Using Smart Meters to Learn Water Customer Behavior

Michele R. B. Malinowski and Richard J. Povinelli

***Abstract*—This article addresses the need to divide a population of water utility customers into groups based on their similarities and differences, using only the measured flow data collected by water meters. After clustering, the groups represent customers with similar consumption behavior patterns and provide insight into "normal" and "unusual" customer behavior patterns for individually metered water utility customers within North America. The contributions of this work not only represent a novel work, but also solve a practical problem for the utility industry. This article introduces a method of agglomerative clustering using information theoretic distance measures on Gaussian mixture models within a reconstructed phase space, designed to accommodate a utility's limited human, financial, computational, and environmental resources. The proposed weighted variation of information distance measure for comparing Gaussian mixture models emphasizes those behaviors whose statistical distributions are more compact over those behaviors with large variation and contributes a novel addition to existing comparison options. We conduct several experiments with both synthetic and real data to show the reasonableness of the clustering results and their consistency.**

***Index Terms*—Artificial intelligence, artificial intelligence for technology management, data analytics, environmental issues in technology management, smart services.**

I. Introduction

THIS article addresses the need to divide a population of water utility customers into groups using only the measured flow data collected by water meters. We introduce a novel method of agglomerative clustering using information theoretic distance measures on Gaussian mixture models (GMMs) within a reconstructed phase space (RPS). The proposed weighted variation of information (wVI) distance measure for comparing GMMs places emphasis upon those behaviors whose statistical distributions are more compact over those behaviors with large variation.

Since 2011, more than 25% of the U.S. has coped with drought conditions. In California, one of the most severely affected areas, over 45% of the state has experienced drought conditions over the same period, increasing to over 90% for 2016 [1]. In response to the long-term drought, municipalities have responded with conservation ordinances introducing severe restrictions of water use including irrigation system flow limits, watering date/time restrictions, and punitive monetary fines for violations [2]. These restrictions and conservation projects require timely water consumption data and processing to enforce the ordinances, as well as educational and targeted communication with the water consumers.

The water utilities need an easy method to identify which customers' behavior is within the accepted normal patterns, and which customers' behavior is wasteful, fraudulent, or in violation of regulations. The unsupervised clustering algorithm presented in this research fills the need for grouping customers by behavior. This assists the utility to determine customers needing additional scrutiny and those that do not. The output of this algorithm is a hierarchical diagram grouping all customers compared with each other using an information-theoretic distance measure based on the temporal behavior patterns observed within the collected flow measurements.

The unsupervised clustering algorithm described in this article addresses the need to divide a population of water utility customers into groups based on their similarities and differences, using only the measured flow data collected by water meters. Two motivations drive this article—a commercial motivation to provide useful segregation of customer data, and an academic motivation to create a new method of comparing two time series. The agglomerative clustering method accommodates the practical limitations of a utility's finances, resources, or staffing. This allows the customer data to be segmented into groups for targeted marketing, conservation campaigns, or incentive programs based on the water usage data. As time-of-use billing is not yet widespread in water systems, the consumption volume, patterns, and flow rates are more critical to identify groups of customers. The academic contribution of this work, the wVI distance measure, presents a novel component-weighting scheme for emphasizing components of GMM with compact distributions.

The remainder of this article is organized as follows. Section II describes the water supply industry and related work. Section III explains the method for clustering customers based on their water consumption. Section IV describes the data used for our experiments. Section V presents the experiments and results. Finally, Section VI concludes this article.

II. Background

While over 15 million American households rely upon private well sources for water [3], the remaining 110 million households

Manuscript received May 30, 2019; revised October 25, 2019 and February 25, 2020; accepted May 13, 2020. Review of this manuscript was arranged by Department Editor T. Hong *(Corresponding author: Richard J. Povinelli.)*

The authors are with the Department of Electrical and Computer Engineering, Opus College of Engineering, Marquette University, Milwaukee, WI 53201-1881 USA (e-mail: michele.malinowski@marquette.edu; richard.povinelli@marquette.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TEM.2020.2995529

are connected to public water supplies. Likewise, most commercial and industrial applications use public water supplies. Public and municipal water utilities must carefully monitor the water they provide for public safety, billing, and resource management.

Over the last few decades, water utility companies have begun installing automated meter reading (AMR) systems to further simplify the process of meter reading, decrease manual labor, and reduce transcription errors within collected data [4]. These systems allow more frequent reporting of measured demand at the individual customers, while simultaneously reducing the manual effort of physically looking at each meter to record the volume measured. In 2018, the American Water Works Association reported 37% of utilities in North America have fully implemented AMR systems, and another 24% are in the process of doing so [5]. Many of the AMR systems support quarter-hourly reads, but battery limitations and data-related costs constrain the data collection to hourly or daily reads. It is from these AMR systems that the data for our proposed algorithm comes.

While there is little work in customer water flow clustering, other research explores clustering of energy customers using smart meter data. Panapakidis *et al.* [6], [7] implement clustering of electric smart meter data. As opposed to creating models such as our algorithm, their work clusters the daily typical load profiles within a customer's dataset. Representatives of those clusters are used to complete the second stage clustering across the population of all customers. Their work illustrates the complex problem of identifying the optimal number of clusters in a diverse dataset. In contrast to the Panapakidis work, the clustering method presented here does not require a definition of an optimal number of clusters.

Bose and Chen [8], [9] track changing cluster populations over time using fuzzy c-means algorithms. Their work focuses upon migratory patterns of cellular phone customers, for the purposes of tracking dynamic market demands and customer retention. Their data exhibit not only customers who migrate from one cluster within the data to another, but also the formation of new clusters and dissolution of others as new behavior patterns emerge within the population.

A related problem arises in clustering music. Genre classification is not an identical problem, as the entirety of the work is available at time of classification. The whole song is already produced and recorded, but similarity exists in the approach to first model the music, and then compare the model with others during the classification step. Logan and Salomon [10] create models using the audio spectrum of the composition and then cluster multiple works using earth movers distance. Jensen *et al.* [11] create GMMs from the Mel frequency coefficients within a work and then cluster those models based on three different distance measures—Kullback–Leibler distance, earth movers distance, and normalized least squares.

Another popular algorithm, spectral clustering, simplifies the problem by reducing the dimensionality in a different manner. First, the similarity matrix is constructed as a representation of the commonalities between every pair of data samples. Then, a graph Laplacian is computed from this similarity matrix. The clustering operates on eigenvectors from this graph Laplacian matrix and some predetermined clustering algorithm such as k-means or c-means. Spectral clustering algorithms vary on the specific details of constructing the graph Laplacian and the clustering step, but the same framework applies [12]–[14].

Statistical modeling of biological time-series has been applied to electrocardiogram data for classifying specific heart rhythms [15], [16]. This work casts the time-series signals into an RPS and further applies GMMs to represent the attractor within the RPS. These models then classify a new time-series as a particular heart rhythm, aiding in medical diagnosis. Our clustering method used to group water meter time series is similar to that of [13] and [17]. We extend their work by clustering different customer models using the VI distance measure.

Some existing research classifies water usage based on metering data. Laspidou *et al.* [18] use quarterly water billing information and self-organizing maps to cluster customers based on consumption. Willis *et al.* [19] and Cardell–Oliver [20]–[22] investigate fixture-level consumption patterns to identify specific end uses of water in a location using high-resolution metering. Other research focuses upon partitioning a utility's entire water distribution network into the optimal district metered areas for processing groups of customers sourced by the same supply mains [23], [24]. Related research using smart meter flow data has produced outlier detection and forecasting algorithms [25]–[27] and leakage detection methods [28]. In contrast to this existing water utility research, our article focuses upon clustering similar customers based on temporal behavior patterns using only the hourly flow measurements recorded in the BEACON AMA system.

III. Method

Our approach transforms the water meter time series into an RPS. A GMM is learned on this space. Next an ellipsoid hull is computed to model each GMM component, and a geometric tessellation of the hull is calculated within the RPS. To compare two customers, the variation of information (VI) distance is calculated between the hulls of each customer. The VI distance between customers is then used to build an agglomerative cluster hierarchy.

When comparing large sets of raw time-series data, the computational and storage space requirements quickly become unmanageable. Reducing the large set of individual measurements to a small set of model parameters for each customer makes the comparison more manageable. One way to study these systems is to cast the time series into an RPS such that any location within the space identifies the system state at that moment. Phase space embedding is an established method to represent a system in a vector space chosen to illustrate the dynamics of the original system [29].

A GMM is learned on the RPS. An ellipsoid hull is then computed to model each GMM component for a customer, and a geometric tessellation of the hull is calculated within the RPS. The volume of this ellipsoid hull estimates the entropy of this component of the GMM. A summation of all GMM component hull volumes estimates the entropy of the customer model. If a customer has perfectly consistent water consumption behaviors, the associated GMM component volumes will be small. As variations in the consumption behavior or temporal patterns increase, the GMM component volumes will also increase.

