Water Supply

IWA PUBLISHING

Water Supply Vol 22 No 1, 1131 doi: 10.2166/ws.2021.210

A use case of anomaly detection for identifying unusual water consumption in Jordan

Samer Nofal *, Abdullah Alfarrarjeh and Amani Abu Jabal
Department of Computer Science, German Jordanian University, P.O. Box 35247, Amman 11180, Jordan
*Corresponding author. E-mail: samer.nofal@gju.edu.jo

SN, 0000-0002-0216-934X

ABSTRACT

We present a use case of anomaly detection for identifying the unusual water consumption of consumers. Unusual water consumption may be due to a faulty water meter, fraudulent tampering with a water meter, or a leak in the water pipes within the consumer's property. We apply several anomaly detection methods to a real dataset of 22,877 mechanical water meters located in Amman, the capital city of Jordan. The dataset is unlabeled such that no discrimination is given for any meter whether it records a normal water consumption or not. The objective of this study is to test the hypothesis that abnormal water consumption (registered by a given water meter) can be identified based on previous records of water consumption measured by the same meter. We tested our hypothesis using well-known anomaly detection methods, namely: z-score (ZS), local outlier factor (LOF), density-based spatial clustering of applications with noise (DBSCAN), minimum covariance determinant (MCD), one-class support vector machine (OCSVM), and isolation forest (*i*FOrest). In the settings of our experiments, we observed that ZS, LOF, OCSVM and *i*FOrest support our hypothesis, contrasting with DBSCAN and MCD.

Key words: anomaly detection, water consumption, water supply network

HIGHLIGHTS

- We give a framework for detecting abnormal water consumption.
- Our dataset is drawn from the Jordan water supply network.
- Our methodology is based on unsupervised machine learning.

INTRODUCTION

Ninety-eight percent of Jordan's people live in water-scarce areas (World Data Lab 2020). Jordan faces extremely high water stress (Hofste *et al.* 2019). With huge efforts being put by the Jordan government and with assistance from the international community, the majority of Jordan people have access to improved-quality water despite the water scarcity. As a consequence of water scarcity, the water supply in Jordan is intermittent so that people have to store water in tanks on the roof of their houses. Although water service is subsidized greatly by the Jordan government, water service is not entirely free. In Jordan, water consumers are charged according to the consumed amount of water. Thus, mechanical water meters are installed within the consumer's property such that water bills are issued four times a year according to the readings of the meters.

Water meters are subject to failures due to natural reasons such as aging. Likewise, accidents, because of maintenance work, for example, may result in damage to water meters. Totally damaged water meters due to accidents might be reported by consumers as it might cause water service to be disconnected from their property. Major damage might also be discovered by the water company inspector upon the regular visits for registering water meter readings. In Jordan, such regular inspections occur four times a year. Conversely, discovering malfunctioning water meters, for natural reasons, for instance, might not be that easy. A malfunctioning water meter may look sound but it registers water consumption incorrectly: either higher or lower than the actual consumption. Unless examined by an expert, visual inspection of malfunctioning meters might not reveal any issues. If the registered water consumption is higher than usual, consumers might complain about that and so the water company may act accordingly to fix the defective meters. If the registered water consumption is lower than usual, consumers might not notify the company about that. Additionally, abnormal water consumption may be due to

tampering with water meters to avoid paying the bill of the actual water consumption. This is because water tariffs in Jordan are differentiated by the volume of consumption. For a certain level of consumption, the water tariff is nominal. If consumption goes beyond that level, the water tariff gets higher accordingly. In some other cases, the readings of the water meters might be unusual due to water leaks caused by faulty terminal pipes within the consumer's property. Such cases might be hard to identify effectively.

In this paper, we present a use case of anomaly detection for identifying the unusual water consumption of consumers. We apply several anomaly detection methods to a real dataset of 22,877 mechanical water meters located in Amman, the capital city of Jordan. The dataset is unlabeled such that no discrimination is given for any meter whether it records a normal water consumption or not. The objective of this study was to test the hypothesis that abnormal water consumption (registered by a given water meter) can be identified based on previous records of water consumption measured by the same meter. We tested our hypothesis using well-known anomaly detection methods, namely: z-score (ZS), local outlier factor (LOF), density-based spatial clustering of applications with noise (DBSCAN), minimum covariance determinant (MCD), one-class support vector machine (OCSVM), and isolation forest (*i*Forest). In the settings of our experiments, we observed that ZS, LOF, OCSVM and *i*Forest supported our hypothesis contrasting with DBSCAN and MCD. As mentioned previously, unusual water consumption may be due to a faulty water meter, fraudulent tampering with a water meter, or a leak in the water pipes within the consumer's property. Such causes need to be discovered and fixed by the operating water utility company. Exhaustive, human-powered inspections are highly costly. Reducing such costs would be of great benefit for a water utility company. The results of our study, reported in this article, suggest a framework for detecting abnormal water consumption. This framework can be viewed as a decision-support mechanism that helps human operators of a water utility company focus inspections on the detected water meters that register abnormal water consumption.

