npj | Clean Water

www.nature.com/npjcleanwater

BRIEF COMMUNICATION OPEN

Check for updates

Quantifying the impact of the COVID-19 lockdown on household water consumption patterns in England

Halidu Abu-Bakar[1], Leon Williams[1]✉ and Stephen H. Hallett[2]

The COVID-19 lockdown has instigated significant changes in household behaviours across a variety of categories including water consumption, which in the south and east regions of England is at an all-time high. We analysed water consumption data from 11,528 households over 20 weeks from January 2020, revealing clusters of households with distinctive temporal patterns. We present a data-driven household water consumer segmentation characterising households' unique consumption patterns and we demonstrate how the understanding of the impact of these patterns of behaviour on network demand during the COVID-19 pandemic lockdown can improve the accuracy of demand forecasting. Our results highlight those groupings with the highest and lowest impact on water demand across the network, revealing a significant quantifiable change in water consumption patterns during the COVID-19 lockdown period. The implications of the study to urban water demand forecasting strategies are discussed, along with proposed future research directions.

npj Clean Water (2021)4:13 ; https://doi.org/10.1038/s41545-021-00103-8

INTRODUCTION

The ongoing COVID-19 pandemic had its first confirmed case in the United Kingdom in late January 2020, but transmission increased rapidly leading to the government imposing a lockdown on the whole population, banning all "non-essential" travel and contact with people outside one's home on 23 March 2020.[1] Globally, the lockdown has caused households to change their typical consumption behaviours drastically across a variety of major categories, resulting in an initial sharp increase in spending, especially in essentials and food items.[2] Studies dedicated to the impact of COVID-19 on water consumption focused on aggregate demand and general demand peaks. For example, in Germany, a significant shift in aggregate demand peak was observed from 07:10 pre-lockdown to 09:40 during lockdown.[3] In a Waterwise[4] report, certain regions in the UK saw a 35% increase in peak daily consumption during the lockdown. In Brazil, analysis of data from 26 days pre-lockdown and 26 during lockdown has revealed an 11% increase in household water consumption attributable to the lockdown.[5] Although this rise can be generally attributed to an increase in diurnal consumption owing to people remaining at home, increase in preventive behaviours such as hand-washing[1] also became contributory factors.

Household water demand in England and Wales is already at an all-time high, constituting 55% of the 32 Cubic Gigametres per year (Gm^3/yr) total UK household water consumption footprint, with southeast England having the highest per capita consumption (PCC) and being already declared as severely water-stressed.[6–8] The impacts of the extended time people stayed at home under the lockdown and the ensuing changes in behaviour arising from this have been an increase in household water demand, exacerbating existing pressure on network water supply.

Water utility companies are increasingly searching for ways to understand the full nature of household water use, how to improve network demand forecasting and achieve effective water efficiency interventions. By presenting a data-driven detailed characterisation of household clusters, including their unique patterns, we have demonstrated how the understanding of the impact of these unique patterns of behaviour on network demand can help in the design of demand forecasting and intervention that targets households on the basis of their shared cluster characteristics. Many demand strategies have relied on existing socioeconomic (SE) and sociodemographic (SD) household variables (e.g., ACORN[9]) and self-reported behaviours through surveys and water use diaries.[10,11] Our work not only significantly enhances the precision of forecasting and intervention when enriched with SE and SD variables, but also provides a scalable framework for the inclusion of ordinary-metered and unmeasured households that share SE/SD characteristics peculiar to particular clusters.

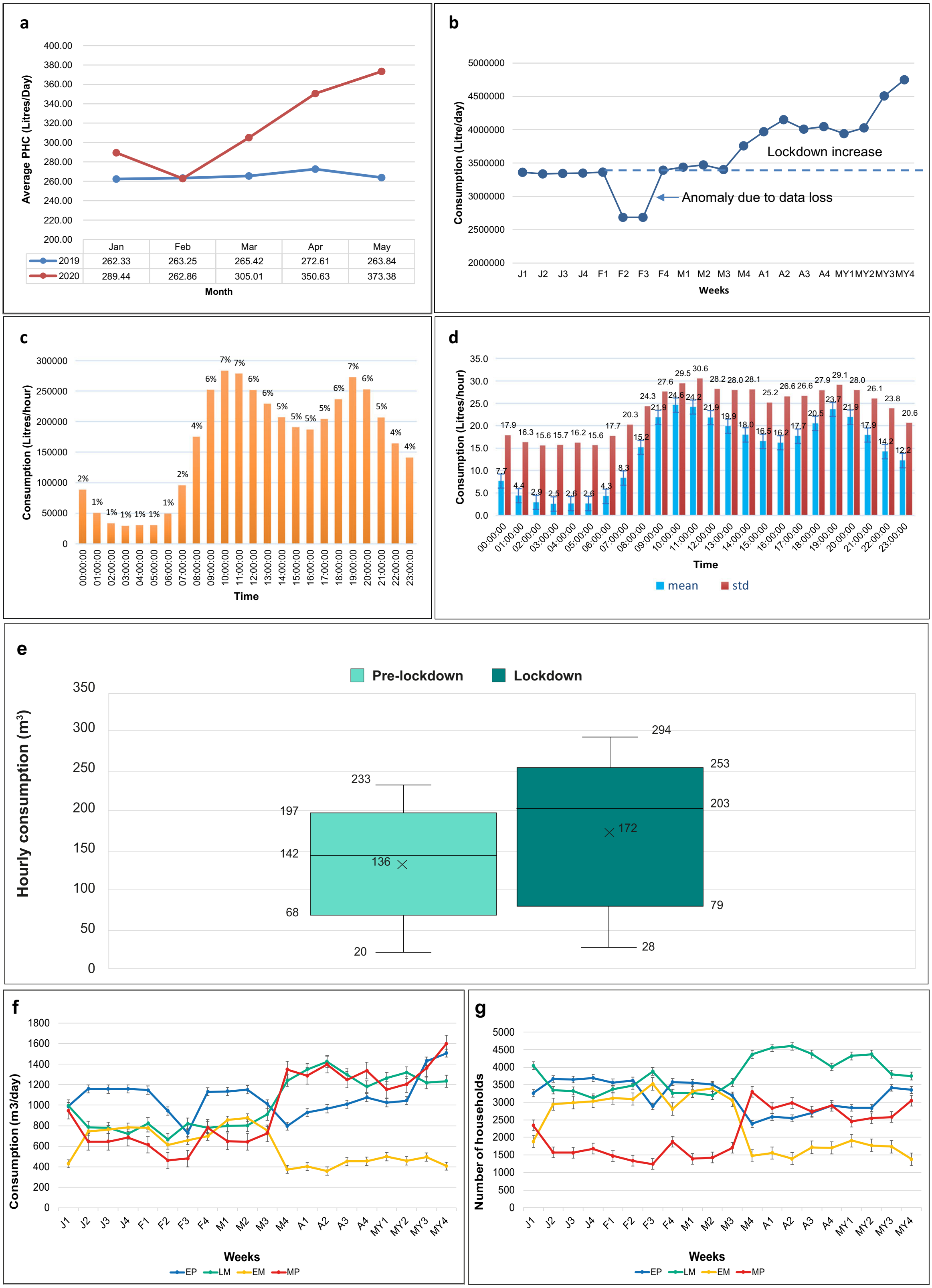
We analysed the weekly water consumption data, at an hourly resolution, for January to May 2020 of 11,528 smart-metered households. We then classified the households according to their temporal patterns of consumption, highlighting their unique characteristics and their respective shares of relative and absolute consumption before and during the COVID-19 lockdown.

