Clustering and classification of aggregated smart meter data to better understand how demand patterns relate to customer type

S. R. Mounce[1], W. R. Furnass[1], E. Goya[2], M. Hawkins[2], J. B. Boxall[1]

[1] Department of Civil and Structural Engineering, University of Sheffield, Sheffield, S1 3JD, UK
[2] The Power House -Reading STW, Island Rd, Reading, Berkshire, RG2 0RP, UK

Abstract

The most widely applied approach to representing a WDS demand pattern is a 24-hour depiction of historically observed values (e.g. for network model use). In recent years, the proliferation of AMR meters has yielded large volumes of demand data. The richness of information contained in this daily cycle behaviour has seldom being used to explore significant variability as regards geographical location, consumer types and their heterogeneity, presence of large consumers and day of the week / season of the year. In this study, after pre-processing/ cleaning the raw AMR data, the data (consisting of over 250 million readings) was aggregated to give several representative daily profiles per meter/property (96 points/profile at 15-min resolution). A set of customer demand profile types was obtained by clustering data from 3428 AMR units (installed in the UK cities of Reading, Swindon and London) logging fifteen minute data over an approximate three year period (meter dependent), using k-means++ (with a correlation distance metric) on the cleaned dataset. All meters had additional metadata including a property type label (corresponding to residential dwelling type or commercial property). This allowed for an evidence-based evaluation of the number and shape of demand profiles and allowed characterisation of the ensuant clusters using property type, in particular whether commercial in nature. Three natural clusters in the data were found to correlate strongly to residential and commercial composition based on the customer type label which was used for post-analysis. This was explored further by developing a classifier using the profiles as predictor variables and ‘commercial’ or ‘residential’ as the target. Classifier models achieved up to 94.5 % overall accuracy and 0.95 ROC AUC with a five K-fold validation.

1 Introduction

The concept of a Smart City places citizens at the centre of services within a city. This involves bringing together hard infrastructure, social capital including local skills and community institutions, and technologies to fuel sustainable economic development and provide an attractive environment for all. For the water sector this means using technologies for optimising water resources and waste treatment, monitoring and controlling water, and providing real-time information to help water companies and households manage their water better. The increasing use of smart water metering technologies for monitoring networks in real-time is providing water utilities with an ever growing amount of data on their business operations and infrastructure. Such metering devices embrace two distinct technologies: meters that record water usage; and

communication systems that can store and transmit real-time water use information (Stewart et al., 2010). The ideal approach for their Smart City application is installing smart water meters at the property boundary in conjunction with intelligent end use pattern recognition algorithms either in-built into the meter software or within a processing module at the utilities data centre. However, such an end goal requires the ability to analyse collected data without human interaction and manual reclassification and this is non-trivial.

Recent work has explored the use and analysis of such data. It has been argued that using actual observed data, demand profiles can be calculated to provide more accurate representations of high-granularity historical data, with potential applications in real-time leakage detection, customer profiling and the provision of network modelling demand patterns (Vitorino et al. 2014). Time series clustering is an active area of research, with the major issues being high dimensionality, temporal order and noise (Rani and Sikka, 2012). Time-series clustering is Temporal-Proximity-Based Clustering if it works directly on raw data either in frequency or time domain. Laspidou et al. (2015) used Kohonen Self-Organising Maps to cluster consumers according to their water consumption (household and business consumer type was available). Non real time quarterly billing data for two datasets (168 and 454 customer meters) was analysed, with 30 data points per meter in total. Features were extracted (means and ratios) for input to the SOM. Distinct clusters in the SOM were found to correspond to higher percentages of residential or commercial properties. McKenna et al. (2014) investigated employing Gaussian Mixture Models (GMM's) as the basis set for representing demand patterns using a data set of hourly demand readings spanning a six-month study period, for 85 service connections within a single DMA. Whilst there was no customer information available for the data set, it was hypothesised after applying k-means that evidence of patterns found may represent both residential and commercial customers. Garcia et al. (2015) demonstrated the potential use of k-means for clustering AMR data based on shape. Hadoop and Spark were used in a Big Data context to provide an unsupervised classification of the demand patterns from smart meters, with hourly interval feature vectors of a weekly profile for 51,117 smart meters over a one-year period (approximately 317 million observed readings). Nine distinctive clusters were identified. However, no additional information (including customer type) was available other than demands.

