Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 70 (2014) 1121 – 1130

Procedia Engineering

www.elsevier.com/locate/procedia

12th International Conference on Computing and Control for the Water Industry, CCWI2013

Water demand pattern classification from smart meter data

S.A. McKenna[a,*], F. Fusco[a], B.J. Eck[a]

[a]*IBM Research, Smarter Cities Technology Centre, Bldg. 3, Damastown Industrial, Mulhuddart, Dublin 15, Ireland*

Abstract

High frequency measurements of water demand at service connections are becoming more common as utilities install smart meter technology. The full range of use for these observations by water suppliers is only beginning to be realized. Potential applications include leak detection, improved demand forecasting, variable water pricing, and improved network operations. Here we develop an approach for the classification of demand patterns and apply this approach to a set of demands collected from smart meters within a single District Metered Area (DMA) of a municipal network. The goal of this work is to develop a robust procedure for classification of demands derived from smart metering and test this procedure on observational data. A fundamental aspect of many feature classification tools is representation of what are often complex and noisy data in a low dimensional feature space that captures the important attributes of the signal. In this work, we employ Gaussian Mixture Models (GMM's) as the basis set for representing demand patterns. GMM's provide a flexible approach to representing the temporal demand patterns with a relatively small number of parameters. The values of these parameters then serve as the feature set for multivariate classification. A data set of hourly demand readings spanning a six-month study period serve as the test case for analysis here. The smart meters record demands to both residential and commercial consumers. Results show that the GMM approach captures variations in the demand patterns between locations. To the first order, the identified patterns appear to be explained by the differences between residential and commercial consumers. The resulting groupings are compared to classifications made using total demand as the sole feature. The stability of the patterns over time is tested by independently clustering each month of data.

Selection and peer-review under responsibility of the CCWI2013 Committee

Keywords: Smart meters; water demand; Gaussian mixture model; clustering;

1. Introduction

Distribution networks provide necessary materials from central locations to consumers located at nodes of the network. For efficient design and operation of the network, it is necessary to estimate the demands that will be placed upon that network. For water distribution networks, nodal demands are often measured monthly or quarterly for billing purposes and the instantaneous demand at any node is simply estimated as the fraction of the total demand over the billing period consistent with the fraction of time of interest. Similarly, natural gas and electricity networks also require estimation of demands at varying time scales.

* Corresponding author. Tel.: +353 018 269 738.
E-mail address: seanmcke@ie.ibm.com

1877-7058
Selection and peer-review under responsibility of the CCWI2013 Committee
doi:10.1016/j.proeng.2014.02.124

Research into methods for accurately predicting network demands has received considerable attention with a variety of approaches being applied including statistical prediction models, machine learning and simple population-based approaches (e.g., Adamowski (2008); Blokker, et al. (2010); Jain and Ormsbee (2002)). Pervasive use of high-resolution meters at service connections provides a large data set of historical demands and shifts the research focus from demand prediction to classification of the observed demands. The goal of classification is to identify a finite set of patterns that are representative of the use types within the network. These classes can then be combined with socio-economic data to better understand water use across all service connections and provide the basis for leak detection, network design and operation and serve as the basis for variable water pricing if desired. Classification of demand patterns has been explored in the past and is a currently active area of research in both water and electricity distribution networks with the electricity networks having received more attention to this point.

The key component in any multivariate classification problem is to define a feature set that efficiently captures the unique characteristics of the demand patterns. While any function can be used to fit the observed demands, probability density functions (pdf's) are preferred as a straightforward means of assigning the probability of occurrence to any demand value. Singh et al. (2009) provide a brief summary of various probability density functions (pdf) used to fit distributions of residential power loads (demands) from electricity consumption data. In these studies, the classification is done first based solely on the known consumer type (e.g., Domestic or Commercial) and then the probability density function (pdf) for all nodes at a junction for that class of consumer are calculated. One goal is to fit these observations with a parametric distribution that can be used to define the variability in loads with a small set of parameters. Works cited by Singh et al. (2009) that have attempted these parametric fits include: Irwin et al (1986) who used the Weibull distribution and Heunis and Herman. (2002) who fit beta distributions. Additional authors (e.g., Ghosh, et al. (1997)) have compared the ability of different distributions to fit a single data set. The conclusion of these works is that load distributions typically do not fit any parametric distribution well and the correct distribution to use is often problem-specific.

