Data Processing Automation for Bulk Water Supply Monitoring

Arno de Coning [1, 2] [0000−0001−7960−3119] and Francois Mouton [3,4] [0000−0001−8804−7601]

[1] University of Pretoria, Pretoria, South Africa
[2] North-West University
arnodeconing@gmail.com
[3] Noroff University College, Oslo, Norway
[4] University of The Western Cape, Belville, South Africa
moutonf@gmail.com

Abstract. Water as a resource is becoming more scarce with South Africa having several provinces being struck with droughts. Up to 30% of water is lost through leaks in water distribution networks. It is common practice to monitor water usage in large water distribution networks. These monitoring systems unfortunately lack the ability to alert on high flow rates and detect water leaks unless the data is reviewed manually. The paper will explore statistical and Artificial Intelligence approaches to test the viability to detect leaks. This will can then be used as an alerting team to improve operational efficiencies of small teams and reduce repair time of leaks and thus reduces water lost through leaks.

Keywords: Artificial Intelligence, Automation, Big Data, Critical Infrastructure, Data Optimization, Leak Detection, Water Management.

1 Introduction

Water as a resource has become more scarce with 40% of the world's population living in water stressed areas - Guppy and Anderson, 2017. Fresh water has reduced by 55% from the 1960's and the forecast is that it will increase by another 50% by 2030 [4]. Economic impact equates to US$ 500 billion per annum due to water insecurity - Guppy and Anderson, 2017. Sustainability Development Goal 6 (SBG6) has been developed due to this scarcity and projections by the United Nations (UN) to work towards water security to the world that is affordable to the masses [8].

South Africa as a region has been struck with droughts over several of its provinces [2, 10]. This in turn has forced introductions of water restrictions with the hope that water supply can be maintained to the communities. The unfortunate fact is that around 30% of water losses occur from leaks in distribution networks [4, 8, 9]. Reduction in these water losses can assist in alleviating water supply in all ready stressed water regions. An additional benefit can be realized on utilities bill savings if leaks is reduced on the client side.

This paper focuses on a specific client site, a University Campus, model development to make use of water monitoring data to detect water leaks and in-crease reaction time. Currently the site has a monitoring system in place, how-ever, in its current state it does not perform early leak detection without excessive manual work. The current monitoring system has been installed to collect usage data per hour and can be displayed on a web interface. This system has around 300 monitoring station reporting to a central server but the large amount of information requires manual intervention to view each site on the system to determine if there are irregularities in the water usage. This manual intervention of the data is required to the fact that it has no alerting mechanism installed. It is unfortunate that current monitoring system still lacks the ability to intelligently alert on high consumption while technology and commonly available techniques can greatly improve the reaction time on water leaks. The authors have previously suggested methods of utilizing the monitoring data to trigger alerts on water leaks while minimizing false alarms. This data driven alerting approach is essential for small teams to manage large water distribution systems. Section 2 addresses an overview of the current system and the data that is available that can be used for the alerting system. Section 3 takes statistical approach to test the effectiveness to improve leak detection from the monitor- ing system. Section 4 covers an artificial intelligence (AI) implementation on the same dataset and performs a comparison of the results with the statistical approach. This section furthermore delves into a detailed discussion on the pro- posed improvement measures and future work that can be considered. Finally, Section 5 concludes this paper with a short summary and the direction of future work that is proposed.

2 System overview

Ongoing repairs to water networks are essential requirement to ensure a continuous reduction and early detection of water leaks. Unfortunately, this is not the case in most of the implementations in industry. The focus of this paper is on the client side, the section of the water network on which the client can exert control over, of the water network. The client unfortunately only has control over their side of the water network and has to entrust the supplier to perform regular maintenance on the other side. The current system receives information from monitoring equipment installed on the water distribution network. Data is sent to a central server and trends is viewed through a web-based system with the minimum, maximum, and average flow information. Figure 1 provides an ex-ample of the system for a week's worth of data. This system has around 300 sites to be viewed manually to potentially detect leaks and this is where improvement is required as small teams cannot analyze this system constantly. The first step initially was to directly run queries on a daily basis to generate a report for nightly leak flows. This report analyses the usage of the sites between 00:00 and 05:00, during which time the site should be empty and the flows that are detected has a high likelihood of being leak flows in buildings or bulk water supply lines. Repairs can then be initiated on these sites and prioritize the actions taken by the severity of the leak. Additionally, the report includes the site name where the monitoring equipment is placed and the hourly kiloliter per hour (kl/h) usage. This is also converted into the South African

Currency (ZAR) per day and the equivalent pipe size in millimeters (mm) that would cause such a leak. The reason for this conversion is due to the water site that was used as a case study for this paper. An example of the report is indicated below in Table 1 with the information available to initiate repairs on.

