BTO 2019.041 | October 2019

BTO report

Implementation of Data Mining at Water Utilities

BTO

Implementation of Data Mining at Water Utilities

BTO 2019.041 | October 2019

Project number
402045

Project manager
Dr. Geertje Pronk

Client
BTO - Verkennend onderzoek

Quality Assurance
Christos Makropoulos

Author(s)
Henk-Jan van Alphen MSc, dr. ir. Dirk Vries, dr. Joost van Summeren, Masja Bronts MSc, Erwin Vonk MSc, dr. ir. Martin Korevaar, dr. Peter van Thienen

Sent to
BTO Directeurenoverleg

Year of publishing
2019

More information
MSc, Henk-Jan van Alphen
T 0655281776
E Henk-Jan.van.Alphen@kwrwater.nl

Postbus 1072
3430 BB Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
F +31 (0)30 60 61 165
E info@kwrwater.nl
I www.kwrwater.nl

Keywords
Hydroinformatics, datamining, data management

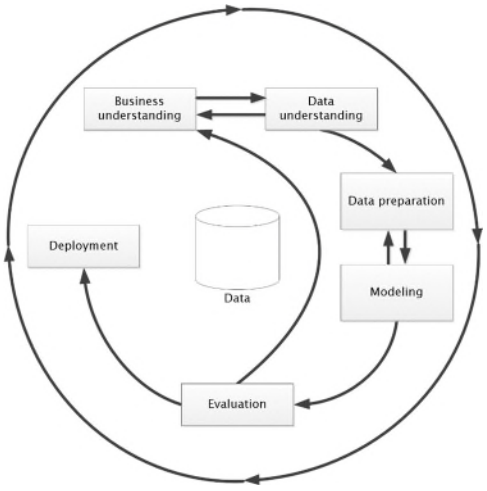
BTO *Managementsamenvatting*

Implementatie van dataminingtechnieken bij waterbedrijven

"Datakwaliteit en multidisciplinaire samenwerking cruciaal voor succesvolle implementatie van dataminingtechnieken bij waterbedrijven."

Auteurs: Henk-Jan van Alphen MSc, dr. ir. Dirk Vries, dr. Joost van Summeren, Masja Bronts MSc, Erwin Vonk MSc, dr. ir. Martin Korevaar, dr, Peter van Thienen.

Dataminingtechnieken staan onder toenemende belangstelling van waterbedrijven als een manier om informatie uit data verkrijgen en besluitvorming te ondersteunen. Uit een eerder onderzoek bleek dat de grootste uitdagingen bij de implementatie van datamining van organisatorische aard zijn. Daarom zijn in dit onderzoek drie pilot projecten uitgevoerd, waarbij in elk van de pilots de volledige dataminingketen bij een waterbedrijf is geïmplementeerd. Hiermee is al doende inzicht verkregen in de succes- en faalfactoren van dataminingprojecten bij waterbedrijven. Uit deze pilots bleek dat datakwaliteit en –beschikbaarheid de meest cruciale factor is voor succesvolle implementatie. Multidisciplinaire samenwerking is eveneens onontbeerlijk, omdat dat leidt tot een beter begrip en interpretatie van de resultaten. De belangrijkste aanbevelingen voor de waterbedrijven zijn het verbeteren van data management en ICT-ondersteuning en het ontwikkelen van een bedrijfsbrede visie op data.



CRISP-DM proces model voor datamining

Methode: Opzetten van dataminingketen bij drie waterbedrijven.

Bij drie waterbedrijven (Oasen, PWN, Waterbedrijf Groningen (WBG)) is een pilotproject uitgevoerd, waarbij de volledige dataminingketen is geïmplementeerd. KWR onderzoekers en externe datawetenschappers richtten zich op datavoorbereiding, methoden en prototyping in samenwerking met domeinexperts die zich op interpretatie, toepassing en verbinding van de resultaten met de behoeften van de organisatie richtten. ICT-medewerkers richtten zich op de implementatie van software binnen de bestaande systemen. De

drie projecten zijn geobserveerd en geëvalueerd door een onderzoeker van KWR, om inzicht te krijgen in de succes- en faalfactoren van data-miningprojecten bij waterbedrijven.

Belang: Datamining kan nieuwe informatie opleveren en beslissingen ondersteunen

Machineleren en datamining worden breed toegepast in diverse sectoren (marketing, gezondheidszorg, mobiliteit, sociale media) om nieuwe informatie te genereren, gedrag te voorspellen en sturen en beslissingen te ondersteunen. Dit zal naar verwachting in de toekomst toenemen door de stijgende hoeveelheid data (bijvoorbeeld van sensoren en robots) en nieuwe methodes om die data te analyseren. Waterbedrijven verzamelen al data uit hun infrastructuur en interacties met klanten (zoals facturering) en zoeken naar manieren om die data te gebruiken om besluitvorming te ondersteunen. Uit het eerste deel van dit onderzoek bleek dat de uitdagingen die de waterbedrijven daarbij ervaren niet zozeer van methodologische aard zijn, maar vooral met organisatorische aspecten te maken hebben. Daarom richt het tweede deel van deze studie zich op de praktische implementatie van de complete keten: van dataverzameling en -validatie tot analyse en implementatie.

Results: Success and fail factors for data mining projects at water utilities

De drie pilots hebben een vergelijkbare aanpak gevolgd, maar de uitkomsten verschillen in de wijze waarop ze strategisch en operationeel bruikbaar zijn. Deze verschillen zijn grotendeels toe te schrijven aan de beschikbaarheid en kwaliteit van data en het inzicht in het onderzochte vraagstuk. In elk van de drie pilots werd multidisciplinaire samenwerking op de locatie van de klant als belangrijkste succesfactor genoemd. Specifiek werd daarbij op de interactie tussen datawetenschappers en domeinexperts gewezen. Het is moeilijk voor te stellen hoe de pilots succesvol waren geweest als alleen de data was gedeeld met de datawetenschappers en de analyse was uitgevoerd zonder interpretatie of feedback door de domeinexperts.

Een cruciale stap in de dataminingketen is het vertalen van de analyseresultaten naar informatie die bruikbaar is voor beslissingsondersteuning. Ten slotte is het belang van de rol van de ICT-afdeling onderstreept. Die kan toegang verlenen tot interne systemen en databronnen en kan een virtuele ruimte creëren om de analyses uit te voeren. Bovendien heeft ze een belangrijke rol bij het implementeren van een eventuele tool in de systemen van het waterbedrijf.

Implementatie: Aanbevelingen voor toekomstige dataminingprojecten

Op basis van de pilots wordt aanbevolen dat de waterbedrijven hun datamanagement verbeteren, zodat de data die ze verzamelen en beheren van consistente kwaliteit is en geschikt is voor een breed scala aan analyses, bijvoorbeeld door het toevoegen van voldoende metadata.
Waterbedrijven zouden verder moeten gaan met het ontwikkelen van een bedrijfsbrede visie op data, met de betrokkenheid van alle delen van de organisatie die op een of andere manier met data te maken hebben. Bij het uitvoeren van concrete dataprojecten moet goed bepaald worden wat het doel van het project is. Als implementatie het primaire doel is, bepaal dan vooraf waar de tool moet worden geïmplementeerd en maak direct de goede verbindingen met de databronnen. Maak ook vooraf een inventarisatie van de datakwaliteit en -beschikbaarheid en bepaal op basis daarvan of het project door kan gaan. Kies bij dit soort projecten een iteratieve aanpak, waarbij op basis van deelresultaten vervolgstappen worden gekozen. Dit in combinatie met een duidelijk overzicht van beschikbare middelen en te besteden tijd. Het is verder aan te bevelen om op de locatie van de opdrachtgever te werken in nauwe samenwerking met domeinexperts, probleemeigenaar, data-eigenaar en eindgebruiker. Dat maakt het makkelijker om problemen op te lossen, informatie te verkrijgen, resultaten te delen en bewustwording te creëren. Zorg daarbij dat onderzoekers van buiten makkelijk toegang hebben tot interne systemen en data en zorg voor een geschikt platform om de analyses uit te voeren.

Report

This research is reported in *Implementation of Data Mining at Water Utilities* (BTO-2019.041)

More information
MSc, Henk-Jan van Alphen
T 0655281776
E Henk-Jan.van.Alphen@kwrwater.nl

KWR
PO Box 1072
3430 BB Nieuwegein
The Netherlands

More information
MSc, Henk-Jan van Alphen
T 0655281776
E Henk-Jan.van.Alphen@kwrwater.nl

KWR
PO Box 1072
3430 BB Nieuwegein
The Netherlands

Contents

1 Introduction

1.1 Context and motivation

Techniques for extracting knowledge from (combinations of) databases, often presented under the flags of ‘machine learning’ and ‘data mining’, have shown significant progress over recent years. Many such techniques are already being used every day in all kinds of contexts (often without our being aware of it). Also, more and more data is being collected, both in general and specifically by the water companies. This is expected to only increase in the future (through, for example, developments in sensors and robotics). Initial attempts have been made at applying these techniques in the water sector with the objective of 'obtaining more insight from the available data'. However, these attempts have not yet produced ‘revolutionary’ results. That is not to say that results to date have not been interesting. These results make clear that there are significant opportunities for the application of data mining techniques in many areas in the drinking water chain, from source to tap.

In the context of the BTO exploratory research project *VO datamining*, we aim to provide an overview of opportunities for the water companies, to offer a perspective on the successful implementation of these techniques and to support the water companies in their choices in this respect. More concretely, this entails 1) scouting approaches in the water sector and other sectors, 2) identifying a number of approaches with the most potential for fruitful application in the water sector, and 3) create three applications/demonstrations within the water sector.

In the first part of this exploratory study, a wide inventory has been made of existing data mining techniques (such as clustering, classification, regression, deep learning) and their applications in other fields (outside the water sector), as well as a broad inventory of internal and external data sources available within the water sector (now and in the near future). Subsequently, promising current and emerging applications (combination of techniques and one or more data sources) have been identified, on technical grounds, but also on the basis of the expected information yield to support decision processes at water companies. The results of both activities have been reported in Van Thienen et al. (2018). An important conclusion of this first part was that the challenges in applying data mining techniques are not so much of a methodological nature, but more in the practical application, presumably in the context of organizational and ICT aspects. Therefore, the second part of this study, focuses in the practical implementation of the complete chain: from data gathering, to validation, analysis and implementation.

1.2 Approach

Three pilot projects were initiated at three water utilities, Oasen, Waterbedrijf Groningen (WBG) and PWN. These companies offered to host and participate in an implementation pilot after an invitation which was sent to the participants of the BTO Hydroinformatics Platform.

For each of the pilot projects one or two cases were selected for analysis and implementation. Also, for each pilot project a project team was formed consisting of

researchers from KWR, domain experts from the water utility, external data scientists, information managers, ICT and a decision maker. The mission of the project teams was to set up a data mining chain with a specific application in mind and with the perspective of implementation as an operational tool. KWR researchers and external data scientists focused on data preparation, methods and prototyping; domain experts from the water utilities focused on interpretation and applicability and aligning with the needs of the organization; ICT focused on software implementation and integration within existing systems. The work was conducted at the respective water utilities offices. Intermediate results were shared during monthly progress meetings with all relevant participants. For the three pilots the following people contributed to this project:

- **PWN**: Martin Korevaar (KWR), Laurens van der Drift (external data scientists, Phineon), Peter Schaap (domain experts, PWN), Arjen Schimmel (product owner, PWN) and Lianne van der Laan (information manager, PWN);
- **Oasen**: Erwin Vonk (KWR), Jurjen den Besten (domain expert water utility, Oasen/); Brian Buitelaar (ICT; Oasen); Bas Bouwman (decision maker, PWN);
- **WBGr**: Dirk Joost Masja Bronts (WLN, external data scientist), Roel Hoekstra (centric, external datascientist)

The three pilot projects were observed and evaluated by a social science researcher from KWR. This evaluation focused on identifying success and fail factors for conducting data science projects in water utilities. Information was obtained by two rounds of interviews with the participants, attendance at monthly progress meetings and observation of activities and communication of the project team.

1.3 How to read this report

Chapter 2,3 and 4 describe the process and results of the different pilots. The first part of each of these chapters is technical in nature, describing the approach and results. The second part of each of these chapters focusses on the process. In chapter 5 the cases are compared, and common success and fail factors are discussed. Finally, Chapter 6 summarized the overarching conclusions and presents a number of recommendations for the successful implementation of data mining techniques at drinking water utilities.

2 Case Waterbedrijf Groningen

2.1 Introduction and research question

Unvalidated water meter registrations and erroneous water meters can cause errors in customer billing administration, the water balance and the calculated non-revenue water (NRW). These errors can result in additional work in customer administration and errors in the NRW. One measure to improve registration is to reduce the risk of water meter failures, an active field of attention for Waterbedrijf Groningen.

The goal of this pilot at Waterbedrijf Groningen is to develop a method and a prototype data mining tool to improve the identification of water meters that are suspected of errors due to a stagnant water meter readout. This pilot investigates the potential of automated identification techniques (i) to reduce the workload of employees that is currently required to identify individual anomalous meters, and (ii) to identify suspicious populations of water meters that have similar characteristics.

To reach the goal we formulated three routes to identify erroneous, stagnant water meters, by:

1. using water consumption statistics
2. using a machine learning (ML) approach on different, meter-specific tags (location, brand, etc.)
3. using the water consumption statistics as an extra input to the ML approach. This route is briefly described in section 2.2, but results are omitted due to data inconsistencies.

For route 1 and 2 we compared the results with the suspicious water meters identified by Waterbedrijf Groningen. We focussed on the distribution of water meter population characteristics and the number of anomalies. Additionally, we present the consumption statistics of the first approach.

2.2 Approach and methods

2.2.1 Data sources and preparation

The evaluation area encompasses the entire province of Groningen, which includes approximately 270.000 water meters. Relevant data for routes 1 and 2 is listed in

Table 1. The primary source of data are the customer consumption data. Consumption records are stored on a yearly basis (1988-2017), on the level of individual water meters.

Table 1. Data sources.