Gaussian Mixture Models (GMM's) use a weighted linear combination of Gaussian pdf's to represent a probability distribution possessing more complexity than could be described by a single Gaussian distribution. An advantage of fitting demand distributions with GMM's is that the multiple Gaussian distributions allow for a single model containing several parametric distributions to flexibly fit a large variety of distribution shapes. Singh et al. (2009) apply GMM's to electricity load distributions with good results. While GMM's have seen some application to electricity load modeling, application of GMM's to water demands is extremely limited. Application of GMM's to water demands with the most relevance to this work is that of Aksela and Aksela (2011). The significant differences of this work and that of Aksela and Aksela (2011) are that here we apply GMM's to the observed data and then use multivariate classification based on the GMM parameters, whereas Aksela and Aksela (2011) do one-dimensional classification using the measured total consumption at each service connection and then apply GMM's to model the demands within each pre-defined class. Additionally, we focus here on classification of daily demand patterns where Aksela and Aksela (2011) apply their approach to weekly demands.

The goal of this paper is to develop and demonstrate a robust approach to classifying water demand patterns as obtained from smart meter observations. The approach developed here uses a GMM to fit the daily demand patterns and the estimated GMM parameters serve as the input to multivariate clustering to identify demand classes. The following section provides a high-level overview of the GMM and clustering algorithms. An example data set is introduced in the next section and the results of applying the GMM-based classification approach are discussed.

2. Methods

Density mixture models can be used to approximate an empirical, often non-parametric, probability density function as a mixture of multiple known parametric distributions. Here we focus on mixtures of Gaussian distributions to fit observed demands throughout a 24 hour period. The estimated parameters of the mixture model define the features used in multivariate clustering.

2.1. Gaussian Mixture Models

Mixture models provide a means of describing a single observed data set as a combination of multiple probability density functions. Given a vector of observational data of length N, $\chi = (x_1, x_2, x_3, ..., x_N)$ where each datum is an i.i.d. sample from a probability density, p, the goal is to estimate the parameters, Θ, such that the likelihood the observations, χ, were drawn from a distribution with parameters, Θ, is maximized:

$$p(\chi|\Theta) = \prod_{i=1}^{N} p(x_i|\Theta) = \mathcal{L}(\Theta|\chi) \tag{1}$$

where Θ contains the parameters that describe each component distribution as well as the weight applied to each distribution. The general expression for a mixture model is:

$$p(x|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(x|\theta_i) \tag{2}$$

See (Bilmes (1998)) where probability densities of M component distributions are mixed together using the weighting coefficients α. The weighting coefficients are constrained such that $\sum_{i=1}^{M} \alpha_i = 1.0$. Each component density, p_i, has parameters contained in θ_i. For multi-dimensional densities that are solely Gaussian, $\theta = (\mu, \sum)$ where μ are the means and $\sum$ is the covariance matrix. Here, demand patterns are estimated with time as the single independent variable, one-dimensional, such that $\sum$ is a vector containing only the M variances. The Gaussian probability density function (pdf) is given by:

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\} \tag{3}$$

Each Gaussian density is fully defined by $\theta_i = [\mu_i, \Sigma_i]$. The number of component distributions, M, is fixed *apriori* and estimation of the parameters in θ and the corresponding weight coefficients, α, are done using maximum likelihood estimation. The non-linear relationship between the likelihood and the estimated parameters requires an iterative approach to estimation, and this is accomplished through the expectation-maximization algorithm as implemented in the *gmdistribution.fit* function (Matlab (2012a)). Once a GMM is fit to the demand pattern for a particular time period, the estimated parameters of the GMM then serve as the feature space for classification of those patterns.

