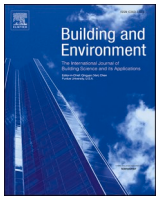
Building and Environment 231 (2023) 110032

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv



A cluster analysis approach to sampling domestic properties for sensor deployment

Tamaryn Menneer [a,c,*], Markus Mueller [b,c], Stuart Townley [b,c]

[a] European Centre for Environment and Human Health, University of Exeter Medical School, Knowledge Spa, Royal Cornwall Hospital, Truro, TR1 3HD, UK
[b] College of Engineering, Mathematics and Physical Sciences, Penryn Campus, University of Exeter, Penryn, TR10 9FE, UK
[c] Environment and Sustainability Institute, Penryn Campus, University of Exeter, Penryn, TR10 9FE, UK

ARTICLE INFO

Keywords:
Cluster analysis
Representative sampling
Sensors
Water
Energy
Environment

ABSTRACT

Sensors are an increasingly widespread tool for monitoring utility usage (e.g., electricity) and environmental data (e.g., temperature). In large-scale projects, it is often impractical and sometimes impossible to place sensors at all sites of interest, for example due to limited sensor numbers or access. We test whether cluster analysis can be used to address this problem. We create clusters of potential sensor sites using factors that may influence sensor measurements. The clusters provide groups of sites that are similar to each other, and that differ between groups. Sampling a few sites from each group provides a subset that captures the diversity of sites. We test the approach with two types of sensors: utility usage (gas and water) and outdoor environment. Using a separate analysis for each sensor type, we create clusters using characteristics from up to 298 potential sites. We sample across these clusters to provide representative coverage for sensor installations. We verify the approach using data from the sensors installed as a result of the sampling, as well as using other sensor measures from all available sites over one year. Results show that sensor data vary across clusters, and vary with the factors used to create the clusters, thereby providing evidence that this cluster-based approach captures differences across sensor sites. This novel methodology provides representative sampling across potential sensor sites. It is generalisable to other sensor types and to any situation in which influencing factors at potential sites are known. We also discuss recommendations for future sensor-based large-scale projects.

1. Introduction

Remote and automatic collection of data has become increasingly viable with the development of the Internet of Things [1,2], smart meters [3,4] and sensor networks [5,6]. Availability of such data opens up new avenues of research for multiple domains in analysing, monitoring or forecasting from measurements of water usage [7], gas usage [8,9], electricity usage [10–12] and usage profiles [3,4,13–15], occupancy monitoring [16–18], energy efficiency across different build types or occupant behaviours [19–21], temperature [22,23], humidity [24], air quality [25–30], effects of ventilation [31–33], water quality and temperature in lakes and streams [34,35], drinking water quality [36], and monitoring of health and environment in housing association homes [37–39].

In parallel with this enhanced access to data, there has been an increasing practical drive for energy savings [40,41], renewable energy utilisation [42], healthy home environments [43], and smart control of domestic systems [44,45]. Establishing the influences on these factors requires research into energy usage, water usage and environmental conditions in domestic settings, which relies on feasible monitoring and effective sensor placement.

The purpose of the current study is to address a core long-standing problem for monitoring sensors. The problem is selecting representative and optimised locations for placing sensors [46,47] when it is not possible to place sensors in all potential locations, for example, due to limited numbers of sensors or difficulties with installation. Suboptimal placement of sensors can result in collecting data across similar situations, thereby providing potentially redundant information. Sensors could instead be more usefully deployed across a variety of sites [48], and by targeting locations through representative sampling.

Research on optimum placement of sensors is often necessarily specific to the type of sensor and the application [e.g., water contamination:

* Corresponding author. European Centre for Environment and Human Health, University of Exeter Medical School, Knowledge Spa, Royal Cornwall Hospital, Truro, TR1 3HD, UK.
E-mail address: t.s.i.menneer@exeter.ac.uk (T. Menneer).

https://doi.org/10.1016/j.buildenv.2023.110032
Received 28 October 2022; Received in revised form 19 December 2022; Accepted 19 January 2023
Available online 29 January 2023

[49], fluid flow: [50]], with placement design requiring solutions for specific spaces and sensors [51]. For indoor air quality or temperature, fluid dynamic modelling of pollutant or heat diffusion can highlight locations that would be most useful to monitor [52], with methods typically adapted to reduce the computational complexity [46,51,53], or incorporate influences from occupants and the building [54]. For studies on households, previous research has placed sensors in a subset of homes and assessed representativeness of the subset using census data, in order to provide a subset of sensor data alongside a full survey dataset [55].

Selection of sensor locations is often aimed at maximising coverage of a two- or three-dimensional known physical space. In these circumstances mathematical and spatial analysis techniques can be used to provide complete or optimised coverage, for example, for gas detection [56], security monitoring [47], and environmental sensors [57]. However, sensor placement is not always driven by spatial location, but by other potential influences, for example, in the current study, household characteristics and road distance.

Methods for optimising sensor placement include minimising uncertainty in the data estimates [58], evolutionary algorithms and neural networks [59], with machine learning as a promising emerging approach [60]. However, given specific applications and tailored methods, the techniques are not often readily accessible to those requiring sensor monitoring, and cannot guarantee a generalisable solution for users such as building owners [61].

Futhermore, the optimal placement of sensors might not necessarily comprise uniform coverage of a known feature space, rather the coverage needs to reflect the values and weightings of features of the potential locations themselves. For example, in the current study, the purpose is to deploy limited sensors across a representative sample from our participants' homes and their locations. We therefore wish to use a data-driven approach to ensure we capture the similarities and variety specifically within our cohort.

In the current study we use cluster analysis in a novel methodology for selecting a representative sample of homes for sensor installation. Cluster analysis is an established technique for grouping individuals according to similar feature values [62], whilst also representing variety across groups. The benefits of this approach for sensor placement are that it is (1) generalisable to new settings and applications; (2) it is data-driven, so that the groups are defined by the set of potential sensor locations rather than being influenced by feature combinations that may not exist; (3) groups are based on known features that are expected to influence the monitored data.

Clustering methods have been widely applied within the field of sensors and monitoring, including selecting locations for sensor placement. However, unlike our study, clustering was performed on sensor data, rather than only using factors believed to influence those sensor data (i.e., in the current study, household and local environment characteristics). In one study, clusters based on sensor data were further refined with spatial clustering, as detailed below [61].

For utility use, clustering has been used to group together similar electricity demand or usage patterns in terms of magnitude and timing in order to determine categories of power station loads [[63], and in conjunction with renewable energy availability: [64–66]] and household usage profiles [67–71]. Clustering of gas and water usage data, has been used to identify usage profiles for categorisation or prediction [7, 72–79], and clusters of energy usage across different building systems have revealed different user behaviour patterns [80].

Clustering by electricity usage has also been used to examine the household characteristics within each cluster [3,4], and supports findings that energy usage is influenced by characteristics such as household size and the time spent at home [81,82].

For environmental measures, clustering has been used to determine patterns in temperature and humidity data [83] and comfort levels across clusters [24], to cluster climate data [84–86], and establish emergent geographical patterns [87]. Clustering of air pollution data groups monitoring stations or cities with similar readings, which can inform network development [88–90] and appropriate pollution reduction methods [91]. Air quality data has also provided a base for testing methods for clustering of time-series data [92,93].

Specifically applied to optimising sensor placement, clustering has also been primarily based on the sensor measurements. For the detection of water leaks, estimated pressure changes across the water distribution network were clustered into different types, and the most informative locations from each type were selected for pressure monitoring [94].

For environmental sensors (temperature, humidity and luminance), sensors were also clustered into groups based on the sensor data [61]. Separate cluster analyses for different areas of the building allowed for the influence of air conditioning. Clusters were refined using spatial clustering results. Strategies were provided and used to place a limited number of sensors for monitoring an office environment. Sensor coverage was validated by comparison with data from a full set of sensors [61].

The purpose of the current study is to apply cluster analysis in a novel methodology for representative sampling from potential sensor sites. We apply this method to sampling across domestic properties in order to inform the placement of gas, water and environmental sensors. In contrast to previous work, cluster analysis is performed using household and environment characteristics, as opposed to sensor data, to create clusters of similar homes. The method is tested with two specific sensor types, but is generalisable to any situation in which influencing factors at potential locations are known. This methodology addresses a resource limitation problem when sensors are limited in number and need to be placed to maximise coverage of the potential dataset. Groups are defined by the features of the potential sensor locations, such that they represent the similarities and diversity within the cohort to be sampled. Sensors are placed into some homes in each cluster to provide a representative sample across all types of home. The resulting clusters and chosen sensor placements are verified using one year of utility usage data and environmental measurements collected at 3–30 min intervals across a maximum of 280 homes.

2. Overview

In the next section we describe the broader project to provide the context for the current aim of representatively sampling potential sensor sites (Section 3). There are four main steps to this study. The first is conduct cluster analyses to provide groups of similar homes from which to sample across (Section 4). Secondly, the most appropriate cluster solution is chosen for each application (utility sensors and environmental sensors) (Section 5). Thirdly, sensors of each type are installed using the chosen cluster solutions to inform installation sites (Section 6). Finally, we use the sensor data collected to assess whether this clustering approach achieved its purpose of representative sampling of potential sensor sites (Section 7). Fig. 1 provides an overview of the four steps, including the datasets used and the number of homes or sensors available at each relevant step.

3. Study context and data collection

3.1. The Smartline project

Over 300 households were recruited to take part in *Smartline*, from domestic properties that are managed by Coastline Housing, a housing association in Cornwall, South West UK. The overarching aim of the *Smartline* project is to investigate opportunities for technology to support healthier and happier living in homes and communities [95–101]. To our knowledge, *Smartline* is the largest domestic project of its kind to date, although non-domestic projects are ongoing [102].

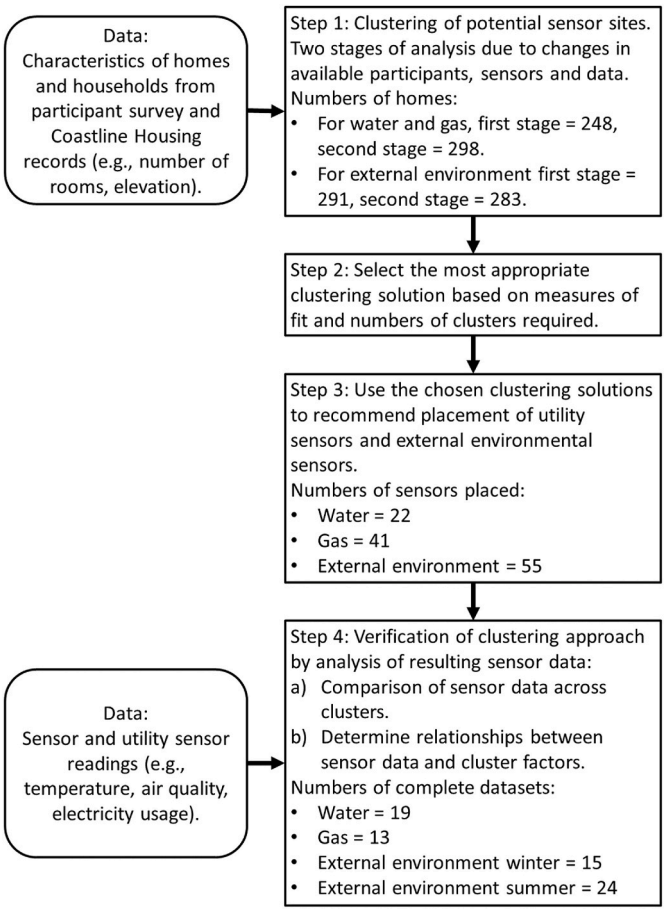


Fig. 1. Overview of the study, showing the sequence of steps, the contributing datasets and the numbers of homes or sensors.

3.2. Smartline data

Survey, sensor and housing data were collected from the participating homes, following informed written consent. The large dataset is a unique combination of cross-sectional and time-series data, including household characteristics and behaviours, environmental readings, and utility usages.

3.3. Surveys

Face-to-face surveys were conducted with 329 participants in September 2017 to November 2018. In the broader *Smartline* project, survey data were collected to provide information about the home, the household, occupant behaviours, community interactions, health and wellbeing.

3.4. Sensors

On the *Smartline* project, utility usage sensors and indoor environmental sensors were installed in up to 280 homes from October 2017 onwards. The broader purpose of the sensors, within *Smartline*, was to provide information on the indoor environment and utility usage, to be considered in relation to occupant health and wellbeing. Environmental sensors external to the home were also installed to provide a context when considering the indoor environment.

Utility readings comprised electricity, gas and water usage for the. They were each installed on the utility supply meter to provide overall measures of usage for whole property. Readings were recorded every 3–7.5 min. Indoor and external environmental measures comprised air temperature, relative humidity (RH), volatile organic compounds (VOCs), equivalent CO2 (eCO_2) and particulate matter up to size 2.5 μm (PM2.5), together with PM10 for the external environmental sensors. Measurements were taken every 3–5 min in the living room and bedroom, and every 30 min for the external sensors.

All sensors were manufactured by Invisible Systems Ltd. and installed by Blue Flame (Cornwall) Ltd. from October 2017 onwards. Table 1 in the provides sensor details. Given the proximity of homes taking part in this study, the sensor gateway in one home can be close enough to transmit readings from sensors in other homes, such that multiple readings can be captured within the update interval. The update interval of 7.5 min is standard for these sensors, providing 8 readings per hour. A shorter interval was chosen for sensors when battery life or mains power allowed (see Table 1). An update rate of 30 min was chosen for external sensors based on estimations for the battery to last for two years.

3.5. Current study

Smartline electricity and indoor environmental sensors were available for installation in 280 homes. However, there were limited numbers of water, gas and external environmental sensors. The aim of the current study was to select sites for the placement of these sensors in order to capture a representative range of homes and environments across the *Smartline* cohort area.

In this study, survey responses are used as factors for the cluster analyses in order to create groups of similar homes to be sampled across. The same survey factors are also used as predictors in regression analyses to verify the cluster solutions. More details are provided about the measures used as factors and predictors in relevant sections below. A reference table for terminology is presented in Table 5.

Data from sensors are used to verify whether the cluster analysis approach was successful in informing representative placement of sensors that were limited in number. The purpose of analysing these data is not to draw conclusions about the measures themselves, but is to verify the clustering method.

4. Clustering methods

This section presents the cluster analysis methods, using known factors to create groups of similar homes, from which sensor sites can be sampled.