Data set	Type of data	Owner	Approach
Meter readings (SAP)	Time series	Waterbedrijf Groningen	1, 2, 3
Reasons for anomalous consumption	Listing	Waterbedrijf Groningen	2
Specifications of water meters based on Meter ID (Mecoms)	Distributed	Waterbedrijf Groningen	2, 3
Water meter population identifiers	List	Waterbedrijf Groningen	For comparison reasons
Water consumption statistics	List	VEWIN	1

Data collection started with the SAP and Mecoms data sets. These files were very large and contained privacy information about customers, so they were obtained via an FTP server. After this we processed (read: joined) the data sets to end up with a single data set where each sample contains

- meter reading information (meter ID, timestamp, current and previous meter readings, estimated and calculated consumption, number of consecutive estimates, communication type: letter, email, etc.),

- corresponding meter specifications (meter brand, meter type, date of installation ,address and postal code), and

- the stated reasons for the anomaly (as labelled by Waterbedrijf Groningen).

Meter brand and meter type were combined to create meter population identifiers, but they didn't join well.

For each of the three approaches mentioned above, we performed specific preparation steps which are explained in Attachment I.

The platform that was used during this project is Dataiku (www.dataiku.com) – a collaborative data science software platform. Dataiku facilitates tools to build a visual flow process (see Figure 1) of concatenated modules that consist of one or more operations for data reading, processing, analysis and visualisation. The user can build the flow process by either clicking and dropping modules or programming scripts in Python or other programming languages. The reason for using Dataiku is that its use was facilitated by Centric, a cooperating partner in this pilot, and that all the programming steps could be stored and visualised in one place.

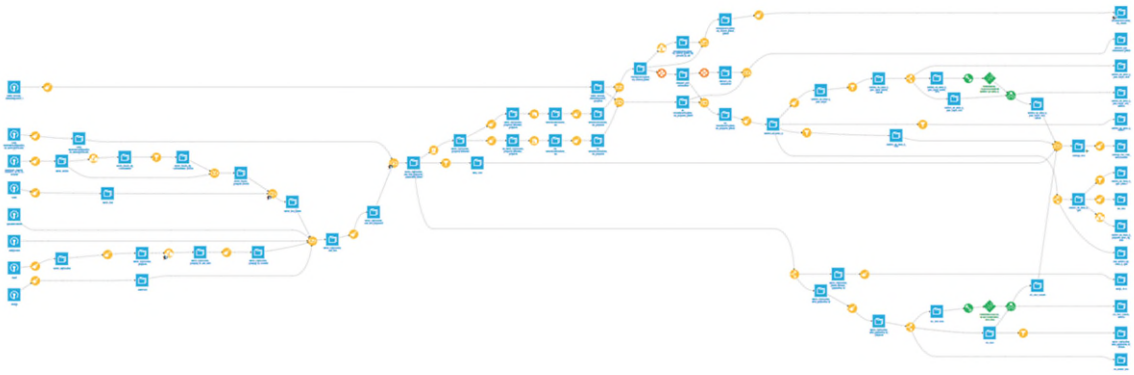


Figure 1: final flow scheme of route 1 to 3 in Dataiku. Yellow and green icons represent modules that operate on the data (represented by blue icons).

2.2.2 Current procedure of identifying suspicious water meters at Waterbedrijf Groningen

Waterbedrijf Groningen has approximately 270.000 water meters in operation. When a meter reading is registered by a customer, the consumption is calculated based on the meter reading of the year before. It is automatically checked whether the consumption lies within a defined range based on an expected consumption (which is an estimation by Waterbedrijf Groningen based on the consumption of the previous year). If this is the case, the invoice is sent immediately to the customer. If the value falls outside the boundaries, an employee evaluates the meter reading in more detail and assigns a label from a checklist with the reason why it is anomalous. The manual evaluation comprises 10% (~26.000) of the meter readings each year. This is a laborious process, which is why it is interesting for Waterbedrijf Groningen to further automate this process.

One of the labels that an employee can fill in is 'suspected of stagnation'. Each year almost 3% of the water meters gets this label. Financially speaking, Waterbedrijf Groningen has an interest in solving this issue, because their policy is to replace all these meters. Customer satisfaction is another aspect that plays a role in correctly identifying anomalous meters.

In 2017 Waterbedrijf Groningen assessed which portion of the 3% suspicious meters was actually stagnating. The conclusion was that 3 out of 10 were *not* stagnating, which gave Waterbedrijf Groningen enough proof to just replace all 3% of the suspicious water meters. Additionally, the labour costs involved in repairing water meters do not weight up against replacing.

The reason for this high percentage (in comparison to other water companies) of anomalous meters is a bulk purchase of a specific water meter brand and type. Waterbedrijf Groningen expects that this water meter population does not meet the required performance. Consequently, replacement of meters within this population is needed in the near future. Note that replacement of all meters at once would become very expensive.

2.2.3 Approach 1 – Anomaly detection of suspicious water meters

The goal of approach 1 is to automate the process of distinguishing between suspicious and non-suspicious meters, using only water consumption data. If this method labels less than 10% of the water meters as anomalous, but still covers the set

of water meters labelled by Waterbedrijf Groningen as 'stagnating', it can reduce the workload of employees.

For this approach, lower bounds on regular consumption deviations with respect to long-term averages are determined. Outliers in data are typically defined as observations that fall below the 25th percentile (Q1) minus 1,5 times the inter quartile range (IQR). The IQR is defined as difference between the 75th (Q3) and the 25th percentile. Hence, the detection threshold for stagnant water meters is defined as Q1 – 1,5*IQR of the deviation from the average yearly consumption Thus, the lower bound is used to identify outliers with respect to regular water consumption. The anomaly detection is tailored to household size-specific bounds and contains the following steps which are described in more detail in Attachment I.1.

1) Prepare a reference data set of consumption history per water meter.
2) Calculate the average yearly consumption for each water meter.
3) Assign household size categories based on average consumption and VEWIN data.
4) Determine consumption deviation thresholds per household category for outlier identification.
5) Apply the consumption deviation thresholds of step 4 on the original data set to identify consumption outliers.
6) Compare outliers to the water meters labelled by Waterbedrijf Groningen as 'suspected of stagnation'.

2.2.4 Approach 2 – Predict whether a meter is suspected of stagnation using ML

The goal of approach 2 is to predict whether a water meter is suspected of stagnation using a machine learning approach of different, meter-specific tags. We use the labels from Waterbedrijf Groningen whether a water meter is suspected of stagnation (anomaly type '12') as our target variable in a ML model. The training dataset comprises of raw meter values observed between 2011 and 2016. Evaluation is performed on data of 2017, this is the so-called test data set. During training another 80/20 split was performed where the model trained the logistic regression on 80% of the data and evaluated the model on 20% of the data. The following features were used for making a prediction using a logistic regression: city, manufacturing date, manufacturer, street, location of the water meter, reading value raw, brand, reading type, recording reason and date of first usage. This way, we predicted the target variable for the year 2017 and compared the results with the labels from Waterbedrijf Groningen for 2017.

2.2.5 Approach 3 – Predict whether a meter is suspected of stagnation using the results of approach 1 together with ML

The goal of approach 3 is to predict whether a water meter is suspected of errors due to a stagnation in meter readout, using the results from route 1 as features (that is, household size, yearly consumption and deviation in yearly consumption) and using tags that are meter-specific (meter location, date of installation, meter identifier, population identifier). Then, machine learning models are trained on the water meter data, where we use the tag for stagnant water meters by Waterbedrijf Groningen (tag 'type12') as our target variable. The results of this approach are omitted due to data inconsistencies.

2.3 Results

Below, the results of each of the three approaches are shown. A more detailed explanation of the results and the methods used can be found in Attachment I.

2.3.1 Results of approach 1

As mentioned before, the goal of this approach was to automate the process of distinguishing between suspicious and non-suspicious meters, using only water consumption data. The consumption data is discussed and evaluated per household size in Section 2.3.1.1. The number of water meters which are labelled as stagnant for each year is shown in **Error! Reference source not found.**. Note that the number of outliers increases significantly by the year 2008 and 2011 compared to the number of outliers in the preceding periods. Section 2.3.1.2 further discusses the differences between approach 1 and the approach as used by Waterbedrijf Groningen.

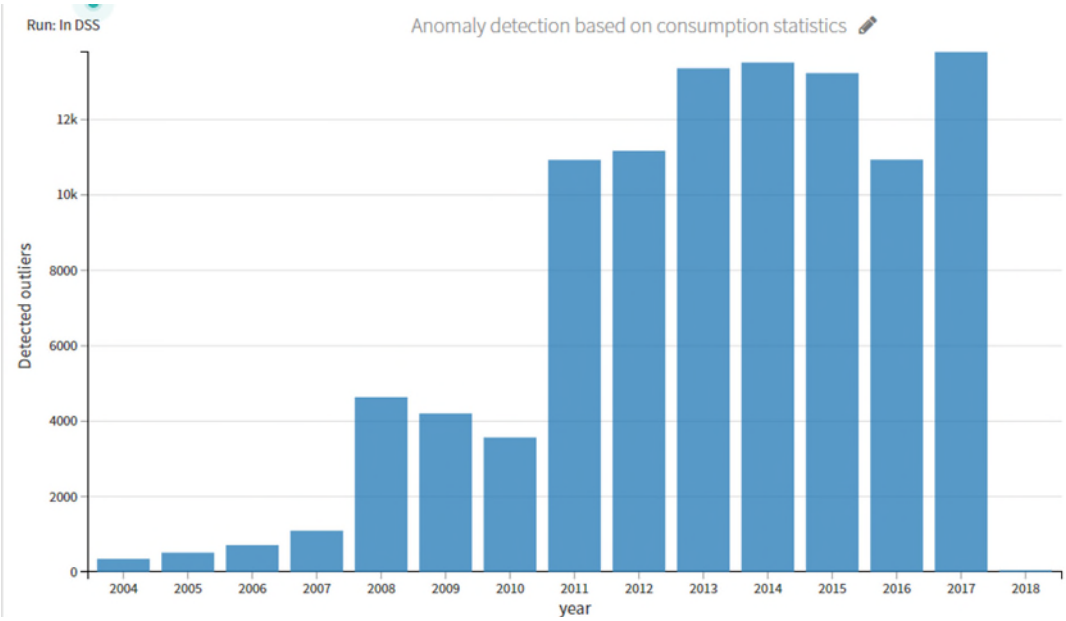


Figure 2: Number of water meters identified as being stagnant per year.

2.3.1.1 Consumption data evaluated per household size

Based on the household size categories (result of step 3), the distribution of yearly consumption data was calculated and plotted as box-whisker graphs for each household size category. Each box spans the 25th and 75th percentile, the band represents the median, and the whiskers depict the 1st and 99th percentiles. Results are presented in Figure 3 and Figure 4 for the un-scaled and scaled data set respectively. For Figure 4, yearly consumption is scaled to the household size. We show results up to household sizes of 6 persons, although categories of larger consumption, up to 1000 m^3/year, were analysed. Interestingly, household sizes 3, 4 and 5 have little variation (small inter-quartile bandwidth), while a household consisting of one person can have a large variation in yearly consumption. This becomes even more apparent in the scaled yearly consumption graph (Figure 4).

The deviation in yearly consumption is shown as box-whisker graphs for the whole data set and for every household size in Figure 5. Notice the general increase of (un-scaled)

consumption deviations with progressively increasing household size (i.e. wider box-whiskers). Larger deviations are indeed expected for households with a bigger demand. The dots below each box-whisker graph in Figure 5 represent the threshold values used for the identification of consumption outliers (step 5 in Section 2.2.3).

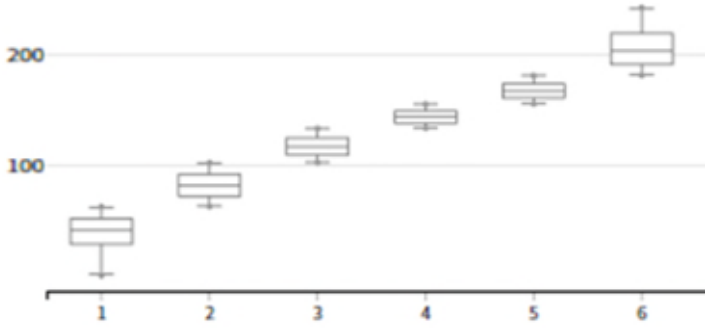


Figure 3: Box-Whisker graph of yearly water consumption (vertical axis in cubic meters per year) for each household size category (horizontal axis, from 1 to 6 persons).

Figure 4: Box-Whisker graphs on yearly household consumption data which are scaled to household size.

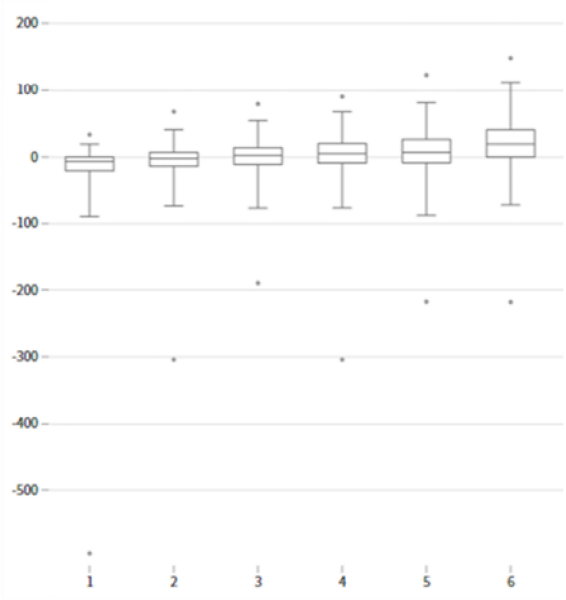


Figure 5: Box-Whisker graphs of the deviation in yearly consumption data for each household size.

2.3.1.2 Comparison of results

The number of detected stagnant water meters of approach 1 and the distribution of brand and type of meters (i.e. the combination of brand, type and year of installation is tagged with a population identifier) have been compared with the stagnant meters as identified by Waterbedrijf Groningen.

Table 2: overview of water meters tagged as being stagnant and population id match for the year 2017.