In the published literature, we found some studies that made use of water meter data to build classification models for discovering malfunctioning water meters or for detecting fraudulent tampering with water meters. Humaid (2012) presented a fraud detection model for water consumption in the Gaza Strip. Using an originally imbalanced training dataset, Humaid implemented three classification models: support vector machine (SVM), neural network, and k-nearest neighbors (KNN). Humaid's three classification models performed better when the training dataset was balanced. Similar to Humaid's study, Al-Radaideh & Al-Zoubi (2018) built two classification models (SVM and KNN) for detecting fraudulent tampering with water meters in Irbid, a major city in Jordan. Although Al-Radaideh and Al-Zoubi used a different dataset, they reached an accuracy comparable to the accuracy of Humaid's models. Differently, Monedero *et al.* (2015) presented three algorithms for detecting tampering with water meters in Spain. Their algorithms were generated after a careful examination of a real dataset and so their algorithms detect three types of consumption patterns: progressive drops, sudden drops, and abnormally low water consumption. Using a similar methodology to the work of Monedero *et al.*, Roberts & Monks (2015) designed an algorithm for detecting faulty water meters in Australia. Their algorithm was constructed after reviewing a dataset of real water meters. More recently, Roccetti *et al.* (2019, 2021) built a deep neural network, based on a carefully filtered real dataset, for detecting water meters failures in Italy. To summarize, related work on water meters tackled two different (but related) problems: the problem of detecting fraudulent tampering with water meters and the problem of discovering malfunctioning water meters. The existing studies (discussed above) use different real datasets and they build either semantic-based algorithms (Monedero *et al.* 2015; Roberts & Monks 2015) or supervised-machine-learning-based classification models (Humaid 2012; Al-Radaideh & Al-Zoubi 2018; Roccetti *et al.* 2019, 2021).

We emphasize that our study reported in this paper is different from related work. Our dataset is original and has not been used in any related work. The dataset is real such that it includes water utility bills for five years for 22,877 consumers living in the capital city of Jordan, Amman. The dataset was given exclusively to us by Miyahuna, the Jordan water utility company operating in Amman. Additionally, we apply numerous *anomaly detection* techniques for revealing abnormal water consumption. The anomaly detection techniques (employed for our study) have not been used in the literature discussed above.

METHODOLOGY

In this section, we describe the original dataset, how it was acquired and subsequently how it was prepared. Likewise, we give the necessary background on the data processing methods used in our study. Lastly, we conclude this section by presenting the settings of our experiments.

Data acquisition

In Jordan, water consumers (i.e. customers) pay for the water service according to the amount consumed, and so a mechanical water meter is installed within the consumer's property to measure the water consumption. The dataset used in this study was obtained from Miyahuna, the water utility company operating in Amman, the capital city of Jordan. The obtained dataset represents the records of the billing system of Miyahuna for the years 2015–2019 covering four areas in the city of Amman. Note that Miyahuna issues four water bills annually based on consumers' water consumption recorded by water meters. Thus, the water bills of consumers hold useful data for our study. After several interviews with Miyahuna's employees, we characterized the type of data contained in the billing system of Miyahuna. Table 1 summarizes the attributes of water bill records. Note that our dataset is a collection of water customer records where every record contains 20 water bills from five years for one customer. The original dataset includes 34,865 customers located in four areas of Amman. In the next subsection, we elaborate on the dataset preparation.

