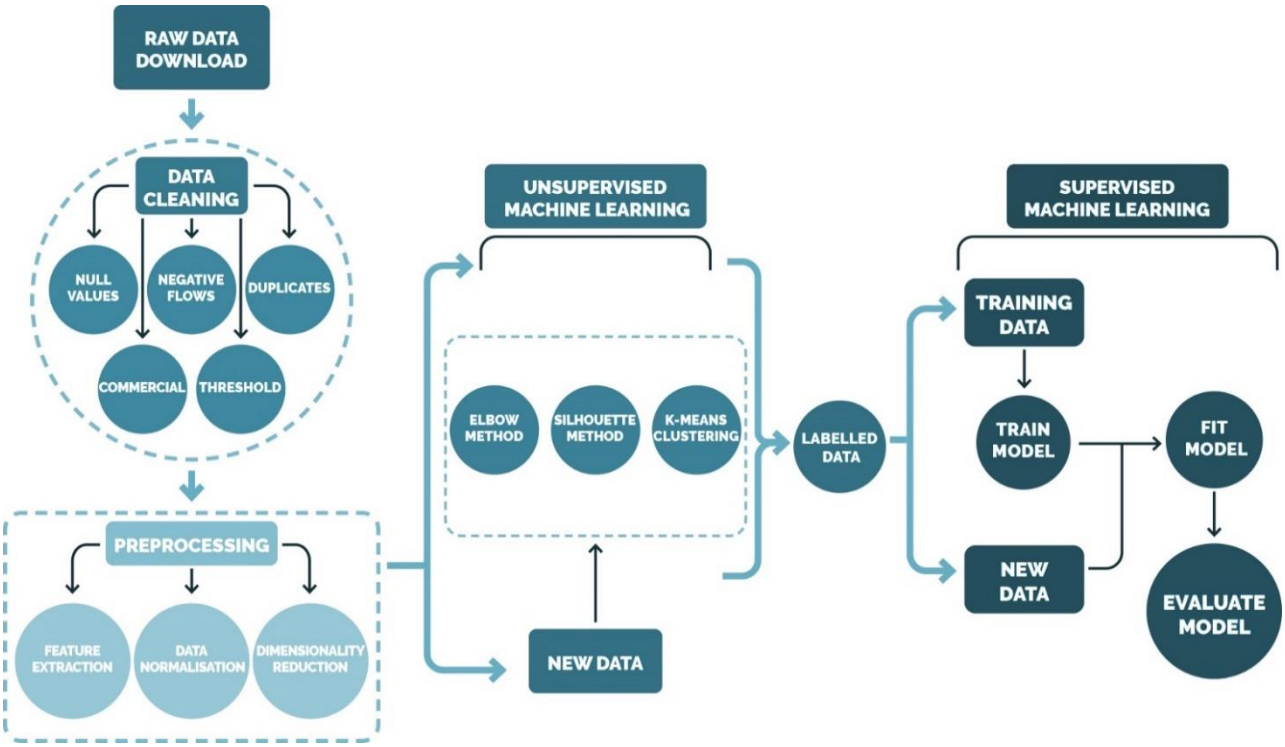


Fig. 1. Machine Learning processes adopted in the processing workflow.

2.1 Data fusion and pre-processing

Three preprocessing techniques were applied: 1) Data cleaning to remove anomalies and outliers and to correct inconsistencies in the data. 2) Data reduction and attribute construction to remove redundant features, create new features and condense the data to improve the efficiency of the data-mining process. 3) Feature scaling and dimensionality reduction to improve the accuracy of the ML algorithms (Ding and He, 2004; Jain and Dubes, 1988).

All null values (Lost data), commercial customers, negative flows, outliers and duplicate records were removed from the datasets and all households in those categories excluded from the study, leaving just ~10,000 households with consistent data across all 12 months. The raw data, consisting of approximately 14 million rows per month and 6 columns (“MeterRef”, “CustomerRef”, “SampleRate”,

“ReadDateTime”, “MeterRead” Vol (l)), was pivoted into 18 thousand rows of mean 24-hour monthly consumption featuring one customer per row with three additional columns appended featuring “REF” for Customer Reference; “DMA” for District Metered Areas and “Postcode”.

Preparing the data for clustering requires normalisation as a means of assigning the data points equal weighting by aligning their attribute distributions to a normal distribution (Han et al., 2012; Kotsiantis et al., 2006). This approach is particularly useful for clustering algorithms such as k-means (Aggarwal and Reddy, 2014). The results of two normalisation algorithms were compared: the “min-max scaler” (which transforms the minimum value of every row into 0, the maximum into a 1, and every other value into a fraction of 1) and the “standard scaler”, also known as “z-score” or “zero-mean normalisation” (which is calculated by subtracting the mean from an individual data point and then dividing the difference by the standard deviation) (Han et al., 2012). Min-max was chosen as it preserves the relationships among the original data values, i.e., all row data points are expressed as a fraction of the row max, which makes it particularly ideal for revealing patterns by, for example, giving all peaks equal weight (Fig. 2). The min-max scaler was applied to achieve a linear transformation on the original data by essentially normalising the data range such that the range was set between 0 and 1:

$$z = \frac{x_i - min(x)}{max(x) - min(x)} \quad (1)$$

Where z is the normalised value; x_i is the raw value of range attribute being normalised; $min(x)$ is the minimum attribute of the range and $max(x)$ is the maximum attribute of the range.

Fig. 2 compares the outcomes of the min-max and z-score normalisation of four households, presenting these in relation to the original data. Households A, B, C and D have PHCs of 978L, 422L, 169L and 117L, max of 68L, 70L, 24L and 24L, and mean of 40.75L, 17.58L, 7.04L and 4.88L respectively. The ‘Original data’ and ‘Z-Score’ subplots show the consumption quartiles in their respective positions, which do not offer a uniform benchmark for pattern comparison. The MinMax subplots, by contrast, show the peak as 1, with the lower quartiles expressed as proportions of the peak while preserving the relationships among the values of the original data.

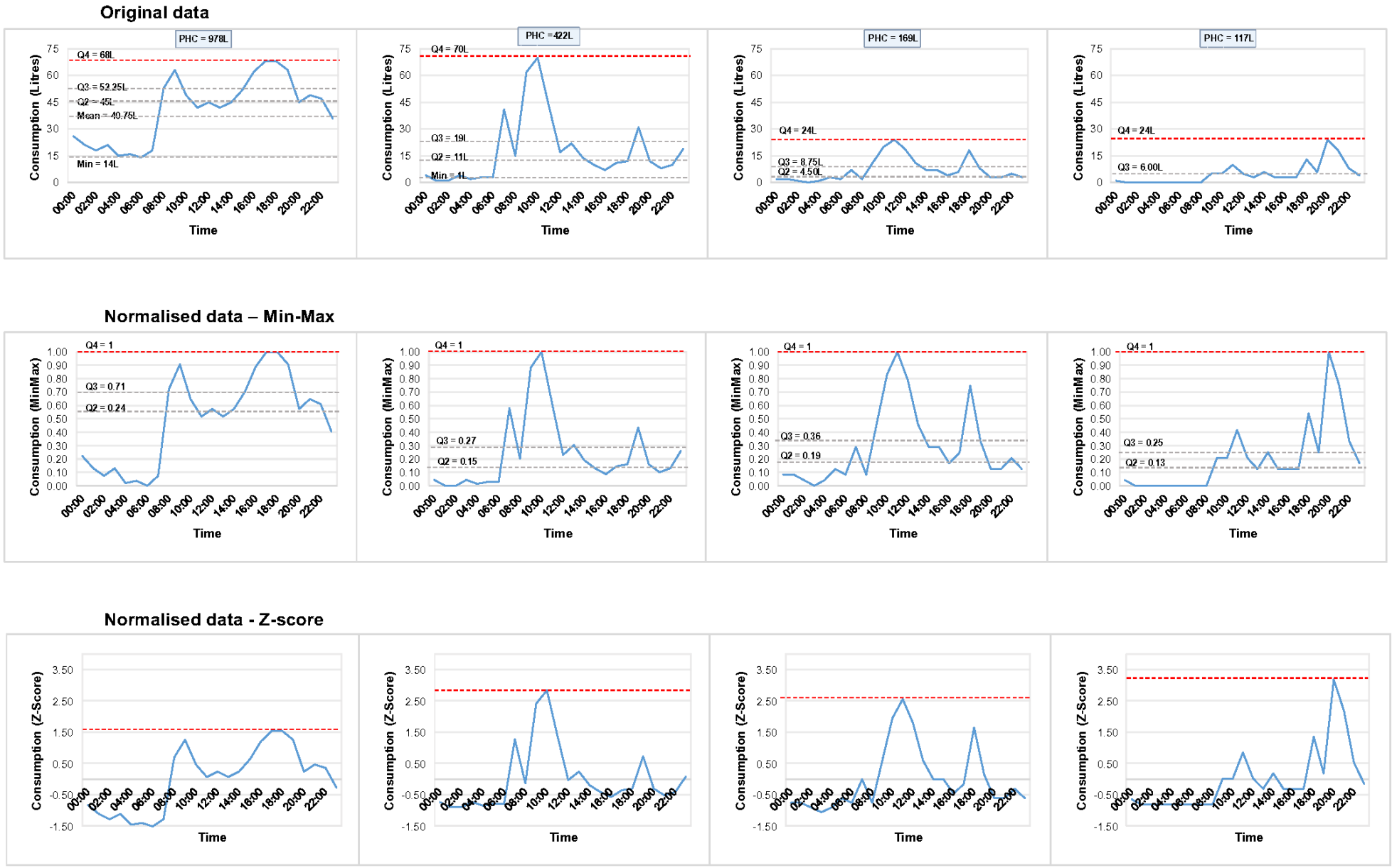


Fig. 2: Comparing normalisation results of a sample of 2019 data with the original data

2.1.1 Dimensionality reduction

The multidimensionality of the dataset presented a common challenge that is faced in exploratory ML. Household temporal water consumption patterns are known to possess many similarities especially in relation to their sub-diurnal characteristics, typically between 01:00 and 06:00 hours (Balacco et al., 2018). Dimensionality reduction was performed on the normalised dataset using the 't-Distributed Stochastic Neighbor Embedding (t-SNE)' (Van Der Maaten et al., 2014; Van Der Maaten and Hinton, 2008) algorithm which, like Principal Component Analysis (PCA) (Jolliffe, 2010), allows for dimensionality reduction of high-dimensional objects into low-dimensional points using feature extraction. This simplified the dimensionality by reducing the 'noise' (Priyadarshini et al., 2019) or the redundant similarities within the dataset, allowing the clustering algorithm to reveal the 'essence' therein and to cluster all observations on the basis of their significant differences. The t-SNE projected the high-dimensional data (the 24-hour features) into low-dimensional space, creating a two-dimensional output, by modelling the probability distribution of neighbouring data points in the dataset while preserving "as much of the significant structure of the high-dimensional data as possible in the low-dimensional map" (Van Der Maaten and Hinton, 2008). In Cominola et al.'s (2018 and 2019) studies, t-SNE was used for feature extraction in their household water demand profiling.

2.2 Machine learning process

Different unsupervised ML techniques were explored for pattern recognition and predictive modelling on the two-dimensional data for unknown patterns and to organise the data into a specific number of clusters based on found patterns. Subsequently, the labelled datasets were used to train a supervised classification model to "predict" the classes of the remaining normalised datasets.

2.2.1 Clustering

The patterns desired are those that are distinctive and peculiar enough to form hard clusters, which in the case of this study, are the significant differences in households' temporal water consumption patterns, achieved through t-SNE. The January pre-processed dataset was run through the k-means (Jain, 2010; Macqueen, 1967) algorithm, a method used widely in water studies (Cardell-Oliver, 2013; Cominola et al., 2019, 2018b), whose goal is to identify and cluster data points based on a similarity metric determined by the Euclidean distance (Macqueen, 1967) while trying to move clusters as far apart

as possible (Yuan and Yang, 2019) such that, each object must belong to exactly one group, where each partition represents a cluster and k ≤ n, where k is the number of specified clusters. That is, it divides the data into k groups (each with a defined centroid) (Han et al., 2012; Zhu et al., 2010).