	Number of stagnant meters	Match with water meter population identifiers
As detected by the model in Route 1	~13600	Complete
As identified by Waterbedrijf Groningen	~7000	Partial (~500)

As mentioned in Section 2.2.2, Waterbedrijf Groningen assessed in 2017 for a small subset of suspicious meters that 70% were indeed stagnant. Such unequivocal information is not available for all meters that Waterbedrijf Groningen identified as suspicious. We therefore used the ~7000 suspicious meters (per year) as our target set. Ideally, the identified outliers include all of the ~7000 suspicious meters and is less than ~26000 meters per year. This is a prerequisite in reducing the laborious process of manual evaluation of meter readings for Waterbedrijf Groningen. Although the outlier number (13600 meters, Table 2) is indeed between 7000 and 26000, it was not possible to assess overlapping outliers and hence determine whether or not the Route 1 approach is overestimating the outliers determined by Waterbedrijf Groningen. This assessment was hampered by the process of population matching, due to water meter population identifiers that are registered in another data set and joined with the water meter data. The join resulted in approximately 500 water meters having a water meter population match out of approximately 7000 stagnant water meters, see also Table 2. The relative proportions of represented water meter population identifiers in stagnant water meters is shown in Figure 6. The relative proportions of population identifiers as detected in route 1 are shown in Figure 7. Since there is a limited match with population identifiers, we only discuss the top 4 population ids of both routes. Interestingly, the results from both approaches are quite similar (although the order is different): the Presikhaaf M3 Qn 1.5 (both of epoxy and non-epoxy type), the Brinck Qn 1.5 Sensus 520 and the Elster Qn 1.5 type V200 appear to be troublesome. Conclusive remarks about overlapping or disjoint sets can only be made in case all detected stagnant water meters have a population match. In a potential follow-up it is recommended to normalise the relative proportions of populations (Figure 6and Figure 7) to their presence in the field, and to apply a balanced accuracy metric[1].

[1] With a balanced accuracy metric, the true positive and true negative predictions are normalized by the number of positive and negative samples, respectively, and their sum is divided by two.

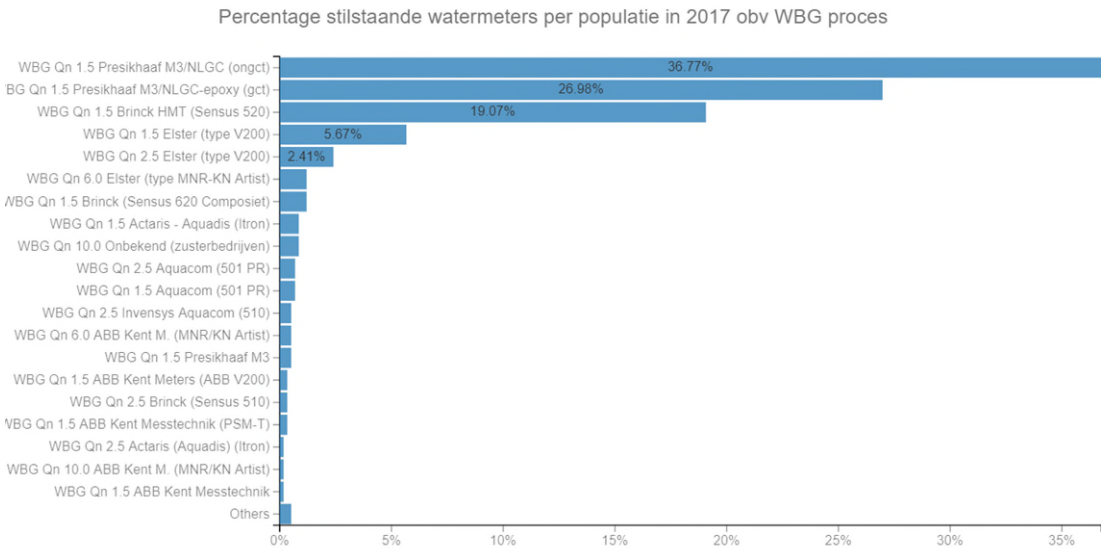


Figure 6: distribution of population identifiers in stagnant water meters as tagged by Waterbedrijf Groningen

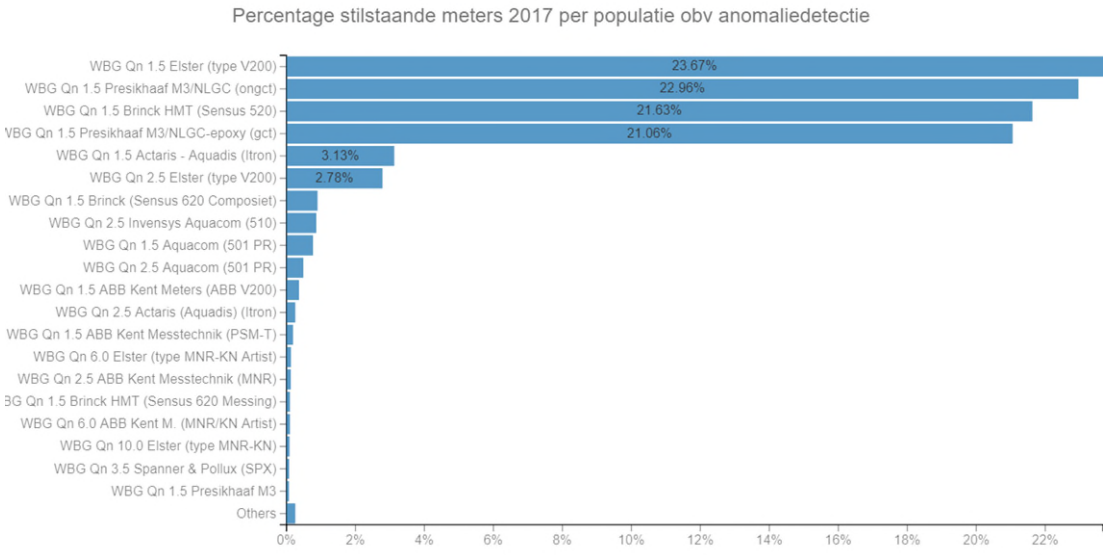


Figure 7: distribution population identifiers in stagnant water meters as identified by Route 1.

2.3.2 Results of approach 2

2.3.2.1 Performance metrics

For training and evaluating the model with the ML-approach, we used the metrics Accuracy, Precision, Recall and F1, see Table 3.

The first training results show an Area Under the Curve (AUC) of 0.973. The AUC is a diagnostic measure for a binary classifier, varying between 0,5 for a random guess and 1 for a perfect classifier. Although results look promising, note that we had to deal with an imbalanced dataset, which means that 93% of the dataset are registered meter values of meters that are still working. In other words, a manual prediction where each registration retrieves a value 0 (i.e. meter is still working) would be correct in 93% of all cases. This is why we do not use the proportion of correct predictions, i.e. the metrics AUC and 'Accuracy', for evaluating performance.

Table 3: training and test scores for different performance metrics in approach 1.

Measurement	Training score	Test score
Accuracy: proportion of correct predictions	100%	99%
Precision: proportion of correct positive predictions	42%	11%
Recall: proportion of positive actually records correctly predicted as positive	21%	25%
F1: Harmonic mean between precision and recall. More informative than Accuracy for unbalanced datasets	28%	15%

The 'precision' measure shows that 42% of the total number of water meters predicted by the model as stagnating, were actually correct in the training set. The 'recall' measure shows a correct prediction of 21% of the total number of water meters identified by Waterbedrijf Groningen as stagnating. The F1 score is in between the precision and recall score, as expected.

Surprisingly the precision is much lower during the test on 2017 data versus the training on 2011-2016 data, whereas the recall is 3% higher. Looking at the precision and recall, a lower F1 score is to be expected here.

Currently, the ML performance would not benefit Waterbedrijf Groningen in identifying stagnant water meters. It is important to investigate the possible reasons, in order to evaluate the potential of the ML-approach more conclusively. First, it is possible that relevant data is still missing in the set of evaluation parameters that was employed. Inclusion of customer consumption data could improve the results. Note that this was explored in approach 3. Other potentially relevant data comprises the water type: Waterbedrijf Groningen supplies drinking water from five production locations, subdividing the supply region into identifiable subregions with specific water types (e.g. Marel & Van der Woerdt, 2014) that potentially influence the performance of water meters. . Finally, it is possible that the performance is hampered by incomplete data cleaning, which is a likely factor given the course of this study. Investigation of the contributing factor of the ML performance is a recommendation of future research.

2.3.2.2 Comparison of results

For the identified stagnant water meters, the proportion of corresponding water meter population ids is again investigated, see Figure 8. Unfortunately, we were unable to find a complete match with population identifiers, hence only a qualitative evaluation is possible. Recall that the distribution of population ids within the set of stagnant meters as tagged by Waterbedrijf Groningen is shown in Figure 6. Again the Presikhaaf Qn 1,5 M3 meters as well as the Brinck Qn 1,5 Sensus 520 appear to be troublesome. Interestingly, the logistic regression model did not tag Elster V200 meters as being stagnant, contrary to the findings of Waterbedrijf Groningen.

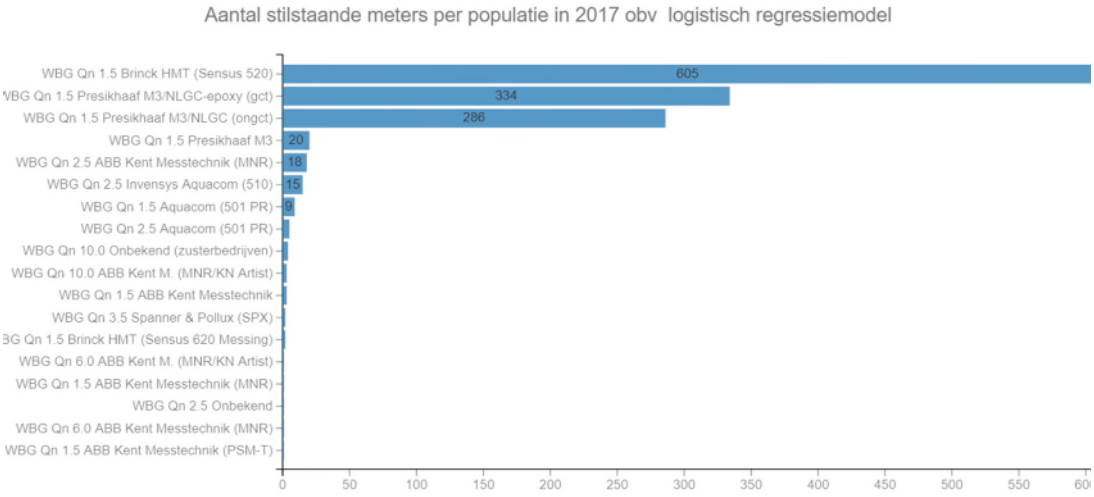


Figure 8: distribution of population identifiers in stagnant water meters as identified by Route 2.

2.3.3 Results of approach 3

The idea of approach 3 is to support the performance of machine learning algorithms by adding calculated features of route 1 (e.g. household size, yearly consumption and deviation in yearly consumption). Due to the very limited amount of labelled stagnant water meters by Waterbedrijf Groningen and data processing, the final train data contained only 78.574 records, with 1173 stagnant water meters for data until 2017. Trained models had a precision metric lower than 0,1; hence, further (prediction) results are omitted. Due to limited resources we were unable to remediate the lack of training data.

2.4 Conclusions and recommendations

2.4.1 Conclusions

- Data management (obtaining, understanding, cleaning, and preparation) was a laborious step in this project. A common rule of thumb in data mining is a 80%-20% division in time between data management and data analysis. This balance was heavily skewed in this project - perhaps 95%-5%. A contributing factor were the incoherencies in data labels across different data sources. This illustrates the importance of data preparation prior to starting a data mining project.
- Although it was possible to identify water meters that could be caused by a stagnant readout using statistics of customer water consumption (approach 1), a proper comparison to outliers by Waterbedrijf Groningen was not possible. At this point and with the current data sets, it is impossible to say whether this approach

is over estimating this number compared to the results by Waterbedrijf Groningen. The statistics based method does result in a smaller amount of possibly erroneous meters than the first step of the approach by Waterbedrijf Groningen. The bottleneck problem can be traced back to the definition of water meter population ID names having inconsistent naming across different data sources. We recommend thorough data cleaning and data preparation with checks from Waterbedrijf Groningen to provide a sound basis for data mining.

- With machine learning using logistic regression (approach 2) it was not possible to convincingly identify meters suspected of stagnation. It is likely that consumption data (not included) is required to improve the performance. Attempts to include consumption data failed because of a lack of stagnant water meter labels, resulting in prediction of low precision (precision metric <0.1). Another complicating factor is that the data is very imbalanced, i.e. few suspicious water meters on a large data set which hampers modelling and gives unreliable results. This issue can be tackled by using a balanced accuracy metric to score the performance of the machine learning algorithms. Another potential improvement is the expansion of the data set with information about the water type received from different production locations, which potentially influences the performance of meters..

2.4.2 Recommendations

The main recommendations concerning the registration process and data management for Waterbedrijf Groningen that follow from this project are:

- Register the failure and failure reason during water meter replacement, specifically and consistently. Allow registration of various reasons of stagnation (freezing, stagnating counter, clogged filter, etc.) or replacement (preventive replacement due to high risk population, etc.). Consistent and specific registration allows the future availability of target labels, and this is required for a proper comparison between the current stagnant meter identification process of Waterbedrijf Groningen and the data mining approaches investigated in this project. This would be a relatively minor effort to improve the identification of stagnant meters.
- Use unique labels that correspond across different data sources, in particular for the population IDs. This will allow the effortless joining of population IDs required for the improvement of the outlier identification (approach 1) and the related machine learning (approach 3).
- Make sure that a server is available to work with large data files. During this project it appeared that the available server using was not powerful enough to train the models: we had to reduce the amount of training data used for modelling.
- Always congegrate with domain knowledge, data science and ICT experts before starting a data science project. This way helps to overlook obstacles beforehand.
- To understand the process of replacing anomalous water meters we came across the ‘Quality Assurance Regulation for Water Meters’ (Dutch: ‘Regeling Kwaliteitsborging Watermeters’ (RKW)). This topic lies outside the scope of this project, but we had some doubts about the determination of the sample size of the process of eliminating water meter populations. We therefore recommend Waterbedrijf Groningen (together with other water companies) to investigate this regulation.