Two sets of cluster analyses were conducted on factors representing characteristics of the potential sensor sites. The purpose of each was to determine a set of homes to provide a representative sample for the placement of utility usage sensors and external environmental sensors, one type per analysis. After assigning homes into clusters, we sampled from each cluster to provide a subset of homes that captured the range of characteristics across all homes.

The first analysis was for the placement of sensors monitoring water and gas usage (m^3). The second analysis was for the placement of external environmental sensors for temperature (°C), relative humidity (RH, %), volatile organic compounds (VOCs, parts-per-billion), equivalent CO_2 (eCO_2, parts-per-million), and particulate matter of sizes 2.5 μm and 10 μm (PM2.5 and PM10, $\mu g/m^3$).

The process for each analysis was the same except different sets of factors were included for the clustering. Because of changes in participating households, availability of data for cluster factors, availability of additional sensors, and practical limitations in installing utility sensors, two stages were conducted for each analysis.

For the utility sensors, the first stage was conducted on 248 homes in November 2017 for the placement of up to 50 water and 50 gas sensors. The second stage was conducted on 298 homes in October 2019. Between the first and second stages 81 households withdrew from the project and 131 joined. In addition, more suitable data became available for the cluster factors, as described below. Installations were restricted

Table 1
Sensor information for each type of utility usage and air measurement.

Sensor data type	Sensor model	Sensor technology	Sampling interval	Measurement accuracy or operation notes
Electricity	ISL067 Smart RF Mk2 single-phase wireless meter (ref: QC0142)	Split core current transformer providing energy and current measurements using the output low AC voltage (0–0.33V AC).	Readings recorded every 3 min.	±10%
Gas	Pulse transmitter (ref: QC0145b)	The sensor takes the pulse output from the utility meter installed in the home, and operates with either rotary or electronic pulse meters.	Count of the number of pulses generated per 7.5 min.	In line with equipment it is connected to.
Water	Pulse transmitter (ref: QC0145c)	The sensor takes the pulse output from the utility meter installed in the home, and operates with either rotary or electronic pulse meters.	Count of the number of pulses generated per 7.5 min.	In line with equipment it is connected to.
Indoor temperature and relative humidity (RH)	Ultra-RF for temperature and relative humidity (ref: QC0160)	A band-gap temperature sensor and capacitive RH sensor.	Readings recorded every 3–5 min. Update rate every 5 min.	±0.5 °C
Indoor VOCs and eCO_2	LoRa VOC Transmitter	A low-power digital gas sensor solution, which integrates a gas sensor solution for detecting low levels of VOCs. Sensor also provides eCO_2 level derived from the VOC levels.	Readings recorded every 3–5 min. Update rate every 5 min.	VOCs detected: Alcohols, Aldehydes, Ketones, Organic Acids, Amines, Aliphatic and Aromatic Hydrocarbons
Indoor PM2.5	LoRa PM2.5 - PM10 Transmitter	A laser-based sensor, using the light scattering method to detect and count particles in the concentration range of 0 μg/m^3 to 1,000 μg/m^3.	Readings recorded every 3–5 min. Update rate every 5 min.	0 μg/m3 to 100 μg/m3: ±15 μg/m3 100 μg/m3 to 1000 μg/m3: ±15%
Outdoor environment	Invisible Systems External AQ Monitor	As the indoor temperature, relative humidity, VOCs and PM2.5 sensors. The sensor also provides measures of PM10 using the same technology as for PM2.5.	Readings recorded every 30 min.	±0.3 °C
Gateway	Boxed ISL058 gateway			RF transmitters connect to sensors and send data to the infrastructure internet connected Gateway on site. The Gateway stores and transfers the data to the Realtime Online cloud server over a secure mobile cellular connection.

to 22 water and 41 gas sensors (see Section 6), so the second-stage clustering also allowed a more appropriate number of clusters to be chosen.

For the external sensors, the first stage was conducted on 291 homes in November 2017 for the placement of 30 sensors. In February 2019, an additional 30 sensors became available. However, at this time, 103 households from the first stage had left the project, while 95 households had joined. To determine sites for the placing the additional sensors, we conducted a second stage to repeat the analysis on the updated cohort of 283 homes.

4.1. Factors

Factors considered relevant for affecting water and gas use were as follows. Previous research has shown that these factors can affect water usage [103,104] and gas usage [105,106].

1. Property-type (flat or house).
2. Property-size (number of bedrooms or rooms).
3. Property-age (years).
4. The time spent inside the home (two factors in the first-stage analysis).

Number of occupants per age-group.

5 0–12 years.
6 13–17 years.
7 18–65 years.
8. 66+ years.

Property-size was represented by number of bedrooms in the first-stage analysis and by number of rooms in the second stage. Time spent occupying the home was represented by numbers of occupants in part-time and full-time employment in the first stage, and by survey responses about the number of hours typically spent inside the home in the second stage.

The factors considered relevant for affecting external air measurements were as follows.

1. Latitude.
2. Longitude.
3. Digital Terrain Model data [DTM; 107] as a measure of elevation above sea level.
4. The difference between Digital Surface Model data [DSM; 108, 109][1] and DTM data as a measure of surrounding cover (e.g., trees, buildings).
5. The distance from the nearest A-road.

It seems unlikely that there will be sufficient variation in latitude and longitude within the area project location to reflect global climate differences. However, latitude and longitude were used in order to capture the relative location of the property within the local context. Variation in latitude and longitude are likely to reflect other underlying groupings or variables that are unknown in advance, such as differences in surroundings that affect wind speed or local microclimates. Latitude and longitude therefore allow sensors to be distributed over the project area and capture differences in local surroundings. We also included other factors that may be more likely to directly affect influences on environmental measures. Previous research has shown influences of elevation on temperature and, in part, on wind speed [111],[2] whilst surrounding cover or shelter is also likely to affect temperature and humidity by mediating wind speed. Finally, distance to the nearest main road can affect air quality [112].

[1] DTM and DSM data freely obtained from the Natural Environment Research Council (Centre for Ecology & Hydrology; British Antarctic Survey; British Geological Survey), converted from ASC to CSV format using PyLidar [110].

[2] The findings were made with larger and higher elevations than in the current study.

4.2. Determining factor values

The factors for the utility-sensor placement analysis (property-type, bedrooms, property-age, part- and full-time employed counts, and age-group counts) were determined using Coastline Housing data in the first stage. In the second stage, rooms (previously bedrooms), time at home (previously employment counts) and age-group counts were replaced by participant survey responses.

Latitude, longitude, Easting and Northing of the properties were determined from the postal code [113]. The DTM and DSM values were calculated as the mean average of the 1-m resolution data across a 9 m × 9 m square surrounding the property's Easting and Northing. Distance to the nearest A-road was measured in metres for each latitude and longitude pair using the distance tool on Google Maps [114].

4.3. Pre-processing

Factors in a cluster analysis can carry different weightings in influencing the clustering calculations due to differences in variance and magnitudes of the factor values. Transformation into *z*-scores standardises values to the same mean and standard deviation, and are calculated by subtracting the mean of the values and dividing by the standard deviation. All factors were transformed into *z*-scores to ensure similar variances and therefore similar weights in the cluster analysis. For utility-sensor factors, given small numbers and high skew in some counts of people, and to maintain relative magnitudes across those count data, we used the standard deviation and mean across all people-count factors to calculate *z*-scores.

4.4. Correlations

In cluster analysis, correlations between factors can be problematic if the factors are representing the same underlying characteristic. A characteristic captured by multiple factors contributes more influence on the cluster process than characteristics that are only captured by one factor. The correlation between each pair of factors was therefore checked. We calculated correlations using values from all homes available.

Figs. 2 and 3 present the correlation coefficients, and show high correlations between some factors for both the utility-sensor and the external-sensor analyses. However, while some of the bases for the factors overlap, it was decided that each of the factors also brings its own quality. For example, number of full-time employed correlates with the number aged 18–65, but the first provides information about daytime occupancy while the second gives the number of adults, both of which may influence utility usage for different reasons.

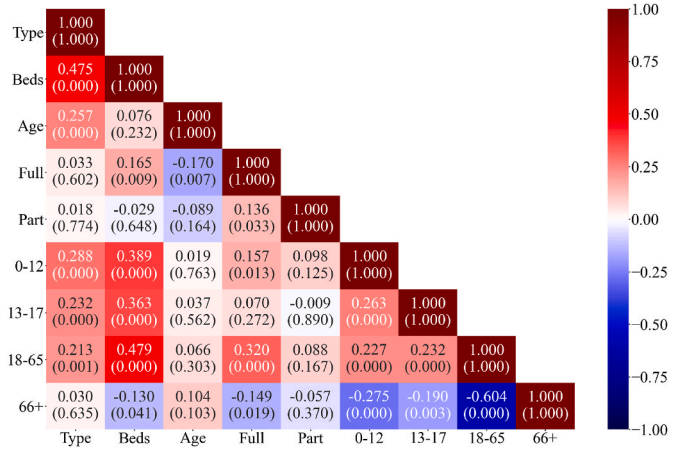


Fig. 2a. Utility-sensor factors, first stage. Pearson correlation coefficients (and p-values) between pairs of utility-sensor factors used in the first stage with 248 homes. The coefficient scale is from −1 (blue) to +1 (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

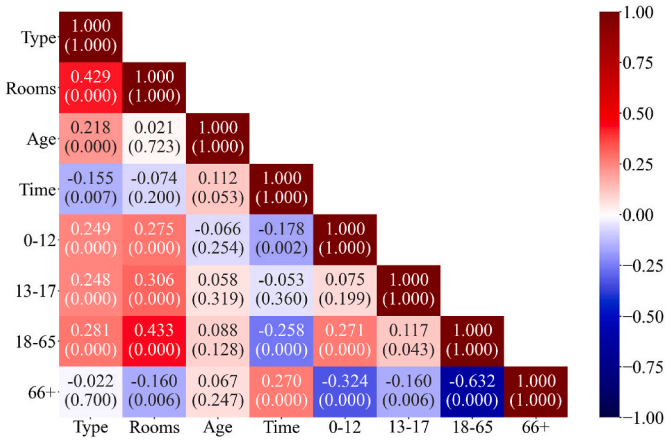


Fig. 2b. Utility-sensor factors, second stage. Pearson correlation coefficients (and p-values) between pairs of utility-sensor factors used in the second stage with 298 homes. The coefficient scale is from −1 (blue) to +1 (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

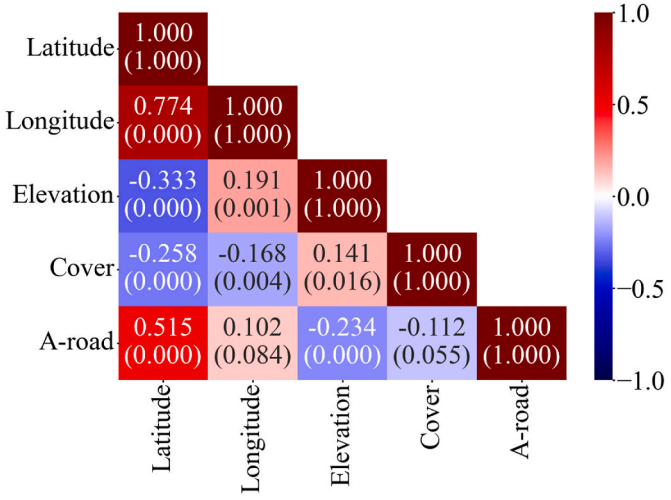


Fig. 3a. External-sensor factors, first stage. Pearson correlation coefficients (and p-values) between pairs of external-sensor factors used in the first stage with 291 homes. The coefficient scale is from −1 (blue) to +1 (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

4.5. K-means clustering

We employed a k-means clustering technique, in which the distance between potential sensor sites and the cluster centre (i.e., centroid) is minimised by iteratively updating the membership of the clusters according to the closest centroid then recalculating the centroid location as the mean of the cluster members [62,115].

K-means was chosen above other clustering methods due to the numerical nature of the factors, and its applicability to this dataset [98]. In this study, other benefits of using k-means are to provide a generic and accessible approach for application in other settings. It is probably the most widely used clustering approach, with extensive documentation, tutorials and tools for implementation, and is applicable to any set of numerical factors.

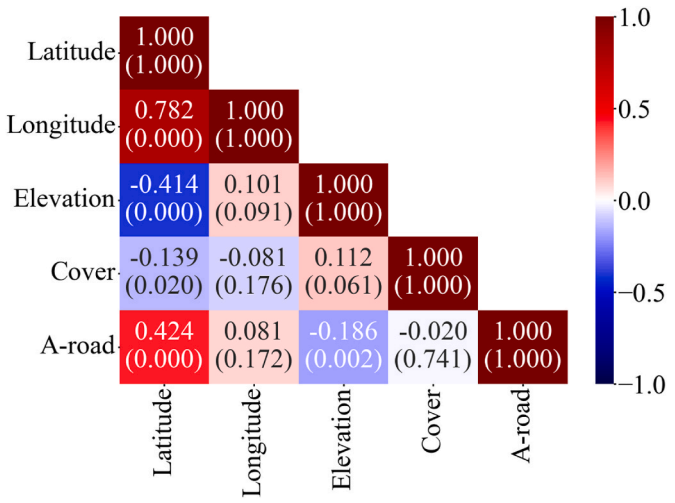


Fig. 3b. External-sensor factors, second stage. Pearson correlation coefficients (and p-values) between pairs of external-sensor factors used in the second stage with 283 homes. The coefficient scale is from −1 (blue) to +1 (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Latitude and longitude data represent fewer than half of the factors in one of the applications we are testing. However, it is worth noting that k-means clustering is sometimes avoided for latitude and longitude data. Distortion can occur due to changes in the distance between longitudes with the curvature of the earth, and locations can be overrepresented or underrepresented due to random selection of starting centroids from the dataset [116]. However, in our study, earth curvature is minimal, and it is advantageous to capture data-driven weightings to ensure representative coverage of our participant homes and locations, rather than obtain uniform coverage of a predefined space.

The method was implemented in Python [version 3.6; [117,118]][3]using the k-means module [Scikit-learn version 0.19.1; [124]] with the triangle inequality option for speed efficiency [125].

The k-means algorithm was set to create 50 models with different initial seeds for centroid locations and return the model with the lowest resulting inertia, which is the sum of squared distances between the homes and the centroid within each cluster. Each model converged when the relative change in the inertia was less than 0.0001 between iterations. Models were created for 5 to 25 clusters to determine the number of clusters (k) that provides the most appropriate solution. The whole process was conducted ten times (hereafter called run 1 to 10) and selection of the final solution was guided by consistency of solutions across different runs, which reflects less susceptibility to noise.