The objective function for the K-means clustering algorithm is the squared error function (Yuan and Yang, 2019):

$$J = \sum_{k=1}^{k} \sum_{i=1}^{n} \|(x_i - \mu_k)\|^2 \qquad (2)$$

Where J is the objective function (sum of the squared error), k is the number of clusters, n is the number of objects (data points), x_i is object i, μ_k is the centroid for $x_i's$ cluster (thus $\|(x_i - \mu_k)\|$ is the Euclidean distance between point x_i and centroid μ_k).

Determining the optimum number of clusters (k) is a crucial element of k-means clustering, which is often attained by measuring the intra-cluster variation also referred to as the sum of the squared distances to the nearest cluster centre aka or inertia. The elbow method (Han et al., 2012), a heuristic for choosing k using the curve of the sum of within-cluster variances showed $k = 4$ (Fig. 3A).

The silhouette method (Rousseeuw, 1987), used to measure the similarity of objects in the dataset to their assigned clusters compared to other clusters, returned a coefficient score of 0.41 at k=4, compared with 0.37, 0.40 and 0.38 at k=2, k=3 and k=5 respectively, making 4 the optimal number of clusters. The equation is as follows:

$$s(o) = \frac{b(o) - a(o)}{max\{a(o), b(o)\}} \qquad (3)$$

Where $s(o)$ is the silhouette coefficient value of data point or object o in the dataset, $a(o)$ the mean distance between object o and the neighbouring objects in the same cluster, and $b(o)$ the minimum mean distance between o and all objects in other clusters.

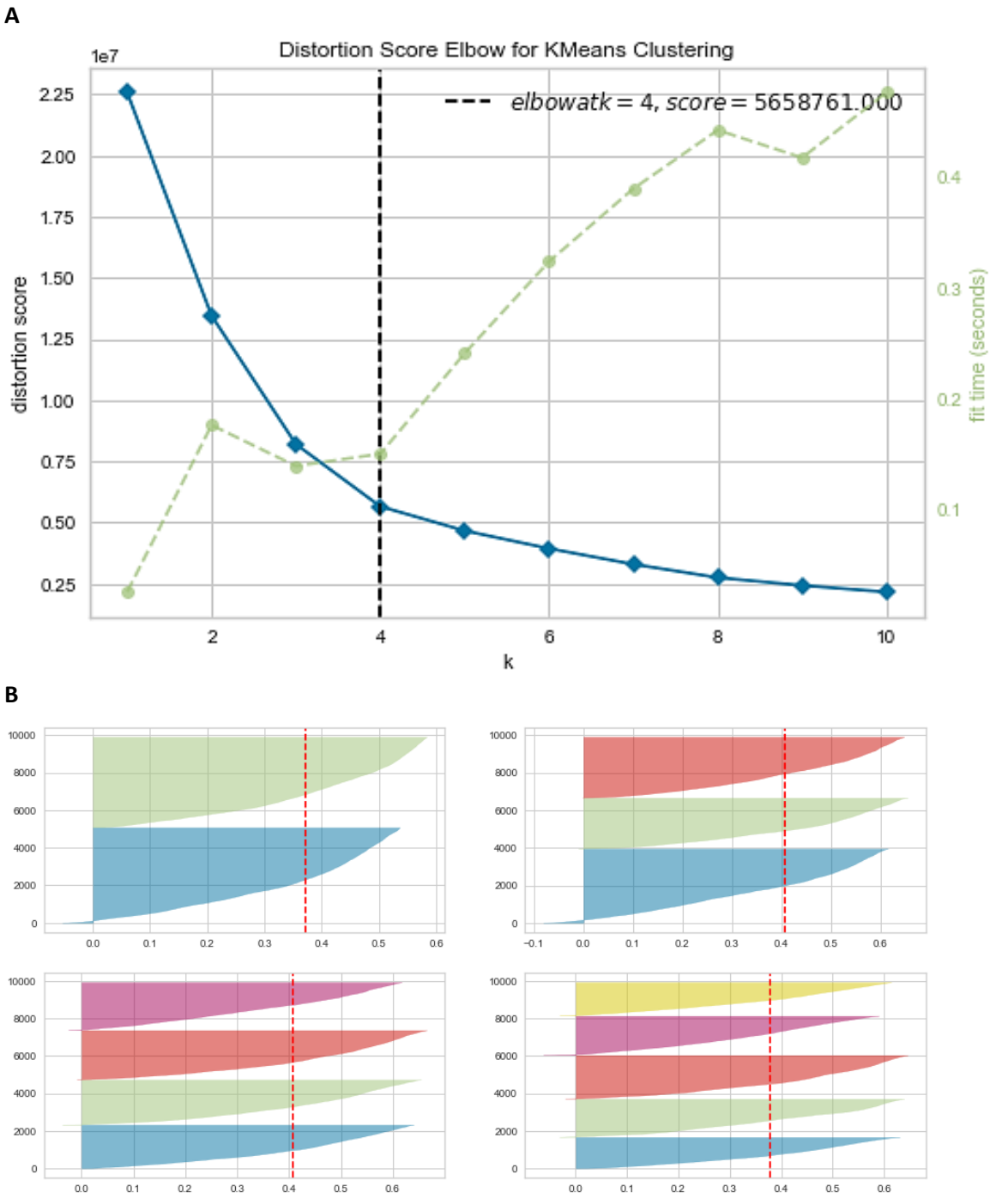


Fig. 3: A. Elbow method to determine k ; B. Silhouette coefficient scores for the optimum number of k

Multiple datasets (February to December 2019) were processed through the elbow and silhouette methods with each returning a coefficient score of 4.0 ~ 4.1 at k=4 and lower scores at all other k values (Fig. 3B). The four clusters were thus labelled 0,1,2 and 3. Cluster 0 is henceforth referred to as 'Evening Peaks' (EP), Cluster 1 'Late Morning' (LM), Cluster 2 Early Morning (EM) and Cluster 3 Multiple peaks (EP).

2.2.2 Classification model

A supervised model was built to classify subsequent input data using the labelled dataset. A “train-test-split” model evaluation procedure was used to compare the classification accuracies of two supervised learning algorithms: K Nearest Neighbor (KNN) – a distance-measuring pattern classification algorithm and Logistic Regression (LR) – a probabilistic pattern classification algorithm (Bishop, 2011; Dreiseitl and Ohno-Machado, 2002; Tsai et al., 2011). Each of the labelled datasets was split into several train test permutations and used in training and testing the model to evaluate how well each algorithm performed in predicting the labels on that same data. Starting with KNN, the accuracy of the model was tested within a “k range” of k=1..k=25, showing a general rise in the testing accuracy as the value of k increases, reaching its peak of between 89 and 91% at k=25, for all the datasets (January – December 2019). Testing accuracy for LR was between 93% and 94% on all the datasets, which makes it a better performing model for the dataset.

2.3 Test for significant difference

Tukey's multiple comparison test (Salkind, 2010) (also known as the Honestly Significant Difference test – or HSD for short) was used to evaluate whether there is any evidence that the means of the multiple features of the four clusters differ significantly, using the following equation:

$$HSD = \frac{M_1 - M_2}{\sqrt{MS_w\left[\frac{1}{n}\right]}} \quad (4)$$

Where HSD is the Honestly Significant Difference; M_1 - M_2 are the mean values; MS_w is the Mean Square Width and n is the Number per Mean. The test compared the differences between each of six sets of pairs (EM vs EP; EM vs LM; EM vs MP; EP vs LM; EP vs MP and LM vs MP) of means for the 12 months, 83% of the compared pairs of means for all four clusters at 00:00 – 02:00, 05:00, 08:00, 10:00 – 11:00, 14:00 and 18:00 have a p-value of $p < 0.05$, with only one pair having insignificant values. The percentage of pairs with $p < 0.05$ for the remaining time features are 03:00 and 06:00, 67%; 04:00 and 06:00, 50% (meaning half of all households across the clusters exhibit similar consumption patterns at this time). 100% HSD is achieved for all pairs at 07:00, 12:00 – 13:00, 15:00 – 17:00 and 19:00 – 23:00. The cumulative results in Fig. 4 (A-F) show the “P-adjust” values of each pair over 24 features (hours).

The heatmap in Fig. 4G shows clusters' temporal patterns, highlighting times of high and low consumption.

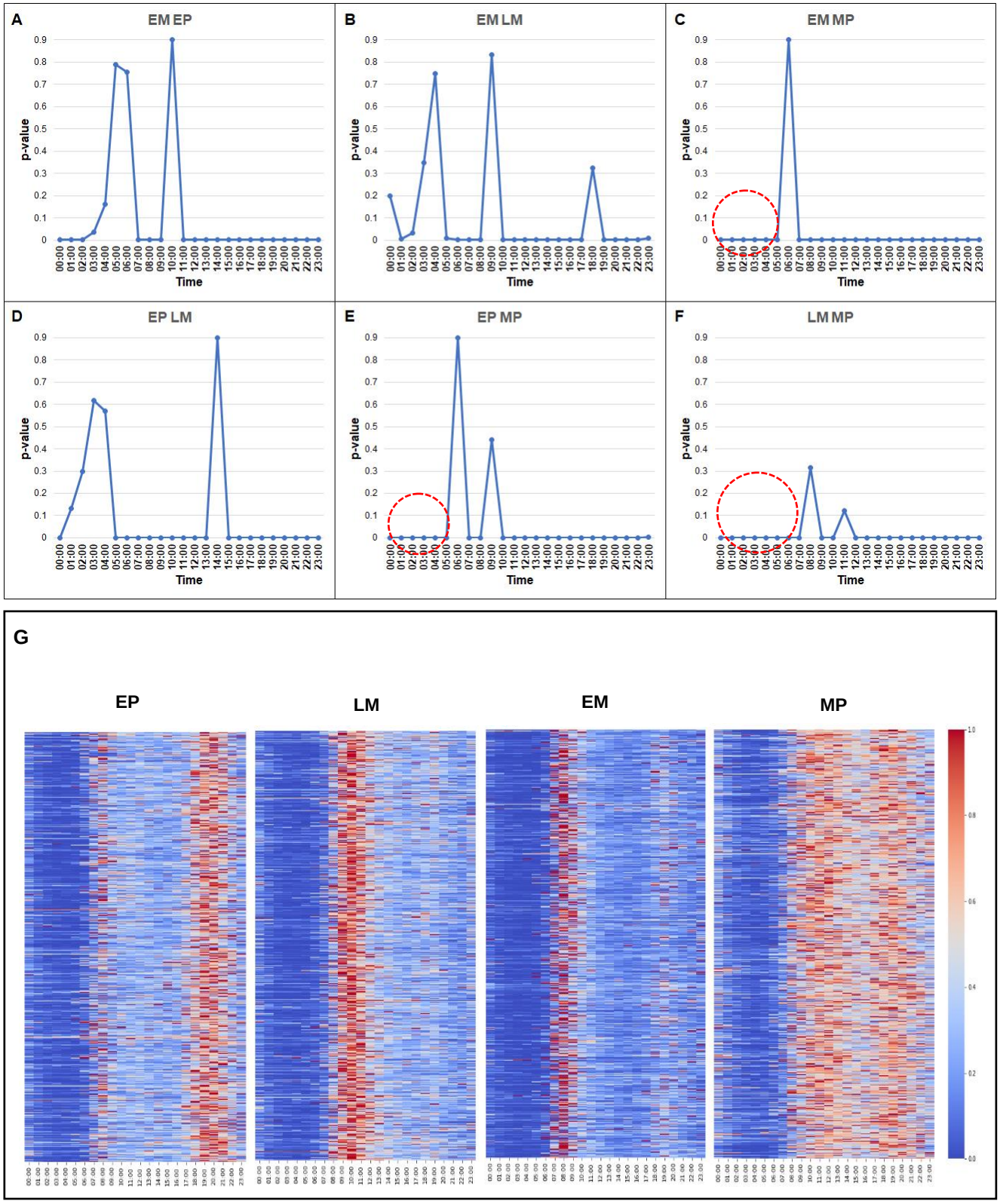


Fig. 4. A-F. p-values for compared pairs of means for the 24 features of all four clusters (Circles indicate the fact that MP's sub-diurnal pattern is significantly different from that of the other three clusters). G. heatmap highlighting clusters' temporal patterns over 24 hours.