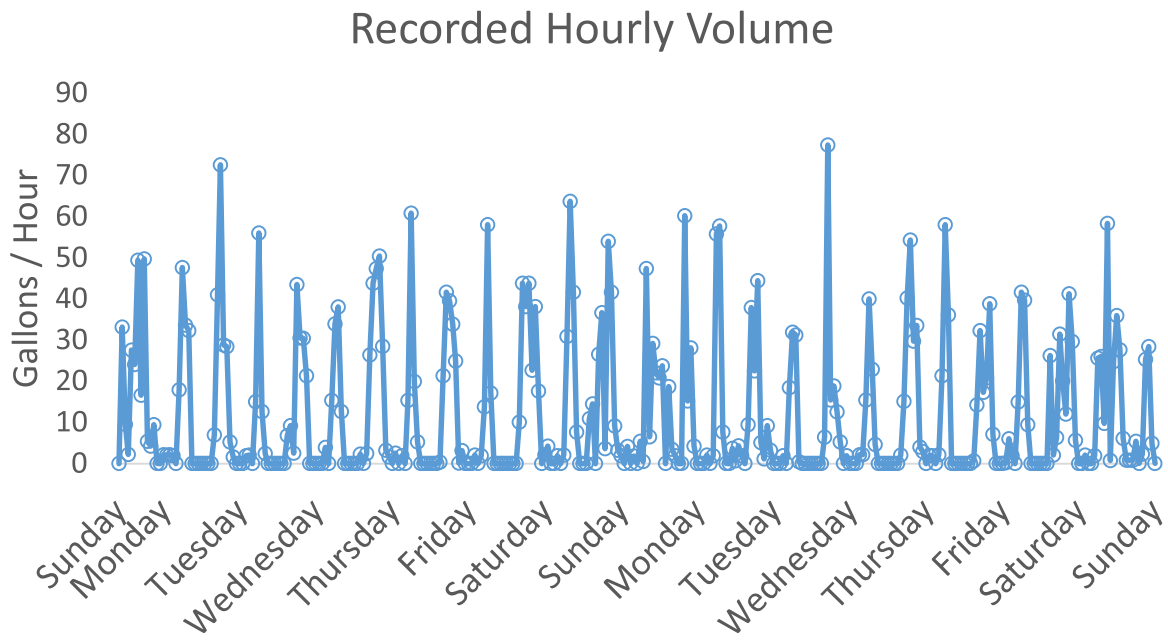

Fig. 1. Recorded hourly consumption values for a residential customer.

VI is a measure of how much information is lost when combing two groups as opposed to keeping the groups separate. Unlike mutual information (MI), VI is a true metric, satisfying the triangle inequality and allowing comparison of clusters with different populations. The VI corresponds to the nonoverlapping volume of two convex hulls in the RPS. If the two hulls are coincidental, the VI is small, and combining the two into a new cluster loses very little information. Conversely, if the two hulls are entirely separate, the VI is large and reflects the large loss of unique information if they are combined. Hulls that overlap partially or touch will fall somewhere in between these two extremes.

Two customer models are compared to each other by computing the points of intersection of the GMM component hulls. When the surface of one hull is located within the enclosed volume of a second hull, an intersection is present. Consider the intersection of a large and a small ellipsoid. A boundary for the intersecting volume is created by first identifying the set of points on the surface of the large ellipsoid that exist within the volume of the smaller ellipsoid. Then, we identify the reciprocal set of points on the surface of the small ellipsoid that exist within the volume of the larger. A new convex hull is created from the combined set of intersecting points.

Since the summation of all model component volumes enclosed within a hull estimates the entropy of the customer model, the summation of all the intersecting volumes between two customer models is the estimated MI between those two customer models. The VI is then the sum of all volumes from both models subtracting double the volume of the MI. The VI distance is then used to cluster customers. We avoid the problem of determining the number of clusters by using an agglomerative cluster hierarchy. Then post hoc the number of clusters can be selected accommodating the utility's resource limitations.

The remainder of this section presents the data normalization and cleaning process. Then the details of the method are presented. This includes the definitions of VI, wVI, RPS, and GMM.

A. *Cleaning and Normalizing Data*

The algorithm is not designed to accommodate large gaps in the data. During the data cleaning process, the longest continuous set of hourly measurements without any missing, aggregated, or negative flow data is selected for evaluation, discarding other data. The BEACON AMA system, which is used for data collection, can disaggregate values naïvely, but it stores an internal flag for those records, permitting the detection and removal of disaggregated data that would otherwise affect the overall performance of the method.

Water meters in residential properties sit idle for many hours of the day while the occupants are at work or asleep, resulting in most recorded data indicating zero volume. Likewise, many commercial meters sit idle during the evening or early morning hours when the business is closed. Fig. 1 shows a few days of recorded hourly consumption values for a residential customer. The periods indicating zero consumption coincide with time spent sleeping or at work during the weekdays. Fig. 2 shows a histogram of the recorded values, as well as the overall median and nonzero median, which are described below.

The zeros themselves are not unusual, but the number of zero measurements can create computational problems. Traditional normalization by median or mean is deceptively small, due to the large quantity of zeros in this data; or erroneous, due to divide-by-zero errors. Instead, the median of nonzero values is used—this is the number associated with the 50th percentile of consumption for all nonzero consumption records. Using

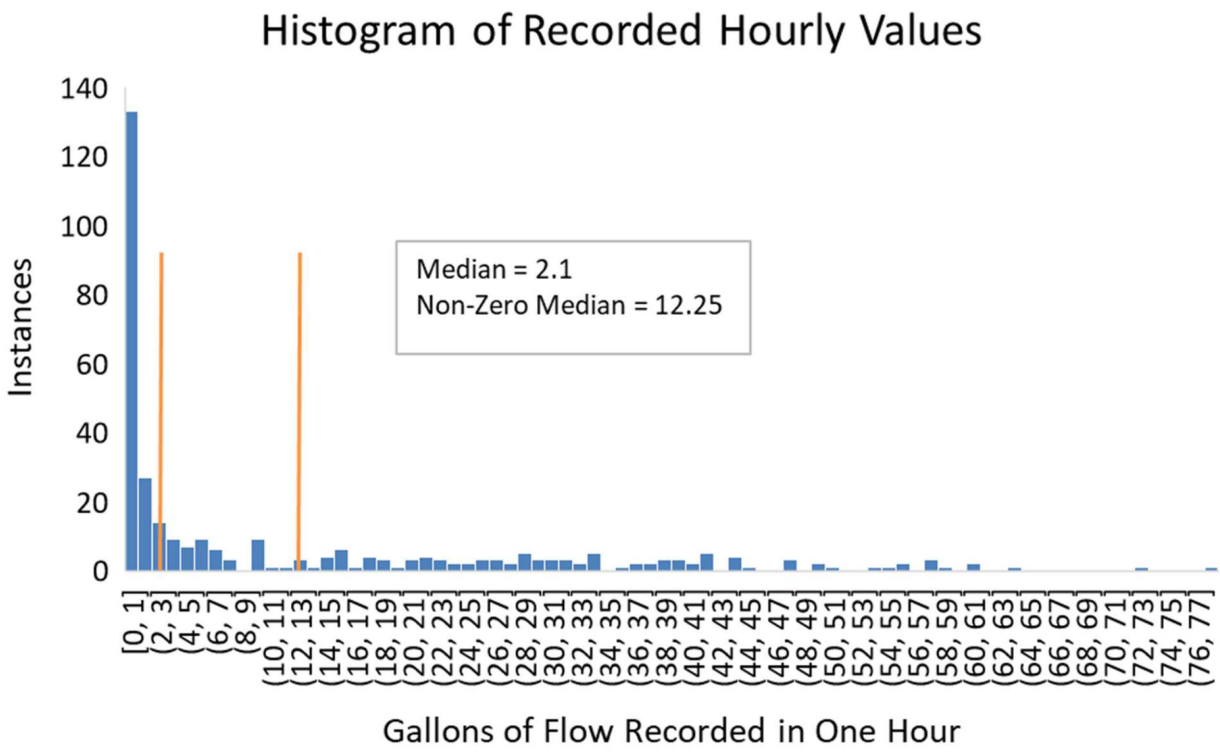

Fig. 2. Histogram of values for a residential customer.

the time series $\mathbf{X} = \{\ldots, x_{t-\tau}, x_t, x_{t+\tau}, x_{t+2\tau}, \ldots\}$, take the median of the set of $\mathbf{X}$, excluding values in $\mathbf{X}$ equal to zero. Then $\mathrm{Med}_{\mathbf{X}\neq 0} = \mathrm{Median}(\mathbf{X} \cap \overline{0})$. $\mathrm{Med}_{\mathbf{X}\neq 0}$ is the nonzero median. Dividing the original data by $\mathrm{Med}_{\mathbf{X}\neq 0}$ produces the normalized data: $\mathbf{X}_{\mathrm{normalized}} = \mathbf{X}/\mathrm{Med}_{\mathbf{X}\neq 0}$.