For the two stages of the utility-sensor analysis, factors differed across the two. We therefore conducted two separate cluster analyses, providing two independent cluster solutions. Locations of sensors, which were placed according to the first stage, were verified in the context of the second-stage clusters.

Across the two stages of the external-sensor analysis, the factors remained consistent. We could have therefore added the new homes to the cluster solution from the first stage. However, we instead conducted the second-stage analysis on the complete set of 283 homes independently of the first stage because some households had withdrawn. In addition, 30 additional sensors were still to be placed, so we verified the location of existing sensors, and compensated for any lack of cluster coverage by the placement of a new sensor.

For both the utility-sensor and external-sensor analyses, the classification of homes by cluster were validated, and the overlap between first- and second-stage solutions was quantified to ascertain a consistent structure between the two stages. The supplementary material provides details.

5. Resulting clusters

This section describes the results of the clustering analyses. Measures of fit are presented, which were used to select the most appropriate cluster solutions.

5.1. Measures of fit: inertia and silhouette

Two measures of fit were plotted across the different numbers of clusters. These plots were used to visually determine the point(s) at which a so-called elbow occurs, reflecting a change in the rate of change of measure. Figs. 4 and 5 provide plots for the run that was ultimately chosen as containing the solution for each analysis.

The first measure used was the inertia, defined earlier [e.g., see Ref. [126]]. The second was the silhouette [127], which is an inverse measure of overlap of the clusters. It is calculated as the normalised difference between the mean distance between members within a cluster and the mean distance between those cluster members and the members in the nearest other cluster.

5.2. Selecting a solution

Plots for all ten runs revealed a change in rate for each measure at $k = 9$ and $k = 6$ for the first and second stages of the utility-sensor analysis respectively, and at $k = 11$ and $k = 7$ for the first and second stages of the external-sensor analysis respectively. In all cases, except the second stage of the utility-sensor analysis, the expected number of sensors supported a larger number of clusters than that identified. In addition, inertia and silhouette generally improve with more clusters. It was therefore decided to choose larger numbers of clusters than indicated by these initial solutions.

For the first stage of the utility-sensor analysis, $k = 16$ was indicated by two of the ten runs, while other values of $k > 9$ were indicated by one run at most. Of those two runs, one solution gave a cluster with only two members, which was considered too small for our sampling purposes, so the other run was chosen. This run had the second largest silhouette and the fourth lowest inertia of all runs for $k = 16$.

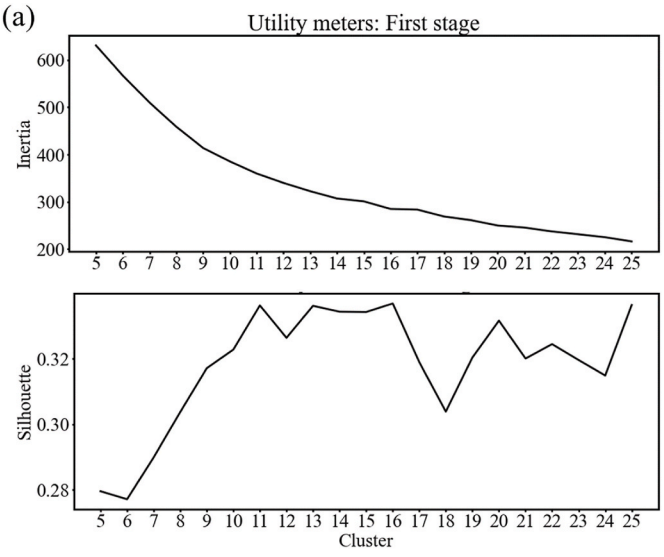


Fig. 4a. Utility-sensor analysis inertia (upper panel) and silhouette (lower panel) as a function of the number of clusters for the first stage.

[3] with NumPy 1.13.3 [119,120], SciPy 0.19.1 [121], Pandas 0.20.3 [122], and Matplotlib 2.1.0 [123].

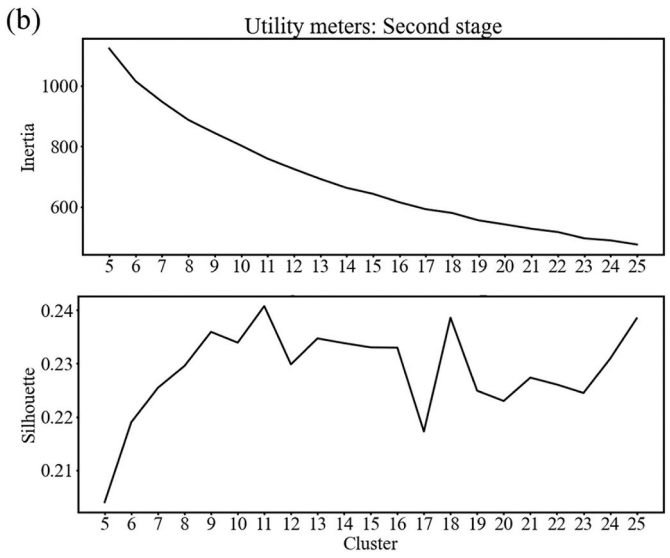


Fig. 4b. Utility-sensor analysis inertia (upper panel) and silhouette (lower panel) as a function of the number of clusters for the second stage.

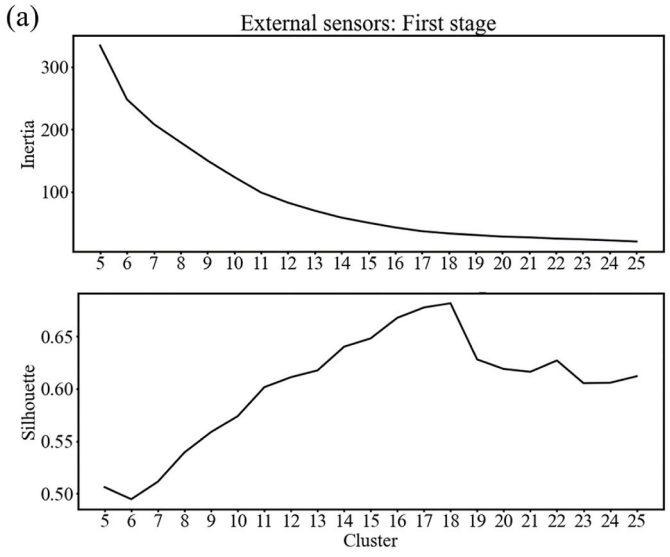


Fig. 5a. External-sensor analysis inertia (upper panel) and silhouette (lower panel) as a function of the number of clusters for the first stage.

Fewer sensors were placed than originally planned (22 for water and 41 for gas). Therefore, in the second-stage analysis, $k = 6$ was chosen to allow at least two sensors per cluster. Four runs gave the lowest inertia, and the third highest silhouette. All four provided identical cluster membership.

For the first stage of the external-sensor analysis, all runs indicated visual elbows in the rate of change for inertia at $k = 17$ and for silhouette at $k = 14$ and $k = 17$. Given a maximum of 30 sensors available at this first stage, the solution with 14 clusters was chosen to allow for more sensors per cluster. All runs gave identical solutions for $k = 14$.

For the second stage of the external-sensor analysis, 30 additional sensors were available, providing a maximum of 60 sensors. We therefore decided to use 15 or more clusters. Elbows were indicated at $k = 15$ in inertia for three runs and in silhouette for five runs, with little consistency for values of $k > 15$. One run indicated $k = 15$ in the elbow for both inertia and silhouette, and the provided lowest inertia and highest silhouette across all runs. This cluster solution was therefore chosen to make recommendations for placing the additional sensors.

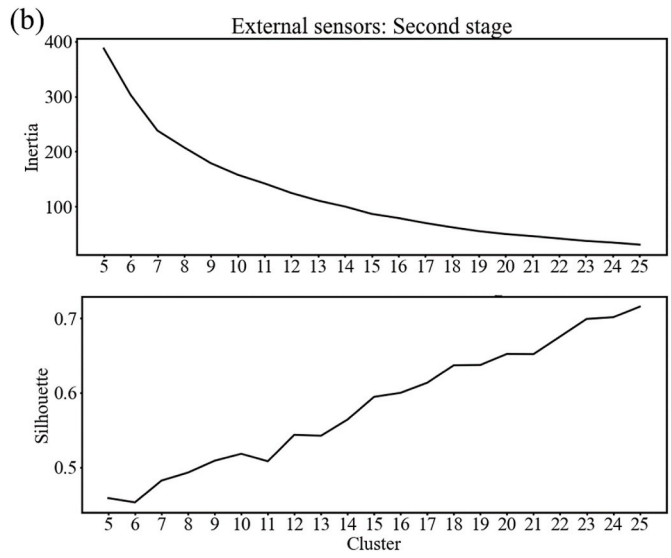


Fig. 5b. External-sensor analysis inertia (upper panel) and silhouette (lower panel) as a function of the number of clusters for the second stage.

6. Sensor placements

Given the clustering solutions selected, homes within each cluster were ordered by Euclidian distance from their cluster centroid to provide recommendations for the placing of sensors at representative sites. However, there were installation restrictions for many of the sites that were recommended following the first-stage cluster analyses, as detailed below. In addition, some households withdrew from the project between the cluster analyses and the installation of the sensors. In all cases, if a recommended site was not available, the next site in terms of distance from the centroid was instead recommended.

Figs. 6 and 7 show the clusters, with each cluster represented by a different colour. Each dot represents one home, with a line connecting to its cluster centroid. Open circles indicate the sites suggested for sensor placement, and open squares represent sites at which sensors were placed. Given nine (first-stage) or eight (second-stage) factors for the utility-sensor cluster analysis and five factors for the environmental sensors, only two factors are plotted for visual clarity. In Fig. 6, principal components of the cluster factors are used [Scikit-learn PCA; [128]]. However, Euclidian distance of homes from the centroid was determined using all cluster factors. In Fig. 7, to ensure anonymity and to increase visual clarity, jitter is applied to individual homes.

Based on the first-stage clustering of the utility sensors, the recommended choice was for two sites close to the centroid and one site distant from the centroid, giving 16 clusters * 3 sites = 48 sensors.

Installation of gas sensors was restricted by some homes having a gas meter that does not produce a pulse output, making it unsuitable for the pulse sensor to be used. Installation of water sensors was also restricted due to some participants preferring to avoid the installation disruption (e.g., having a hole cut into the back of the kitchen cupboard). Of the 48 recommended homes, no gas homes were suitable, and 5 water were installed as planned. When a home was not available, the next home in terms of distance from the centroid was instead approached for install.

22 water sensors and 41 gas sensors were successfully installed. However, 11 of the gas sensors were placed in homes not in the first-stage clustering due to being recently recruited participants. For both water and gas, there were clusters without any sensors placed due to the installation restrictions. See first-stage clusters without open squares in Fig. 6 (water cluster numbers 6, 10, 13, 14, 15; gas cluster numbers 0, 6, 8, 13, 14, 15).

For second-stage clusters, some households with sensors installed following the first stage had withdrawn from the project or had missing

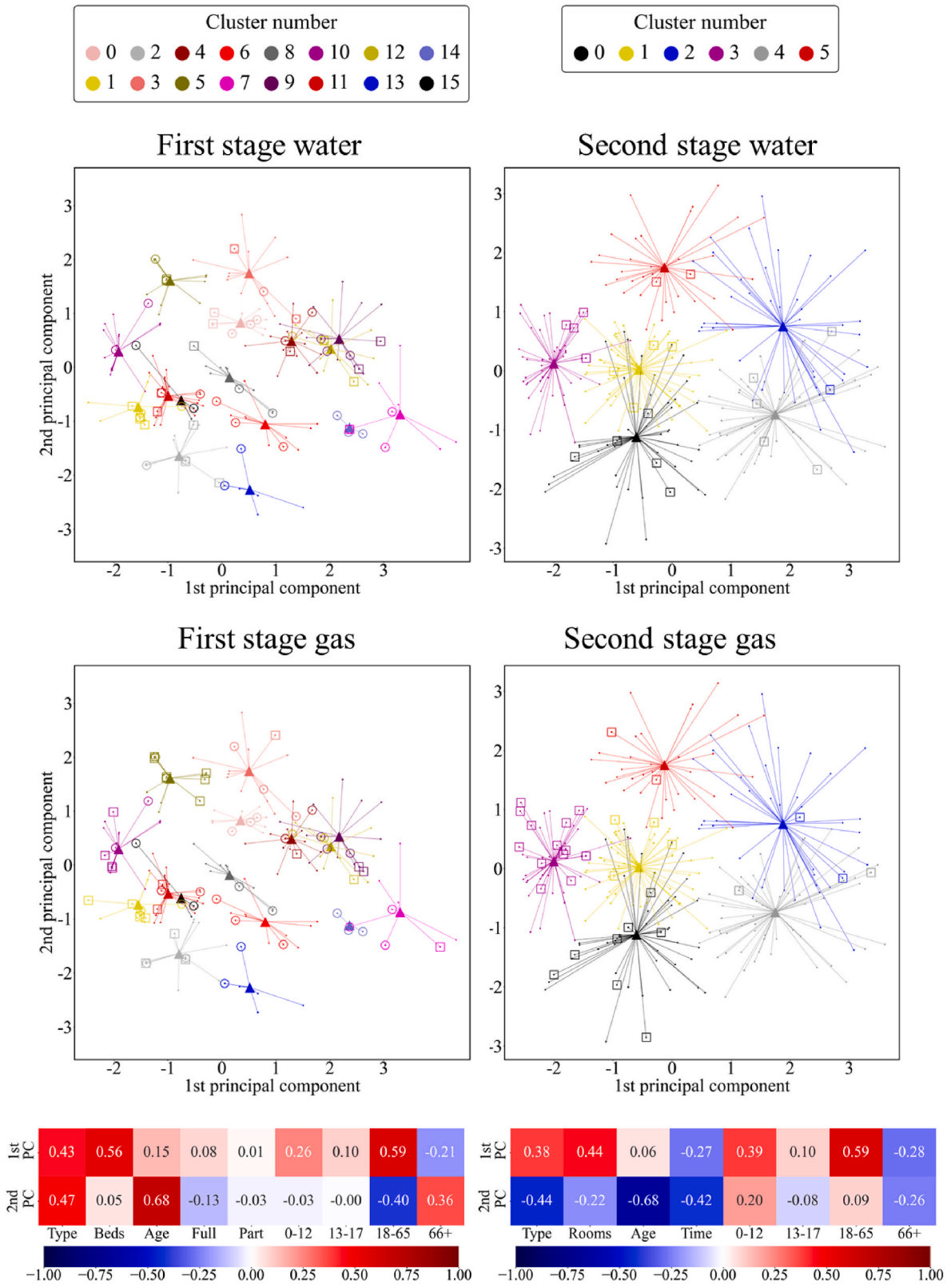


Fig. 6. Utility-sensor clusters for the first (right panels) and second (left panels) stages, for water (upper panels) and gas (middle panels), and coefficient weight for "1st" and "2nd" principal components (PC) for each cluster factor (lower panels). See text for details.

survey data, leaving 21 homes with water sensors and 32 with gas sensors. Given all sensors had already been installed following the first-stage analysis, no recommendations were made following the second stage.