The main recommendations for further research are:

- Re-evaluate the outlier identification results (approach 1) by joining population IDs (when matching IDs become available). This allows for a proper comparison to the identification approach currently employed by Waterbedrijf Groningen.

- Explore performance improvement by a more thorough data cleaning procedure.
- Perform the anomaly detection method on the data collected for 2018 as a test (as suggested by Waterbedrijf Groningen). A start was made in this project, but because of data quality issues and lack of time the first results are not yet reliable. This would be a useful exercise for a follow-up project.
- The anomalous water detection problem can be solved by answering two questions (1) is the timeseries suspicious, and (2) if so, is the time series coming from a suspicious meter? To solve this problem, we suggest to use a combination of statistics and machine learning as e.g. proposed in approach 3. Due to the nature of the problem, random forest trees or similar other algorithms are worthwhile to explore. The use of additional environmental data , like water quality data, could improve the prediction accuracy further.

2.5 Evaluation of the process

This evaluation is based on participation in progress meetings, spending time with the project team and two rounds of interviews (see Table 4) with the main participants in the project.

Table 4: Interviewees

Name	Organization
Bernard Enthoven	WBG
Masja Bronts	WLN
Eddy Postmus	WBG
Sybo van Scheltinga	WBG
Anne Fijma	WBG
Jacob van Dijk	WBG
Dirk Vries	KWR
Joost van Summeren	KWR
Roel Hoekstra	Centric
Arie Martens	Centric
Ernst van Aagten	WBG

This case was initiated by Peter van Thienen (KWR) as part of the Exploratory Research Data Mining. Waterbedrijf Groningen was asked for potential cases to be studied as part of this project. Waterbedrijf Groningen had several project ideas and the final selection was made by WLN, WBG and KWR in a joint meeting.

At that point it was not yet clear which data was needed for the analysis and whether that data was (readily) available. The project team noted in hindsight that a quick scan of the data necessity and availability could have been useful to get an impression of the feasibility of this project, also given the limited time available.

A project team was formed, consisting of four people: two researchers from KWR, one from WLN and one from Centric. The team had a good mixture of knowledge and competences (modeling skills, statistics and domain knowledge) and the atmosphere in the team was good.

The project team had weekly one-day sessions at WBG to do the project work. First interviews were conducted with experts from WBG to get familiar with the issue at hand, the relevant processes within WBG and to start acquiring the necessary data. In this

phase it became apparent that the data availability as well as the quality of the data was challenging. Several issues with the data became apparent in the course of the project. Most important issue being the lack of data on stagnating water meters (only 'suspected of stagnation' was available). It was noted in the interviews that if this was known before selecting this topic, another topic may have been selected.

Also, WBG had a change of systems (SAP to MECOMS). In the migration of data, some of the labeling of brand names for water meters changed, which made integration of the different datasets more difficult. After some iterations of data preparation it was decided, together with the project leader from WBG, to proceed with the data analysis, instead of trying to further improve the data.

At the outset of the project, there was no platform available within WBG to collect the data and conduct the analyses (which was known to the people involved). If this platform would have been available, then giving access to people from outside WBG would also have been an issue. Centric did provide a platform, DataIku, which proved very useful for the project work. However, this platform is not owned by WBG, which makes transferability of the models and continuation and implementation of the results more difficult. Also, the project team used a free version of the platform, with limited functionalities.

Finally, a beginning was made with the Machine Learning techniques (see 2.2.2 and further), but the project team had about 5% of its time left for this part. This means the project team could not come to substantial outcomes that were to their own satisfaction. This does not mean the project was not successful. The project generated important insight in what it means for a company such as WBG to conduct these kinds of analyses, primarily regarding data management. Also, it initiated a closer focus on the process of collecting data from water meters. Closer examination of the process in 2018 revealed some odd situations, such as individuals submitting their meter data 500 times in 10 minutes.

2.6 Success and fail factors

In the following paragraphs, a number of success and fail factors are identified, based on observations and interviews with the people involved in the project.

2.6.1 Success factors

At an early stage all stakeholders were involved to explore the issue and get an understanding of the relevant processes within WBG. This created a thorough understanding of the relevant issues and related processes within WBG. Stakeholders within WBG were easily accessible and were willing to share time and information with the project team. In this way, the project team could get an idea of the data necessity, availability and what specific issues the data mining project could address.

There was a good atmosphere and good cooperation within the project team. The members of the project team enjoyed working together, which, according to one of the members, proved particularly valuable when difficulties occurred with the project work. The cooperative atmosphere may have made the difference between trying to overcome difficulties or letting the project muddling through. Each of the project members has invested time beyond the project hours to get the best results possible, given the constraints of the data availability and quantity.

There were close connections with professionals within WBG, working at the site of WBG. The project team had their own working space at the 6th floor of the WBG building. In this way, they were slightly separate from the day to day business of WBG, but close at hand for obtaining information as well as data and discussing progress with the project leader. This made it easier to reiterate steps, get additional information from experts within WBG and keeping WBG aware of the project.

Dataiku was very useful to combine and visualize all the scripts used in the process. It was more suitable and practical than the tools that the other parties could provide. The visualization made it easier to cooperate on the platform and keep track of the complex iterations necessary to clean up, combine and prepare the data for analysis. Possible drawback is the lack of an export function to continue the use of the scripts by WBG.

Clear assignment and proper adjustment along the way. At the outset, the goals for the project were clear, as well as the available time and resources. After the first assessment of the data situation, the project goal was slightly reformulated to focus on the identification of water meters suspected of stagnation, instead of erroneous water meters. Although the goals were clear, there was no long term project planning. The project team worked through several iterations to collect and prepare the data. After consultation with the project leader, the decision was made at some point to proceed to the analysis phase. This adjustment prevented the project from continuing in more cycles of data collection and preparation.

2.6.2 Fail factors

Data availability was not suited to the research problem. This was mentioned by most respondents as the key factor for the outcome of the project. Most notable the 'target column' (stagnating water meters), was not available, only 'suspected of stagnation'. There were also questions on how consistent the label 'suspected of stagnation' was assigned.

Data quality was not sufficient for analysis purposes. Primarily, the labelling of data across different data sets is not consistent. This made it difficult to compare water meter populations across different data sets.

No overview of the data situation at WBG. There is no single overview of the available data within WBG. Data sets are owned and maintained by different people within WBG. This can cause inconsistencies in labelling and makes the process of collecting and combining data for analysis purposes very time consuming.

Project planning. There was a beginning-to-end project plan with time allocated for the different phases. However, the project team took an iterative approach, going through different iterations of a step before going to the next step. Although this is a very justifiable choice given the uncertainty with regard to the data situation, it meant in this case that the analysis phase was pushed ahead a few times, to do a few more cycles of data collection and preparation. Finally, together with the project leader, it was decided to proceed to the analysis phase. It is by no means suggested that doing more cycles of data preparation was a wrong decision, but it left much less time for the analysis than was allocated in the project plan.

Limited time for a complex project. Available time proved a limiting factor in this project, particularly given the large amount of time necessary for data preparation.

Some participants noted that two months of time extra, could make a substantial difference in the end-result. Thereby, the timing of the project overlaps with the timing of the primary process to which this project is related (collecting water meter scores), which made the availability of important stakeholders limited. This does not mean that data mining projects always require lots of time. It means that in this case, the available time did not match the required time for doing the work, and also, that in these projects, the time needed to come to an end result is difficult to predict of there is not a clear view on the data availability and data quality.

Too little involvement of domain experts in structuring the algorithm. Some of the domain experts indicated that they would have liked to be more involved in building the ML algorithm. This was more a matter of timing than a lack of willingness to collaborate. The domain experts had no time reserved for this project and the project was conducted at the same time that WBG collects the measurements from their 260.000 water meters, which is a time consuming process.

2.6.3 Conclusion on the process

Although some stakeholders involved did not find the outcomes satisfactory, most agreed that it was a useful learning experience which has generated important insights, primarily regarding the need for consistent data management practices, systematic labelling of data and the consistent use of meta data. The focus should not be on collecting data for a specific purpose, but to manage the data in such a way that it is useful for a broad range of (yet unknown) purposes. Improving this is not done overnight and requires substantial commitment of people throughout the whole organisation.

This project was framed both as a research project and an implementation project (with a go/no-go based on the outcomes of the prototype. However, these two are quite different in nature. Research requires a certain amount of tinkering, trial and error and exploring, while implementation is focused on standardisation and integration in current practices.

When implementation is the goal of a data mining project, this should be the focus early on. The environment in which the model is developed has to be suited for implementation later. All the steps from the data source to the outcome of the model have to be documented so that the model (including data preparation) can eventually be run on the required platform. This requires continuity and consistency in the way data is gathered and stored and transferability of the necessary scripts from the research environment (like DataIku in this case) to the server that runs the final application.

The go/no-go that was put between the prototyping and the implementation could be put earlier in the project planning as well. For instance, one could select 5 – 10 potential cases for a project such as this, then let the project team make a first assessment of the data availability and quality on the basis of which a selection is made. Then, a prototype could be developed for that selection (let's say 2 or 3), after which a go/no-go for implementation is made. This would of course require more time and effort than was available in this project, but may reduce the overall time invested in data management (by the project team).

If a model has to be developed from scratch, including data preparation, much more time has to be reserved for gathering and preparing the data. In this project, the time

reserved for preparation (including data gathering, data preparation, literature review of possible methods) amounted to about the same number of hours as the time reserved for analysis. In a typical data analysis project, this ratio is 80:20 and in the case of this project it turned out to be about 95:5.

The close cooperation between researchers, professionals and private sector data scientist in this project is in itself an interesting outcome. The iterations between scientists and professionals in validating results and identifying next steps based on a joint understanding of these results makes for a good example of knowledge co-creation. This, of course, puts new requirements on both professionals and researchers. Professional need to step out of their daily practice and engage in the more exploratory and abstract scientific discourse. Researchers need to relate their findings (throughout the project) to the needs and challenges faced by the professionals.

The complexity of data analysis projects requires an increased effort from WBG in its role as client. Substantial knowledge and understanding of data techniques is required to ask the right questions and to evaluate the outcomes of the project.

3 Case Oasen

3.1 Introduction and research question

The goal of the pilot at Oasen was to develop a web-based dashboard that shows an up-to-date prediction of the extremes in water demand (peaking factor) for various future scenarios (horizon 2050), as well as additional information such as model accuracy scores, and visualizations of the predicted water demand. Under the hood, this dashboard contains a machine learning model, named EDWARD, that relates driving factors (daily holiday absence and weather) to the water demand. With EDWARD, water utilities can estimate the impact of climate change and variations in holiday behaviour on the 'peaking factor', which is a useful metric for estimating future infrastructure capacity requirements. This model has previously been developed in the BTO research programme (Vonk et al., 2017), but so far had not yet matured into an operational application. Hence, when reviewing this process from a datamining perspective (Figure 9), the modeling effort prior to this project was stuck at the 'evaluation' stage: the model works and is documented, but it is not yet deployed for everyday use.

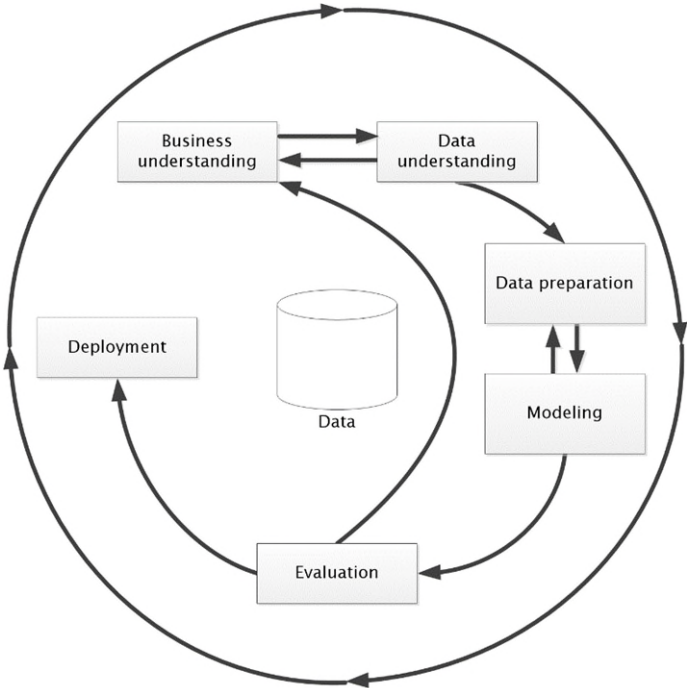


Figure 9: CRISP-DM process model for datamining.

This project was guided by the following research questions:

1. How accurately can we predict the aggregated water demand in the supply area of Oasen with the data sources that are available to feed the model (in relation to data quality and quantity)?
2. How can we integrate the model within the existing software landscape of Oasen?
3. How can the model results be presented so that they add value to the end user?

3.2 Approach and methods

3.2.1 Prediction accuracy

Data sources required to feed the model are:

- Daily weather measurements; precipitation, evaporation, radiation, average temperature, maximum temperature.
- Weekly or daily holiday absence rate (percentage of the population that is on holiday on that particular day).
- Daily water demand for the supply area.

Weather measurements can be refreshed on a daily basis by using the web API of the Dutch Meteorological Service (KNMI). Since the supply area of Oasen lies close to multiple weather stations located in the vicinity, we used Thiessen interpolation to obtain spatially averaged measurements. Holiday absence statistics were provided by the CBS. For the daily water demand we used the water production volumes as recorded in data acquisition and storage system of Oasen, Osisoft PI. Underlying assumption here is that the volume of non-revenue water (leakages) is small and more or less constant, and the production is always sufficient to meet the actual water demand. Both assumptions hold in practice for supply areas in the Netherlands. At Oasen, aggregated production volumes for the total supply area are manually checked and validated on a daily basis. We used those validated volumes as model input.

3.2.2 Integration in existing software landscape

Integration with the current software landscape was achieved by direct involvement of the IT-department of Oasen. Initial talks quickly revealed a set of software tools that were currently in use. Whenever possible, those tools would be used in this project to ensure that employees feel comfortable working with it and to avoid unnecessary ‘reinventing of the wheel’. In addition, Oasen shared their roadmap of upcoming developments in their software landscape, so that design decisions for this project could take that into account.