Data preparation

For our experimental study, we used only water consumption values (attribute 13 in Table 1). The attributes 1–8 and 10 & 14–15 are obviously not useful for our study, and so, we did not include these attributes in our experiments. For the other attributes, now we give a justification why we do not see them as helpful for testing our hypothesis. For subscription type (attribute 9 in Table 1), the purpose of this attribute is to distinguish the subscription for selecting the water tariff, which is different according to the subscription type. For attributes 11 and 12 in Table 1, by subtracting the previous water meter reading from the current water meter reading we get the water consumption amount (attribute 13 in Table 1). For reading status (attribute 16 in Table 1), it has three possible values: actual reading, estimated reading, or defective meter. These statuses respectively indicate whether the meter reading was taken by an inspector, the meter reading was estimated (without referring to any reason), or the meter reading was estimated because the respective water meter was obviously defective. Miyahuna did not give us any specific procedures for the estimation process or any further reasons behind such estimations. Although the original dataset has occasionally reported on defective meters, many cases of defective meters might not be discovered in reality. As affirmed by Miyahuna, an apparently sound water meter might be defective actually. This is because some kinds of defectiveness are hard to spot without a thorough examination by an expert technician. For 'tampering detected' (attribute 17 of Table 1), we noticed that the original dataset includes little information on fraudulent tampering

Table 1 | The attributes of a water bill record

No.	Attribute Name	Description
1	Customer number	To identify customers
2	Address of customer	Area name
3	Area code	To identify the area of the customer
4	Bill number	To identify the bill
5	Meter number	To identify the water meter
6	Payment number	For electronic payment
7	Inspector number	To identify the employee issuing the bill
8	Device number	To identify the device issuing the bill
9	Subscription type	Domestic or commercial
10	Reading date	The date of reading the meter
11	Current reading	Current reading of the meter in cubic meters
12	Previous reading	Previous reading of the meter in cubic meters
13	Water consumption	Consumed amount in cubic meters
14	Due payment	For the consumed water in the last quarter
15	Previous bills	Overdue payments
16	Reading status	Actual or estimated reading, or defective meter
17	Tampering detected	Tampering with meter (true or false)

with water meters. In fact, the original dataset include 596 fraud cases out of the whole dataset that has 34,865 customers. Many cases of fraudulent tampering with water meters might not be discovered. This is because many of the fraud cases, included in the original dataset, were found by chance as confirmed by Miyahuna. Additionally, the original dataset did not indicate any further details about the reported fraud cases, such as describing the discovered fraudulent activities and the consequent actions taken to ensure that these activities have been ended.

Recall, the original dataset (provided to us by Miyahuna) included 34,865 water meter records (i.e. customers) in four areas of Amman, where each record included 20 water bills for the years 2015–2019. The original dataset contained many missing water bills. There are 11,988 customers with one or more missing water bills. We removed every water meter record that had a missing water bill. Regardless of the reason behind those records with missing bills, we believe that we do not need them given the objectives of our study. Recall, our hypothesis under question is that abnormal water consumption (registered by a given water meter) can be identified based on previous records of water consumption measured by the same meter. Thus, including water meter records with missing bills will not help to investigate our hypothesis. In summary, our dataset (used in our experiments) includes 20 water consumption values recorded from the year 2015 to the year 2019 for 22,877 customers located in four different areas in Amman. As mentioned earlier, the company (Miyahuna) annually issues four bills based on water consumption recorded by water meters. Thus, we stress again, our dataset is a table of 22,877 water meter records where each record includes 20 water consumption values (recorded over five years). Every water consumption value of our dataset represents water consumption for three months.

Data processing methods

In this subsection, we review, in general, the selected methods for processing our dataset. Specifically, we will overview the z-score (ZS), local outlier factor (LOF), minimum covariance determinant (MCD), density-based spatial clustering of applications with noise (DBSCAN), one-class support vector machine (OCSVM), and isolation forest (*i*Forest). For an in-depth presentation of these methods, the reader may consult the cited references for each method.

Z-score

The z-score (ZS) is a standard statistical measure. Let μ be the arithmetic mean of a set, L_n, of n values, and σ be the standard deviation of L_n. Then, the ZS of a value $x \in L_n$, denoted by $z(x)$, is defined as $z(x) = \dfrac{x-\mu}{\sigma}$. In other words, $z(x)$ measures how far x is from the mean μ. The ZS test is a parametric method where it is used with the assumption that the tested dataset follows a normal distribution. Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about the mean, showing that data close to the mean happen more frequently than data distant from the mean.

Minimum covariance determinant

The essence of the minimum covariance matrix (MCD) (Rousseeuw & Driessen 1999; Hubert *et al.* 2018) is to discover k p-dimensional points (out of a set of n p-dimensional points) whose covariance matrix has the minimum determinant. The intuition behind MCD is that the arithmetic mean and the covariance matrix are computed hopefully using normal points only. Let x denote a p-dimensional point. Then, MCD measures an anomaly score of x by calculating a statistical distance equal to

$$\sqrt{(x-\bar{x})^T \Sigma^{-1} (x-\bar{x})}$$

using the covariance matrix, Σ, and the arithmetic mean, $\bar{x}$, of the discovered k points.