2.2. Multivariate Classification

The goal of multivariate clustering is to use the measurements in the feature space to separate demand patterns into K clusters that minimize intra-cluster variability and maximize inter-cluster differences (separation). Following Xu and Wunsch (2009), a set of objects $x_j \in \Re^d$, $j = 1, ..., N$ to be organized into K clusters $C = [C_1, ..., C_K]$ must minimize the sum of squared errors defined as:

$$J_s(\Gamma, A) = \sum_{i=1}^{K}\sum_{j=1}^{N} \gamma_{ij} \parallel x_j - a_i \parallel^2 \tag{4}$$

where $A = [a_1, ..., a_K]$ is the matrix of cluster centroid means (averages) within the d-dimensional space with the sample mean for the *ith* cluster containing N_i objects being:

$$m_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \gamma_{ij} x_j \tag{5}$$

$\Gamma = \gamma_{i,j}$ is the cluster partition matrix,

$$\gamma_{ij} = \left\{\begin{matrix} 1 \; if x_j \in cluster_i \\ 0 \; otherwise \end{matrix}\right\} \tag{6}$$

with

$$\sum_{i=1}^{K} \gamma_{ij} = 1, \forall j \tag{7}$$

The k-means algorithm requires that K be pre-specified. Random starting locations for the cluster centroids are used to initialize the clustering process. The k-means clustering is done using the implementation in the R package (R Core Team (2012)).

3. Application

Water meters were installed at approximately 250 service connections within a single district metered area (DMA). Hourly water use measurements were recorded over a six-month period beginning on January 1st, 2011. The customers within this DMA are a mix of residential and commercial entities.

3.1. Example Data Set and Preparation

The data examined here are obtained by recording the total cumulative amount of water consumed with a one-hour sampling frequency. The difference between the measured cumulative volumes at the current and previous time steps provide the measured flow rate (demand) at the service connection: $Q_t = (V_t - V_{t-1})/\Delta t$. The data are noisy with individual demands occurring on a finer time scale than the one-hour sampling interval. A median filter was used to interpolate values of for sample times with missing observations and to smooth the values of some outliers.

The one-hour sampling rate for these data is relatively coarse, and in order to make the parametric model fitting more stable, the observations are expanded to a 30 minute sampling rate using a piecewise cubic Hermite interpolating polynomial (PCHIP) as implemented in the Matlab function *interp*1 (Matlab (2012a)). Interpolation with the PCHIP algorithm was chosen as this expansion does not change the character of the observations, it merely adds more observation points at intermediate time steps with values contained inside the range of measured values to improve stability of the GMM fits.

A subset of the total data set collected is used here for analysis. This subset consists of 85 service connections that represent the locations with the fewest missing data due to meter malfunctions. An example week of these data are shown in Figure 1. A number of features regarding the data set can be seen in Figure 1: 1) There is a strong diurnal pattern to the water demands with relatively low use during the early morning hours; 2) A number of the sensors show a difference between weekend and weekday use – the first two days of the week in Figure 1 are Saturday and Sunday; and 3) There are some sensors with anomalous readings, for example, sensor 36 exhibits daily minimum demands that are considerably greater than zero as evidenced by the colored line extending across the one-week time period (Figure 1)

A practical consideration for classification is definition of the '"water day" as the 24 hour period beginning and ending at the time of daily lowest flow. Lowest flows for residential demands typically occur in the early morning hours. For the week shown in Figure 1, 05:00 appears to be the lowest flow period across all sensors. Visual examination of observations from other weeks supports 05:00 as the time of lowest flow and classification of daily demands are done using the water day between 05:00 and 05:00. This average low flow time is later than observed for other systems and is thought to be influenced by early morning filling of on-site storage at the majority of service connections.