Normalizing the recorded values for each customer in this manner allows comparison between customers of different size (number of household members or size of business). The comparison then identifies common behavioral patterns, regardless of the volume of the consumption pattern.

The next section describes the process of converting a time series to an RPS and using a GMM to model the RPS.

B. Gaussian mixture models (GMMs) in reconstructed phase space (RPS)

One way to study customer flow is to cast the time series into a vector space such that any location within the space identifies the system state at that moment [30]. Phase space embedding [30], [31] is an established method to represent a system in a vector space chosen to illustrate the dynamics of the original system most clearly. Fig. 3 illustrates the embedding of a few data points as an example. Two groups of repeated behaviors are shown, red dots indicate behaviors occurring on a 24-hour schedule, and blue dots indicate behaviors occurring on a weekly, 168-hour schedule. The embedding process, indicated by the colored arrows, shows how groups form within the vector space with axes corresponding to 0-, 24-, and 168-hour time lags.

When the time series is embedded into the phase space, a single point is defined as a vector $\mathbf{Y}$ of points from the original series $\mathbf{X}$ each separated by time lags T_m to produce the dimensions within the space. The subscript m indicates the particular dimension associated with that lag T_m: $\mathbf{Y} = [x_{t+\mathrm{T}_1}, x_{t+\mathrm{T}_2}, \ldots, x_{t+\mathrm{T}_m}]$. These vectors are plotted in the newly defined phase space as a topological embedding of the original system [31].

A GMM is learned over the RPS, where a GMM of k component Gaussians in d dimensions is

$$\hat{\mathbf{X}} = \sum_{i=1}^{k} \left(\mathcal{N}_d\left(\mu_k, \mathbf{\Sigma}_k\right)\right)_i. \tag{1}$$

The central location vector $\mu_k = [\mu_1, \mu_2, \ldots, \mu_d]$ describes each Gaussian component along with a $d \times d$ covariance matrix

$$\mathbf{\Sigma}_k = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix}. \tag{2}$$

The GMM means and variances are estimated using expectation maximization [32], where the initial GMM centers are generated using k-means clustering. The initialization process is stochastic, yielding different GMMs in each trial of the process. The consistency of the resulting GMMs is evaluated in the experiments and results section of this paper.

C. Variation of Information

To measure the similarity of two customers' GMMs the VI measure is used. VI is a measurement of how much information

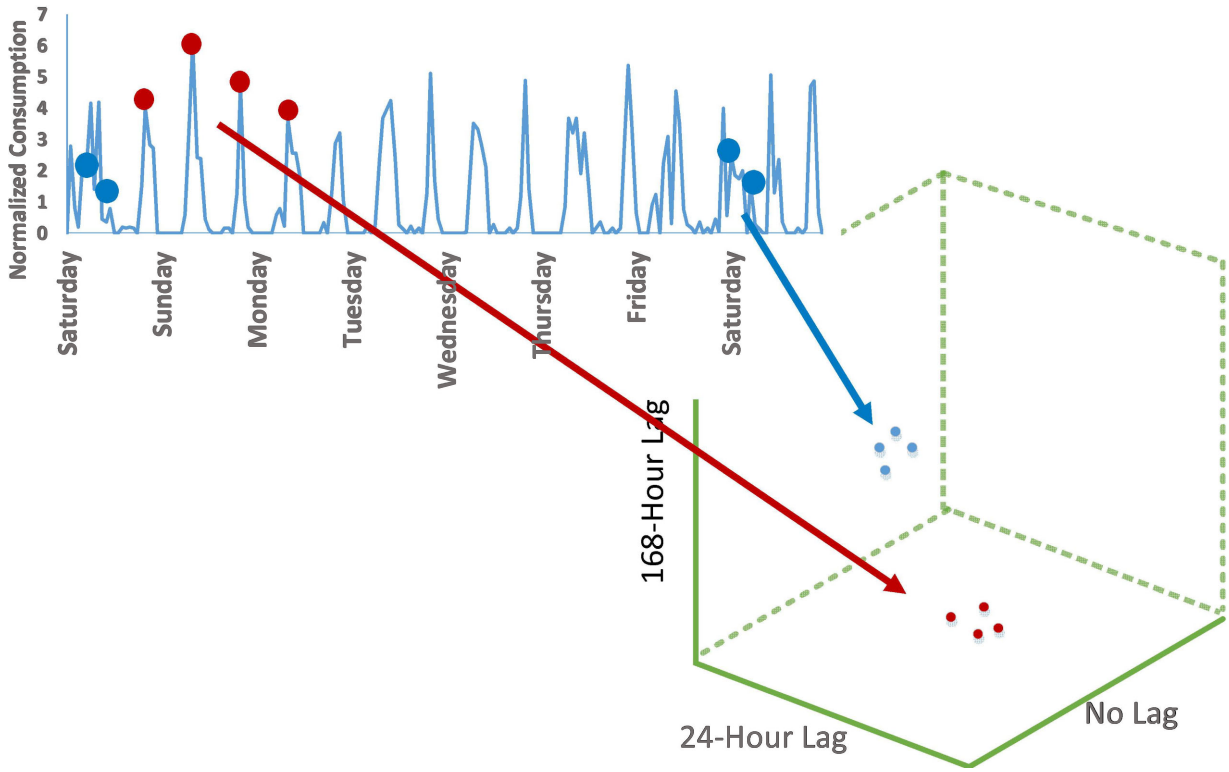

Fig. 3. Embedding from a time series into a vector space forces groups of points to form based on repetitive behaviors corresponding to the time lag.

is lost when combing two groups as opposed to keeping the groups separate. Unlike MI, VI is a true metric [33], satisfying the triangle inequality and also allowing comparison of clusters with different populations. We chose VI as a distance measure because of these desirable properties.

MI describes the amount of information known about one probabilistic model through knowledge of a second probabilistic model. MI is the information shared by the two models, defined as

$$MI\left(A,B\right)=\sum_{i}\sum_{j}P\left(a_i,b_j\right)\log\left(\frac{P\left(a_i,b_j\right)}{P\left(a_i\right)P\left(b_j\right)}\right). \quad (3)$$

The models A and B each contain one or more components, a_i and b_j. MI is illustrated in Fig. 4 and is used to compute the VI [14], [34], [35].

The VI distance between two sets is the sum of unique information that would be lost if the two sets are combined. With the MI as defined by (3) and individual entropy of each set

$$\begin{aligned} H\left(A\right) &= \sum_{i}P\left(a_i\right)\log_2\left[P\left(a_i\right)\right] \text{ and} \\ H\left(B\right) &= \sum_{j}P\left(b_j\right)\log_2\left[P\left(b_j\right)\right] \end{aligned} \quad (4)$$

the nonintersecting parts of all sets are collectively

$$VI\left(A,B\right)=H\left(A\right)+H\left(B\right)-2\left[MI\left(A,B\right)\right]. \quad (5)$$

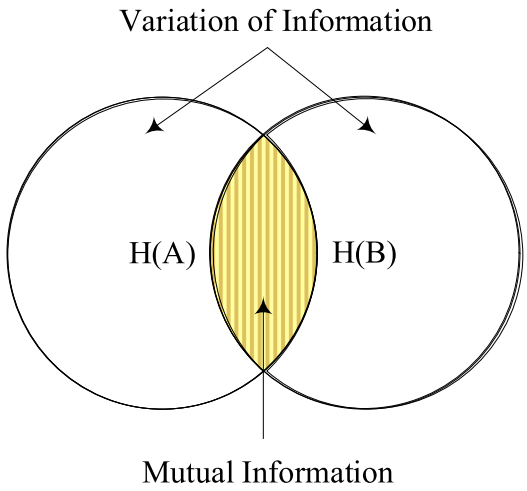

Fig. 4. Venn diagram describing relationships between entropy, MI, and VI.

The relationships between the individual entropy of each set, the intersection of the two sets (MI), and the VI are clarified in Fig. 4 Venn diagram.