2 Case Study

SmartWater4Europe (SW4E) is a four-year FP7 demonstration project (2014-17) funded by the European Commission (Demonstration of integrated smart water supply solutions at four sites across Europe, grant 619024). The four demo sites (of varying in scale and located in the United Kingdom, Spain, The Netherlands and France) are allowing demonstration of solutions incorporating sensors, data processing, modelling and analytics technologies.

The UK demo site (TWIST) is focusing on leak management in particular the use of Advanced Metering Infrastructure (AMI) smart metering. This Thames Water demonstration site in Reading has been investigating how new and emerging

technologies can be used to create a 'smart network' with real time notification of performance and even asset condition to enable proactive management and intervention. For leakage applications, the main focus is on two DMAs. Thames Water have instrumented these two DMAs with Sensus water meters (SWM) that communicate over AMI, in particular using the Flexnet system. In the Flexnet system, the SWM incorporates a low power short range radio that transmits to its associated Local Communication Equipment (LCE). This is located very close to the meter itself; typically the meter and the LCE are less than 500 mm apart and each meter has its own LCE. The LCE then transmits over a long range, typically up to 1 km, to a radio base station. Each radio base station will usually receive data from many LCEs, communications between these being two way. The SWMs collect 15 minute interval data and transmit once per day.

3 Application

3.1 Data set, preparation and meta-statistics

A large AMR dataset was obtained (as a 12GB CSV anonymised set of readings and associated metadata, including property type classification) covering areas contained in three UK cities, with AMR data principally from five DMAs including the two main DMAs from the Reading TWIST demo site. 98.1% of Reading meters were active during the data set period. Much of the initial low level data cleaning is only summarised in brief here, and included dealing with such issues as meters without readings, meters not defined, properties without meters, dealing with one to many / many to one / many to many relationships between properties, meter and supply as well as duplicates and conflicts. It is often the goal when data mining large, dirty datasets to filter to a best quality subset. Once data is cleaned from such duplicates and conflicts there are approximately 250 million readings for 4,108 meters for all three areas. The longest continuous period is for 891 days of raw AMR data. The pre-processing was conducted using Jupyter Notebook, the pandas Python library and awk for chunking/sorting and filtering the 12GB AnonReadings CSV into set of per-meter CSVs as follows:

Step 1: Remove readings with null meter IDs

Step 2: Create a metadata table by:

`AnonMeters` inner join to `AnonRelationships` on `AnonMeterID`

Left join results to `AnonProperties` on `AnonPropertyID`

Step 3: Ensure each reading can be associated with one property (type), dropping where they do not exist:

- `AnonReadings` where `AnonMeterID` not in metadata
- `metadata` rows where `AnonMeterID` not in `AnonReadings`
- `metadata` and `AnonReadings` rows where meter to property relationship is not 1:1

Step 4: Process readings per meter/property

For each meter/property (each `metadata` row):

4.1: Keep only one of duplicate non-conflicting readings (same meter, timestamp and value)

4.2: Drop conflicting readings (same meter, timestamp, all estimated or not estimated, different readings)

4.3: Remaining conflicting readings should be an estimated and a non-estimated reading for a given meter and timestamp; sort (partially) by EstimatedReading then take the first (the non-estimated value) of each pair

4.4: Drop meters (readings and metadata) if have no non-null readings

4.5: If have >1 reading with same meter ID and timestamp then drop all metadata and readings associated with this meter and record the ID

4.6: Sort by timestamp (by setting the dataframe index to timestamp)

4.7: Resample at 15-min resolution

4.8: Calculate flow as 1st order difference of readings / 900s.