All households in the study are from a single water provider, collected across two geographical areas 50 miles apart consisting of 24 District Metering Areas (DMAs). As the aim of this study was to quantify the impact of the Covid-19 lockdown on aggregate water demand while highlighting household clusters' underpinning temporal demand patterns, only anonymised smart meter data was utilised. Data on SD/SE or occupancy variables of the participating households were not available.

Overall temporal water consumption patterns

The analysis revealed an average consumption of 3256 cubic metres per day (m^3/d) for the 11,528 households across the network for the period before the COVID-19 lockdown, equating to a per household consumption (PHC) of 284 litres per day (l/h/d), as per the UK average.[12] Consumption remained even between the first week of January (J1) and the first week of February (F1) averaging 350 m^3/d (291 l/h/d), followed by a 20% decline in

[1]Cranfield Centre for Competitive Creative Design; School of Water, Energy and Environment Cranfield University, Bedford, UK. [2]Centre for Environmental and Agricultural Informatics, Cranfield University, Bedford, UK. ✉email: l.williams@cranfield.ac.uk



February week 2–3 (F2–F3), before returning to average values in February week 4 (F4) to March week 3 (M3) as in Fig. 1b. A sharp increase was recorded in March week 4 (M4), the week of the COVID-19 lockdown, to 3756 m^3/d (326 l/h/d), a rise of 10% on the previous week, reaching 4747 m^3/d (411 l/h/d) in May week 4 (MY4), some 46% above pre-lockdown average. The cause of the 20% drop in consumption in the second week of February 2020 remains unknown. The water utility did report the loss of four days of data in that period owing to equipment power outage. The absence of any other plausible cause is suggestive that this may

Fig. 1 Households' consumption patterns and trends before and during the COVID-19 lockdown in the UK. **a** Differences in per household consumption (PHC) for January–May 2019 and 2020. **b** Weekly average 24-hour consumption for all households–January week 1 (J1) to May week 4 (MY4)—showing normal consumption trend, anomaly due to data loss and increase in consumption during lockdown period. **c** Hourly consumption patterns, showing households' average proportion of hourly consumption to their daily average. **d** Households' hourly mean and standard deviation consumption in litres. **e** Boxplots illustrating the comparison between pre-lockdown and lockdown total consumption in cubic metres (m^3). Value at the top of whisker is the maximum consumption; bottom of whisker is the minimum consumption; 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 median and the *x* in the centre of the box is the mean ($\bar{x}$). **f** Weekly cluster consumption trend showing how much each of the four clusters consumes per week in m$^{3.}$ The error bars indicate standard deviation (σ). **g** Weekly number of households per cluster. The error bars indicate standard deviation (σ).

have resulted from "Storm Ciara",[13] which brought heavy rain and very strong winds to the region on 9 February 2020, causing widespread power issues.

Comparison between this study and similar data from the previous year, January–May 2019, for the same households revealed similar patterns of consumption and cluster behaviours. However, analysis of the data revealed a respective rise in PHC across the network of 13%, 22%, and 29% in March, April and May 2020 (Fig. 1a).

To examine the temporal (hourly) consumption patterns, four quartiles (Q1–Q4) were assigned to the values between the lowest and highest consumption range, revealing a consistent 24-hour pattern throughout the period irrespective of the volume of consumption (Fig. 2a–d). Households Q1 represents 1–2% (per hour) of daily consumption and occurs invariably between 00:00 and 06:00. Q2, representing 3–4% of daily consumption, occurs principally between 14:00 and 15:00 and Q3, 5–6%, occurring at different times, particularly 12:00–13:00 and 21:00. The Q4 (peak) occurs at 9:00–11:00 and 19:00–20:00. The daily mean network water demand was 27% higher during lockdown than pre-lockdown, median 43% higher and Q4 26% higher. Figure 1e presents a comparison of consumption before and during the lockdown.