D. Weighted Variation of Information

Adding component weighting to the VI yields better clustering consistency. This section describes the wVI variant of VI. Let the average of the trace of the covariance matrix from (2) be

$$\sigma_{avg}=\frac{\sum_{i=1}^{d}\frac{1}{\sigma_{dd}}}{d}. \quad (6)$$

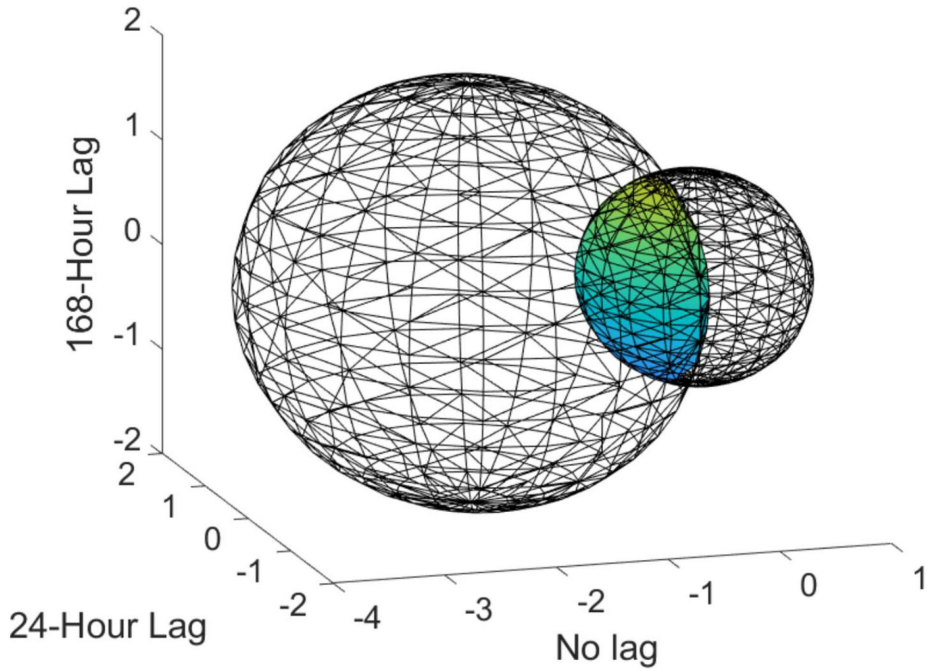

Fig. 5. Visualizing MI (intersecting volume) of two customer models.

Then the weight for the *k*th GMM component described in (1) is

$$w_k = \frac{\frac{1}{\sigma_{k\,avg}}}{\sum_{i=1}^{k} \frac{1}{\sigma_{i\,avg}}}. \tag{7}$$

Finally, apply these weights when computing the entropy, MI, and VI distance between two cluster models (cluster A and cluster B)

$$wH\left(A\right) = \sum_{i=1}^{n} H\left(A_i\right) w_{Ai} \tag{8}$$

$$wH\left(B\right) = \sum_{k=1}^{m} H\left(B_k\right) w_{Bk} \tag{9}$$

$$wMI\left(A,B\right) = \sum_{k=1}^{m}\sum_{i=1}^{n} MI\left(A_i,B_k\right) w_{Ai} w_{Bk}. \tag{10}$$

And the *wVI* distance is

$$wVI\left(A,B\right) = wH\left(A\right) + wH\left(B\right) - 2\left[wMI\left(A,B\right)\right]. \tag{11}$$

This weighting allows the clustering algorithm to emphasize smaller cluster components, those with a tighter standard deviation, and reduces emphasis of the components with a large average standard deviation.

E. Estimating VI

The VI distance metric describes the amount of information lost when two models are combined into a new model. As an example in three dimensions, Fig. 5 illustrates that the VI corresponds to the nonoverlapping volume of two convex hulls. If the two hulls are coincidental, the VI is small, and combining the two into a new cluster loses very little information. Conversely, if the two hulls are entirely separate, the VI is large and reflects the large loss of unique information if they are combined. Hulls that overlap partially or touch will fall somewhere in between these two extremes.

An ellipsoid hull is computed to model each Gaussian mixture component for a customer, and a geometric tessellation of the hull is generated within the RPS. The volume of this ellipsoid hull

$$V_k = \left(\frac{4}{3}\pi\right)\prod_{j=1}^{d} r_j \tag{12}$$

estimates the entropy of this component of the GMM, with r_j being the radius of the ellipsoid for any axis. The volume is computed based on the Qhull method [36]. A summation of all GMM component hull volumes

$$\hat{H} = \sum_{i=1}^{k}\left(\left(\frac{4}{3}\pi\right)\prod_{j=1}^{d} r_j\right)_i \tag{13}$$

estimates the entropy of the customer model. If a customer has perfectly consistent water consumption behaviors, the associated GMM component volumes will be small. As variations in the consumption behavior or temporal patterns increase, the GMM component volumes will also increase.

Two models are compared to each other by computing the points of intersection of the GMM component hulls. Fig. 5 shows a simple model with spheres, illustrating the intersection between the two hulls as a solid volume. Since the summation of all model component volumes enclosed within a hull estimates the entropy $H$of the particular customer model, the summation of all the intersecting (filled) volumes between two

customer models is the estimated MI between those two customer models. The VI is the sum of all volumes from both models (A and B), subtracting double the volume of the MI. Thus $\hat{V}I(A,B) = \hat{H}(A) + \hat{H}(B) - 2[\hat{M}I(A,B)]$. To perform hierarchical clustering, two models need to be joined. This is done by computing the hull of the combined customers' GMMs.

IV. Water Flow Data

Research data have been drawn from the cloud-based Badger Meter, Inc. BEACON Advanced Metering Analytics (AMA) system. This database of hundreds of utilities maintains historical records for meters with equipment details, measured flow, time stamps, and status information throughout the life of the meter. Hundreds of thousands of endpoints are tracked daily in the system. Unique identifiers for the meter, radio endpoint, and customer label each record within the system, but have been anonymized for this research, and no personally identifiable information is presented here. All records are used with permission from their respective owners.

The source data used in this study comprise a group of 99 m from a Midwestern utility with approximately four years of historical records archived within the BEACON AMA system. The utility was selected due to the longevity of the records and the approval for research purposes by both the company and utility. Experiments and examples using a single customer have been drawn from this collection as well. For this research, all customers are assumed to have resided in their homes for the entirety of the sample period, with no changes in ownership of a property. This is a naïve approach, and future work should investigate methods to identify changes related to ownership or commercial usage of a property.

All collected water records for this study have a 1-hour reading interval. Data collected from the BEACON AMA system includes the flow volume as well as status alarms from the meter, radio, and collector. These status alarms may include continuous flow, no reported flow, naïve disaggregation, and communication errors. Data with naïve disaggregation and communication errors are excluded from the study, while those with continuous flow and no reported flow are included. Only the recorded flow volume and time stamps are preserved as inputs to the clustering process. The other status alarms regarding flow are not used.

V. Experiments and Results

Unsupervised clustering operates without any data labels to confirm the results. Thus, there is no ground truth. To evaluate the cluster method, we look at the behavior of various types synthetic data, including phase shifts, simulated leaks, and randomization of the time series, to evaluate the effectiveness of the method.

Further evaluation is done on real customer data, where we look at consistency of the results. Consistency is used here to describe the stability of a particular outcome when the same data are clustered multiple times. Recall the stochastic nature of learning a GMM. For a clustering method to be valuable to utilities, the cluster populations must remain stable if the underlying behavior has no changes. This is determined by stability of individual customer cluster assignments with respect to other individuals and is discussed in the literature as cluster membership or migration of individuals within the data [8], [9], [37], [38]. The wVI distance shows consistent clustering results when run on the same original data multiple times, using a new GMM for each trial. Even the customers showing the most volatile placement and those customers with the furthest distance from the remainder of the data remained consistent.

A. Evaluation Using Synthetic Data

Despite the lack of labeled data, unsupervised clustering algorithms must still be tested. One approach is to create synthetic data with known labels as a substitute for identifying specific groups within the dataset. As the customer-grouping problem does not have specific labels without the reference to other customers, various data processing methods create synthetic customers who will be assigned "near" and "far" distances from their source data. Starting with actual customer data, we manipulate the individual hourly meter readings to represent customers that have similar behavior, different behavior, and leaks. Descriptions of the individual customer data used in all synthetic experiments are provided in Table I.

1) Synthetic "Similar" Customers: Shifting the entire time series forward or backward in time creates synthetic similar customers. This is equivalent to taking all the recorded meter data from a household and changing the time—instead of waking at 0745 and showering, the household now wakes at 0545. All behavior maintains the same volumes and temporal patterns. These customers will appear nearly the same when plotted in the RPS, as the method extracts behavioral time patterns, not specific times of use. The VI distance of these synthetic similar customers will be very close to the original customer. Fig. 6 illustrates the synthetic customer (orange) generated from the original customer data (blue) by shifting the time axis by approximately 30 h without changing any of the hourly flow values.