4.9: Set invalid flows to null; invalid if a) Negative (should not get backflow at customer properties) or b) Exceed max theoretical flow predicted by orifice equation

4.10: Infill gaps of <3 readings in flow series using linear interpolation

4.11: Calculate several aggregate profiles for meter reading data, each of 96 readings:

I. filter all reading data for current meter by whether weekday, whether weekend or no filter
II. group by minute of the day
III. aggregate flow values by taking mean or std dev
IV. repeat filtering and aggregation for all combinations of week subset and aggregation function resulting in six 96-pt profiles per meter

4.11: Since most statistical and machine learning algorithms can't handle nulls; drop all aggregate profiles with >0 null values

4.12: Write outputs (as CSVs) and statistics

Figure 1 provides some high level statistics for the averaged profiles (all data). The Seaborn package's 'tsplot' function has been used to calculate confidence intervals: it uses bootstrapping to estimate the distribution of the mean value at each time point and then finds the low and high percentile values (corresponding to the confidence interval being used) from these distributions. The default confidence interval is 68% – equivalent to ± one standard deviation of the mean, assuming normal distribution. The respective low and high percentiles are 16% and 84%. In Figure 1 we see the typical strong diurnal demand pattern. The difference between weekday and weekend use is also apparent, with weekday use starting earlier and weekend use being larger to the first peak. The distribution of property type classification labels comprises around 88% residential types (subdivided into detached, semi-detached, terraced and flat housing) and approximately 12% commercial, which includes all non-residential types. For some analyses, all the residential types are combined into type RS.

3.2 Methodology

Clustering aims to discover structure in a complex data set and is useful when natural groupings are suspected but there are many competing patterns in the data. k-means clustering is a popular method of vector quantization, that is popular for cluster analysis in data mining. k-means clustering partitions data observations into k clusters in which each observation belongs to the cluster with the nearest mean; it is a so-called centroid method. These clusters are then used to characterise the

dataset. Previous work by Garcia et al. (2015) has demonstrated the potential use of k-means for clustering AMR profiles based on shape. The k-means++ algorithm (Arthur and Vassilvitskii, 2007) was used (MATLAB implementation) which chooses initial centers in a way that gives a provable upper bound on the within cluster sum of squares objective.

