Analysis of Advanced Meter Infrastructure Data of Water Consumption in Apartment Buildings

Einat Kermany, Hanna Mazzawi, Dorit Baras, Yehuda Naveh
Analytics Department, IBM Research
Haifa University Campus, Israel
einatke@il.ibm.com,
hannam@il.ibm.com, doritb@il.ibm.com,
naveh@il.ibm.com

Hagai Michaelis
Arad Technologies
Yokneam Elite, Israel
Hagai.Michaelis@aradtec.com

ABSTRACT

We present our experience of using machine learning techniques over data originating from advanced meter infrastructure (AMI) systems for water consumption in a medium-size city. We focus on two new use cases that are of special importance to city authorities. One use case is the automatic identification of malfunctioning meters, with a focus on distinguishing them from legitimate non-consumption such as during periods when the household residents are on vacation. The other use case is the identification of leaks or theft in the unmetered common areas of apartment buildings. These two use cases are highly important to city authorities both because of the lost revenue they imply and because of the hassle to the residents in cases of delayed identification. Both cases are inherently complex to analyze and require advanced data mining techniques in order to achieve high levels of correct identification. Our results provide for faster and more accurate detection of malfunctioning meters as well as leaks in the common areas. This results in significant tangible value to the authorities in terms of increase in technician efficiency and a decrease in the amount of wasted, non-revenue, water.

Categories and Subject Descriptors

G.4 [**Mathematics of Computing**]: MATHEMATICAL SOFTWARE—*Algorithm design and analysis*

Keywords

Machine Learning, Advanced Meter Infrastructure, Water, Leaks, Malfunction

1. INTRODUCTION

In recent years, water has become an increasingly scarce resource in many parts of the world. This trend, driven by the increase of water use in developing countries, the pollution of water sources, and desertification, is of utmost importance from both environmental and societal perspectives. In addition, it has led to a steep increase of water prices and thus forms a new business problem in many parts of the world that previously did not regard water supply as a significant business challenge.

Both the environmental and business aspects of the problem are currently at the forefront of concerns faced by local or municipal governments, and by the local water authorities responsible for actual distribution and supply of water to residents and to commercial consumers. Given the current level of environmental awareness, any significant leak of water, or other mishandling of this perceived scarce resource, is likely to result in bad public relations and could become a political nightmare for the authorities. In addition, with the frequent hikes in water prices, customers are becoming much more aware of their water bills and are much less likely to tolerate over-charges due to leaks within and outside their properties. On the other hand, so called non-revenue water (NRW), i.e., water that is supplied but is not billed for, has a large effect on the bottom line profitability of the water distribution companies. Common sources of NRW are non-metered supplies of water, leaks from pipes, theft, fraudulent tempering of meters, and malfunctions or inaccuracies of meters. It is estimated[1] that NRW ranges from less then 5% of the total water production in places like Copenhagen and Singapore, 20% in the UK and France, 50% in Mexico, to more than 70% in sub-Saharan Africa. The World Bank estimates that the total annual NRW reaches 48 billion cubic meters, with an accompanied loss of revenue of 14 billion USD [14].

All this has created strong pressure on government and private authorities in charge of handling water to more closely analyze their entire operations and to find novel ways to reduce water waste and losses. One of the many ways to reduce NRW and increase customer care is by implementing advanced metering infrastructure (AMI) systems. In such systems, all meters in a municipality are read at short time intervals, and the current reading at each interval is transmitted electronically (typically by wireless means) to a data center. Then, in the most naive usage, a strong leak (e.g., a burst of a large pipe), can be identified as an abrupt and abnormal increase in the water consumption as measured by

[1]See multiple references within en.wikipedia.org/wiki/Non-revenue_water.

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

the meter. In this work, we report on much more delicate scenarios, which can be identified only by deep analysis of AMI data, but nevertheless have a similarly large impact on the business of water distributers, as well as on the environment.

We focus on two major aspects of municipal water operations: reduction of NRW, and better utilization of the authority's resources, both human (e.g., technicians) and machinery. Obviously, minimizing NRW reflects directly on the authorities' revenues and cost balances. Similarly, better resource utilization means less wasted time of technicians, ultimate cost reduction to the authorities, and better service to the public. While several works have dealt with NRW [2, 10, 17, 18] and resource utilization [8] in the public infrastructure outside the building domain, this is the first time that such a work is done on issues related to the individual building property. As we will see, those issues are of high stakes to the authorities, while posing unique challenges which do not appear in the public infrastructure case.

The rest of the paper is organized as follows: In Section 2 we provide a detailed description of the AMI data and the challenges related to it. Section 3 summarizes previous related work. Section 4 describes a preprocessing phase which we claim is generic to AMI data. In Section 5 we describe the machine learning algorithms we used and the design decisions leading up to them. In Section 6 we present our results, and in Section 7 we conclude and discuss this work.

2. PROBLEM AND DATA SPECIFICATION

In this work, we are interested in the analysis of AMI data originating at apartment buildings. An apartment building setting introduces unique complexities due to the large number of independent meters in each building, the correlations between them, the interaction and tensions among the residents in the building, and the shared resources (and shared bills) of the building. For example, misreadings or deliberate fraud in one of the residential meters may affect all other residents in the form of higher bills for their own properties or for the shared properties.