3. Analysis of results

The 10,000 households in the study (henceforth referred to as the network) consumed a total of 31,788 Cubic metres (m^3), equating to a monthly mean ($\bar{x}$) of 2,649m^3 with a standard deviation (σ) of 316m^3, amounting to an average per household consumption (PHC) of ~270 litres per household per day (l/h/d), between January and December 2019. The next sections present a detailed analysis of the four clusters' representation of the network households, their share of network demand and how their temporal and seasonal dynamics affect aggregate demand.

3.1 Clusters' significant differences

The significant differences of the four clusters are defined, primarily, by the diversity of their consumption characteristics, temporal patterns, relative and absolute consumption profile (including PHC). Relative consumption profile refers to the intracluster ratio of hourly consumption to PHC (e.g. peak : PHC or lowest quartile : PHC ratio), whereas the absolute consumption profile represents the ratio of individual clusters consumption profile to network demand (e.g. Cluster's peak : network demand at peak ratio). Although most households across the four clusters exhibit a lot of similarities in consumption patterns, especially in their sub-diurnal events, there exist some substantial variations in peak demand characteristics in terms of their values, temporal dynamics, average frequency and how they relate to relative and absolute demand. Table 1 presents the monthly summary of the clusters' absolute and relative consumption profiles.

Table 1. Monthly cluster profiles detailing their network representation and share of aggregate demand for 2019

	Clusters' proportion of network households				**Clusters' mean network consumption (m^3/H/D)**				**Clusters' share of network demand**				**Clusters' $\bar{x}$ PHC (L/H/D)**				**Clusters' σ PHC (L/H/D)**			
Month	**EP**	**LM**	**EM**	**MP**	**EP**	**LM**	**EM**	**MP**	**EP**	**LM**	**EM**	**MP**	**EP**	**LM**	**EM**	**MP**	**EP**	**LM**	**EM**	**MP**
January	28%	28%	22%	22%	828.0	593.4	533.6	636.1	32%	23%	21%	25%	295.4	214.4	250.9	291.9	160.6	129.4	143.9	168.3
February	28%	30%	20%	22%	801.9	644.8	478.1	675.4	31%	28%	18%	26%	293.8	216.0	241.1	309.8	155.5	120.2	126.3	202.0
March	28%	28%	23%	20%	837.4	599.5	562.4	622.3	32%	23%	21%	24%	299.1	215.9	245.5	309.8	153.7	125.8	133.4	205.6
April	28%	29%	18%	25%	832.7	659.1	418.6	782.3	31%	24%	16%	29%	298.2	231.4	231.8	321.8	160.7	138.6	125.1	203.5
May	30%	26%	21%	23%	861.8	558.1	476.2	710.2	33%	21%	18%	27%	290.4	216.8	231.2	312.0	159.5	129.0	139.6	216.2
June	30%	26%	21%	23%	910.5	571.5	506.6	757.1	33%	21%	18%	28%	306.9	224.9	244.6	329.5	166.1	132.4	152.6	236.3
July	33%	23%	20%	23%	1033.4	543.3	493.1	792.6	36%	19%	17%	28%	315.0	237.2	244.1	346.7	172.1	143.8	153.8	252.2
August	26%	30%	18%	26%	762.0	710.3	401.7	872.1	28%	26%	15%	32%	294.1	241.0	228.3	338.2	165.7	139.7	132.2	238.9
September	30%	27%	21%	22%	887.4	575.2	501.2	690.5	33%	22%	19%	26%	297.7	215.0	241.9	321.3	157.2	128.6	138.5	239.3
October	29%	28%	21%	22%	817.1	560.1	502.0	653.5	32%	22%	22%	32%	286.0	205.2	237.4	300.4	143.0	114.6	126.9	214.2
November	28%	28%	22%	22%	807.2	580.7	532.3	634.3	32%	23%	21%	25%	292.1	209.2	241.1	297.7	147.4	120.3	129.9	201.1
December	21%	34%	14%	31%	577.1	749.3	303.1	949.1	22%	29%	12%	37%	280.2	222.9	223.7	305.7	139.5	126.4	122.6	184.5
Average	**28%**	**28%**	**20%**	**23%**	**829.7**	**612.1**	**475.7**	**731.3**	**31%**	**23%**	**18%**	**28%**	**295.7**	**220.8**	**238.4**	**315.4**	**156.8**	**129.1**	**135.4**	**213.5**

3.2 Cluster characterisation

As observed in Abu-Bakar et al's (2021b) characterisation, the temporal patterns of the four clusters remain the same irrespective of the volume of consumption, seasonal and other variables. The terms “Q1”, “Q2”, “Q3” and “Q4” were used to describe consumption range in the lowest quartile, second quartile, third quartile and fourth quartile, respectively. The consumption quartiles are calculated thus:

$$Q4 = C_h \geq 0.75(max_i); \quad (5)$$

$$Q3 = 0.5(max_i) \geq C_h < 0.75(max_i); \quad (6)$$

$$Q2 = 0.25(max_i) \geq C_h < 0.5(max_i); \quad (7)$$

$$Q1 = C_h \leq 0.25(max_i); \quad (8)$$

where C_h is the hourly consumption value and max_i is the maximum daily consumption.

Consumption quartiles provide a means of quantifying the intensity of households’ water use and offers a way of characterising peak consumption, lowest consumption and what occurs in between. The cluster’s main significant differences lie in their Q4 consumption patterns underpinning demand peak. The significant differences begin to wane in the lower quartiles with all clusters exhibiting near-identical Q1 characteristics, with the exception of MP whose Q1 differs both in terms of intensity and frequency. For example, 99% of households across the network record their lowest consumption at 04:00, with the exception of MP whose representation is 90% of households at 04:00. All households’ consumptions between 00:00 and 06:00 are in the Q1 range.

This section presents a detailed characterisation and quantification of (i) clusters’ temporal peculiarities and (ii) clusters’ consumption profiles.

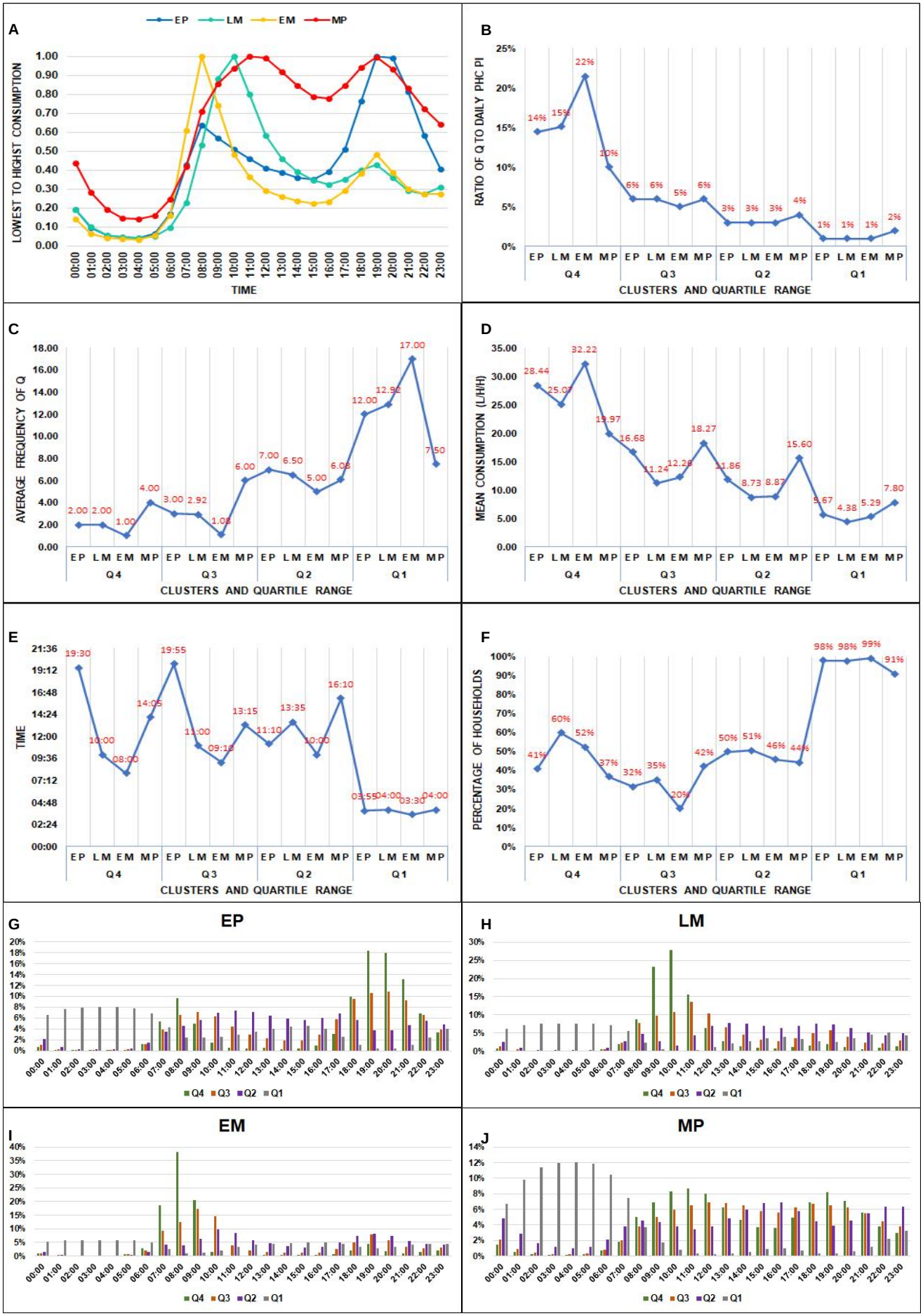

A
EP
LM
EM
MP
LOWEST TO HIGHST CONSUMPTION
TIME
B
RATIO OF Q TO DAILY PHC PI
CLUSTERS AND QUARTILE RANGE
Q4
Q3
Q2
Q1
C
AVERAGE FREQUENCY OF Q
CLUSTERS AND QUARTILE RANGE
D
MEAN CONSUMPTION (L/H/H)
CLUSTERS AND QUARTILE RANGE
E
TIME
CLUSTERS AND QUARTILE RANGE
F
PERCENTAGE OF HOUSEHOLDS
CLUSTERS AND QUARTILE RANGE
G
EP
H
LM
I
EM
J
MP

Fig.5. Characterisation of clusters' patterns of consumption. A. Clusters' typical hourly consumption patterns as a fraction of their max. B. Clusters'quartile ranges expressed as a ratio of their relative PHC per Instance (PI). C. Clusters' average relative frequency of consumption quartile instances. D. Mean relative consumption per hour per quartile. E. Average time when most households in cluster reach quartile. F. Average relative percentage of households reaching their quartiles at (E). G. Typical profile of EP's consumption quartiles. H. Typical profile of LM's consumption quartiles. I. Typical profile of EM's consumption quartiles. J. Typical profile of MP's consumption quartiles.