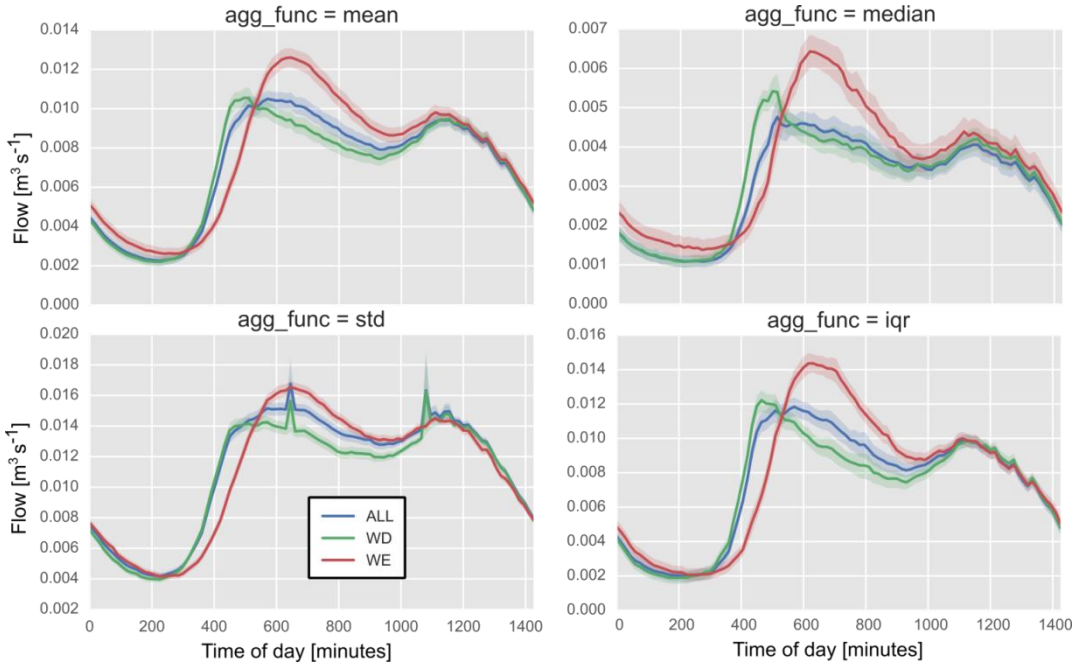

Figure 1: Mean, standard deviation, median and iqr (with confidence intervals) for all daily average profiles, and separated into weekday and weekend

An iterative approach with replicates was used to find what proved to be a good clustering of the mean diurnal profiles of the type illustrated in Figure 1. A silhouette plot using the cluster indices output from k-means can be used to assess how well-separated the resulting clusters are. The plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters. This measure ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, Note that k-means cannot be used with NaNs – either imputation (such as feature average/median or using more sophisticated approaches) or marginalisation (leaving the data out) must be used. Marginalisation was used in this analysis (for one set of the CSVs generated in the data processing step). Rather than the usual squared Euclidean distance, a correlation distance metric was used calculated as one minus the sample correlation between points (treated as sequences of values). Each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to zero mean and unit standard deviation. The k-Nearest Neighbour (k-NN) pattern matching method is widely used in data clustering, classification and prediction. Based on a specific distance metric or similarity measure, k-NN examines vector distances to determine the nearest neighbours (Cover and Hart, 1967). MATLAB was used to explore variants of k-NN and other classification algorithms to evaluate the ability to classify profiles into residential and commercial types for all meters.

4 Results and discussion

4.1 Clustering and customer type

Figure 2a provides a silhouette plot for k-means applied with 2 clusters, which provided the best separation for n = 2 to 8 when considering commercial only meters in the full data set. Figure 2b and 2c provide the two dominant centroids for commercial properties.

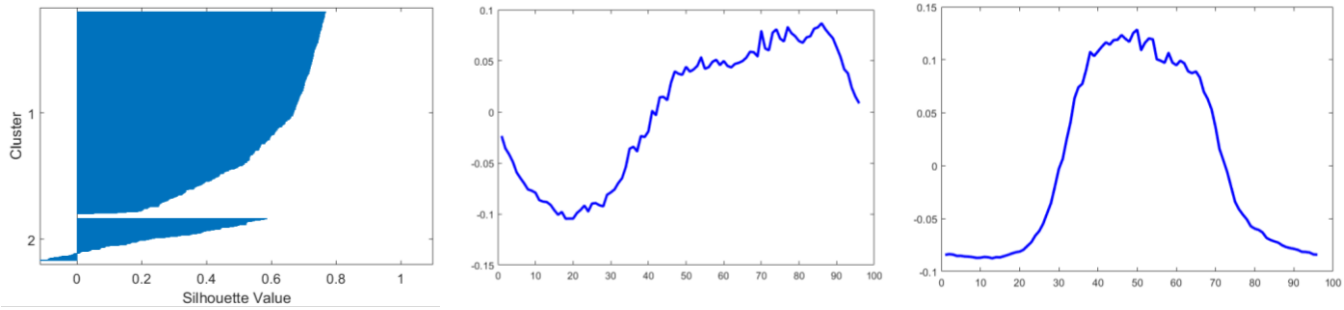

Figure 2a, b, c: Silhouette plot of k-means++ with two commercial clusters, Centroid 1 (cluster 1), Centroid 2 (cluster 2)