The results of clustering synthetic similar customers show a short join distance between the donor customer and the synthetic generated customers. Fig. 7 is a dendrogram illustrating the hierarchical clustering. To select the number of clusters, the dendrogram is cut at the join distance. If we want two clusters, we would cut at a VI distance of 1000. Alternatively, if we wanted three clusters, we would cut at a VI distance of 500. Fig. 7 illustrates clustering with four synthetic customers, all generated from the Customer 1 dataset. The labels indicate the type of operation used to create the synthetic data. Customers 002—005 are actual customers from collected data.

2) Synthetic "Different" Customers: Creating dissimilar customers requires changing the temporal behavior patterns within the recorded meter data. The simplest method is to draw a random permutation from the existing hourly data records, as illustrated in Fig. 8. Repeating this process multiple times creates a group of random customers with the same recorded consumption volumes as the original customer, but no discernable schedules associated with the time of day or day of week. Within the RPS, these random permutations have no obvious structure. In the hierarchical clustering, these three random permutations are

TABLE I.
CUSTOMER DESCRIPTIONS FOR SYNTHETIC CUSTOMERS EXAMPLES

Customer Title	Description
x001	Original data set from one customer, used as donor data to create synthetic data sets
x001 Backward Shift	Donor data has been shifted backward in time
x001 Forward Shift 1	Donor data has been shifted forward in time
x001 Forward Shift 2	Donor data has been shifted forward in time by a different number of hours
x001 Medium Leak	A leak of 2.3 gallons per hour has been applied to 200 consecutive hours with a random start time
x001 Random 1, 2, 3	Three different random permutations of the donor data
x001 Small Leak	A leak of 0.75 gallons per hour has been applied to 500 consecutive hours with a random start time
x002-x007	Original data sets from various customers

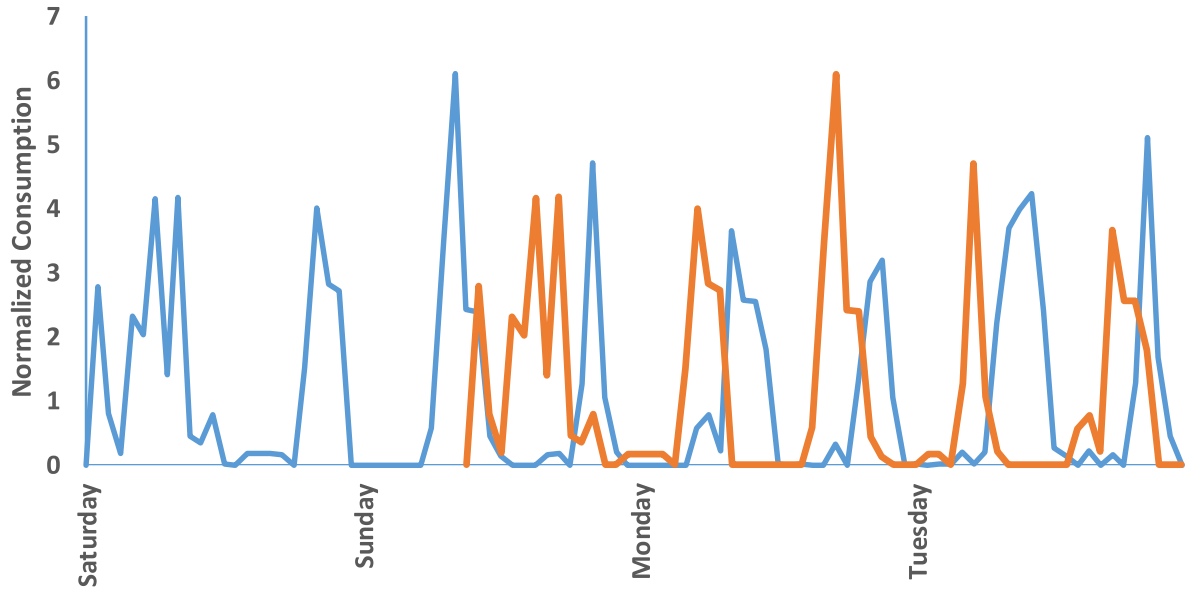

Fig. 6. Generating a synthetic "similar" customer through temporal shift.

expected to have small VI distances to each other, but large VI distances to the original customer who has daily or weekly behavioral patterns.

As the name implies, synthetic different customers tend to be grouped randomly far from the donor dataset. Fig. 9 shows these results. One of the random permutation synthetic meters is grouped near to the donor meter, while the other two are grouped further away. These results are not surprising, as random permutations occasionally form similarities that resemble the source. Descriptions of the individual customer data used for Fig. 9 are provided in Table I.

3) Synthetic "Leak" Customers: In the water industry, a leak is any unintended loss of water from the pressurized distribution system [39]. While much of the focus in the water industry has been on distribution network leakage [28], [40]–[43], consumer-side (after the meter) leakage is important to the individual residents and commercial accounts, as they must pay for the lost water and the maintenance caused by water damage [40], [44]. Leaks may occur when a mechanical failure has occurred in a fixture or pipe, such as a leaky valve on a commode, a failed weld on a pipe joint, or a worn seal on a faucet. Human error may also appear as a leak from the perspective of measured flow—forgetting to turn OFF an irrigation system. Due to the low

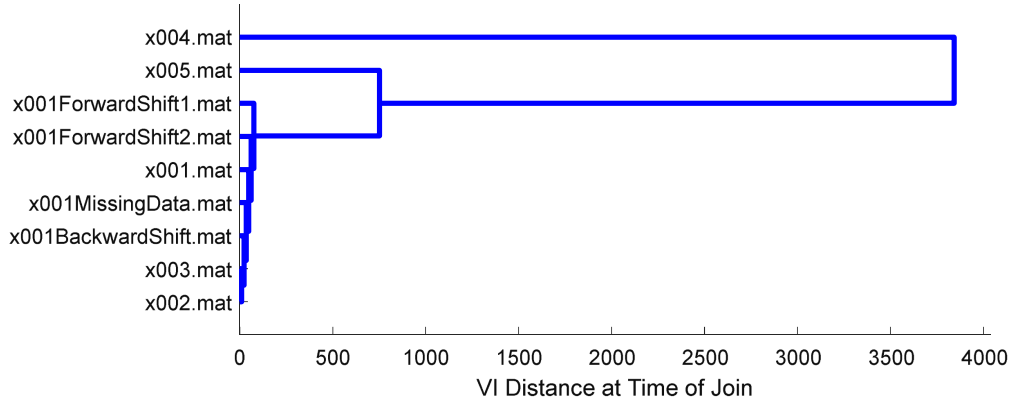

Fig. 7. Clustering of four synthetic similar customers and five actual customers.

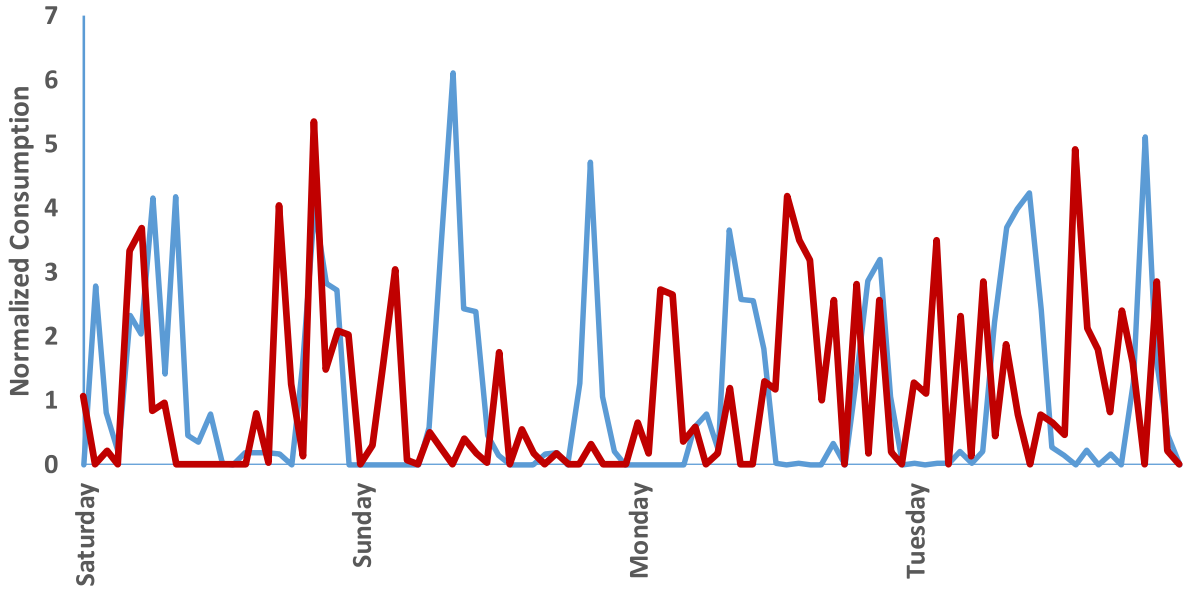

Fig. 8. Generating a synthetic "different" customer through random permutation of hourly flow measurements.