2.1 Leaks and Theft in the Common Area

A typical apartment building consists of any number of apartments and a common area that includes a relatively large garden, staircases, the building roof, and similar shared spaces. We refer to water consumption in the common area of a building as common consumption. The total consumption of the building is measured by a large meter at the entry to the building, known as the main meter. In addition, the residential consumption is measured by smaller meters at the entry to each apartment. The common consumption is not directly measured, because it consists of multitude of pipes built at different times and connected to the main pipelines at different, sometimes chaotic or even piratic, places in the building. Therefore, the common consumption needs to be computed by subtracting the amount of water measured by all individual meters from the total amount of water that passes through the main meter. However, this calculation is not necessarily as straightforward as it may appear, due to several reasons. First, as the data from each meter are inherently noisy and incomplete, the arithmetic calculation that involves all meters only amplifies the uncertainty. Second, the calculated common consumption is composed of several elements: legitimate consumption (e.g., cleaning the stairs, watering the common garden); malfunctions in individual apartments (e.g., stopped meters, sabotage, or fraud); inaccuracies of any of the meters; and real losses of water that may be due to leaks or theft of water in the common area. All of these reasons cause the calculated common consumption to be extremely noisy. For example, the calculated common consumption frequently results in negative values, which are of no use[2].

Detecting losses in the common area is of particular importance compared to losses in the residential areas. First, the pipes in the common area are larger, and losses can be much more significant. Second, transferring the cost of the loss to the residents may result in frustration and objections, as none of the residents feel responsible for losses outside their own apartment meters. Lastly, losses in the common area may indicate criminal behavior of any of the residents or of an outside party, which is in the interest of the community to identify quickly.

To illustrate the significance of the problem, Figure 1a shows a histogram of the consumption[3] in the common area of 590 buildings in January 2012. A crude estimate determined by regular usage suggests that the normal monthly consumption should be several cubic meters, but we can see that in most of the buildings, the consumption is much higher. Figure 1b shows an example of a leak; in this case the leakage continued almost 40 days before fixed and the lost water was more then 470 cubic meters.

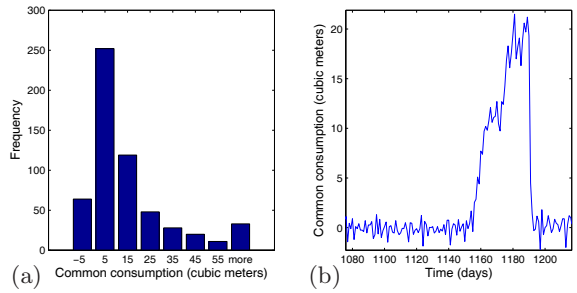

Figure 1: (a) Common consumption histogram and (b) a leakage example.

The current practice for detecting water loss in common consumption is to define a threshold consumption beyond which the authority suspects a problem. However, fixed-threshold setting is problematic, because inherent differences exist among buildings. Moreover, the common consumption of any building varies greatly with time. This means that either the threshold is large enough to detect only the most severe losses or that it is bound to create many false alarms due to legitimate uses of water in the common area. Therefore, our first goal was to identify leaks or theft of water in common areas of buildings in a more accurate and reliable way.

2.2 Identification of Faulty Meters

As with any mechanical element, water meters are prone to malfunction and may measure and report less water than

[2]Negative values can indicate occurrence of one of the problems described, but it can also be an artifact of the quantization of meter readings, see Section 2.5.

[3]Common consumption is measured by subtracting individual apartment meter readings from the main meter readings.

actually consumed. In addition, a common way to defraud the water authority is by tampering with the meter such that not all water is measured. Whether due to malfunction or fraud, such faulty meters are obvious sources of NRW. The authority may react by either taking the loss on itself or by billing the resident with a presumed consumption value, a practice that invariably results in a conflict between the resident and the authority. Hence, identifying and fixing faulty meters as soon as possible is of utmost priority to water authorities.

Many of the faulty meters, either malfunctioning or fraudulent, do not report any consumption. Therefore, a naive procedure could be to send a technician to examine and possibly replace any zero-consumption meter in the municipality, where a zero-consumption meter is defined as a meter that reports the same reading for a predefined length of time. However, some meters report the same readings for legitimate reasons, most often in cases in which all household residents are on vacation away from home. Empty properties and storehouses also result in zero-consumption over a long period of time. With this in mind, sending a technician to all zero-consumption meters after short periods would lead to much wasted technician time. Conversely, waiting for longer periods would reduce the number of false alarms, but would dramatically increase the time for which water is not billed in cases of illegitimate zero-consumption meters, resulting in larger losses to the authority and more unpleasant interactions with the residents.

Our goal regarding the identification of faulty meters is therefore to statistically distinguish between illegitimate and legitimate zero-consumption meters. Being able to do so shortens the period necessary before sending a technician to illegitimate zero-consumption meters. It also increases the technician pool efficiency by decreasing the number of visits to legitimate zero-consumption meters.

2.3 Consumption Prediction

The ability to predict consumption is useful for planning water distribution systems and new neighborhoods. In addition, it can be useful as another technique to detect leaks and faulty meters. When a predicted consumption value is significantly higher (or lower) than the actual metered value, the probability of either a faulty meter or a leak increases. Our goal in this regard is to predict the consumption for each meter on a daily basis, focusing on building main meters. The prediction can be performed in both on-line and off-line settings, depends on the exact need.

2.4 AMI Data

AMI data were the main source of data we analyzed for this work. Each meter in the AMI system can provide up to one data point every 15 minutes[4]. Each such data point consists of: (1) a unique identification number, (2) a current meter value, (3) a current meter status, and (4) a timestamp. The meter status is a set of flags identifying the state of the meter, reporting indications about tampering attempts, suspected leakages, etc. The transmitted data are received by local devices and almost immediately transferred to central data servers.

[4]We had access only to daily consumption during the course of this work.