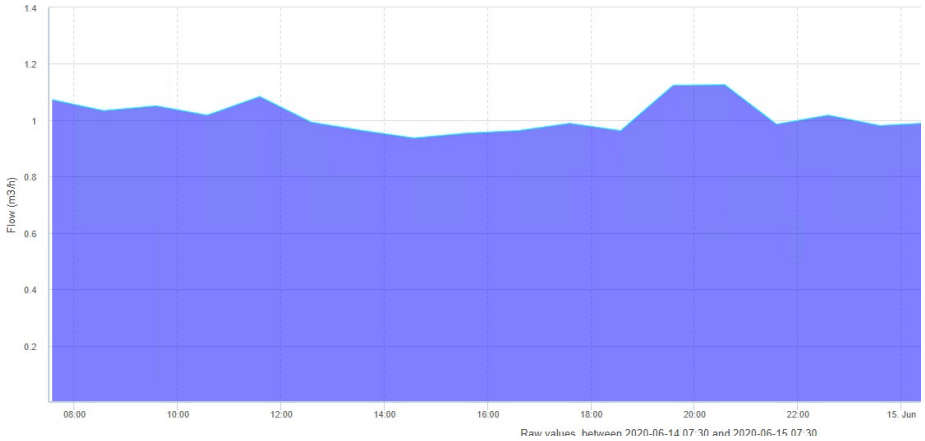

Fig. 1. Monitoring System Overview

Table 1. Example of night leak flow report

Site Number	Night leak flow (kl/h)	ZAR per day	Equivalent pipe size (mm)
Site 1	1.3148	993.36	13.64
Site 2	1.2695	959.13	13.4
Site 3	1.2533	946.89	13.32
Site 4	0.1199	90.59	4.12
Site 5	0.1011	76.38	3.78
Site 6	0.9863	745.17	11.81
Site 7	0.9626	727.26	11.67

The reported depicted in table 1 does have a positive impact on the detection of leak flows and have an extremely positive impact to reduce the reaction time of addressing the detected leaks. The problem still remains though that at the current moment, it only takes a specific time period into account and leaks outside of this period is missed. Improvement is thus required to attempt to detect leaks and send alerts to decrease manual intervention required to react. An additional aspect to take into account is that water usage trends change during the day and even time of year. Trend changes are a common occurrence in several sectors and is known as seasonality. The first step is to take a statistical approach on the data in section 3 to attempt to detect leaks. An Artificial intelligence (AI) approach is then implemented to detect anomalies in section 4.

Development of these models require some insight into the data available from the system. A site has been selected to test the leak flow detection during the years worth of data. This specific site has been selected due to a large leak that was detected with

the nigh leak flow report. An example of the data can be seen in table 2. The data for this study has been downloaded for the year of 2019 and should cover the seasonality aspects as well. A classification of the academic year can be split into six distinct sections that are used to address the seasonality of the data and the seventh to include public holidays. These current set of identified classification are:

- Class Weekday
- Class Weekend
- Exam Weekday
- Exam Weekend
- Recess Weekday
- Recess Weekend
- Public holiday

Table 2. Data from sites example

Date & Time	Flow (kl/h)
2019/01/01 00:43	0.785
2019/01/01 01:43	0.79
2019/01/01 02:43	0.795
2019/01/01 03:43	0.835
2019/01/01 04:43	0.781
2019/01/01 05:43	2.607

The system overview indicates that the monitoring system has useful in- formation but requires intelligence to adapt the system to alert on leak flows. Section 3 investigates the statistical approach to determine from the data if a leak is present.

3 Statistical model development

The first method to test is to the average the flow rate of 2019's data. This can then be used as the threshold to test the statistical approach performance in section 3.1 and is indicated as Year average in the scenario tests. Average over the dataset is 0.678 kl/h for this specific site. This average seems fairly low as time of the day is not taken into account. An average is thus calculated for each hour of the day with the result varying between 0.29 kl/h and 1.2 kl/h with the scenario indicated as Non classifier average. Results of these approaches can be seen in figure 2 and labeled as with the respective scenario names.