According to our findings, households' proportion of total hourly water demand depends upon the clusters they belong to (Fig. 2g, h), although the ratio of their hourly consumption to their daily demand is largely consistent irrespective of the time of year or volumes consumed (Fig. 2a–d).

Household cluster characterisation before and after lockdown

The results reveal four distinct clusters of household water consumers characterised by unique diurnal and night-time consumption patterns. The clusters are named Evening Peak (EP), Late Morning (LM), Early morning (EM), and multiple peak (MP).

EP

Households in EP typically use ~6% of their daily consumption between 07:00 and 08:00 but their most significant consumption occurs between 19:00 and 20:00, which invariably constitutes ~10% of their daily demand in just 1 hour. Their Q2 constitutes ~4% of their relative daily consumption per hour and occurs between 09:00 and 16:00, with Q1, occurring between 00:00 and 05:00, representing ~1–2% (Fig. 2a). During the pre-lockdown weeks, this cluster constituted ~30% of the households across the network and has been responsible for over 50% (~76 m^3/hr) of the relative consumption between 19:00 and 22:00 (Fig. 2g) and ~33% (1065 m^3/d) of the total daily consumption, with a mean ($\bar{x}$) of 39 m^3/hr, standard deviation (σ) of 25 m^3/hr and maximum (max) of ~92 m^3/hr (Fig. 2e).

During the lockdown, the percentage of households in EP dropped to a network average of 25% and their dominance of consumption between 19:00 and 22:00 decreased to an average 45%, being ~135 m^3/hr (see Fig. 2h) along with their share of the total daily consumption, which also decreased to 26% (~1087 m^3/d). Lockdown hourly mean, standard deviation and maximum for EP were, respectively, $\bar{x}$ 63 m^3/hr, σ 44 m^3/hr and max 165 m^3/hr (Fig. 2f).

LM

LM describes households whose peak (Q4) occurs typically at ~10:00, representing ~11–12% of their relative daily water consumption in just 1 hour, with their next significant water use activities (~5% of daily consumption/hour) occurring at 19:00. Q2 for this cluster constitutes ~4% of their relative daily consumption per hour and occurs between 14:00 and 17:00, with a Q1 being identical to EP and EM (Fig. 2b). On average, this cluster has the highest relative consumption between 10:00 and 12:00, constituting 38% (~63 m^3/hr) pre-lockdown (Fig. 2g.), was represented by ~30% of households and had a 25% (808 m^3/d) share of the total daily consumption, with $\bar{x}$ 28 m^3/hr, σ 19 m^3/hr and max 76 m^3/hr (Fig. 2e).

The percentage of households in LM increased to an average of 37% across the network during the lockdown weeks but their consumption between 10:00 and 12:00 remained at an average of 38%–~74 m^3/hr (see Fig. 2h). Their share of the total daily consumption increased to ~31% (~1281 m^3/d). Lockdown being respectively $\bar{x} = 51$ m^3/hr, $\sigma = 36$ m^3/hr and max = of 134 m^3/hr (Fig. 2f).