Across the final six clusters, there were 1–5 water sensors per cluster, with four clusters having at least four sensors. The water-sensor home that was closest to the cluster centroid was positioned as follows. In two clusters, it was at the closest home to the centroid, and in three it was

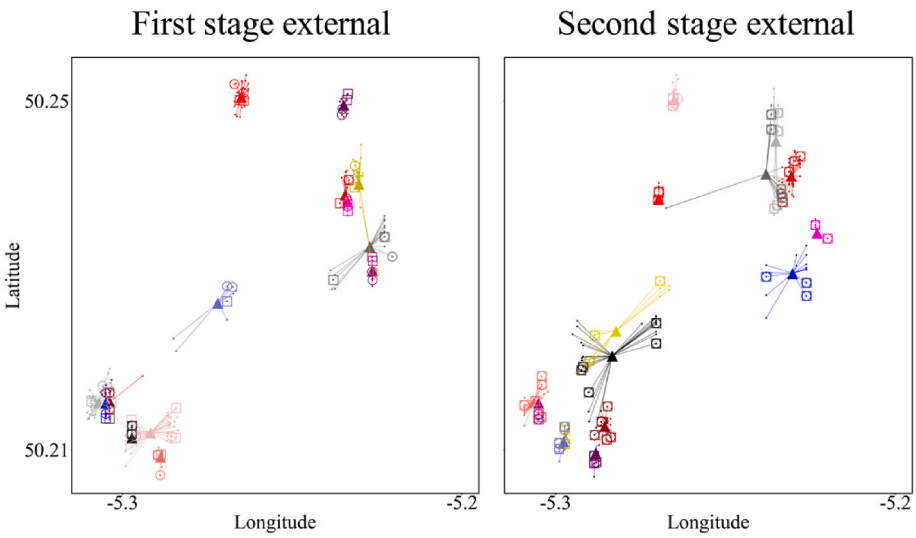


Fig. 7. External-sensor clusters for the first (left panel) and second (right panel) stages. See text for details.

tenth closest at most. In one cluster (number 3), with 49 unique members and four water sensors, it was at the 36th home from the centroid. There were 2–14 gas sensors per cluster. In four clusters, a gas sensor was within eight homes of the centroid, and in the remaining two clusters (numbers 2 and 4) it was 27th.

For the first stage of the external sensors, the recommended choice was for one site close to the centroid and one distant from the centroid, giving 14 clusters * 2 sites = 28 sensors. The installation of external sensors was restricted by requiring a suitable mount for the sensor. Of the 28 recommended sites, 7 had sensors installed. As for gas and water, when a site was not available, the next site in terms of distance from the centroid was instead considered. Of the 30 sensors planned, 27 sensors were successfully installed including three installed at sites not in the first-stage cluster analysis.

Recommendations for second-stage placements were made under constraints from the 27 sensors that had already been placed. Two of the 15 clusters contained no existing external sensors, and five had only one. Distribution of the additional 30 sensors was recommended across the clusters to give 2–6 sensors per cluster. Twenty-eight additional sensors were successfully placed, achieving 2–6 sensors per cluster, and 55 sensors altogether. Of the 55 sites recommended, sensors were placed at 53, with two of the homes each having two sensors at different orientations. Four clusters had only two sensors. Three of these clusters contained at most eight sites. In the other, all sites in the cluster were located on adjacent roads. In 14 of the 15 clusters, an external sensor was positioned at the closest distance to the cluster centroid, and in the remaining cluster it was placed at the second closest distance.

7. Clustering verification

In this section, we verify the appropriateness of the clustering methodology to provide a representative sample of sensor sites from the full set of potential sites. If the clusters created are successful in capturing a range of sensor sites, then we would expect the resulting sensor data to vary across clusters, and vary with the factors used to define the clusters. Such variation would suggest that the clusters are meaningful with respect to the sensor data being collected.

Overall, we wanted to test for differences in sensor data between clusters. Water, gas and external sensors were deliberately distributed across clusters, resulting in limited numbers per cluster, and providing limited statistical power for the effects of cluster. However, other related measures across all homes can be used to test for an effect of cluster. First, we test for relationships between measures that could be related, for example between internal and external temperature, and between electricity and gas usage. Second, we test for differences between clusters for these related measures, which should occur if the clusters are meaningful for these types of measures. Thirdly, we test for relationships between the cluster factors and the data from water, gas and external sensors.

For each sensor, readings were used from 1st November 2018 to 31st October 2019, taking the mean hourly usage for utilities and the mean average reading for environmental measures. To allow for any differences in the interval between the readings, all readings were interpolated to a resolution of 1 min before means were calculated. For utilities, the usage was summed for each hour before calculating the mean hourly usage. Sensors were excluded if readings did not span the date-range, and outliers were also excluded. For utilities, values were excluded that were more than 12 standard deviations from the mean and dates were excluded if zero usage was recorded. Nine gas sensors were excluded for a mean hourly usage below 0.01 m^3 or for zero usage on 25% of days or more. 32 electricity sensors were excluded for a mean hourly usage below 0.08 kWh or for a visually abrupt change in usage during the date-range.

Most external sensors did not capture data during the full year. Therefore, to maximise the number of valid external sensors, we used two restricted date-ranges: Winter, from 1st December 2018 to 28th February 2019, and summer, from 1st May 2019 to 31st July 2019. Table 2 provides the final numbers of sensors for each type. Table 3 provides descriptive statistics for the factors used for the cluster analysis.

7.1. Correlations between related sensor measurements

The factors used for the utility-sensor cluster analysis should also influence electricity usage. We therefore tested for correlations between

Table 2
Number of sensors with complete data for each sensor type.

	N
Bedroom temperature and RH	237
Living room temperature and RH	231
VOC	150
eCO_2	150
PM2.5	180
Electricity	110
Water	19
Gas	13
External temperature, RH, VOC, eCO_2, PM2.5, PM10 winter	15
External temperature, RH, VOC, eCO_2, PM2.5, PM10 summer	24

Table 3
Descriptive statistics for each clustering factor for homes with water, gas, and external sensors.

Clustering factor	Water (N = 19) Mean (SD)	Min, max	Gas (N = 13) Mean (SD)	Min, max
Type (0 = flat)	0.53 (0.51)	0, 1	0.38 (0.51)	0, 1
Rooms (count)	6.26 (1.28)	4, 9	5.31 (1.18)	4, 8
Age (years)	41.21 (25.99)	6, 68	37.08 (25.13)	6, 66
Time (hours)	19.55 (3.02)	13, 24	20.79 (2.58)	15, 24
0-12 (count)	0.21 (0.54)	0, 2	0.00 (0.00)	0, 0
13-17 (count)	0.11 (0.32)	0, 1	0.00 (0.00)	0, 0
18-65 (count)	1.00 (0.94)	0, 3	0.38 (0.51)	0, 1
66+ (count)	0.52 (0.70)	0, 2	1.00 (0.58)	0, 2

Clustering factor	External sensors winter (N = 15) Mean (SD)	Min, max	External sensors summer (N = 24) Mean (SD)	Min, max
Latitude (°)	50.227 (0.015)	50.209, 50.250	50.227 (0.014)	50.210, 50.250
Longitude (°)	−5.264 (0.035)	−5.309, −5.224	−5.266 (0.033)	−5.309, −5.220
Elevation (metres)	104.915 (20.219)	81.654, 153.103	106.308 (19.349)	79.476, 158.651
Cover (metres)	4.781 (2.987)	0.036, 13.429	3.935 (3.235)	0.002, 13.429
A-road (metres)	451.1 (397.1)	16, 1680	374.4 (332.4)	16, 1680

water and gas usages and electricity usage in homes with the relevant utility sensors in place. Fig. 8 shows significant positive relationships between utility usages.

For the external sensors, the external air could influence the air internal to the home. We tested for correlations between external-internal pairs of measures using only those homes at which external sensors were installed. Relative humidity is dependent on air temperature because it determines the maximum amount of water that air can hold. To compare external and internal humidity, we therefore compared absolute humidity as well as RH. Absolute humidity was calculated for each reading from temperature (T) and RH sensors as [129]:

$$\frac{6.112 * e^{\frac{17.67*T}{243.5+T}} * RH * 2.1674}{273.15 + T}$$

There were no significant correlations between external and indoor measures. See Fig. 9.

There are significant correlations between the electricity and other utility usages, and there is reason to expect that the factors used for the water and gas cluster analysis could also affect electricity. There were no correlations between the internal and external environmental measures. However, the lack of significant relationships could reflect the complexity of the indoor environment, as discussed in the Discussion section below.

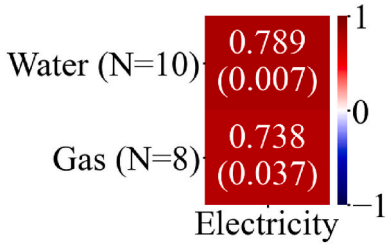


Fig. 8. Utility usages. Pearson correlation coefficients (and p-values) between water, gas and electricity usages. The coefficient scale is from −1 (blue) to +1 (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

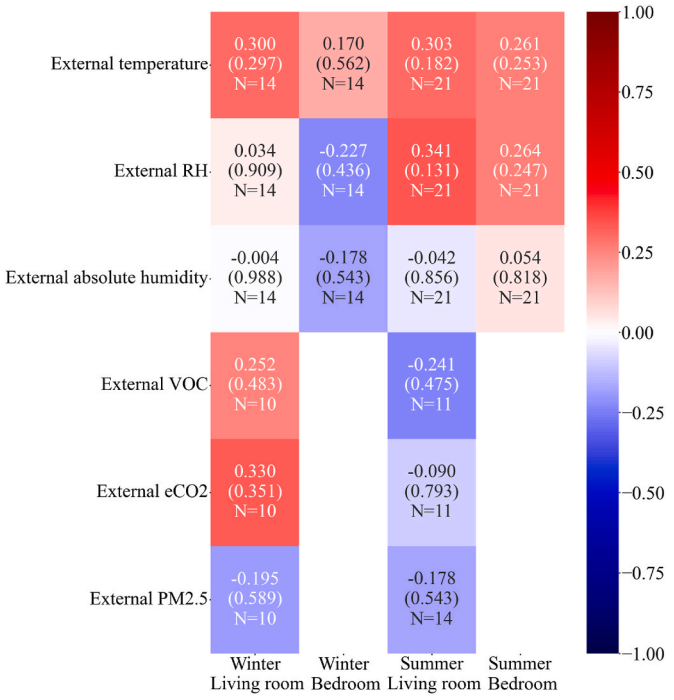


Fig. 9. Environmental measures. Pearson correlation coefficients (and p-values) between external and internal values during winter and summer. The coefficient scale is from −1 (blue) to +1 (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

7.2. Relationships between clusters and sensor data

In this section, we wish to test whether sensor readings vary across clusters. To have sufficient statistical power we used the measures that have large numbers of sensors in each cluster, namely the electricity and the indoor environmental measures.

Each household was assigned to the cluster with the nearest centroid for each of the utility-sensor clustering and the external-sensor clustering, both using the second-stage clusters.

For the utility-sensor clusters, an effect of cluster for the electricity usage would suggest that the correlated water and gas measures also vary across cluster.

For the external-sensor clusters we test for an effect of cluster for each internal environmental measure. Despite no significant correlations between the indoor and external environmental measures, an effect of cluster would suggest that the factors used for the external-sensor clustering do influence environmental conditions.

Data were analysed using a one-way *k*-level (number of clusters) ANOVA for each dependent measure. When Bartlett's test for equal variances was violated ($p < 0.05$), a Kruskal-Wallis test was performed instead.

For the utility-sensor clusters, there was a significant effect of cluster for electricity usage, N = 110, $\chi 2(5) = 29.979$, $p < 0.001$. For the

Table 4
F for ANOVA and χ^2 for Kruskal-Wallis (and p-values), for the effect of external-sensor cluster for each of the internal-sensor measures.

Temperature living room	N = 231 $\chi 2(14)$	23.828 (0.048)
RH living room	N = 231 F(14, 216)	2.72 (0.001)
Temperature bedroom	N = 237 F(14, 222)	1.51 (0.110)
RH bedroom	N = 237 F(14, 222)	1.83 (0.035)
VOC	N = 150 F(14, 135)	1.07 (0.386)
eCO_2	N = 150 F(14, 135)	1.09 (0.372)
PM2.5	N = 180 $\chi 2(14)$	20.769 (0.108)

external-sensor clusters, effects of cluster are provided in Table 4. There were significant effects of cluster for temperature in the living room and RH in both rooms.

The analyses have revealed effects of cluster, but interpretation of influences on the differences is limited given that the cluster factors are not represented in the analysis. Linear regressions were therefore conducted to establish which cluster factors were predictors of the sensor readings. Regressions were only conducted for those measures that revealed an effect of cluster. Variance inflation factors for all predictor variables in all regressions were below 4, except latitude (13.3–13.7) and longitude (10.0–10.2). The plots of residuals against fitted values showed no evidence of heteroscedasticity for all regression models.

The utility-sensor cluster factors significantly predicted the electricity usage overall ($F(8, 101) = 12.29$, $p < 0.001$). Electricity usage increased with all age-group counts (coefficients = 0.070 to 0.115, $ps \leq 0.002$), and there was a trend towards property-age being a significant predictor (coefficient = 0.001, $p = 0.070$). There were no other significant predictors (all $p > 0.156$).

The external-sensor cluster factors were used as predictors in three separate regressions to predict temperature in the living room and RH in each room,. The models for RH in the bedroom was not significantly different from the null model with no predictors ($F(5, 231) = 1.45$, $p = 0.207$). The regression models for temperature and RH from the living room exhibited a significant overall relationship between the predictors and the outcome (both $F(5, 225) \geq 2.76$, $p < 0.020$). Temperature showed a strong trend towards increasing further south (latitude coefficient = −54.113, $p = 0.055$) and increased with cover (coefficient = 0.104, $p = 0.025$). RH increased further north (latitude coefficient = 225.270, $p = 0.025$) and further west (longitude coefficient = −95.977, $p = 0.023$), and decreased with cover (coefficient = −0.369, $p = 0.025$).