For each service connection, observations were averaged in time to represent a mean demand pattern for that connection. The averaging was done over a calendar month and only weekday demands are considered resulting in 20 weekdays being used in each average for the January data. Figure 2 shows three example sensors with 20 days of daily demands and the average demands. A GMM is then fit to these average demand patterns using the models and EM estimation as described above. Figure 3 shows results of fitting GMM's to the three average demand patterns shown in Figure 2. For the results in Figure 3, M, the number of Gaussian distributions, is fixed at three.

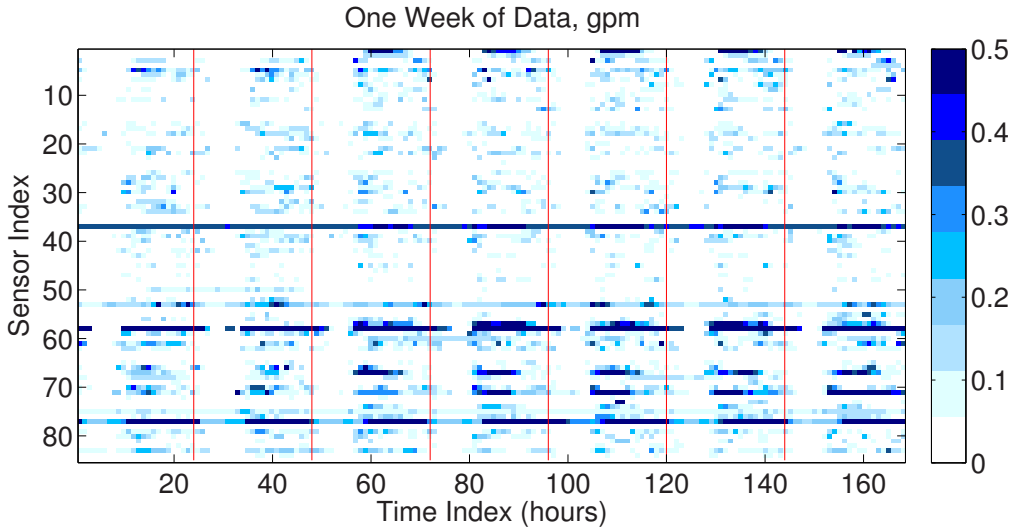

Fig. 1: Example demands for all sensors for the week of January 1st, 2011

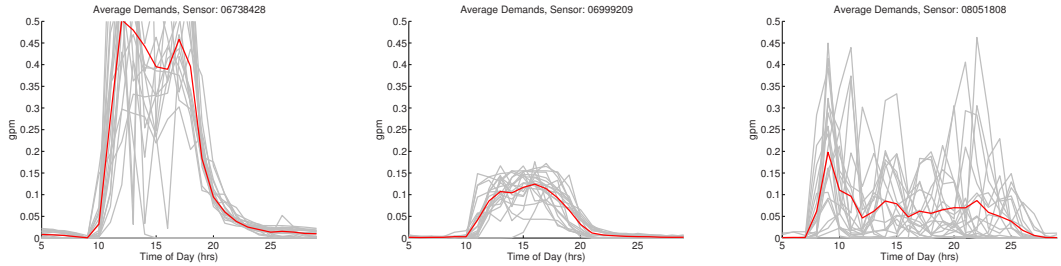

Fig. 2: Individual daily demands for 20 weekdays (grey) and the average demand (red) are shown for three different sensors.