Our analysis is based on consumption data from a medium-size city of around 100,000 residents, most living in apartment buildings of 4-40 apartments. The data consisted of:

1. Daily readings from each meter
2. Meter technical data for each meter
3. Consumer data for each meter (e.g., number of residents in the household)
4. Apartment buildings water network (identification of main meters and their sub meters)
5. Lists of meter replacements and meter cleanings as reported by technician logs

All data related to this paper were collected by an Arad Technologies system[5].

2.5 Data Challenges

The water management tools we developed are based on the supplied AMI data. We faced several generic challenges when attempting to mine the data. Such challenges prevail in any significant amount of data produced by AMI systems, and where acute enough to require specific handling. These include the following:

- Relatively large amounts of missing data due to communication problems, empty transmitter batteries, etc.
- Time discrepancy: Different meters reporting at different times of the day.
- Metering quantization: Each meter reports only a floor value of the actual reading. The floor resolution depends on the type and the capacity of the meter. Typically the resolutions are 0.1, 1, and 10 cubic meters for small (apartment), medium (main building), and large (infrastructure) meters, respectively[6].
- Meter thresholds: Each water meter has a threshold flow below which it does not measure any consumption. This is different from the previous item in that flow below the threshold value is neither recorded nor transmitted at any subsequent time. Threshold flow is typically 10-15 liters/hour for apartment meters, and higher for main meters.
- Meter inaccuracies: Meters are presumed to have a non-systematic measurement margin of error. The estimated error is 5% for flows below 200 liters/hour and 2% above that.
- Faulty reports: Many cases of faulty reports occurred, originating from different causes. These included meters that advanced backwards causing apparent negative consumption, faulty meters with unpredictable reports, and large differences between two subsequent readings due to replacement of the meter.

3. RELATED WORK

Several papers address the detection of anomalies in water distribution networks [2, 10, 17, 18]. Unlike our work, which focuses on buildings and uses AMI data only, these papers focus on large-scale networks that are equipped with sensors that measure water pressure, flow, and more. The papers use a variety of machine learning techniques, such as K-PCA

[5]More information regarding Arad's measurement technology can be found at http://www.aradtec.com.

[6]Future systems will have smaller quantization levels, which will reduce the quantization error, but it will not completely vanish.

and neural network, to detect anomalies and mainly focus on leak detection, water quality monitoring, etc.

Other recent papers address the problem on an apartment level. In [8], A. Hampapur et al. described a water management system they developed. Their system uses AMI readings and machine learning techniques to analyze water consumption in apartments. They focus on detecting leaks, but only at the apartment level, and their work does not offer a solution for malfunction of meters or for leaks at the common grounds. A few other papers [6, 13, 7] suggest installing additional sensors in the apartments, such as microphone-based sensors, water pressure sensors, and others to address the same problems.

In an additional effort to conserve water on an apartment level, Chen et al. [3] provided an algorithm for disaggregating apartments' water consumption signals into components. Disaggregation of the water consumption signal helps residents gain insights, identify, and change bad consumption behavior.

Many other research efforts were performed in the area of water consumption prediction using different approaches and algorithms. Many of these works deal with prediction of large populations, such as cities [1, 4, 12], campuses [11] and water plants [19]. Additionally, in some of the works the prediction is required in resolutions of months (e.g., [1, 12]) or weeks (e.g., [11]). Even in works in which the resolution of prediction is daily [12, 19], the input features include not only the past consumption but also additional features related to weather conditions (temperature, humidity). In [4], Cutore et al. attempt to predict daily consumption based on past consumption and additional features (day of the week and whether or not it was a working day). In [16], Liu et al. predict a domestic model, but the input features include the consumption as well as features related to water pricing, household income, and property size.

None of the works related to prediction specifically mention AMI data. Only two works referred to a preprocessing stage: in [11] Jain et al. mention missing values but use only part of the samples with complete values and in [12] Jowhar et al. use linear transformation to change the range of values. No other work has dealt with missing values, alignment in sampling times, or noise. We are not aware of other works similar to ours in the sense of prediction of domestic (or building-level) consumption on a daily basis, based on past data only.

4. DATA PREPROCESSING

As a first step in dealing with the data challenges we described in Section 2.5, we performed an extensive phase of preprocessing of the data. This phase contained three stages: cleaning the data, fixing existing values, and estimating missing values. The stages were performed sequentially (usually in the above order) before running the main data mining algorithms. Note that the raw data for the preprocessing phase is pairs of timestamps and readings from each meter. In this section, we describe each stage and in the subsequent sections, we show a comparison of those methods and present their effects on the performance of our algorithms.

4.1 Cleaning the Data

In the cleaning stage, we removed all unreasonable readings from the historical data. Unreasonable readings are defined as very high consumption and negative consumption, both of which occur due to various known malfunctions.

The raw data provided were particularly noisy in nature. An initial analysis revealed that the noise occurs at all frequencies, and that the information is also spread over the entire spectrum, in the sense that different consumers have different frequency components. Therefore, filtering or other signal-processing related preprocessing procedures were not performed.

4.2 Alignment of the data

Most of the historical data were at a daily rate. Thus, we relied on daily consumption when analyzing for leaks and faulty meters. On most days, any given meter was sampled at 7am. However, on some days the meters were sampled at a different time. In this section, we present two main methods for making our samples aligned.

- **Linear Interpolation.** We used linear interpolation to create a continuous signal from the samples. It was done by linear connection between every two nearest points. Then we sampled this signal every day at 7am. This method changes only samples that were not taken at 7am.
- **Average Consumption.** We used the average consumption to fix each sample. A sample $s_j = (t_j, x_j)$ is a pair of positive numbers in which t corresponds to the time and x is the actual reading from the meter. For fixing the ith sample $s_i = (t_i, x_i)$, we looked at samples $s_{i-k} = (t_{i-k}, x_{i-k})$ and $s_{i+k} = (t_{i+k}, x_{i+k})$, where k is a fixed positive constant. Assume that the average water consumption between times t_{i-k} and t_{i+k} is c liters per hour; that is, $c = (x_{i+k} - x_{i-k})/(t_{i+k} - t_{i-k})$. Denote the difference in hours between t and 7am of that day by Δ (note that Δ can be negative). The algorithm changed the ith sample s_i to be $\tilde{s_i} = (t - \Delta, x - \Delta \cdot c)$.