Centroid 1 for the most numerous cluster manifests a raised evening consumption, cluster 2 has a more humped profile with usage dominating in daylight hours. The k-means clustering was repeated for the Reading data set (all property types) which was known to have more commercial property types, contains two DMAs in the SW4E demo site area and has superior data quality. Three clusters provided the best separation (based on silhouette plots) for n = 2 to 8. Figure 3a, 3b and 3c provide the centroids. Figure 3 gives the calculated composition of property types for the three clusters (post-analysis use of type codes). Subsequent plots of groups of meters from the clusters confirmed the three predominant shapes of the centroids provided in Figure 3. Evidently the data set is dominated by three types of pattern: the typical WDS diurnal flow demand, a profile with larger secondary maxima after about 6pm (peaking around 9pm) and finally a more 'humped' profile. Centroid 1 equated to the standard diurnal residential pattern where water consumers are gone for a significant portion of the day at predictable hours for work and educational reasons. This results in a sharp morning peak and a more dispersed evening peak with a minimum between the two peaks (Figure 3a). Centroid 3 seems to indicate a possible alternative residential pattern with a householder generally at home throughout the day, whilst still possessing morning and evening demand peaks, but without a well delineated minimum in between them (Figure 3c). Demographics such as age and employment status in particular regions could contribute to this cluster. Centroid 2 reveals a cluster of patterns in which the secondary peak is greater than the morning peak due to demand throughout business hours (the commercial customer types may be predominantly bars, restaurants etc.). This corresponds to centroid one shown in Figure 2 when analysing commercial meters only. Figures 3d, 3e and 3f provide the composition of clusters when utilising the property type code. A key distinction is residential vs commercial (the commercial type (CM) is the exploded slice). Cluster 1 is 91% residential, cluster 3 is 94% residential but cluster 2 is 62% commercial. Notice also how cluster 3 has a significantly higher proportion of terraced housing.