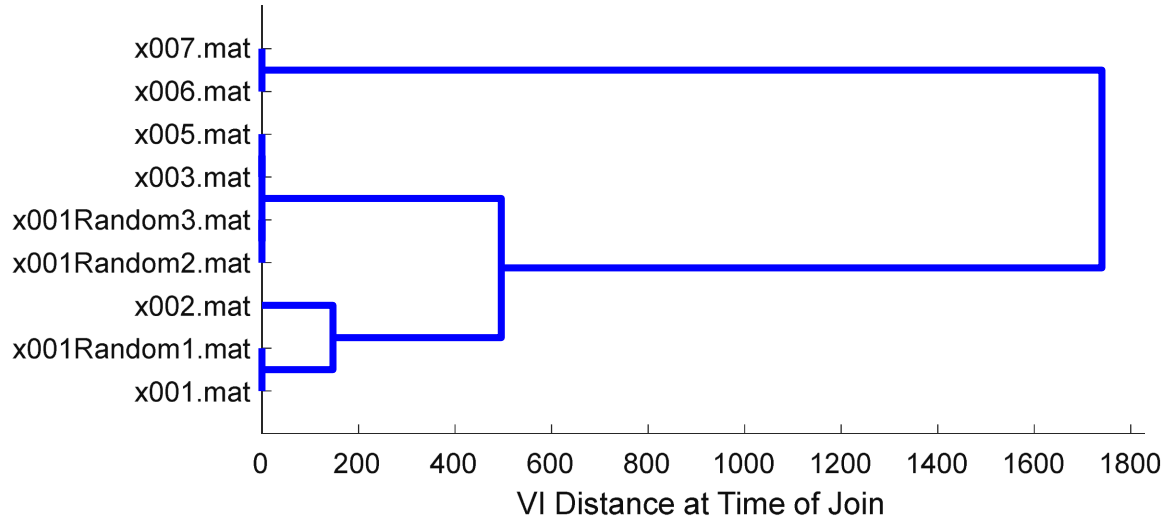

Fig. 9. Clustering of three synthetic different customers generated through random permutations of the time series, with six actual customers.

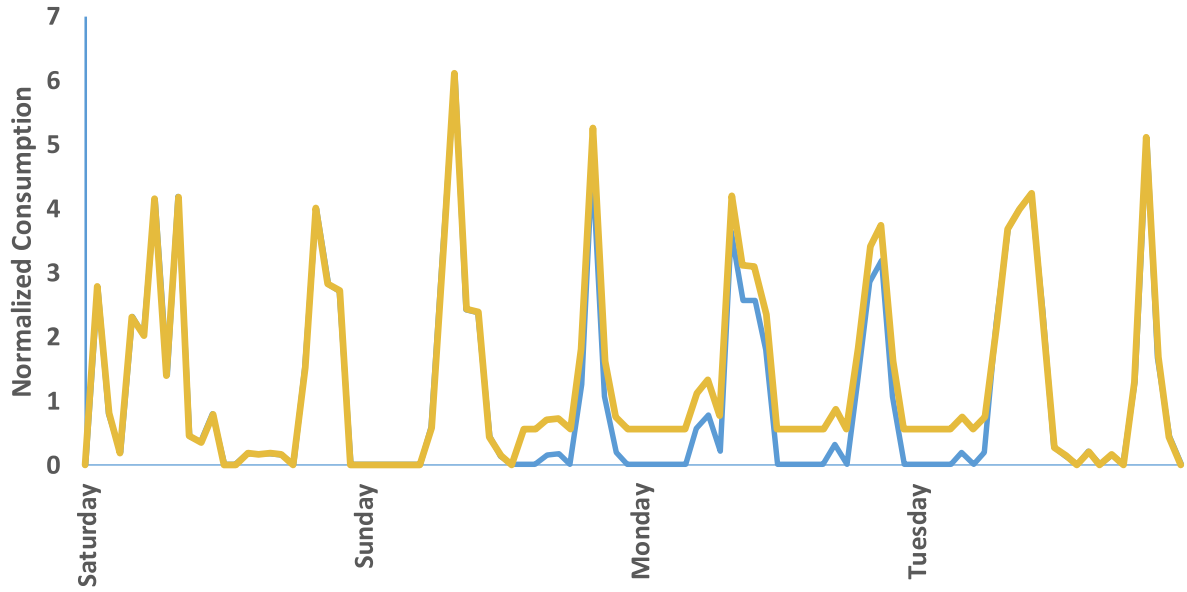

Fig. 10. Generating a synthetic "leak" customer by adding a fixed flow volume for a random duration.

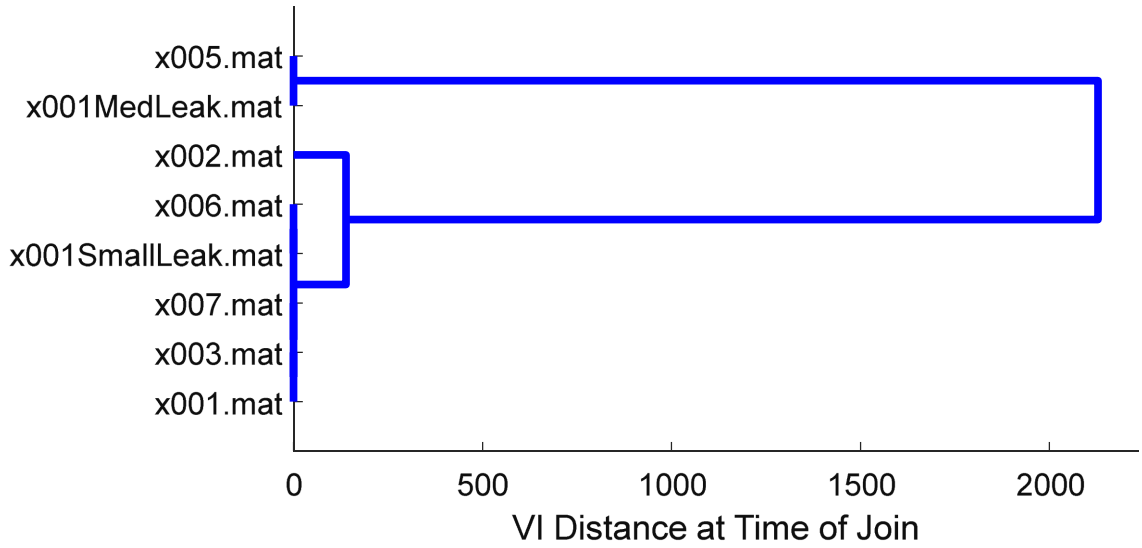

Fig. 11. Clustering of two customers with synthetic leak events and six actual customers.

probability of detection, small volume leaks may run for weeks or months before repair, contributing to the total volume lost.

To test the ability of the clustering algorithm to separate leaks from typical behaviors, synthetic leaks are created by choosing a duration the leak is present and a volume per hour of the recorded leak flow. At a random time, the leak begins, and the leak volume is added to every hourly data point for the duration, as illustrated in Fig. 10 for a very short duration leak (blue) compared to original customer data (gold). This assumes a fixed-volume leak, which is not entirely accurate. Future improvements to this algorithm should represent more realistic leaks—a small initial flow rate, increasing over time, sometimes progressing to a rupture with high flow rate [39].

The synthetic leak customers have been generated from Customer 001 by creating either a small volume of 0.75 gallons per hour for a duration of 500 consecutive hours or a medium volume of 2.3 gallons per hour for 200 consecutive hours. This does not imply leaks follow these volumes and durations but provided a case for supporting future work to investigate these results. Fig. 11 illustrates the output of the clustering algorithm for the leak customers as compared to six actual customers, including the donor data. The medium leak of 2.3 gallons has been grouped much further from the original customer than the small leak. Descriptions of the individual customer data used for Fig. 11 are provided in Table I.

