Smart, Big, and Dirty

Experiment Findings *in* Molecular Interventions · January 2017

DOI: 10.13140/RG.2.2.30330.31689

2 authors:

Milad Khaki
Western University
18 PUBLICATIONS **58** CITATIONS

Nasim Mortazavi
Western University
15 PUBLICATIONS **5** CITATIONS

Smart, Big, and Dirty
A Case Study on Peak Water Consumption from Smart Meter Data

MILAD KHAKI, University of Waterloo
NASIM MORTAZAVI, University of Western Ontario

ACM Reference format:
Milad Khaki and Nasim Mortazavi. 2017. Smart, Big, and Dirty
A Case Study on Peak Water Consumption from Smart Meter Data. *ACM J. Data Inform. Quality* 9, 4, Article 39 (March 2017), 26 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

ABSTRACT

Many municipalities have recently installed wireless (‘smart’) water meters that allow functionalities such as demand response, leak alerts, identification of characteristic demand patterns, and detailed consumption analysis. These benefits require the meter data to be error-free, which is not true in practice, due to *dirty data.* We focus on the impact of dirty data on identifying customers contributing to a load peak, using a case study obtained from the City of Abbotsford, British Columbia, Canada. We identify the sources of errors in this dataset, find systematic ways to remove these errors and compare the results obtained from both unprocessed and cleaned data. The contributions of our work are a systematic study of the errors existing in a large-scale smart water meter deployments, a careful study of the impact of dirty data on peak load attribution, and classifying techniques to deal with errors from dirty water consumption measurements.

LIST OF ABBREVIATIONS AND ACRONYMS

SEEGTS	Smart Electrical Energy Generation and Transmission Systems
SMI	Smart Metering Infrastructure
WSS	Water Supply System
AMI	Advanced Metering Infrastructure
AMID	AMI long-term/Archive data table
MIND	Meter information data table
BILD	Billing information data table

1 INTRODUCTION

As a cost-saving measure, many municipalities have recently installed wireless ('smart') water meters that allow remote meter reading, such as Toronto and Saskatoon, in Canada, and Baltimore and Pittsburgh, in the United States. Essential characteristics of these meters are the ability to read at time intervals, as low as a minute, rather than a month(s), and providing a bi-directional communication channel between the provider and consumer. Although frequent meter reading is unnecessary for billing customers on a monthly basis, analyses of these high-frequency water consumption data permit functionalities such as

- demand response, in which customers responsible for a short-term demand peak may be asked to reduce consumption [15],
- identifying characteristic demand patterns to allow more accurate forecasting [6], and
- providing detailed consumption analysis, including suggestions on how to reduce the water bill, to the customer [15].

The effects of the mentioned functionalities mostly to add more effective water preservation strategies and also a more efficient prediction of future demands. As the water shortage is a global problem and is getting considerable attention in the literature and industries [], the smart meters can be used as means to generate raw data that can be utilized in these cases.

We know that these benefits can only be realized if the meter data stream is free from significant errors. It is well known that a considerable fraction of data obtained from virtually *all* large-scale meter deployments can be incorrect (for some examples refer to [33], [36], [22], [3], [17], and [9]). This issue is often called the problem of *dirty* data. Dirty data reduces the potential benefits of a smart meter deployment, for example, the additional cost of addressing customer complaints of over-billing. We also focus on another negative consequence of dirty data: incorrect decision-making. In particular, we study the impact of dirty data on identifying customers contributing to a peak load, using a case study dataset obtained from the City of Abbotsford, British Columbia, Canada. We identify the sources of errors in this dataset, find systematic ways to remove the errors and compare the results obtained from both unprocessed and cleaned data. Our primary conclusion is that data cleaning must precede any use of smart meter data, especially for extremal statistics (such as peaks). The contributions of our work are

- a systematic study of the errors existing in a large-scale smart water meter deployments,
- a careful study of the impact of dirty data on peak load attribution, and
- introducing and classification of techniques available removing errors from dirty data, including those produced for this study

The remainder of the paper is structured as follows. Section 2 provides a basic understanding of smart water meter networks. A generalized infrastructure of such network is demonstrated, and possible errors that can occur are introduced. In addition, the cause of the errors are discussed and are individually correlated to infrastructure building blocks. In Section 3, we present related literature in the field of smart water systems and compare them with similar studies in the field of smart meters in electrical energy. In Section 3.1, background information about the case study, City of Abbotsford, are provided and the definition of peak water consumption, which is one of City's issues, are presented, as well. The second part of this section discusses the structure of the employed dataset and its schema. Furthermore, the data quality issues that were particularly encountered in the current study are introduced together with the adopted or produced solutions. Next, the context-dependent errors, which were detected in our study, are introduced in Section 3.5, and procedures for handling them are provided. As the final part of the case study, the results of using the cleaned dataset for peak contribution analysis are presented in Section 3.6, and the sensitivity of these results to the different types of error are examined, as well. Finally, Section 4 provides conclusions and future work.

2 BACKGROUND

2.1 Smart Metering Infrastructure

During the past decade, a worldwide rising trend in adopting smart metering infrastructures (SMI) has emerged. This movement is mainly a result of proved and potential advantages of such systems, in comparison with their traditional counterparts [14]. Because of the bi-directional communication channels in SMIs, these systems are now more aware of their current state such as automatically detecting possible faults, i.e. water leaks, in the network. Possessing such capabilities prevents SMIs to have similar vulnerabilities of traditional networks, such as unnecessary losses due to technical problems or thefts that remain undetected for extended periods of time. From the data analytics perspective, a more significant feature of a SMI is generating a valuable source of information-rich data that can further improve the quality of service and overall performance of the system.

Smart electrical and water supply networks share the same concept of smart measurement infrastructure; however, as the inherent properties of the two are not similar, their configurations are slightly different. Figure 1 shows a general configuration of a smart meter infrastructure in the case of water supply networks. The proposed figure is based on the current case study. In addition, to keep it generalized, it is influenced by the diagrams suggested by the following articles, as well: [37], [29], [33], [18], [27], [38], and [14]). Currently, the main international companies are producing the equipment for smart infrastructures. Each manufacturer has its particular hardware specifications. Therefore, for better understanding, manufacturer-specific details are removed from the diagram and provided as examples. The details presented in the current part of the section are based on the specifications of smart water meter devices published by a company called: "Itron," a major manufacturer of smart metering infrastructure solutions [20]. The block diagram in Figure 1 is composed of the following parts:

Block (A), wireless smart meters that are distributed around the city and measure water consumption in a standard unified unit, e.g. cubic meters. As stated previously, there is a bi-directional communication channel between the meter and the utility. Through this medium, the meter can transfer its status and consumption values, and the utility can perform actions such as re-configure the meter, read the instantaneous value of the meter register, and reset it. Each meter contains a counter register that records the consumption in the cumulative format as an unsigned integer value at every specified interval. Therefore, they are also known as cumulative interval meters. As an example, if the state of a register at 11:00 am is 14390 with the interval of one hour, and consumption of 30 Litres between 11:00 am and 12:00 pm, the state of the register at 12:00 pm would be 14420. These meters have the capability of temporarily storing the measurement logs, e.g. 45 days. The process of billing consumers starts from this point by downloading the logs of the consumptions.

Block (B), wireless data collectors are hardware-specific data collection servers that are responsible for collecting the readings from meters at every interval (i.e. in CSV format and are programmable down to intervals of minutes or seconds) and transferring them wirelessly/wired to the data warehouse (for an example of such unit refer to [19]). As these units require wireless communication with various endpoints, which are meter transmitters and are low powered, they usually have the flexibility of mounting on towers, buildings, or utility poles. CCU transfers data packages, containing the measurements of the meters under their communication group gathered during the last 24 hours (depending on the configuration), and sends them to the data warehouse, Block D.

Block (C), is the control center of the utility infrastructure. Commands to reconfigure the meters (such as reset the meter, changing interval length, and one-time read requests) or collectors (such as adding/removing a meter, flushing the buffer, and status check) are relayed through this block.

In addition, this block has a local data server that stores customer specific information and meter information together with archived billing records, which will be described as MIND and BILD tables in this paper.

Block (D) is the Temporary Measurement Data Storage; it receives the raw measurement data, i.e. hourly cumulative readings for the entire city, from the collectors and provides outputs for block E and F. The outputs of this block are the same received raw data or slightly revised version of it that is suitable for storage (In the current case study outputs are hourly instantaneous readings and daily cumulative ones). The performance of this block, which directly affects the status of the system, is closely monitored by Block C.

Block (E) is the long-term storage or archive of the network and stores the data for future analyses. Depending on the network configuration the archived data can be raw meter readings or transformed data with a pre-storage treatment scripts. Any further access or modification to the archived data is provided through Block C.

Block (F) is the billing system and can join the raw meter readings which are received from the meters, with the meter specific unit information. A copy of meter unit table is kept locally on this block apart from its source, at infrastructure administration (Block C). Therefore, Block F is responsible for generating bills, the payments process, and being accountable to consumer concerns. Each billing record reflects the consumption during that period for a specific consumer. From this perspective, block F acts as a substitute for manual meter reading process and would not cause any other alteration to the billing records.