Local outlier factor

Local outlier factor (LOF) (Breunig *et al.* 2000) is actually an anomaly score that measures how isolated is a data point from its surrounding neighborhood. To see how LOF is computed we recall some definitions from (Breunig *et al.* 2000). Let k be a positive integer, and D be a dataset. The *k-distance* of an object $p \in D$, denoted by k-distance (p), is defined as the distance $d(p, o)$ between p and an object $o \in D$ such that

- for at least k objects $q \in D\backslash\{p\}$ it holds that $d(p, q) \leq d(p, o)$, and
- for at most $k-1$ objects $q \in D\backslash\{p\}$ it holds that $d(p, q) < d(p, o)$.

Given the k-distance of p, the *k-nearest neighbors* of p, denoted by $N_k(p)$, is defined as $N_k(p) = \{q \in D \backslash \{p\} | d(p, q) \leq k\text{-distance}(p)\}$. The *reachability distance* of an object $p \in D$ with respect to an object $o \in D$ is defined as

$$rd_k(p, o) = \max\{k\text{-distance}(o), d(p, o)\}.$$

The *local reachability density* of an object $p \in D$ is defined as

$$lrd_\alpha(p) = \frac{|N_\alpha(p)|}{\sum_{o \in N_\alpha(p)} rd_\alpha(p, o)}$$

where α is a positive integer parameter decided by the user of the algorithm. The *local outlier factor* of an object $p \in D$ is defined as

$$lof_\alpha(p) = \frac{\sum_{o \in N_\alpha(p)} \frac{lrd_\alpha(o)}{lrd_\alpha(p)}}{|N_\alpha(p)|}$$

Density-based spatial clustering of applications with noise

Density-based spatial clustering of applications with noise (DBSCAN) identifies clusters in a given dataset based on a distance function defined by the user of the algorithm (Ester *et al.* 1996). To give the notion of 'clusters' and 'noise' (outliers) of DBSCAN, we recall some definitions from (Ester *et al.* 1996). Let ϵ and φ be positive integers, and D be a set of d-dimensional points. The *ϵ-neighborhood* of a point $p \in D$, denoted by $N_\epsilon(p)$, is defined by $N_\epsilon(p) = \{q \in D | d(p, q) \leq \epsilon\}$ where $d(p, q)$ denotes a distance function between $p \in D$ and $q \in D$. Note that any appropriate distance function can be chosen for a given application. A point $p \in D$ is *directly density-reachable* from a point $q \in D$ with respect to ϵ and φ if $p \in N_\epsilon(q)$ with $|N_\epsilon(q)| \geq \varphi$. A point p is *density-reachable* from a point q with respect to ϵ and φ if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i. A point p is *density-connected* to a point q with respect to ϵ and φ if there is a point o such that both p and q are density-reachable from o with respect to ϵ and φ. A *cluster* C with respect to ϵ and φ is a non-empty subset of D satisfying the following conditions:

- $\forall p, \quad q \in D$: if $p \in C$ and q is density-reachable from p with respect to ϵ and φ, then $q \in C$.
- $\forall p, \quad q \in C$: p is density-connected to q with respect to ϵ and φ.

Let $C_1, \ldots, C_k$ be the clusters of D with respect to parameters ϵ_i and φ_i, where $i = 1, \ldots, k$. Then, the *noise* of D is defined by $noise = \{p \in D | \forall i : p \notin C_i\}$.

One-class support vector machine

One class support vector machine (OCSVM) (Schölkopf *et al.* 1999) is a well-known anomaly detection method. We recall the definition of OCSVM from (Schölkopf *et al.* 1999). Let $\mathcal{X}$ be some dataset. OCSVM estimates a decision function f which is positive on a subset of $\mathcal{X}$ and negative elsewhere. Let $x_1, x_2, \ldots, x_n \in \mathcal{X}$ be data points. Let Φ be a feature map $\mathcal{X} \to \mathcal{F}$, i.e. a map into a dot product space $\mathcal{F}$ such that the dot product in the image of Φ can be computed by evaluating some kernel function $k(x, y) = (\Phi(x) \cdot \Phi(y))$. Then, OCSVM estimates a decision function $f(x) = sign((w \cdot \Phi(x)) - \rho)$ by solving the following quadratic program

$$\min_{w \in \mathcal{F}, \rho \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n} \quad \frac{1}{2}\|w\|^2 + \frac{1}{vn}\sum_{i=1}^{n} \xi_i - \rho$$

$$\text{subject to} \quad w \cdot \Phi(x_i) \geq \rho - \xi_i, \xi_i \geq 0, \quad \text{for all } 1 \leq i \leq n$$

where $\boldsymbol{\xi} = \{\xi_1, \xi_2, \ldots, \xi_n\}$, and $v \in (0, 1]$ is an *a priori* specified parameter representing the probability that a point drawn from the probability distribution of $\mathcal{X}$ lies outside of $\{x_1, x_2, \ldots, x_n\}$.