The practical integration was achieved by a step-wise approach in which the existing model was incrementally coupled to workflows and infrastructure at Oasen:

1. Initially apply the existing model to a static input dataset, to get a model performance baseline and quickly show first results.
2. Develop a data model for the data warehouse, then fill the data warehouse with dummy measurements to test model database connections.
3. Connect the data warehouse to its underlying data sources.
4. Develop the visualization frontend (dashboard).

3.2.3 Presentation of model results

Active discussions during project meetings revealed a number of potential ‘information users’. It became clear that a simple bar diagram visualization representing the calculated peaking factor per scenario was the most useful and easiest to interpret, along with a separate tab for visualizing the model prediction accuracy.

3.3 Results

3.3.1 Prediction accuracy

The initial model was run on a static dataset from 2002 till 2015 and yielded a training score (R^2) of 0.85 and a test score (R^2) of 0.78. This is fairly accurate and more than sufficient for our goal of assessing future scenario impacts (Figure 10). EDWARD later on is fed with climate-transformed timeseries to yield predictions for 2050.

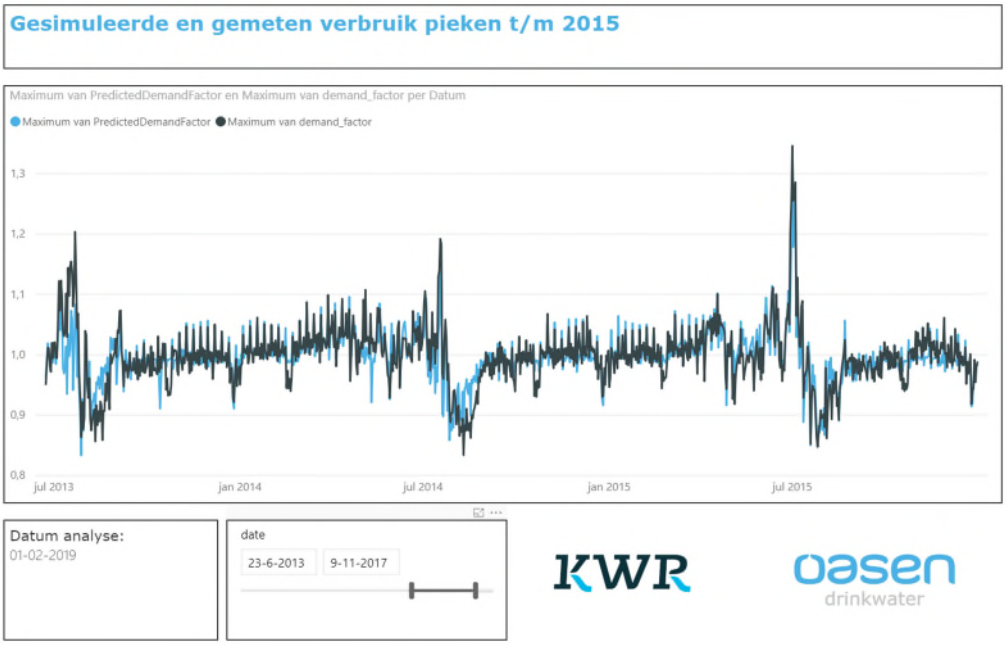


Figure 10 Simulated and measured demand peaks

3.3.2 Integration in existing software landscape

To ensure that the model can be run periodically, in a stable environment and without manual interference, an in-house server has been configured. Model calculations are automatically triggered annually on the 5th of January, but can also be triggered manually at any time, if required.

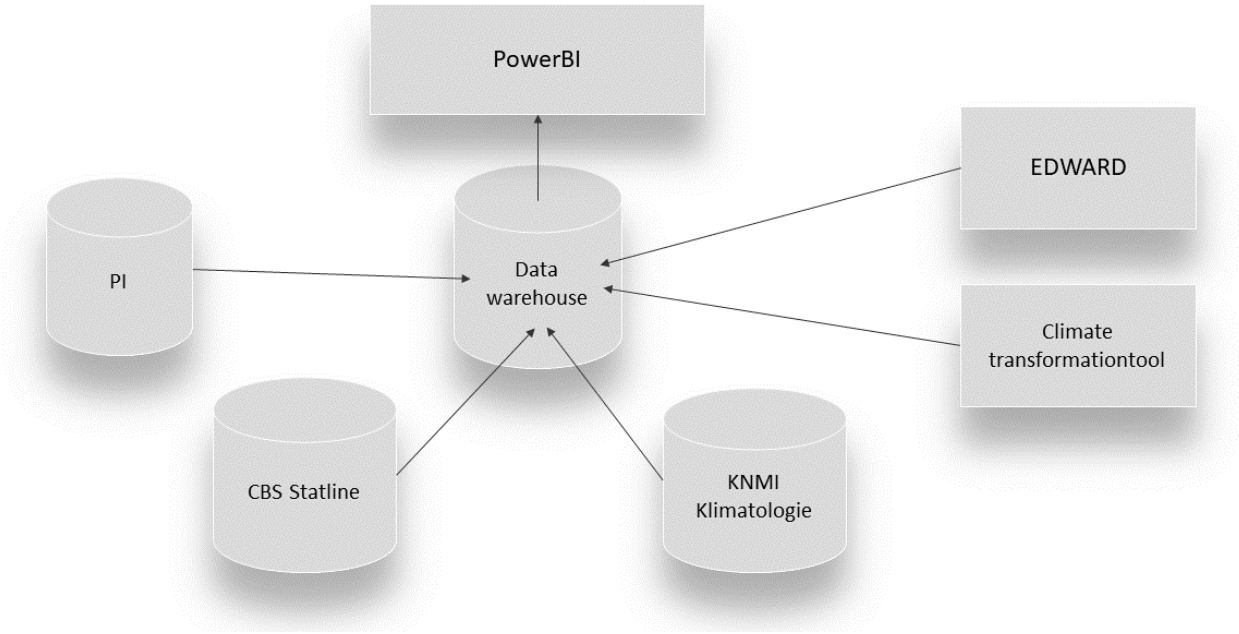


Figure 11 Software landscape. External (CBS and KNMI) and internal (PI) datasources are connected to a central data warehouse, which in turn feeds the EDWARD model and ultimately the PowerBI dashboard (for visualizations).

To ensure availability of model results to a broad audience within Oasen, a web-based dashboard was chosen as preferred presentation method. Since Microsoft PowerBI is already in use as data visualization tool at Oasen, this tool was also chosen for presentation of the model results.

The EDWARD model also integrates with an in-house data warehouse at Oasen (at the moment a PostgreSQL server, but to be migrated later on to Microsoft SQL Server), where all input data sources are stored. Incoming weather measurements, production volumes and holiday absence are synchronized periodically. This structure allows Oasen to use the same data sources for any other data driven models that may be developed in the future. By using a database-agnostic coupling, the model can be coupled to other database types in the future, if necessary.

3.3.3 Presentation of model results

Visualization of the results in PowerBI is displayed in Visualization of results in PowerBIFigure 12

Figure 12 Visualization of results in PowerBI

3.4 Evaluation of the process

This evaluation is based on participation in progress meetings, spending time with the project team and two rounds of interviews (see Table 5Table 4) with the main participants in the project.

Table 5: Interviewees

Name	Organisation
Jurjen den Besten	WBG
Brian Buitelaar	Oasen
Bas Bouwman	Oasen
Erwin Vonk	KWR

After talks between KWR, Oasen and Jurjen den Besten, a case for Oasen was selected. Based on a prior study for Texel, KWR would, in close cooperation with Oasen, develop a tool to estimate the day peak in water demand based on behavioural data (when do people go on holiday) and climate change scenarios.

The goal of this project was not to develop a new method, but to implement an existing method as a chain of data gathering-validation-interpretation-decision support. Through this implementation both KWR and Oasen would gain valuable insights in what is needed both technically and in terms of organisational capacity to successfully implement a data driven decision making tool.

In conjunction with the other cases, it was decided that the work would be done at the office of Oasen, with a KWR researcher working in situ on Thursdays and Fridays. Most of the work was shared between KWR and Den Besten, the latter being a former employee of Oasen and now an independent data scientist. The project team was

supported by an Information Analyst and a Team Leader Engineering. Each working day would start with a stand up meeting with the project team in which progress would be discussed. The project team would report back to the project manager in monthly progress meetings.

Since the methodology was already developed, the first part if the project was focused on exploring the specifics of the Oasen case and gathering and validating the necessary data. This phase went remarkably smooth. It was clear from the start what kind of data was needed, it was known where that data was available and the data quality was very good.

There was no data warehouse available at the outset of the project, so a temporary server was used to do the calculations. That meant the existing Python scripts had to be translated to R. It took quite a long time for the ICT-support to organize access to that server for the KWR researcher, which was addressed in the December progress meeting.

In addition to data from Oasen, also data sources from outside were used, namely climate data from KNMI and holiday-behaviour data from CBS. The latter was not part of the standard available public data, but due to this project, the CBS decided to make this data available in the future as well.

The prototype was finished by the end of the project and the resulting estimate of the day peak matched with the current expectations of Oasen. The prototype was linked to the existing data sources and runs on the separate server designed for the project. The goal of embedding the prototype in the data warehouse and maintain it in that context is not yet realised. This has mainly to do with capacity issues at the ICT department, due to a big ERP transfer project that absorbs much of the ICT capacity and takes longer than expected. This also causes delay in implementing the data warehouse.

3.5 Success and fail factors

3.5.1 Success factors

This project is considered a success by all of the participants and it provides a good example of what data driven decision making means for an organisation such as Oasen. A number of factors contributing to this success were identified.

Working on site, close connections with the organisation. This was mentioned by all participants as the primary driver for success. Having a KWR researcher around created awareness of the project among Oasen staff. It also helped in overcoming small barriers and kept all the participants informed of the progress. This can go as far as the design of the workplace, whether external researchers share facilities, such as coffee machines or are located on the same floor which means passing by each other's (preferably open) doors.

Good match between members of the project team. Both on the personal level as well in terms of skills there was a good match between the members of the project team. This contributed to a positive atmosphere and the ability to overcome challenges. All of the necessary skills (domain knowledge, hacking skills and statistics) were present in the two project team members that did most of the work. The position of Den Besten in this case is quite unique. He has long experience in working at Oasen and knows the domain very well; he knows which data is available within Oasen and

who owns it; and he has advanced data analytical skills, which is a rare combination. This proved an excellent match with the KWR researcher who also has advanced analytical skills, specific knowledge of the model and general knowledge of the water domain.

Good overview of available data and good data quality. At the level of data, the conditions for this project were comparatively good. It was known in advance which data would be needed for the project. Although there was no centralized data infrastructure (such as a data warehouse), the project team knew where and from whom to get the necessary data. Also, the data was of very high quality, making the data validation part of the research much briefer than anticipated.

An existing model was used. By using an existing model, the project team could save a lot of time and focus more resources on developing a prototype. The model already proved to give adequate results with the right data and part of the datasets (KNMI and CBS) were known from developing the model, although the CBS-data had to be acquired for a different region. From the model it was also clear which data from Oasen was required and whether this data was available in the right quality.

Clear planning, stand ups, progress meetings. There was a clear planning from start to end and there were frequent (weekly) stand ups with the project team and monthly progress meetings. In these meetings challenges could be addressed and the planning could be adjusted according to the circumstances. Partly because of the high data quality, multiple iterations of data validation, which can be really time consuming, could be avoided.

3.5.2 Fail Factors

ICT capacity and uncertainty about design of data warehouse. The ICT department at Oasen is not very well equipped to support these kind of data analysis projects. This is partly incidental, due to a transition to a different ERP-system, there was not enough capacity to develop a data warehouse and uncertainty to where this tool should be placed in a future data warehouse. These issues have not been resolved within the time frame of this project, but will be worked on in the future.

On a more structural level, it is difficult to get outside access to Oasen's servers, since this has to be arranged at different parts of the ICT department and generally does not have a high priority. Also, the ICT department is not very flexible in developing small scale solutions for specific issues. The tendency is to approach issues very thoroughly and structurally, which may lead to a complicated solution that requires the involvement of third party ICT specialists.

Data management at Oasen. Although in this particular case, the data was readily available and of high quality, this is not always the case. Respondents identified three potential issues with the current data management at Oasen: the data is scattered around the organization, usually owned by application-holders; the data quality is not optimal, because the data is not validated in a standardized way; the data is not easily accessible, because it is often not clear which data is stored were and how it can be accessed. This puts a ceiling on what can be achieved with data analysis, since few people within the company have the knowledge, skills and contacts within the organization to overcome these challenges.

3.6 Conclusions on the process

This case was considered successful by all participants. The results were ready on time and according to expectation. The only part that could not be realized was full implementation into a data warehouse. This is partly due to capacity issues, but also due to the fact that ICT at Oasen is not yet designed for a data driven organization.

Data at Oasen is often owned by application owners, who collect and maintain data for the purpose of the application. To be a data driven organization, Oasen needs to centralize its data management, so that all data is available to everyone, there is consistency across data sets in labeling and meta data and the data is well validated and maintained.

According to some respondents, this project shows that closer connections between ICT and asset management are needed. The initiative to become a data driven organization should not come from asset management, with ICT just acting on specific requests, but requires active involvement from ICT at a strategic level. According to one respondent, Oasen needs people that understand “both the bytes and the business”.

This process also shows that data science at Oasen is currently relying on a few individuals (perhaps only Den Besten), who knows his way around the Oasen data landscape, understands the specific challenges that Oasen faces (domain knowledge) and has strong competencies in data science. Combined with the scientific competencies of KWR, this makes for a strong, but also fragile, team.

Concluding, this case provides a good example of what data science can contribute to decision making, even at the very fundamental level of the capacity of the water infrastructure. It also highlights potential issues for asset management and ICT at Oasen on the road to becoming a data driven organization.

4 Case PWN

4.1 General introduction

In this pilot, two different topics are investigated. The reason to treat two instead of one is given in by the following reasoning that *it is better to start many smaller projects of which a few may be successful than one big project that may fail entirely*. We also argue that even small projects that improve the actual practice of people in the business will have a significant effect in advocating the benefits of a data-driven Water Company.

The two cases addressed in this pilot are:

- gaining insight in the food chain of the organisms that are found in the bag filters in the distribution network;
- relate the water quantities as measured at the treatment plant to those measured at the customers tap.