B. *Evaluation Using Real Customer Data*

The section describes experiments and results on 99 real customers. Fig. 12 illustrates ten runs of the clustering algorithm

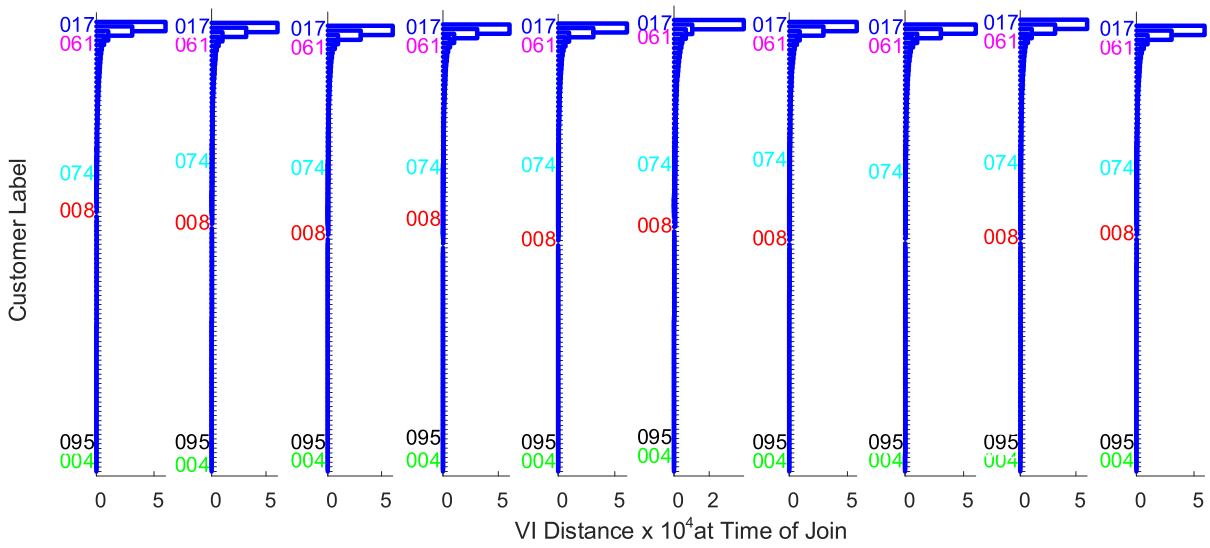

Fig. 12. Consistency experiment results using wVI distance—consistent customers.

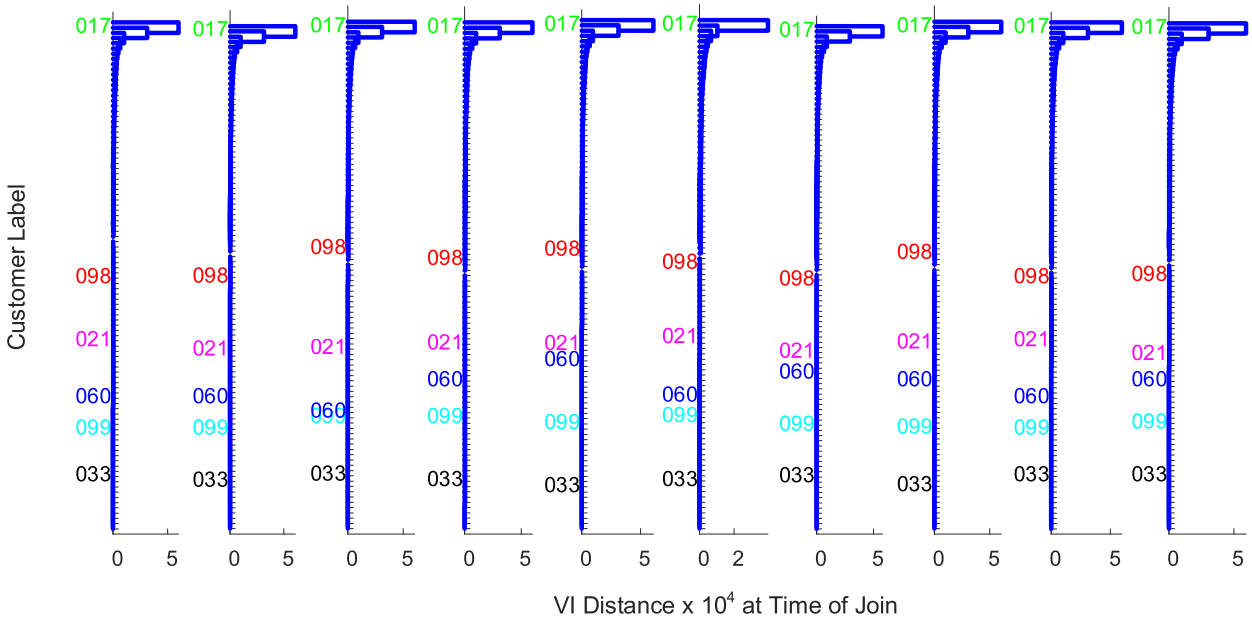

Fig. 13. Consistency experiment results using wVI distance—inconsistent customers.

for the most consistent customers. The six most consistent customers are highlighted in color. Here we see that the customers are very consistent in how they are joined across runs. For example, customer 095 (yellow) is joined consistently and quickly with customer 004 (green). This indicates that customer 095 and customer 004 have similar consumption patterns, and can be treated as a group for marketing, incentive, and conservation programs. Having a consistent algorithm able to segment the customers by usage allows utilities to focus their resources upon other aspects of the business while still allowing for communication and outreach tailored for the different use cases. Another example is customer 017 (blue) is always the last customer to be joined with a large wVI distance. Customer 017 is likely an outlier and worth investigating further, and treated individually for customer contact and targeted programs to accommodate their unique use case within this group.

The least consistent customers are highlighted in Fig. 13. As Fig. 13 shows clearly, using the wVI distance measure for clustering the GMMs has low volatility of all individual customers across many trials. This improves the consistency seen when running the clustering multiple times with the same data, and reduces the volatility caused by random differences in the GMMs. For a practical application, the repeatability of results is critical to performance. A utility must be confident the same data produces nearly the same clusters, regardless of the randomness within the models.

VI. Conclusion

In this section, we summarize our method, discuss the results, and propose future work. As data were collected from water meters, the measurements of flow in gallons were recorded at hourly intervals. The records were stored as time series entries in the Badger Meter, Inc. BEACON Advanced Metering Analytics system. Prior to any clustering, the data required preprocessing to eliminate anomalies and errors that would invalidate the clustering results. The data also were normalized per customer by the nonzero median value, leaving only the behavioral patterns and relative magnitude of flow.

Upon completion of the preprocessing step, dimensional reduction was performed using a GMM of the data within an RPS with time lags of 0, 24, and 168 h. The purpose of the RPS was to generate areas within the space related to daily and weekly habitual water consumption behaviors. The GMM reduced the space required to store a representation of a single meter and allowed the direct comparison of multiple meters with different quantities of historical data.

Following the dimensional reduction and preprocessing, the hierarchical clustering process could begin. The wVI distance measured the distance between two customers GMMs. Customers were combined using a hull-based approach. Distance measures were defined and compared, with supporting examples to illustrate advantages and shortcomings.

Two sets of experiments were performed—one using synthetic data and the other real customer data. The synthetic data experiments showed expected results, while the real customer data showed consistency across runs. To judge the veracity of the results, we had evaluated two sets of data. The first was seen in Fig. 7, 9, and 11. We had designed simulated customers which should cluster in an expected way. These figures showed that our method does cluster them as expected. The second was seen in the consistency of the clustering in Fig. 12 and 13.

This article could be expanded through enhancements to pre and postprocessing methods, exploring the RPS further, identification of evolutionary customer behavior, and practical improvements for commercial application. Additionally, the wVI distance measure could be generalized further for more flexibility in applications. The data cleaning implemented in this article extracted the longest consecutive set of measurements with no gaps, disaggregated records, or negative values. This brute-force approach simply excluded data that would cause the clustering method to crash. A more robust method would identify the anomalies and modify the clustering method to accommodate them. The customer model currently stored only a single time series as input to generate the model. A different approach could accommodate multiple, nonconsecutive time series to accommodate large gaps in the recorded data. This would involve making several initial GMMs for a customer, one for each time series segment. The collection of submodels would need to be combined into one master-model used for clustering among all customers, weighting the contribution to the master by the amount of data used to create that particular submodel.

In another implementation, the preprocessing of large gaps in the data may include disaggregating the sum of consumption over the missing time. The system would determine the expected value during the missing periods, based on previously collected data. For instance, the weekday expected value pattern composed of two Gaussian distributions could be used as the function to scale the known missing volume. This pattern is unique to every customer and day types (day of week, weekday, weekend, or other schedules). Once the scaled expected values are recreated, the disaggregated data can be entered into the former gap in the time series. This will not provide additional insight in the model (creating a model from itself is moot) but will allow the data system to handle a single continuous time series rather than several smaller time series. The advantage will be simplified implementation, data storage, and handling of the data by the program compared to the previous suggestion of storing many time series individually for one meter.