EP emerged as the largest cluster, both in terms of the proportion of network households it represents and its share of both monthly mean and total network consumption. With an average representation of 28.34% of network households, EP has a more than proportionate total annual consumption of 9,956m^3 – some 31.25% of network consumption, a monthly $\bar{x}$ of 830m^3 and σ of 101m^3, equating to an average daily $\bar{x}$ PHC of 296l/h/d and σ of 157l/h/d (Table 2). ~50% of consumption in the Q4 range for this cluster occurs between 19:00 and 21:00, with ~10% occurring at 08:00 and 18:00. ~41% of households in EP typically reach their peak (consumption in the Q4 range) at 19:00, 40% at 20:00 and 26% at 21:00. On average, every instance of Q4 equates to ~28l/h/h, has a daily frequency of 2 and constitutes ~14% of PHC per instance (PI). Q3 has an average daily frequency of 3, occurring between 18:00 and 21:00, a mean of ~16.7l/h/h, constituting 6% of PHC PI. Q2 has an average frequency of 7, a mean of ~11.9l/h/h and constitutes 3% of PHC PI, while Q1 occurs 12 times on average at ~5.7l/h/h and only constitutes 1% of the daily amount of water consumed PI. This means that, on average, 50% of the frequency of daily water use for this cluster is in the lowest quartile. Consumption in the Q1 and Q2 ranges are relatively temporally ubiquitous. On average, 51% of all households in this cluster record 0 water consumption at 04:00 (Fig.5).

LM represents a similar proportion of network households as EP – 28.07%. However, households in this cluster have a less than proportionate share of network consumption of 7,345 m^3, constituting only 23.43% of network demand, a monthly $\bar{x}$ 612m^3 and σ of 62m^3, equating to an average daily $\bar{x}$ PHC of 221l/h/d and σ of 129l/h/d (Table 2). ~60% of households in LM typically reach Q4 at 10:00, 54% at 09:00 and 32% at 11:00. This cluster records its highest concentration of Q4 (~67%) events between 09:00 and 11:00 and has an average Q4 frequency of 2 with each instance equalling ~25l/h/h, constituting ~14% of PHC PI. LM's Q3 has an average frequency of 3, makes up 6% of PHC at an average of 11l/h/h PI; Q2's average frequency is 7 and constitutes 3% of PHC at 8.711l/h/h PI, while

their Q1 is recorded 12 times on average at ~4l/h/h and only constitutes 1% of PHC PI, also meaning that, on average, 50% of their average daily water use is in the lowest quartile. The diurnal prevalence of Q1 and Q2 is similar to that of EP. On average, 53% of all households in this cluster record 0 water consumption at 04:00 (Fig.5).

Comprising only 20.14% of households, EM represents the fewest of network households. It also has a less than proportionate portion of network demand of 5,709m^3, that is 18.17%, a monthly $\bar{x}$ of 476m^3 and σ of 67m^3, equating to an average daily $\bar{x}$ PHC of 238l/h/d and σ of 135l/h/d (Table 2). ~54% of households in EM reach Q4 at 08:00, 29% at 09:00 and 27% at 07:00. EM records ~77% of its daily Q4 events between 07:00 and 09:00 and has an average Q4 frequency of 1 – making it the cluster with the fewest hourly peak events – with each instance equalling ~32l/h/h, constituting ~22% of PHC PI. Consumption in the Q3 range for this cluster also has an average frequency of 1 and makes up 5% of relative PHC at 12l/h/h PI. Q2 has a frequency of 6.5 and constitutes 3% of PHC at 8.73 while Q1 is 1% of PHC at ~5l/h/h with an average frequency of 17. This means that, on average, ~71% of the number of times in 24 hours that households in the EM cluster use water occurs in their Q1 range with ~21% occurring in Q2. On average, 56% of all households in this cluster record 0 water consumption at 04:00 (Fig.5).

Lastly, MP represents 23.45% of network households. It has a more than proportionate share of network consumption of 8,775m^3, some 28.25% of the annual network value, a monthly $\bar{x}$ of 731m^3 and σ of 98m^3 (Table 2). MP, as the name suggests, not only has the most temporally ubiquitous high quartile events but also the lowest frequency of Q1. Only 34% of all households in this cluster record 0 water consumption at 04:00. Although Q4 only constitutes 10% of PHC and only equates to 20l/h/h PI, households in this cluster record, on average, 4 instances of Q4 events throughout the day. ~36% of households reach Q4 at 11:00 and 19:00, ~34% at 10:00 and 12:00, ~31% at 09:00 and 18:00, 28% at 20:00, 25% at 13:00 and 23% at 08:00. Consumption in Q3 has an average daily frequency of 6 at ~18l/h/h, constituting 6% of relative PHC PI. Q2 has an average frequency of 6 and the highest proportion of PHC of all four clusters at ~16l/h/h (6% of PHC) PI. Q1for this cluster is at 8l/h/h, also the highest, constituting 2% of PHC PI (Fig.5).

4. Discussion

4.1 The implication for demand management:

The accurate and reliable modelling of water usage can help water utilities understand the complete patterns of household water use, aiding the improvement of demand forecasting and development of both proactive and responsive water conservation strategies. This approach also facilitates peak demand management, a key component of strategic DSM (Beal et al., 2016). The disaggregation of sub-minute smart meter data remains a very effective means of revealing consumption patterns and modelling water usage in microcomponent details. However, longitudinal observation (Moglia et al., 2018) of households' water use is necessary for the complete understanding of consumption patterns that underpin demand peak, as well as the effectiveness of intervention strategies such as targeted feedback. Incorporating a network-wide routine micro-component analysis via sub-minute data disaggregation is not currently readily feasible (Abu-Bakar et al., 2021a). The consumer segmentation and consumption pattern characterisation presented here can, however, play a pivotal role in allowing the reliable targeting of households according to their shared characteristics – by utilising medium resolution data from ordinary smart meters.

4.2 Cluster consistency and migration:

The model enables a longitudinal tracking of individual households' detailed consumption patterns, establishing the extent to which these patterns shift, how they underpin peak consumption and their influence on aggregate demand. Analysis suggests that, while cluster tenure (time households spend in a given cluster) is dynamic both temporally and spatially, there is a reliable degree of consistency in the extent to which households stay in or "switch" clusters. The number of months out of 12 that households stay in a cluster is hereby referred to as "consistency block". On average, 19% of network households stay within the same cluster throughout the year, 33% stayed in 11 out of 12 months, 46% 10 months, 58% 9 months, 70% 8 months, 83% 7 months and 94% are consistent for 6 out of 12 months. This gives the clusters a "tenure consistency factor" (TCF) of 13% – i.e. on average, tenure consistency increases by 13% with every decremental consistency block. Fig. 6 presents a breakdown of tenure consistency on a cluster level by revealing cluster representation at each consistency block. The significance of this to

targeted intervention is that utilities can tailor the frequency and duration of appropriate conservation feedback reliably to the TCF of individual clusters. For instance, as MP has the least 12/12 consistency, the duration of intervention targeted at this cluster should reflect that. Conversely, LM, being the most consistent cluster could be targeted with a longer duration of intervention.

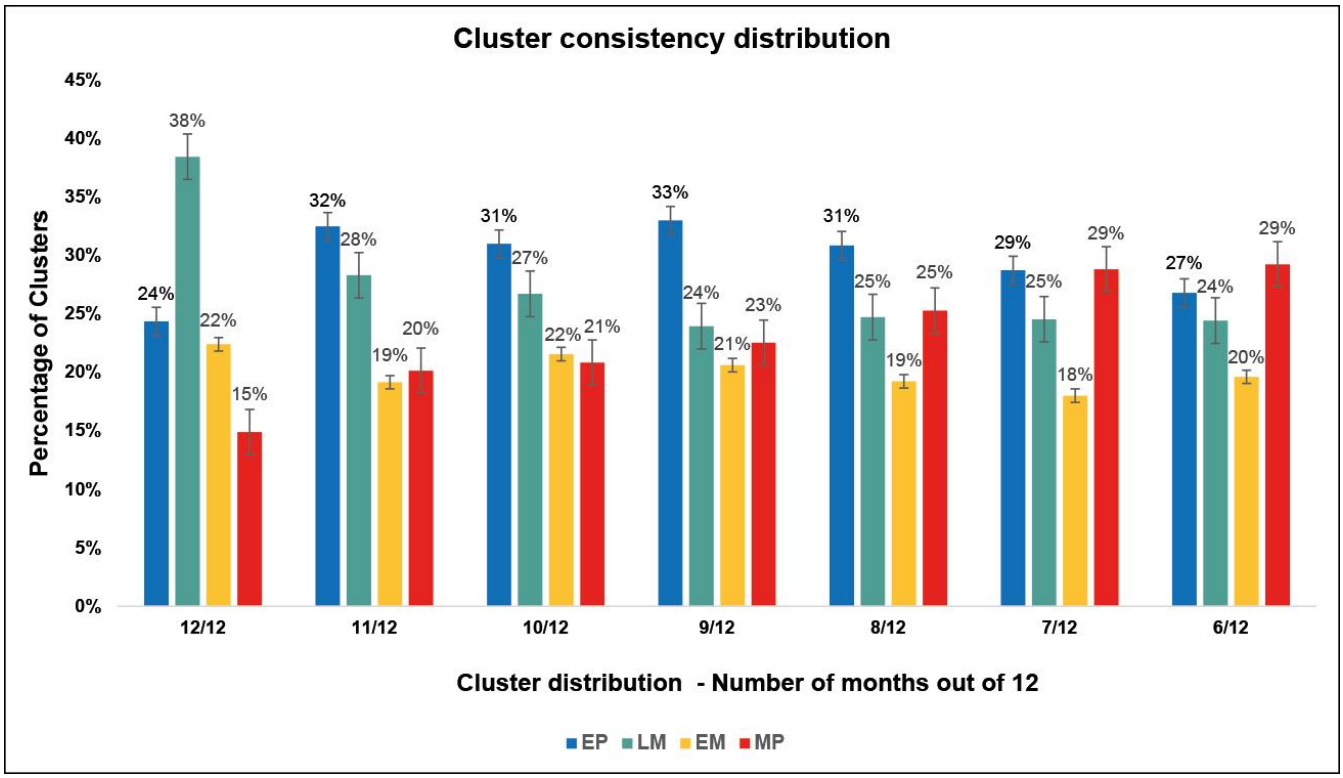


Fig. 6. Cluster consistency distribution.

Tenure consistency can also be ascertained by looking at the precise number of households that remain in their clusters from one month to another and their distributed migration to the other clusters. Fig. 7 presents the monthly cluster migration, revealing the proportionality of households' flow from one cluster in one month to another in another month. On average, LM retains a greater proportion of its households – ~73% – compared to ~70% for EM, ~69% for EP and ~64% for MP, giving households in LM the highest tenure consistency.