Both these approaches do not take the seasonality into account and the following statistical approach will be to determine the average flow rates for the specific time of day combined with the classification. The calendar for the aca-demic year is used to determine the specific dates for this classification. An hourly average is then calculated for each of the seven-day classifiers and the trends can be seen in figure 2 combined with the Year average and Non classifier average trends. A comparison of the performance is discussed in section 3.1 after the AI implementation in section 4. The accuracy

of the models require testing to determine its accuracy and if false alarms will occur or leaks will not be detected.

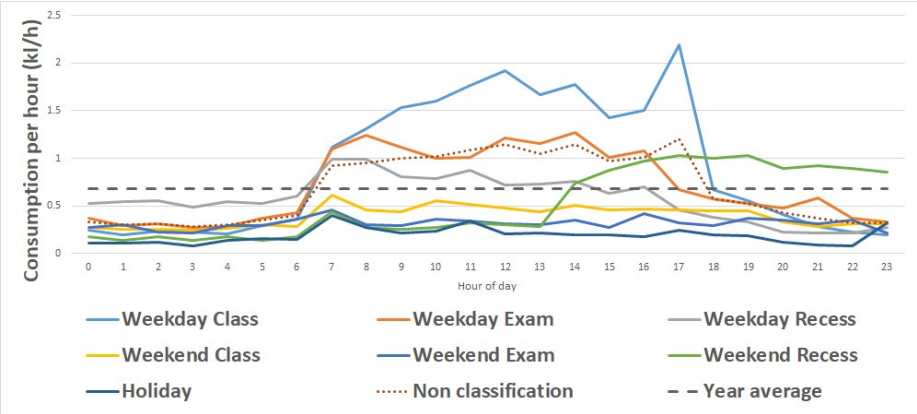

Fig. 2. Site flow statistical analysis

3.1 Model performance testing

The performance of the different models is also tested to determine the impact that they individually have on the leak flow reporting and the subsequent impact on false alarms. During testing it is required to determine how each method will cause false alarms and report on positive leak flow results when comparing against the seasonality classifier hourly data. Each test result can visually be interpreted from figure 2 with the spaces between each dataset as the leaks that would have generated false alerts on or not alerted on. The scenario tests are as follows:

- Scenario 1: Year average where it triggered above threshold
- Scenario 2: Year average where it triggered above threshold and below classification hourly rate. This is then a false alarm in the test
- Scenario 3: Year average where it did not trigger threshold and above the classification hourly rate. This is then a false positive in the test
- Scenario 4: Year average where it triggered above threshold and above classification hourly rate. This will then be a positive result for leak flow.
- Scenario 5: Non classifier average where it triggered above threshold
- Scenario 6: Non classifier average where it triggered above threshold and below classification hourly rate. This is then a false alarm in the test
- Scenario 7: Non classifier average where it did not trigger threshold and above the classification hourly rate. This is then a false positive in the test.
- Scenario 8: Non classifier average where it triggered above threshold and above classification hourly rate. This will then be a positive result for leak flow.

– Scenario 9: Classifier where it triggered above threshold that should equate to positive results for leak flow

A total of 8090 data points was available for the specific site. Each of the scenarios are tested on the available 8090 data points and the results of the testing is shown in Table 3.

Table 3. Scenario performance results

Scenario Number	Data Points	Percentage of Total Data Points
Scenario 1	2477	30.62%
Scenario 2	834	10.31%
Scenario 3	1169	14.45%
Scenario 4	1643	20.31%
Scenario 5	2900	35.85%
Scenario 6	726	8.97%
Scenario 7	641	7.92%
Scenario 8	2174	26.87%
Scenario 9	2815	34.80%

Scenario 9 indicated a total of 34.8% of the data as leak flows when compared to the statistical data and the models are compared to this approach as it include more classifier into the statistics averages. The year average approach would have incorrectly detected 10.31% as false positives and did not report on 14.45% of leaks above the classifier. This approach reported correctly on only 14.45% of the 34.8%. The non classifier approach had a decrease on the false positives with only 8.97% and a decrease on the amount it did not detect to 7.92%. This approach also increased the correct detection to 26.87% of the 34.8%.

