Assessment of Apparent Losses Due to Meter Inaccuracy using an Alternative, Validated Methodology

Mthokozisi Ncube[1, 2], **Akpofure E. Taigbenu**[2]

[1] Development Bank of Southern Africa, P O Box 1234, Halfway House, Midrand, 1685, South Africa. Email: mthokozisin@dbsa.org,

[2] School of Civil and Environmental Engineering, University of the Witwatersrand, Johannesburg, Private Bag 3, WITS, 2050, South Africa. Email: akpofure.taigbenu@wits.ac.za

Abstract

Despite wide acceptance of the IWA water balance as the basis of managing water losses, experience suggests that there are difficulties with its application. This is in part due to the required effort in assessing its various components, of which apparent water loss has been one of the most difficult to estimate. The traditional approach of deriving consumption profiles through field work and testing water meters in a laboratory exceeds the resources of many utilities. While a few studies have explored alternative methodologies, these have largely not been validated and are susceptible to reproducibility and interpretation difficulties.

This paper introduces an improved comparative billing analysis method that combines data preparation techniques, clustering analysis and classical regression analysis on monthly billing data of a water utility in Johannesburg, South Africa. Using the method, an average estimate of apparent losses due to metering errors of 8.2% was found against the best-case scenario of 9.4% using field investigations and laboratory tests, which also measure meter under-registration that the proposed methodology does not cater for. The validated results were possible at a fraction of the cost and effort, while also providing better insight of the underlying consumption patterns. The results show that data-driven discovery processes are viable alternatives for the improved assessment and management of water losses.

Keywords
Apparent losses, data mining, metering errors, water loss management, non-revenue water (NRW)

INTRODUCTION

Water Loss Management

Water loss control remains a major challenge in the sustainability of water utilities and the promotion of efficient use of water as a finite natural resource (Loureiro *et al.*, 2014), and regarded as one of the top ten global risks (World Economic Forum, 2017). Reducing municipal water losses is therefore a key opportunity that can unlock significant resource and financial benefits (Green Cape, 2017). In South Africa, recent estimates (2015/16), using the standard International Water Association (IWA) water balance, put non-revenue water (NRW) at 41%, an increase from 34.6% in 2013/14. The increase in water losses within South Africa was not due to system degradation, but is attributed to improved quantification of water losses in rural municipalities (Department of Water and Sanitation, 2017). Yet, the NRW estimate for South Africa is only from 107 out of a total of 152 water services authorities. This demonstrates that the ability to quantify, and therefore manage, water losses remains a challenge. This is not unique to the developing world. Despite the wide acceptance of the IWA water balance, a review of its application shows that very few utilities produce complete water balances owing to the effort involved and the lack of adequate and efficient methods for determining the various components (Klingel and Knobloch, 2015). Developing new methods and streamlining existing methods for enhanced water loss management stands to contribute towards better quantification and management of water losses and its two main components of real and apparent losses. In particular, the apparent water loss component is largely based on guideline percentage values, such as in Seago et al. (2004) and McKenzie et al. (2012), and its estimation remains at infancy. Without the proper quantification of apparent losses, or any of the other water loss components, any implementation of water loss interventions may not be comprehensive and adequately informed and might therefore not necessarily be optimum.

While a number of authors (Arregui, Cabrera Jr. and Cobacho, 2006; AWWA, 2009; Criminisi *et al.*, 2009; Mutikanga, Sharma and Vairavamoorthy, 2011; Ncube and Taigbenu, 2018) have used the extensive traditional empirical field- and laboratory-based method, it has proved to be very resource intensive and out of reach for many utilities. Alternative approaches to bridge this gap have been suggested, including the use of a meter replacement database (Couvelis and van Zyl, 2015) and the use of a comparative analysis of a billing database (Mbabazi et al., 2015). There has, however, been limited published application of such methods and the few in literature have not been validated and packaged for a typical utility user. Ncube & Taigbenu (2017) demonstrated that while these methods show great promise, particularly the comparative analysis of billing database, further refinements that make such methods efficient, reproducible and accurate are required to make them more relevant and usable. The use of data mining, as applied in various sectors, is one promising alternative for water loss management.

Data Mining

The use of smart metering infrastructure, databases, and information systems has provided an opportunity to apply data mining and computational intelligence in the analysis of water and electricity consumption (Monedero *et al.*, 2016). Despite the increasing application of smart water metering, it is undeniable that conventional metering will remain a reality for many water utilities, particularly in the developing world. It therefore remains relevant to evaluate the applicability of data mining and computational intelligence on existing consumption databases to fulfil the need for methods and tools to support water utilities in establishing complete water balances (Klingel and Knobloch, 2015) with fewer guestimates on various components.

Data mining is a process of discovering valuable information such as patterns and non-trivial extraction of implicit information from large amounts of data (Yin *et al.*, 2011) using computational techniques, machine learning, artificial intelligence and pattern recognition (Gorunescu, 2011). This has been applied in sectors such as financial services, health care services, supply chain management, telecommunications (Gorunescu, 2011), customer relations management (Ngai, Xiu and Chau, 2009), electricity (De Silva *et al.*, 2011), water resources (Solomatine, 2003; Abrahart, See and Solomatine, 2008; Kalteh, Hjorth and Berndtsson, 2008; Dhanya and Nagesh Kumar, 2009), water asset management (Babovic *et al.*, 2002) and water consumption and metering related applications (Humaid and Barhoum, 2013; Monedero *et al.*, 2016).

Applications of data mining within the water and electricity consumption and metering related applications include fraud detection (Chauhan and Rajvanshi, 2013; Humaid and Barhoum, 2013) and time series clustering and classification (Aghabozorgi, Shirkhorshidi and Wah, 2015). Examples of such applications include Nguyen, Stewart & Zhang, (2013) who used a hybrid combination of the Hidden Markov Model (HMM), Dynamic Time Warping (DTW) and gradient vector filtering for pattern recognition of smart metering flow data into predefined water end use events; Räsänen, Ruuskanen & Kolehmainen (2008) used Self-Organising Maps (SOM) together with k-means clustering and fuzzy logic to create personalised information to consumers. There are no known applications of data mining techniques for the quantification of apparent losses. The water loss management future can therefore benefit from combining the traditional problem solving and the use of theory-driven, understanding-rich processes with the emerging data-driven discovery processes (Babovic, 2005).