4.3 Estimation of Missing Values

In this section, we present the various methods we used to predict the values of missing samples.

- **Linear Interpolation.** Similar to Section 4.2 except that we sampled the continuous signal everyday at 7am.
- **Polynomial Regression.** We tried to estimate meter readings using a polynomial $P_d(t) = \theta_0 + \theta_1 t + \cdots + \theta_d t^d$ of limited degree d. Given a set of examples

$$(t^{(1)}, x^{(1)}), (t^{(2)}, x^{(2)}), \ldots, (t^{(m)}, x^{(m)}) \in (\mathbb{R}^n, \mathbb{R}),$$

  the polynomial regression algorithm chooses a polynomial P_d such that the L_2 norm of the error on the training set is minimized. That is, $\sum_i \left(P_d(t^{(i)}) - x^{(i)}\right)^2 + \lambda \sum_{j=2}^{d} \theta_j$, is minimized. Here λ is the regularization constant and θ_j is the jth coefficient of the polynomial. Now, to predict the reading in time t_i, we ran a polynomial regression algorithm. We provided the algorithm with the training set $S = \{(t_\ell, x_\ell), \ldots, (t_{\ell+k}, x_{\ell+k})\}$, where S corresponds to the k samples closest in time to t_i[7]. After training the algorithm with S, we used its hypothesis $P_d(t)$ to predict the reading in time t_i to be $(t_i, P_d(t_i))$.
- **Polynomial Interpolation.** In this method, we used polynomial regression with no regularization ($\lambda = 0$) and full degree (that is, if the size of the training sample set is k then, we look for a polynomial of degree $k-1$). In other

[7]We make sure that $k/2$ of the samples were taken before t_i, and $k/2$ were taken after t_i.

words, let $S = \{(t_{\ell+1}, x_{\ell+1}), (t_{\ell+2}, x_{\ell+2}), \ldots, (t_{\ell+k}, x_{\ell+k})\}$, be the set of k samples closest in time to t_i. We found a $k-1$ degree polynomial P such that for all $j \in \{\ell+1, \ldots, \ell+k\}$ we had $P(t_j) = x_j$. Our prediction for the reading in time t_i was $P(t_i)$. That is, we added the predicted sample $(t_i, P(t_i))$ to the data.

- **Estimating using Median.** To estimate the consumption in a given day — in this method we estimated the consumption and not the actual read from the meter — we looked at the water consumption on nearby days (past and future). Our estimated consumption was the median of these values. Note that this is the only method that produces quantized values.
- **Collaborative Estimation — Principal Component Analysis (PCA).** In this method we used PCA components. The idea is to use water consumption vectors with only a few missing samples[8]. We performed PCA on these vectors and found the principal components from which we could recover the original vectors with minimal error. That is, we found vectors $u_1, u_2, \ldots, u_m$, for which every water consumption vector v with few missing values there exists coefficients $a_1, \ldots, a_m$ where $||v - (a_1u_1 + a_2u_2 + \cdots + a_mu_m)||$ is minimized. Here m is the smallest integer for which we retain 95% of the variance. After we had identified these vectors, when constructing a vector with many missing entries w, we used regression to find coefficients $\theta_1, \ldots, \theta_m$ that minimize the cost function $||w - (\theta_1u_1 + \theta_2u_2 + \cdots + \theta_mu_m)||_2$. Obviously, we minimized the distance between w and $(\theta_1u_1 + \theta_2u_2 + \cdots + \theta_mu_m)$ in entries for which w i known. The vector w is replaced with the vector $(\theta_1u_1 + \theta_2u_2 + \cdots + \theta_mu_m)$, which does not contain missing entries.

5. ALGORITHMS AND DESIGN DECISIONS

In this section we describe the main structure of the data mining algorithms that we used in order to analyze the three use cases described in Sections 2.1, 2.2, and 2.3, respectively,and the design decisions leading to them.

5.1 Leaks and Theft in the Common Area

We used an unsupervised anomaly detection method to detect irregular behavior in the common area of the buildings. With this method, the data were cleaned and corrected according to the preprocessing methods described in Section 4 before running the main algorithm. The common consumption was calculated as described in Section 2.1. Then the relevant features were extracted. These features were created separately for each building and for each day[9]. The features we selected were the current day of the week, the current month, the daily common consumption, and the weekly common consumption. The weekly common consumption was calculated by moving windows of seven days each on the daily consumption. The weekly consumption is important because the common consumption is very noisy and inaccurate, while the weekly consumption significantly smooths this noise. The following distance measures were used for the various features: for consumption — the arithmetic difference; for the day in the week — delta function (namely 1 if the elements are identical and 0 otherwise); for the month — the gap between two months.