The statistical approaches can improve the leak detection rate and the per- formance increased when taking the classifiers into account that gives a better reflection of the seasonality. Section 4 investigates the AI application on the same dataset to determine the potential improvement on the statistical models.

4 AI implementation and results

AI implementation has increased in the recent years. There are several use cases in industry and this section tests the accuracy when implementing AI methods on the flow data to detect leak flows. The training is based on supervised training techniques to tests the best performance. These models make use of the time series data as used in section 3 with the day classification as the first classifier and the time of day as the second classifier data for the model input. The following approaches are implemented, and performance measured to test the viability of future implementation:

- Approach 1: Supervised training with only the classifiers to predict water usage for the specific hour. The alert trigger will then be if the current usage is above the predicted value.
- Approach 2: Supervised training with the statistics approach average usage per classifier as inputs and predicting the output value. The alert trigger will then be if the current usage is above the predicted value.
- Approach 3: Artificial Neural Network (ANN) classifier implementation if the values are above the statistic seasonality data it is classified as a leak and below not a leak.

4.1 Data preparation

Data preparation is an essential step prior to input into any AI model. All three of the approaches require the classifiers as input to ensure the seasonality is taken into account for the model performance. These columns are label encoded to take the 7 day classifiers and 24 hour classifiers into a integer value to be used [3]. Output of this step results in a 2-column array with day classifiers values from 0 to 6 and the hours from 0 to 23 as indicated in table 4.

Table 4. Label encoded result

Label of day classifier	Label of hour classifier
0	20
0	21
0	22
0	23
3	0
3	1
3	2

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Figure 3 OneHotEncoded results

The values are then OneHotEncoded to split these classifier values into its own column and this output results is a 31 column array. This step assists in the model not adding a higher importance to the higher label encoder data value as each column can only be zero (0) or one (1) as the output [3]. Figure 3 indicates the hour values in each of its individual columns with the fist column value as hour 00:00 and then followed by hour 01:00 for the next index.

Approach 1 and 2 make use of the flow values as output that the model predicts. Approach 2 takes the statistics data as input with the classifier data. Approach 3 takes the same values as Approach 2 but instead of the flow data as output it has a list of zeros (0) and ones (1) where the zero (0) occurs if the flow value is below the statistic value and one (1) if it is above the value.

4.2 Supervised training implementation

The supervised training implementation is tested with several *sklearn* models on the dataset. Approach 1 and 2 makes use of regression models to predict the expected output and compare the current usage to determine if a leak is present. The models tested is linear regression, Gradient Boosting regression, Random Forest regression, KNeighbor Regression, Support Vector Regression (SVR) [5, 7, 3]. The regression models are tested for both approach 1 and 2 as they have different input datasets. Accuracy of the prediction is calculated by the r^2 test with 1 being the best accuracy. The results of the model testing can be seen in table 5. The accuracy scoring is quite low on these implementations and is discussed in section 4.4.

Table 5. Supervised regression results

Model	Approach 1 - r^2 score	Approach 2 - r^2 score
Linear regression	0.156	0.205
KNeighbor regression	0.021	0.021
Random Forest regression	0.178	0.167
Gradient Boosting regression	0.186	0.186
Support Vector regression	0.183	0.183

4.3 Classifier implementation

The classifier implementation requires a output result set with specific false (no leak) or true (leak) values. A model is then trained to predict this output value where the previous regressors predicted the actual flow data for the specified day and hour classifiers. The data output change is split between the false and true output when the flow in the input dataset is above the statistical values in section 3.1. Two different approaches are tested to compare the performance. The first approach is to implement a Support Vector Classification (SVC) and then an ANN implementation, as a second approach. This accuracy is then determined by generating a confusion matrix which indicates *True positives* (top left), *True negatives* (bottom right), *False positives* (bottom left), and *False negatives* (top right) [1, 5, 7]. The SVC implementation had an accuracy of 90.6% with the output confusion matrix in equation 1. The ANN implementation had a higher accuracy at 97.78% with the confusion matrix in equation 2.

$$\begin{matrix} & \mathbf{1} & \\ 1046 & & 12 \\ 140 & & 420 \end{matrix} \tag{1}$$

$$\begin{matrix} & \mathbf{1} & \\ 1041 & & 17 \\ 19 & & 541 \end{matrix} \tag{2}$$

4.4 Results discussion and future work

The implementation of the classifier models had a theoretical improvement on the reaction time. The classifier led to a water leak detection accuracy of 97.78%. It is specifically stated as a theoretical improvement based on the fact that several improvements can be made to the model to give a true reflection of the leak flows. Implementation of the regressors has had very low accuracy and this can be attributed to multiple factors that also needs to done to further improve upon the classifier approach.