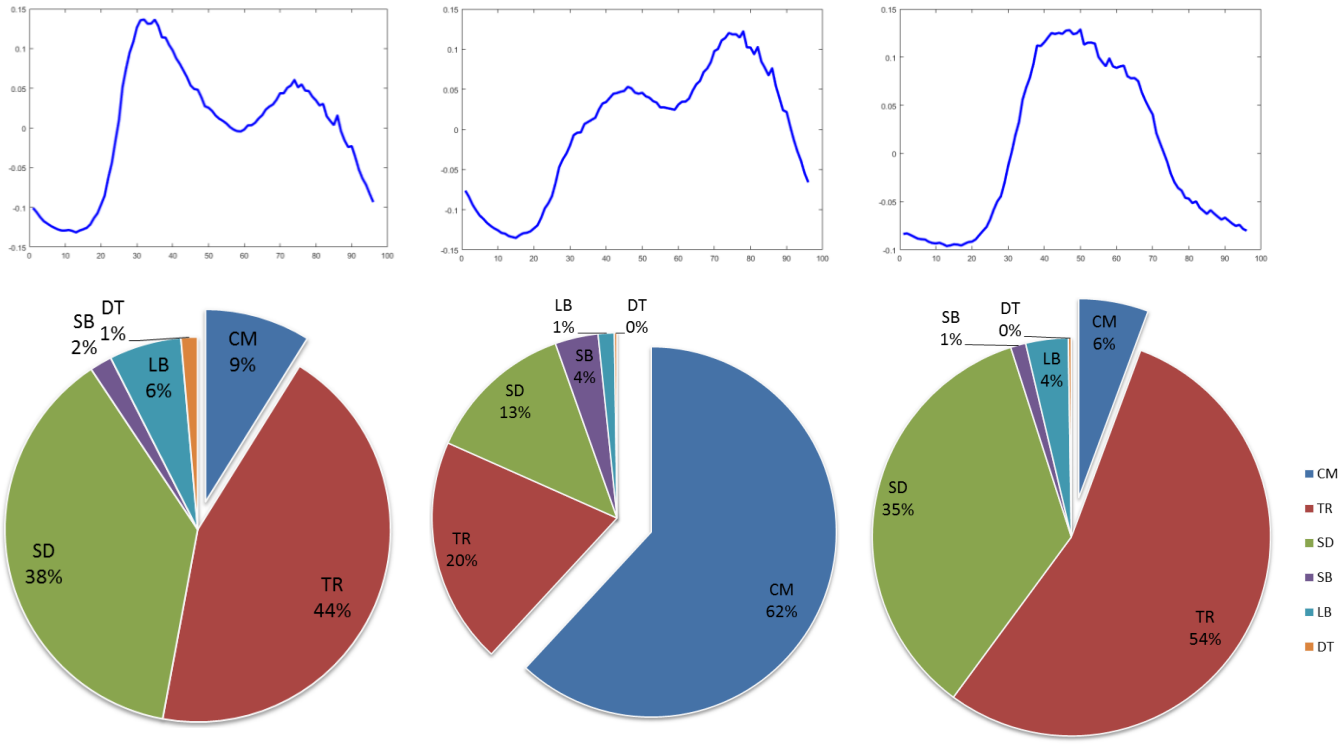

Figure 3a, b, c, d, e, f: Centroid 1 (cluster 1), Centroid 2 (cluster 2), Centroid 2 (cluster 2), composition of customer types for cluster 1, cluster 2, cluster 3 (CM: Commercial, TR: Terraced, SD: Semi-detached, D: Detached, SB/LB: Flats/ Apartments)

4.2 Classification for residential and commercial customer type

A complementary approach to clustering is to explore the use of the mean profiles and associated customer type in a classifier. Since k-means suggests the presence of features relating to customer types, it should be possible to classify unseen meter averages into residential or commercial categories. Several classifiers were applied. A K-fold cross validation approach was applied (Kohavi, 1995). The data are broken into K-blocks (five were used here, resulting in an 80/20 split). Then, for K=1 to X, the Kth block becomes the validation (or test) block with the remaining data becoming the training data. The K-folds were randomly selected from all meters for all three cities, 'RS' denotes residential and is all the combined residential types. Table 1 provides the results including overall accuracy, True Positive Rates and the Area Under Curve value for the ROC curve for the best 5 performing models. High percentage accuracy was achieved of up to 94.5% on the k-fold validation. The overall classifier accuracy is a not always the best indication of superior models in the case of imbalanced numbers of class labels. Note that only 12% of meters are classed as commercial in the full meter dataset. When examples of one class greatly outnumber examples of the other class(es), traditional machine learning algorithms tend to favour classifying examples as belonging to the overrepresented and dominating (majority) class. The RUSBoost (Random Under Sampling) algorithm is designed to classify when one class has many more observations than another (Seiffert et al. 2010). The majority of class-imbalance learning techniques currently implemented, including RUSBoost, have been

designed for two-class problems. An ensemble approach using RUSBoost (Model 2) provided the best TPR (84%) for commercial property, whilst maintaining over 90% overall accuracy for unseen data. For the Reading subset, where commercial properties account for 25% of meters, the results were similar but with a higher commercial percentage accuracy without using RUSBoost – for example Model 1 provided TPR of 96% for residential and 76% for commercial.