7.3. Relationships between cluster factors and utility-sensor and external-sensor data

In this final verification, we test for variation of the sensor data with the factors used for clustering. We used the cluster factors in regressions as predictors for the measures from the sensors that were placed as a result of the clustering process. Sensors were purposely distributed across clusters to capture data from a range of homes. An ANOVA was not therefore appropriate given small numbers per cluster (after outlier removal: 0 to 6). Separate regressions were performed for homes with water sensors, homes with gas sensors, and the external sensors.

For homes with gas sensors, there were no occupants in the 0–12 and 13–17 age-groups, and the variation inflation factors were high for age-group 66+ (13.0), property-age (34.2) and number of rooms (12.5), we therefore removed property-age from the predictors and summed the number of occupants into a single measure.

Variance inflation factors for all predictor variables in all regressions were below 4, except for latitude (15.67 or 15.84) and longitude (8.54). The plots of residuals against fitted values showed no evidence of heteroscedasticity for all regression models.

Water usage was successfully predicted by the model, while the gas usage model only revealed a trend towards significance over the null model with no predictors ($F(8, 10) = 7.51$ $p = 0.002$ and $F(4, 8) = 2.99$ $p = 0.087$ respectively).

Fig. 10 provides the regression coefficients and p-values.[4] Water usage increased with the numbers of 13–17 and 66+ year-olds, and the strongest predictor for gas usage was number of rooms. There were also multiple trends towards significant predictors for water. Effects of factors differed between utilities, suggesting that usages of all utilities are not necessarily affected by the same factors.

For the external sensors, each home was assigned the readings from the nearest external sensor. Only homes that had a unique set of cluster factor values were included in the analysis, giving 38 homes for winter and 50 for summer. All measures were successfully predicted by the respective model in both winter and summer, except VOCs in winter, which was not significant, and summer particulate matter, which showed trends towards significance. Figs. 11–13 provide the Fs, regression coefficients and p-values.

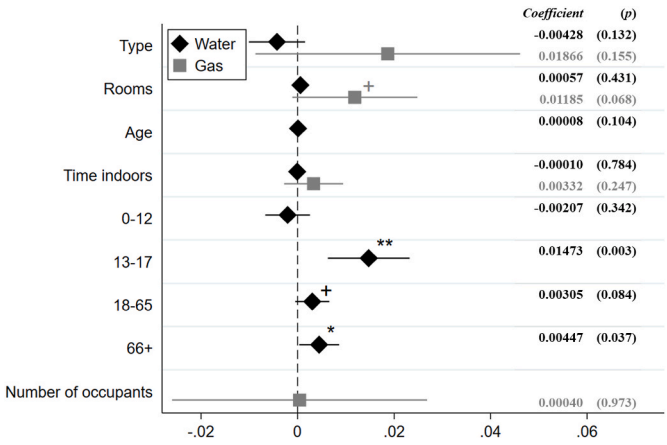


Fig. 10. Coefficients (and p-values[5]) for the predictors of water (N = 19) and gas (N = 13) usage.

In winter, temperature increased further south, further east and with distance from an A-road, and decreased with elevation. In summer, temperature also decreased with elevation. RH in winter showed patterns opposite to temperature, increasing further north and further west, and with elevation. In summer, there was a trend towards an increase in RH with elevation.

In summer, VOCs increased further west, and decreased with surrounding cover and with distance from an A-road. In winter, eCO_2 increased further west and with elevation,. In summer, eCO_2 also decreased with distance from an A-road. Particulate matter (PM) increased further north in both winter and summer, and further west in winter. PM levels increased with elevation in both winter and summer. PM decreased with distance from an A-road, although these relationships only revealed a trend towards significance in summer. These results demonstrate that the external measures have relationships with the factors used to define the clusters.

8. Discussion

We present a methodology to achieve representative sampling for placement of a limited number of sensors. The methodology used cluster analysis to segment potential sensor sites into similar groups, and then selected sites from across those groups to provide a representative sample over the variety of potential sites. Meaningfulness of the clusters with respect to the sensors was verified using the sensor data collected. These verification results provide evidence that the clusters exhibit differences across sensor data, and that the cluster factors selected for clustering the potential sensor sites have relationships with relevant sensor measurements.

The aim of this study was to develop, implement and test the cluster-based methodology. While not a direct aim of the study, it is first interesting to discuss and interpret the relationships found in the verification analyses. We then outline the pitfalls revealed by our study. The overall aim of the study is then considered, and finally future examples for use of the sensor data are presented.

We tested for correlations between the different types of sensor data. For utility usage, significant correlations were observed between electricity and water usages and between electricity and gas usages. These relationships are to be expected given that water and gas usage are

[4] Graphs in Figs. 10–13 were created using CoefPlot [130].

[5] Significance $^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

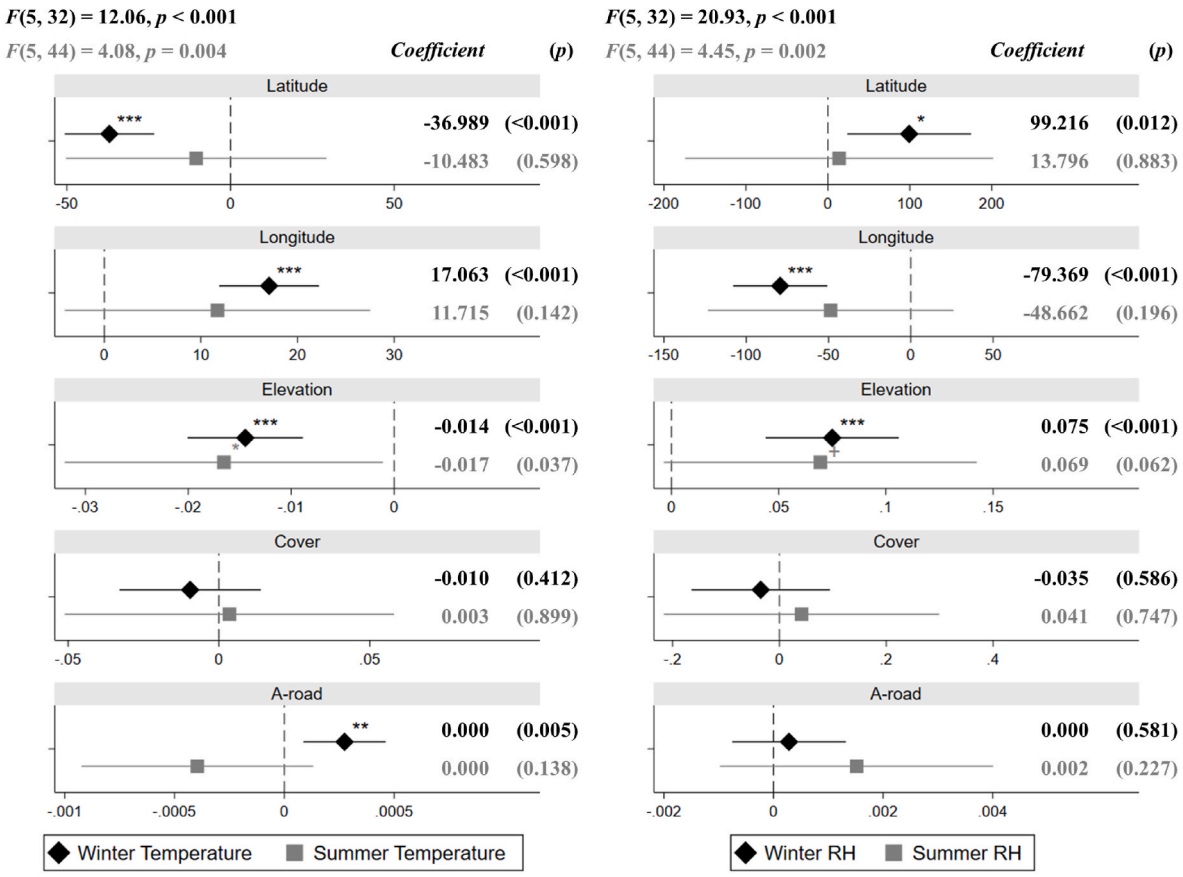


Fig. 11. Coefficients (and p-values[5]) for the predictors of temperature and RH in winter (N = 38) and summer (N = 50).

influenced by similar factors as those found to influence electricity usage [131].

No significant correlations between indoor and external air measures were observed. The lack of relationships is perhaps not surprising given that the indoor environment results from complex interaction of the built environment (e.g., building thermal properties, permeability, solar exposure, insulation levels) and human behaviours (occupancy, heating, ventilation)[6] [e.g., Ref. [132]]. We might have expected stronger relationships during the summer, reflecting more window-opening behaviour in response to warm weather. However, during the winter, it seems more likely that the indoor environment would arise from internal sources due to lack of ventilation and use of heating.[6] In addition, correlations between the indoor and outdoor environments can be lagged, such that internal changes follow external changes, with external air quality levels and window-opening behaviours mediating that relationship [133]. Such dependencies may have been missed in the current analyses given correlations were assessed using mean average values rather than between individual streams of sensor data.

Our verification analyses revealed significant relationships between the factors used to create the clusters and the sensor data. We found that water usage is related to the number of teenagers in the home [104], while gas usage is weakly related to number of rooms [106]. The relationships we observed are in line with previous work showing relationships between energy usage and household size, but not with time spent at home [81]. These relationships between the cluster factors and sensor readings demonstrate that the factors are meaningful with respect to water and gas usage, and provide evidence that the clustering methodology for sampling achieved its purpose.

For the external sensors, latitude and longitude were included to distribute sensors across the locality and capture local variation in environmental influences that were unknown (e.g., microclimates). The analyses showed significant relationships between some external environmental measures and latitude or longitude, despite other potentially stronger influences from the surrounding environment (e.g., sensor orientation).

Increased elevation was associated with decreased external temperature and increased RH, eCO_2, and particulate matter, with stronger patterns in winter than summer. Cover showed no significant relationships with any external environmental measure, except decreased VOCs in the summer. Increased distance from a main road was associated with decreases in VOCs and particulate matter, in line with previous findings [112].

Despite the lack of correlations between indoor and external environmental measures, three indoor measures did exhibit differences across the clusters, and the factors used to create the clusters did show significant relationships with the external sensor data. These results are consistent with meaningful factors and clusterings for informing the deployment of the external sensors.

The process highlights two main pitfalls and subsequent recommendations for sampling using cluster analysis. First, the installation of utility sensors was limited such that some recommendations could not be implemented, resulting in some clusters with no sensors installed. This sparsity was resolved by repetition of the cluster analysis. However, we recommend that the cluster analysis is constrained such that each cluster includes a minimum number of sites that are known in advance to be suitable for sensor installation. Alternatively the technical feasibility of installing a sensor at each location could be assessed, and clusters created to include a range of different levels of feasibility.

Second, the participant base underwent changes between the cluster analyses and the sensor data analysis. Such changes are to be expected,

[6] We thank an anonymous reviewer for raising these points.

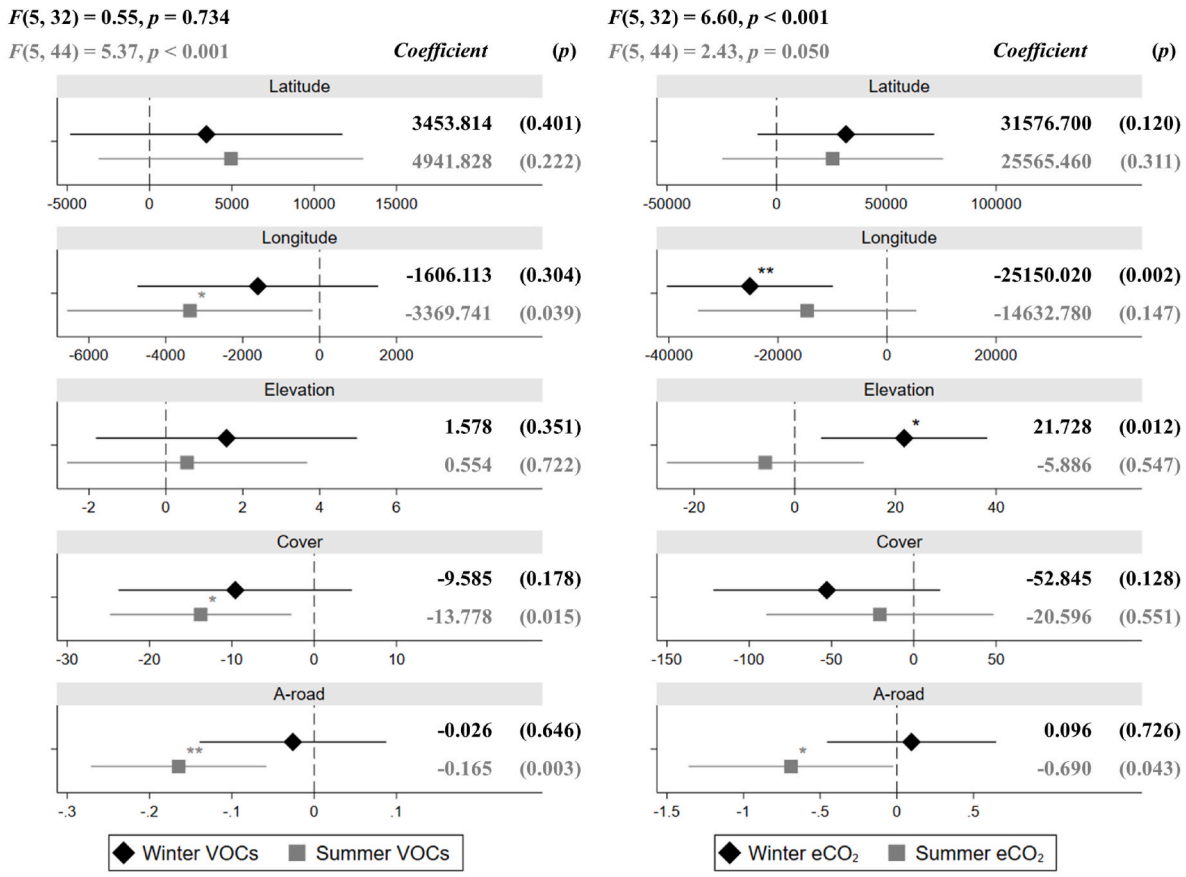


Fig. 12. Coefficients (and p-values[5]) for the predictors of VOCs and eCO_2 in winter (N = 38) and summer (N = 50).

given that time is required to allow sensor data to be collected. However, the impact could be attenuated in a variety of ways, depending on suitability for the project. The number of clusters could be decreased, such that the loss of participants has a reduced impact on each cluster. The entire set of potential participants could be known and clustered in advance to ensure that participants of all types are represented, such that new recruits can be assigned to existing clusters. Time could be allowed for stabilisation of the participant base. For example, some of our participants withdrew when the survey was conducted, at the beginning of the project. Allowing such a milestone to pass before performing the cluster analysis would have reduced the number of homes subsequently lost.