This paper presents findings of research undertaken to develop a data-driven assessment of apparent water losses using monthly water consumption data and successfully validated against the traditional assessment methodology, with clear recommendations for application.

METHODOLOGY

Ncube & Taigbenu (2017) had previously applied the methodology of Mbabazi et al. (2015) to determine water meter accuracy degradation using the dataset from Johannesburg Water (JW), the largest water utility in South Africa. On preliminary validation, the method performed better than another alternative method evaluated, but not adequate to match the gold standard of apparent loss measurement based on field and laboratory measurements. Some of the gaps included:

- The lack of clarity in the selection of the final dataset used and other various sub-steps of data cleansing and analysis. This introduced a lack of reproducibility in methodology application on any dataset leading to subjectivity and verification challenges.
- The failure to maximise the utilisation of the consumption dataset, with only 0.4% of the available data used and a targeted twelve months of consumption records usage. This therefore did not consider the full water consumption time series.
- While the analysis catered for different meter models and types, the requirement is that they be labelled accordingly, which is a challenge in many utilities. Coupled with the minimal use of available data, the use of small sample size of meters per model introduced uncertainties.
- There is omission on how the increasing consumption over time, as found in Ncube & Taigbenu (2015), is handled in the analysis, with only reference to declining consumption.

To minimise and counteract these deficiencies, the methodology comprising data preparation followed by clustering analysis and lastly accuracy degradation analysis was developed and applied on JW's monthly reading records spanning from July 2003 to June 2015. The records were received as 144 monthly flat files of meter reading data submitted for billing purposes. The number and actual meters read within each individual month varied from month to month. The files were in two different formats owing to a billing system change in June 2010, and all personal data were stripped from the dataset to focus only on the meter ID, the property ID, consumer category and the meter reading in any particular month.

Data Preparation

Each monthly reading file was processed with Microsoft SQL Server 2016 and SQL Server Integration Services (SSIS) for data extraction, transformation, and loading into a new database. The database was subsequently cleaned using the SQL Data Quality Client to combine and/or remove duplicate meter numbers and property IDs, resulting in slightly over 600,000 meter reading records. Because the dataset had several irregularities that precipitated a billing crisis widely reported in the local media, stringent criteria had to be introduced to minimise data cleansing time and increase accuracy. After a couple of trial runs, only records which met the following preliminary criteria were retained, in order of priority;

- a sufficiently long record of at least 60 individual readings i.e. 5 years of data;

- Less than 20% patching of missing data required anyway within the dataset, i.e. less than 1 year of data was missing;
- Meters that did not clock-over (i.e. did not start again from the beginning) and whose minimum/maximum meter reading was at the beginning/end of the record;
- Records without constant readings, such as zero consumption or entirely negative consumption, and records without abnormally high readings.

The dataset was reduced to 185,000 records (31%) and statistical analysis was done to categorise each record according to its average consumption quantiles and other variables such as the initial meter reading on record. The data was thereafter processed and patched using the “zoo package” (Zeileis and Grothendieck, 2005) within the R Statistical Software (R Core Team, 2017).

An important variation to the original analysis was the use of running annual total consumptions to ensure that the time series property of consumption is retained at a scale that is not susceptible to monthly and seasonal variations. The running annual consumption for all records was generated and used for subsequent analyses.

A visual first-order assessment was performed on the dataset by plotting the average running annual consumption for the first 100 months as shown in Figure 1. The plot clearly indicates unusual variations from month 60 which was about the time when the billing system was changed.

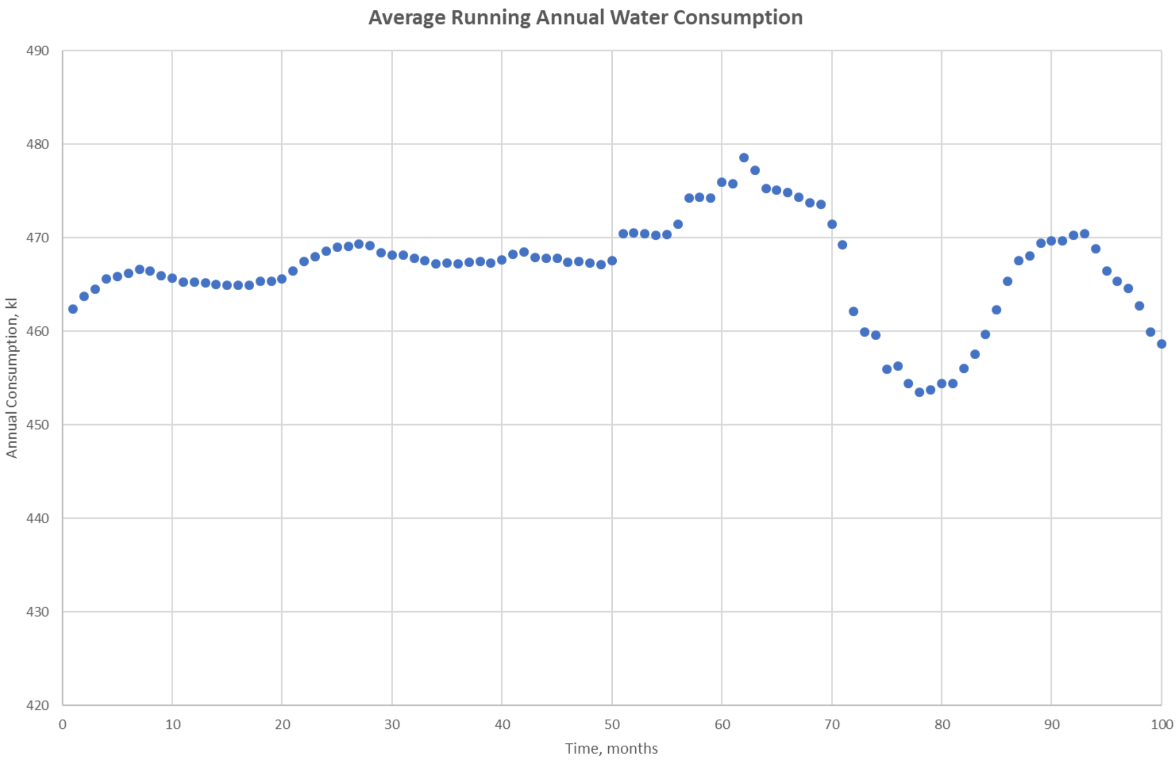

Figure 1: Average Running Annual Consumption

The anomalies of Figure 1 were not conducive for method development and therefore all post-2009 data were therefore excluded from further analysis and the following additional criteria were imposed in the data cleansing algorithm:

- Individual meters had to have at least 60 months prior to 2009;
- Individual meters had to have less than 10% of the annual running consumption readings falling outside of the statistical interquartile range to ensure outliers are removed.