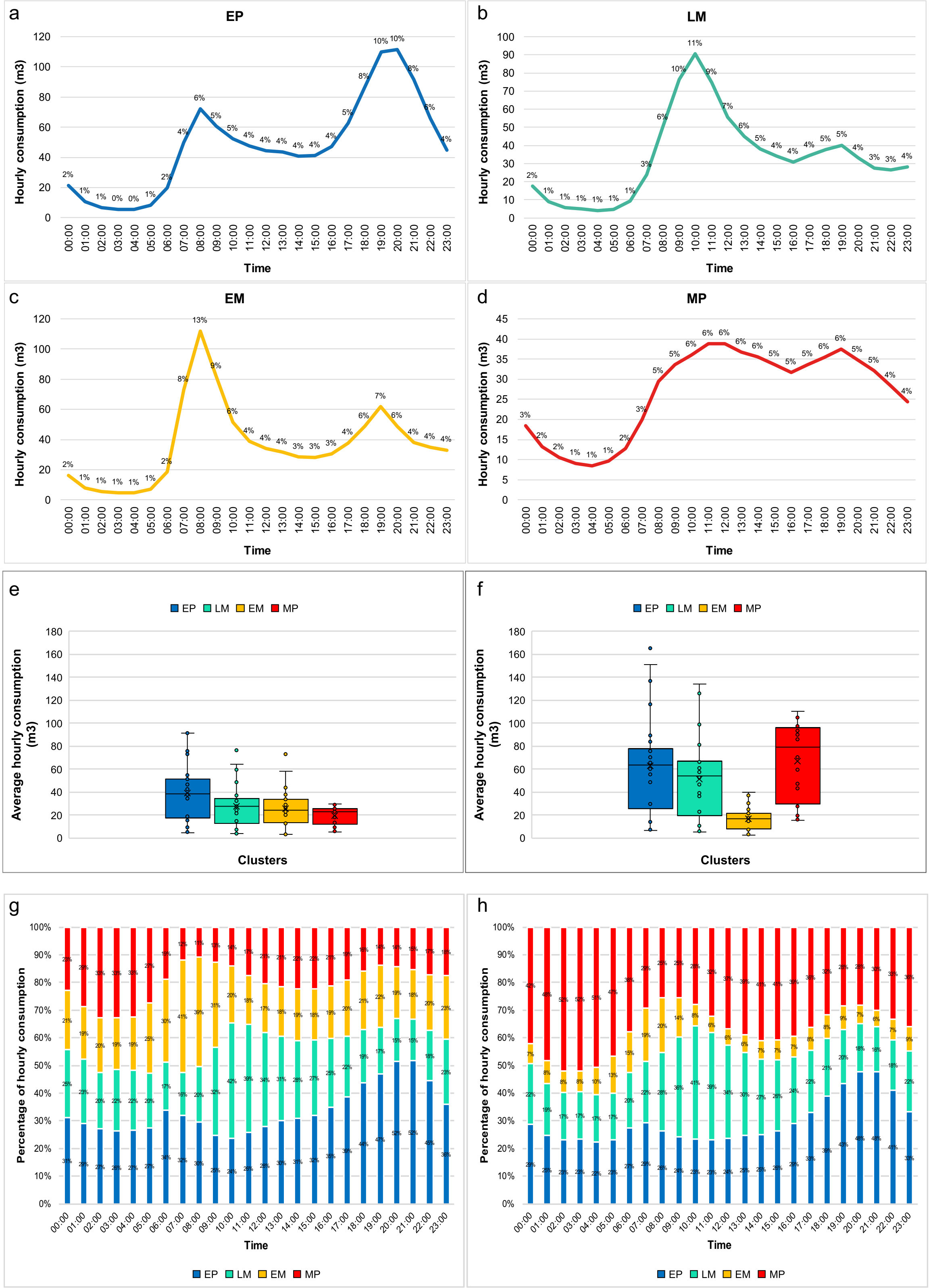
EM

Households in EM have the fewest instances of peaks which constitute ~12–13% in 24 hours and occurs between 07:00 and 08:00. Q3 for this cluster, ~7% of their relative daily consumption, occurs at 19:00, Q2 between 10:00 and 17:00, constituting ~3–4% of their relative daily consumption per hour, and Q1 identical to EP and LM (Fig. 2c). On average, this cluster, made up of 26% of household, was responsible for 40% (~59 m^3/hr) of pre-lockdown consumption occurring between 07:00 and 08:00 (Fig. 2g.) and 22% (723 m^3/d) of the total daily consumption, with $\bar{x}$ of 26 m^3/hr, σ of 17 m^3/hr and max of 73 m^3/hr (Fig. 2e).

EM experienced the sharpest decrease in the number of households during the lockdown period—an average of 12% across the network, resulting in a significant drop in their share of relative consumption between 07:00 and 08:00 to from 40% to 20%–~38 m^3/hr (see Fig. 2h). Their share of the total daily consumption also fell to 12% (433 m^3/d). Lockdown being, respectively, $\bar{x}$ 17 m^3/hr, σ 10 m^3/hr and max of 40 m^3/hr (Fig. 2f).

MP

MP has the highest instances of Q4s within 24 hours (about seven instances of 6–7% of their relative daily consumption). They also have multiple instances of Q3s and Q2s at 5% and 4% of relative daily consumption, respectively. Their Q1, like the other clusters, resides between 00:00 and 06:00, constituting ~2–3% of relative daily consumption (Fig. 2d). During the pre-lockdown weeks, this cluster represented 14% of the households across the network. MP dominates consumption between 00:00 05:00—at an average of 32% (~8 m^3/hr) (Fig. 2g) and about 20% (661 m^3/d) of the total



daily consumption, with $\bar{x}$ 19 m^3/hr, σ 8 m^3/hr and max 29 m^3/hr (Fig. 2e).

MP experienced the most significant increase in the number of households during the lockdown period—a 93% increase between M3 and M4 maintaining an average of 26% of all households during the lockdown period. This has resulted in an increase in their share of hourly consumption between 00:00 and 07:00 to an average of 45%–~24 m^3/hr; between 12:00 and 17:00 to an average of 39%–~98 m^3/hr and 23:00 to 36%–~60 m^3/hr (see Fig. 2h). Their share of the total daily consumption also rose to

Fig. 2 Clusters' hourly consumption patterns and comparison of clusters' share of consumption before and during the COVID-19 lockdown in the UK. a Cumulative pattern and percentage of hourly consumption for households in the "Evening Peak (EP)" cluster. Consumption is in (m^3). **b** Cumulative pattern and percentage of hourly consumption for households in the "Late Morning Peak Peak (LM)" cluster. Consumption is in (m^3). **c** Cumulative pattern and percentage of hourly consumption for households in the "Early Morning Peak (EM)" cluster. Consumption is in (m^3). **d** Cummulative pattern and percentage of hourly consumption for households in the "Multiple Peak (MP)" cluster. Consumption is in (m^3). **e** Average daily consumption per cluster pre-lockdown. **f** Average daily consumption per cluster during the lockdown. **e**, **f** Value at the top of the whisker is the Maximum consumption; bottom of the whisker is the minimum consumption; 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 median and the *x* in the centre of the box is the mean ($\bar{x}$). **g** Clusters' share of total hourly consumption pre-lockdown. **h** Clusters' share of total hourly consumption during the lockdown.

~32% (~1326 m^3/d). Lockdown being, respectively, $\bar{x}$ 67 m^3/hr, σ 34 m^3/hr and max of 110 m^3/hr (Fig. 2f).

