Assessment of apparent losses due to meter inaccuracy using an alternative, validated methodology

Mthokozisi Ncube and Akpofure E. Taigbenu

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. For apparent losses assessment, the traditional approach of deriving consumption profiles and testing water meters 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 into the underlying consumption patterns. The results show that data-driven discovery processes are viable alternatives for improved assessment and management of water losses.

Key words | apparent losses, data mining, metering errors, non-revenue water (NRW), water loss management

INTRODUCTION

Water loss management

Water loss control remains a major challenge in the sustainability of water utilities and the promotion of the efficient use of water as a finite natural resource (Loureiro et al. 2014), and is 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 (GreenCape 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 is attributed to improved quantification of water losses in rural municipalities as opposed to system degradation (Department of Water & Sanitation 2017). Yet, this 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 and is not unique to the developing world. A review of the IWA water balance 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 & Knobloch 2015). Developing and improving methods for enhanced water loss management stand 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 of either the billed volume or the non-revenue water, such as in Seago et al. (2004) and McKenzie et al. (2012). Some authors, such as Mousavi et al. (2017), omit...
apparent losses estimates in the water balance, even when some of its components such as unauthorized connections are quantified. Such studies clearly show that apparent loss estimation remains a challenge, with the unavailability of data regarding meter errors, unauthorized use and unbilled authorized use being cited as the leading constraint (Kanakoudis et al. 2015). Without the proper quantification of apparent losses, or any of the other water loss components, water loss interventions may not be adequately informed or optimally implemented.

While a number of authors (Arregui et al. 2006; AWWA 2009; Criminisi et al. 2009; Mutikanga et al. 2011; Ncube & 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 mitigate this difficulty have been suggested and they include the use of a meter replacement database (Couvelis & van Zyl 2015) and that of a comparative analysis of a billing database (Mbabazi et al. 2015). There has, however, been limited published application and validation of such methods. Ncube & Taigbenu (2018) compared these alternatives with the empirical method and demonstrated that further refinements are required to make the methods efficient, reproducible and accurate. In particular, application of the methodology proposed by Mbabazi et al. (2015) was applied to determine water meter accuracy degradation for Johannesburg Water (JW), the largest water utility in South Africa. However, comparison of the results from the method and those based on field and laboratory measurements showed a huge difference. Additional gaps identified include the following:

- The lack of clarity in the selection of the final dataset, which introduces subjectivity and lack of reproducibility.
- The failure to maximally utilize the consumption dataset, with only 0.4% of the available data used over a limited 12-month period.
- Lack of differentiation of consumption clusters such as that of increasing consumption over time, as found in Ncube & Taigbenu (2015). Reference is only made to declining consumption.

This paper therefore extends the work of Mbabazi et al. (2015) by proposing additional analysis to mitigate against the identified gaps and also validates the methodology against the results previously obtained from the empirical field- and laboratory-based method as detailed in Ncube & Taigbenu (2018).

**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 metering, 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 support water utilities in establishing complete water balances (Klingel & Knobloch 2015).

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 and health care services, supply chain management, telecommunications (Gorunescu 2011), customer relations management, electricity, water resources, water asset management and water consumption and metering-related applications such as in Monedero et al. (2016). Water loss management can therefore also 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). The use of data mining, as applied in various sectors, is one such promising alternative for water loss management.

A critical component of apparent water loss analysis is the derivation of errors for different devices that are in use for different consumption categories. However, many utilities have inadequate information on meter-type characteristics and consumer categories to achieve such disaggregation. As such, unsupervised clustering is well suited to deriving distinct groupings of consumption profiles without classification a priori. 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 & Flynn 1999). Time-series clustering is a special type which is of interest due to its ubiquity in various applications (Aghabozorgi Shirkhorshidi & Wah 2015). In particular, water loss consumption data are, by their very nature, a typical example of a time series. Examples of time-series algorithms that are well suited for the analysis of large dataset sets include the following:

1. **k-Shape Clustering**: a partitional clustering algorithm developed 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 $z$-normalization requirement. SBD is given by the formula (Sard 2015):

$$SBD(x, y) = 1 - \frac{\max(NCCc(x, y))}{||x||_2||y||_2}$$  \hspace{1cm} (1)

where $||·||_2$ is the Euclidean norm of the series. Shape extraction relies on NCCc and uses it to optimally match any two series with random selection of the centroid series. Alignment can be done between series with different lengths, but the length of the resulting prototype depends on a chosen reference length.