The outcome of application of the above data cleansing criteria resulted in a sample of 87,589 (about 15%) records. This sample size was considered adequate for method development as it is not only a significant population of active consumers within the JW conventionally metered consumers (about 375,000 consumers) but is also comparable to the number of consumers in many small to medium sized utilities.

There was an expected higher proportion of meters with low meter readings at the start of the individual records. There was also a sizeable proportion of old meters with high initial meter readings which was expected as the starting point of the database was different from the installation date of the meters. This necessitated bringing all the records to a common arbitrary start date for the clustering analysis, and factoring the initial meter reading of the individual records for the degradation analysis. The record of the start readings for each meter was retained in a separate dataset to enable the comparison of the consumption profiles with similar start readings at a later stage.

Clustering Analysis

A critical component of apparent water loss analysis is being able to derive errors for the different consumption categories and the different devices in use. From the authors' experiences, many utilities have inadequate information relating to meter-type characteristics and the consumer categories. As such, unsupervised clustering is well suited to deriving distinct groupings of consumption profiles without requiring prior classification. Clustering is a process of grouping data items based on a measure of similarity and is useful in data mining for database segmentation, predictive modelling, and visualization of large databases (Jain, Murty and Flynn, 1999). Time-series clustering is a special type which is of interest due to its ubiquity in various areas and sectors (Aghabozorgi, Shirkhorshidi and Wah, 2015).

The "dtwclust package" of the R Statistical Software (R Core Team, 2017) provides a common platform on which classical and new clustering algorithms can be evaluated and compared against each other (Sard, 2015). This package was selected because of its convenience and range of supported implementations. However, only partitional clustering algorithms were evaluated in this study due to the time and memory complexity of O(N^2) of hierarchal clustering, where N is the total number of records in the dataset. While several algorithms were tested, k-Shape and Fuzzy clustering were finally selected for implementation due to their relative speed and ease of implementation.

k-Shape Clustering

k-Shape is partitional clustering algorithm created by Paparrizos & Gravano (2015) consisting of a custom centroid function (shape extraction) and a custom distance measure (shape-based distance, SBD). The function is stochastic in nature and requires z-normalization of the input data. SBD is based on the cross-correlation with coefficient normalization (NCCc) sequence between two series, and it is thus sensitive to scale, hence the z-normalisation requirement. SBD is given by the formula (Sard, 2015);

$$SBD(x,y) = 1 - \frac{\max(NCCc(x,y))}{||x||_2 ||y||_2} \qquad \textbf{\textit{Equation 1}}$$

where $||\cdot||_2$ is the Euclidean norm of the series. Shape extraction relies on NCCc and uses it to match any two series optimally with the centroid series selected in random. Alignment can be done between series with different lengths, but the length of the resulting prototype will depend on the length of the chosen reference. This was not seen as a handicap as the objective of clustering in this case was to identify similar groupings of time series only.

Fuzzy Clustering

Fuzzy clustering does not create “crisp” clusters but rather fuzzy or soft partitions in which each member belongs to each cluster to a certain degree. This is a particularly useful property in water consumption analysis as it is likely that a consumer may belong to a few clusters at the same time. The Fuzzy clustering implementation in the “*dtwclust package*” utilises fuzzy c-means described in Bezdek (1981) by default and uses the Euclidean distance as a distance measure. Defining $\mu_{c,i}$ as the *i*-th element of the *c*-th centroid, and $x_{p,i}$ as the *i*-th data-point of the *p*-th object in the data, the centroid function is expressed as (Sard, 2015);

$$\mu_{c,i} = \frac{\sum_{p=1}^{N} u_{p,c}^{m} x_{p,i}}{\sum_{p=1}^{N} u_{p,c}^{m}} \qquad \textbf{\textit{Equation 2}}$$

From Equation 2, it is clear that all the time series data are required to have the same dimensionality, which had to be addressed for the current dataset through re-interpolation of the series to match the longest time-series. The re-interpolation presupposes the underlying trend remains the same and extrapolates the time-series using its statistical properties. The non-crisp partitions of the algorithms can also be made crisp by taking the maxima of each cluster.

A non-trivial aspect of time-series clustering is the determination of the number of clusters, *k*, and evaluating the clustering algorithm performance. To this end, the best performing cluster validity indices (CVI) identified in Arbelaitz et al. (2013) were used to evaluate the output of the algorithms, with the number of clusters determined through a majority vote of the different CVIs used (Sard, 2015). The indices used for this study are the Silhouette index (Sil), the Dunn index (D), COP index, Davies-Bouldin index (DB), Calinski-Harabasz index (CH) all covered in detail in Arbelaitz et al. (2013) together with the Modified Davies-Bouldin index (DBstar) (Kim and Ramakrishna, 2005) and the Score Function (SF) (Saitta, Raphael and Smith, 2007). While different CVIs most suited for the Fuzzy algorithm could also have been implemented, this was not deemed necessary based on the objective of this study and the output of a few trial runs.

Initially, due to the high dimensionality of the dataset, the first step of the analysis was to compile a random sample of 5,000 records. Each of the clustering algorithms was run and all above CVIs for the sample dataset with *k=2:10 (inclusive)*, where *k* is the number of clusters. The clustering algorithms were thereafter implemented on the entire dataset using the selected optimum number of clusters.