We applied the k-nearest neighbor (KNN) [5] algorithm on the common consumption of each building on a given day to detect anomalous behavior. Note that we can only detect malfunctions that started relatively recent to the historical data. The idea is to check if the common consumption on the given day is abnormal compared to the historical behavior of this building. We did this by comparing the extracted feature's vector of the given day with the vectors of its closest neighbors. In particular, we chose the k vectors that were closest to the given day and then calculated the average distance between those instances and the vector of the current day. A low average distance indicates that the vector of the current day is similar to its neighbors, which means that the consumption on the given day is normal. A high average distance indicates that the vector of the current day is different from its closest neighbors, meaning that the consumption behavioral is changed and therefore should cause concern of a leak or malfunction. The chosen k in our solution was 30. We calculated KNN twice, once before applying the estimation of missing values procedure and again after estimation, and we took the maximum value of the two. The KNN algorithm requires a certain amount of data in order to work correctly. Therefore, if not enough historical data exist, we used a simple statistical model to represent the daily consumption and the weekly consumption. We then used Grubb's test for outliers [9] for each feature (daily and weekly consumption) and the anomaly score of the current day as the average value.

Since no labeled data existed in the case we analyzed, we had to use an unsupervised learning technique. The KNN algorithm is suitable to our problem because it is entirely data-driven and no additional feedback or labeling is required [9]. In addition, KNN suits our setting in which no assumptions related to the distribution of the data exist-the data are unknown and vary among different buildings.

Using our approach, once we had the results for each building, we created a ranked list in descending order of building identifications and scores, indicating the level of anomalous consumption in the common area. Note that we reported only cases of positive common consumption; in cases that the common consumption is negative the algorithm applies a score of zero. In addition, the function returned a status for each building including information such as whether enough historical data exist or whether the common consumption is high only on the given day, etc. A water authority can then use the scored list and the additional indicators to handle the cases of abnormal common consumption in buildings. Typically, an authority looks at the most highly-scored anomalous cases, and either alerts the building administration of the potential loss of water or directly investigates the case. Figure 2 summarizes the solution.

5.2 Identification of Faulty Meters

Our solution used a classifier to distinguish between cases of legitimate, such as residents on vacation, and illegitimate, as malfunctioning or tampered, zero-consumption meters. Before using the classifier, the algorithm cleaned and corrected the data as described in Section 4. Following this, the algorithm created samples, and calculated the features. The samples were created as follows: The historical data

[8]In a consumption vector v, the ith coordinate v_i equals $x_{i+1} - x_i$, where x_i are the meter readings.

[9]For small leaks, looking at smaller intervals can not distinguish legitimate and illegitimate usage.

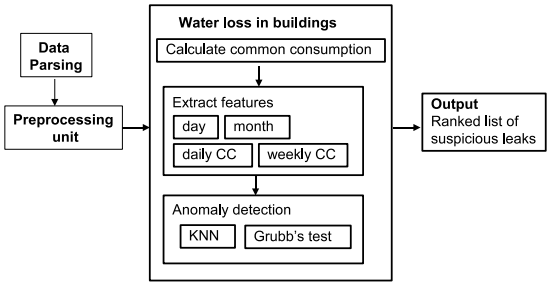

Figure 2: Algorithm's scheme for detection of losses in common areas of buildings.

were parsed to find the events in which meters report zero-consumption during a predefined period of time. For each event, we calculated the start of the event — the first day that the meter started to report zero and the end of the event — the last day that the meter reported zero, and we then set labels. If the meter was replaced or cleaned by a technician, we labeled it faulty, otherwise it was labeled legitimate.

One main problem in the sample creation stage is distinguishing between two similar cases that can occur for a specific meter. Once case is a non zero-consumption meter that progresses slowly and therefore reports rarely. The second is the case of two or more different zero-consumption events for the same meter. Due to meter quantization and inaccuracy, such cases may appear quite similar from a data point of view, yet they have different origins and may thus confuse the classifier. Our solution handles such cases by merging two events when the consumption between the events was less then a predefined number, typically one half of a cubic meter, and removed the second event if the gap between the events, between the end of the first event and the beginning of the second event, was less then 30 days. In our analysis, accurately recognizing such cases was particularly important, especially the start time, because there were several features that represent the consumption behavior during the last few days before the meter started to report the same reading and several features that compared between behavior before the event start and during the event.

After the algorithm defined the samples, it calculated the following features:

- Features that describe the behavior during the days prior to the beginning of the event of zero-consumption (e.g., average, variance, and slope of the consumption)
- Consumption features (e.g., average and entropy of the daily consumption)
- Features that describes the history of the meter (e.g., number of zero-consumption events in the past)
- Meter's and property's features (e.g., age, usage type)
- Features of the event (month, day of the week, and difference in the common consumption of a building before and during the event)

We checked a few classifiers — random forest [15], Naïve Bayes [5], and support vector machine [5] — and found empirically that random forest gives the best results. In Section 6.3 we present the full details of this.

Specifically, in our sample, and generally, in reality, many more events of legitimate zero-consumption occur than events of faulty meters. This ratio between the two types of events depends on the duration in which the meter reports zero-consumption. In our historical data, for example, around 10 percent of the events were faulty when we set this period as one month. For this reason, the faulty meter events got a weight that is 10 times more then the legitimate zero-consumption events. If the classifier did not accept a weight parameter as input, we simulated this by replicating the faulty events 10 times. In order to improve prediction results, we learned separate models for various intervals of zero-consumption. In particular, we learned a model for meters that were reported zero-consumption four-five weeks, a model for meters that were reported zero-consumption five-six weeks etc. The minimal duration for learning a model (four-five weeks in our case) can be determined by the user (the water authority). The decision to take durations of one week obtained the best performance.

Once we had the results for each zero-consumption meter, we created a ranked list in descending order of meter identifications and scores, indicating the level of confidence that a given zero-consumption meter is faulty.

Figure 3 summarizes the algorithmic stages in this part of the work.