The main aim of this study was to provide a simple (non-specialist) and generalisable solution to the deployment of sensors when numbers are limited. This resource limitation presents a problem for selecting representative sites at which to place sensors [46,47], whilst also avoiding redundancy in resource deployment [48]. As reviewed earlier, previous approaches to selecting locations for a limited number of sensors are usually specific to the type of sensor or setting, and can involve application-specific techniques. Clustering approaches use sensor data to provide similar groupings [61,94], which requires collecting or estimating the data before recommendations can be made. While the specificity of approaches are beneficial for some applications, generalisability and accessibility for future users can be limited [61]. Our novel methodology offers the following benefits.

Cluster analysis is an accessible approach that can readily implemented, facilitated by freely available software packages, and online tutorials and resources. It results in groups of similar items, which can then be sampled in a targeted and strategic manner to provide representation across the diverse range of potential sensor sites.

In this study, we present two applications of this clustering method. The factors we chose to cluster potential sensor locations in the current research were specific to households and the outdoor environment. However, the method does not rely on using these factors, rather they reflect information that we had available in advance that might affect sensor readings. Other applications may have different potentially influencing factors. For example, the relevant factors differ between our two example applications (utilities and environment). In other applications these factors could be replaced by any potential explanatory variables (e.g., Fig. 14).

This methodology is therefore applicable to any type of sensor as long as at least some of the factors influencing sensor measurements are known. This data-driven approach allows creation of similar groups of possible locations that together cover the sensor space, allowing targeted yet unbiased selection of representative locations.

Biased sampling could be accounted for after data collection. However, sufficient variation within the sample would still be required to be able characterise the bias. For example, if all homes selected had four rooms, then the effect of number of rooms on the data could not be estimated without strong assumptions, so extrapolation to homes with a different numbers of rooms would not be possible.

Furthermore, cluster analysis for representative sampling can be applied to other domains than sensor placement. For example, large cohorts of participants can be categorised into archetypes or personas that capture the core characteristics of the cohort without preconceptions or bias in categorisation, given a bottom-up data-driven approach [98]. Such personas can then be used to sample from the cohort for more detailed investigation, such as in-depth qualitative interviews, while being confident that different ranges of multidimensional characteristics are being represented [e.g., Ref. [134]].

Beyond this methodology, and in line with the aims of the broader project, ongoing and future studies will use the data collected from these sensors to investigate relationships between energy usage, indoor and external environments, including temperatures and air quality. There

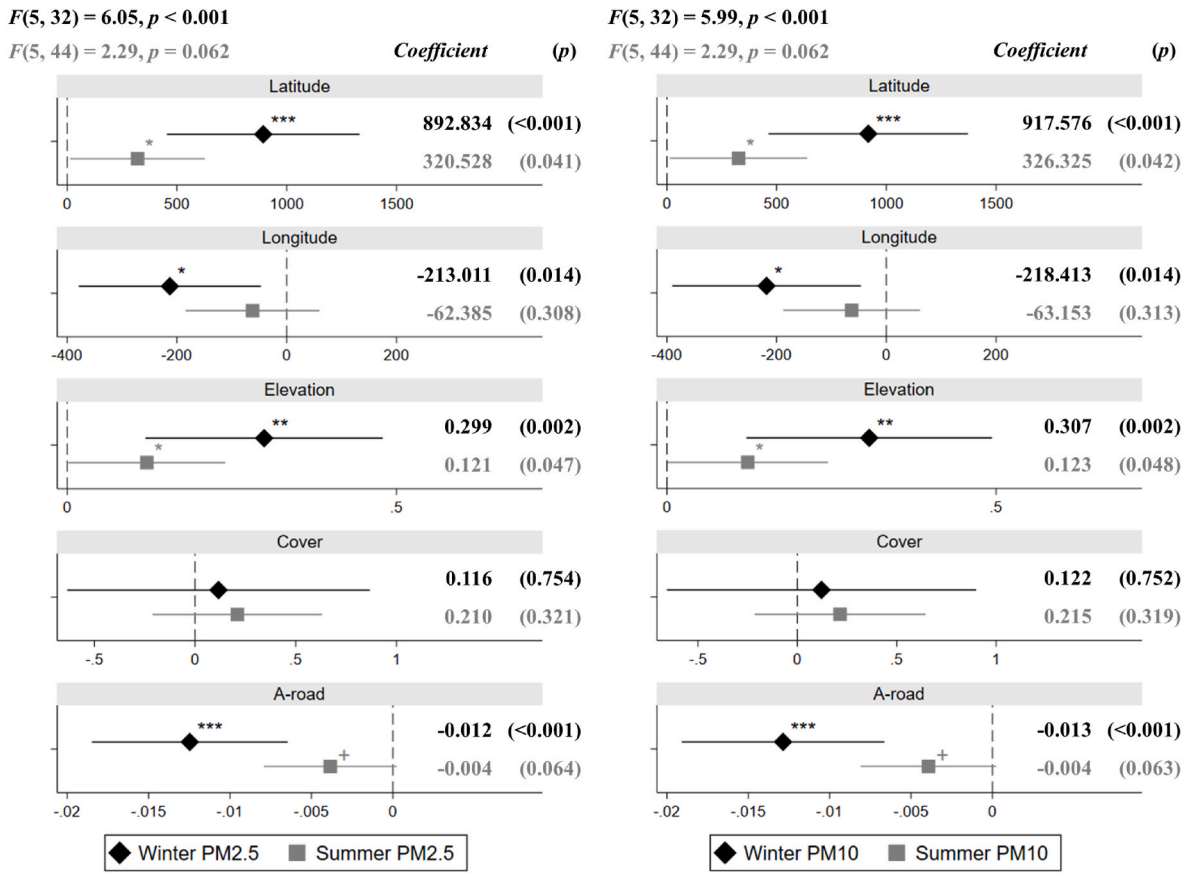


Fig. 13. Coefficients (and p-values[5]) for the predictors of PM2.5 and PM10 in winter (N = 38) and summer (N = 50).

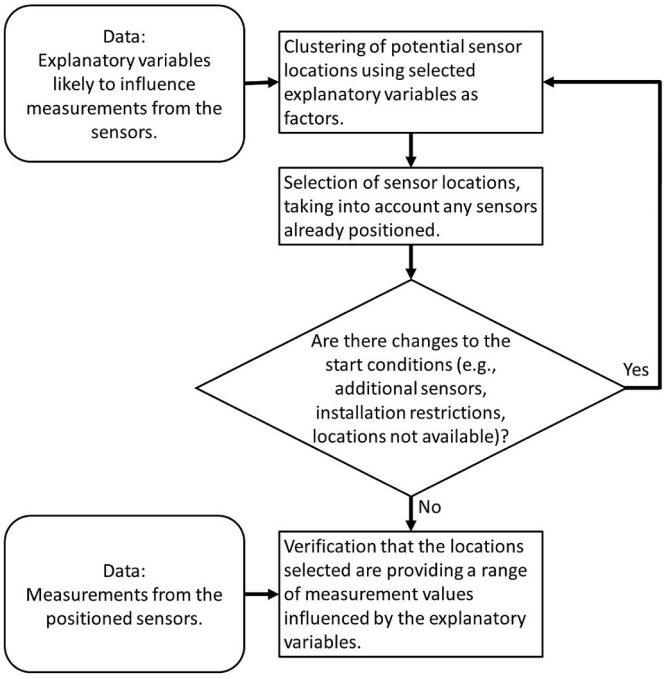


Fig. 14. Generalised graphical representation of the methodology.

are many research questions that could be addressed with these data. For example, differences between indoor and external temperature data could be used to assess thermal performance, and in conjunction with energy usage could be used to assess energy efficiency. Separately, changes over time in relative humidity or air quality could be assessed in response to environmental interventions (e.g., positive pressure units). The sensor dataset will also be published for future researchers.

9. Conclusion

The aim of this study was to test the use of cluster analysis to provide a representative sample of sensor sites when sensors cannot be installed at all available sites. Clusters were successfully created for two types of sensors in separate analyses. Sites for sensor installation were then sampled from each cluster according to the distance of the site from the cluster centre, in order to capture typical and atypical members of each cluster group. Results of analyses to verify the clusterings showed that clusters did capture differences across sensor data, and that the cluster factors selected for segmenting and sampling the potential sensor sites had relationships with relevant sensor measurements. These results suggest that the clusters were meaningful, and successfully captured a range of homes and sensor sites. In conclusion, when sensor deployment is limited, for example, by sensor numbers or access issues, this cluster-based methodology can provide a representative subset of sensor sites by sampling across clusters in order to capture the variety across potential sites.

10. Terminology

Please see Table 5.

CRediT authorship contribution statement

Tamaryn Menneer: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation,

Table 5
Terminology and abbreviations.

Term	Description
Factor	In this study, factor denotes a variable or characteristic of a home, household or potential sensor site. It represents a numerical property. The combination of factor values describing a given item is used to calculate distance in the cluster analysis. Items with similar factor values will be grouped together.
Cluster analysis	A process to create groups (i.e., clusters) of items that are similar to each other and that differ across groups. First cluster centres are randomly created, then a two-step iterative process is used to minimis the inertia. (1) Items are assigned to their nearest cluster. (2) The cluster centres are recalculated as the average of their item members.
Inertia	A measure of how well the items are clustered into similar groups. It is calculated as the sum of squared distances between the homes and the centroid within each cluster. Distance is the distance between factors used to define the clusters.
Silhouette	A measure of overlap of the clusters, with a larger value representing less overlap. It is calculated as the normalised difference between the mean distance between members within a cluster and the mean distance between those cluster members and the members in the nearest other cluster.
VOC	Volatile organic compounds. These gases can be emitted by substances such as paints, cleaning products, furnishings and cosmetics.
eCO_2	Equivalent carbon dioxide. It can be used for measuring carbon footprints. It expresses the impact of each different gas in terms of the amount of CO_2 that it would create. The VOC sensor reports equivalent CO_2. eCO_2 is not actual CO_2 present in the area. The VOC can be used as an eCO_2 sensor to in real world environments, where the main cause of VOCs is from humans.
PM2.5	Fine particulate matter. Particles or liquid droplets in the air that have a diameter up to 2.5 μm PM2.5 can be naturally occurring, such as dust and sea salt, while some is human made, such as particulates from vehicle exhausts and combustion in the home.
PM10	Particulate matter of diameter up to 10 μm.
DSM	Digital Surface Model. The heights of the surface of the earth, including features such as vegetation and buildings [108]
DTM	Digital Terrain Model. Based on the DSM with surface objects removed [107,109].
A-road	Distance from the home to the nearest A (main) road.

Conceptualization. **Markus Mueller:** Writing – review & editing, Visualization, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Stuart Townley:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

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.

Data availability

Data will be made available on request.

Acknowledgements

The Smartline Project (05R16P00305) has received £3,740,920 and the Smartline Extension Project (05R18P02819) is receiving up to £3,307,703 of funding from the England European Regional Development Fund as part of the European Structural and Investment Funds Growth Programme 2014–2020. With additional funding of £25k from the Southwest Academic Health Science Network and £200k from Cornwall Council. For more information on our funding please see the Smartline website - https://www.smartline.org.uk/about.

The Smartline project (www.smartline.org.uk) is a partnership between University of Exeter, Coastline Housing, Volunteer Cornwall, Cornwall Council, and the South West Academic Health Science Network.

The authors would like to acknowledge the use of the University of Exeter's Advanced Research Computing facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.buildenv.2023.110032.

