Pattern extraction and profiling of historical water network demand patterns

submitted in partial fulfillment for the degree of master of science

Marten Staghouwer
12851868

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on Fill In The Date In Format DD.MM.YYYY

	UvA Supervisor
Title, Name	Viktoriya Degeler
Affiliation	UvA Supervisor
Email	v.o.degeler@uva.nl

ABSTRACT

Write your abstract here.

KEYWORDS

pattern extraction, clustering, water network, time series

GITHUB REPOSITORY

https://github.com/Mjnstag/IS-thesis-pattern-extraction

1 INTRODUCTION

Water distribution networks play an important role in making sure everyone has access to water. Due to their size and complexity it can be difficult to have a complete manual overview of the network. To help combat this issue, automated sensors can be used to measure different aspects like pressure and flow in different parts of the network. The data of these sensors can be used to gather insights in the current state of the network, but could also be used in predicting what the next state of the network will be. Having this information in advance could help the network adapt to changes more easily.

Having a lot of different data points and input variables, could pose problems for getting accurate classifications when using clustering methods [1]. There are different ways to help with this like using less inputs, though this might not be favorable. Another method could be reducing the data's dimensionality using methods such as principal component analysis in which patterns are tried to be found and data inputs which do not correlate or influence much, be removed or condensed into less variables.

A water network which uses sensors to capture its current state will most likely have too many inputs. This study therefore wants to see if the aforementioned method of reducing dimensionality can help in forecasting how the network will behave by answering the question: To what extent can pattern extracting help in forecasting historical water network data.

To help answering this question, multiple sub questions are needed to make clear certain parts like how methods like principal component analysis (PCA) and k-means work and if there are any gaps or disadvantages to using these methods. Due to the many available clustering methods, it is important to better understand the nature of these methods and which are actually relevant and usable in this context, while also knowing how the different methods compare to each other. Due to the graph structure of the data, it will also be important to find a way in which the data can work with clustering methods.

This results in the following sub questions:

- How do already existing pattern extraction methods work and are there any gaps.
- Which clustering methods can capture the behaviour profiles from the water network?
- How do the different methods compare?
- How can a water network's graph structure be transformed to a format suitable for clustering methods?

2 RELATED WORK

3 METHODOLOGY

There have been different uses for clustering in research related to water networks, from using clustering to calibrate models to dividing water networks in different sectors [2, 3]. This paper builds on already existing use-cases of these methods to propose a framework in which clustering methods are used to extract patterns from historical water network data. The goal is to find out if this use-case of clustering if viable and feasible in predicting changes and trends in water networks. For this, multiple different unsupervised clustering algorithms were used in an experimental setup where the model's hyperparameters were automatically optimized and the final results compared against each other and a base model.

3.1 Experimental Setup

3.1.1 Data Gathering.

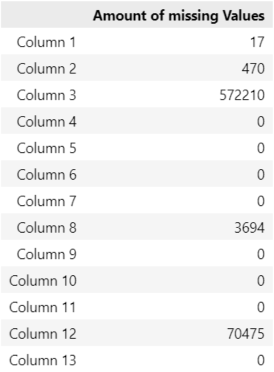
For this paper one dataset was used. The dataset comes from Vitens[1], a Dutch water network company, and contains data about a subset of the total water network they manage. This dataset is the one primarily used and tested with. The dataset has 17 columns and mainly contains sensor data from a few different points of the network. The measurement interval for the sensors was 1 minute, which when combined with a collection period between 2017 and 2022, results in a total of 3.067.200 rows.

3.1.2 Data Description.

3.1.3 Data transformation.

Even though the dataset has 17 columns, not all of them are useful and contain irrelevant data. This is due to the dataset also being used for other projects and purposes. After filtering out these columns, 14 columns remained.

Due to the data having the Central European Time (CET) timezone, the data was first converted to Coordinated Universal Time (UTC) to avoid conflicts with regards to daylight saving time.



	Amount of missing Values
Column 1	17
Column 2	470
Column 3	572210
Column 4	0
Column 5	0
Column 6	0
Column 7	0
Column 8	3694
Column 9	0
Column 10	0
Column 11	0
Column 12	70475
Column 13	0

Figure 1: Amount of missing values per column.