Accuracy Degradation Analysis

For each combination of the derived clusters, the average consumption quantile, the consumer category and a classification of whether a meter is new or old were established. The temporal variation of the annual running consumption was evaluated using the relationship (Mbabazi *et al.*, 2015);

$$y = \beta_0 + \beta_1 x \qquad \textbf{\textit{Equation 3}}$$

Where y is the average annual running consumption, x is the monthly time-step, with β_0 and β_1 as constants to be determined. Other than the differences in approach thus far, the other main difference between this research and that of Mbabazi *et al.* (2015) is the use of the average annual running total every month instead of the annual consumption for meters of various calendar years. The method adopted in this research maximises the use of each individual time series without taking only a small component for analysis. The degradation rate d is calculated for each combination as;

$$d = \frac{\beta_1}{\beta_0} \qquad \textbf{\textit{Equation 4}}$$

The accuracy degradation analysis also took cognisance of a categorical classification of whether a meter is new or old. New meters were assumed to have an initial reading of up to twice the average annual consumption of that category while old meters were those with an initial reading of at least five times the average annual consumption. The degradation rates for new meters and that of old meters was thereafter compared to evaluate the progression of the degradation, or the lack thereof.

RESULTS AND DISCUSSION

The surviving dataset, together with the composition of the consumer categories from the billing data are summarised in Table 1. While there is an over-representation of the residential customers and an under-representation of the business category, the overall distribution results were deemed to be adequately indicative.

Table 1: Surviving Dataset

Consumption Category	No. of	%	% in JW
Residential	81,211	92.7	85.9
Multiple Residential Dwelling	2,278	2.6	2.5
Business	3,563	4.1	10.8
Utility (Public Benefit Organisation)	537	0.6	0.8
Total	87,589	100	100

The variation of the average running annual consumption is shown in Figure 2. An expected result was that all non-residential categories have a much higher annual consumption, up to an order of a magnitude than the residential sector. This therefore does emphasise the importance of z-normalisation in the clustering algorithm to remove differences in amplitude and focus on shape similarity.

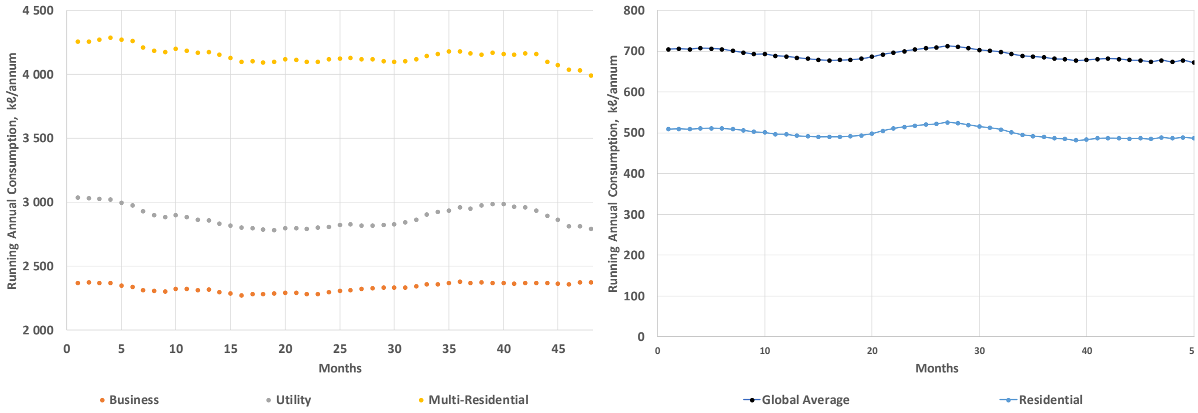

Figure 2: Variation of the running average annual consumption per category

The effect of the residential consumption on the global average is also very clear and emphasises the importance of managing the accuracy of this segment despite the lower consumption per individual consumer

The results of the clustering analysis showed that two clusters, out of up to ten clusters evaluated, were the “purest” number of clusters. These two clusters consist of records with decreasing annual consumption over time while the other one had increasing consumption as depicted in Figure 3. It must be noted that even with up to ten clusters, the two main characteristics remained a combination of clusters with increasing and decreasing consumption, albeit at different amplitudes. Due to the stochastic nature of the algorithms there were slight differences in the composition of clusters, but the results were similar. There are also some peculiar variations at the tail end of both the clusters owing to very few records with more than five years of data, thereby skewing the distribution.

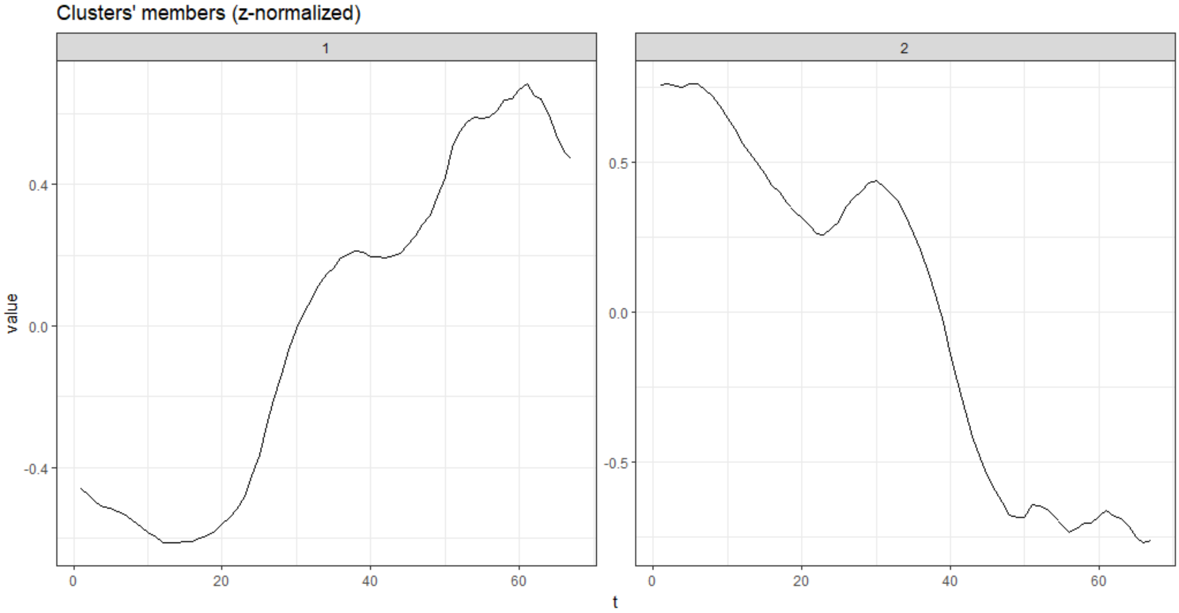

Figure 3: Centroids for Running Annual Consumption Clusters (Fuzzy c-means algorithm)