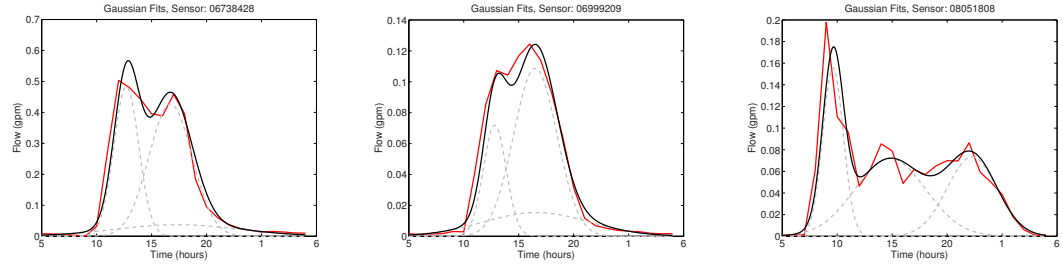

Fig. 3: Gaussian Mixture Model fits to average demands for three sensors. The red line is the observed data (20 day average), the black line is the GMM fit and the grey lines are the three component Gaussian models. Note the variation in the Y-axis range

3.2. Classification Parameters

The number of Gaussian distributions necessary to adequately fit the daily demand patterns is unknown and the patterns for some days may be well fit with a smaller number than the patterns for other days. Here we are looking for a constant number of Gaussian components that can adequately fit all days. This value is determined by fitting GMM's with, 2, 3, and 4 components and then using the corrected Akaike Information Content (AICc) as a quantitative

Table 1: Summary of AICc distributions for varying numbers of Gaussian components

Gaussian Components	2	3	4
k	6	9	12
AICc Median	11,888	11,808	11,811
AICc IQR	805.4	759.9	770.3

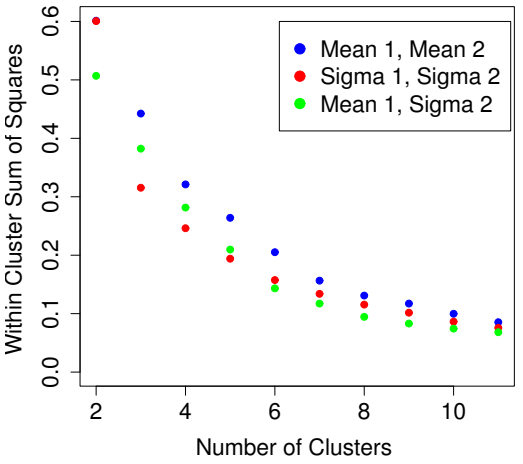

Fig. 4: Within Cluster Sum of Squared Errors as a function of the number of clusters.

measure. The AIC is defined as: $AIC = 2k - 2ln(L)$ where k is the number of parameters, here three times the number of Gaussian components, and L is the maximized value of the likelihood function. AICc is the corrected AIC (see Burnham and Anderson, 2002) to account for smaller sample sizes:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{8}$$

where n is the sample size. Here $n = 48$ to account for the additional interpolated points used in the GMM fits. The weekday data from January 2011 were used to create an average demand pattern for each service connection and these patterns were fit with GMM's having 2, 3 and 4 components. Table 1 shows a summary of the AICc values across all 85 sensors. Using three Gaussian components provides the lowest median AICc value and the tightest inter-quartile range (IQR) and, therefore, GMM's with three components are used for the analyses here.

For the three-component GMM, there are nine dimensions to the feature space comprised of μ, σ and α for each component. The feature vectors have been ordered in increasing value of the μ's for each service connection such that the Gaussian component that occurs earliest in the day is number 1 and the latest is number 3. The ability of subsets of these features to separate the demand patterns between the different service connections was examined visually and quantitatively. In summary, high-dimensional subsets of the features did not improve the classification process significantly and it was determined that a two-dimensional feature space would suffice.

The questions of which pair of features to select and how many clusters to use are answered simultaneously. The normalized total sum of squared errors within the clusters is shown as a function of the number of clusters in Figure 4. In an ideal setting, the normalized total sum of squared errors as a function of the number of clusters would show a distinct *elbow* beyond which the addition of more clusters would only provide minimal decrease in normalized total

sum of squared errors. That ideal case does not appear in Figure 4, where the normalized total sum of squared errors continues to decrease smoothly with increasing numbers of clusters. Given the relatively small data set of 85 service connections, the decision was made to use four clusters. From Figure 4, the feature pair that minimizes the normalized total sum of squared errors for four clusters are the standard deviations of the first two Gaussian components: σ_1 and σ_2. Other feature pairs beyond those seen in Figure 4 were also examined.