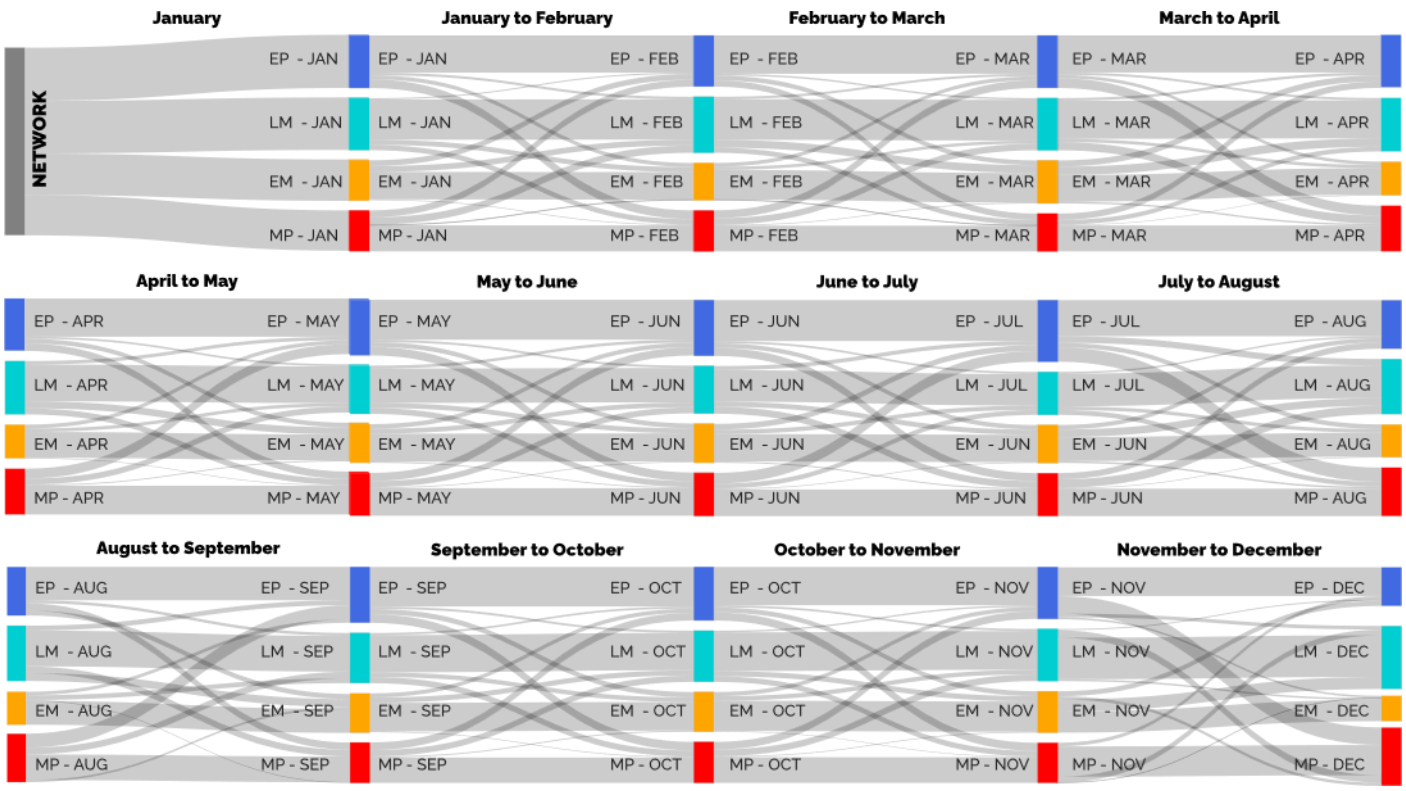


Fig. 7. Sankey diagram illustrating monthly cluster migration

4.2.1 Cluster correlation

The analysis of the relationship between the trends of pairs of clusters across three categories over time, namely (i) clusters' network representation, (ii) clusters' mean network consumption, (iii) clusters' mean PHC, revealed that EP/LM, EP/MP, LM/EM and EM/MP are correlated negatively while EP/EM and LM/MP are correlated positively (Fig. 7). Cluster correlation can serve as a useful tool for predicting cluster behaviour. For instance, planners can use the peak characteristics of one cluster as a predictor of the activities of another cluster, where they are positively correlated.

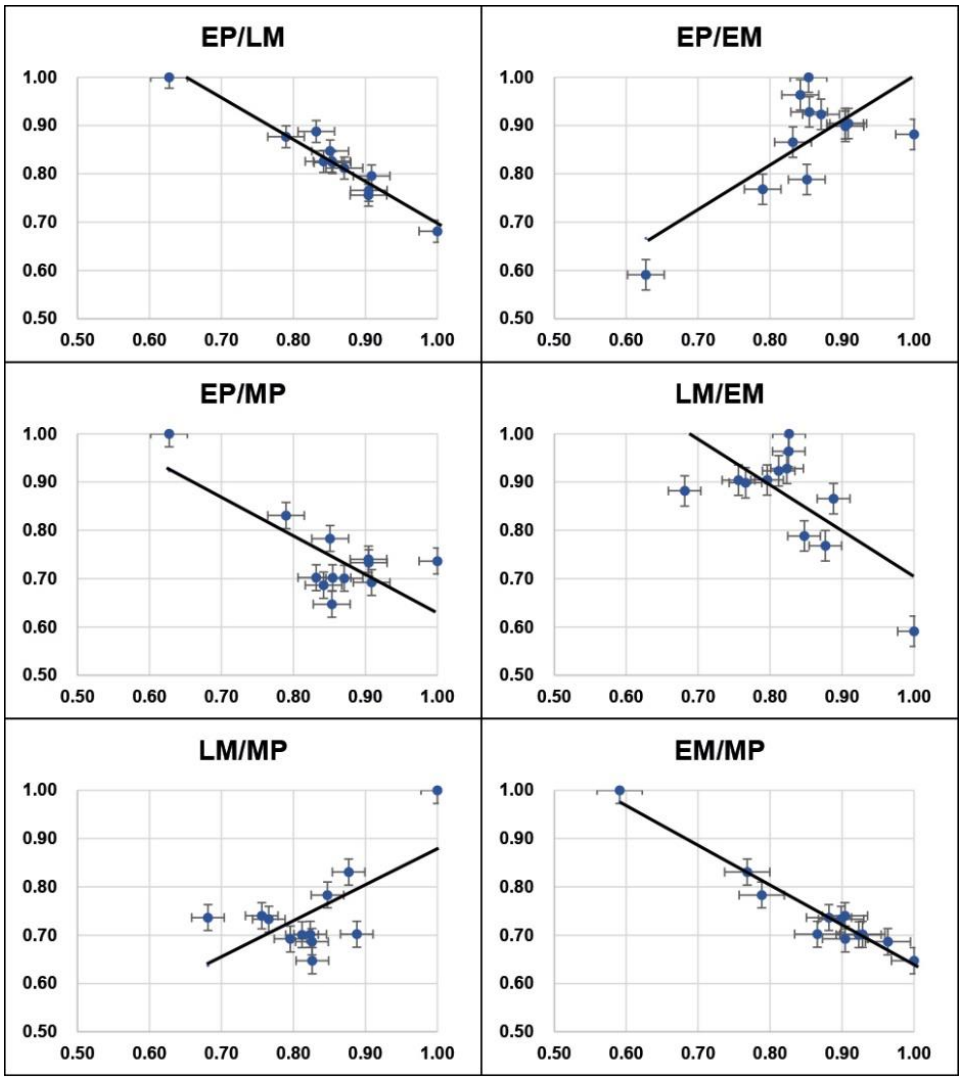


Fig. 8. Correlation between pairs of clusters.

4.3 Peak management and shifting using clusters' relationship with absolute demand

The segmentation and quartile classification of consumption patterns provide an avenue for peak shifting both as a DSM tool to foster consumer water conservation and for network load monitoring and shifting. Water planners would have the capability to predict peak events at multiple resolutions (hourly or daily) and to reduce the peak efficiently by applying peak-shifting DSM instruments to several households, with a significant impact on aggregate demand at peak. For instance, some of the successful peak demand "shaving" tools used in the energy industry, such as incentives to households for cutting back consumption during peak times, can be borrowed by the water sector to target particular end-users, e.g. heavy outdoor water users (Mayer and Smith, 2017). This could mark a paradigm shift in the water industry by enabling peak demand reduction without adversely impacting supply to customers. It could lead to water conservation, cost reduction and the reduction of pressure on the network water resources.

4.4 Internal leak management:

Water utilities are increasingly dedicating resources to internal leakage detection (ILD), which accounts for ~10% of network household water supply (Britton et al., 2013). Many network leakage detection systems rely on the analyses of minimum night flow (MNF) patterns by monitoring when hourly flow value surpasses a fixed threshold at the same time interval (Britton et al., 2013; Luciani et al., 2019). The quartile-based cluster characterisation presented here can be adapted as a means of setting a threshold for ILD. A Q1 threshold of ≤1% of PHC PI can be set for MNF as a trigger for ILD so that a segmentation based on households' Q1 events can be used to isolate households whose Q1 PI values successively exceed the normal threshold. Analysis of existing internal leakage data for 2019 reveals the MP cluster (whose MNF is twice that of the other clusters) to have a disproportionately high representation of households with confirmed internal leaks compared with the other clusters: 23% of network households are in this cluster while they make up 37% of network households with leaks (Fig. 9). This means, while only 1 in every 5 network households is an MP, 2 in every 5 MP households have a confirmed internal leak, compared with a 1:5 cluster/network households and 1:5 internal leak/cluster households ratio for the remaining 3 clusters.

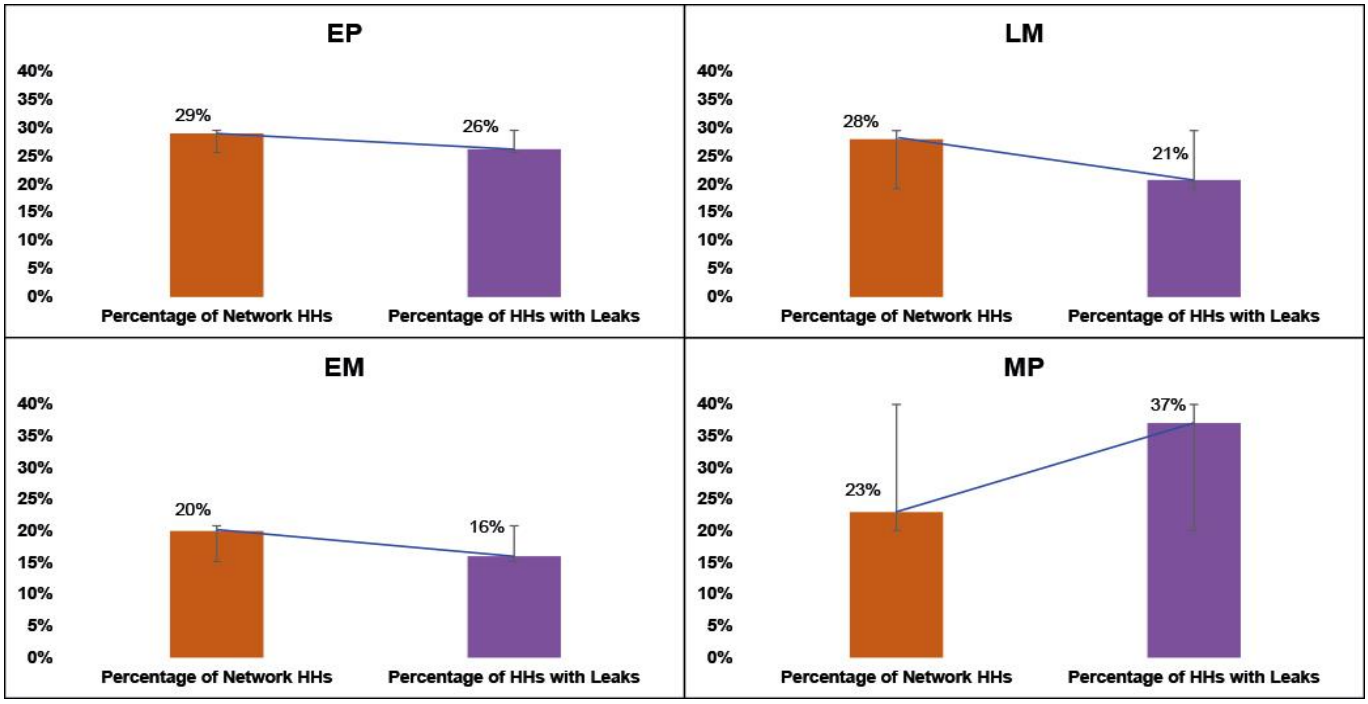


Fig. 9. The proportion of network households per cluster vs the proportion of cluster households with a confirmed internal leak.

4.5 Implications for demand forecasting:

Accurate water demand forecasting allows water utilities to plan, optimise and reduce the pressure on water resources, improve the design and management of water supplies and deliver a successful demand management strategy. Utilities routinely alternate forecast horizons and periodicities to achieve different DSM and operational outcomes (Donkor et al., 2014). In doing this they must take into account the complexities of demand components to better inform decision optimisation about water allocation. Short term forecasting, such as hourly, is integral to optimizing and managing system operations. The characterised hourly consumption patterns in this study (Fig. 10) provide a unique basis for short term forecasting using Brekke et al's (2002) "unit water demand analysis" method – forecasting for individual clusters after disaggregating consumption into household segments and then aggregating it.