Using the clustering results, the final sample dataset (without z-normalisation) was accordingly classified into the two clusters for accuracy degradation analysis. The running annual consumption for each of the clusters is shown in Figure 4 confirming the output of the cluster analysis of declining and increasing consumption clusters. The graph also shows the downside of calculating the degradation rate at only a point (i.e. for only a year), particularly if the results are not validated, as the annual consumption is not static, a fact not properly accounted for in prior studies.

The average annual degradation rate for each of the categories of consumers was calculated using Equations 3 and 4 from all consumption quantiles with an $R^2 > 0.7$, which excluded a few and insignificant quantiles for consumption categories that had a few data points within that quantile. The results of the degradation rates are tabulated in Table 2. New and Old refers to the categorical classification based on the initial reading of the meter while the average aggregates all meters regardless of initial meter reading and the decline columns is the difference between the rate for new and old meters.

The properties with a declining consumption show a clear trend of decrease in consumption ranging from 0.7% to 1.1% per annum, with the residential sector (small meters) having a consistent rate for both old and new meters. The categories with higher consumption have a higher degradation rate, which is expected due to the higher wear and tear of mechanical meters that are exclusively used. The multi-residential sector also shows the largest differential between new and old meters and this again is very logical as these are meters that would typically be the most used at varying flowrates due to the many permutations of users using water at various times and in varying volumes.

What has previously been missing in such analysis is how to handle increasing consumption, which were either entirely excluded from the analysis or considered as an anomaly and hence ignored. From Table 2, the properties with increasing consumption show that between new and old meters, there is a meaningful difference between the rates of increase of consumption. A closer scrutiny shows that the decline in the increasing consumption is similar, albeit lower, than the overall decline experienced by meters with declining consumption. This suggests that what could be more relevant for properties with increasing consumption is how the rate of increase decreases from the time meters are new to when they are old.

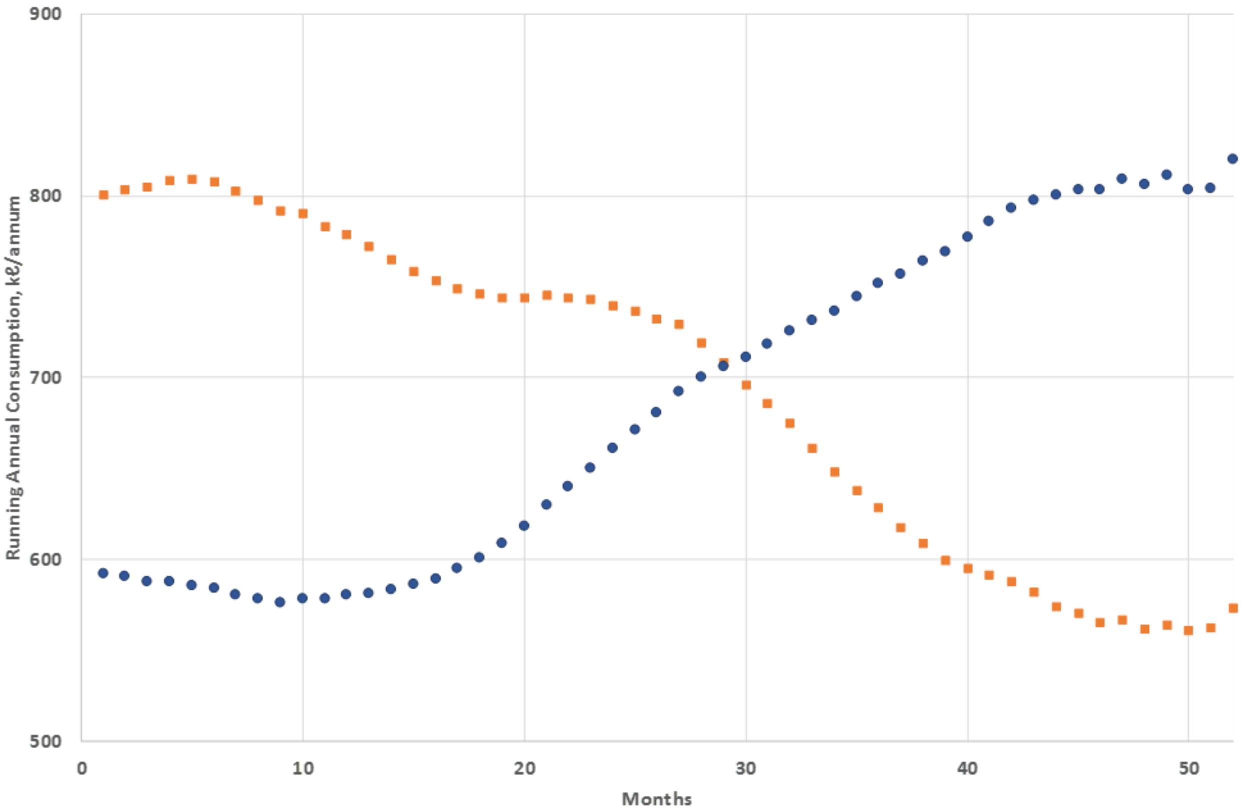

Figure 4: Cluster Average Consumption

Table 2: Degradation Rates

Consumption Category	**Declining Consumption**				**Increasing Consumption**			
	New	Ave	Old	Decline	New	Ave	Old	Decline
Residential	-0.7%	-0.7%	-0.7%	0.0%	1.4%	1.1%	0.8%	0.5%
Multiple Residential Dwelling	-0.7%	-0.7%	-1.1%	0.3%	1.4%	1.3%	0.6%	0.8%
Business	-0.9%	-0.9%	-1.0%	0.1%	1.9%	1.7%	1.4%	0.5%
Public Benefit Organisation	-0.9%	-0.9%	-0.9%	0.1%	1.8%	1.8%	1.2%	0.7%
Combined	-0.7%	-0.7%	-0.8%	0.1%	1.5%	1.1%	0.8%	0.7%