References

[1] K. Ashton, That 'Internet of Things' thing, RFiD Journal 22 (7) (2009) 97–114.
[2] A. Verma, S. Prakash, V. Srivastava, A. Kumar, S.C. Mukhopadhyay, Sensing, controlling, and IoT infrastructure in smart building: a review, IEEE Sensor. J. 19 (20) (2019) 9036–9046, https://doi.org/10.1109/JSEN.2019.2922409.
[3] F. McLoughlin, A. Duffy, M. Conlon, A clustering approach to domestic electricity load profile characterisation using smart metering data, Appl. Energy 141 (2015) 190–199, https://doi.org/10.1016/j.apenergy.2014.12.039.
[4] J.D. Rhodes, W.J. Cole, C.R. Upshaw, T.F. Edgar, M.E. Webber, Clustering analysis of residential electricity demand profiles, Appl. Energy 135 (2014) 461–471, https://doi.org/10.1016/j.apenergy.2014.08.111.
[5] J.K. Hart, K. Martinez, Environmental Sensor Networks: a revolution in the earth system science? Earth Sci. Rev. 78 (3–4) (2006) 177–191, https://doi.org/10.1016/j.earscirev.2006.05.001.
[6] J.K. Hart, K. Martinez, Sensor networks and Geohazards, Environ. Earth Sci. (2020), https://doi.org/10.1016/B978-0-12-818234-5.00037-7.
[7] M. Leyli-Abadi, A. Same, L. Oukhellou, N. Cheifetz, P. Mandel, C. Feliers, O. Chesneau, Predictive classification of water consumption time series using non-homogeneous Markov models, Proc. Int. Conf. Data Sci. (2017) 323–331, https://doi.org/10.1109/Dsaa.2017.32.
[8] S. Yamagami, H. Nakamura, A. Meier, Non-Intrusive Submetering of Residential Gas Appliances, American Council for an Energy-Efficient Economy (ACEEE), 1996.
[9] S.L. Dong, S.H. Duan, Q. Yang, J.L. Zhang, G.G. Li, R.Y. Tao, MEMS-based smart gas metering for Internet of Things, IEEE Internet Things 4 (5) (2017) 1296–1303, https://doi.org/10.1109/Jiot.2017.2676678.
[10] A. Veit, C. Goebel, R. Tidke, Household electricity demand forecasting - benchmarking state-of-the-art methods, in: ACM E-Energy, 14, Cambridge, United Kingdom, 2014. Jun 11-13, 2014.
[11] K. Gram-Hanssen, K.N. Petersen, Different everyday lives: different patterns of electricity use, in: Proc 2004 American Council for an Energy Efficient Economy, 2004, pp. 1–13.
[12] J. Widen, E. Wackelgard, A high-resolution stochastic model of domestic activity patterns and electricity demand, Appl. Energy 87 (6) (2010) 1880–1892, https://doi.org/10.1016/j.apenergy.2009.11.006.
[13] B. Anderson, S. Lin, A. Newing, A. Bahaj, P. James, Electricity consumption and household characteristics: implications for census-taking in a smart metered future, Comput. Environ. Urban Syst. 63 (2017) 58–67.
[14] F. McLoughlin, A. Duffy, M. Conlon, Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: an Irish case study, Energy Build. 48 (2012) 240–248, https://doi.org/10.1016/j.enbuild.2012.01.037.
[15] A.G. Riddell, K. Manson, Parametrisation of domestic load profiles, Appl. Energy 54 (3) (1995) 199–210.
[16] W. Kleiminger, C. Beckel, S. Santini, Household occupancy monitoring using electricity meters, Proc ACM Int Joint Conf Pervasive and Ubiquitous Computing (UbiComp 2015) (2015) 975–986, https://doi.org/10.1145/2750858.2807538.
[17] Y. Zhao, W. Zeiler, G. Boxem, T. Labeodan, Virtual occupancy sensors for real-time occupancy information in buildings, Build. Environ. 93 (2015) 9–20, https://doi.org/10.1016/j.buildenv.2015.06.019.
[18] R.P. Zhang, M. Kong, B. Dong, Z. O'Neill, H. Cheng, F. Hu, J. Zhang, Development of a testing and evaluation protocol for occupancy sensing technologies in building HVAC controls: a case study of representative people counting sensors, Build. Environ. 208 (2022), https://doi.org/10.1016/j.buildenv.2021.108610.
[19] A. Lofti, L.I. Jalil, A. Al-Habaibeh, Investigating occupant behaviour to improve energy efficiency in social housing, IEEE 9th Int. Conf. Intell. Comput. 9 (2013) 124–128.
[20] Z. Ren, G. Foliente, W. Chan, D. Chen, M. Ambrose, P. Paevere, A model for predicting household end-use energy consumption and greenhouse gas emissions in Australia, Int J Sustain Build 4 (3) (2013) 210–228.
[21] M. Royapoor, T. Roskilly, Building model calibration using energy and environmental data, Energy Build. 94 (2015) 109–120, https://doi.org/10.1016/j.enbuild.2015.02.050.
[22] B. Chun, J.M. Guldmann, Impact of greening on the urban heat island: seasonal variations and mitigation strategies, Comput. Environ. Urban 71 (2018) 165–176, https://doi.org/10.1016/j.compenvurbsys.2018.05.006.
[23] S. Bhandari, N. Bergmann, R. Jurdak, B. Kusy, Time series data analysis of wireless sensor network measurements of temperature, Sensors 17 (6) (2017), https://doi.org/10.3390/s17061221.

[24] J. Yun, K.H. Won, Building environment analysis based on temperature and humidity for smart energy systems, Sensors 12 (10) (2012) 13458–13470, https://doi.org/10.3390/s121013458.

[25] S. Devarakonda, P. Sevusu, L. Hongzhang, R. Liu, L. Iftode, B. Nath, Real-time air quality monitoring through mobile sensing in Metropolitan areas, in: UrbComp '13 Proc 2nd ACM SIGKDD Int Workshop on Urban Comput, 2013, pp. 1–8, https://doi.org/10.1145/2505821.2505834.

[26] T. Elbir, Comparison of model predictions with the data of an urban air quality monitoring network in Izmir, Turkey, Atmos. Environ. 37 (15) (2003) 2149–2157, https://doi.org/10.1016/S1352-2310(03)00087-6.

[27] G.M. Solomon, T.R. Campbell, G.R. Feuer, J. Masters, A. Samkian, K.A. Paul, No breathing in the Aisles: diesel exhaust inside school buses, Nat. Resour. Defense Council, Coalition Clean Air (2001). January 2001.

[28] H.R. Bohanon Jr., J.J. Piade, M.K. Schorp, Y. Saint-Jalm, An international survey of indoor air quality, ventilation, and smoking activity in restaurants: a pilot study, J. Expo. Anal. Environ. Epidemiol. 13 (5) (2003) 378–392, https://doi.org/10.1038/sj.jea.7500284.

[29] I. Filella, J. Penuelas, Daily, weekly, and seasonal time courses of VOC concentrations in a semi-urban area near Barcelona, Atmos. Environ. 40 (40) (2006) 7752–7769, https://doi.org/10.1016/j.atmosenv.2006.08.002.

[30] A. Mohammadshirazi, V.A. Kalkhorani, J. Humes, B. Speno, J. Rike, R. Ramnath, J.D. Clark, Predicting airborne pollutant concentrations and events in a commercial building using low-cost pollutant sensors and machine learning: a case study, Build. Environ. 213 (2022), https://doi.org/10.1016/j.buildenv.2022.108833.

[31] Z.T. Ai, C.M. Mak, D.J. Cui, On-site measurements of ventilation performance and indoor air quality in naturally ventilated high-rise residential buildings in Hong Kong, Indoor Built Environ. 24 (2) (2015) 214–224, https://doi.org/10.1177/1420326x13508566.

[32] S.M. Dutton, M.J. Mendell, W.R. Chan, M. Barrios, M.A. Sidheswaran, D. P. Sullivan, E.A. Eliseeva, W.J. Fisk, Evaluation of the indoor air quality minimum ventilation rate procedure for use in California retail buildings, Indoor Air 25 (1) (2015) 93–104, https://doi.org/10.1111/ina.12125.

[33] H.S. Shin, Y.C. Ahn, J.K. Lee, T.W. Kang, K.G. Lee, H.S. Park, Measurement of indoor air quality for ventilation with the existence of occupants in schools, in: Proc 10th Int Conf on Indoor Air Quality and Climate, 2005, pp. 2762–2766.

[34] A. Gnauck, Interpolation and approximation of water quality time series and process identification, Anal. Bioanal. Chem. 380 (3) (2004) 484–492, https://doi.org/10.1007/s00216-004-2799-3.

[35] H.G. Stefan, E.B. Preudhomme, Stream temperature estimation from air-temperature, Water Resour. Bull. 29 (1) (1993) 27–45.

[36] X.S. Yan, J.Y. Gong, Q.H. Wu, Pollution source intelligent location algorithm in water quality sensor networks, Neural Comput. Appl. 33 (1) (2021) 209–222, https://doi.org/10.1007/s00521-020-05000-8.

[37] iOpt, iOpt Smart Devices to Combat Fuel Poverty. Housing Technology, 2019. Available online: https://www.housing-technology.com/iopt-smart-devices-to-combat-fuel-poverty/. (Accessed 17 October 2019).

[38] Cross Keys Homes, CityFibre, Social Housing Landlord Undertakes IoT Trial across City-wide Fibre. SmartCitiesWorld, 2019. ; Available online, https://www.smartcitiesworld.net/news/news/social-housing-landlord-undertakes-iot-trial-across-city-wide-fibre-4306. (Accessed 17 October 2019).

[39] G. Tu, K. Morrissey, R.A. Sharpe, T. Taylor, Combining self-reported and sensor data to explore the relationship between fuel poverty and health well-being in UK social housing, Wellbeing, Space and Society 3 (2022), https://doi.org/10.1016/j.wss.2021.100070.

[40] Cambridge Architectural Research, Loughborough University, and Element Energy, Further Analysis of the Household Electricity Survey. Early Findings: Demand Side Management, 2013. Available online, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/275483/early_findings_revised.pdf. (Accessed 29 May 2018).

[41] J. Palmer, I. Cooper, United Kindgom Housing Energy Fact File, Department of Energy and Climate Change, London, UK, 2013. Available online, https://www.gov.uk/government/statistics/united-kingdom-housing-energy-fact-file-2013. (Accessed 12 September 2018).

[42] Flatline. Unlocking the value from flexibility in housing: Phase 1 feasibility report. Accessed on 22 January 2019; Available online: http://www.seroenergy.com/flatline-project/.

[43] UK Parliament, UK Indoor Air Quality: November 2010. POSTnote, 2010. Available online: https://researchbriefings.parliament.uk/ResearchBriefing/Summary/POST-PN-366. (Accessed 14 October 2019).

[44] G. Laput, Y. Zhang, C. Harrison, Synthetic sensors: towards general-purpose sensing, Proc. Conf. Hum. Factors Comput. Syst. (CHI 17 (2017) 3986–3999, https://doi.org/10.1145/3025453.3025773.

[45] Wales & West Utilities, Freedom Project: Interim Findings (2018). Available online, https://www.wwutilities.co.uk/about-us/our-company/publications/the-future-of-energy-research/. (Accessed 18 April 2019).

[46] J.W. Ding, S.J. Cao, Identification of zonal pollutant diffusion characteristics using dynamic mode decomposition: towards the deployment of sensors, Build. Environ. 206 (2021), https://doi.org/10.1016/j.buildenv.2021.108379.

[47] A.T. Murray, K. Kim, J.W. Davis, R. Machiraju, R. Parent, Coverage optimization to support security monitoring, Comput. Environ. Urban 31 (2) (2007) 133–147, https://doi.org/10.1016/j.compenvurbsys.2006.06.002.

[48] J.C.M. Pires, S.I.V. Sousa, M.C. Pereira, M.C.M. Alvim-Ferraz, F.G. Martins, Management of air quality monitoring using principal component and cluster analysis - Part I: SO2 and PM10, Atmos. Environ. 42 (6) (2008) 1249–1260, https://doi.org/10.1016/j.atmosenv.2007.10.044.

[49] O.S. Adedoja, Y. Hamam, B. Khalaf, R. Sadiku, A state-of-the-art review of an optimal sensor placement for contaminant warning system in a water distribution network, Urban Water J. 15 (10) (2018) 985–1000, https://doi.org/10.1080/1573062x.2019.1597378.

[50] R. Paris, S. Beneddine, J. Dandois, Robust flow control and optimal sensor placement using deep reinforcement learning, J. Fluid Mech. 913 (2021), https://doi.org/10.1017/jfm.2020.1170.

[51] A.D. Fontanini, U. Vaidya, B. Ganapathysubramanian, A methodology for optimal placement of sensors in enclosed environments: a dynamical systems approach, Build. Environ. 100 (2016) 145–161, https://doi.org/10.1016/j.buildenv.2016.02.003.

[52] C. Chen, C. Gorle, Optimal temperature sensor placement in buildings with buoyancy-driven natural ventilation using computational fluid dynamics and uncertainty quantification, Build. Environ. 207 (2022), https://doi.org/10.1016/j.buildenv.2021.108496.

[53] L.J. Zeng, J. Gao, L.P. Lv, R.Y. Zhang, L.Q. Tong, X. Zhang, Z.H. Huang, Z. F. Zhang, Markov-chain-based probabilistic approach to optimize sensor network against deliberately released pollutants in buildings with ventilation systems, Build. Environ. 168 (2020), https://doi.org/10.1016/j.buildenv.2019.106534.

[54] J.C.P. Cheng, H.H.L. Kwok, A.T.Y. Li, J.C.K. Tong, A.K.H. Lau, BIM-supported sensor placement optimization based on genetic algorithm for multi-zone thermal comfort and IAQ monitoring, Build. Environ. 216 (2022), https://doi.org/10.1016/j.buildenv.2022.108997.

[55] G.M. Huebner, I. Hamilton, Z. Chalabi, D. Shipworth, T. Oreszczyn, Comparison of indoor temperatures of homes with recommended temperatures and effects of disability and age: an observational, cross-sectional study, BMJ Open 8 (5) (2018), https://doi.org/10.1136/bmjopen-2017-021085.

[56] S.S.V. Vianna, The set covering problem applied to optimisation of gas detectors in chemical process plants, Comput. Chem. Eng. 121 (2019) 388–395, https://doi.org/10.1016/j.compchemeng.2018.11.008.

[57] C.C. Castello, J. Fan, A. Davari, R.X. Chen, Optimal sensor placement strategy for environmental monitoring using wireless sensor networks, in: 42nd Southeastern Symp On System Theory (SSST), 2010.

[58] C. Papadimitriou, Optimal sensor placement methodology for parametric identification of structural systems, J. Sound Vib. 278 (4–5) (2004) 923–947, https://doi.org/10.1016/j.jsv.2003.10.063.

[59] J. Williams, O. Zahn, J.N. Kutz, Data-driven sensor placement with shallow decoder networks, arXiv Dynamical Systems (2022), https://doi.org/10.48550/arXiv.2202.05330.

[60] Y. Tan, L.M. Zhang, Computational methodologies for optimal sensor placement in structural health monitoring: a review, Struct. Health Monit. 19 (4) (2020) 1287–1308, https://doi.org/10.1177/1475921719877579.

[61] D. Yoganathan, S. Kondepudi, B. Kalluri, S. Manthapuri, Optimal sensor placement strategy for office buildings using clustering algorithms, Energy Build. 158 (2018) 1206–1225, https://doi.org/10.1016/j.enbuild.2017.10.074.

[62] D. Arthur, S. Vassilvitskii, k-means plus plus: the Advantages of Careful Seeding, Proc Annu ACM-SIAM Symp (2007) 1027–1035.

[63] B.F. Hobbs, Y.D. Ji, Stochastic programming-based bounding of expected production costs for multiarea electric power systems, Oper. Res. 47 (6) (1999) 836–848, https://doi.org/10.1287/opre.47.6.836.

[64] H. Kile, Evaluation and grouping of power market scenarios in security of electricity supply analysis (PhD Thesis), in: Department of Electric Power Engineering, Norwegian University of Science and Technology, Trondheim, 2014.

[65] P. Nahmmacher, E. Schmid, L. Hirth, B. Knopf, Carpe diem: a novel approach to select representative days for long-term power system modeling, Energy 112 (2016) 430–442, https://doi.org/10.1016/j.energy.2016.06.081.

[66] S. Wogrin, P. Duenas, A. Delgadillo, J. Reneses, A new approach to model load levels in electric power systems with high renewable penetration, IEEE Trans. Power Syst. 29 (5) (2014) 2210–2218, https://doi.org/10.1109/Tpwrs.2014.2300697.

[67] T. Rasanen, D. Voukantsis, H. Niska, K. Karatzas, M. Kolehmainen, Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data, Appl. Energy 87 (11) (2010) 3538–3545, https://doi.org/10.1016/j.apenergy.2010.05.015.

[68] Y. Amri, A.L. Fadhilah, Fatmawati, N. Setiani, S. Rani, Analysis clustering of electricity usage profile Using K-Means algorithm, Iop Conf Ser-Mat Sci. 105 (2016), https://doi.org/10.1088/1757-899x/105/1/012020.