Table 1: Classifier results for k-fold validation for all meters

Model	Type	Accuracy	TPR (CM)	TPR (RS)	AUC
1	KNN (10 neighbours, Euclidean distance)	94.1%	62%	98%	0.93
2	**Ensemble of RUSBoosted decision trees**	**91.3%**	**84%**	**92%**	**0.95**
3	KNN (10 neighbours, Cosine distance)	93.1%	67%	97%	0.92
4	Ensemble of subspace KNN (30 learners, subspace dimension 48)	94.1%	68%	97%	0.91
5	Ensemble of boosted decision trees (AdaBoost)	94.5%	68%	98%	0.95

5 Conclusions and further work

Increasing amounts of Smart Network data is now being collected by WSPs, however the data is only of real business value if this valuable resource is ultimately used to inform and support decision making. The full range of uses for these observations is only beginning to be realised and exploited. This paper has presented, using a case study of approximately 250 million readings, a workflow for cleaning and pre-processing AMR data and then clustering average daily demand patterns using the k-means ++ algorithm with a correlation distance metric. Three natural clusters in the data (confirmed by using silhouette plots) were found to correspond strongly to a residential and commercial composition based on customer type which was used for post-analysis. Further, a classification approach was also presented, comparing five classification models with K-fold cross validation, in order to classify into residential and commercial customers. When using an ensemble of RUSBoosted decision trees (for a 5 fold) the overall accuracy was 91.3% (TPR 92% for residential and 84% for commercial) confirming dominant patterns of usage. Potential application areas for further work using this form of cluster and classification analysis as more smart data becomes routinely available include:

- data mining: understanding how businesses and households use water and whether and where unique patterns in this use exist is essential for proactive management (including weekday vs. weekend analysis)
- applying customer segmentation based on consumption data for customer loading and variable water pricing in a similar manner to energy (some industrial customers already have tariffs based on time of day usage)
- for leak detection activities based on detecting pattern changes (deviation from cluster centroids/ distributions); data-driven models of demand could also help identify atypical customers or unusual changes in consumption
- filling missing data for audits/ regulatory purposes using cluster centroids/ typical usage perhaps allowing volumetric usage and flow profiles to be estimated for unmetered customers.
- more accurate demand profiles for hydraulic modelling.

Acknowledgements

The authors wish to thank Thames Water for data provision and assistance. This research was funded under the EU FP7 SmartWater4Europe demonstration project, grant 619024.

References

Arthur, D. and Vassilvitskii, S., 2007. K-means++: The Advantages of Careful Seeding. In: SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. 2007, pp. 1027–1035.

Cover, T., Hart, P. (1967). Nearest neighbour pattern classification. IEEE Transactions on Information Theory, 13(1), pp. 21–27.

García, D., Gonzalez, D., Quevedo, J., Puig, V. and Saludes, J. (2015). Water demand estimation and outlier detection from smart meter data using classification and Big Data methods. In: Proceedings of 2nd IWA New Developments in IT and Water Conference, Rotterdam, February 2015.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, 20–25 August, Montreal, Quebec, Canada, 2, pp. 1137–1143.

Laspidoua,C., Papageorgioub, E., Kokkinos, K., Sahud, S., Guptae, A., and Tassiulas, L. (2015). Exploring patterns in water consumption by clustering. Procedia Engineering 119 (2015) 1439 – 1446.

McKennaa, S. A., Fuscoa, F. and Ecka, B. J. (2014). Water demand pattern classification from smart meter data. Procedia Engineering 70 (2014) 1121 – 1130

Rani, S. and Sikka, G. (2012). Recent Techniques of Clustering of Time Series Data: A Survey. International Journal of Computer Applications (0975 – 8887), Volume 52– No.15, August 2012.

Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V. and A. Napolitano, A. "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance," IEEE Transaction on Systems, Man and Cybernetics-Part A: Systems and Human, vol. 40, no. 1, January 2010.

Stewart, R.A., Willis, R., Giurco, D., Panuwatwanich, K., Capati, G., 2010. Web-based knowledge management system: linking smart metering to the future of urban water planning. Australian Planner 47, 66e74.

Vitorinoa, D., Loureirob, D., Alegreb, H., Coelhob, S., Mamadeb,A. (2014). In: Defense of the Demand Pattern, a Software Approach. Procedia Engineering 89 (2014) 982 – 989.