The lower rates can be explained by the combined effect of increasing losses and decreasing accuracy that mask the actual expected loss in accuracy. Considering the high extent, greater than 60% occurrence, of on-site leakage within the study area (Lugoma, Van Zyl and Ilemobade, 2012; Ncube and Taigbenu, 2016), with leaks tending to get bigger with time, the increasing consumption cluster is very plausible. The use of the average rate in this case would be very misleading. The decline in the degradation is also the highest for multi-residential consumers and will be attributed to the same reasons as for the declining consumption cluster. However, both degradation rates are much lower than the 1.45% to 6.67% in Mbabazi et al. (2015) and 2.1% per year in Arregui, Cabrera E. & Cobacho (2006) which both used comparative billing analysis. Both studies however concede that such rates are much higher than those found using the weighted meter accuracy method with ranges from 0.1% to 0.7% in Arregui, Cabrera E. & Cobacho (2006) and 0.1% to 0.6% in Noss *et al.* (1987). The results of the developed method are therefore comparable with degradation rates from weighted accuracy methodology, which is considered the gold standard for measuring meter errors.

For validation purposes, the field based results of Ncube and Taigbenu (2018) for the same study area were used, with the average age of the meters being 11.5 years. The field assessment was based on 408 properties using highly accurate R800 meters with a low starting flow rate that captured water flows that would normally not be read by conventional domestic meters (Ncube and Taigbenu, 2016). The field estimates therefore inherently include meter under-registration, while this will not be measured by any data-mining based approach. Additionally, unlike this current study which includes all meter sizes for all consumers, the field assessment concentrated on domestic meters of up to 25mm and its results excluded bulk meters which have been shown to be oversized within the study area (Ncube and Taigbenu, 2016). Most non-domestic consumers included in the 2016 study were essentially high-volume consumers with small meters. The validation results are reproduced in Table 3, together with the estimates of the data-mining method described in this paper and presented in columns three and four for the respective clusters. For simplicity, the average degradation rates for the declining consumption cluster and the decline in the increasing consumption clusters were used.

Table 3: Estimated Apparent Water Losses

Consumption Category	Field Estimates (Best Case), %*	Declining Consumption	Increasing Consumption
Residential	11.2	8.1%	6.2%
Multiple Residential Dwelling	6.5	8.2%	9.8%
Business	8.3	10.6%	6.0%
Public Benefit Organisation	6.4	10.2%	7.5%
Average	9.4	8.2%	7.9%

* From Ncube and Taigbenu (2018)

In line with the approach of the field measurements, the residential category offers the best basis for comparison of results as both methods considered only small meters as opposed to the other categories. The overall average scenarios should also be sufficient for comparison purposes. On average, the differences between the declining and increasing consumption clusters is very small and both are comparable with field estimates. These results are an improvement on prior estimates in Ncube & Taigbenu (2017) where the method had a much higher estimate of 14%, which was similar to the worst case scenario of assuming all consumers having 20mm meters instead of 15mm meters. This would therefore be an indication of the improved accuracy of the developed methodology compared to the previous one.

For the residential category, both the declining consumption and the increasing consumption clusters understate the losses as determined through the field-based approach by 3.1% and 5% respectively. This difference, particularly for the declining consumption cluster, can be attributed to meter under-registration that is accounted for in the field-based approach. Arregui, Balaguer & Soriano (2017) have shown that the expected initial error of ISO velocity meters, under the most favourable working conditions, ranges from an excellent -0.71% to -3.87%, while oscillating pistol meters range from 0% to -1%. Such errors will not be captured by the data-mining methodology as they will not be reflected on a meter reading. It would therefore be reasonable to assume that the difference of 3.1% for the decreasing consumption cluster could be accounted for by the initial meter error, validating the results for the residential category. The bigger differences for the increasing consumption cluster is potentially the complex compounding effect of meter under-registration and increasing level of leakages with time. It is therefore recommended that the increasing consumption cluster be used only to provide some form of sanity check for similarities with the other cluster and the traditional method, without necessarily adopting its degradation rates.

As expected and due to the inclusion of bulk meters for the error estimates from the developed method, all other categories have higher estimates of apparent losses than from the field-based method, particularly when looking at the declining consumption cluster. This also tends to agree with field observations of meter sizing challenges within the Johannesburg Water metering fleet. However, due to the lack of adequate and accurate information, these bulk consumer results should only be taken as indicative.

Generalised Equation for Apparent Loss Estimation

For utilities having similar water meters and consumption characteristics as Johannesburg, this methodology and its results offer typical degradation rates of 0.7% per annum of meter accuracy for domestic consumers. Consideration for meter under-registration needs to be factored in, and in the absence of actual data estimates such as those found in Arregui, Balaguer & Soriano (2017), can be adopted with allowances of 4% - 5% for initial meter errors. It has also been shown in Ncube & Taigbenu (2017) that upsizing a 15mm to 20mm increases the metering error by up to 50% and therefore utilities with a legacy of incorrect meter sizing should allow for another margin of error – an additional allowance of up to 5% is suggested in such cases. In keeping with other generalised values of apparent loss (Seago, Bhagwan and McKenzie, 2004; Mutikanga, Sharma and Vairavamoorthy, 2011) the following relationship for apparent losses due to metering errors is proposed;

$$Apparent\ Loss\ Estimate = \frac{0.7a+4b+5c}{100} \qquad \textbf{\textit{Equation 5}}$$

Where ***a*** is the average meter age, ***b*** is the estimated probability of meter under-registration and ***c*** is the probability of meter oversizing. These probabilities relate to the users' judgement on the prevalence of meter under-registration and meter oversizing. The factor for ***a*** relates to the degradation rate of 0.7% per year, while those for ***b*** and ***c*** relate to the allowances for meter under registration and oversizing, respectively. A factor of one for either of the variables defaults to the suggested values for Johannesburg, with 100% oversized meters for residential consumers. No allowance has been made for illegal or unmetered connections, data handling errors, and additional errors due to the effect of the use of storage tanks as these have not been evaluated in this study.