In another study,[14] segmentation was based on heterogeneous micro-component consumption patterns and behaviour regularities and temporal characteristics. This work, unlike our study, performed a disaggregation of sub-minute smart meter data into end-use events, subsequently clustering households based on their end-use similarities. One difference of this study to the one here reported, however, is the consumer household sample size. In our study, some 11,528 households were assessed, and currently, sub-minute smart meters are unavailable across such a large region. Our segmentation was derived from normalised hourly smart meter data, being based on temporal patterns of consumption. The silhouette coefficient value, when 't-distributed stochastic neighbour embedding' (t-SNE)[15] was used for dimensionality reduction (as opposed to PCA), improved slightly from 3.9 to 4.1 for n_cluster = 4. However, this improvement only marginally enhanced the k-means results (by < 1%). Household consumption patterns over a 24 hour period[16] are generally typified by minimal activities during the early hours (00:00–06:00), followed by a significant diurnal rise (07:00–10:00), a diurnal drop (11:00–16:00) and then another significant rise from 16:00 to 20:00. The significant differences in the clusters are underpinned by the time and percentage of both relative and absolute peak consumption (Q4) and time and percentage of the next most significant event (Q3). Our results demonstrate conclusively that, whilst many intercluster pattern similarities exist in Q1s, Q2s and Q3s (accounting for the relatively low silhouette coefficient score), for EP, LM and EM, 94% of households share the same demand peak times with their intracluster neighbours (Fig. 3), the remaining 6% being within ±1 hour of the typical peak times. MP, however, has no typical demand peak time and households in this cluster are characterised by several peak and near peaks events at different times.

This paper provides a reliable quantified characterisation of household water consumption patterns across the network. It has revealed the extent to which the COVID-19 lockdown has impacted materially on these patterns of consumption, leading to a marked change in demand. Table 1 shows the summary of the COVID-19 impact on water demand at PHC level.

Peak demand pattern management forms the core of water DSM strategy[17] and as such the accurate identification of household consumption patterns underpinning peak temporal demand[18] and detailed characterisation of these patterns represents an important step towards the development of strategies aimed at reducing consumption and improving demand forecasting. Studies have established how both exogenous (e.g., seasonal) and endogenous (e.g., SE/SD) variables[18] contribute to peak demand, thereby enabling predictions. Our characterisation of household clusters encapsulates multiple levers such as the times of peak demand for each cluster, their relative proportion of water consumed at the peak, their absolute proportion of the peak demand as well as their absolute shares of hourly consumption. This allows a precise identification and prediction of geographical consumption hotspots at multiple resolutions (hourly, daily, monthly or annually) and sets parameters for more precise forecasting for the utility and shifting the peak demand in households.[17] For example, our model allows the utility to determine not only the proportion of peak demand between 07:00 and 08:00 to total network demand but also to identify the households underpinning that and the relative proportion of the households' daily demand that this constitutes.

Household water demand is characterised by marked temporal variation, with diurnal patterns of demand being the most diverse.[19] For peak demand management, it is vital to keep track of times of peak demand and the groups of households contributing to that. One of the difficulties faced by water companies is the ability to track effectively individual household patterns of behaviour, ascertaining the extent to which these patterns change and their impact on aggregate demand. Previous studies have clustered households based on shared micro-component patterns to examine the likely impact of alterations in consumption of particular water use events and other cluster characteristics, coupled with seasonal variability and occupancy, on aggregate demand.[20] Our work allows the quantification of the impact that each cluster has on aggregate demand at any given time and further allows the tracking of household movements from one cluster to another, as well as the associated peak characteristics. This makes it possible to quantify the impact of change in household circumstances such as a change in work patterns, tenure, household appliance use, occupancy, SD or SE variables and leakage.

Our work also provides a basis for identifying water consumption hotspots within the distribution network and supports seasonal and regional temporal load monitoring in both normal and unusual times, such as during pandemics, to help mitigate water stress and the resultant impact on the environment and water resources.

All households assessed in the study derive from a single water provider, drawn across two geographical areas 50 miles apart consisting of 24 DMAs. As the aim of this study was to quantify the impact of the Covid-19 lockdown on aggregate water demand while highlighting household clusters' underpinning temporal demand patterns, only anonymised smart meter data were utilised. Data on sociodemographic/SE or occupancy variables of the participating households were not available.

January week 2 (J2) recorded a sharp increase of 57% of households in EM, ostensibly signifying a return to work (characterised by 07:00–08:00 peaks) after the Christmas break, and a simultaneous decrease in LM and MP by 17% and 33% (characterised by diurnal and MPs), respectively (Fig. 1e, f). The next significant cluster migration event occurred between March week 3 and 4 (M3 and M4), which saw a 93% increase in MP and a 52% decrease in EM, ostensibly signifying more people working from home and assuming MP characteristics (Fig. 1e, f).

Our study addresses fundamental policy issues relating to the Demand-Side Management (DSM) of water resources. First, it focuses on the categorisation of households by their respective temporal as well as the aggregate impact on network demand, as a basis for demand forecasting. Focusing on peak patterns of daily water consumption can lead to both relative and absolute water conservation.[21] Metering alone is capable of reducing PCC by