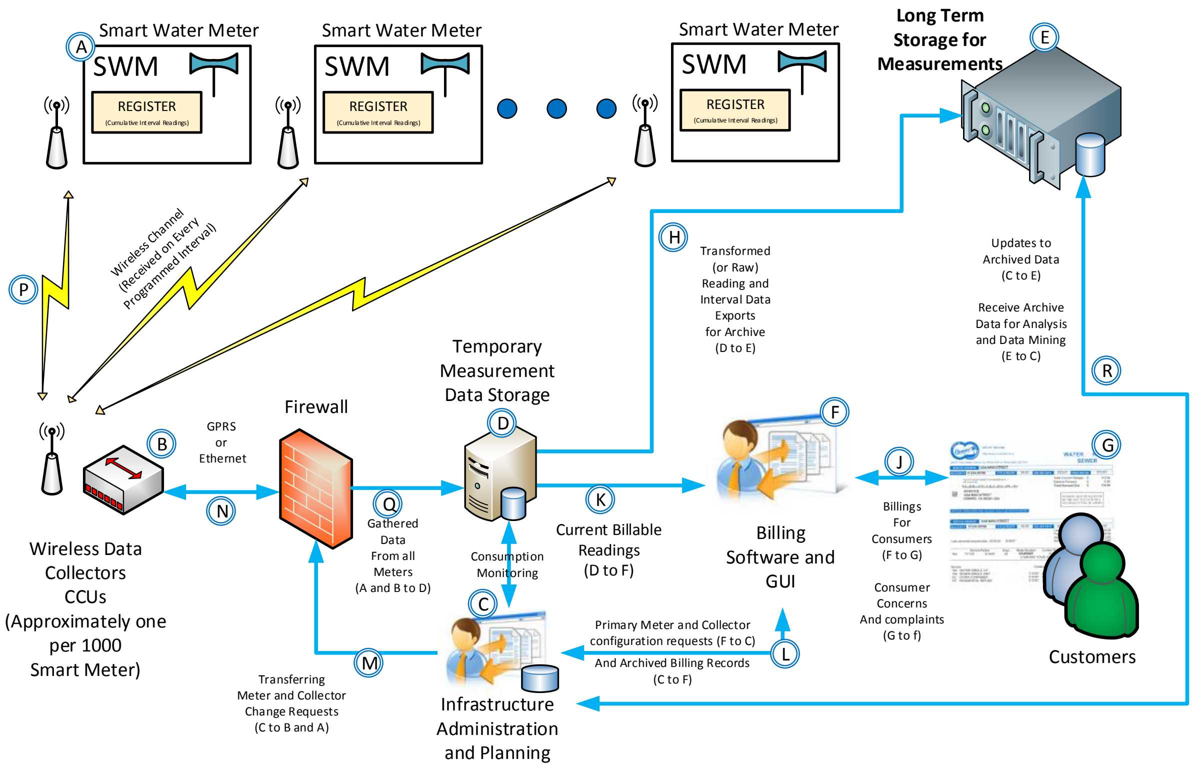

Fig. 1. Block diagram of the wireless water metering infrastructure

2.2 Difficulties of Data Analysis in SMI

As we are currently passing through a worldwide installation phase of the SMI, the focus of most researchers and interested industrial partners are the immediate advantages, such as time-of-user pricing [17], efficient automatic billing instead of the manual process [25], and early fault detection in the network [18]. In contrast, few studies have focused on the data analysis aspects and how the infrastructure can be modified to accommodate these requirements.

At this part of the survey, our intention is to raise the concern that despite the benefits of a SMI, it is essential to verify the validity of its outputs. In Section 3.1, the case study, errors encountered while analyzing such data are described in details, their characteristics are analyzed, and solutions for removing or reducing their effects are provided. In the remainder of this part of the section, similar data quality challenges will be introduced by a list of the errors, and the possible starting points on Figure 1 are discussed. In addition, Table 1 provides a summary of all issues analyzed here and shows the correlation between each error and different sources demonstrated in Figure 1.

2.2.1 Duplicate Records. Because of the communication channel problems, Paths (P) or (N), the server might ask the collector or the meter to retransmit the data. A possible error is that the retransmitted readings are not all missing. Therefore, some records are registered as duplicates and should be removed.

2.2.2 Missing Records. Similarly, because of the communication channel problems, some recordings would be irreversibly lost. The main sources of this issue are the communication channel problems between Blocks A and B, or an interrupt in the storage services of Blocks D, E, or F.

2.2.3 Measurement Granularity Errors. In some cases, a meter can have coarse grain resolution and cause this error, which is restricted to Block A. For instance, a meter that can measure consumptions as low as a fraction of a Litre, could be programmed to have the minimum step size of one cubic meter. As a result, the accuracy of the meter would be virtually reduced. It is the responsibility of administration, Block C, to ensure that the temporary data stored at Block D do not have such problems and handle the exceptions by updating meter configuration according to its specifications.

2.2.4 Spikes. Are defined as abrupt and short-duration changes in the consumption pattern of a consumer's profile that are not a valid representation of the actual consumption. The current sources of spikes that we have observed and are mentioned in the literature are mechanical faults of the meter or storing multiple inconsistent readings for the same timestamp (i.e. instead of having one consumption value of 11.3, two 38965 readings and one 11.3 reading would be recorded with similar time stamps. If the consumption value of 38965 is selected because of more probability of being true, it will cause a spike that is clearly erroneous, which results in a discrepancy and faulty values misrepresented as spikes, in Block F.)

2.2.5 Meter Unit Inconsistencies. This error can be originated by meter unit changes that are not back-propagated in the archive records. In such cases, Block C's decisions are affecting Block A's configuration. However, this error type would not necessarily affect Block E's billing records as, at the time of calculating corresponding billing values, there is no discrepancy between meter readings and its respective unit. A simple workaround to avoid the generation of this error could be storing the meter units along with the consumption values.

2.2.6 Meter Counter Resets. The smart meters usually accommodate a counter that registers the consumption at every interval cumulatively. In general, the meter only communicates these cumulative readings to the server. Therefore, if the server reconfigures the meter, it can also cause a reset on its register with a faulty command, as well. In Figure 1, this inconsistency is caused by Block C and affects Block A.

2.2.7 Meter Under/Non-Registration Errors. A popular belief is that a smart meter has high precision and would not be prone to measurement errors. In fact, smart meters are the next generation of traditional ones and merely benefit from the ability to store and transmit measurements at very short intervals ([11] and [25]).

Therefore, the accuracy problems existing in the traditional meters also occur in them, as well. Mukheibir et al., Fantozzi, and Arregui et al. have examined the changes of accuracy a water meter through its lifetime ([31], [13], and [2]). According to Mukheibir et al., the meter's failure to register low flow consumptions accurately is called under-registration. However, if the meter is completely incapable of detecting the flow below a certain threshold, it is known as non-registration [31]. Mukheibir et al. also show that because of these characteristics, the meters are calibrated to slightly over-register constantly for the first few years of their effective life-time. The increased offset would cause the meter to readjust reading inaccuracies roughly and after a period of time or consumption amount, more than 3500 cubic meters for some instances, it would start to under-register again. In the presented Figure 1, this error only involves Block A.

Table 1. Root
causes of errors cross correlated by the blocks and paths in Figure 1
ror Acronyms: Duplicate Records (DUP), Missing Records (MIS), Measurement Granularity Errors (GRN), Spikes (SPK), Meter Unit Inconsistencies (MUI), Meter Counter Resets (MCR), Meter Under/Non-Registration Errors (UNR)

Root Cause of the Error	**DUP**	**MIS**	**GRN**	**SPK**	**MUI**	**MCR**	**UNR**
Communication Channels	×	×		×			
Meter Hardware				×			×
Database					×		
Operator Mistakes			×			×	
Post-Error Correction Issues				×			
Path P	×	×		×			
Path N	×						
Block A			×	×			×
Block B							
Block C			×		×	×	
Block D		×					
Block E		×					
Block F		×		×			

Our analysis of the current literature in SMIs for Water systems shows that most of the studies do not evaluate the quality of data against the mentioned errors. However, data quality errors have impeded gaining the expected results in the majority of these studies. In addition, there are few papers in the field of electrical engineering based smart meter infrastructures that have focused on these errors either. To the best of our knowledge, only Quilumba et al. and Shishido, a technical report, have acknowledged the existence of some the mentioned errors in their study and provided some solutions for handling them ([33] and [36]). The fact that Quilumba et al. has published this article recently indicates that such concerns in the smart meter academic community are being raised and are expected to increase, as well.

3 RELATED WORK

In this section, benefits of smart water meter data are presented with the related work; next, the cause of each data quality issue is examined, and last, those articles that are dealing with data quality issues, which are in the field of smart water meter data, are provided. Lastly, similar studies in SEEGTS are provided and analyzed, as well. As previously stated, two established advantages of measurement using smart water meters are demand prediction and customer billing. Some instances of these works can be seen as the pioneers in using smart meter data, such as An et al. that applied machine learning and AI techniques for predicting demand and managing consumption [1]. They used 300 data streams, obtained from water distribution pumps, to formulate decision rules to manage consumption [1]. Similarly, [6] used a disaggregated consumption dataset of 252 households to train an artificial neural network-based demand predictor.

Another previously discussed fact is that water meters are prone to data quality errors, such as over- and under-registration, directly proportional to length and amount of usage [31]. Despite the various reports of errors in smart water meter measurements among the growing body of studies, such as [31], we are not aware of any "systematic" efforts to model and correct them. It should be acknowledged that a few articles merely mention these effects, which summarized as follows. The factors that influence the data quality of a water meter readings are discussed by [31] and [2]. These include noisy communication channels that would lead to corruption of the incoming data messages. An et al. also mentions that minor inconsistencies in the meter data input result in substantial uncertainty in the results [1]. We now outline the approaches to deal with data quality issues in the literature.

As one of the contributions of the current paper, we provide a summary of the state-of-art methods for evaluating and improving the data quality of water meter data in the literature. In general, five approaches to deal with data quality issues are presented, which are outlined in the remainder of this section.

The first approach to deal with errors is simply to ignore the detrimental effect of errors because of their proportion to clean data. For example, [3] and [4] use meter data to model the energy usage and carbon emissions resulting from water consumption. Their study provides considerable detail about the procedures for installing the smart meters and gathering data. However, as the data quality is not discussed, it can be inferred that the collected data was assumed to be error-free.

The second approach is to remove data streams that are suspected to have errors or missing data. For example, [16] explored the determinants of water consumption and developed a methodology to monitor end-uses of water. The study was performed using twelve household data streams, of which two had some missing data points, because of various meter failure issues. To confirm that the remaining data streams were correct representatives of data, in addition to the presented aggregated results, each household measurement time series that did not contain missing data was analyzed individually, as well. Similarly, Fielding et al. recognized the adverse effect of excessive missing data on the results, and so removed 17% of the streams, which did not contain at least ten weeks of daily water data use. Makki et al. encountered the problem of missing data while using smart water data to find the determinants of shower water consumption and removed the affected household measurements, [29]. In all aforementioned cases, despite the reported problems with the missing data, no analysis on the nature of the errors or how to deal with them directly, other than adding more accurate hardware that would improve future data, is provided [15]. The advantage of using the above approach is its simplicity. However, it can merely benefit the time series that only an insignificant percentage of their data is affected by errors. In such cases, the omission of erroneous data most probably would not cause loss of valuable information. Otherwise, it is not recommended.

The third approach is to approximate the missing or corrupted data based on the readings that are in temporal proximity of that specific point. This method is adopted by Machell et al. in their study by replacing the missing or invalid data, falling outside of the specified limits for each parameter. The replacement candidate is calculated using one of the following choices: a default predefined value, average over the previously valid data points, or replacing the value from a similar location of another data stream [28]. The stated approach is used as a pre-processing phase of data analysis application, which is simulating an on-line hydraulic model of a water distribution network in the UK. The authors also talk about another data quality issue: "data skew", which is an issue occurring when the reading process is done by polling the meters and not simultaneously. In such cases, the timestamps of all readings do not align properly, and data suffers from inconsistency in the temporal alignment of samples in different meters. Likewise, Umapathi et al. also used linear projection approximation method to estimate the annual consumption of households that had missing data to evaluate the effectiveness of plumbed rainwater tanks in reducing consumption [39].