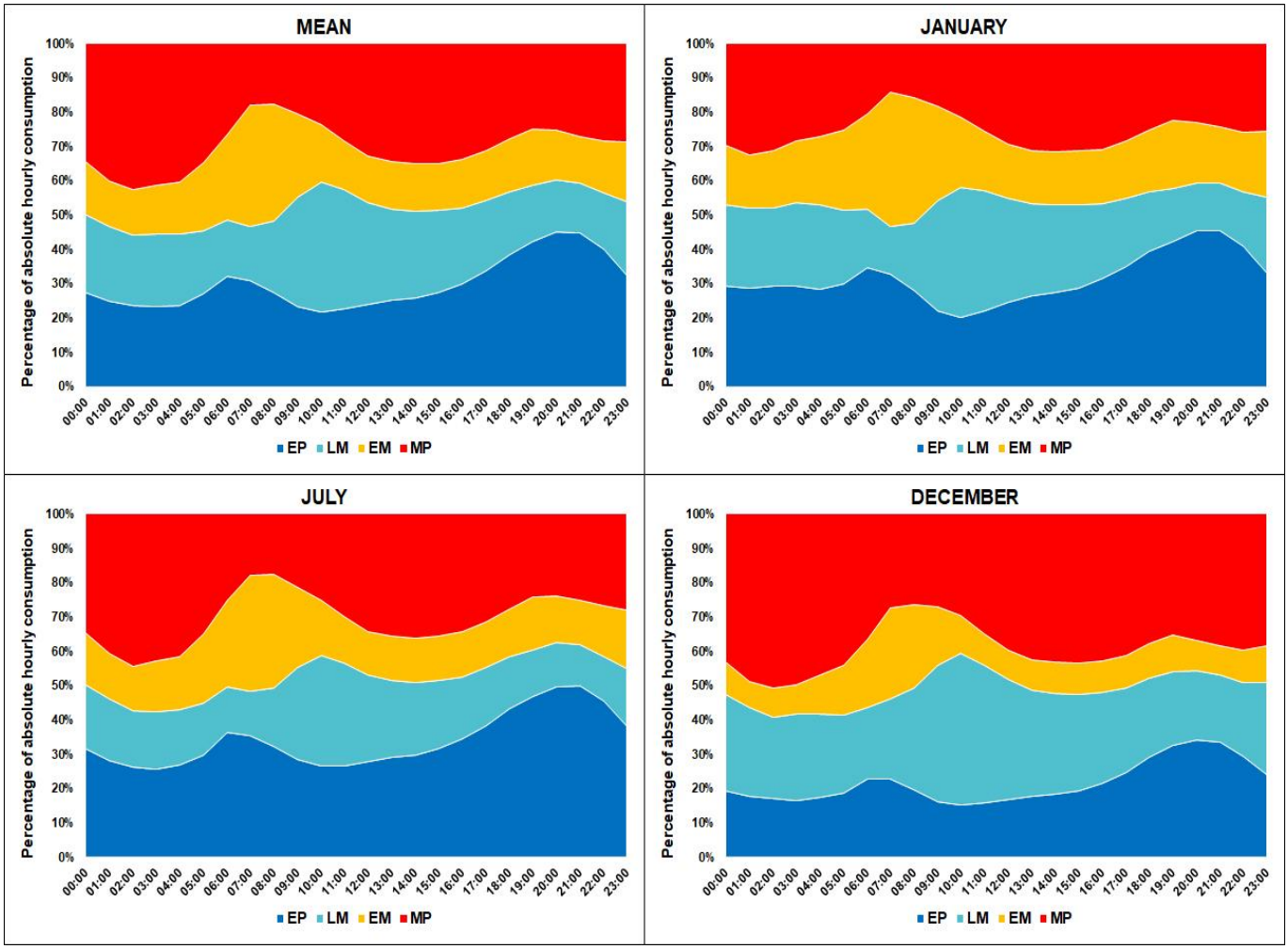


Fig. 10. Clusters' dominance of hourly demand

4.6 Seasonal variability:

Seasonal variability affects clusters differently, making some clusters more active and others less active. This study has revealed significant seasonal variability in January, July and December in several categories including hourly patterns, monthly averages and cluster tenure.

4.6.1 Seasonal hourly patterns

Fig. 10 compares clusters' mean hourly consumption patterns with patterns for January, July and December, revealing clusters' dominance of hourly demand across the network. On average, MP has the highest hourly network demand at 00:00- 05:00 and 12:00-16:00; EP dominates at 17:00 and 23:00; LM between 09:00 and 11:00 with EM only dominating at 07:00 and 08:00. In December, however, MP had the highest proportion of network demand for every hourly period (by over 50% in the Q1 ranges) except 08:00-11:00 which was dominated by LM. EM's proportions are significantly reduced (down to ~11-8% in the Q1-Q3 ranges). December patterns are consistent with the clusters' inherent behaviours with MP's increased dominance ostensibly being the result of the temporal ubiquity of household water consumption associated with Christmas.

In January, EM significantly increased dominance at 07:00-08:00 to ~40%, ostensibly denoting a return to typical routines and work after the holiday periods. LM maintains dominance of 09:00-11:00, making them the cluster least sensitive to seasonal variability. EP dominates in 13 out of 24 hours making them the most active in January.

4.6.2 Seasonal monthly averages

Temperatures of up to 39 °C were recorded across central and southern England (including the region of this study) in July of 2019, making it the hottest month on record (Kendon, 2020). This has been reflected in both global (network) consumption and cluster dynamics. Consequently, total consumption across the network in July was 11% higher than the monthly mean for the year, PHC was 8% higher, average max was 4% higher and min was 10% higher. Fig. 10 illustrates the network monthly averages, highlighting the month with the highest demand. Such seasonal consumption patterns translate into cluster characteristics and provide an indication as to the behaviours underpinning these seasonal

variabilities. While on average, EP, LM, EM and MP's respective share of network consumption is 31%, 23%, 18% and 28%, their respective shares of July consumption are 36%, 19%, 17% and 28%, equating to a monthly mean of 43m^3, 23m^3, 21m^3 and 33m^3. This means that EP alone is accountable for the increase in July demand. This increase is resonated across all categories including PHC, max, min and proportion of network households as illustrated in figs. 12 and 13.

December records indicate a significantly higher than average consumption in MP (37%) and LM (29%), with EM reducing to 12% and EP to 22%. However, LM's average max in December was 85m^3 compared to MP's 62m^3, while their mean for the month was respectively 31m^3 and 40m^3. EM and EP recorded slightly higher than average figures in January, with MP being 3% below average consumption while LM retained average values.

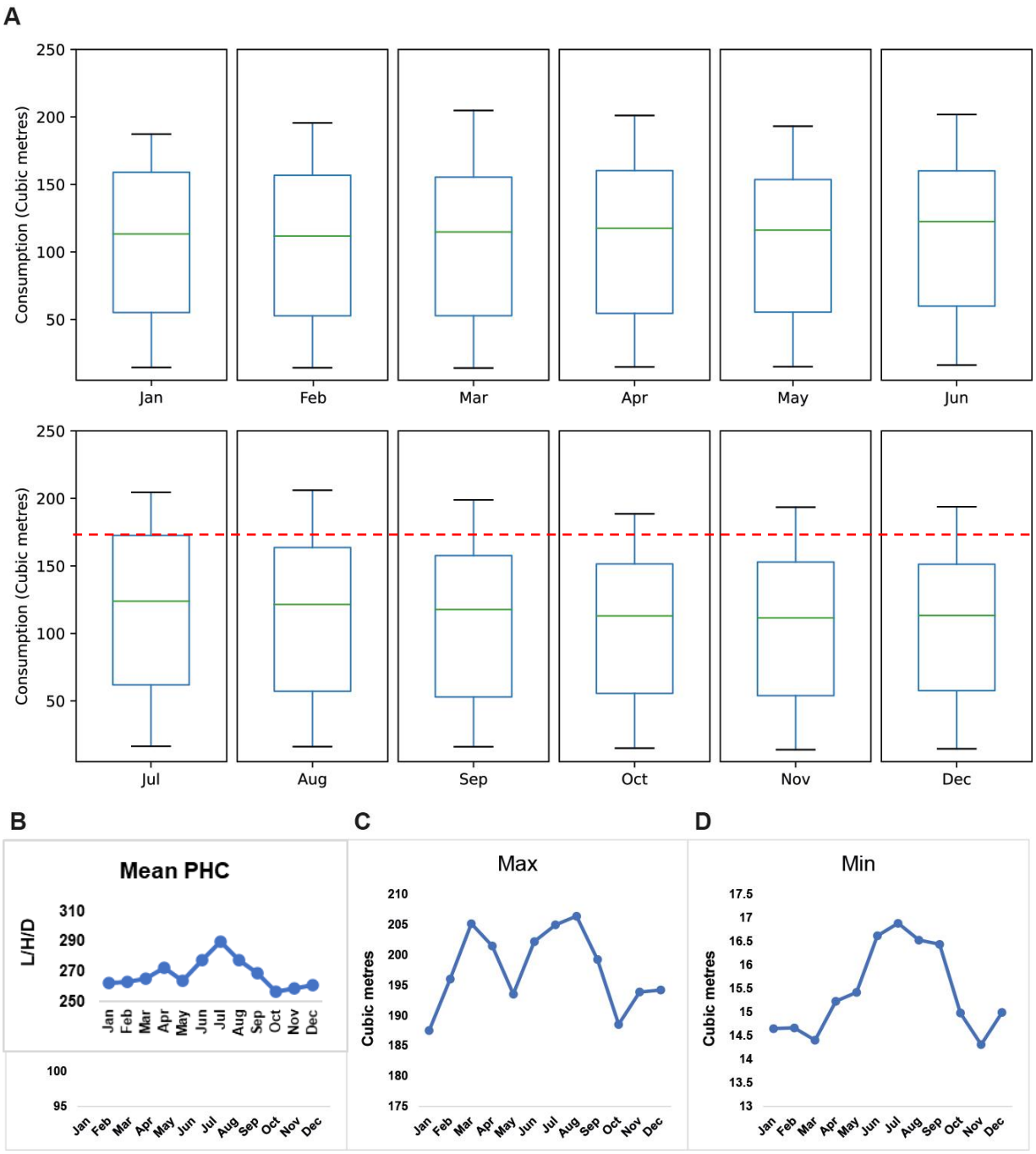


Fig. 11. A. Box plots describing Network monthly consumption variations (Value at the top of the whisker is the Maximum consumption; bottom of the whisker is the minimum value; top of the whisker is the maximum value; top bound of the box is the upper quartile value; bottom bound of the box is the lower quartile value; the line in the centre of the box is the mean (**$\bar{x}$**). B. Mean monthly PHCs in litres per household. C. Average monthly network maximum values in m^3. D. Average monthly network minimum values in m^3. Red dashed-line across highlighting highest consumption.