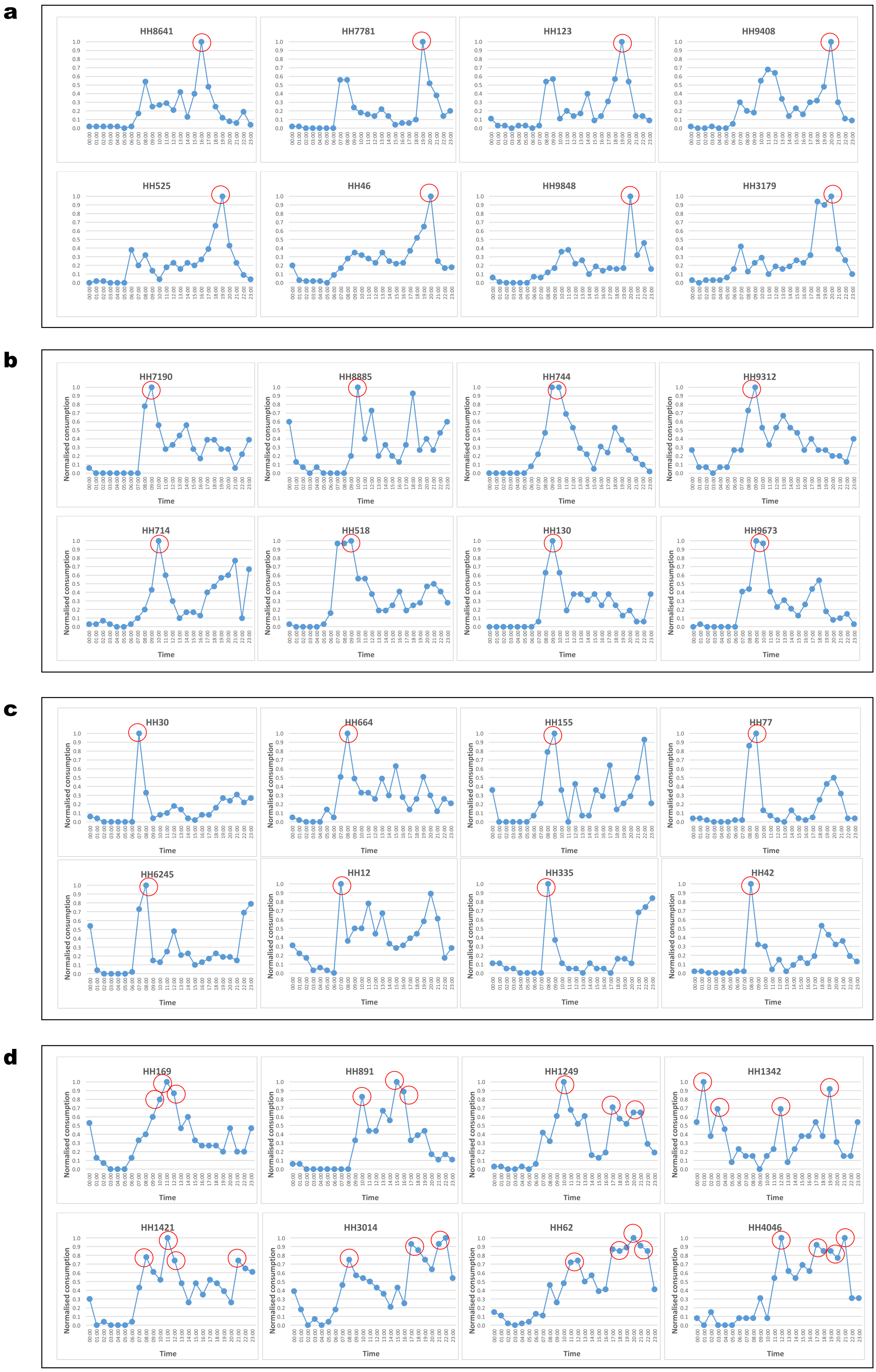


Fig. 3 24 h patterns of individual households in clusters. **a** 24-hour patterns for a random sample of households in the EP cluster. **a** 24-hour patterns for a random sample of households in the LM cluster. **a** 24-hour patterns for a random sample of households in the EM cluster. **a** 24-hour patterns for a random sample of households in the MP cluster. **a**–**d** Circles indicate observed peaks.

Table 1. Comparison between pre-lockdown and lockdown impact on household clusters' water demand.

	EP	LM	EM	MP
Average number of households pre-lockdown (J1–M3)	3470	3442	3016	1600
Percentage of households	30%	30%	26%	14%
Average number of households during lockdown (M4–MY4)	2842	4239	1627	2820
Percentage of households	25%	37%	14%	24%
Pre-lockdown PHC (l/h/d)	307	235	240	413
During lockdown PHC (l/h/d)	383	302	266	470
Percentage increase	25%	29%	11%	14%

15%,[8,22,23] which, as reported by UKWIR,[24] could be significantly improved by the clustering of households into categories of water consumption characteristics, achieving marked accuracy improvements in demand forecasting.

Second, our analysis of the unique patterns of behaviour of the household clusters before and during the lockdown provides the water sector with an insightful basis for the assessment of network demand load and its distribution attributable to the impact of the COVID-19 lockdown.

Third, the approach presented offers a methodology to support examination of resilience for future unforeseen events capable of placing immense stress on network demand in particular and the water resources in general. The approach adopted is aligned with UK water legislation (the Water Act 2014)[25] as it (a) sets the grounds for long-term water resource planning by informing the provision of a range of measures to sustainably manage water resources and (b) helps to improve household water use efficiency, reduce aggregate demand and reduce the pressure on water resources.

We conclude that whilst the scope of this study focuses on the characterisation of consumption patterns attributable to the COVID-19 pandemic lockdown, the work has also identified significant implications for future research. First, our segmentation method and characterisation of consumption patterns can be adapted to support demand management strategies by delivering intervention feedback to households according to their specific cluster characteristics. Second, the model can play a role in seasonal and temporal (diurnal and night-time) water consumption dynamics. Third, it is estimated that households' internal leakage constitutes ~10% of the total network water consumption.[26] Many network leakage detection systems rely on the analyses of night flow patterns by monitoring when hourly flow value surpasses a fixed threshold at the same time interval.[26,27] Our cluster characterisation delivers a framework that would facilitate the tagging of network households that fall in this category (typically MP), for a targeted leakage intervention. Drawing from this work, we further consider that to gain improved insight into consumer patterns, future research should incorporate data from high temporal resolution smart meters, that can offer greater discrimination than the hourly data available to us here. This would have significant cost implications and thus such work may need to wait to benefit from new generations of future metres as technology develops. Future work should also cross-correlate electricity supply smart meters as well as other corollary data such as prevailing weather patterns.

