Pattern extraction and profiling of historical water network demand patterns

Submitted on: **19-02-2023**

Marten Staghouwer
marten.staghouwer@student.uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Viktoriya Degeler
v.o.degeler@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Github

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 [3]. 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 PCA 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

2.1 Clustering Methods

Clustering can both be used as a supervised as well as an unsupervised method, though during this thesis the focus will be on the unsupervised versions. The method can be used to do a multiple of things including machine learning, data mining, and pattern recognition. It works by grouping, and therefor splitting, data into different subsets. When a new point of data falls within the range of such a group, it is labelled as a member or subset of that group [4].

There are 2 main types of clustering: hierarchical clustering and partitional clustering techniques. Hierarchical clustering consist of top-down and bottom-down methods where top-down clustering starts with only 1 cluster which all data points fall under, after which it is divided into smaller clusters. Bottom-down does the reverse, starting with each data point as a different cluster, merging them together to form bigger clusters [4]. An example of this is the BRICH algorithm.

The second type of clustering methods, partitional clustering, start by dividing the data into number of clusters set using a parameter. They divide the data by minimizing a specified function like mean squared error [5]. A well known algorithm of this kind is the K-means clustering technique.

2.2 Dimensionality Reduction

The rising amount of data points and features, also called the dimensionality of the data, can pose challenges for clustering methods and can lead to computational and accuracy problems [3]. In the case where the data is highly dimensional, reducing this dimensionality can help in alleviating the aforementioned problems. One of the methods in which this can be achieved is by using the principal component analysis (PCA). The PCA is an algorithm which reduces the dimensionality, while preserving as much variety of the data, by creating different new variables, called components, which each represent a part of the original features. By using a few of these components, it is possible represent most of the original features, thus reducing the dimensionality [6].

Another method closely related to PCA is common factor analysis (CFA). It differs from PCA in that CFA only analyzes the common variance of data, while PCA analyzes all the variance of data. Due to this difference, running both on the same dataset can give different results.

In python's scikit-learn package there also is a dimensionality reduction function which works on the basis of clustering, named

Unsupervised dimensionality reduction. This function uses clustering to group similar features, resulting in a decrease of the number of features and dimensions.

2.3 Explanatory Data Analysis (EDA)

During this thesis 2 datasets will be used. The first of these datasets is a public dataset called LeakDB [9]. The data in this dataset is artificially created but is an accurate representation of reality. It includes 1000 different scenarios from a water network in Hanoi under varying conditions. This water network is represented in a geographical graph structure which contains pipes, joints, reservoirs, and consumption points, see image 1 [7]. These data points have a time series csv file which contains data about each node and pipe, connection between nodes, in the network. This includes demand, pressure, and flow. The time series contains a data points with 30 minute intervals over the course of a year, for a total of roughly 17.500 data points for each time series.

The second dataset is a private dataset from Vitens [8] and contains historical sensor data from the water network in the Dutch province of Gelderland. This dataset closely matches the public dataset with regards to content, behaviour, and structure. This means that minimal revisions will be needed when adapting the processes from the public dataset to this private dataset.

Due to the open-source nature of the LeakDB dataset, access to this data is already available. With regards to the private dataset, access will be gained on February 24th 2023. This will be combined with a workshop during which there will be a more in-depth explanation of the data itself as well as different tools which can be used to visualize and interact with the data.

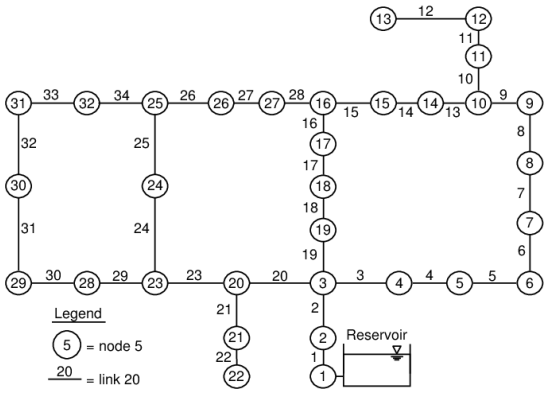

Figure 1: Hanoi water network's graph structure.

3 METHODOLOGY

Within this thesis project there will be multiple different steps. Each of these steps have their own associated software, data, and methods.

3.1 Understanding the Data

For getting a deeper understanding of the data, a few different pieces of software can be used. These are EPANET and python with a few different libraries. EPANET is a piece of software on which can be used to visually inspect the layout and location of the different sensors. It also enables running different simulations on the network and showing how parameters like demand and pressure for each node change. This feature can be used to create more synthetic data if needed. Python can be used for the same purpose while also giving the possibility to better visualize and combine data. A python library called epynet is available to be able to connect and interact with the models [2].

3.2 Reducing the Data Dimensionality

The next step is researching and understanding the different methods which can be used for clustering and reducing the dimensionality of the data. These methods will probably mostly come from python's scikit-learn package, which includes different clustering and dimensionality reduction methods [1]. Using the understanding of the data, a plan can be made on which clustering methods should be considered for testing.

Since there are many different features in the data, a reduction in the data's dimensionality is assumed to be needed. As shown in 2.2, different methods for this exist. Because of this, tests will be done which will compare different methods and which parameters give the best results. During this section of the project, PCA and its variations will be tested as well as similar methods like CFA. These will be tested using python and the scikit-learn package.