4.2 Bag filter case study

4.2.1 Introduction and research question

In the distribution network of PWN, bag filters are placed to prevent complaints by customers; they filter out larger sediment particles as well as some larger organisms. These filters foul over time and need replacing every one to two weeks. For over ten years, PWN is analysing samples of these filters to keep track of changing conditions that may be related to problems in the network. The sediments in the filter are analysed according to a strict protocol with which the number of certain organisms and type of sediments per volume are determined. Furthermore, the different fractions of the sediment are classified and their volume measured.

The question addressed in this case study is whether it is possible to gain insight in the food chain of the organisms found in the bag filters. This enhances understanding of factors influencing biological stability of the water because more is learned on the (micro)biology in the network.

4.2.2 Approach and methods

The data has been collected in several Excel sheets. In this project, they are combined to one file to enable machine learning. The file contains the following measurements for every 1 or 2 weeks:

- Physical quantities
    - temperature;
    - Metered volume of filter in certain time;
    - cumulative temperature since January;
    - number of weeks since last moment of flushing pipes;
    - temperature above 13 degrees Celsius;
    - week number;
- Biology
    - asellidae;

    - chironomidae;
    - cladocera;
    - cyclopoida;
    - parts of asellidae;
    - other organisms;
    - gastropoda;
    - harpacticoida;
    - hydrachnellae;
    - naupliuslarven;
    - nematode;
    - oligochaeta;
    - ostracoda;
    - prostoma;
    - turbellaria;
- sediment
    - asbestos fibers;
    - biofilm
    - detritus
    - iron bacteria
    - chalk
    - coal
    - pipeline material
    - other sediment
    - plant material
    - plug material
    - rust
    - rust detritus
    - shield of water flea
    - feces of asellus
    - fibers
    - sand
    - black material

There exist also a log file with information of calamities in the distribution network near the bag filters. It also contains the timestamps of the moments the network is flushed. Because the flushing disturbs the environment in the pipeworks this may be relevant information and is therefore added as metadata to each measurement; the time since the last flush is added

The different statistics and machine learning algorithms applied to answer the research question are:

- univariate regression;
- multivariate regression with gradient boosting regression;
- pattern recognition.

All these approaches try to find a relation between a measured quantity and one (or more) of the other measured quantities. For example, the measured number of asellidae is related to each other organism separately (univariate regression) or all other organisms together (multivariate regression and pattern recognition).

Univariate regression assumes a linear relation between to measurements and finds the line that resembles this relationship best. Multivariate regression considers a (non-linear) relation between one measurement and a linear combination of (some of) the other measurements.

When organisms are related in a food chain, it is expected that there is a correlation between their abundance at different moments; the one that eats the other will be more abundant when the other has become abundant some time before. This hypothesis is tested by a univariate regression with the measurements at a certain time and the measurements two and four weeks before. If some organisms show a strong correlation with organisms at two or four weeks before, this may indicate one of them is eaten by the other.

4.2.3 Results

In the univariate regression analysis, a regression is made between every biological quantity and every physical, biological and sediment quantity. The same is done for every sediment quantity. For every combination, this results in a value that describes the goodness of fit (the r^2); a good fit means a strong (linear) correlation. The measure r^2 is maximally 1 in which case 100% of the variance in the data can be described by the model. If it is 0.6, 60% of the variance in the data can be described by the mode. If the value is lower than 0.6 it is generally considered that the quantities have a poor linear relation with each other.

The r^2 of all the combinations can be visualized in a correlation matrix where rows and columns are spanned by the quantities. Every element in the matrix shows the correlation between the quantity of its row and its column. This matrix becomes very large so not all rows are shown here. But because the full matrix is symmetric, (almost) all information is presented; the matrices are shown in Table 6, Table 7 and Table 8.

The tables clearly show that the biology has positive correlation with temperature which corresponds to the observations that biology grows faster with higher temperature. On the other hand, no correlation is shown with any other physical quantity. Furthermore, it is shown that all biology correlate positive with each other. This means that if one organism is abundant, the others are as well (and if one is absent the others are absent as well). Finally, when looking at the correlation between biology and sediment quantities, a strong correlation is only found with sediment fractions formed by biology: shields of waterflee; feces of asellus and detritus.

A similar matrix is created that shows the correlation between the quantities at some time and the quantities at 2 or 4 weeks before. This might reveal a time lag exists between the rise and decline of different species. These matrices look similar to the ones shown in Table 6, Table 7 and Table 8 except that all values are smaller; this weaker correlation suggests no time lag between rise and fall of species.

Table 6: Correlation matrix between biological and physical quantities.

R Value	AverageWeekVolume	temperatuur	Cumulativetemperatu	Recodedtemperature	Spoelmoment	weeksSinceLastSpo
asellidae	-0.21	0.59	0.16	0.29	0.08	-0.14
chironomidae	-0.11	0.60	0.23	0.41	0.06	-0.16
cladocera	-0.04	0.51	0.09	0.30	0.07	-0.06
cyclopoida	-0.13	0.37	0.16	0.22	0.00	0.01
delen.asellidae	-0.11	0.22	-0.01	-0.13	0.06	-0.11
organisms misc.	0.06	0.30	0.20	0.17	0.03	-0.06
gastropoda	-0.06	-0.01	0.07	-0.08	0.09	-0.11
harpacticoida	0.04	0.33	0.57	0.07	-0.04	0.11
hydrachnellae	0.03	0.08	0.33	-0.11	0.00	0.03
naupliuslarven	-0.03	0.21	0.25	0.02	0.05	0.05
nematoda	-0.02	0.31	0.07	0.06	0.10	-0.11
oligochaeta	0.00	0.18	0.01	0.10	0.05	0.00
ostracoda	0.05	0.08	0.01	0.10	0.00	0.00
prostoma	-0.06	0.00	-0.03	0.03	0.02	0.06
turbellaria	-0.02	0.25	0.06	0.10	-0.04	0.08
organisms total	-0.11	0.57	0.14	0.27	0.07	-0.06

Table 7: Correlation matrix between biological and sediment measurements.

R Value	asellidae	chironomidae	cladocera	cyclopoida	delen.asellidae	organisms misc.	gastropoda	harpacticoida	hydrachnellae	naupliuslarven	nematoda	oligochaeta	ostracoda	prostoma	turbellaria	organisms total
asellidae	1.00	0.50	0.52	0.45	0.45	0.23	0.09	0.35	0.18	0.28	0.35	0.27	0.03	0.07	0.34	0.64
chironomidae	0.50	1.00	0.53	0.39	0.29	0.26	0.07	0.33	0.17	0.26	0.41	0.26	0.00	0.06	0.34	0.63
cladocera	0.52	0.53	1.00	0.57	0.36	0.23	0.17	0.50	0.23	0.37	0.47	0.34	0.06	0.08	0.46	0.81
cyclopoida	0.45	0.39	0.57	1.00	0.35	0.21	0.08	0.37	0.23	0.39	0.42	0.28	0.08	0.10	0.32	0.64
delen.asellidae	0.45	0.29	0.36	0.35	1.00	0.13	0.14	0.24	0.14	0.24	0.27	0.23	-0.01	0.08	0.26	0.55
organisms misc.	0.23	0.26	0.23	0.21	0.13	1.00	0.16	0.26	0.22	0.16	0.20	0.02	0.13	0.08	0.16	0.27
gastropoda	0.09	0.07	0.17	0.08	0.14	0.16	1.00	0.17	0.14	0.12	0.19	0.05	-0.02	0.10	0.14	0.14
harpacticoida	0.35	0.33	0.50	0.37	0.24	0.26	0.17	1.00	0.47	0.39	0.39	0.23	0.09	-0.02	0.33	0.52
hydrachnellae	0.18	0.17	0.23	0.23	0.14	0.22	0.14	0.47	1.00	0.30	0.31	0.15	0.08	0.06	0.20	0.34
naupliuslarven	0.28	0.26	0.37	0.39	0.24	0.16	0.12	0.39	0.30	1.00	0.20	0.13	0.09	0.01	0.27	0.41
nematoda	0.35	0.41	0.47	0.42	0.27	0.20	0.19	0.39	0.31	0.20	1.00	0.37	0.09	0.06	0.34	0.59
oligochaeta	0.27	0.26	0.34	0.28	0.23	0.02	0.05	0.23	0.15	0.13	0.37	1.00	0.09	0.12	0.33	0.42
ostracoda	0.03	0.00	0.06	0.08	-0.01	0.13	-0.02	0.09	0.08	0.09	0.09	0.09	1.00	-0.01	0.10	0.09
prostoma	0.07	0.06	0.08	0.10	0.08	0.08	0.10	-0.02	0.06	0.01	0.06	0.12	-0.01	1.00	0.02	0.13
turbellaria	0.34	0.34	0.46	0.32	0.26	0.16	0.14	0.33	0.20	0.27	0.34	0.33	0.10	0.02	1.00	0.51
organisms total	0.64	0.63	0.81	0.64	0.55	0.27	0.14	0.52	0.34	0.41	0.59	0.42	0.09	0.13	0.51	1.00

Table 8: Autocorrelation matrix of biological quantities

The multivariate analysis tries to fit a model that can predict a target quantity based on other quantities. The algorithm used here is Gradient Boosting Regression and the dataset is split in a train (80%) and test (20%) set. First it is investigated whether each respective biology and sediment quantity can be predicted by the physical quantities. The r^2-scores of these analysis were above 0.6 for some train sets, but always below 0.35 for all test sets. This shows that the training of the model was only possible for a few quantities, but in those cases the model is overtrained. A model is called overtrained when it described noise instead of the real underlying model.

It is also investigated for each quantity whether it can be described by all other quantities *together*. This yielded high training scores but only two quantities yielded test scores above 0.5; they are shown in Table 9. Interestingly, both the quantities are from the sediment, not the biology. In the same table it is also shown which quantities were most important in the trained model. The higher their value, the more important they are. The value should only be used to compare the importance of the quantities among each other; it is not a percentage of the total weight of all quantities. All the important quantities that can predict both feces assellus as well as detritus are biology and sediment quantities. While based on the univariate regression, temperature was expected to be important as well. On the other hand, it supports the observation of the univariate regression that abundance of biology and sediment is related to the abundance of the other biology species and sediment fractions.

Table 9: The R^2 score of the two target quantities feces of asselus and detritus. The column 'important quantities' lists in descending order the quantities that contribute most to the prediction of the target. Note that the listed value is not a fraction of the total but should only be used to compare which features are more important than others.

Name of target	R^2 score	important quantities	
		name	value
feces.asellus	0.78	detritus	0.1635
		overig.sediment	0.087891
		cladocera	0.084223
		asellidae	0.076544
		schildjes.van.watervlooien	0.067138
		nematoda	0.065758
		oligochaeta	0.047319
		hydrachnellae	0.046793
		zwart.materiaal	0.036137
detritus	0.68	ijzerbacteriën	0.147852
		feces.asellus	0.147686
		schildjes.van.watervlooien	0.097236
		oligochaeta	0.096329
		asellidae	0.068456
		turbellaria	0.064026
		propmateriaal	0.062613
		overig.sediment	0.047397
		cyclopoida	0.044509

Finally, a classifier algorithm has been applied to the data. This algorithm finds patterns in the relation between the different quantities. A selection of the results is shown in table 4.4. It shows for different quantities its maximum and minimum value given a temperature and number of weeks since the last flush. It shows for all quantities that if temperature is below 15 °C (13.7 °C or 14.6 °C) their maximum is (much) lower than when temperature is above 15 °C. Apparently, the system is asleep for temperatures below 15 °C. For Asellidae the maximum occurrence is lower if flushing is longer than four weeks ago, apparently the lack of flushing reduces aselidae counts. On the other hand,feces of asselus and sand increase when flushing takes longer than 7 or 9 weeks; postponing flushing increases them. Chladocera is not influenced by the flushing (like 75% of the quantities tested). Note however, that the lower limit of all four quantities is barely affected by the different conditions; this makes it hard to make very strong conclusions based on this analysis.

Quantity	Condition temperature	Condition weeks Since Last flush	Lower limit	Upper limit
Asellidae	<= 14.6	> 4	0.01	0.15
	<= 14.6	<= 4	0.03	0.58
	> 14.6		0.00	1.31
Feces asellus	<= 14.6		0.00	0.43
	> 14.6	<= 7	0.00	0.68
	> 14.6	> 7	0.20	0.84
zand	<= 13.7		0.00	0.01
	> 13.7	<= 9	0.00	0.06
	> 13.7	> 9	0.00	0.15
cladocera	> 11.3 ; <= 12.9		0.00	0.58
	> 12.9 ; <= 14.6		0.00	1.58
	> 14.6 ; <= 17.3		0.00	6.18
	> 17.3		0.01	19.18

The weak relations found in the above analysis are not in line with some strong presumption at PWN based on their experience. Therefore, the data is visualized to reveal seasonal effects; see **Error! Reference source not found.Error! Reference urce not found.**. It shows that values for both sediment and organism increase in spring, reach their maximum in summer and their minimum in winter. In autumn their values are widely scattered indicating that their relation with temperature is not strong in that time of year. This is also shown in Figure 14 where total number of organisms is shown as function of temperature for the different seasons. In winter, their count is always low; in spring their count increases with temperature and variation between the measurements is limited; in summer and winter their amount increases with temperature but the variation between the measurements is large. Indeed, the scatter in autumn is much higher than in spring while the same temperatures are reached. This can be explained by principles lend from chaos theory: in every spring the initial condition is very much the same because all biology dies or sleeps in winter time. Then, as time and temperature increase, they grow every year in a similar way *because they start from the same initial state.* However, in summer and even more so in autumn, the path towards their initial state differs because every spring differs. Due to their

difference in initial state, more outcomes are possible thus more scatter in the data is expected. So it seems that number of organisms can be predicted very well in spring based solely on temperature, but for summer and autumn it cannot. For those seasons more information is needed, possibly the course of the temperature in spring and autumn.