[69] T. Chen, A. Mutanen, P. Jarventausta, H. Koivisto, Change detection of electric customer behavior based on AMR measurements, IEEE Trans. Power Syst. (2015).

[70] K. Gajowniczek, T. Zabkowski, Data mining techniques for detecting household characteristics based on smart meter data, Energies 8 (7) (2015) 7407–7427, https://doi.org/10.3390/en8077407.

[71] A. Satre-Meloy, M. Diakonova, P. Grunewald, Cluster analysis and prediction of residential peak demand profiles using occupant activity data, Appl. Energy 260 (2020), https://doi.org/10.1016/j.apenergy.2019.114246.

[72] N. Cheifetz, Z. Noumir, A. Same, A. Sandraz, C. Feliers, V. Heim, Modeling and clustering water demand patterns from real-world smart meter data, Drink. Water Eng. Sci. 10 (2017) 75–82, https://doi.org/10.5194/dwes-10-75-2017.

[73] C. Laspidou, E. Papageorgiou, K. Kokkinos, S. Sahu, A. Gupta, L. Tassiulas, Exploring patterns in water consumption by clustering, Computing and Control for the Water Industry (CCWI2015) 119 (2015) 1439–1446, https://doi.org/10.1016/j.proeng.2015.08.1004.

[74] K. Aksela, M. Aksela, Demand estimation with automated meter reading in a distribution network, J Water Res Pl-Asce 137 (5) (2011) 456–467, https://doi.org/10.1061/(Asce)Wr.1943-5452.0000131.

[75] N. Avni, B. Fishbain, U. Shamir, Water consumption patterns as a basis for water demand modeling, Water Resour. Res. 51 (10) (2015) 8165–8181, https://doi.org/10.1002/2014wr016662.

[76] S.A. McKenna, F. Fusco, B.J. Eck, Water demand pattern classification from smart meter data, Comput. Control Water Ind. 70 (CCWI2013) (2014) 1121–1130, https://doi.org/10.1016/j.proeng.2014.02.124.

[77] M.P. Fernandes, J.L. Viegas, S.M. Vieira, J.M.C. Sousa, Segmentation of residential gas consumers using clustering analysis, Energies 10 (12) (2017), https://doi.org/10.3390/en10122047.

[78] A. Franco, F. Fantozzi, Analysis and clustering of natural gas consumption data for thermal energy use forecasting, J. Phys. Conf. Ser. 655 (2015), https://doi.org/10.1088/1742-6596/655/1/012020.

[79] O. Laib, M.T. Khadir, L. Mihaylova, A Gaussian process regression for natural gas consumption prediction based on time series data, in: 2018 *21st* Int Conf on Inf Fusion, 2018, pp. 55–61.

[80] A. Zakovorotnyi, A. Seerig, Building energy data analysis by clustering measured daily profiles, Enrgy Proced 122 (2017) 583–588, https://doi.org/10.1016/j.egypro.2017.07.353.

[81] M. Hayn, V. Bertsch, W. Fichtner, Electricity load profiles in Europe: the importance of household segmentation, Energy Res. Social Sci. 3 (2014) 30–45.

[82] W. Abrahamse, L. Steg, How do socio-demographic and psychological factors relate to households' direct and indirect energy use and savings? J. Econ. Psychol. 30 (5) (2009) 711–720, https://doi.org/10.1016/j.joep.2009.05.006.

[83] S.S. Badhiye, P.N. Chatur, B.V. Wakode, Temperature and humidity data analysis for future value predcition using clustering technique: an approach, Int. J. Emerg. Technol. Adv. Eng. 2 (1) (2012) 88–91.

[84] S. Kalaivani, K.S. Kumar, Cluster analysis: temperature data, Int. J. Pure Appl. Math. 119 (7) (2018) 779–785.

[85] A. Kumar, A.K. Swamy, Comparison of clustering approaches on temperature zones for pavement design, in: A. Nikolaides (Ed.), Bituminous Mixtures & Pavements VI, Taylor & Francis Group, London, 2015.

[86] J. Hidalgo, R. Jougla, On the use of local weather types classification to improve climate understanding: an application on the urban climate of Toulouse, PLoS One 13 (12) (2018), https://doi.org/10.1371/journal.pone.0208138.

[87] M. Bador, P. Naveau, E. Gilleland, M. Castella, T. Arivelo, Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe, Weather Clim. Extrem. 9 (2015) 17–24, https://doi.org/10.1016/j.wace.2015.05.003.

[88] A. Iizuka, S. Shirato, A. Mizukoshi, M. Noguchi, A. Yamasaki, Y. Yanagisawa, A cluster analysis of constant ambient air monitoring data from the Kanto region of Japan, Int. J. Environ. Res. Publ. Health 11 (7) (2014) 6844–6855, https://doi.org/10.3390/ijerph110706844.

[89] S. Saksena, V. Joshi, R.S. Patil, Cluster analysis of Delhi's ambient air quality data, J. Environ. Monit. 5 (3) (2003) 491–499, https://doi.org/10.1039/b210172f.

[90] J. Soares, P.A. Makar, Y. Aklilu, A. Akingunola, The use of hierarchical clustering for the design of optimized monitoring networks, Atmos. Chem. Phys. 18 (9) (2018) 6543–6566, https://doi.org/10.5194/acp-18-6543-2018.

[91] G. Tuysuzoglu, D. Birant, A. Pala, Majority voting based multi-task clustering of air quality monitoring network in Turkey, Appl Sci-Basel 9 (8) (2019), https://doi.org/10.3390/app9081610.

[92] Y.L. Chen, L.Z. Wang, F.Y. Li, B. Du, K.K.R. Choo, H. Hassan, W.J. Qin, Air quality data clustering using EPLS method, Inf. Fusion 36 (2017) 225–232, https://doi.org/10.1016/j.inffus.2016.11.015.

[93] R.K. Grace, R. Manimegalai, M.S.G. Devasena, S. Rajathi, K. Usha, N.R. Baseria, Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data, IEEE Reg 2016 Annu Int Conf Proc/TENCON (2016) 1945–1949.

[94] J. Li, C. Wang, Z. Qian, C. Lu, Optimal sensor placement for leak localization in water distribution networks based on a novel semi-supervised strategy, J. Process Control 82 (2019) 13–21, https://doi.org/10.1016/j.jprocont.2019.08.001.

[95] T. Menneer, Z. Qi, T. Taylor, C. Paterson, G. Tu, L.R. Elliott, K. Morrissey, M. Mueller, Changes in domestic energy and water usage during the UK COVID-19 lockdown using high-resolution temporal data, Int. J. Environ. Res. Publ. Health 18 (13) (2021), https://doi.org/10.3390/ijerph18136818.

[96] T. Menneer, M. Mueller, R.A. Sharpe, S. Townley, Modelling mould growth in domestic environments using relative humidity and temperature, Build. Environ. 208 (2022), https://doi.org/10.1016/j.buildenv.2021.108583.

[97] T. Walker, T. Menneer, C. Leyshon, M. Leyshon, A.J. Williams, M. Mueller, T. Taylor, Determinants of volunteering within a social housing community, Voluntas (2020), https://doi.org/10.1007/s11266-020-00275-w.

[98] A.J. Williams, T. Menneer, M. Sidana, T. Walker, K. Maguire, M. Mueller, C. Paterson, M. Leyshon, C. Leyshon, E. Seymour, Z. Howard, E. Bland, K. Morrissey, T. Taylor, Fostering engagement with health and housing innovation: development of participant personas in a social housing cohort, JMIR Public Health and Surveillance 7 (2) (2021), https://doi.org/10.2196/25037.

[99] Smartline, Smartline (2020). Available online: https://www.smartline.org.uk/. (Accessed 17 February 2020).

[100] L. Moses, K. Morrissey, R.A. Sharpe, T. Taylor, Exposure to indoor mouldy odour increases the risk of asthma in older adults living in social housing, Int. J. Environ. Res. Publ. Health 16 (14) (2019), https://doi.org/10.3390/ijerph16142600.

[101] A.J. Williams, K. Maguire, K. Morrissey, T. Taylor, K. Wyatt, Social cohesion, mental wellbeing and health-related quality of life among a cohort of social housing residents in Cornwall: a cross sectional study, BMC Publ. Health 20 (1) (2020), https://doi.org/10.1186/s12889-020-09078-6.

[102] Glasgow Open Data. (n.d.). Glasgow Open Data. Accessed on 23 January 2020 and 5th January 2021; Available online: https://data.glasgow.gov.uk/https://futurecity.glasgow.gov.uk/data/.

[103] E. Domene, D. Sauri, Urbanisation and water consumption: influencing factors in the Metropolitan region of Barcelona, Urban Stud. 43 (9) (2006) 1605–1623, https://doi.org/10.1080/00420980600749969.

[104] A.A. Makki, R.A. Stewart, C.D. Beal, K. Panuwatwanich, Novel bottom-up urban water demand forecasting model: revealing the determinants, drivers and predictors of residential indoor end-use consumption, Resour. Conserv. Recycl. 95 (2015) 15–37, https://doi.org/10.1016/j.resconrec.2014.11.009.

[105] F. Fuerst, D. Kavarnou, R. Singh, H. Adan, Determinants of energy consumption and exposure to energy price risk: a UK study, Zeitschrift für Immobilienökonomie. 6 (2020) 65–80.

[106] J. Harold, S. Lyons, J. Cullinan, The determinants of residential gas demand in Ireland, Energy Econ. 51 (2015) 475–483, https://doi.org/10.1016/j.eneco.2015.08.015.

[107] DTM, LiDAR Based Digital Terrain Model (DTM) Data for South West England, 2014. Available online: https://data.gov.uk/dataset/lidar-based-digital-terrain-model-dtm-data-for-south-west-england. (Accessed 1 November 2017).

[108] DSM, LiDAR Based Digital Surface Model (DSM) Data for South West England, 2014. Available online: https://data.gov.uk/dataset/lidar-based-digital-surface-model-dsm-data-for-south-west-england. (Accessed 1 November 2017).

[109] F. Ferraccioli, F. Gerard, C. Robinson, T. Jordan, M. Biszczuk, L. Ireland, M. Beasley, A. Vidamour, A. Barker, R. Arnold, M. Dinn, A. Fox, A. Howard, LiDAR Based Digital Terrain Model (DTM) Data for South West England, NERC Environmental Information Data Centre (Dataset), 2014, https://doi.org/10.5285/e2a742df-3772-481a-97d6-0de5133f4812.

[110] K.J. Pizzey, PyLidar: Python package for loading LIDAR geospatial Digital Surface Models (DSM) (2017). Available online, www.pypi.org/project/pylidar/, www.github.com/Ffisegydd/pylidar/. (Accessed 27 November 2017).

[111] M.H. McCutchan, D.G. Fox, Effect of elevation and aspect on wind, temperature and humidity, J. Clim. Appl. Meteorol. 25 (1986) 1996–2013.

[112] A.A. Karner, D.S. Eisinger, D.A. Niemeier, Near-roadway air quality: synthesizing the findings from real-world data, Environ. Sci. Technol. 44 (14) (2010) 5334–5344, https://doi.org/10.1021/es100008x.

[113] C. Bell, Doogal: Postcodes, maps and code (2019). Available online, https://www.doogal.co.uk/. (Accessed 10 December 2019).

[114] Google, Google Maps (2017; 2019), 5th February 2019; Available online, www.google.co.uk/maps/. (Accessed 30 November 2017).

[115] S.P. Lloyd, Least-squares quantization in PCM, Ieee T Inform Theory 28 (2) (1982) 129–137, https://doi.org/10.1109/Tit.1982.1056489.

[116] G. Boeing, Clustering to reduce spatial data set size, arXiv (2018), https://doi.org/10.48550/arvix.1803.08101.

[117] Python, Python Software Foundation (2017). Available online, www.python.org. (Accessed 23 January 2020).

[118] G. van Rossum, Python Reference Manual, 1995. CWI Report CS-R9525.

[119] NumPy. Numerical Python. Accessed on 23 January 2020; Available online: www.numpy.org.

[120] T.E. Oliphant, Guide to NumPy, Trelgol Publishing, USA, 2006.

[121] E. Jones, T. Oliphant, P. Peterson, and others., SciPy: Open source scientific tools for Python (2001). Available online, http://www.scipy.org/. (Accessed 23 January 2020).

[122] Pandas. Pandas: Python Data Analysis Library. Accessed on; Available online: https://pandas.pydata.org/.

[123] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95, https://doi.org/10.1109/Mcse.2007.55.

[124] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[125] C. Elkan, Using the triangle inequality to accelerate k-means, Proc. 20th Int. Conf. Mach. Learn. (2003) 147–153.

[126] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. Roy. Stat. Soc. B 63 (2001) 411–423, https://doi.org/10.1111/1467-9868.00293.

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

[128] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, J. Roy. Stat. Soc. B 61 (1999) 611–622, https://doi.org/10.1111/1467-9868.00196.

[129] Carnotcycle, How to convert relative humidity to absolute humidity, Available online: https://carnotcycle.wordpress.com/2012/08/04/how-to-convert-relative-humidity-to-absolute-humidity/, 2012. (Accessed 9 December 2022).

[130] B. Jann, COEFPLOT: Stata module to plot regression coefficients and other results (2013). Available online, http://ideas.repec.org/c/boc/bocode/s457686.html. (Accessed 8 March 2022).

[131] A. Kavousian, R. Rajagopal, M. Fischer, Determinants of residential electricity consumption: using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior, Energy 55 (2013) 184–194, https://doi.org/10.1016/j.energy.2013.03.086.

[132] R.A. Sharpe, T. Taylor, L.E. Fleming, K. Morrissey, G. Morris, R. Wigglesworth, Making the case for "whole system" approaches: integrating public health and

housing, Int. J. Environ. Res. Publ. Health 15 (11) (2018), https://doi.org/10.3390/ijerph15112345.

[133] C. Lu, H. Xu, W. Meng, W. Hou, W. Zhang, G. Shen, H. Cheng, X. Wang, X. Wang, S. Tao, A novel model for regional indoor PM(2.5) quantification with both external and internal contributions included, Environ. Int. 145 (2020), 106124, https://doi.org/10.1016/j.envint.2020.106124.

[134] Walker, T., Menneer, T., Morrissey, K., Tu, G., Mueller, M., Leyshon, C., Leyshon, M., and Bland, E. (Submitted for publication). Adoption of Indoor Environment Sensor Technology for Health: a Social Housing Case Study.