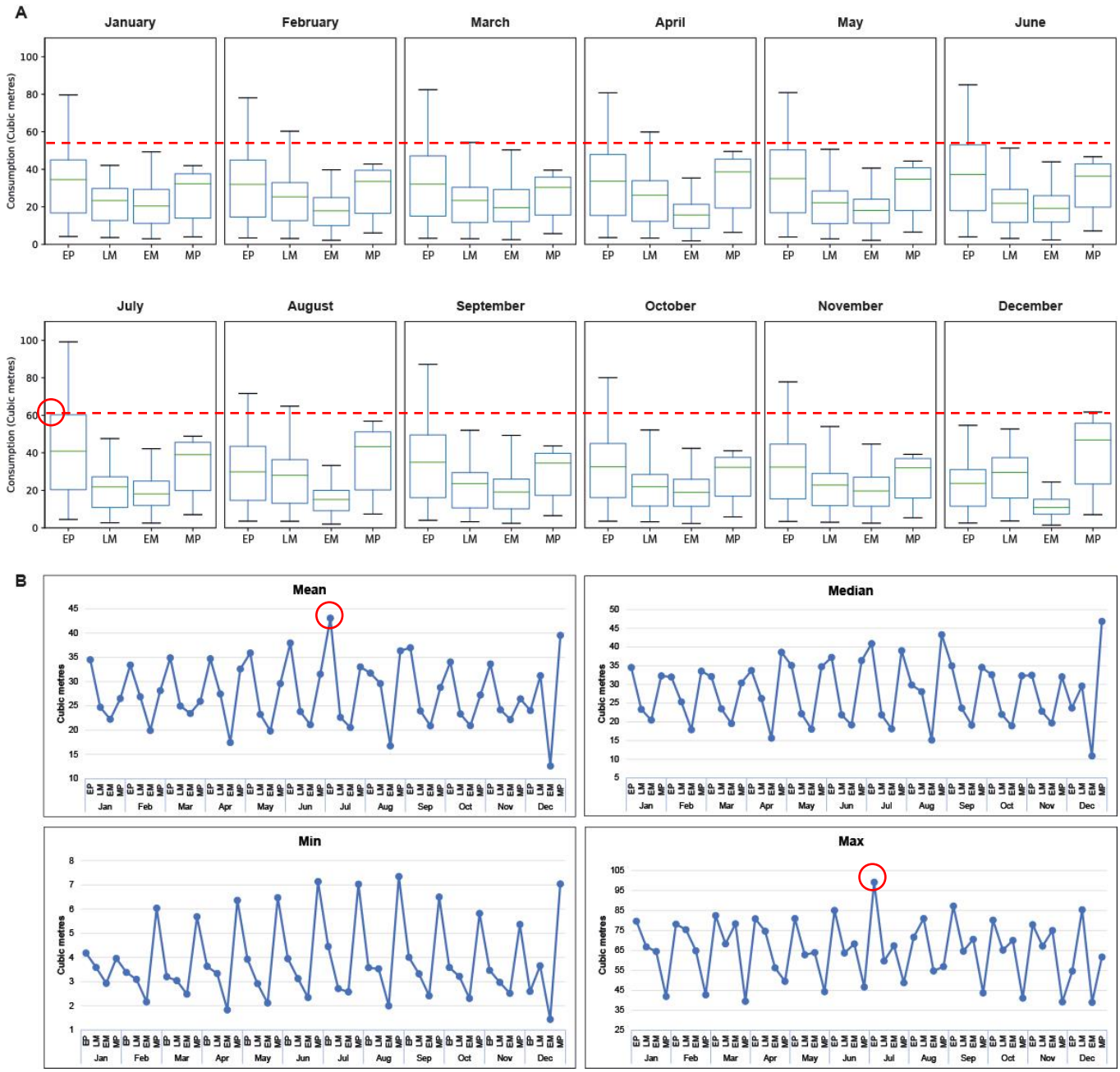


Fig. 12. A. Box plots describing clusters' monthly consumption variations (Value at the top of the whisker is the Maximum consumption; bottom of the whisker is the minimum value; top of the whisker is the maximum value; top bound of the box is the upper quartile value; bottom bound of the box is the lower quartile value; the line in the centre of the box is the mean ($\bar{x}$). B. Mean, median, minimum and maximum monthly consumption per cluster.

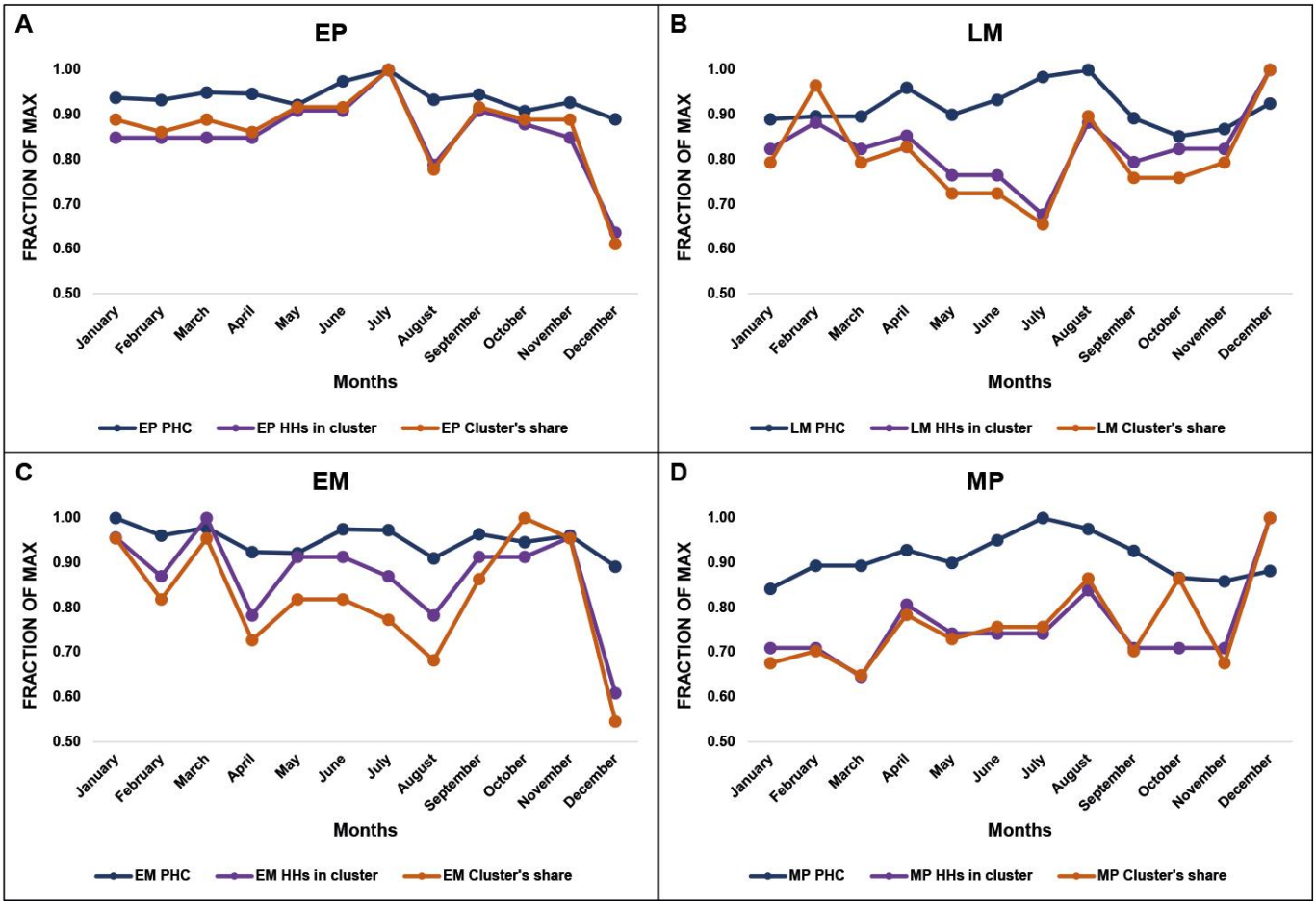


Fig. 13 Clusters' monthly trends in PHC, the proportion of households and share of network demand.

4.6.3 Seasonal cluster tenure

Seasonal factors are major determinants of cluster migration. Marked changes in clusters' proportion of network households attributable to seasonal variability have been observed in July and December. Households in EP grew by 10% in July while LM and EM decreased by 11% and 2% respectively. No change was observed with MP. December records revealed a 34% and 63% decline in EP and EM respectively, while MP and LM grew by 31% and 17%, respectively.

4.7 Geographical implications

Another significant implication of the findings is that it provides the basis for identifying regional consumption hotspots within the distribution network, aiding seasonal and regional load monitoring and measures to mitigate water stress, having a resultant positive impact on the environment and water resources. The data for this study were split into two adjacent geographical regions of similar sizes (Region 1 and Region 2). Fig.13 shows Region one to have LM as it's dominant cluster followed by EP, MP and EM in that order, whilst Region 2 is quite decisively dominated by EP, followed by MP, LM and

MP. This can be vital in the design of DSM instruments as well as demand forecasting since each cluster has a fundamentally different impact on peak and aggregate demand.

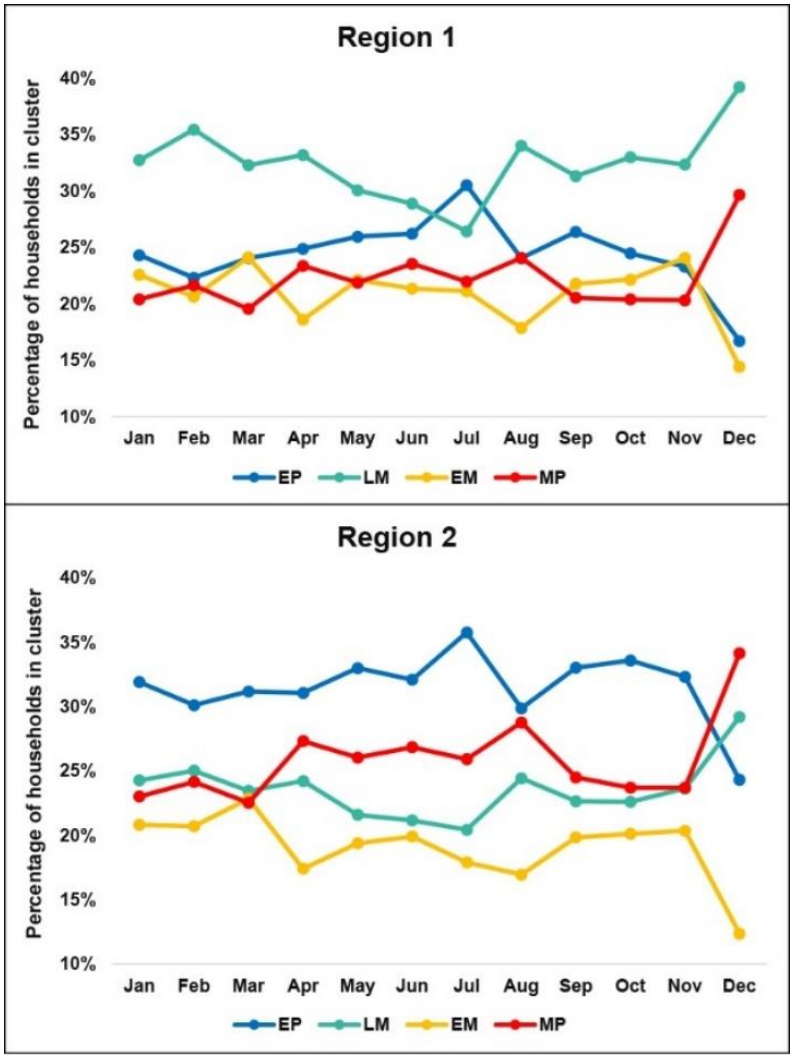


Fig. 14 Regional cluster dominance

5. Conclusion:

This study offers a compelling and practical alternative to micro-component analysis as a basis for household water DSM intervention. One of the significant drawbacks of current micro-component analysis methods is scalability and network-wide applicability. Methodologically, this study delivers both reliable and scalable approach to consumer segmentation with wide-ranging implications. Besides providing a basis for multi-horizon demand forecasting and the prediction of geographical and seasonal consumption hotspots, this work can be part of a dynamic smart meter DSM feedback system that

targets and delivers cluster-specific intervention to households on a regular basis, as opposed to the current one-size-fits-all approach adopted by the water sector.

The findings could have more profound consequences if future research is dedicated to determining the SD, SE and behavioural attributes of households associated with these clusters. That would help explain how the prevalence of a particular cluster in an area is influenced by some of these attributes thereby providing a degree of richness and diversity to any targeted intervention. For example, as two regions in this study have different dominant clusters, it would be interesting to determine whether there are any significant differences in their SD/SE attributes underpinning the trend.