METHODS

In all, 20 weeks' smart meter data for January to May 2020 were obtained from 20,000 anonymised households and processed using exploratory machine learning (ML) for unsupervised pattern recognition and supervised pattern classification.

To achieve a robust classification model, the clustering and labelling process was performed on each week's normalised data set. Several models were then constructed using each of the labelled data sets. The classification accuracy of models drawing on pre-lockdown data sets was similar to that of lockdown-derived data sets. The process of data normalisation, however, removed all scale properties from the data set such that all data points are between 0 and 1.

Data preprocessing

Data cleaning, integration, feature reduction, and transformation were conducted to improve the accuracy of the ML algorithms.[28] The data were pivoted into 24 hour columns so that each datapoint was a weekly mean, derived using Eq. 1:

$$\bar{x} = \frac{\sum x_i}{n} \qquad (1)$$

where $\bar{x}$ is the monthly mean consumption for each of the 24 hour columns; $\sum_i^x$ is the sum of each of the 24 hour columns and n is the number of each of the 24 hour columns. Outliers such as commercial customers ($n = 830$), null values and inconsistent customers were removed leaving 11,528 consistent households comprising 891 unique postcodes across 25 DMAs, in two areas in East Anglia, England (Town A and Town B). The weekly data were labelled thus: January week 1 to week 4 = J1..J4; February week 1–week 4 = F1..F4; March week 1–week 4 = M1..M4; April week 1–week 4 = A1..A4 May week 1–week 4 = MY1..MY4.

Data normalisation

The data set values presented with differing ranges. A min–max scaler[28] was applied to achieve a linear transformation on the original data by normalising the data range such that the range was set between 0 and 1 (Eq. 2):

$$z = \frac{x_i - \min(x)}{\max(x) - \min(x)} \qquad (2)$$

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.

This normalisation preserves the relationships among the original data values within a range,[28] i.e., all range attributes are expressed as a fraction of the range max, which makes it particularly ideal for revealing patterns by highlighting the times of peak, near peak or lowest consumption in a uniform fashion.

Dimensionality reduction

Dimensionality reduction using 't-distributed stochastic neighbour embedding'[15] was applied to transform the 24-hour features of the normalised data into two dimensions (TA1 and T2), which was a crucial process for the improvement of clustering and classification accuracy, the goal being to preserve the meaningful structure of the features while using fewer attributes to represent them.[28,29]

ML process

ML pattern-mining techniques were used to explore the normalised data sets for unknown patterns, perform cluster analysis on all observations (households) in the data set based on found patterns, label the observations with the assigned clusters, train a classification model with the labelled data and "predict" the clusters for the remaining normalised data set.

Pattern recognition (clustering)

The "k-means algorithm", a "centroid-based partitioning clustering method" was chosen, owing to the nature of the data points (n) in the data set and the fact that his algorithm typically employs exclusive segregation of clusters such that each object must belong to exactly one group, where each partition represents a cluster and $k \leq n$, where k is the number of specified clusters. That is, it divides the data into k groups (each with a defined centroid) such that each group contains at least one object.[28,30]

The objective function for the K-means clustering algorithm is the squared error function[31] (Eq. 3):

$$J = \sum_{k=1}^{k}\sum_{i=1}^{n}\left\|(x_i - \mu_k)^2\right\| \tag{3}$$

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 optimal number of clusters (or the k value) in the normalised data set became our fundamental challenge. We applied the Elbow method, which uses the "Within-Cluster-Sum-of-Squares (WCSS)"[28] algorithm, to the data set and arrived at $k = 4$.

The silhouette coefficient algorithm,[28,31] used to measure the similarity of objects in the data set to their assigned clusters compared with other clusters, revealed that $k = 4$ has the highest coefficient (0.41), making 4 the optimal number of clusters (Eq. 4).

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}} \tag{4}$$

Where $s(o)$ is the silhouette coefficient value of data point or object o in the data set, $a(o)$ 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.

The k-means algorithm was applied to the data set with the following parameters: $k = 4$; maximum iteration = 300 and random state = none. These steps were repeated on the remaining data sets (February–May), ending the unsupervised machine learning process.

Pattern classification

The purpose of our pattern classification is to use training samples from the labelled data set to predict the clustering of our input data using its features.[32] A supervised learning technique was used to perform pattern classification on the remaining input data set to aid comparison of the classification with the clustering outputs. To achieve a robust classification model, the clustering model was applied successively to each week's normalised data and values labelled. Several models were then constructed using each of the labelled data sets. A "train-test-split"[28] model evaluation procedure was used to compare the classification accuracies of two supervised approaches: K Nearest Neighbour (KNN)—a distance-measuring pattern classification algorithm and Logistic Regression (LR)—a probabilistic pattern classification algorithm.[32–34] Each of the labelled data sets was split into 60% train and 40% test 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, we tested the accuracy of the model within a "k range" of $k =$ 1..$k = 25$. Testing accuracy for LR was between 93% and 94% on all the data sets, which makes it a better performing model for the data set. A further LR classification model was built for the assignment of cluster labels to the remaining normalised data set, achieving a classification accuracy of 94%.