2. **Fuzzy Clustering**: a clustering algorithm that outputs soft partitions with each record belonging to all clusters to a certain degree. In particular, the fuzzy $c$-means clustering implementation of Bezdek (1981) 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):

$$\mu_{c,i} = \frac{\sum_{p=1}^{N} u_{p,c} x_{p,i}}{\sum_{p=1}^{N} u_{p,c}}$$  \hspace{1cm} (2)

From Equation (2), all the time-series data are required to have the same dimensionality, achieved through the re-interpolation of the data to match the longest time series. The non-crisp partitions of the algorithms can be made crisp by taking the maxima of each cluster.

Both of these fall under partitional clustering algorithms, which was well suited for this study as opposed to hierarchal clustering due to its time and the memory complexity of $O(N^2)$, where $N$ is the total number of records in the dataset. This paper proposes and applies data mining algorithms to the specific area of assessing apparent water losses due to metering errors using monthly water consumption data. This is the first known attempt to apply such tools for the express purpose of quantifying apparent losses.

### City of Johannesburg

The City of Johannesburg is the largest metropolitan in South Africa with a population of 4.4 million and with 1.4 million households as per the 2011 Census. Through its municipal entity, Johannesburg Water (JW), the city supplies about 1,600 ML/day of potable drinking water to an estimated 561,000 connections of which up to 81% are metered. Of the metered connections, up to 325,000 are conventional post-payment meters.

In a related study, the authors assessed the apparent losses due to metering errors using the field-based approach for the conventional metered connections and found it to be, on average, 9.4% of the metered consumption (Ncube & Taigbenu 2018). That study was implemented over a period of two years. In this paper, the monthly billing data of the same conventional metered consumers are used to derive an estimate for apparent losses and compared with the prior results.

### METHODOLOGY

The developed methodology comprises data preparation, clustering and accuracy degradation analyses that were applied on JW’s monthly meter-reading records from July 2003 to June 2015. The 144 monthly flat 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 month.

Data preparation

Each monthly 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 remove duplicate meter numbers and property IDs, resulting in slightly over 600,000 records. Stringent criteria had to be introduced to minimize data-cleansing requirements and increase prediction accuracy. After a couple of trial runs, only records which met the following criteria were retained, in order of priority:

- Readings after 2009 were ignored due to the significant errors introduced by the change of the billing system.
- A long record of at least 60 individual readings.
- Less than 20% patching of missing data required.
- Meters that did not clock-over.
- Records without constant readings or abnormally high readings.

An important variation in the analysis was the use of running annual total consumption to ensure that the time-series property is retained but at a scale that is not susceptible to monthly and seasonal variations. This maximized the use of the entire record of each individual time-series in the analysis. The date/time signature was removed from all records and replaced with a count of the months for the clustering analysis as the actual date was not required for the pattern assessment. After applying the data-cleansing criteria, a sample of 87,589 (about 15%) records remained, which was considered adequate for method development. The number is a significant population of active consumers but is also comparable to the number of consumers in many small to medium sized utilities. The data were thereafter patched using the ‘zoo package’ (Zeileis & Grothendieck 2005) within the R Statistical Software (R Core Team 2017) and also processed to produce an annual consumption time series for all the records.

Clustering analysis

The clustering analysis was implemented within the R Statistical Software (R Core Team 2017) using the dtwclust package (Sard 2015). The dtwclust package was chosen as it provides a common platform on which classical and new clustering algorithms can be evaluated and compared against each other (Sard 2015). The k-Shape and fuzzy clustering algorithms were chosen due to their relative speed and ease of implementation.

A non-trivial aspect of time-series clustering is the determination of the number of clusters, \( k \), and evaluation of 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 (Sard 2015). The indices used for this study were the Silhouette index (Sil), the Dunn index (D), COP index, Davies–Bouldin index (DB), and Calinski–Harabasz index (CH) that are covered in detail in Arbelaitz et al. (2013) together with the Modified Davies–Bouldin index (DBstar) (Kim & Ramakrishna 2005) and the Score Function (SF) (Saitta et al. 2007).

Initially, due to the high dimensionality of the dataset, the first step was to perform preliminary clustering using only 5,000 time-series records that were randomly selected. The CVIs were found for the sample dataset with \( k = 2:10 \) (inclusive). The clustering algorithms were thereafter implemented on the entire dataset using the selected optimum number of clusters.