Isolation forest

Isolation forest (*i*Forest) (Liu *et al.* 2012) is a well-known anomaly detection method. We recall a description of *i*Forest from (Liu *et al.* 2012). The term 'isolation' means separating an instance from the rest of the instances. Basically, *i*Forest builds random binary trees by partitioning a given set of instances recursively. Then, *i*Forest average path lengths over the generated trees to find the expected path length. Hence, anomaly instances are those that have path lengths substantially shorter than the expected path length. Let T be a node of an isolation tree. T is either an external node with no child, or an internal node with one test and exactly two daughter nodes (T_l, T_r). A test at node T consists of an attribute q and a split value p such that the test $q < p$ determines the traversal of an instance to either T_l or T_r. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a given dataset. A sample of ψ instances $\mathcal{X}' \subset \mathcal{X}$ is used to build an isolation tree (*i*Tree). Then, $\mathcal{X}'$ is recursively divided by randomly selecting an attribute q and a split value p, until either: (i) the node has only one instance or (ii) all instances at the node have the same values. An *i*Tree is a proper binary tree, where each node in the tree has exactly zero or two daughter nodes. Assuming all instances are distinct, each instance is isolated to an external node when an *i*Tree is fully expanded, in which case the number of external nodes is ψ and the number of internal nodes is $\psi - 1$. The task of anomaly detection is to provide a ranking that reflects the degree of anomaly. Using *i*Trees, instances are sorted according to their average path lengths; and anomalies are instances that are ranked at the top of the list. Path length of an instance x is measured by the number of edges x traverses an *i*Tree from the root node until the traversal is terminated at an external node.

Settings of experiments

We give a description of the framework of the anomaly detection techniques that we applied in our experimental study. Particularly, we present the settings of z-score (ZS), local outlier factor (LOF), minimum covariance determinant (MCD), density-based spatial clustering of applications with noise (DBSCAN), one-class support vector machine (OCSVM), and isolation forest (*i*Forest). We utilize the available implementation of these techniques from Python known libraries: pandas (version 1.0.5) and scikit-learn (version 0.23.1). We specify in the following subsections the methods of these libraries that implement the aforementioned anomaly detection techniques. For the parameters of the Python library's methods of anomaly detection, we use the default values unless it is necessary to do otherwise as we explain in the following subsections. In our experiments, we consider a water meter record anomalous if it has at least one water consumption value that is identified as abnormal by the applied anomaly detection technique. We conducted three experiments, denoted by E_5, E_{10} and E_{16}, with the assumption that respectively, 5, 10, and 16% of our water meters is anomalous. Assuming a certain level of an anomaly in the dataset is necessary for conducting our experiments to determine a threshold that will differentiate between normal and abnormal water consumption values as we elaborate in the coming subsections. Using a different anomaly percentage for each trial, we ran three trials (i.e. E_5 & E_{10} & E_{16}) to see if the performance of the applied methods changes as the anomaly percentage increases. We do not mean by our experiments to make any conclusion about our assumed anomaly percentages. Generally, one might choose a different set of anomaly percentages other than the selected ones for our experiments. Given the lack of studies in the literature in this regard, the right percentage of an anomaly in a water meters dataset remains an open issue. However, assuming an anomaly percentage not exceeding 16% is sensible for the objectives of our study, at least, from a practical point of view where the water company needs to take actions concerning the detected abnormal consumption values; thus revealing too many unusual consumption cases might not add further benefits to the company, which may still have a limited capacity for investigating the detected cases.