The statistical models with day and time classification has improvement over the fixed threshold alerting but the years worth of data could have leak con- stantly skewing the data. A process is thus required to log specific leaks and time span for a site to be used in conjunction with flow data to either exclude the data or reduce the flow by the leak amount to improve accuracy of the sta- tisti-cal approach. This will further benefit the AI approaches as the statistical mod-els are used to predict the leak flows. An additional step can be taken to reduce the flow data by the night leak flow data to ensure better accuracy to predict the leak flow data. The regressor models have the low accuracy as it has to few input variables to predict the usage. Water usage is commonly attributed to the amount of people in a building and this can greatly assist in the prediction process. Future work should thus be to introduce occupancy data as input to the model to improve accuracy of predictions and in turn leak detection. In addi-tion, one should also have a look how social engineering attacks could have an impact on water monitoring systems [6].

5 Conclusion

Water as a resource has become more scarce with 40% of the world's population living in water stressed areas [4]. The unfortunate fact is that around 30% of water losses occur from leaks in distribution networks [8]. Current monitoring systems lack the ability to intelligently alert on leaks within the system without large amount of manual intervention to review the data. A data driven approach is thus proposed to analyze the data to detect leaks from the monitoring system and then to alert relevant personnel to take action to repair the leak. The automation of data analysis will assist in improved

reaction times on water leaks correction by small management teams that would have required several man hours per day to detect.

The first approach that was performed was a statistical approach that deter- mines the average flow over a year dataset for a year. This was further adapted determining the average flow per hour of the day as the trends change during the day. Finally, the statistical analysis approach was adapted to take the hour of the day and the time of year classifier into account as seasonality also has severe impact on the usage. The second approach was to test AI models to firstly predict the usage for the hour to determine if a leak is present. Regressor implementation was used to predict the usage based on the flow data from the monitoring system with hour of the day and time of year information. This was then further adapted to test the implementation of an ANN classifier model to determine if a leak is present.

The model that performed the best with current testing was the ANN model classifier with 97.78% accuracy when combining the statistical data that includes the time of day and time of year classifier information. This model has room for improvement as the statistical model currently may include leak flows in that can potentially skew the results. The models can also benefit by the inclusion of additional input parameters such as building occupancy data. Expansion of data sets will assist in improving model performance while minimizing potential class imbalance.

It is proposed that the current water management policy should be enforced that would have assisted in accurate logs of leak flows as they are detected with the duration and severity. This will then assist in model training while this leak can then be removed the data to improve the input data to the model as well. The authors are planning to conduct a further study on the impact of non-compliance on current water management policy.

References

1. Alwis,R.: Introduction to confusion matrix [classification modeling]. https://medium.com/tech-vision/introduction-to-confusion-matrix-classification-modeling-54d867169906
2. Baker, A.: What it's like to live through cape town's massive water crisis, https://time.com/cape-town-south-africa-water-crisis/
3. Boschetti, A., Massaron, L.: Python data science essentials. Packt Publishing (2015)
4. Guppy, L., Anderson, K.: Global water crisis: The facts. University Institute for Water, Environment and Health pp. 1–16 (September 2017)
5. Joshi, P.: Artificial Intelligence with Python. Packt Pub-lishing (2017), https://books.google.no/books?id=O1AoDwAAQBAJ
6. Mouton, F., Teixeira, M., Meyer, T.: Benchmarking a mobile implementation of the social engineering prevention training tool. In: Information Security for South Africa (ISSA). pp. 106–116 (Aug 2017). https://doi.org/10.1109/ISSA.2017.8251782 (2017)
7. Raschka, S.: Python machine learning. Packt Publishing (2015)
8. The United Nations: Goal 6: Ensure access to water and sanitation for all., https://www.un.org/sustainabledevelopment/water-and-sanitation/
9. The Water Project: Water in crisis - south africa., https://thewaterproject.org/water-crisis/water-in-crisis-south-africa

10. Welch, C.: Why cape town is running out of water, and who's next., https://www.nationalgeographic.com/news/2018/02/cape-town-running-out-of-water-drought-taps-shutoff-other-cities/