CONCLUSIONS

In an era where there is an urgent need for efficient methods that can assist in the estimation of water losses and in particular, the apparent loss component, the improved comparative billing methodology developed in this study could be very handy. Using a combination of data preparation techniques, clustering analysis and classical regression analysis, the methodology uses commonly available monthly billing data and has been validated for a water utility in Johannesburg, South Africa producing comparable results at a fraction of the cost and effort of the traditional field-based method. Central to this novel methodology, is the disaggregation of underlying consumption patterns into similar clusters that are analysed separately to eliminate the averaging effect. Apparent losses due to metering error are estimated to be 8.2% of billed consumption compared to 9.4% found using the field-based method, with an error degradation rate of 0.7% per annum. The difference in the values are attributed to meter under-registration which is not accounted for with the new developed method and can be higher than 4% for new meters. Average results are even closer with only a difference of 1.2% which is also better that the previous estimates using similar alternative methods.

About 15% of the available data was used in development of the method, which is a significant improvement from previous attempts of related work. With better data quality, it is possible to further enhance the method for detailed classification purposes to capture consumer segment identification, different metering technologies, and other categories of interest, which was not possible in this study.

Data-driven discovery processes therefore do offer viable alternatives that can complement traditional water loss assessment processes as part of a toolset that utilities can use to better manage water losses. This can come in handy in counteracting the difficultly and costs involved in field assessments and laboratory work. However, to be able to leverage on these emerging tools, utilities must place greater emphasis in curating reliable and clean datasets that con be mined for useful trends and information.

REFERENCES

Abrahart, R. J., See, L. M. and Solomatine, D. P. (eds) (2008) *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*. Springer-Verlag Berlin Heidelberg.

Aghabozorgi, S., Shirkhorshidi, A. S. and Wah, T. Y. (2015) 'Time-series clustering - A decade review', *Information Systems*. Elsevier, 53(May), pp. 16–38. doi: 10.1016/j.is.2015.04.007.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., P??rez, J. M. and Perona, I. (2013) 'An extensive comparative study of cluster validity indices', *Pattern Recognition*, 46(1), pp. 243–256. doi: 10.1016/j.patcog.2012.07.021.

Arregui, F. J., Balaguer, M. and Soriano, J. (2017) 'Quantifying measuring errors of new residential water meters considering different customer consumption patterns', *Urban Water Journal*. Taylor Francis, 9006(October), pp. 1–13. doi: 10.1080/1573062X.2014.993999.

Arregui, F. J., Cabrera Jr., E. and Cobacho, R. (2006) *Integrated Water Meter Management*. IWA Publishing.

AWWA (2009) *Water Audits and Loss Control Programs - M36*. Third. American Water Works Association.

Babovic, V. (2005) 'Data mining in hydrology', *Hydrological Processes*, 19(7), pp. 1511–1515. doi: 10.1002/hyp.5862.

Babovic, V., Drécourt, J.-P., Keijzer, M. and Friss Hansen, P. (2002) 'A data mining approach to modelling of water supply assets', *Urban Water*, 4(4), pp. 401–414. doi: 10.1016/S1462-0758(02)00034-1.

Bezdek, J. C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer, Boston, MA. doi: https://doi.org/10.1007/978-1-4757-0450-1.

Chauhan, A. and Rajvanshi, S. (2013) 'Non-Technical Losses in Power System: A Review', in *Proceedings of the 2013 International Conference on Power, Energy and Control (ICPEC)*, pp. 558–561. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6527720 (Accessed: 14 January 2014).

Couvelis, F. a and van Zyl, J. E. (2015) 'Apparent losses due to domestic water meter under-registration in South Africa', *Water SA*, 41(5), pp. 698–704. doi: http://dx.doi.org/10.4314/wsa.v41i5.13.

Criminisi, A., Fontanazza, C. M., Freni, G. and La Loggia, G. (2009) 'Evaluation of the apparent losses caused by water meter under-registration in intermittent water supply.', *Water Science and Technology*, 60(9), pp. 2373–82. doi: 10.2166/wst.2009.423.

Department of Water and Sanitation (2017) *Benchmark of Water Losses, Water Use Eficiency and Non Revenue Water in South African Municipalities (2004/05 - 2015/16)*. Pretoria, South Africa.

Dhanya, C. T. and Nagesh Kumar, D. (2009) 'Data mining for evolution of association rules for droughts and floods in India using climate inputs', *Journal of Geophysical*

Research, 114(D2), p. D02102. doi: 10.1029/2008JD010485.

Gorunescu, F. (2011) *Data Mining: Concepts, models and techniques*. Edited by J. Kacprzyk and L. C.Jain. Springer-Verlag Berlin Heidelberg.

Green Cape (2017) *Water – 2017 Market Intelligence Report* –. Cape Town, South Africa. Available at: https://www.greencape.co.za/assets/Uploads/GreenCape-Water-MIR-2017-electronic-FINAL-v1.pdf (Accessed: 7 August 2017).

Humaid, E. H. and Barhoum, T. (2013) 'Water Consumption Financial Fraud Detection: A Model Based on Rule Induction', *2013 Palestinian International Conference on Information and Communication Technology*. Ieee, pp. 115–120. doi: 10.1109/PICICT.2013.28.

Jain, A., Murty, M. and Flynn, P. (1999) 'Data clustering: a review', *ACM computing surveys (CSUR)*, 31(3), pp. 264–323.

Kalteh, A. M., Hjorth, P. and Berndtsson, R. (2008) 'Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application', *Environmental Modelling & Software*, 23(7), pp. 835–845. doi: 10.1016/j.envsoft.2007.10.001.