Settings of the Z-score

Using ZS, we consider a water meter record anomalous if it includes a water consumption with a ZS greater than a threshold t, where $t = 3.66$ for E_5, $t = 3.28$ for E_{10}, and $t = 3.0$ for E_{16}. To apply the ZS test to our dataset, we used the pandas library. More specifically, we applied the mean and std methods of the pandas DataFrame class. We applied the default parameters of the two methods with one exception: the std method has been invoked with the *delta degrees of freedom* (ddof) being set to zero, whereas the default value of ddof is 1. Observe, the divisor used in calculating the standard deviation is n – ddof, where $n = 20$ is the number of water consumption values of a given water meter. For a given water meter, we compute the mean water consumption using the whole set of 20 water consumption values registered by the water meter; and hence, it is sensible to use the value of 20 as the divisor employed in calculating the standard deviation of the water meter record.

Settings of minimum covariance determinant

We used the scikit-learn implementation of MCD. Specifically, we first constructed an instance, ee, of sklearn.covariance.EllipticEnvelope with the default parameters. Then, for each water meter record m, we executed ee.fit(m) (passing on m as an instance of pandas.DataFrame). Then we checked the ee.score_samples(m): we consider m anomalous if it has a water consumption with an anomaly score less than t, where $t = -212$ for E_5, $t = -90$ for E_{10}, and $t = -53$ for E_{16}. To avoid a run-time error (division-by-zero) of the ee.fit(m), we check every water meter record m: if m has more than 10 duplicates of water consumption, then we skip executing ee.fit(m) and subsequently we label m as *undecided* indicating that it cannot be decided whether m is anomalous or not.

Settings of local outlier factor

To utilize LOF, we used the implementation of LOF of the scikit-learn library. Particularly, we constructed an instance, lof, of sklearn.neighbors.LocalOutlierFactor with the parameters n_neighbors = 18 and novelty = True. These parameters are selected to achieve the assumed percentages of an anomaly in our three experiments E_5, E_{10}, and E_{16}. The default value of n_neighbors is 20, which is inapplicable for our dataset. Recall that each water meter record has 20 water consumption values. Thus, it is unreasonable to use all water consumption values of a water meter in computing the LOF scores of the water meter; otherwise, if we use n_neighbors = 20 or n_neighbors = 19, all water consumption values of every water meter will be normal. Setting novelty = True enables us to access the calculated anomaly scores of water consumption values as we explain now; for each water meter record m, we execute the method lof.fit(m) (passing on m as an instance of pandas.DataFrame). Then, we check the lof.score_samples(m): we consider m anomalous if m includes a water consumption with an anomaly score less than t, where $t = -2.29$ for E_5, $t = -1.73$ for E_{10}, $t = -1.47$ for E_{16}.

Settings of density-based spatial clustering of applications with noise

To apply DBSCAN to our dataset, we used the scikit-learn library. We created an instance, dbscan, of sklearn.cluster.DBSCAN with the parameter eps = 0.92 for E_{10} and eps = 0.66 for E_{16}. Note, eps is the maximum distance between two data points for one to be considered as in the neighborhood of the other. The other parameters of sklearn.cluster.DBSCAN are left with the default values. For E_5, the dbscan instance could not retrieve the 5-percent-anomaly. We tried eps with different values up to 0.99 but still we achieved much greater than 5% anomaly. Since the default behavior of dbscan uses a Euclidean distance function, we normalized our dataset using the method sklearn.preprocessing.StandardScaler.fit_transform. Using the normalized dataset, for each water meter record m, we applied the method dbscan.fit_predict(m) (passing on m as an instance of pandas.DataFrame). This method classifies each water consumption value in m as -1 (abnormal) or as 1 (normal).

Settings of one-class support vector machine

To apply OCSVM, we used the scikit-learn library. We constructed an instance, ocsvm, of sklearn.svm.OneClassSVM with the default parameters. Then, for every water meter record m, we invoked ocsvm.fit(m) (passing on m as an instance of pandas.DataFrame); afterward, we checked the ocsvm.score_samples(m): we consider m anomalous if it has a water consumption with an anomaly score less than or equal to t, where $t = 1.125$ for E_{16}, $t = 1.025$ for E_{10}, and $t = 1.0013$ for E_5.

Settings of isolation forest

To utilize *i*Forest, we used the scikit-learn implementation of *i*Forest. Using the default parameters, we created an instance, x, of sklearn.ensemble.IsolationForest. Then, for each water meter record m, we apply x.fit(m) (passing on m as an instance of pandas.DataFrame). Subsequently, we check the x.score_samples(m): we consider m anomalous if it includes a water consumption with an anomaly score less than or equal to t, where $t = -0.812$ for E_5, $t = -0.790$ for E_{10}, and $t = -0.772$ for E_{16}.