The fourth approach is to clean the data at the time of collection. For example, [17] use 'Concentrators' to improve data quality. Concentrators are on-site devices, at the meter's side, with 90 days' worth of buffer to shield the measurement data from packet loss during communication to the server. Moreover, the concentrators use an encoding protocol with redundancy to provide resiliency against communication channel's noise. The extra transmission packet size is removed when received by the server and data is reconstructed as much as possible. However, the algorithm that is used for improving data quality is not described in detail. It is merely mentioned that the algorithm would remove minor errors from the slightly corrupted messages affected by noise. Nevertheless, the extent of quality improvement, in the data after using Concentrators, is not examined.

Finally, it is possible to select meters that are known to be less susceptible to measurement errors. For instance, [8] used the smart meters of roughly 3000 residential and light commercial customers to analyze peak consumption. They chose the meters that had a pipe diameter higher than a certain threshold, which according to the authors are less prone to measurement errors [8].

Tables 2 and 3 summarize the mentioned data quality issues and the solutions, which are suggested by the studies mentioned before, respectively.

The data quality issues are not limited to the field of water supply systems. In fact, the electricity supply systems have progressed more in the SMI field and have gained more in-depth analysis. One of the main motives comes from the issue that the resource in question, electrical energy, cannot be easily stored. As a result, in comparison to WSS, electricity supply industry has always been more forthcoming in investment in research and implementation of smart meters.

Fortunately, the previously learned lessons in SEEGTS can be applied to similar conclusions in WSSs, as well. A majority of the efforts in analysis and improvement of data quality in SEEGTS are done by the industries involved in this field. As an example, Albert et al., Shishido, and Quilumba et al. mention concerns about errors occurring in the measurement data that affect data quality that are quite similar to the current study (such as missing data, reading errors, lack of demographic survey data, zero readings, spikes and duplicate readings) and provide preliminary analysis for them ([36] and [33]). In both studies by Shishido and Quilumba et al., these errors can propagate to results and would deteriorate them. Confirming this issue, a set of preliminary results are presented in Shishido's report, which is only illustrated to introduce the possibility of data quality problem as a concept. Moreover, Quilumba et al. present more details of the errors' nature and discuss an application of consumer profile classification by k-means clustering with the semi-cleaned data as training and test inputs.

To sum up, the main goals of a smart infrastructure are to both analyze various states of the system and make it more optimized and provide a bi-directional communication channel with the consumers for further advancement of the first goal. However, compromised data quality would directly affect analysis results. Therefore, the main concerning point for every WSS is to find out in what ways do data quality issues could affect the behavior of the system and how to avoid them. In this regard, Jia et al. have studied the results of bad data, of both analog and digital measurement sources, on SEEGTS and demonstrate how it would affect decision-making results. The results are formulated based on the hypothesis that the error in data comes in a nature of noise or misreading of the actual measurement values. In addition, a metric is defined to quantify the effect of bad data on real-time price, which is called Average Relative Price Perturbation. The authors have concluded that errors in topographical data are more detrimental for the pricing schemes, comparing to the measurement data [21]. Similar examples of topographical data in WSS can be the state of water reservoirs and enabled/disabled status of pressure pumps. Given the presented facts, our goal is to explore the possibility of using software and algorithmic-based approaches for evaluating data quality status of a sample of smart water measurement data. Furthermore, we intend to employ the measurement data features to detect and remove errors, which is a considerably inexpensive solution to keep valuable data. In addition, we will introduce metrics that can be used to evaluate the effect of dirty data on different analysis results in the context of smart meters.

Table 2. Papers mentioning data quality issues

Data Quality Category	**Details**	**Mentioned by**
Missing Data	Communication channel noise	[2], [31], and [17], SEEGTS: [33]
	Due to the meter failure	[16], and [39] SEEGTS: [12] and [33]
	Daylight saving transition	SEEGTS: [33]
	The source is not mentioned	[15], [29], and [28]
Zero Readings	meters's inaccuracy at low consumption	[8]
	The source is not mentioned	SEEGTS: [33]
Duplicate Records		SEEGTS: [36] and [33]

Table 3. Solutions to DQ problems and the studies they were employed by

Solution to the Data Quality problem	**Employed by**
Assume that the errors are negligible	[3], and [4]
Discarding suspicious records	[16], [15], [39], [29] SEEGTS: [12] and [33]
Replace errors by the expected values	[28] SEEGTS: [33]
Improving future measurements' by improving communication protocol & hardware	[17]
Improving future measurements using more accurate measurement devices	[8]

3.1 Problem Definition

In this part of the section, first some detailed information about the City of Abbotsford is presented and subsequently the peak concept, and peak contribution is introduced. The City of Abbotsford is located in the Lower Mainland region of British Columbia, Canada. Although British Columbia has abundant water resources, the water supplies that are close to Abbotsford and can be economically treated are limited, making it imperative to manage the available resources carefully.

With the main intention of eliminating or minimizing the need for manual meter reading, recently the City installed wireless digital meters. The consumption profiles that are recorded by these meters, theoretically would allow us to answer the following questions [5]

- Which days (or weeks) of the year have peak demand?
- Which customers contribute most to these peaks?
- Which consumer sectors contribute most to these peaks?

Answers to these questions and similar ones are important because they would allow the City to plan its water requirement needs by implementing intelligent and adaptive demand response schemes. These schemes can be targeted towards specific customers who can best reduce peak demand, and avoid expensive infrastructure upgrades. In fact, one of the most valuable sources of knowledge, regarding the behavior of a water supply network, is its “peak consumption.” It enables us to ensure that the network is capable of handling the expected volume of water at any period of peak usage. Additionally, for water planning purposes, the system should be prepared for long-term peak consumption of the entire network. Therefore, this criterion should be considered during the infrastructure design and maintenance.

To find the actual peak contributors, we need to find the highest consumption over a specific period. After finding the temporal location of the peak period, we need to run a top-k query analysis to identify the main contributors. Once the peak consumers are found, we manually inspect their raw consumption profiles to verify the validity of peaking behavior.

However, the existence of the errors in the collected data (‘dirty data’) would cause incorrect calculation and consequently, incorrect decision-makings, which will be discussed in the remainder of the paper.

The significance of the peak contribution analysis is that it enables us to sift through the entire dataset and find the profiles that are the best candidates for being affected by errors, which are the focus of our study. By exhaustively examining those consumption profiles, it is possible to characterize the error types and to devise an error identification mechanism. In other words, the peak contribution filter enables us to find a starting point for more complex data quality analyses.

During this study, some data quality problems were detected, and solutions were devised. These data quality concerns can be grouped into three main categories: (1) errors that were recognized while importing and reorganizing raw data into a structured format, (2) gaps in the dataset that were captured after successful imports (missing data), and (3) context-dependent errors, detected while analyzing the data of water supply systems. The second half of the following section will discuss items one and two. As the third item is more complex and entails a considerable part of this work, it is discussed in more detail in Section 3.5.

3.2 Dataset Schema

The current part of the section describes the dataset schema that is used for storing the smart meter measurements. The datasets that we received from the described system in Section 3.1 consist of two parts: the exports of long-term archived records and a dataset containing consumer, meter, and billing data together. From the stated smart meter measurement data and meta-data, we managed to generate the following relational tables for our study: (1) anonymized meter and consumer information data (MIND), (2) billing data (BILD), and (3) Advanced Metering Infrastructure data (AMID).

MIND mostly contains time-invariant meta-data for each installed meter, including, but not limited to, measurement unit, installation date, physical characteristic, latitude, longitude, and pipe size. Also, it includes customer specific information, such as name (anonymized for analysis), address, postal code, and the category of each user, which is quite valuable. The categories for most water supply systems are Single-Family Residence (SFR), Multi-Family Residence (MFR), Industrial (IND), Commercial (COM), Institutional (INS), and Agricultural (AGR). In the current case, each category is further divided down to more specific sub-categories, which has resulted in 6 main and 127 subdivisions. In Figure 1, MIND can be located as a part of block F, which acts as a lookup table for meter information and consumer demographics for other tables. Each MIND record contains the mentioned fields together with the primary composite key that will be discussed in the next part of the section.

The long-term data storage acts as a data source for AMID, for each particular meter and its programmed interval, it contains a measurement record. All readings for a specific meter are identified by the primary composite key that will be described in the next part of the section. The reading interval of each meter is one hour; therefore, each time-stamped record in AMID is an indicator of the meter’s registered consumption for the past 60 minutes (instantaneous consumption of the consumer). AMID includes measurements for more than 25,000 individual smart meters that cover the entire city for the period of September 2012 to August 2013. It also includes a lower-resolution daily consumption data, which is cumulative, and similar to instantaneous hourly data, is generated based on raw meter readings. Each AMID record is created based on the archive exports that contain the following fields: meter ID, account ID, recording device ID (meter's digital serial number), customer ID (anonymized for analysis), channel type (interval daily or instantaneous hourly), consumption value, and end of interval timestamp). To reduce redundancy, the meter and consumer information are stripped from the record, and only the following fields are kept in AMID's relational table: primary composite key, consumption value, end of interval timestamp, and daily or hourly data flag.

The raw cumulative hourly readings are temporarily stored in Block D of Figure 1. After they are used to calculate the monthly billing records (instantaneous hourly, and cumulative daily), they are discarded. As we do not have access to the raw "cumulative" hourly data, the next viable alternative would be the "instantaneous" hourly records in AMID, which also have similar high-resolution information about consumption measurements.

About the possible advantages of the daily data, we performed a preliminary analysis that indicated they are affected by meter reset errors that were not transferred to instantaneous hourly records. Therefore, we decided to limit our focus to AMID's hourly instantaneous data in the remainder of the paper.

The billing records in BILD have an essential role in our study. As further would be discussed, we use them as means to evaluate if the data warehouse has performed the raw data transformation to archive records flawlessly. To our surprise, various types of data quality issues were introduced between Block D, the temporary raw data warehouse, and Block F, the archive. After confirming with Abbotsford that billing records were directly generated from raw data, we assumed that they are immune to any modification that can affect and reduce the data quality during transformation and storage in the long term fine-grained archive (AMID). Each BILD record contains the following fields: primary composite key, billing measurement unit, billing start date, billing read integer count, billing end date, billing end integer count, and consumption in cubic meters.

3.3 Data Quality Issue, Primary Composite Key

As a part of the importing smart meter data, each meter is required to be identified uniquely across all tables; therefore, as we did not have access to the original primary key, a join operation was required. Ideally, the join should be performed on a single primary key or a composite one, which is constructed by combining more fields. The MIND and BILD equivalent dataset existed in Abbotsford's computer system before implementing the smart metering infrastructure and its particular dataset (AMID). Therefore, as they were previously aligned finding a primary key that joins MIND and BILD was not a complex task.

In theory, the primary key used by the server, Blocks D, E, and F in Figure 1, would unify all datasets (AMID, BILD, and MIND). However, as it would disclose personal information and might cause a breach of customer information confidentiality, we were not provided with this primary key. As a result, we were needed to reverse engineer a primary composite key.