Accuracy degradation analysis

From the output of the clustering analysis, the final sample dataset (without z-normalization) was accordingly classified for accuracy degradation analysis. The temporal variation of the annual running consumption for each cluster was evaluated using the relationship (Mbabazi et al. 2015):

\[
y = \beta_0 + \beta_1 x
\]  

(3)

where \( y \) is the average annual running consumption, \( x \) is the monthly time-step, with \( \beta_0 \) and \( \beta_1 \) as constants.
to be determined. The degradation rate $d$ was calculated as:

$$d = \frac{\beta_1}{\beta_0}$$

(4)

Within each cluster, quantile categories of water consumption were also determined and the degradation analysis performed per consumption quantile. The categorical classification of whether a meter is new or old was incorporated into the analysis using the initial meter reading for each record. This classification was based on the average consumption per quantile, which was in multiples of ten and the additional 5% and 95% quantiles. New meters were assumed to have an initial reading of up to twice the average annual consumption of that consumption quantile 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 those of old meters were thereafter compared to evaluate the progression of the degradation, or the lack thereof.

**RESULTS AND DISCUSSION**

The results of the clustering analysis showed that two clusters, out of ten evaluated clusters, were the ‘purest’ number of clusters. These two clusters consist of records with decreasing annual consumption over time and of records with increasing consumption, as depicted in Figure 1 from the fuzzy algorithm. Due to the stochastic nature of the evaluated algorithms there were slight differences in the composition of clusters, but the results were similar for both algorithms. 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 1 shows the downside of calculating the degradation rate at only selected points (such as only for a year), particularly if the results are not validated, as the annual consumption is not static, a fact not properly accounted for in previous studies. This therefore underscores the importance of using the entire time-series record wherever possible.

The clustering algorithms provided the two main consumption patterns in the data and they could be used to explore additional information that drills down into the subcategories of consumers. This demonstrates the capability of the method to meet one key requirement of apparent loss assessment of disaggregating data to homogeneous consumer groupings before subsequent analysis. The results of the degradation rates are tabulated in Table 1. ‘New’ and ‘Old’ refer to the categorical classification based on the initial reading of the meter while ‘Average’ aggregates all meters regardless of initial meter reading and the ‘Decline’ columns are the difference between the rates for new and old meters.

The properties with a declining consumption show a trend of decrease in the rates of consumption which range from 0.7% to 1.1% per annum, with the residential sector having a consistent rate for both old and new meters. This was a rather surprising finding and might be related to the predominantly linear degradation of meter error for residential consumers who comparatively do not use a lot of water. The other consumption categories with higher consumption have correspondingly higher degradation rates, which is expected due to the greater wear and tear associated with the use of mechanical meters for high volumes. 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 with higher variability of flowrates.

From Table 1, the properties with increasing consumption show that there are significant differences between the rates of increase of consumption for new and old meters. A closer scrutiny of the results shows that the magnitude of decline in the increasing consumption is slightly lower than the overall declining rate of meters with declining consumption. This suggests that what could be more relevant for properties with increasing consumption is how the rate of increase decreases with the age of the meters. The lower rates can be explained by the combined effect of increasing losses and decreasing accuracy that masks the actual expected loss in accuracy. Considering the high extent of on-site leakage within the study area (Lugoma et al. 2012; Ncube & Taigbenu 2016), with leaks tending to get larger with time, the increasing consumption cluster is very plausible. The highest decline in the degradation is also observed for multi-residential consumers and
this is attributed to the same reason adduced for the declining consumption cluster.

The degradation rates for increasing and decreasing consumptions are much lower than those of comparable studies that were based on alternative methodologies such as the 1.45% to 6.67% in Mbabazi et al. (2015) and 2.1% per year in Arregui et al. (2006). Both the comparable studies conceded that their rates were much higher than those found using the weighted meter accuracy method with ranges from 0.1% to 0.7% in Arregui et al. (2006) and 0.1% to 0.6% in Noss et al. (1987). However, the results of this study are comparable with degradation rates from the traditional weighted accuracy methodology, demonstrating the improved estimation of the current methodology.