RESULTS AND DISCUSSION

Ground truth, regarding whether, in reality, a water meter record is anomalous or not, is not available within our dataset. Hence, we adopted a peer evaluation strategy such that for each conducted experiment we assumed that one of our applied anomaly detection methods is the reference point for assessing the efficiency of the other anomaly detection methods. That is, for each experiment we used the outcome of an anomaly detection method as a benchmark to evaluate the F_1 score of the other anomaly detection methods. As we mentioned earlier, we conducted three experiments (E_5 and E_{10} and E_{16}) assuming respectively that 5, 10, and 16% of the dataset was anomalous. Because of these assumed percentages, the *recall* of an applied anomaly detection method is expected to be close to the *precision* of the method. However, we report on the traditional F_1

score observed in all of our experiments. Note, the F_1 score was calculated assuming that the *positive* class is *anomalous*, and the *negative* class is *normal*. Tables 2–7 summarize the F_1 score under each reference point respectively: *i*Forest, LOF, OCSVM, ZS, MCD, and DBSCAN. The highest F_1 scores are shown in bold text in Tables 2–7. Regarding DBSCAN, recall that we could not retrieve the assumed anomaly of 5 percent, and so in Tables 2–6, the F_1 score for DBSCAN is left empty for E_5. Similarly, in Table 7, where DBSCAN is the reference point, there is no record for E_5.

Table 2 | The F_1 score with *i*forest being the reference point

Experiment	MCD	LOF	OCSVM	ZS	DBSCAN
E_5	0.40	**0.87**	0.83	0.82	–
E_{10}	0.46	**0.87**	0.86	0.82	0.43
E_{16}	0.51	**0.87**	0.86	0.81	0.45

Table 3 | The F_1 score with LOF being the reference point

Experiment	*i*FOREST	MCD	OCSVM	ZS	DBSCAN
E_5	0.87	0.38	**0.90**	0.87	–
E_{10}	0.87	0.48	**0.90**	0.89	0.44
E_{16}	0.87	0.52	**0.88**	0.87	0.47

Table 4 | The F_1 score with OCSVM being the reference point

Experiment	*i*FOREST	MCD	LOF	ZS	DBSCAN
E_5	0.83	0.37	**0.90**	0.85	–
E_{10}	0.86	0.45	**0.90**	0.86	0.44
E_{16}	0.86	0.49	**0.88**	0.84	0.47

Table 5 | The F_1 score with ZS being the reference point

Experiment	*i*FOREST	MCD	LOF	OCSVM	DBSCAN
E_5	0.81	0.42	**0.87**	0.85	–
E_{10}	0.82	0.53	**0.89**	0.86	0.45
E_{16}	0.81	0.59	**0.87**	0.84	0.49

Table 6 | The F_1 score with MCD being the reference point

Experiment	*i*FOREST	LOF	OCSVM	ZS	DBSCAN
E_5	0.40	0.38	0.37	**0.44**	–
E_{10}	0.46	0.48	0.45	**0.53**	0.38
E_{16}	0.50	0.50	0.49	**0.53**	0.47

Table 7 | The F_1 score with DBSCAN being the reference point

Experiment	*i*FOREST	LOF	OCSVM	ZS	MCD
E_{10}	0.43	0.44	0.44	**0.45**	0.38
E_{16}	0.45	0.47	0.47	**0.49**	0.44

Overall, our purpose of this study was to test our hypothesis that abnormal water consumption (registered by a given water meter) can be identified based on previous records of water consumption measured by the same meter. Having a closer look at the figures of Tables 2–7, we note that LOF, OCSVM, *i*Forest, and ZS supported our hypothesis because their F_1 scores were greater than 0.80 with respect to most of the reference points (except for DBSCAN and MCD). In contrast, DBSCAN and MCD seem to undermine our hypothesis as their F_1 scores were less than 0.60 with respect to all reference points. To the extent we checked, we did not see a definite explanation for the observed behavior of the DBSCAN and MCD. It might be due to the default parameters of the scikit-learn library; or, it might be due to the underlying actions of DBSCAN and MCD. We did not investigate this issue any further since we already reached sufficient support for our hypothesis. As a side observation, regarding the time efficiency, we observed that the (elapsed) running time of most of the applied anomaly detection methods was less than 5 minutes. The only exception is *i*Forest: for which more than one hour was required to finish processing our dataset.