In addition, it was observed that for some meters, their units did not match among different tables. Standardizing those meters that were defined in ft^3, $gallons$, or $liters$ and converting them to m^3 also required precisely matching the records that, again, shows the significance of a global primary key. To ensure that a system-wide one-to-one match existed, some additional necessary steps were taken that will be described in the rest of part of the section.

Three individual fields that are shared among imported datasets and are the most probable candidates for reconstruction of the primary key are: (1) AccountID, which is unique number that is assigned to each consumer, (2) MeterID, which associates each meter to a representative number, (3) and RecordingDeviceID, which is a unique number for the digital hardware of the smart meter, which is responsible for recording and transmission of consumption measurements.

As the first trial, the AccountID field was used as the primary key. It resulted in more than 1600 records that could not be uniquely matched between AMID and other datasets (unmatched records). To resolve the above issue on the second trial, the fields AccountID, MeterID, and RecordingDeviceID were combined to be used as a primary composite key. As a result, the number of unmatched records considerably reduced to 1300. On the third attempt, we analyzed the inconsistencies throughout the tables and recognized that some fields in MIND and BILD had partial string matches with their AMID counterparts. Therefore, the join process was changed to accept the strings with partial matches as well as the complete ones (e.g. AccountID 789889 in BILD and MIND is the same as AccountID 789889_W in AMID).

Although we used partial matching for the three fields above, 67 records remained unmatched. As an amendment to the final solution, we manually implement the joining script, and, after finding the exact and partial three-field matches, paired the records with less than three (semi-)identical fields. The results contained only 13 unmatched records that were discarded as negligible errors.

3.4 Data Quality Issue, Missing and Duplicate Records

As the issue of missing and duplicate data are interrelated, this section would address both these issues together. At this phase, to find the cause of inconsistencies and errors, more understanding of the data specifications was required. The next step was to analyze missing and temporal availability of the data streams in AMID. Preliminary analysis of AMID showed that it should contain readings for approximately 27,000 data streams for every hour of the entire period of 365 days. A more precise examination revealed that, unanimously, all data streams missed recording for four days' worth of data. Abbotsford confirmed that their data servers were not functional during those days, and no recordings were gathered. Furthermore, we expected that AMID would contain separate readings for every individual meter/customer for the entire city. However, only 70% of the data streams contained the exact number of expected readings: 361 days multiplied by 24 hours.

About 25% of the data streams contained missing data and had fewer records while the remaining had duplicate data and had cases with multiple readings for one timestamp, which was the cause of excessive records. Among the data streams that had duplicate timestamps, some had equal reading values, as well. Some recent studies have focused on improving the quality of smart meter measurement data in electrical systems by dealing with missing and duplicate records ([36] and [33]).

In the cases of multiple similar readings, removing the redundant copies could be done by keeping one record and discarding the rest of the duplicate copies. The problem with the smart meter data in electricity is addressed by Shishido, as well, which is dealt with using the same approach. Nonetheless, a small portion of the multiple records with similar timestamps might not have equal reading values, as we encountered in this case study. The way that these duplicates were dealt with will be discussed in the "context-dependent errors" part of the section in detail. Shishido reports that the duplicate data can be caused by several factors, such as meter resolution inconsistency (recording measurement in multiple units as duplicates), the different number of significant digits of meters, and daylight saving duplicates. However, in the case of our study on water data, mostly the duplicates are caused by discrepancies in dataset operations. In the current study, some rare cases of duplicate records due to the daylight saving were observed, as well.

Because of the existence of duplicate records in an individual data stream, we hypothesized that the same issue might exist in different data streams as well. In other words, some meter data streams might be copies of each other, which by calculating checksum values we confirmed the existence of such duplicate data streams and removed them from AMID. Further investigations revealed that the duplicates were the exact copies of one meter that erroneously misrepresented as two different ones. This error could have originated at some step between recording consumption values and organizing them as raw streams of data. As an example, the field RecordingDeviceID is entered incorrectly in different instances of the record; therefore, the resulting join would contain two duplicate readings for both the correct and incorrect values.

After the second join process, the number of distinct data streams in AMID dropped from about 27,000 streams to approximately 25,500, and only 30 unmatched records remained. A unique primary key, UWID, was introduced to the relational tables, as well. We believed that at this point the first barrier of obtaining access to error-free data was passed. However, our further analysis showed that the data was still not clean enough to undergo any data mining and analysis. From this point on, all efforts on cleaning the data were focused on AMID's high-resolution fine-grained measurement data streams. Concluding the duplicate records analysis, our strong belief is that discrepancies between blocks E and F of the Figure 1 is the cause of duplicate records problem.

Next, the focus was to analyze the missing records. Answers to the following questions can provide us with the valuable insight of the nature of the problem: 'How are the missing records distributed in different data streams?' moreover, 'Do these missing incidents are more probable for customers with higher annual consumption rates, categorized as high consuming clients or do they occur uniformly?' Figure 2 shows the percentage of missing data for customers in a range of low consuming (left) to high consuming (right). As the figure indicates, for most customers in the normal range, the missing data has a relatively uniform distribution. Two small peaks for low and high ends of the plot can be seen, as well. Higher probability of missing records occurring for high consuming customers can be troublesome because they are the focus of the study and most likely are the main contributors to the peak. Therefore, the issue of missing data should be considered with more caution, while identifying the peak contributors. The low end of the plot in Figure 2 (left) is not as critical as the high end (right), as it is only an indicator of customers who are consuming hardly any water or even have virtually zero consumption.

One of the preferred approaches to deal with missing data in a dataset with large sample size is to reduce the size of the dataset [23] simply by removing sample time series that include excessive missing data. However, this approach is only acceptable in the case that the proportion of missing data records to the size of the dataset is negligible. Otherwise, it would lead to unexpected misrepresentation of information. This approach was previously adopted for smart meter data in electrical grids, and if the requirements are met, is shown to provide reliable results [33]. Therefore, because of the above reasons, it is adopted in the current study, as well. The missing data issue is examined in the results part of the section and is handled accordingly. However, we have to reiterate that re-examining the missing data issue, once the quality of the entire database is determined with higher confidence, is essential.

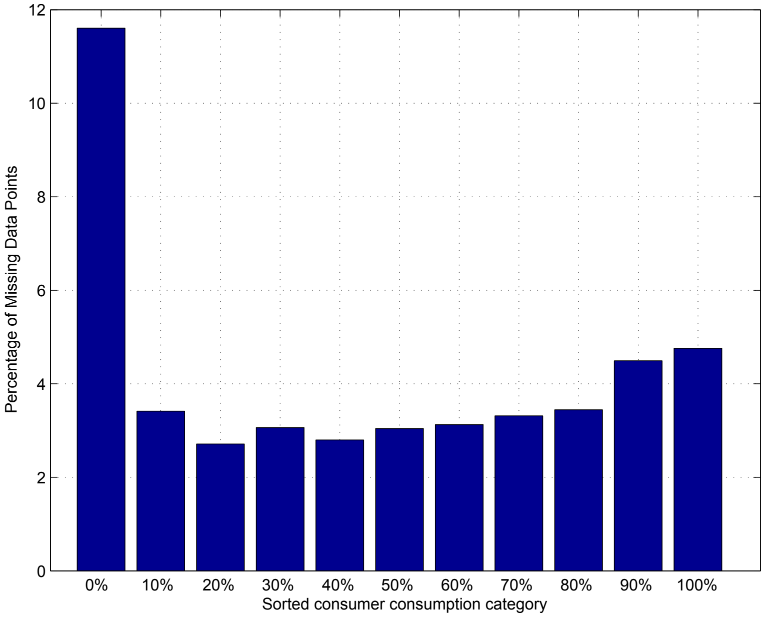

Fig. 2. The percentage variations of the missing data for consumption values ranging from lowest (left, 0%) to highest (right, 100%). According to the plot, more than 95% of the recorded measurements of the high consuming customers are available.

3.5 Context-Dependent Errors

Context-dependent errors have a particular feature that easily distinguishes them from the ones discussed previously: *They are not caused by issues arising from the data transformations.* For example, a meter can fail abruptly and record incorrect readings. As a result, it might record a spike in consumption that is not valid. While analyzing readings to verify the expected features of the consumption profiles, such as demographic features of consumptions, the inconsistencies led to detect each specific error type. Each of the following subsections highlights a context-dependent error that exists in the AMID and provides possible solutions, as well.

3.5.1 Context-Dependent Error, Quantized Meter Readings. The first unexpected behavior encountered in the AMID was a Quantized Meter Reading error. Data streams with this condition only had quantized recordings. In other words, the readings are rounded to a lower resolution. For instance, Figure 3 shows an example of this error in a consumption profile.

This pattern can lead to spiking behaviour in consumption and trailing zero readings for some instances. However, further study revealed that the recording resolution, which is usually 1 Litre, is considerably larger for the affected data streams. After reporting a sample set of these errors to Abbotsford, it was confirmed that an out-of-date setting in some meters would cause reading in bigger quantized steps, e.g. five cubic meters. As this problem is caused at the measurement point, consequently, higher-frequency information was not recorded and completely restoring the data to its original state is impossible.

This anomaly would also affect other aspects of the analysis. Some examples of the caused problems are: 1) the peak duration should be considerably larger than the minimum time to consume one quantization step; otherwise, the results might miss some peak contributors. 2) because of the imposed coarse data resolution, the load disaggregation at the household level would not be possible. 3) spikes and high-frequency noise components are introduced to data without carrying additional information.

We examined several methods to improve data quality for the affected data streams; however, none of those methods enhanced the quality of data substantially. Therefore, the final decision at this stage of the process was to keep the quantized meter readings as they are and make sure that their condition would not considerably affect further analysis steps.

3.5.2 Context-Dependent Error, Unexpected Spikes. Another observed error type is the *unexpected spike.* A spike error is defined as a short-duration high-amplitude negative or positive change in the consumption profile that does not reflect a real consumption measurement. However, some spikes can be legitimate consumption patterns, and the differences between faulty instances from real ones may not be clear, as shown in Figure 4. The left spike's validity is confirmed by Abbotsford experts, while the right spike is invalid and the recorded instance is faulty. According to Abbotsford experts, genuine spikes are acceptable water consumption patterns and can be distinguished by the fact that they are usually comparable to the average consumption in value. One way to repair the detected errors is replacing them with the mean value of windows of correct readings in the time series close neighborhood. The number of readings used for finding the value is determined by data variability; and, in this study, eight hours was the minimum required duration. In the current study, the reading values that are considerably larger than the maximum spike-free consumption (1000 times higher regardless of the sign) are considered to be erroneous.