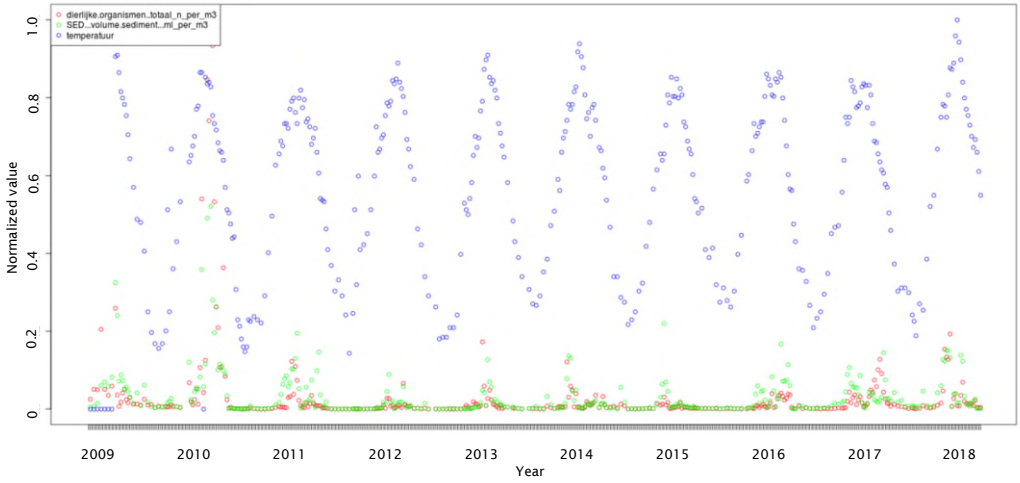


Figure 13 - The temperature (blue), total number of organisms (red) and total volume of sediment (green) plotted against time and normalized to their respective maximum. Note how temperature can be used to locate the seasons as it is maximal in summer and minimal in winter.

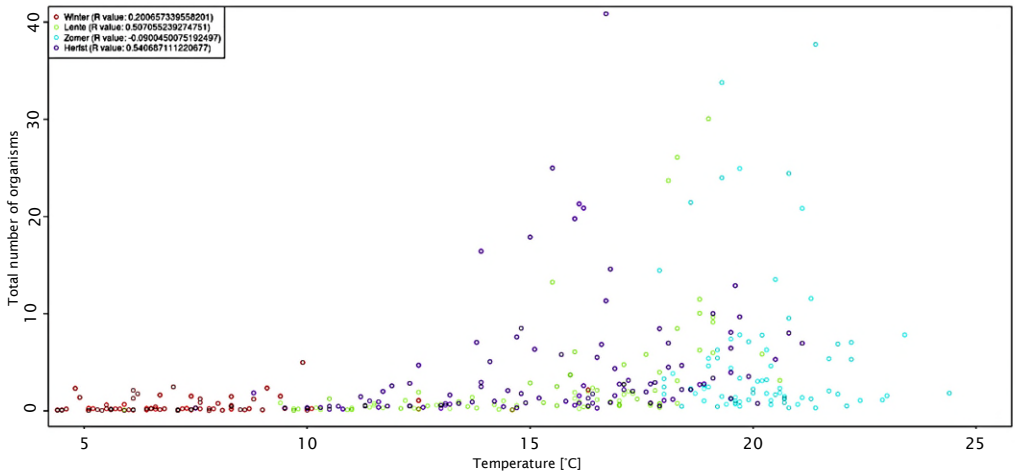


Figure 14 - The total number of organisms found as function of the temperature. The colors red, green, cyan and purple indicate the seasons winter, spring, summer and autumn respectively. Note that the temperature is an excellent proxy for the time of year. So the values that lie on the border of two seasons can be identified by their temperature; e.g. a value in spring with a temperature of 19 is at the end of the spring, while a temperature of 10 would indicate the begin of spring.

4.3 Relate treatment plant to tap measurements case study

4.3.1 Introduction and research question

The water in the distribution network changes as it travels from the treatment plant to the customer's tap; pH and iron concentration can change due to interaction with the pipe material while microbiological quantities change under the influence of several factors, including for example, temperature. It is important that the water meets all (legal) health requirements when it reaches the customers. Therefore, water companies in The Netherlands have a legal obligation to monitor the quality of their drinking water at the tap of their customers. For PWN this is done by HWL (Het WaterLaboratorium). They have a routine in which different quantities are measured regularly at different locations in the supply area of PWN. Some locations are measured with a fixed interval, others are only measured when they are randomly selected; the locations are evenly distributed over the area supplied by the different treatment plants.

Sometimes measured values exceed legal or company thresholds and measures need to be taken to prevent it from happening again or to restore to normal conditions. This is not always easy because it is not clear why thresholds are exceeded. Sometime, a calamity took place for which flushing of the network was required and thus temporally increased iron concentrations. Or when the residence time and/or temperature is/are high, biological quantities will exceed thresholds. In an ideal world, one would like to know what trajectory of the water from which the sample is taken, tracing it from the tap back to the treatment plant. However, this is currently not possible (real-time water quality modelling or digital twins are not available yet) . It is only possible to get a most likely path of the water based on calculations of hydraulic models. However, these models rely on a model of the network which is never exactly up to date with respect to the actual network. Furthermore, these models are often only pressure calibrated for standard situations with a given demand of the customers and often only as a day average. In reality, water demand changes over the course of the day and differs between days. For some zones in the network this means that depending on the time of day, they are supplied by different treatment plants, so-called mixing zones (see Figure 15).

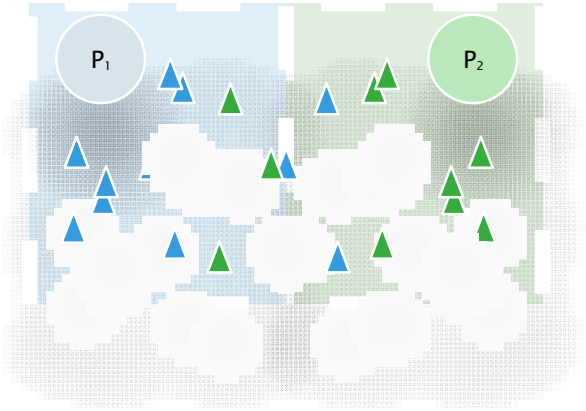


Figure 15 – Schematic picture of how two treatment plants P_1 and P_2 pump their water in the same network. On average, they will supply a fixed area; P_1 will supply the blue area and P_2 the green one. However, when demand is not average, one can find water that comes from P_1 in the area that is normally supplied by P_2; i.e. the blue triangle in the green rectangle. These triangles are in the so called mixing zone.

On the other hand, for most of the network, most of the time, hydraulic models give reliable results for residence time and traveled distance. When these models are coupled to the HWL measurements described above, one can relate residence time and traveled distance to the increase and decline of quantities like the bacteria counts and pH. If such relations exist, one can try to train machine learning models to predict the bacteria growth based on measured properties like residence time and temperature. This will increase the insight on the risk of bacterial growth in the network, which might be used as an indictator for the biological stability of the water leaving the treatment plant.

The research questions in this case study are:

- is it possible to identify the supplying treatment plant based solely on the HWL measurements;
- does the theoretic residence time of hydraulic models relate to the growth of microbiology in the network;
- is it possible to predict physical and biological quantities based on other quantities regularly measured at the pumping station (e.g. NH4, temperature or Aeromonas) .

The first two are addressed in this study while the last one could not due to budget constraints.

4.3.2 Approach and methods

To retrieve the treatment plant that produced each measured water sample, a clustering algorithm is used to assign each measurement to a cluster; the algorithm applied is k-means. Such an algorithm puts every measurement in an n-dimensional space where n is the number of measured quantities considered. For example, if only the color and EGV of a measurement are taken into account in the clustering, every measurement will correspond to a dot on two-dimensional plane with EGV and color as its axis (see Figure 17). For this study, it is assumed that each cluster will correspond to the supply area of a certain treatment plant. This is schematically depicted in Figure 15. The available

quantities retrieved from the database are shown in Table 10; the ones that are conservative are viable candidates to use with the cluster algorithm. Except for the measurements, the only input required for this algorithm is the number of clusters it should use.

To test whether the theoretic residence time relates to microbiological growth, the HWL data needs to be coupled to a computed hydraulic model of the PWN network. This is done based on the distance between the RD-coordinate of the measurements and the RD-coordinate of the nearest node in the hydraulic model. Furthermore, the HWL measurements at the tap need to be linked to the measurements of its supplying treatment plant. This way the difference between plant and tap can be calculated which reveals the increase and decrease of certain quantities. The hydraulic model is used to determine the supplying treatment plant and to compensate for residence time in the network. Not all the measurements are used in this analysis because the hydraulic model is only calculated for a certain supply share of each treatment plant to a supply region. Only measurements are used for the days when the shares of the supplying plants are the same as in the hydraulic model. Furthermore, only measurements where the model predicts a 100% supply of a certain treatment plant are used; measurements in mixing zones are neglected.

Table 10: The quantities extracted from the HWL database and whether it can be used to determine the supplying treatment plant, that is if it is conservative. Furthermore, it is indicated whether the quantity is relevant for microbioloiogical growth.

Type	Name	Conservative in the network	Relevant for growth of microbiology
Physical quantity	NH4	no	maybe
	EGV	yes	no
	Iron	sometimes, changes with iron cast pipes	no
	Color	yes	no
	Temperature	no	yes
	Turbidity	no	no
	pH	no, especially changes in contact with cement lined pipes	maybe
Biological quantity	DCT	no	yes
	Aeromonas	no	yes
	Colony number	no	yes

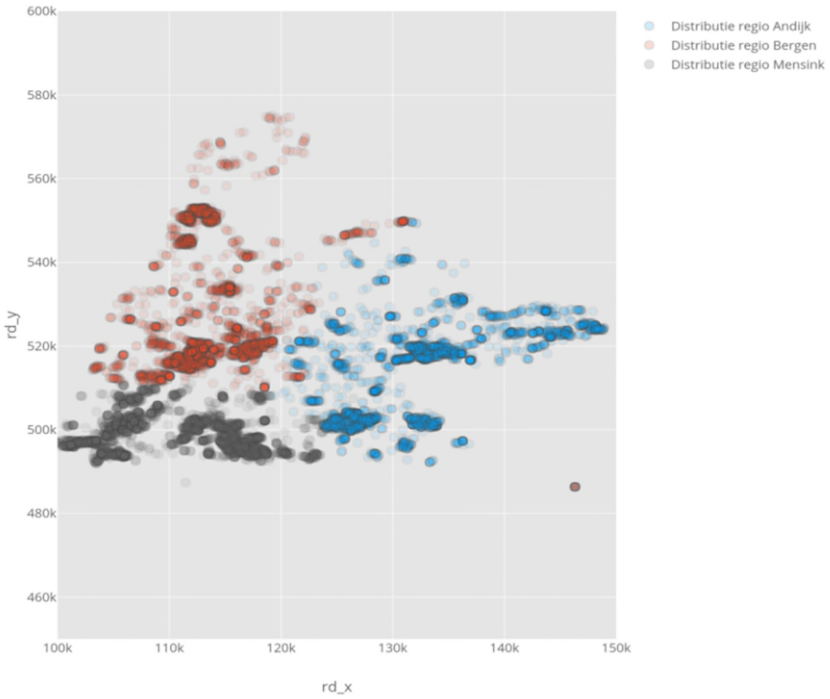


Figure 16: The locations of the measurements of the HWL database. Horizontal and vertical axis are RD-coordinates in x- and y-direction. The north of the province of Noord-Holland (The Netherlands) can be recognized although the image is a little stretched horizontally. The more opaque the marker, the more measurements are taken at that area. The colors black, red and blue indicate the supply area of treatment plant Wim Mensink, Bergen and Andijk respectively. Note that these supply areas are used by HWL to determine where to take the measurement and do not nessecarily comply to the actual supplying treatments plants

4.3.3 Results

4.3.3.1 Determine supplying pumping station by clustering

The result of the clustering with 4 clusters and EGV and color as quantities is shown in Figure 17. The clusters are not well separated except for the long tail of higher color values. The quality of the clustering can be assessed by looking at the portion of the datapoints that lies at the border of two clusters; those data points do not clearly belong to one or the other cluster. A metric that describes this is the silhouette score; the higher the score the better the separation. An overview of silhouette scores for different number of clusters and sets of quantities is given in Table 11. It can be observed that the highest score is found when 4 clusters are selected with color and EGV as quantities; using iron gives the same score. Apparently, adding iron does not give better cluster results. Furthermore, when adding pH as quantity, silhouette score drops; pH makes it harder to make separate clusters. Finally, in all cases, 4 clusters give the highest score. Based on these results, the remainder of the analysis is done with four clusters and EGV and color as quantities.

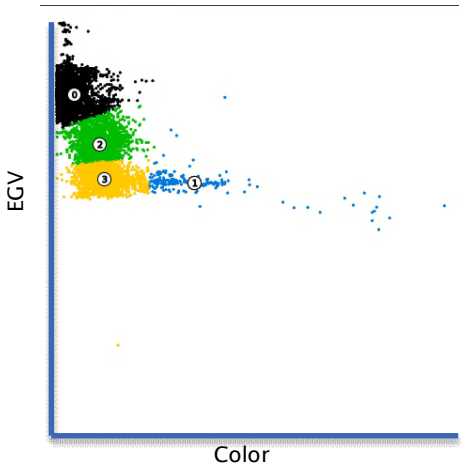


Figure 17: graphical representation of the clustering when 4 clusters are used and EGV and color are used as quantities.

Table 11: Overview of the silhouette score when using different sets of quantities and different number of clusters for the clustering. The higher the score, the better the separation between clusters.

Quantities	Number of used clusters	Silhouette score
Color, EGV	3	0.47
Color, EGV	4	0.49
Color, EGV	5	0.42
Color, EGV, pH	3	0.4
Color, EGV, pH	4	0.41
Color, EGV, pH	5	0.34
Color, EGV, iron	3	0.47
Color, EGV, iron	4	0.49
Color, EGV, iron	5	0.42
Color, EGV, iron, pH	3	0.41
Color, EGV, iron, pH	4	0.42
Color, EGV, iron, pH	5	0.34

Plotting the measurements of the four clusters in geographical space gives an idea whether the clusters coincide with supply regions. Indeed, Figure 18 shows how the clusters have their own distinct geographical region. By comparing this figure to Figure 16 one can assign the orange cluster to Bergen (purple dot), blue to Wim Mensink (black dot) and green to Hoorn and Andijk (big red and blue dot); Hoorn and Andijk both receive their water from Andijk treatment plant. The red cluster corresponds to the cluster with distinct color values; it is suspected that this is water coming from Bergen. The distinct color occurs because sometimes extra water is required and extracted from the dunes which has more humic acids that gives the water a different color.