3.3. Identifying Demand Patterns

The parameters identified in the previous section are applied to data from January 2011. The weekends are excluded from the analysis and the remaining 20 days are clustered into four classes with the resulting patterns shown in Figure 5. For each pattern, the number of members and the total daily demand are shown above the pattern. The four patterns here contain between 16 and 32 of the 85 total service connections. The overall pattern is shown in black with the first, second and third Gaussian components shown in red, green and blue, respectively. it is noted here that all GMM's are fit to the observed patterns in a normalized space that only uses information on the pattern of demand and does not use any information on the actual amount of demand. After the clustering, the average daily demand of the service connections within each cluster is calculated and the area under the pattern curve is set to be equal to that average demand amount.

There is no customer information available for the service connections in this data set, so confirming the types of customers that belong to a given pattern is not possible. However, the patterns shown in Figure 5 may represent both residential and commercial customers. A residential pattern where the water consumers are gone for much of the day to jobs and school will show a sharp morning peak and a more dispersed evening peak with a minimum between the two peaks such as demonstrated by Cluster 4 (lower right, Figure 5). A different residential pattern with someone generally at home throughout the day still has the morning and evening demand peaks, but without a well-formed minimum in between them as possibly represented by patterns 1 and 2 (top row, Figure 5). A commercial customer with more regular water consumption will show a relatively uniform distribution of demand throughout the business hours as may be represented by pattern 3 (lower left, Figure 5).

The locations of the cluster means and the individual cluster members within the feature space are shown in the left image of Figure 6. The solid triangle symbols show the location of each cluster mean. Figure 6 (left image) shows that three of the four clusters are characterized by the standard deviation of the first (earliest) Gaussian component (Sigma 1) being relatively small, ≤ 2.0. This result corresponds to the relatively sharp early peak in demand as seen in clusters 1, 2 and 4 in Figure 5, Only cluster 3 in Figure 5 has the first Gaussian component with standard deviations > 2.5. Figure 5 also shows that the means of the Gaussian components range from 10:00-12:00, 14:00-17:00 and 20:00-22:00 across the four patterns.

An obvious feature for classification is the total demand at each service connection, and previous authors (e.g., Aksela and Aksela, 2009) have done one-dimensional demand classification using the total demands. For comparison, we also use total demand as the only dimension in the classification. A log10 transform of the total demand at each service connection was made and three uniformly spaced thresholds within the log10 transformed space: 1.25, 1.75 and 2.25, were used to classify the service connections into four classes. These thresholds correspond to 17.8, 56.2 and 177.8 gpm, respectively. The resulting four classes are shown in the original two-dimensional feature space in Figure 6 (right image). Clearly, the patterns identified by the GMM fits are quite different than those identified by thresholding the total demand at each node.

The stability of the identified clusters over time is examined by applying the same GMM approach along with k-means clustering to all 6 months of data (February through June) and then comparing the position of the cluster centroids within the feature space for each month. Due to some missing records in the May and June, the number of sensors used in the analysis was decreased from 85 to 77 and this change has a significant impact on the cluster centroids for January. The four clusters formed with the reduced sensor set include a second cluster with a Sigma 1 coordinate > 2.0 (Figure 7). The results in Figure 7 show that the centroids of the four clusters are stable over time with the convex hull of any set of cluster centroids being able to be drawn without including any centroids from other clusters. The closest any two convex hulls come to intersecting is between clusters 2 and 4 (red and cyan in Figure 7) with the January centroid for cluster 2 (red) being close to the cluster 4 group (cyan).