Kim, M. and Ramakrishna, R. S. (2005) 'New indices for cluster validity assessment', *Pattern Recognition Letters*. North-Holland, 26(15), pp. 2353–2363. doi: 10.1016/j.patrec.2005.04.007.

Klingel, P. and Knobloch, A. (2015) 'A Review of Water Balance Application in Water Supply', *Journal - American Water Works Association*, 107(July), pp. E339–E350. doi: 10.5942/jawwa.2015.107.0084.

Loureiro, D., Alegre, H., Coelho, S. T., Martins, A. and Mamade, a. (2014) 'A new approach to improve water loss control using smart metering data', *Water Science & Technology: Water Supply*, 14(4), p. 618. doi: 10.2166/ws.2014.016.

Lugoma, M. F. T., Van Zyl, J. and Ilemobade, A. A. (2012) 'The extent of on-site leakage in selected suburbs of Johannesburg', *Water SA*, 38(1), pp. 127–132. doi: 10.4314/wsa.v38i1.15.

Mbabazi, D., Banadda, N., Kiggundu, N., Mutikanga, H. and Babu, M. (2015) 'Determination of domestic water meter accuracy degradation rates in Uganda', *Journal of Water Supply: Research and Technology - AQUA*, 64(4), pp. 486–492. doi: 10.2166/aqua.2015.083.

McKenzie, R., Siqalaba, Z. and Wegelin, W. (2012) *The State of Non-Revenue Water in South Africa (2012)*. Gezina, Pretoria, South Africa: Water Research Commission.

Monedero, I., Biscarri, F., Guerrero, J. I., Peña, M., Roldán, M. and León, C. (2016) 'Detection of Water Meter Under-Registration Using Statistical Algorithms', *Journal of Water Resources Planning and Management*, 142(1), p. 4015036. doi: 10.1061/(ASCE)WR.1943-5452.0000562.

Mutikanga, H. E., Sharma, S. K. and Vairavamoorthy, K. (2011) 'Assessment of apparent losses in urban water systems', *Water and Environment Journal*, 25(3), pp. 327–335. doi: 10.1111/j.1747-6593.2010.00225.x.

Ncube, M. and Taigbenu, A. (2015) 'Meter Accuracy Degradation and Failure Probability based on Meter Tests and Meter Change Data', in *Proceedings of the 4th YWP-ZA Biennial Conference and 1st African YWP Conference*. Pretoria, South Africa.

Ncube, M. and Taigbenu, A. E. (2016) 'Consumption Characterisation and On-Site Leakage in Johannesburg, South Africa', in *Proceeding of the IWA Water Loss Conference 2016*.

Ncube, M. and Taigbenu, A. E. (2018) 'Assessment of Apparent Water Losses - A Comparative Approach', *In Press*.

Ngai, E. W. T., Xiu, L. and Chau, D. C. K. (2009) ‘Application of data mining techniques in customer relationship management: A literature review and classification’, *Expert Systems with Applications*. Elsevier Ltd, 36(2), pp. 2592–2602. doi: 10.1016/j.eswa.2008.02.021.

Nguyen, K. a., Stewart, R. a. and Zhang, H. (2013) ‘An intelligent pattern recognition model to automate the categorisation of residential water end-use events’, *Environmental Modelling & Software*. Elsevier Ltd, 47, pp. 108–127. doi: 10.1016/j.envsoft.2013.05.002.

Noss, R. R., Newman, G. J. and Male, J. W. (1987) ‘Optimal Testing Frequency for Domestic Water Meters’, *Journal of Water Resources Planning and Management*, 113(1), pp. 1–14. doi: 10.1061/(ASCE)0733-9496(1987)113:1(1).

Paparrizos, J. and Gravano, L. (2015) ‘k-Shape: Efficient and Accurate Clustering of Time Series’, *ACM Sigmod*, pp. 1855–1870. doi: 10.1145/2723372.2737793.

R Core Team (2017) ‘R: A Language and Environment for Statistical Computing’. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org.

Räsänen, T., Ruuskanen, J. and Kolehmainen, M. (2008) ‘Reducing energy consumption by using self-organizing maps to create more personalized electricity use information’, *Applied Energy*, 85(9), pp. 830–840. doi: 10.1016/j.apenergy.2007.10.012.

Saitta, S., Raphael, B. and Smith, I. F. C. (2007) ‘A Bounded Index for Cluster Validity’, in Perner, P. (ed.) *Machine Learning and Data Mining in Pattern Recognition: 5th International Conference, MLDM 2007, Leipzig, Germany, July 18-20, 2007. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 174–187. doi: 10.1007/978-3-540-73499-4_14.

Sard, A. (2015) ‘Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package’. Available at: https://cran.r-project.org/web/packages/dtwclust/index.html.

Seago, C., Bhagwan, J. and McKenzie, R. (2004) ‘Benchmarking leakage from water reticulation systems in South Africa’, *Water SA*, 30(5), pp. 573–580.

De Silva, D., Yu, X., Alahakoon, D. and Holmes, G. (2011) ‘A data mining framework for electricity consumption analysis from meter data’, *IEEE Transactions on Industrial Informatics*, 7(3), pp. 399–407. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5893960 (Accessed: 27 August 2013).

Solomatine, D. P. (2003) ‘Applications of data-driven modelling and machine learning in control of water resources’, in Mohammadian, M., Sarker, R. A., and Yao, X. (eds) *Computational Intelligence in Control*. Hershey, PA: Idea Group Publishing, pp. 197–217.

World Economic Forum (2017) *The Global Risks Report 2017: 12th Edition*. Geneva, Switzerland. Available at: http://www3.weforum.org/docs/GRR17_Report_web.pdf.

Yin, Y., Kaku, I., Tang, J. and Zhu, J. (2011) *Data Mining: Concepts, Methods and Applications in Management and Engineering Design*. Edited by R. Roy. Springer-Verlag London.

Zeileis, A. and Grothendieck, G. (2005) ‘zoo: S3 Infrastructure for Regular and Irregular Time Series’, *Journal of Statistical Software*, 14(6), pp. 1–27. doi: 10.1017/CBO9781107415324.004.