DATA AVAILABILITY

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

Received: 26 June 2020; Accepted: 12 January 2021;
Published online: 18 February 2021

REFERENCES

1. Jarvis, C. I. et al. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC Med.* **18**, 124 (2020).
2. Baker, S. R., Farrokhnia, R. A., Meyer, S., Pagel, M. & Yannelis, C. *How Does Household Spending Respond to an Epidemic?* Consumption During the 2020 COVID-19 Pandemic. *Rev. Asset. Pricing Stud.* **10**, 834–862 (2020).
3. Aquatech. Case study: Data links COVID-19 lockdown to consumption change. Available at: https://www.aquatechtrade.com/news/utilities/covid-19-lockdowns-impact-water-consumption/. (2020)
4. Artesia. *The effect of the coronavirus lockdown on water use*. (2020).
5. Kalbusch, A., Henning, E., Brikalski, M. P., Luca, F. Vde & Konrath, A. C. Impact of coronavirus (COVID-19) spread-prevention actions on urban water consumption. *Resour. Conserv. Recycl.* **163**, 105098 (2020).
6. DEFRA. *Future Water Strategy for England*. (2008).
7. HR Wallingford. *Updated projections for water availability for the UK.* (2015).
8. Lawson, R. *et al. The long term potential for deep reductions in household water demand*. (2018).
9. Parker, J. M. & Wilby, R. L. Quantifying household water demand: a review of theory and practice in the UK. *Water Resour. Manag.* **27**, 981–1011 (2013).
10. Giurco, D. et al. *Residential end-use measurement guidebook a guide to study design, sampling and technology*. (Institute for Sustainable Futures, UTS, 2008).
11. Willis, R. M., Stewart, R. A., Panuwatwanich, K., Williams, P. R. & Hollingsworth, A. L. Quantifying the influence of environmental and water conservation attitudes on household end use water consumption. *J. Environ. Manag.* **92**, 1996–2009 (2011).
12. Nawaz, R. et al. Long-term projections of domestic water demand: a case study of London and the thames valley. *J. Water Resour. Plan. Manag.* **145**, 1–17 (2019).
13. Kendon, M. *Storm Ciara. Met Office National Climate Information Centre* (2020).
14. Cominola, A. et al. Data mining to uncover heterogeneous water use behaviors from smart meter data. *Water Resour. Res.* **55**, 9315–9333 (2019).
15. Van Der Maaten, L., Courville, A., Fergus, R. & Manning, C. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
16. Carragher, B. J., Stewart, R. A. & Beal, C. D. Quantifying the influence of residential water appliance efficiency on average day diurnal demand patterns at an end use level: a precursor to optimised water service infrastructure planning. *Resour. Conserv. Recycl.* **62**, 81–90 (2012).
17. Beal, C. D., Gurung, T. R. & Stewart, R. A. Demand-side management for supply-side efficiency: modeling tailored strategies for reducing peak residential water demand. *Sustain. Prod. Consum.* **6**, 1–11 (2016).
18. Beal, C. D. & Stewart, R. A. Identifying residential water end uses underpinning peak day and peak hour demand. *J. Water Resour. Plan. Manag.* **140**, 1–10 (2014).
19. McDonald, A., Butler, D. & Ridgewell, C. Water demand: estimation, forecasting and management. in *Water Distribution Systems* 49–71 (Thomas Telford Ltd, 2011). https://doi.org/10.1680/wds.41127.049.
20. Browne, A. L., Medd, W. & Anderson, B. Developing novel approaches to tracking domestic water demand under uncertainty-a reflection on the 'up scaling' of social science approaches in the United Kingdom. *Water Resour. Manag.* **27**, 1013–1035 (2013).
21. Cardell-Oliver, R. Water use signature patterns for analyzing household consumption using medium resolution meter data. *Water Resour. Res.* **49**, 8589–8599 (2013).
22. Boyle, T. et al. Intelligent metering for Urban. *Water.* **5**, 1052–1081 (2013).
23. EA. *International comparisons of domestic per capita consumption Prepared for the Environment Agency by Aquaterra*. (2008).
24. UKWIR. *WRMP19 METHODS – HOUSEHOLD CONSUMPTION FORECASTING: GUIDANCE MANUAL* (2015).
25. Legislation.gov.uk. Water act 2014, 1–243 (2014).
26. Britton, T. C., Stewart, R. A. & O'Halloran, K. R. Smart metering: enabler for rapid and effective post meter leakage identification and water loss management. *J. Clean. Prod.* **54**, 166–176 (2013).
27. Luciani, C., Casellato, F., Alvisi, S. & Franchini, M. Green Smart Technology for Water (GST4Water): water loss identification at user level by using smart metering systems. *Water (Switzerland)* **11**, 405 (2019).
28. Han, J., Kamber, M. & Pei, J. *Data Mining: Concepts and Techniques*. *Data Mining: Concepts and Techniques* (Elsevier, 2012). https://doi.org/10.1016/C2009-0-61819-5.
29. Munzner, T. *Visualization Analysis & Design*. *Taylor & Francis Group* (A K Peters/CRC Press, 2015).
30. Zhu, S., Wang, D. & Li, T. Data clustering with size constraints. *Knowl.-Based Syst.* **23**, 883–889 (2010).
31. Yuan, C. & Yang, H. Research on K-value selection method of K-means clustering algorithm. *J* **2**, 226–235 (2019).
32. Tsai, C.-F., Lin, W.-Y., Hong, Z.-F. & Hsieh, C.-Y. Distance-based features in pattern classification. *EURASIP J. Adv. Signal Process.* **2011**, 1–11 (2011).
33. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.* **35**, 352–359 (2002).
34. Bishop C. M. Pattern Recognition and Machine Learning. *Religion und Konflikt*. https://doi.org/10.13109/9783666604409.185 (2011).

ACKNOWLEDGEMENTS

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

AUTHOR CONTRIBUTIONS

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

COMPETING INTERESTS

The authors declare no competing of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to L.W.

Reprints and permission information is available at http://www.nature.com/reprints