For validation purposes, the field-based results of Ncube & Taigbenu (2018) for the same area with an average meter age of 11.5 years were used. These field estimates inherently measure meter under-registration through laboratory testing which is not evaluated by this data-mining-based approach. This is because the billing records that underpin the current approach do not include such flows as by definition they are not recorded on the meter register but can only be found through meter testing. Additionally, the field assessment was mostly for domestic meters of up to 25 mm while the

Table 1 | Degradation rates

<table>
<thead>
<tr>
<th>Consumption category</th>
<th>Declining consumption</th>
<th>Increasing consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New</td>
<td>Ave</td>
</tr>
<tr>
<td>Residential</td>
<td>-0.7%</td>
<td>-0.7%</td>
</tr>
<tr>
<td>Multiple Residential Dwelling</td>
<td>-0.7%</td>
<td>-0.7%</td>
</tr>
<tr>
<td>Business</td>
<td>-0.9%</td>
<td>-0.9%</td>
</tr>
<tr>
<td>Public Benefit Organization</td>
<td>-0.9%</td>
<td>-0.9%</td>
</tr>
<tr>
<td>Combined</td>
<td>-0.7%</td>
<td>-0.7%</td>
</tr>
</tbody>
</table>
The current study included all meter sizes. The validation results are reproduced in Table 2, together with the estimates of the data mining method of this study. For simplicity, the average degradation rates for the declining consumption cluster and the decline in the increasing consumption cluster were used over the 11.5-year period to estimate the apparent losses.

On average, the differences between the declining and increasing consumption clusters are very small and they are comparable with field estimates. These results are an improvement on prior estimates in Ncube & Taigbenu (2018) where the application of the methodology of Mbabazi et al. (2015) produced a much higher estimate of 14%, which was similar to the worst-case scenario that is considered to be an overestimation. This is an indication of the improved accuracy of the new methodology.

For the residential category, which presents the best basis for comparison due to the common use of small meters, both the declining and 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 only accounted for in the laboratory-based approach. Arregui et al. (2018) indicate that the initial (new) meter error of ISO meters can be up to −5%. The difference of −3.1% for the decreasing consumption cluster is within this range and can therefore be attributed to the initial meter error, validating the results for the residential category. The larger differences for the increasing consumption cluster is potentially the complex combined effects of meter under-registration and increasing level of leakages with time, and should therefore not be used other than as a sanity check for similarities.

### Generalized equation for apparent loss estimation

In keeping with other studies which provide a generalized form for the estimation of apparent losses, such as Mutikanga et al. (2011) and Seago et al. (2004), a generalized equation is proposed. For utilities with similar meters and consumption characteristics as Johannesburg, the results of the new methodology are typical degradation rates of 0.7% per annum. Factoring in meter under-registration, in cases where no data are available for the weighted accuracy of new meters, the estimates of Arregui et al. (2018) can be adopted with allowances of 4–5% for the initial meter error. It has also been shown in Ncube & Taigbenu (2018) that upsizing meters, such as increasing from 15 mm to 20 mm, increases the error by up to 50%. Utilities with meter-sizing challenges should allow for another margin of error found to add about 5% to the error. The following relationship for apparent losses due to metering errors is proposed:

\[
\text{Apparent Loss Estimate} = \frac{0.7a + 4b + 5c}{100} \tag{5}
\]

where \(a\) is the average meter age, \(b\) is the estimated probability of meter under-registration and \(c\) is the probability of meter oversizing. The recommended default value of \(4b\) has been used for the second term in Equation (5) where no data are available, but where there are available data it should be replaced with its actual value. The third term is a subjective estimate of how pervasive meter oversizing is in the utility with the proposed maximum of 5% being found as the worst-case scenario for Johannesburg.

### CONCLUSIONS

Apparent losses due to metering error are estimated to be, on average, 8.2% of billed consumption compared with 9.4% found using the field-based method, with an error degradation rate of 0.7% per annum. These estimates are better than previous estimates of comparable alternative methods. For the residential category, the estimated error was 8.1% compared with 11.2% from the field and laboratory method. The difference in the values is attributed to...
meter under-registration which is not accounted for in the current method and can be higher than 4% for new meters as found in other studies. In an era where there is a need for efficient methods to estimate apparent losses, the improved comparative billing methodology developed and validated in this study over a few months is a valuable contribution as it gives comparable results at a fraction of the cost and effort of the traditional field-based method which took two years to complete.

The developed method used 15% of the available records and used the entire time series per record, which is a significant improvement from the previous attempts of related work. In addition, unsupervised identification of consumption clusters was achieved through the clustering algorithms leading to an improved assessment of the degradation rates. With better data quality, it is possible to further enhance the method for the classification of consumer segments, different metering technologies, and other categories of interest, which was not possible in this study.

Data-driven discovery processes, as demonstrated in this study, 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 comes in handy for utilities which do not have to contend with the difficulties and high costs associated with field assessments and laboratory work. However, to leverage on these emerging tools, utilities must place greater emphasis on curating reliable and clean datasets that can be mined for useful trends and information.

REFERENCES