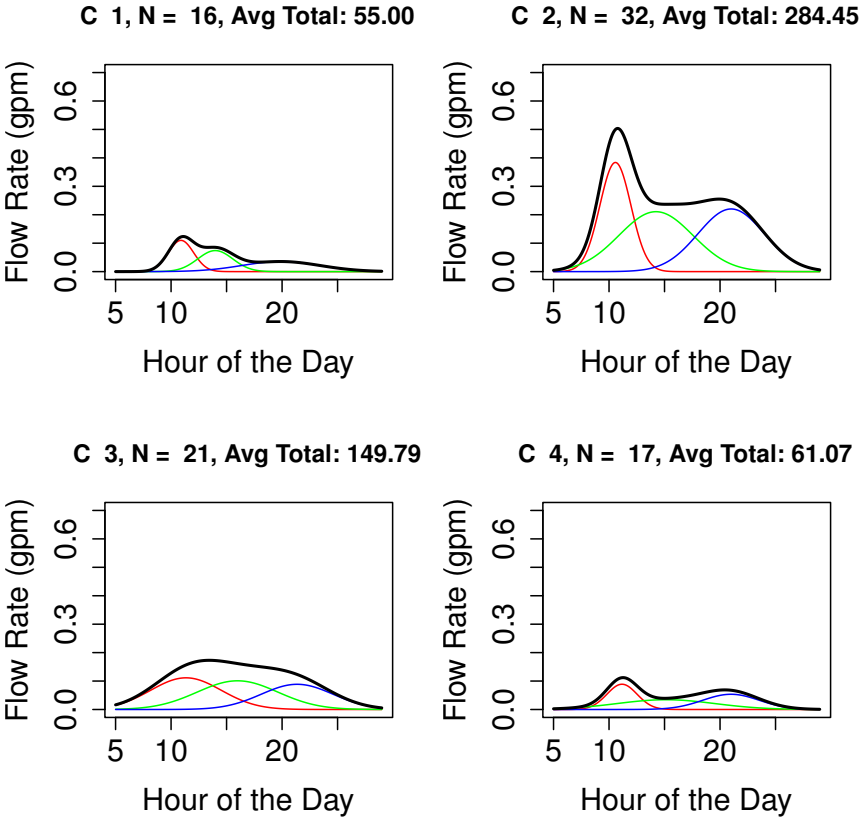

Fig. 5: Four clusters and their three-component GMM's for the January 2011 weekday data.

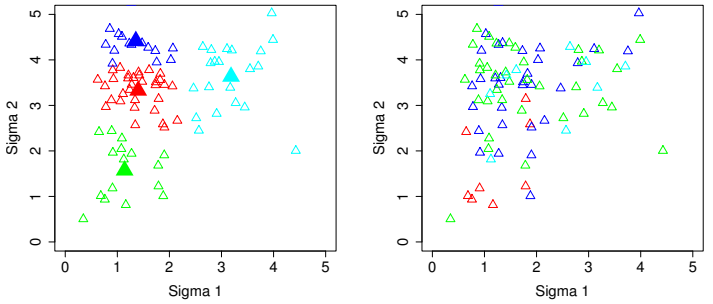

Fig. 6: January data clustered with GMM's (left) and thresholding total demand (right).

4. Conclusions

This work presents a new means of classifying water demand data using GMM's and k-means clustering. The combination of these approaches provides a flexible approach to fitting demand data as obtained with relatively high

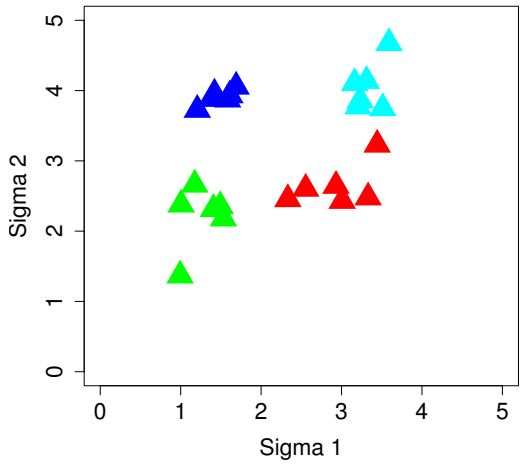

Fig. 7: Monthly cluster centroids within the feature space for January through June.