Even with all the correcting measures, some noisy spikes inevitably remain unnoticed. The missed spike errors are usually in the ordinary range of valid spikes and require more domain knowledge to be classified correctly. The existence of such residual errors, along with the other remaining error types, should be taken into consideration while analyzing and mining the smart meter data.

In general, various malfunctions of the hardware or software can lead to a spike in meter recordings, which should be handled more cautiously.

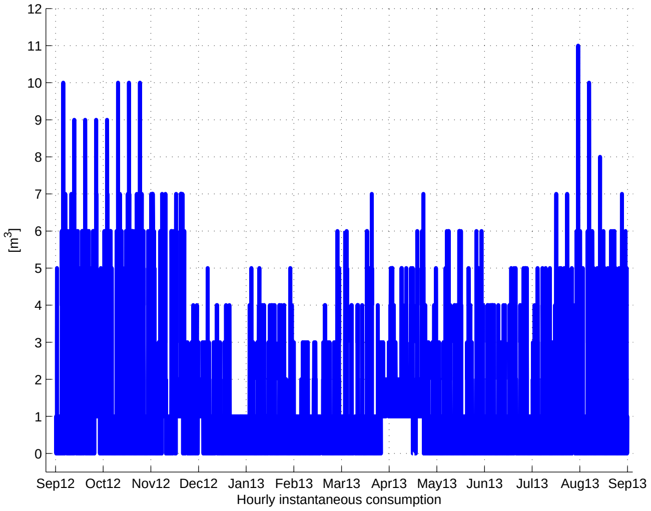

Fig. 3. An instance of a consumer load profile with quantized levels of instantaneous meter reading.

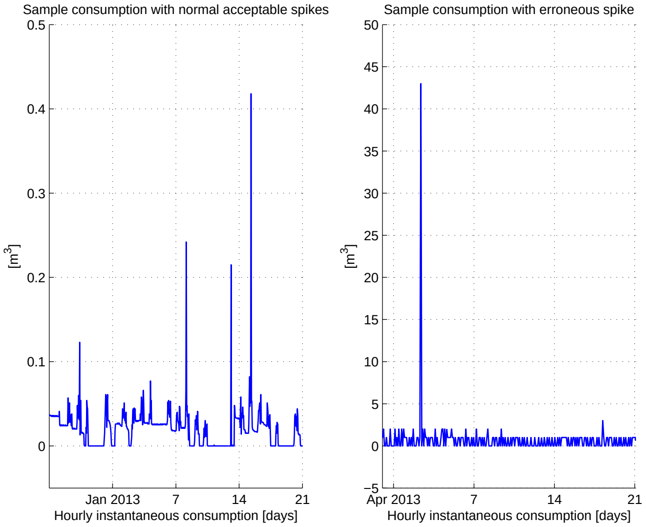

Fig. 4. Comparing acceptable spike example (left plot) and spike as error example (right plot)

3.5.3 Context-Dependent Error, Meter Unit Inconsistencies. The errors highlighted so far have been presented in the data quality literature, and their characteristics have been discussed, to some extent by [33], [36], and [23].

The *Meter Unit Inconsistency (MUI)* error has the highest impact in the current study and has not been previously considered extensively. Figure 5 shows a data stream affected by this error. The plot clearly indicates that there has been a considerable sudden decrease of consumption on May 6th. The maximum consumption values drop from 10 cubic meters to 1 cubic meter, which is the result of reprogramming the meter. In this case, the left side of the profile is represented by an invalid unit of measurement and the maximum of one cubic meter is correct. This unusual behavior was discovered in AMID while attempting to find the peak contributors after cleaning the unexpected spikes. From the perspective of peak contribution for this customer, the main cause of peaking can be the sudden change in the consumption pattern. To verify the validity of this assumption, we discussed this specific profile with Abbotsford experts. Apparently, the step in the consumption is an error and not a reflection of a real phenomenon. Therefore, a technical issue in the datasets could be the source of this problem.

Another possible explanation could be at least one part of the profile, either before or after the jump, should be correct and the increase may be the result of a multiplicative error. Investigating the new hypothesis revealed that, if the two segments of the plot were normalized to correct maximum and minimums, both sides could be valid consumption profiles. Consequently, the cause of the error is that the second part is multiplied by a constant factor, approximately 219, which is the coefficient required to convert cubic meters to imperial gallons.

To repair the MUI error, it is necessary to verify the consumption of a customer in both segments of the profile. Therefore, without verification with some representation of the untransformed data, unlike other types of errors described so far, it is not feasible to recover accurate measurements. In the current study, bi-monthly records in BILD were used as pre-transformation evidence data and provided a way to verify the consumption values. Although it is not possible to prove that all bi-monthly recordings are correct readings, billing dataset records are least affected by data transformations that are involved in the system, refer to Figure 1. Additionally, the validity of these records is examined by most consumers evaluating their bills and reporting any discrepancy between readings and their actual consumption. In any case, to resolve the MUI errors can be quite effective as this error is caused by the data transformation not actual measurement error.

The comparisons showed that the left side of the current example conforms to the standard unit, m^3, so the other part requires reconstruction. Analyzing the dirty data streams to find the cause of MUI error indicated that the origin could be the operator re-configuring the meter's unit and fail to update the affected records. This hypothesis was presented to Abbotsford, and its validity was confirmed. It is worth mentioning that this error type would not necessarily affect Block E's billing records. It is the result of calculating monthly billing records based on the meter units at the time of reading. In contrast, the meter unit change command does not back-propagate or update the historical measurements in Block F and would cause this issue. Further analysis showed that most of these configuration updates were performed to change the meter's resolution in an attempt to remove the quantized meter errors, discussed previously.

Approximately 5% of the records were affected by this type of error. Comparing the dirty records also revealed that the required units and conversions, necessary to resolve MUI error, were not uniform among different data streams. For instance, conversions from gallons, cubic feet, and liters to cubic meters were observed. Because of this variety in MUI error, and the fact that the time stamp of the multiplication event was not fixed throughout AMID, the only possible solution was to perform this task with the help of a human expert.

The first step of removing MUI errors is to find the affected data streams. In most cases, there is a significant change in the signal level before and after the error event. Thus, it is expected that the statistical features related to each part would change considerably. By calculating the monthly standard deviation of each data stream, 12 values would be generated. A second-degree standard deviation (STD2M) of the 12 mentioned results can be calculated, as well. As this criterion exaggerates the discontinuities in the data stream, STD2M is mostly effective for finding MUI errors, spikes, and resets. However, for the quantized meters, it does not work as expected. The following formula defines the threshold value for STD2M.

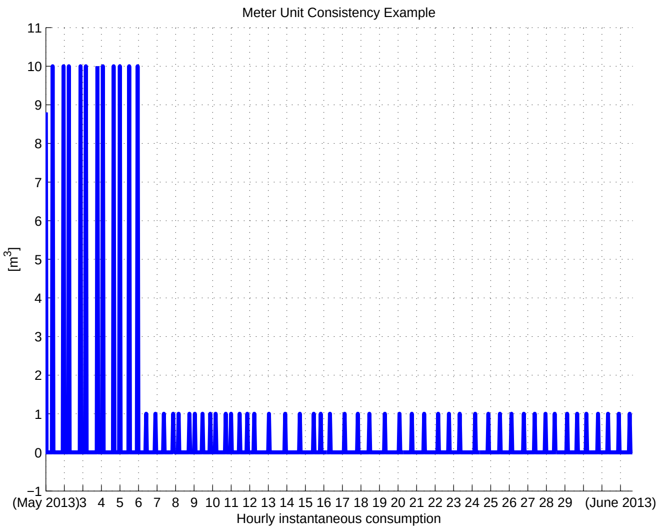

Fig. 5. An example of meter unit inconsistency error

$$\textit{Let } C_{d,i} \textit{ be the consumption of data stream } d \textit{ over hour } i \textit{ in cubic meters}.$$

$$STDM(d, m_k) = \sqrt{\frac{1}{m_k - m_{k-1}}} \sqrt{\left(\sum_{h=m_{k-1}+1}^{m_k} (C_{d,h})^2 \right) - \left(\sum_{h=m_{k-1}+1}^{m_k} C_{d,h} \right)^2}, \tag{1}$$

and

$$STD2M(d) = \frac{1}{\sqrt{12}} \sqrt{\left(\sum_{q=1}^{12} (STDM(d, m_k))^2 \right) - \left(\sum_{q=1}^{12} STDM(d, m_k) \right)^2}, \tag{2}$$

where d is the index of each smart meter generating a data stream, m_k is the index of first day of each month in the measurement streams multiplied by the number of hourly readings per day ($0 \times 24 + 1 = 1$ for Jan 2012 and $244 \times 24 + 1 = 5857$ for Sep 2012), and k is the month number offset from Jan 2012 that is valid between 9 and 21 (mid 2012 to mid 2013). In a clean data stream, the STD2M would remain lower than a specific threshold, which experiments showed it could only go as high as 60 for a normal data stream. However, a data stream that is contaminated by a unit multiplication error would naturally result in a considerably higher second-degree standard deviation, the minimum observed value was 200 in our experiments. To have a stable and robust threshold, the records with STD2M values of lower than 40 and greater than 250 are considered, with high confidence, to be clean and dirty streams respectively. Those rare streams with values between 40 and 250 should be inspected by the expert for higher precision, which only ten instances were observed in our study. It should be noted that 40 and 250 are the values for cubic meter consumption units and can change by modifying any parameter in the system. By using the STD2M criterion, it is possible to detect the potential candidates for this error and to ask the expert to judge whether it is a genuine MUI error.

To conclude, the process of removing the MUI error consists of two distinct steps: constructing a mechanism capable of finding the error patterns in the data stream and devising a repair method that eliminates the error from the data and replaces it with a suitable substitute. The performance of the process is mostly dependent on correctly recalling errors as much as possible and less importantly on the number of false positives it finds. To eradicate MUI errors, it is required to inspect each consumption profile individually and compare them to the ground truth, which is done automatically. Afterward, the segment of the profile that does not match with the billing records is manually selected and multiplied by a correction value. Therefore, there is a trade-off between finding a threshold that ensures the errors will be detected and generating excessive unwanted false positives. It would ensure that the manual inspection task would not impose an unnecessary amount of workload on the expert.

3.6 Results and Sensitivity Analysis

Thus far, different types of errors were introduced, their origins were described, and solutions were provided to deal with them. We summarize this process in Table 4.

Table 4. AMID Dataset state after each phase of cleaning. The symbol "⇒" indicates that the errors in the data streams were carried forward to the next phase.