CONCLUSIONS

Our experimental results reported in this article indicate that previous water consumption values (recorded by a water meter) are likely to be a key factor for discovering abnormal water consumption registered by the meter in the future. Nonetheless, for new water meters (i.e. new consumers), with no historical records of water consumption, abnormal water consumption might be discovered by some other means. This is to be explored by conducting further research. It is important to take our results with care. This is because our suggested framework for detecting abnormal water consumption depends on the extent to which a water meter record (in our dataset) is linked with the residents of a property rather than being connected with the property itself. In Jordan, water service (and so the respective water meter) is subscribed under the name of the landlord. If a property is leased frequently or is used by different people at different times, then it might be hard to determine whether a water consumption value is normal or not due to the fluctuations associated with the change of the residents of the property. It is sensible to consider that water consumption is linked to the property's residents: their number, age, lifestyle, and the nature of their job (i.e. whether they work from home or not, and for how long they are away home). Information about the actual residents of a property was not available in our dataset. The water utility company Miyahuna (the dataset provider) has basic information only about landlords (such as full name and identity number). Nevertheless, even in Jordan, our findings are still practical especially for the water meters that are used constantly by the same residents for several years. Detecting abnormal water consumption is a decision-support mechanism meant to help in shortlisting abnormal water consumption cases. Thus, using such a decision-support mechanism, human inspectors will be more productive in conducting on-site investigations, and if an abnormality is confirmed, they will act accordingly to resolve the causes by replacing a faulty water meter, fixing a water leak within the consumer's property, or stopping tampering with the water meter.

ACKNOWLEDGEMENTS

This study is supported by the deanship of scientific research at the German Jordanian University (project number SEEIT 05/2020). The authors thank Abedlrahman M. Kanan (from the Jordan company of water utility Miyahuna), Aya Aljaloudi, and Jack Abuzulof (from German Jordanian University) for the help in providing the dataset of this study.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Al-Radaideh, Q. A. & Al-Zoubi, M. M. 2018 A data mining based model for detection of fraudulent behaviour in water consumption. In; *2018 9th International Conference on Information and Communication Systems (ICICS)*. pp. 48–54.

Breunig, M. M., Kriegel, H.-P., Ng, R. T. & Sander, J. 2000 Lof: identifying density-based local outliers. *SIGMOD Rec.* **29** (2), 93104.

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. 1996 A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD* **96**, 226–231.

Hofste, R. W., Reig, P. & Schleifer, L. 2019 *17 Countries, Home to One-Quarter of the World's Population, Face Extremely High Water Stress*. Available from: https://wri.org/insights (accessed 3 May 2021).

Hubert, M., Debruyne, M. & Rousseeuw, P. J. 2018 Minimum covariance determinant and extensions. *WIREs Comput. Stat.* **10** (3), e1421.

Humaid, E. 2012 *A Data Mining Based Fraud Detection Model for Water Consumption Billing System in MOG*. Master's Thesis, Islamic University of Gaza, Gaza Strip.

Liu, F. T., Ting, K. M. & Zhou, Z.-H. 2012 Isolation-based anomaly detection. *ACM Trans. Knowl. Discovery Data (TKDD)* **6** (1), 1–39.

Monedero, I., Biscarri, F., Guerrero, J. I., Roldán, M. & León, C. 2015 An approach to detection of tampering in water meters. In: *19th International Conference in Knowledge Based and Intelligent Information and Engineering Systems* (Ding, L., Pang, C., Leong, M.-K., Jain, L. C. & Howlett, R. J., eds), Singapore, Elsevier.

Roberts, S. E. & Monks, I. R. 2015 Fault detection of non-residential water meters. In: *21st International Congress on Modelling and Simulation*. pp. 2228–2233.

Roccetti, M., Delnevo, G., Casini, L. & Cappiello, G. 2019 Is bigger always better? a controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures. *J. Big Data* **6**, 70.

Roccetti, M., Casini, L., Delnevo, G. & Bonfante, S. 2021 Dimensionality reduction and the strange case of categorical data for predicting defective water meter devices. In: *Human Interaction, Emerging Technologies and Future Applications III* (Ahram, T., Taiar, R., Langlois, K. & Choplin, A., eds). pp. 155–159, Paris, Springer.

Rousseeuw, P. J. & Van Driessen, K. 1999 A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** (3), 212–223.

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J. & Platt, J. 1999 Support vector method for novelty detection. In: *Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS'99*, Cambridge, MA, USA. MIT Press, pp. 582–588.

World Data Lab 2020 *Water Scarcity Clock*. Available from: https://www.worldwater.io (accessed 23 December 2020).

First received 9 April 2021; accepted in revised form 20 June 2021. Available online 2 July 2021