3.3 Clustering and Results

While reducing the data and clustering can be seen as separate phases, they are closely related. Because of this, there are a few different possibilities in how this step will go, depending on how the clustering step goes. One of the ways is that it follows each other, meaning that first the dimensionality reduction will be done, after which the clustering will be done. Another way is that they are done simultaneously so that the effects of different dimensionality reduction methods can be seen on the chosen clustering methods. As mentioned before, these different methods will be gotten from python's scikit-learn package.

By testing a combination of different data dimensionality reduction methods and different clustering methods, it is important to have an understanding of how well they perform in relation to each other, but also with a ground truth. For this evaluating step, a ground truth has to be created. Such a model could be a naive base model which does not use any extra methods which help in forecasting.

4 RISK ASSESSMENT

There will be a few risks associated with this thesis. It is assumed that these potential delays will mostly come from the use of multiple datasets.

One of the problems which could arise lies in the assumption that there are no major differences between the content and the behavior of the public and private datasets. Depending on which phase of the project this is discovered, this might mean different things. If it is discovered early into the process, both datasets can either be used at the same time, or a choice can be made on which dataset the focus should be. This probably will result in general timetable changes. When it is discovered later into the process,

the choice might need to be made to only focus on the public data depending on how much time is left. While this will negate part of the thesis, the thesis still would be useful for this more general set of realistic scenarios.

No problems are expected to arise with regards to access to the data. The public dataset has already been downloaded, so even if access online gets removed, a copy still exists. As mentioned before, a workshop will be given to explain the private dataset. While the chance exists that this workshop will be delayed or cancelled, it is not needed to get access to the data. This access can be gained at any time with help from the thesis supervisor, but has been chosen to be delayed to the workshop. A minor problem which could follow a cancellation or delay to the workshop is that it could delay the understanding of the data, though this would most likely result in at most a delay in the data exploration and transformation stages of the project.

Another problem could lie in finding the correct algorithms to use. This could be problematic due to the relatively high volume of algorithms available both in python packages, or in the literature. While it is assumed that many can be discarded based on use case and specialty when the data is better understood, it cannot be excluded that this process might take longer than expected. This potential problem can be mitigated to take longer researching the different methods available to get a manageable selection.

5 PROJECT PLAN

5.1 General Overview

Starting from the final deadline for handing in the thesis design. February 19th 2023, there are 19 weeks until the final deadline for handing the final version of the thesis, June 30th. 2 mandatory courses are are still being attended during the remainder of the 4th period of this masters. This results in a period of roughly 6 weeks where the thesis will be worked on part-time with the remainder intended as full-time.

5.2 Mandatory Milestones

During the thesis there are several mandatory milestones which have been set by the university. These can be seen as bold text in table 1. The deadlines for these milestones cannot be changed. The point of these moments is to make sure that you are on schedule as well as getting feedback from other students. For that reason, a moment after most of these milestones has been made to rework any section as needed based on feedback. For the self-set milestones, while a rough planning has been made, it should be expected that this can change depending on the current situation, potential challenges and delays, and progress with regards to the project.

Week	Date	Achievement
0	19-2-2023	**Deadline thesis design**
1		Data exploration
	24-2-2023	Private data access + workshop private data
2	5-3-2023	Data exploration
3		Writing EDA
4	15-3-2023	Writing EDA
		Buffer writing EDA
		Data processing & transformation
	19-3-2023	**Deadline EDA**
5	24-3-2023	Data processing & transformation
		Writing Methodology
6	1-4-2023	Writing Methodology + buffer
7	6-4-2023	**Draft Methodology section**
	10-4-2023	Rework methodology section based on feedback
8		Clustering
9	23-4-2023	Clustering
10		Evaluation
11	7-5-2023	Evaluation
		Write results section
12		Write results section
13	16-5-2023	Done result section
	21-5-2023	**Draft results section**
14	24-5-2023	Rework results section based on feedback
	28-5-2023	Discussion done
15	1-6-2023	Conclusion & future works done
		Buffer for all writing
16		Buffer for all writing
17	18-6-2023	**Draft complete thesis**
18	23-6-2023	Rework thesis using feedback
		Buffer for making changes
19	30-6-2023	**Final deadline thesis**

Table 1: Project timeline with achievements.

REFERENCES

[1] [n. d.]. scikit-learn: machine learning in Python — scikit-learn 1.2.1 documentation. https://scikit-learn.org/stable/. (Accessed on 15-02-2023).

[2] [n. d.]. Vitens/epynet: Object-oriented wrapper for EPANET 2.1. https://github.com/Vitens/epynet. (Accessed on 02/08/2023).

[3] 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.

[4] T. Soni Madhulatha. 2012. An Overview on Clustering Methods. (2012).

[5] Mahamed G.H. Omran, Andries P. Engelbrecht, and Ayed Salman. 2007. An overview of clustering methods. *Intelligent data analysis* 11, 6 (2007), 583–605.

[6] Markus Ringner. 2008. What is principal component analysis? *Nature biotechnology* 26, 3 (2008), 303–304.

[7] Lina Sela, Ariel Krapivka, and Avi Ostfeld. 2009. Single and multi-objective optimal design of water distribution systems: Application to the case study of the Hanoi system. *Water Science & Technology: Water Supply* 9 (10 2009). https://doi.org/10.2166/ws.2009.404

[8] Vitens. [n. d.]. *Homepage.* (Accessed on 14-02-2023).

[9] Stelios G Vrachimis, Marios S Kyriakou, et al. 2018. LeakDB: a Benchmark Dataset for Leakage Diagnosis in Water Distribution Networks:(146). In *WDSA/CCWI Joint Conference Proceedings*, Vol. 1.