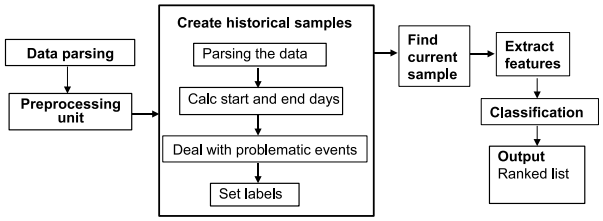

Figure 3: Algorithm's scheme for identification of faulty meters.

5.3 Consumption Prediction

We trained models to predict the daily consumption for the main meters in buildings. A model was learned for each building based on its past data. The prediction of a certain day was based on the consumption in the preceding seven days. Since water consumption patterns can change over time, the predictor was retrained every predefined period of time using the additionally gathered data. We used linear regression [5] with several regularization methods, polynomial regression, and artificial neural networks [5]. A simple average over past data was used as baseline. The results were compared when the missing data were estimated with various data estimation methods.

Previous research on daily consumption show that using additional information (e.g., temperature, humidity, and rainfall) can significantly improve results [12, 16, 19]. We plan to extend our model to include such additional features.

6. EXPERIMENTAL RESULTS

In this section we show (1) the results of the comparison of different methods in the preprocessing stage; (2) the results of our algorithm for water losses in the common area; (3) the results of our zero-consumption meter algorithm, including its performance as calculated by 5-fold cross validation, and comparison of a few classifiers; and(4) the results of our predictor and a comparison with a simple solution. In addition, we checked the effect of the different methods of estimation

of missing values (part of the preprocessing stage) on the performance of all the algorithms.

When presenting and discussing the results, we used the following classical measures: Precision, Recall, F1 score, and AUC [area under the receiver operating characteristic (ROC) curves]. In the zero-consumption scenario, we calculated the mean AUC, which is the mean area under the five ROC curves created by the five-fold cross-validation method. We also used the percent of samples to be checked for obtaining recall of 0.8, namely the proportion of the samples that should be checked in order to detect 80% of the positive values. (Recall that the algorithm's result is a ranked list in decreasing order.) This measure was the most effective in our communication with the water authority.

6.1 Preprocessing

In order to estimate the error of each method we removed intervals of known readings and estimated those missing readings by using the various estimation methods. The length of the interval was chosen using a distribution that represents the lengths of actual missing readings intervals in our data.

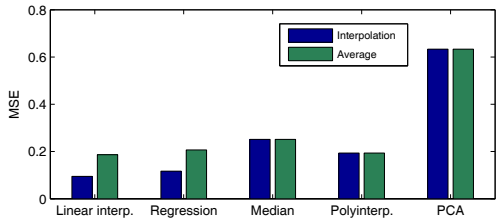

Figure 4: The mean square error for the preprocessing methods. Each couple represents a estimation method and the two colors represent the two fixing discrepancy methods.

Figure 4 displays the mean square error (MSE) of every preprocessing method. That is, we measured the mean square distance between real values and estimated values. The figure clearly shows that the linear interpolation method has the minimal MSE.

In order to further understand our results, we examined the data and monitored the algorithms' runs. When applying polynomial regression, we noticed that the coefficient of the linear component is the most dominant. Other coefficients, that correspond to non-linear features, were small in magnitude. This suggests that many of our meters transmitted a function that is close to linear with high frequency noise.

As for collaborative estimation (PCA method), the noisiness of the water consumption vectors makes this method problematic. We used dimensionality reduction to search for a linear space that approximates consumption vectors with minimal loss of variance. Many meters experienced malfunctions and unreasonable behavior during the observed period of time. This caused many unexpected consumption values that changed the principal component to include non-typical behavior causing the error in estimation to increase.

Finally, we recall that different meters have different resolutions. Given a meter with resolution r (in which r can be 100 Liters or 1000 Liters), and given that the amount of water consumed is equal to c, the meter transmits $\lfloor \frac{c}{r} \rfloor \cdot r$. In an attempt to lower the MSE, we tried to apply a quantization function similar to the one the meter uses on the estimated values. The idea is simple: performing quantization on the estimated consumption using the meter resolution should decrease the MSE, since our real values are quantized. In practice however, the error only increased. Given a real value from the meter x, and predicted values $\tilde{x}$, the error is equal to the difference squared $(x - \tilde{x})^2$. When performing quantization on $\tilde{x}$, the error equals zero for all predictions in $[x, x + r)$. However, for values in $[x + r, x + 2r)$ and $[x - r, x)$, the error equals r^2. That is, we have a very large error when the predicted consumption is lower than the actual consumption. Since this event happens frequently, the MSE increased after performing quantization. Similar analysis showed that the error does not decrease when we round our estimated values using the meter's resolution.

6.2 Leaks and Theft in the Common Area

We checked our new solution with data from 590 buildings. We ran the algorithm on a specific day and compared the new solution with the currently used solution, which is based on a defined threshold. The first 25 buildings with the highest score from our solution were checked by technicians. The check method included talking to the house committee or other residents to find about known leaks, checking for above ground leakage, checking for unmeasured water consumption (other than irrigation), and finally, if no other explanation was found, checking for underground hidden leaks with leak detection tools. Of the 25 cases, 13 were actually checked on-site and the other 12 were not checked due to technical problems (3) or inability to coordinate the check with the house committee (9). The technicians found that 12 cases had leaks or other malfunctions such as faults in the irrigation system or improper connections of pipes. Only one case was a false alarm. Of those 13 cases, the existing solution detected only nine cases and had the same false alarm. The unchecked cases caused due to coordination problems were examined by a domain expert. The domain expert found that eight cases had leaks or other malfunctions and one case was a false alarm. The existing solution detected six cases and had the same false alarm.

In addition, the expert classified 42 cases that the existing system classified as leaks and the new solution gave them low priority. Of those cases 15 were proper, nine cases were prolonged malfunction which our solution does not claim to reveal because it does not yet have historical data that goes that far into the past, 17 cases could not be classified and one case was classified as a new malfunction.