frequency sampling rates. GMM's with three components were used here to fit a wide variety of daily demand patterns. The parameters of these GMM's then serve as the basis for multivariate clustering using the k-means algorithm. In the application here, a relatively low-dimensional space utilizing the standard deviations of the first two Gaussian models was adequate to classify the demands into distinct patterns. These patterns remained as distinct monthly groupings over the the six months of the study period.

This approach to demand classification is based on the pattern of use throughout the day and therefore provides some insight into how the actual customers use water at different times of the day. This approach is quite different from that of classification based on total demand at each connection, which provides no information on when that use occurs during the day. Depending on the goals of the demand classification, one approach or the other may best. For example, in designing the size of a new tank, the fraction of each class as determined by the total demand classification may be best, while predicting the amount of water that needs to be available at any time of the day maybe better served by the demand patterns modeled with the GMM's.

The raw data examined here are noisy with considerable variation between demand patterns recorded at the same connection on different days. Here, we averaged all weekday demands at each connection over a full month to get a less noisy pattern for model fitting. Questions still remain as to how much averaging over time is necessary and whether or not different weekdays and/or weekend days should be included in the same average. These questions will be examined in future work.

Acknowledgements

The authors are grateful to the Dublin City Council (DCC) for providing the test data used in this study.

References

Adamowski, J.F., 2008. Peak Daily Water Demand Forecast Modeling Using Artificial Neural Networks, Journal of Water Resources Planning and Management, 134 (2), 119-128.

Aksela, K. and Aksela, M., 2011. Demand Estimation with Automated Meter Reading in A Distribution Network, Journal of Water Resources Planning and Management, 137 (5), pp. 456-467.

Bilmes, 1989. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, TR-97-021, International Computer Science Institute, Berkeley, California, 247 pp.

Blokker, E.J., Vreeburg, J.H.G., and van Dijk, J. C., 2010. Simulating Residential Water Demand with a Stochastic End-Use Model, Journal of Water Resources Planning and Management, 136 (1), pp. 19-26

Burnham, K.P. and Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd ed.), Springer-Verlag.

Ghosh, A.K., Lubkeman, D.L., Downey, M.J. and Jones, R.H., Distribution Circuit State Estimation Using a Probabilistic Approach, IEEE Trans. Power Syst., 17 (3), pp. 621-625.

Heunis, S.W. and Herman, R., 2002. A probabilistic Model for Residential Consumer Loads. IEEE Trans. Power Syst. 17 (3), pp. 621-625.

Irwin, G.W., Monteith, W. and Beattle, W.C. 1986. Statistical Electricity Demand Modeling from Consumer Billing Data. Proc. Inst. Elect. Eng., Gen., Transm., Distrib., vol. 133, no. 6, pp. 328-335.

Jain, A. and Ormsbee, L.E., 2002. Short-term Water Demand Forecast Modeling Techniques - Conventional Methods versus AI. Journal of American Water Works Association. 94 (7), pp. 64-72.

Matlab, 2012. MATLAB and Statistics Toolbox Release 2012a, The MathWorks, Inc., Natick, Massachusetts, United States.

R Core Team, 2012. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Singh, R., Pal, B.C., Jabr, R.A. 2010. Statistical Representation of Distribution System Loads Using Gaussian Mixture Model. IEEE Transactions on Power Systems 25(1), 29-37.

Xu, R and Wunsch, D.C. 2009. Clustering. John Wiley & Sons Inc., Hoboken, New Jersey.