[1] https://www.vitens.nl/

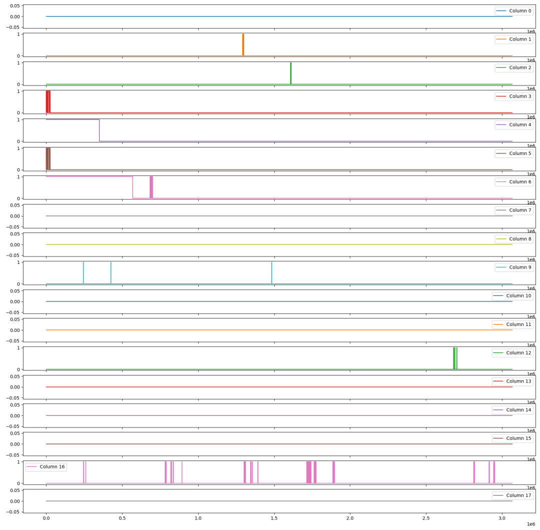


Figure 2: Missing values in the data where a spike means a missing data point.

Throughout the data there are many instances of missing values. Following the EDA, 2 different kinds of missing data were found: Leading missing data and sporadically missing data throughout the entire length of the time series, see image 2.

In both of these cases an attempt was made to fill the missing values by interpolating them linearly. To not change the behaviour of the data too much, a limit of 20 was set on how many consecutive missing values could be interpolated. In the case that there were more than 20 consecutive missing values, these extra missing values were not interpolated and filled.

To prepare the data for clustering, the columns were split into new data frames. This resulted in 14 new data frames where there was 1 column which contained all that column's rows from the original dataset. To enable clustering, for each of these data frames, the rows were split into segments of 1 day. Each of these segments were added as new columns resulting in a dataframe with 1440 rows of sensor data and an amount of columns based on the amount of data being transformed.

After further splitting the dataframes, all columns were checked if there were any missing values remaining. If this was the case, that column was removed from the dataframe and not used.

Finally, one of the limitations of the models are the potential computational resources needed to use the models. To lessen this concern, the amount of data in the final data frames was reduced by only having a data point 5 minutes instead of every 1 minute. This was done by taking the first of these datapoint while discarding the rest. This limits the data needing processing and thus reducing the computational resources needed for running the different models. After this, the remaining columns were then scaled using a min-max scaler.

3.1.4 Dimensionality Reduction and Clustering Methods.

As mentioned, the goal of this paper was to find out if clustering algorithms can help in extracting patterns in water network data. This was done by using and testing different clustering algorithms like: Principal Component Analysis (PCA), K-means, Self-organising maps (SOM), and DBSCAN. These algorithms were used through python's SKlearn library as well as various other affiliated libraries. Due to this, it is not needed to create these algorithms from scratch.

After clustering the data, an analysis was done on the composition and the distribution of the final clusters. This was done by looking at the distribution of columns per cluster and, if needed, filtering cluster which consists of only 1 column due to outliers in the data.

3.1.5 Evaluation.

To evaluate and compare the different clustering algorithms, a base algorithm was chosen: K-means. This algorithm was chosen due to its widespread usage and its simplicity. To evaluate the baseline and other chosen algorithms, multiple different aspects will be used. This includes the accuracy, precision, recall, and F1 as a basis to be combined with the parameters of the clusters itself. This includes the final amount of clusters as well as the distribution of all the columns in the final clusters.

To train, test, and validate the models, the data is split in 80% of the data in the training set, 10% in the validation set, and finally 10% used for the test set. The data being split is comprised of the new columns created from the complete dataset.

4 RESULTS

5 DISCUSSION

6 CONCLUSION

REFERENCES

[1] Shaeela Ayesha, Muhammad Kashif Hanif, and Ramzan Talib. 2020. Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information fusion* 59 (2020), 44–58.

[2] Xinran Chen, Xiao Zhou, Kunlun Xin, Ziyuan Liao, Hexiang Yan, Jiaying Wang, and Tao Tao. 2022. Sensitivity-Oriented Clustering Method for Parameter Grouping in Water Network Model Calibration. *Water Resources Research* 58, 5 (2022), e2021WR031206.

[3] Armando Di Nardo, Michele Di Natale, Carlo Giudicianni, Dino Musmarra, Giovanni Francesco Santonastaso, and Antonietta Simone. 2015. Water distribution system clustering and partitioning based on social network algorithms. *Procedia Engineering* 119 (2015), 196–205.

Appendix A FIRST APPENDIX

Put your appendices here.