To summarize those results, the new solution detects faulty cases that the existing solution did not reveal. Even more importantly, it gave low priority to many cases that were incorrectly classified as malfunctions by the solution of the meter company.

For more extensive testing we succeeded to label 326 buildings from this sample. The labeling was done by the results of the technicians and by the help of the domain expert. In those cases, 35 were leaks or other malfunctions and the rest were proper. In the existing solution 62 buildings were specified as having malfunctions but only 26 of them were actually malfunctions. The precision was 42% and the recall was 74%. We checked the new algorithm using the following estimation methods: linear interpolation, regression, median and polynomial interpolation. Table 1 summarizes the comparison result. The precision, recall, and F1 mea-

sures calculated for threshold 10%, meaning that 10% of the highest samples were classified as leaks.

Table 1: Leaks results.

	AUC	Percent of samples to be checked for recall = 0.8	Precision	Recall	F1
Lin. interp.	90	15	73	69	71
Poly. reg.	90	13	70	66	68
Median	76	34	30	29	29
Poly. interp.	88	23	64	60	62
PCA	84	24	42	40	41

The best results were achieved by linear interpolation and regression. The polynomial interpolation method was slightly worse, the PCA method was much worse for this algorithm, and the median was the worst method. The best $F1$ measure was 71, while the $F1$ of the existing solution was 53.6.

6.3 Identification of Faulty Meters

We checked the algorithm using different methods of preprocessing. We collected all the events in which the meter was reporting zero for two weeks or more from 2010 and 2011 and used 5-fold cross validation to check the performance. In our comparison of methods, we got similar results to those in the previous section, namely that linear interpolation and regression achieved the best results. In addition, we compared a solution using linear interpolation as a estimation method with a solution that does not estimate missing data. Figure 5(a) shows the comparison between the ROC curves of the two solutions. We can see that the estimation stage significantly improves the results (we used a two-sample Kolmogorov-Smirnov test on 30 runs for non-estimation and linear interpolation cases and got that they are from different continuous distributions at a very high significance level with p-values in the range of $10^{-14} - 10^{-12}$). We also tried different classifiers and found that the random forest gave the best result. See Table 2 for the comparable result.

Table 2: Zero-consumption meter results, a comparison between random forest, naïve Bayes, and linear SVM.

	mean AUC	Percent of samples to be checked for recall = 0.8	Precision	Recall	F1
R. forest	83	33	29	50	37
NB	72	71	22	38	28
SVM	73	51	19	32	24

In the beginning of 2012, we tested our solution by sending technicians to all the meters that reported zero-consumption for at least four weeks. We used historical data from 2010 and 2011 to train the model and then tested its performance on 283 new events. The classifier results were verified by correct labels from the technicians who checked each meter directly. The consumption data was estimated by linear interpolation. Recall that the existing solution was to send technicians to all zero-consumption events. The goal of the meter company was to reduce the number of meters that were inspected, while capturing most, e.g., 80%, of the faulty meters. Our solution predicted that to cover 80% of the meters that were actually faulty, inspect only 91 out of the 283 zero-consumption meters, which are 33% of the suspicious meters, would be sufficient. Figure 5(b) shows the predicted proportion of faulty meters among all zero-consumption meters as a function of the proportion of the sample that is inspected by a technician.

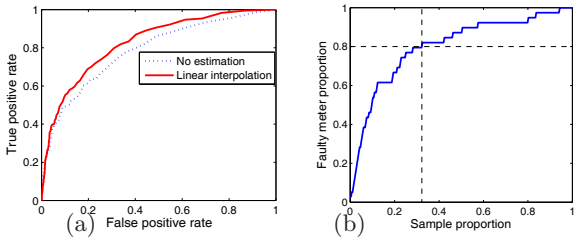

Figure 5: Detection of zero-consumption meters.

6.4 Consumption Prediction

We compared the consumption predictor with a simple algorithm that predicts the consumption of the next following by averaging the preceding seven days. The predictors learned the consumption in 2011 and predicted the consumption in the first six months of 2012. After prediction of each month, the models were retrained using a training set that contained the previous samples and the data of the additional month. Learning the predictors was performed for each data estimation method. We got similar results for the various estimation methods as those mentioned in the previous sections. Therefore, the results we present here are achieved by the linear interpolation method, which consistently showed the best results. The performance measure is the MSE averaged over all buildings. The polynomial regression method provided the best results (mean MSE of 1.87), which is an average improvement of 9% as compared to the baseline (mean MSE of 2.09). The linear regression method (regardless of the regularization method) was better than the baseline but worse than the polynomial. The neural networks gave poor results compared to the baseline, on all the architectures that we tried. We believe that this is due to the noisy nature of the data, and we suspect there was over-fitting to the noise. In order to demostrate the benefits of our solution, we present Figure 6(a) to show the MSE of the polynomial regression predictor vs. the MSE of the baseline. This shows that in most of the buildings, the polynomial regression performed better. Figure 6(b) shows an example from a single building (slightly changed, due to privacy restrictions). Notably, the polynomial regression follows the real consumption more accurately than the simple baseline algorithm.

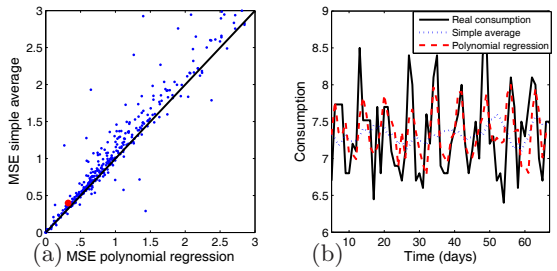

Figure 6: Consumption prediction example. The MSE results of the building that is shown in figure b is marked by a red dot in figure a.