Phase	0	1	2	3	4	5
Operation (Related Section)	**Duplicate Data stream removal (Section 3.4)**	**Datasets Unification (Section 3.3)**	**Duplicate records elimination (Section 3.4)**	**Peak analysis attempts (Section 3.6.3)**	**Statistical rule-based error filtering (Section 3.5)**	**Manual Repair using ground truth (Section 3.5)**
Duplicate Data Streams	Solved	Solved	Solved	Solved	Solved	Solved
Duplicate records in streams	⇒	⇒	Solved	Solved	Solved	Solved
Unexpected Spikes	⇒	⇒	⇒	Some resolved	Solved	Solved
Quantized meter readings	⇒	⇒	⇒	⇒	⇒	⇒
Unexpected Unit Multiplications	⇒	⇒	⇒	Some resolved ⇒	Some resolved ⇒	Solved
Missing Records	⇒	⇒	⇒	⇒	⇒	Possible to Approximate
Data Streams Count	27,000	25,500	25,500	25,500	25,500	25,500

Different phases of the cleaning and treatment process are shown in this table, and the related section of each process are also referenced. The most important observation based on this table is the sequential treatment of the errors. In fact, each cleaning step provided us with means to work on the next type of error, which was more complex in nature. This part of the section analyzes the Abbotsford smart meter data to determine peak contributors and how their order and ranking would change due to each type of error.

To get a general sense of the water consumption behavior in this dataset, Figure 6 shows the annual consumption of the entire city, calculated using the cleaned hourly data, and aggregated across all customers. Note that there is a strong weekly pattern in consumption, reflecting the workweek, modulated by an annual pattern, reflecting the seasons.

3.6.1 Peak Definition and Peak Contributors. In this part of the subsection, a brief definition of peak and peak contributors in the context of water distribution systems will be provided. Because of the inherent characteristics

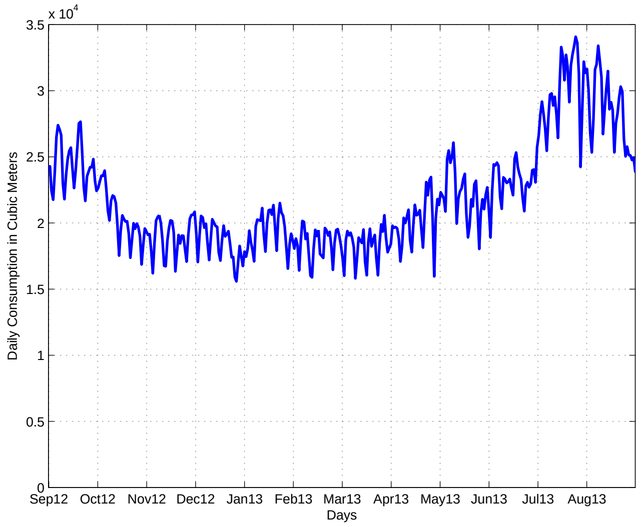

Fig. 6. Annual consumption profile of the entire city. The weekly period consumption period behaviour can be observed.

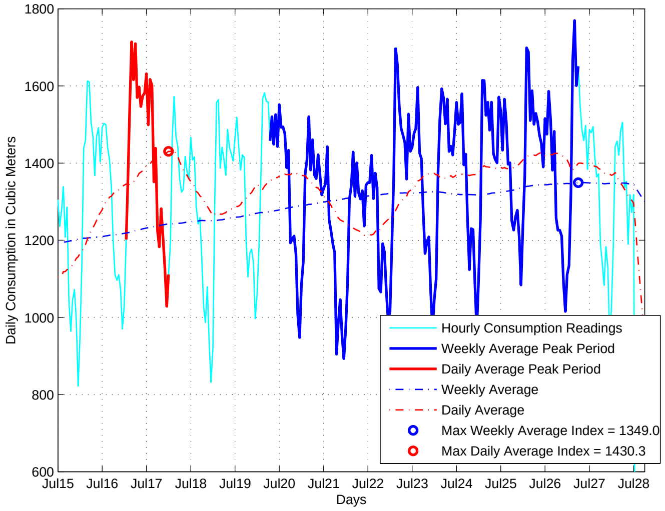

Fig. 7. A close-up view of the consumption during the peak period of July 2013. Weekly and daily peaks are highlighted in the plot, and it is shown why they do not overlap.

of water supply systems, instantaneous peak consumption does not have a significant practical value. Thus, in the context of such large-scale systems, the peak value is described as *the maximum average consumption of a consumer (or group of consumers), during a specific time range R (in hours or days), for a predefined constant window* *size (W in hours or days)*. The peak duration is formulated in the Eq. 3 as:

$$Let\ C_{d,i}\ be\ the\ consumption\ of\ data\ stream\ d\ over\ hour\ i\ in\ cubic\ meters.$$
$$W_{peak}^{hours} = \begin{matrix} max \\ W \in R \end{matrix} \sum_{d \in D} \sum_{i \in W} C_{d,i},$$

where D is a set of data streams (water meter readings) that we are interested in. The proper width of the window, W, is always specified in hours (168 hours for a seven-day-long window, for example). For this paper, as we are interested in the peak load of the entire city, the set D will consist of all the smart water meters across the city. Furthermore, R is a range of hours, W is the constant size of a window moving over the range of hourly readings, i is the index of each consumer, and d is each data point in that window. Depending on the application, the peak averaging window (W) can take values of a few hours to a few weeks (depending on the natural lag and physical size of the water transmission network in question). We only focus on the peak values for window sizes of 24 and 168 hours in the current study, which is based on the requirement of the specific water supply system defined by the City of Abbotsford. Similarly, the average consumption of peak contributor(s) during the peak period can be defined by the Eq. 3.

$$PeakLoad_{peak}^{hours} = \frac{\sum_{d \in D} \sum_{i \in W} C_{d,i}}{|W|} \quad s.t.\ \ W = W_{peak}^{hours} \tag{3}$$

3.6.2 Ranked List of Peak Contributors and Comparing Them. Section 3.6.3 provides a ranked list of contributors to the demand peak. (*peak contributors*) determined from both the clean and dirty datasets. We would like to quantify the effect of each meter error on data quality by comparing the corresponding ranked lists. To do so, a suitable ranking evaluation method is required.

The ranking algorithm proposed by Kendal et al. is extensively used to compare an erroneous permuted or partially permuted list with a given (correct) reference [24]. This algorithm, which builds on prior work by Spearman [7], is widely used in applications such as biomedical data mining and the Internet page ranking (see [32] and [34]). Several variants of this method are well-known [10], [30], [35], and [26]. We use the variant that permits weights for each rank suggested by [35] and [10].

In the unweighted version of the metric, an error in the 100th position on the list is as significant as an error in the first position. We would, instead, like to give less importance to an error associated with a customer with lower water consumption. It motivates our use of the weighted version of Kendall's Tau. In the current study, the natural choice of weight is the water consumption of each customer, defined as $w(.)$, as determined after data cleaning. Weighted Kendall's Tau, K_w, can be defined as

$$K_w(\sigma, \pi) = \sum_{1 \leq d' < d \leq n} \frac{w(d') + w(d)}{2} [\pi(d') > \pi(d)], \tag{4}$$

$$k_w = 1 - \frac{2K_w(\sigma, \pi)}{\sum_{\{d', d \in \sigma \cup \pi : d' < d\}} \frac{w(d') + w(d)}{2}} \tag{5}$$

In equation 4, σ and π are permutations of length n and rankings of element i are defined by $\sigma(i)$ and $\pi(i)$. Permutation σ is assumed to be the reference and π is the erroneous ranking list whose performance is to be evaluated, that is, the highest contributors to the peak after different stages of data cleaning. The expression $[\pi(i) >]B(j)]$ is equal to one if the ranking condition holds and otherwise zero.

Section 3.6.4 will provide and analyze the results that are generated by comparison of clean and erroneous rankings, generate by different stages of noisy data, in detail.

3.6.3 Peak Contribution Results. The City of Abbotsford informed us that the highest peak consumption record occurred on July 24, 2013. To find those consumers who most contributed to this peak, a peak length is required that can accommodate the natural lag existing in water supply networks. Therefore, two peak window periods are selected for the current study: 24-hours and one week, as representatives of short and medium term consumption peaks. Additionally, to emphasize the effect of noise and data errors, results are generated using both clean and dirty datasets. Dirty data contains errors that were described previously; while, clean data is produced by removing the errors, which was performed semi-automatically under expert supervision.

Tables 5 and 6 compare the results of calculating peak windows of length 24 and 168 hours. As indicated in Table 5, the peak event occurs in a 24-hour period starting on July 16, 2013, at 3:00 pm. However, the respective peak event for the dirty original dataset started at Feb 19, 2013, at 12:00 am. Not only does the detected time do not match the correct peak, which exactly overlaps with the Abbotsford's report, but also no justifiable reason exists for a peak occurring in winter. A similar pattern can be observed in the case of the weekly peak in Table 6. By comparing the peak values in correlating clean and dirty cases, a significant inconsistency can be observed, which is caused by both enlarging and deforming the distribution by associating high consumption to a small set of customers.

Both tables provide the top ten consumers and their categories. The correct ranking of dirty data, calculated using clean results, is also provided. Only two consumers in the clean top ten are detected correctly in dirty data, although with the wrong order, and the remaining are not valid peak contributors.

The share of each category in top ten peak is also significantly different. The dirty top ten are mostly multi-family residences (MFRES), while the real contributors are linked to the agricultural and industrial sectors. Another unexpected observation is that the first peak contributor in dirty data for 24-hour window size, Table 5, has real consumption of zero. This behavior is explained by the fact that the peak period of dirty data is in a different season comparing to the actual period. Therefore, it is understandable that the consumer has high consumption in one season and none in another one. As for this specific customer, the consumption during winter is zero; it might be an indicator that the customer has gone on a vacation. In electrical systems, actual zero consumption would be an alarming symptom of a fault in the system; this is because of the electronic gadgets and their standby consumption it is hardly possible to reduce the consumption to absolute zero. However, in smart water meter based systems it is an acceptable state of the meter and tolerable.

Also, to ensure that the missing data points would not introduce any unwanted distortion to the data, the peak periods were examined for missing records. The 24-hour peak period is clean and the seven-day peak period only has less than 2% missing records. To compensate for this error, the weekly averages of customers with missing records were scaled up.

In comparison with current results, Abbotsford's reported highest consumption day, Jul 24, 2013, falls exactly into the range of the results of seven-day peak contribution, confirming the cleaned dataset results. For illustrating why the 168 hours peak and 24-hours peak do not overlap; a selected range of the peak days of July 2013 are shown in Figure 7. The red and blue dotted lines are the daily and weekly averages, respectively, and their maximum points are highlighted with two non-overlapping circles.

3.6.4 Weighted Kendall's Tau. The next step is employing Weighted Kendall's Tau correlation factor for evaluation the performance of ranking methods. The weight coefficient of each contributor in both datasets is the clean monthly consumption during the highest peak of July 2013 with the peak duration of one month. As the final correlation coefficient is normalized to the range [-1,1], the weight vector does not require to be normalized.