This article contributed a method for processing water meter time series data as well as a novel approach to weighing components within a model. The method of unsupervised hierarchical clustering using information-theoretic distance measures was flexible enough to accommodate different numbers of clusters as the individual application requires and needs no training set of labeled customers to determine which individuals have similar behavior to each other. While the data were taken only from one utility, the methods could be applied to other systems with hourly AMR data collection. Findings were tied to the usage patterns of individuals, rather than geographical location or size of utility. These advantages make the method appropriate for implementation in water utilities where resources of time, finances, equipment, or staff are limited. The wVI distance measure presented here improved the clustering consistency to engender confidence in the results, with customers assigned similarly throughout multiple experiment trials. The wVI focused on flow event behaviors with a tight variation in time and volume and relies less upon behaviors that vary widely from day to day.

References

[1] "U.S. Drought Portal." Accessed: Aug. 7, 2014. [Online]. Available: https://www.drought.gov/drought/

[2] *Water Conservation Ordinance*. City of Visalia, CA, USA: Municipal Code § 13.20.

[3] American Housing Survey for the United States: 2009 Current Housing Reports, U.S. Census Bureau, Washington, DC, USA, Series H150/07, 2008.

[4] J. W. Ferguson, "Replacing water meters to cut costs across Texas," *The New York Times*, Aug. 2012.

[5] State of the Water Industry Report, Amer. Water Works Assoc., Denver, CO, USA, 2018.

[6] I. P. Panapakidis, M. C. Alexiadis, and G. K. Papagiannis, "Three-stage clustering procedure for deriving the typical load curves of the electricity consumers," in *Proc. IEEE Grenoble Conf.*, 2013, pp. 1–6.

[7] I. P. Panapakidis, M. C. Alexiadis, and G. K. Papagiannis, "Electricity customer characterization based on different representative load curves," in *Proc. 9th Int. Conf. Eur. Energy Market*, 2012, pp. 1–8.

[8] I. Bose and X. Chen, "Detecting temporal changes in customer behavior," in *Proc. Int. Elect. Eng. Congr.*, 2014, pp. 3–6.

[9] I. Bose and X. Chen, "A fuzzy clustering based analysis of migratory customer behavior," in *Proc. Int. Conf. Comput. Inf. Sci.*, 2011, pp. 480–483.

[10] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2001, pp. 952–955.
[11] J. H. Jensen, D. P. W. Ellis, M. G. Christensen, and S. H. Jensen, "Evaluation of distance measures between Gaussian mixture models of MFCCs," *Int. Conf. Music Inf. Retrieval*, vol. 2, pp. 107–108, 2007.
[12] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 849–856.
[13] T. Jebara, Y. Song, and K. Thadani, "Spectral clustering and embedding with hidden Markov models," in *Proc. 18th Eur. Conf. Mach. Learn.*, 2007, vol. 4701, pp. 164–175.
[14] A. Albert, "Problems, models, and algorithms in data-driven energy demand management," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 2014.
[15] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, and F. M. Roberts, "Statistical models of reconstructed phase spaces for signal classification," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2178–2186, Jun. 2006.
[16] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, and J. Ye, "Time series classification using Gaussian mixture models of reconstructed phase spaces," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 6, pp. 779–783, Jun. 2004.
[17] M. Qiao and J. Li, "Two-way Gaussian mixture models for high dimensional classification," *Statist Anal. Data Mining ASA Data Sci. J.*, vol. 3, no. 4, pp. 259–271, 2010.
[18] C. Laspidou, E. Papageorgiou, K. Kokkinos, S. Sahu, A. Gupta, and L. Tassiulas, "Exploring patterns in water consumption by clustering," *Procedia Eng*., vol. 119, pp. 1439–1446, 2015.
[19] R. M. Willis, R. Stewart, D. P. Giurco, M. R. Talebpour, and A. Mousavinejad, "End use water consumption in households: Impact of socio-demographic factors and efficient devices," *J. Cleaner Prod.*, vol. 60, pp. 107–115, 2013.
[20] R. Cardell-Oliver and G. Peach, "Making sense of smart metering data: A data mining approach for discovering water use patterns," *Aust. Water Assoc. Water J.*, vol. 40, pp. 124–128, 2013.
[21] R. Cardell-Oliver, "Discovering water use activities for smart metering," in *Proc. IEEE 8th Int. Conf. Intell. Sensors, Sensor Netw. Inf. Proc.*, 2013, pp. 171–176.
[22] R. Cardell-Oliver, "Water use signature patterns for analyzing household consumption using medium resolution meter data," *Water Resour. Res.*, vol. 49, no. 12, pp. 1–11, 2013.
[23] S. A. McKenna, F. Fusco, and B. J. Eck, "Water demand pattern classification from smart meter data," *Procedia Eng.*, vol. 70, pp. 1121–1130, 2014.
[24] A. Di Nardo, M. Di Natale, G. F. Santonastaso, V. Tzatchkov, and V. H. Alcocer Yamanaka, "Divide and conquer partitioning techniques for smart water networks," *Procedia Eng*., vol. 89, pp. 1176–1183, 2014.
[25] A. Candelieri and F. Archetti, "Identifying typical urban water demand patterns for a reliable short-term forecasting —The icewater project approach," *Procedia Eng*., vol. 89, pp. 1004–1012, 2014.
[26] A. Candelieri, D. Soldi, and F. Archetti, "Layered machine learning for short-term water demand forecasting," *Environ. Eng. Manag. J.*, vol. 14, no. 9, pp. 2061–2072, 2015.
[27] D. Garcia, D. González Vidal, J. Quevedo, V. Puig, and J. Saludes, "Water demand estimation and outlier detection from smart meter data using classification and big data methods," in *Proc. 2nd New Develop. IT Water Conf.*, 2015, pp. 1–8.
[28] A. Candelieri, D. Soldi, D. Conti, and F. Archetti, "Analytical leakages localization in water distribution networks through spectral clustering and support vector machines. The icewater approach," in *Proc. 16th Conf. Water Distrib. Syst. Anal.*, 2014, vol. 89, pp. 1080–1088.
[29] T. Sauer, A. Yorke, M. Casdagli, J. A. Yorke, and M. Casdagli, "Embedology," *J. Statist. Phys.*, vol. 65, no. 3, pp. 579–616, 1991.
[30] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[31] F. Takens, "Detecting strange attractors in turbulence," *Dyn. Syst. Turbulence Warwick 1980*, vol. 898, pp. 366–381, 1981.
[32] B. Gold and N. Morgan, *Speech and Audio Signal Processing*. New York, NY, USA: Wiley, 2000.
[33] M. Meila, "Comparing clusterings—An information based distance," *J. Multivariate Anal.*, vol. 98, no. 5, pp. 873–895, 2007.
[34] L. Franek and X. Jiang, "Ensemble clustering by means of clustering embedding in vector spaces," *Pattern Recognit*., vol. 47, no. 2, pp. 833–842, 2014.
[35] A. K. Jain, "Data clustering: 50 Years beyond k-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.
[36] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Trans. Math. Softw.*, vol. 22, no. 4, pp. 469–483, 1996.
[37] G. Ver Steeg, A. Galstyan, F. Sha, and S. DeDeo, "Demystifying information-theoretic clustering," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, vol. 32, pp. 1–27.
[38] C. Li and G. Biswas, "Applying the hidden Markov model methodology for unsupervised learning of temporal data," *Int. J. Knowl.-Based Intell. Eng. Syst.*, vol. 6, no. 3, pp. 152–60, 2002.
[39] G. Kunkel, *M36 Water Audits and Loss Control Programs*. Denver, CO, USA: American Water Works Association, 2016.
[40] T. Britton, R. Stewart, and K. O'Halloran, "Smart metering: Enabler for rapid and effective post meter leakage identification and water loss management," *J. Cleaner Prod.*, vol. 54, pp. 166–176, 2013.
[41] S.-C. Hsia, Y.-J. Chang, and S.-W. Hsu, "Remote monitoring and smart sensing for water meter system and leakage detection," *IET Wireless Sens. Syst.*, vol. 2, no. 4, pp. 402–408, 2012.
[42] J. Almandoz, E. Cabrera, F. Arregui, E. Cabrera Jr., and R. Cobacho, "Leakage assessment through water distribution network simulation," *J. Water Resour. Planning Manag.*, vol. 131, no. 6, pp. 458–466, 2005.
[43] R. Gomes, J. Sousa, and A. Sá Marques, "Influence of future water demand patterns on the district metered areas design and benefits yielded by pressure management," in *Proc. 12th Int. Conf. Comput. Control Water Ind.*, 2013, pp. 744–752.
[44] T. Britton, R. Stewart, and K. O'Halloran, "Smart metering: Providing the foundation for post meter leakage management," *J. Cleaner Prod.*, vol. 54, no. 2013, pp. 166–176, 2009.