7. DISCUSSION

We have presented a new data mining application that holds great promise for tackling issues of prime importance to the operation of municipal water supply. The importance stems from the effect of the results on the regular business aspects of water authorities, such as increased revenues, resource efficiency, and customer care, but also from the effect on environmental aspects related to water conservation and sustainability. In fact, with water scarcity becoming a significant issue in more and more parts of the world, this second aspect is gaining growing interest in virtually all municipalities in both the developing and industrialized world.

While the results presented in this paper show the significant value of the work already done, much work remains for the future, in terms of both new functionality, and stronger technology. Together with water authorities, we are working on identifying and formalizing new business cases addressing further major pain points of the customers. In addition, we are working on strengthening and improving the algorithmic methods used to approach AMI data related to apartment buildings. One promising direction that we have only started to experiment with is analyzing correlations between neighboring buildings, and among buildings and neighborhoods of similar characteristics. Another direction we intend to investigate is extending our algorithms to use hourly resolution. This was not done under the scope of this work due to lack of historical data at this resolution.

We are certain that this work and others like it will serve to accelerate the transition of municipalities to modern metering infrastructures. With this, demand for analytics and data mining will only increase and will drive for growing complexity and sophistication of the methods. We are looking forward to this future.

Acknowledgments

We thank Merav Aharoni, Tal El-Hay, Amir Ronen, Michal Rozen-Zvi, Lavi Shpigelman, Chen Yanover, and Tal Zur for comments and useful discussion.

8. REFERENCES

[1] A. Altunkaynak, M. Ozger, and M. Cakmakci. Water consumption prediction of Istanbul city by using fuzzy logic approach. *Water Resources Management*, 19(5):641–654, 2005.

[2] L. Camarinha-Matos and F. Martinelli. Application of machine learning in water distribution networks: An initial study. *ICML '97*, pages 49–57, 1997.

[3] F. Chen, J. Dai, B. Wang, S. Sahu, M. Naphade, and C.-T. Lu. Activity analysis based on low sample rate smart meters. *KDD '11*, pages 240–248, 2011.

[4] P. Cutore, A. Campisano, Z. Kapelan, C. Modica, and D. Savic. Probabilistic prediction of urban water consumption using the SCEM-UA algorithm. *Urban Water Journal*, 5(2):125–132, 2008.

[5] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2 edition, 2000.

[6] J. Fogarty, C. Au, and S. E. Hudson. Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition. *UIST '06*, pages 91–100, New York, NY, USA, 2006. ACM.

[7] J. E. Froehlich, E. Larson, T. Campbell, C. Haggerty, J. Fogarty, and S. N. Patel. Hydrosense: infrastructure-mediated single-point sensing of whole-home water activity. *Ubicomp '09*, pages 235–244, New York, NY, USA, 2009. ACM.

[8] A. Hampapur, H. Cao, A. Davenport, W. S. Dong, D. Fenhagen, R. S. Feris, G. Goldszmidt, Z. B. Jiang, J. Kalagnanam, T. Kumar, H. Li, X. Liu, S. Mahatma, S. Pankanti, D. Pelleg, W. Sun, M. Taylor, C. H. Tian, S. Wasserkrug, L. Xie, M. Lodhi, C. Kiely, K. Butturff, and L. Desjardins. Analytics-driven asset management. *IBM Journal of Research and Development*, 55(1.2):13:1 –13:19, Jan.-March 2011.

[9] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126, 2004.

[10] J. Izquierdo, P. A. Lopez, F. J. Martinez, and R. Perez. Fault detection in water supply systems using hybrid (theory and data-driven) modelling. *Mathematical and Computer Modelling*, 46(3-4):341–350, 2007.

[11] A. Jain, U. C. Joshi, and A. K. Varshney. Short-term water demand forecasting using artificial neural networks: Iit Kanpur experience. *ICPR'00*, pages 2459–2462, 2000.

[12] H. Jowhar R. Mohammed. Hybrid wavelet artificial neural network model for municipal water demand forecasting. *ARPN Journal of Engineering and Applied Sciences*, 7(8):1047–1065, 2012.

[13] Y. Kim, T. Schmid, Z. M. Charbiwala, J. Friedman, and M. B. Srivastava. NAWMS: nonintrusive autonomous water monitoring system. *SenSys '08*, pages 309–322, New York, NY, USA, 2008. ACM.

[14] B. Kingdom, R. Liemberger, and P. Marin. The challenge of reducing non-revenue (NRW) water in developing countries. How the private sector can help: A look at performance-based service contracting. In *Water Supply and Sanitation Sector Board Discussion Paper Series*, number 8, 2006.

[15] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[16] J. Liu, H. H. Savenije, and J. Xu. Forecast of water demand in Weinan city in China using WDF-ANN model. *Physics and Chemistry of the Earth, Parts A/B/C*, 28.

[17] A. Nowicki, G. Michai, and D. Kazimierz. Data-driven models for fault detection using kernel PCA: A water distribution system case study. *International Journal of Applied Mathematics and Computer Science*, 22:939–949, 2012.

[18] M. Osborne, R. Garnett, K. Swersky, and N. de Freitas. A machine learning approach to pattern detection and prediction for environmental monitoring and water sustainability. In *Proc. Workshop on Machine Learning for Global Challenges*, 2011.

[19] Y. Tachibana and M. Ohnari. Prediction model of hourly water consumption in water purification plant through categorical approach. In *Systems, Man, and Cybernetics, 1999*, volume 2, pages 569 –574 vol.2, 1999.