Figure 8 shows the correlation results of using Weighted Kendall's Tau method using ranking lists of top hundred customers. The ranking list is extended to cover the entire list of consumers in both cases. As the extended lists are mostly similar, top 100 peak rankings would have approximately 25,000 matched rankings comparing to the other one.

Table 5. Comparison of the top ten peak contributors for both clean and dirty states of data for the peak window length of 24 hours. The data streams with missing records were scaled to compensate. Categories (CAT) are abbreviated as Agricultural (AGR), Commercial (COM), Industrial (IND), Institutional (INS), Multi-Family Residences (MFR), and Single-Family Residence (SFR).

	Clean Data (24 Hour Peak) Start: Jul 16, 2013,3:00pm End: Jul 17, 2013,3:00pm		*Dirty Data (24 Hour Peak) Start: Feb 19, 2013, 12:00am End: Feb 20, 2013, 12:00am*				
Rank	**Clean Data Cat.**	**Cons. in [m^3]**	*Dirty Data Cat.*	*Cons.* in [m^3]	*Rank in Clean Data*	*Real Cons.* [m^3]	*Cons. Error* $\frac{Dirty-Clean}{Clean}$
1st	IND	1,355.0	*SFR*	*511,528.0*	*23891*	*0*	*NaN*
2nd	IND	1,338.0	*SFR*	*8,104.0*	*1633*	*2*	*405100%*
3rd	IND	1,003.0	*COM*	*5,000.0*	*919*	*5*	*99900%*
4th	COM	775.0	*COM*	*2,426.0*	*443*	*17*	*14171%*
5th	AGR	611.6	*MFR*	*2,000.0*	*1034*	*4*	*49900%*
6th	IND	536.0	*IND*	*1,500.0*	*3*	*1003*	*50%*
7th	IND	523.0	*MFR*	*1,389.0*	*501*	*15*	*9160%*
8th	IND	519.2	*MFR*	*1,071.0*	*569*	*12.16*	*8708%*
9h	IND	467.0	*IND*	*1,000.0*	*2*	*1338*	*25%*
10th	INS	395.0	*INS*	*659.0*	*1589*	*2*	*32850%*

Table 6. Comparison of the top ten peak contributors for both clean and dirty states of data for the peak window length of 168 hours (a week). The data streams with missing records were scaled to compensate. Categories (CAT) are abbreviated as Agricultural (AGR), Commercial (COM), Industrial (IND), Institutional (INS), Multi-Family Residences (MFR), and Single-Family Residence (SFR).

	Clean Data (7 Day Peak) Start: Jul 19,2013,7:00pm End: Jul 26,2013,7:00pm		*Dirty Data (7 Day Peak) Start: Feb 18,2013,4:00pm End: Feb 25,2013,4:00pm*				
Rank	**Clean Data Cat.**	**Cons. in [m^3]**	*Dirty Data Cat.*	*Cons.* in [m^3]	*Rank in Clean Data*	*Real Cons.* [m^3]	*Cons. Error* $\frac{Dirty-Clean}{Clean}$
1st	IND	8,367.0	*SFR*	*511,531.0*	*2832*	*5*	*10230520%*
2nd	IND	7,539.0	*COM*	*20,000.0*	*1170*	*20*	*99900%*
3rd	IND	4,569.1	*MFR*	*17,000.0*	*1105*	*22*	*77173%*
4th	COM	4,480.0	*IND*	*11,738.0*	*1*	*8367*	*40%*
5th	AGR	4,373.7	*MFR*	*9,748.0*	*497*	*96.24*	*10029%*
6th	IND	4,030.0	*MFR*	*8,500.0*	*441*	*110*	*7627%*
7th	IND	3,765.0	*SFR*	*8,117.0*	*1223*	*19*	*42621%*
8th	IND	3,340.0	*INS*	*6,000.0*	*1341*	*16*	*37400%*
9h	IND	3,044.0	*IND*	*5,305.0*	*2*	*7539*	*30%*
10th	AGR	2,743.0	*COM*	*4,526.0*	*487*	*99.9*	*4431%*

The next observation is that the spikes would cause less difference of correlation comparing to the meter unit inconsistencies. However, by increasing the duration of the peak, the former would converge to correct results. In contrast, the dataset with the spikes is virtually transparent to the changes in the peak duration.

To explain the sudden changes of the correlation amount in almost all plots, Figure 9 shows the variations of detected peak day, based on data type and peak window size. The Y-axis is the offset of the detected peak day from reference day of Jan 1, 2012 and X-axis is the length of the peak window duration. As the plots also indicate, the sudden changes of the correlation in Figure 8 is always accompanied by a sudden change of the detected peak day, as well. Another important factor that would deteriorate the quality of a ranking is that in most cases, except the situation of meter unit inconsistencies with peak duration of more than 50 days, the detected peak day is at least four months away from the correct peak day. This phenomenon would contribute to decreasing the correlation of rankings even further.

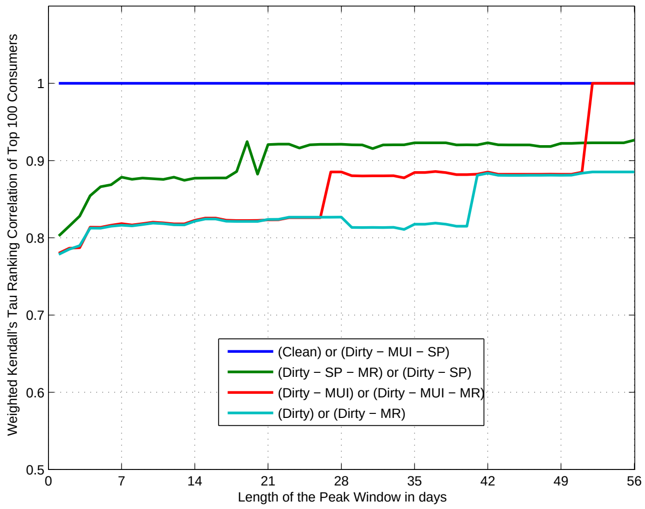

Fig. 8. Comparison of the effect of different errors on the Weighted Kendall's Tau correlation coefficient, between top 100 peak contributors calculated using dirty and clean data

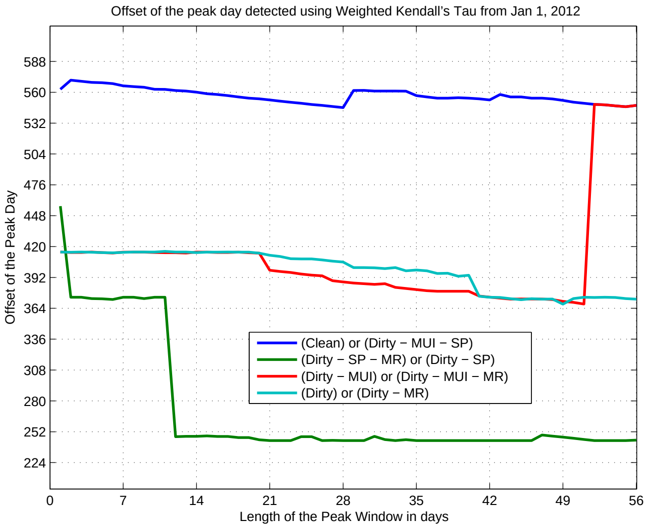

Fig. 9. The effect of peak window duration on the detected maximum consumption day for both clean dataset and affected by different types of error

3.6.5 Recall Percentage Metric for k-top Contributors. Another method, of analyzing the effects of errors in the data, is to calculate the percentage of correctly recalled customers in a set of k-top contributors of dirty data. ‘In other words, this metric answers the following question: *How many dirty top contributors should be analyzed to ensure that the correct 100 top contributors are covered?*. Figure 10 shows the related profiles for 24 hours respectively. According to the Figure, roughly 200 top peak contributors of the dirty data would include $110 = 200 \times 0.55$ correctly recalled clean peak contributors. As another example, in Figure 10, at $x = 500$, y is approximately (0.7). Therefore, it means that the top-500 list based on dirty data includes $500 \times 0.7 = 350$ of the true top-500 consumers according to clean data.

Based on the definition of this plot, the correlation of clean and dirty top peak contributors would increase as it gets closer to the clean line, constant 100% recall. In this case, clean data is used as the reference, which should have the constant recall rate of 100%. In comparison, the effect of different errors is illustrated by the other three plots. The effect of the meter reset error has been negligible in all cases, and as a result, those plots were omitted from the graph and grouped with their approximated counterparts.

From the highest order of influence descending, spikes, meter unit inconsistencies, and meter resets would affect the data quality and recall rate. Although the recall percentage profile should reach 100% with increasing k to the entire dataset, by passing 70%, the increase rate drops significantly with the knee point of 400 customers.

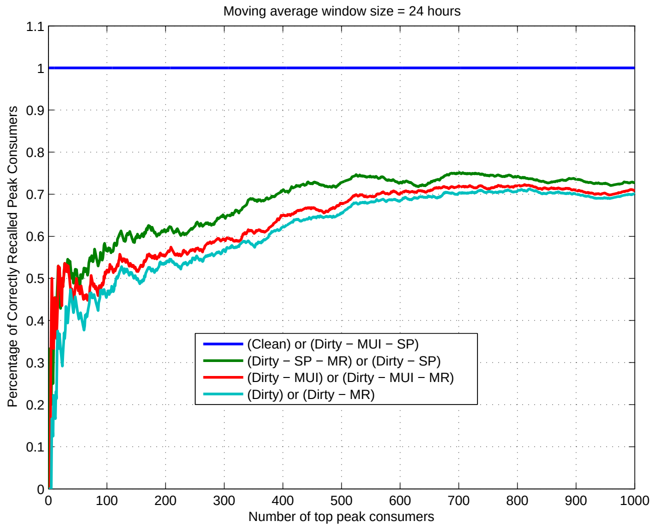

Fig. 10. The percentage profile of correctly recalled top contributors for dirty data. The duration of peak contribution window is 24 hours. The error abbreviations are: MUI = Meter Unit Inconsistencies, SP = Spikes, MR = Meter Resets

4 CONCLUSIONS

Wireless smart meter infrastructures are being adopted by an increasing number of municipalities. These meters are equipped with various measurement and transmission features and would enable the service providers to have demand-supply flexibility, such as: dynamic demand response, to send alerts to consumers of possible leaks in their residences, to characterize the demand profile of each category of customers and model them, and to analyze consumption in more detail.

However, to perform these tasks, as an essential part of the process, measurement data needs to be error-free. Studies have found that in a majority of the cases data is not in the normal condition, and measurements mixed with various kinds of errors are generated by the meters.