Disaggregating consumption into the four quartiles allows planners to predict or infer the microcompenent events underpinning consumption within the quartiles. For instance, It could be inferred that EM's single daily instance of Q4 in the morning is underpinned by a relatively intensive microcomponent event such as shower, suggestive of working households accomplishing their most significant water consumption routine before leaving for work. Studies could be dedicated to understanding the underlying microcomponents epitomising the more temporally ubiquitous quartiles, such as Q2, to help design the appropriate intervention strategy.

Utilities would find it beneficial to not only be able to determine why households belong in a particular cluster but also to be able to predict when households are likely to migrate from one cluster to the other and their likely duration in that cluster. The results, confirm that cluster patterns are affected by exogenous factors such as seasonal variability. Future research could focus on the longitudinal implications of endogenous factors such as a change in occupancy, household setup, dwelling type, demographics or income.

Finally, the rollout of smart meters in England is still in its trial stages with the vast majority of households still on ordinary meters, making the wide-applicability of this study unlikely in the immediate term as it relies on medium resolution data. A segmentation metric-based framework can be developed in the future using a combination of the results of this study and a comprehensive SD/SE survey. The framework will allow the classification of both ordinary metered and unmetered households based on the SD/SE metrics.

Author contributions

The final manuscript has been approved by all authors. Halidu Abu-Bakar undertook the research and compiled the manuscript with inputs and guidance from Leon Williams, Stephen Hallett. All authors discussed the results and contributed to the final manuscript.

Declaration of competing interest

The authors declare that they have no known competing
financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC Ref: EP/N509450/1) and Anglian water plc. The authors are grateful for their support.

Data Statement

The network data have not been made publicly available by the water company. However, the data can be shared upon request.

References:

Abu-Bakar, H., Williams, L., Hallett, S.H., 2021a. A review of household water demand management and consumption measurement. J. Clean. Prod. 292, 125872. https://doi.org/10.1016/j.jclepro.2021.125872

Abu-Bakar, H., Williams, L., Hallett, S.H., 2021b. Quantifying the impact of the COVID-19 lockdown on household water consumption patterns in England. npj Clean Water 4, 1–9. https://doi.org/10.1038/s41545-021-00103-8

Aggarwal, C.C., Reddy, C.K., 2014. DATA CLUSTERING Algorithms and Applications. Taylor & Francis Group.

Avni, N., Fishbain, B., Shamir, U., 2015. Water consumption patterns as a basis for water demand modeling. Water Resour. Res. 51, 8165–8181. https://doi.org/10.1002/2014WR016662

Balacco, G., Carbonara, A., Gioia, A., Iacobellis, V., Piccinni, A.F., 2018. Investigation of Peak Water Consumption Variability at Local Scale in Puglia (Southern Italy). Proceedings 2, 674. https://doi.org/10.3390/proceedings2110674

Beal, C.D., Gurung, T.R., Stewart, R.A., 2016. Demand-side management for supply-side efficiency: Modeling tailored strategies for reducing peak residential water demand. Sustain. Prod. Consum. 6, 1–11. https://doi.org/10.1016/j.spc.2015.11.005

Beal, C.D., Stewart, R.A., 2014. Identifying Residential Water End Uses Underpinning Peak Day and Peak Hour Demand. J. Water Resour. Plan. Manag. 140, 04014008. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000357

Britton, T.C., Stewart, R.A., O'Halloran, K.R., 2013. Smart metering: enabler for rapid and effective post meter leakage identification and water loss management. J. Clean. Prod. 54, 166–176. https://doi.org/10.1016/j.jclepro.2013.05.018

Candelieri, A., 2017. Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection. Water 9, 224. https://doi.org/10.3390/w9030224

Cardell-Oliver, R., 2013. Water use signature patterns for analyzing household consumption using medium resolution meter data. Water Resour. Res. 49, 8589–8599. https://doi.org/10.1002/2013WR014458

Cardell-Oliver, R., Wang, J., Gigney, H., 2016. Smart Meter Analytics to Pinpoint Opportunities for Reducing Household Water Use. J. Water Resour. Plan. Manag. 142, 04016007. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000634

Cominola, A., Giuliani, M., Castelletti, A., Rosenberg, D.E., Abdallah, A.M., 2018a. Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. Environ. Model. Softw. 102, 199–212. https://doi.org/10.1016/J.ENVSOFT.2017.11.022

Cominola, A., Nguyen, K., Giuliani, M., Stewart, R.A., Maier, H.R., Castelletti, A., 2019. Data Mining to Uncover Heterogeneous Water Use Behaviors From Smart Meter Data. Water Resour. Res. 55, 9315–9333. https://doi.org/10.1029/2019WR024897

Cominola, A., Spang, E.S., Giuliani, M., Castelletti, A., Lund, J.R., Loge, F.J., 2018b. Segmentation analysis of residential water-electricity demand for customized demand-side management programs. J. Clean. Prod. 172, 1607–1619. https://doi.org/10.1016/j.jclepro.2017.10.203

Ding, C., He, X., 2004. K -means clustering via principal component analysis, in: Twenty-First International Conference on Machine Learning - ICML '04. ACM Press, New York, New York, USA, p. 29. https://doi.org/10.1145/1015330.1015408

Donkor, E.A., Mazzuchi, T.A., Soyer, R., Alan Roberson, J., 2014. Urban Water Demand Forecasting: Review of Methods and Models. J. Water Resour. Plan. Manag. 140, 146–159. https://doi.org/10.1061/(asce)wr.1943-5452.0000314

Duerr, I., Merrill, H.R., Wang, C., Bai, R., Boyer, M., Dukes, M.D., Bliznyuk, N., 2018. Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A Comparative study. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2018.01.002

Garcia, D., Puig, V., Quevedo, J., Cugueró, M.A., 2017. Big Data Analytics and Knowledge Discovery Applied to Automatic Meter Readers, in: Advances in Industrial Control. Springer International Publishing, pp. 401–423. https://doi.org/10.1007/978-3-319-50751-4_20

Giurco, D., Carrard, N., McFallan, S., M, N., Inman, M., Thornton, N., White, S., 2008. RESIDENTIAL END-USE MEASUREMENT GUIDEBOOK A GUIDE TO STUDY DESIGN, SAMPLING AND TECHNOLOGY.

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, Data Mining: Concepts and Techniques. Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. 31, 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data, Prentice Hall. Prentice Hall Englewood Cliffs, New Jersey 07632, New Jersey.

Jolliffe, I.T., 2010. Analysis , Second Edition Springer Series in Statistics, Second Edi. ed, Design. Springer, New York.

Kendon, M., 2020. Exceptionally hot day, 31 July 2020 - The Met Office.

Kotsiantis, S.B., Kanellopoulos, D., Pintelas, P.E., 2006. Data Preprocessing for Supervised Leaning. Int. J. Comput. Sci. 1, 111–117. https://doi.org/doi.org/10.5281/zenodo.1082415

Lavee, D., Danieli, Y., Beniad, G., Shvartzman, T., Ash, T., 2013. Examining the effectiveness of residential water demand-side management policies in Israel. Water Policy 15, 585–597. https://doi.org/10.2166/wp.2013.146

Luciani, C., Casellato, F., Alvisi, S., Franchini, M., 2019. Green Smart Technology for Water (GST4Water): Water loss identification at user level by using smart metering systems. Water (Switzerland) 11. https://doi.org/10.3390/w11030405

Macqueen, J., 1967. SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS, in: Proceedings Ofthe Fifth Berkeley Symposium on Mathematical Statistics and Probability , Oakland, CA, USA. pp. 281–297.

Mayer, P., Smith, R., 2017. Peak Day Water Demand Management Study.

McDonald, A., Butler, D., Ridgewell, C., 2011. Water demand: estimation, forecasting and management, in: Water Distribution Systems. Thomas Telford Ltd, pp. 49–71. https://doi.org/10.1680/wds.41127.049

McKenna, S.A., Fusco, F., Eck, B.J., 2014. Water Demand Pattern Classification from Smart Meter Data. Procedia Eng. 70, 1121–1130. https://doi.org/10.1016/j.proeng.2014.02.124

Moglia, M., Cook, S., Tapsuwan, S., 2018. Promoting Water Conservation: Where to from here? Water 10, 1510. https://doi.org/10.3390/w10111510

Nguyen, K.A., Stewart, R.A., Zhang, H., Sahin, O., Siriwardene, N., 2018. Re-engineering traditional urban water management practices with smart metering and informatics. Environ. Model. Softw. 101, 256–267.

https://doi.org/10.1016/j.envsoft.2017.12.015

Orr, P., Papadopoulou, L., Twigger-Ross, C., 2018. Water Efficiency and Behaviour Change Rapid Evidence Assessment.

Owen, L., Bramley, H., Tocock, J., 2009. Public Understanding of Sustainable Water Use in the Home: A research report completed for the Department for Environment , Food and Rural Affairs (DEFRA).

Padulano, R., Del Giudice, G., 2018. A Mixed Strategy Based on Self-Organizing Map for Water Demand Pattern Profiling of Large-Size Smart Water Grid Data. Water Resour. Manag. 32, 3671–3685. https://doi.org/10.1007/s11269-018-2012-7

Parker, J.M., Wilby, R.L., 2013. Quantifying Household Water Demand: A Review of Theory and Practice in the UK. Water Resour. Manag. 27, 981–1011. https://doi.org/10.1007/s11269-012-0190-2

Ponce Romero, J., Hallett, S., Jude, S., 2017. Leveraging Big Data Tools and Technologies: Addressing the Challenges of the Water Quality Sector. Sustainability 9, 2160. https://doi.org/10.3390/su9122160

Priyadarshini, K.N., Sivashankari, V., Shekhar, S., Balasubramani, K., 2019. Comparison and Evaluation of Dimensionality Reduction Techniques for Hyperspectral Data Analysis. Proceedings 24, 6. https://doi.org/10.3390/IECG2019-06209

Rahim, M.S., Nguyen, K.A., Stewart, R.A., Giurco, D., Blumenstein, M., 2020. Machine Learning and Data Analytic Techniques in Digital Water Metering: A Review. Water 12, 294. https://doi.org/10.3390/w12010294

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salkind, N., 2010. Encyclopedia of Research Design. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States. https://doi.org/10.4135/9781412961288

Sønderlund, A.L., Smith, J.R., Hutton, C.J., Kapelan, Z., Savic, D., 2016. Effectiveness of Smart Meter-Based Consumption Feedback in Curbing Household Water Use: Knowns and Unknowns. J. Water Resour. Plan. Manag. 142, 04016060. https://doi.org/10.1061/(asce)wr.1943-5452.0000703

UKWIR, 2015. WRMP19 METHODS – HOUSEHOLD CONSUMPTION FORECASTING: GUIDANCE MANUAL.

Van Der Maaten, L., Courville, A., Fergus, R., Manning, C., 2014. Accelerating t-SNE using Tree-Based Algorithms, Journal of Machine Learning Research.

Van Der Maaten, L., Hinton, G., 2008. Visualizing Data using t-SNE, Journal of Machine Learning Research.

Willis, R.M., Stewart, R.A., Panuwatwanich, K., Williams, P.R., Hollingsworth, A.L., 2011. Quantifying the influence of environmental and water conservation attitudes on household end use water consumption. J. Environ. Manage. 92, 1996–2009. https://doi.org/10.1016/j.jenvman.2011.03.023

Yuan, C., Yang, H., 2019. Research on K-Value Selection Method of K-Means Clustering Algorithm. J 2, 226–235. https://doi.org/10.3390/j2020016



https://dspace.org/

An empirical water consumer segmentation and the characterisation of consumption patterns underpinning demand peaks

Abu-Bakar, Halidu; Williams, Leon; Hallett, Stephen H.

2021-07-18
Elsevier
https://doi.org/10.1016/j.resconrec.2021.105792
Abu-Bakar H, Williams L, Hallett SH. (2021) An empirical water consumer segmentation and the characterisation of consumption patterns underpinning demand peaks. Resources, Conservation and Recycling, Volume 174, November 2021, Article number 105792

Downloaded from a DSpace repository. https://dspace.org/