The current study focused on the impact of data errors on the performance of peak consumer identification. The adopted case study is the infrastructure of the City of Abbotsford in British Columbia, Canada. Various sources of errors, such as mistakes made by operators, hardware failures, and context-dependent errors were identified as well. As well, systematic ways of removing the main contributing errors (meter unit inconsistencies, meter resets, spikes, duplicated records, and duplicated data streams) were provided and more complex errors were characterized, as well.

The results of cleaning data and performing peak detection tasks were presented, and the significance of the cleaning process was demonstrated. Also, the sensitivity of the outputs to the errors in the data and the parameters of the peak detection filter were examined.

To conclude, data cleaning is an essential part of smart meter measurement analysis, and there is a possibility of extracting valid information out of semi-clean data. However, prior knowledge of the sensitivity of the results to different types of error is required.

Smart meter data analysis is in its early stages and can benefit considerably from further research. Some possible extensions of the work were provided in the paper. Most importantly, we need to evaluate the data quality further using other physical characteristics of the water supply infrastructure, assuming feasibility of acquiring them, such as pressure information of various key nodes, mass balancing of the consumption and production, using bulk meter data of the network. With the current error-free state of data, other filters can be used to detect possible errors such as: does the hourly consumption profile of different customer categories follow the expected minimum and maximum load?

The other extension is to examine the effect of quantized meters on data quality further and to devise cleaning methods that can deal with such error types. In addition, missing data points, an inevitable aspect of every smart system, were only analyzed and their effects were compensated. Similar to the procedure performed for errors, missing data can be characterized with more systematic techniques, as well.

REFERENCES

[1] Aijun An, Ning Shan, Christine Chan, Nick Cercone, and Wojciech Ziarko. 1996. Discovering rules for water demand prediction: An enhanced rough-set approach. *Engineering Applications of Artificial Intelligence* 9, 6 (1996), 645 – 653. DOI:http://dx.doi.org/10.1016/S0952-1976(96)00059-0

[2] Francisco Arregui, E Cabrera, Ricardo Cobacho, and Jorge García-Serra. 2005. Key factors affecting water meter accuracy. In *Leakage 2005*. Leakage 2005, Portugal, 1 – 10.

[3] Cara Beal, Rodney A. Stewart, T Huang, and E Rey. 2011. SEQ residential end use study. *Australian Water Association* 38(1) (2011), 80–84.

[4] Cara Beal, Rodney A. Stewart, T Huang, and E Rey. 2011. South East Queensland residential end use study: Final Report. *Journal of the Australian Water Association* 38, 1 (2011), 80–84.

[5] Cara D. Beal and Rodney A. Stewart. 2013. Identifying Residential Water End-Uses Underpinning Peak Day and Peak Hour Demand. *J. Water Resour. Plann. Manage.* ja, Article 04014008 (2013), 10 pages. DOI:http://dx.doi.org/10.1061/(ASCE)WR.1943-5452.0000357

[6] Christopher Bennett, Rodney A. Stewart, and Cara D. Beal. 2013. ANN-based residential water end-use demand forecasting model. *Expert Systems with Applications* 40, 4 (2013), 1014 – 1023. DOI:http://dx.doi.org/10.1016/j.eswa.2012.08.012

[7] Spearman C. 1906. A footrule for measuring correlation. *British Journal of Psychology* 2 (1906), 89–108.

[8] Graham Cole and Rodney A Stewart. 2013. Smart meter enabled disaggregation of urban peak water demand: precursor to effective urban water planning. *Urban Water Journal* 10, 3 (2013), 174–194.

[9] Martin Courtney. 2014. *How utilities are profiting from Big Data analytics.* Engineering and Technology Magazine. http://eandt.theiet.org/magazine/2014/01/data-on-demand.cfm

[10] Paolo D'Alberto and Ali Dasdan. 2010. *On the Weakenesses of Correlation Measures used for Search Engines' Results.* Cornell University Library. http://arxiv.org/pdf/1107.2691v1.pdf Access on: 2014-12-15.

[11] C. Efthymiou and G. Kalogridis. 2010. Smart Grid Privacy via Anonymization of Smart Metering Data. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on.* 238–243. DOI:http://dx.doi.org/10.1109/SMARTGRID.2010.5622050

[12] Shu Fan and Luonan Chen. 2006. Short-term load forecasting based on an adaptive hybrid method. *Power Systems, IEEE Transactions on* 21, 1 (Feb 2006), 392–401. DOI:http://dx.doi.org/10.1109/TPWRS.2005.860944

[13] M Fantozzi. 2009. Reduction of customer meter under-registration by optimal economic replacement based on meter accuracy testing programme and Unmeasured Flow Reducers. (2009).

[14] H. Farhangi. 2010. The path of the smart grid. *Power and Energy Magazine, IEEE* 8, 1 (January 2010), 18–28. DOI:http://dx.doi.org/10.1109/MPE.2009.934876

[15] Kelly S. Fielding, Anneliese Spinks, Sally Russell, Rod McCrea, Rodney A. Stewart, and John Gardner. 2013. An experimental test of voluntary strategies to promote urban water demand management. *Journal of Environmental Management* 114, 0 (2013), 343 – 351. DOI:http://dx.doi.org/10.1016/j.jenvman.2012.10.027

[16] Matthias Heinrich. 2007. *Water End Use and Efficiency Project (WEEP): Final Report.* Technical Report. BRANZ Ltd., Judgeford, New Zealand. BRANZ Study Report 159.

[17] Lon House. 2011. *Time of Use Water Meter Impacts on Customer Water Use.* Technical Report. California Energy Commission.

[18] S.-C. Hsia, S.-W. Hsu, and Y.-J. Chang. 2012. Remote monitoring and smart sensing for water meter system and leakage detection. *Wireless Sensor Systems, IET* 2, 4 (December 2012), 402–408. DOI:http://dx.doi.org/10.1049/iet-wss.2012.0062

[19] Itron 2014. *Itron's Fixed-Network collector, CCU 100* (1 ed.). Itron, Washington, USA. https://www.itron.com/PublishedContent/CCU%20100.pdf

[20] Itron 2015. Itron, a world-leading technology and services company dedicated to the resourceful use of energy and water. https://www.itron.com. (2015).

[21] Liyan Jia, Jinsub Kim, R.J. Thomas, and Lang Tong. 2014. Impact of Data Quality on Real-Time Locational Marginal Price. *Power Systems, IEEE Transactions on* 29, 2 (March 2014), 627–636. DOI:http://dx.doi.org/10.1109/TPWRS.2013.2286992

[22] Stephen Kaisler, Frank Armour, J Alberto Espinosa, and William Money. 2013. Big data: Issues and challenges moving forward. In *System Sciences (HICSS), 46th International Conference on.* IEEE, Hawaii, 995–1004.

[23] Mehmed Kantardzic. 2002. *Data Mining: Concepts, Models, Methods and Algorithms.* John Wiley & Sons, Inc., New York, NY, USA.

[24] Maurice George Kendall. 1948. *Rank correlation methods.* Griffin.

[25] T. Khalifa, K. Naik, and A. Nayak. 2011. A Survey of Communication Protocols for Automatic Meter Reading Applications. *Communications Surveys Tutorials, IEEE* 13, 2 (Second 2011), 168–182. DOI:http://dx.doi.org/10.1109/SURV.2011.041110.00058

[26] Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized Distances Between Rankings. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10).* ACM, New York, NY, USA, Article 978-1-60558-799-8, 10 pages. DOI:http://dx.doi.org/10.1145/1772690.1772749

[27] David J Leeds. 2009. The smart grid in 2010: market segments, applications and industry players. *GTM Research, July* (2009).

[28] J. Machell, S. R. Mounce, and J. B. Boxall. 2010. Online modelling of water distribution systems: a UK case study. *Drinking Water Engineering and Science* 3, 1 (2010), 21–27. DOI:http://dx.doi.org/10.5194/dwes-3-21-2010

[29] Anas A. Makki, Rodney A. Stewart, Kriengsak Panuwatwanich, and Cara Beal. 2013. Revealing the determinants of shower water end use consumption: enabling better targeted urban water conservation strategies. *Journal of Cleaner Production* 60, 0 (2013), 129 – 146. DOI:http://dx.doi.org/10.1016/j.jclepro.2011.08.007 Special Volume: Water, Women, Waste, Wisdom and Wealth.

[30] Massimo Melucci. 2009. Weighted Rank Correlation in Information Retrieval Evaluation. In *Information Retrieval Technology*, Gary-Geunbae Lee, Dawei Song, Chin-Yew Lin, Akiko Aizawa, Kazuko Kuriyama, Masaharu Yoshioka, and Tetsuya Sakai (Eds.). Lecture Notes in Computer Science, Vol. 5839. Springer, Berlin, Heidelberg, 75–86. DOI:http://dx.doi.org/10.1007/978-3-642-04769-5_7

[31] P Mukheibir, Rodney A. Stewart, D Giurco, and Kelvin O'Halloran. 2012. *Understanding non-registration in domestic water meters: Implications for meter replacement strategies.* Australian Water Association.

[32] Vasyl Pihur, Susmita Datta, and Somnath Datta. 2009. RankAggreg, an R package for weighted rank aggregation. *BMC bioinformatics* 10, 1 (2009), 62.

[33] F.L. Quilumba, Wei-Jen Lee, Heng Huang, D.Y. Wang, and R. Szabados. 2014. An overview of AMI data preprocessing to enhance the performance of load forecasting. In *Industry Applications Society Annual Meeting.* IEEE, Vancouver, Canada, 1–7. DOI:http://dx.doi.org/10.1109/IAS.2014.6978369

[34] Michael G Schimek, Alena Myšičková, and Eva Budinská. 2012. An inference and integration approach for the consolidation of ranked lists. *Communications in Statistics-Simulation and Computation* 41, 7 (2012), 1152–1166.

[35] Grace S. Shieh. 1998. A weighted Kendall's tau statistic. *Statistics & Probability Letters* 39, 1 (1998), 17 – 24. DOI:http://dx.doi.org/10.1016/S0167-7152(98)00006-6

[36] Juan Shishido. 2012. Smart meter data quality insights. *ACEEE Summer Study on Energy Efficiency in Buildings* 12 (2012), 277–288.

[37] Rodney A Stewart, Rachelle Willis, Damien Giurco, Kriengsak Panuwatwanich, and Guillermo Capati. 2010. Web-based knowledge management system: linking smart metering to the future of urban water planning. *Australian Planner* 47, 2 (2010), 66–74.

[38] Ltd Technology Partners Co. 2013. *Next Generation Smart Meters and AMI Communications.* Virginia, USA.

[39] Shivanita Umapathi, Meng Nan Chong, and Ashok K. Sharma. 2013. Evaluation of plumbed rainwater tanks in households for sustainable water resource management: a real-time monitoring study. *Journal of Cleaner Production* 42, 0 (2013), 204 – 214. DOI:http://dx.doi.org/10.1016/j.jclepro.2012.11.006