The clustering has some obvious errors. It should be noted that water from Wim Mensink can never reach the distribution region of Bergen and Andijk because the networks are physically separated. Nevertheless, at coordinate (125,000, 550,000) that

lies in the distribution region of Bergen and Andijk, some measurements are assigned to cluster 0, which is the cluster associated with Wim Mensink water. As said, this is physically impossible. But the clustering algorithm is not aware of this physical limitation, so it can make this error. The reason of this error can be understood from the clustering result shown in Figure 17; the clusters of Wim Mensink (green) and Bergen (yellow) are adjacent. Consequently, measurements with EGV values at the border of the clusters are easily assigned to the wrong cluster (in this case assigned to Wim Mensink instead of Bergen). On the other hand, the cluster Bergen (yellow) and Andijk/Hoorn (black) are far away from each other which makes it easy for the algorithm to choose the right cluster. The model would obviously improve when it us fed with extra data such that it knows for each measurement which pumping stations is physically able to supply the location of that measurement.

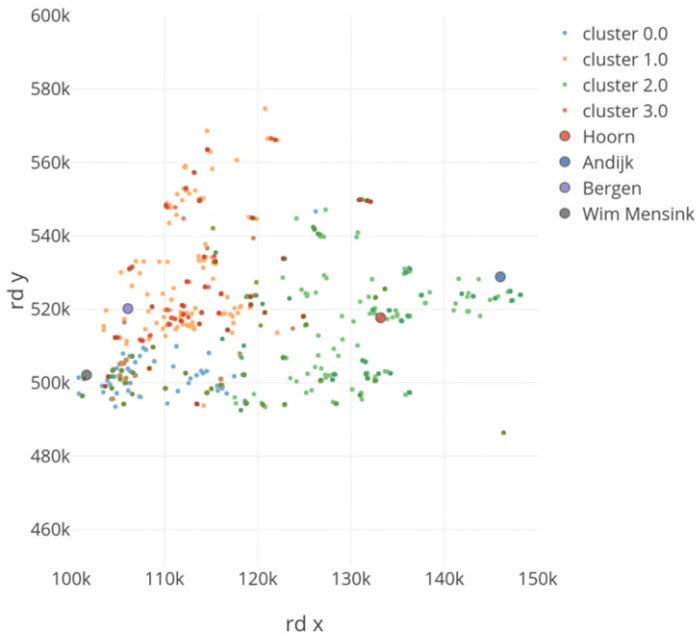


Figure 18: each measurement is plotted here as a small dot; the north of Noord-Holland can be distinguished although it is little inflated in horizontal direction. The color of the dots indicates the clusters the measurement belongs to based on the clustering described above. The four pumping stations are shown here as well: Hoorn (red), Andijk (Blue), Bergen (purple) and Wim Mensink (black). Horizontal and vertical axis depict the x- and y-RD-coordinates.

In the preceding, clusters are assigned to pumping stations. The accuracy of this clustering can be qualitatively investigated by looking at different ratios of the amount of water the pumping stations deliver to a certain area. The north part of the whole PWN supply area is supplied by Bergen, Hoorn and Andijk, while the more south part is supplied by Wim Mensink and Hoorn (which has the same water as Andijk). When this ratio changes, it is expected that the water of one pumping station penetrates deeper into the supply region. Indeed, this is shown in Figure 19 and Figure 20 for the area that is supplied by Bergen and Andijk respectively. The figures show that if a station supplies a larger portion of the water, its measurements are found further away. Even to areas that are supplied by the other pumping station on an average day. Also, the peaks of the PDF's in those figure decreases when their portion increases, meaning the

PDF is more spread out. This shows that the clusters are well suited to determine where the water sample has been treated.

The same analysis is also done for the region that is supplied by Wim Mensink and Hoorn. The results are similar (not shown here) but less pronounced. This is partly explained by the fact that these clusters share a border in the clustering space; values at the border are easily assigned to the wrong one. Because the clusters of Bergen and Andijk have more distinct values for color and EGV they can be distinguished better.

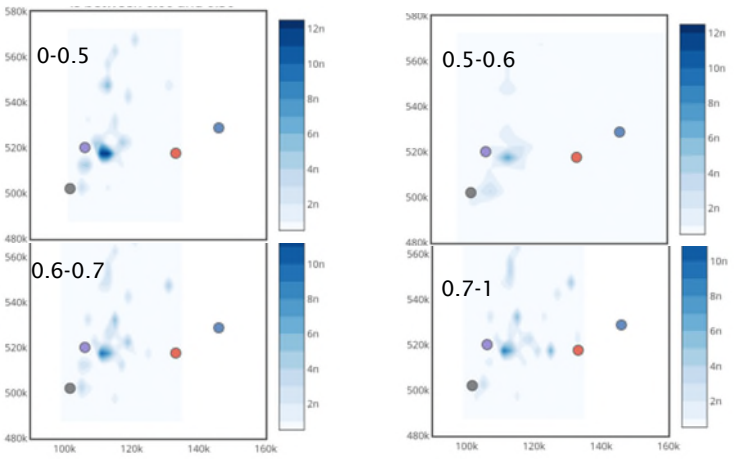


Figure 19: Contour maps of the probability that a measurement of the cluster associated with treatment plant Bergen (larger purple dot) occurs at a certain RD-coordinate; the darker the blue, the higher the likelihood. Each contour plot is based on the measurements at days where Bergen supplied a certain fraction of the water to this region; the fraction is shown in the top-left of each plot. The probability is shown as a Probality Density Function so the peaks can be compared to each other. The other station supplying this region is Andijk; its results are shown in Figure 20. Note that the maximum of the color bar at the top two figures is slightly larger than the bottom two.

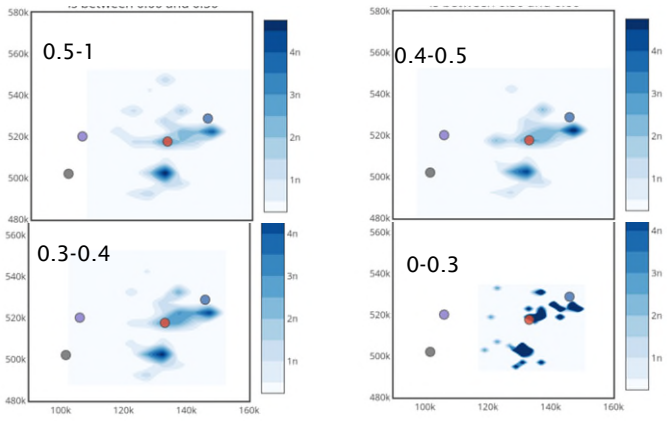


Figure 20: same as Figure 19 but now for the Pumping station Andijk (larger red blue). Bergen (purple dot) and Andijk supply the same region. The supply fraction of Andijk to this region is given at the top left of every figure. Note that the maximum of the color bar at the top two figures

is slightly larger than the bottom two. The bottom right figure suffers from lack of data points to determine a smooth PDF.

The above results show that the clustering can be reliable used to determine the treatment plant of origin of the water sample. This enables the investigation of the differences in the water samples between the different treatment plants. From a practical point of view, the most relevant quantities are the biological ones. It is for example interesting to see whether the distribution of the measurements differ between the treatment plants. This is shown in Figure 21. There the Probability Density Functions (PDF) of different biological quantities are shown. A PDF is a histogram of the data, while normalized such that the area-under-the-curve is 1. This allows comparison of different distribution with each other despite a different number of measurements; peaks with the same heights have equal probability. Furthermore, a lower peak always implies a broader distribution. In Figure 21 a PDF is made for every treatment plant for different biological quantities, namely, cell counts of living Low Nucleic Acid content (LNA) bacteria; cell counts of living High Nucleic Acid content (HNA) bacteria; and ATP concentrations. These quantities are chosen because they yielded interesting PDF's.

For these quantities, the different treatments plant are distinct. This supports the above conclusions that clustering is performed properly. The reason for the distinct PDF's is most likely due to the different treatment methods used at the different treatment plants, although this cannot be concluded from this data.

Another observation is that the distributions seem to follow a Poisson distribution or possibly a log-normal distribution. This is quite surprising giving the number of quantities that are of influence like pipe material, temperature and residence time that are similar for the measurements in all regions. Apparently, the combinations of these factors that determine the DCT and ATP values are very distinct among the supplying treatment plant and its network. Consequently, one can look at each region supplied by a treatment plant as a black box model of which the outcome of the biological quantities are given by a certain PDF. If the distribution is either a Poisson or log-normal distribution it could tell something about the underlying system. But that needs further study and is out of the scope of this subject.

The PDF of the biological quantities describe the system. If a measurement falls in the first or last 5% (or some other threshold) of the distribution, one could consider this as a suspicious measurement because it is not following the normal behaviour of the system. Detecting these measurements can lead to an early detection that something is wrong in the system, that is, the treatment plant or network. This can be done *before* some critical values are reached and damage is done; it allows for pro-active maintenance.

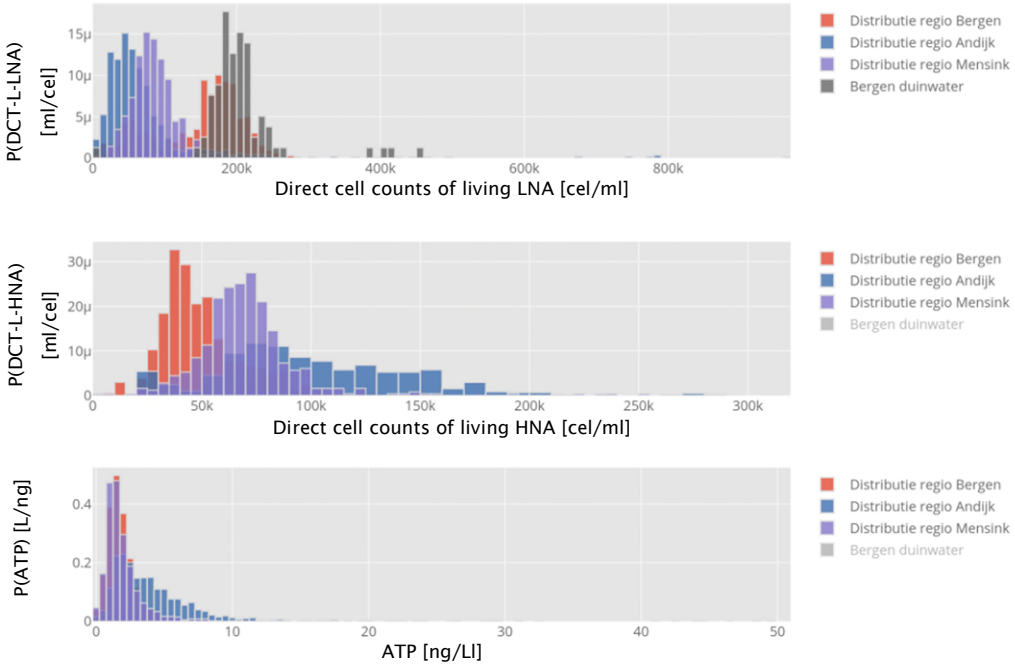


Figure 21: The probablity density function (PDF) of differen biological quantities. From top to bottom: Cell counts of living (L) Low Nucleic Acid content (LNA) bacteria (DCT-L-LNA); Cell counts of living (L) High Nucleic Acid content (HNA) bacteria (DCT-L-LNA); and ATP concentrations. One PDF is derived for each treatment plant: Bergen (red), Andijk (Blue), Wim Mensink (purple) and Bergen dune water (Black). The Bergen dune water is not shown in the lower two figures to keep the figures clearer. By defintion, PDF have a area-under-the-curve of one; this means that a higher peak means a more compact distribution.

4.3.3.2 Coupled measurements and hydraulic model

For this analysis, the measurements are selected that

- have supplying fractions similar to the one used in the hydraulic model;
- supplying fraction of any of the pumping station are 0.9 or more (no mixing zones).

For these measurements, it is safe to say that the pumping station can be reliable determined from the model. This enables comparison of the quantities at pump station and tap and thus their increase and decrease in the network. Furthermore, the residence time of the water can reliable be determined by using values calculated by the model. Consequently, it can be investigated whether biological quantities increase with residence time.

In Figure 22, the relation between measurements at treatment plant and tap are shown for EGV and color. This shows a strong linear correlation for EGV and to a lesser extent for color. Apparently, EGV does not change in the network which is in-line with expectations. This confirms results from the previous section in which clustering based on EGV and color yield reliably the supplying treatment plant of the water sample. As a matter of fact, Figure 22 shows that values of EGV and color for Bergen and Andijk are

well separated as was also seen in the clustering analysis above. It should be noted that the water from Wim Mensink has a larger variance around the direct proportional relation. This is interesting, because it is not expected that the EGV of the water changes during transport of the distribution network. Because it is not likely that any ions will dissolve or precipitate in the network (note that EGV is a measure of ion-concentrations). Another, more likely, explanation might be that it is harder for the Wim Mensink data to couple the correct value at pumping station to a measurement at the tap. The wider deviation from the direct proportional relation also explains why the clustering result on the supply area shared by Hoorn and Wim Mensink are not as clear as the results for the supply area of Andijk and Bergen. Finally, Figure 22 shows that the clustering results in the previous section can be improved by not clustering on the *values* of EGV and color, but on their difference with the supplying region.

In Figure 23, it is illustrated how different biological quantities change with their residence time in the network. Most samples seem to be centred around zero which mean they do not grow in the network; this seems to be independent of the residence time. Except for the data from Andijk water. The increase of the colony number seems to have maximum at 50 hours of residence time. The increase of DCT-L-LNA and DCT-L-HNA increases with residence time. Apparently, Andijk water is less biological stable compared to the other locations. This can be caused by the difference in treatment or differences in the network. It should be noted that the samples of Bergen, Wim Mensink and Andijk do not have the same spread in residence time; the samples with residence time above 100 hours are all from Andijk. This make the aforementioned observations qualitative and preliminary; more research is required to draw solid conclusions.

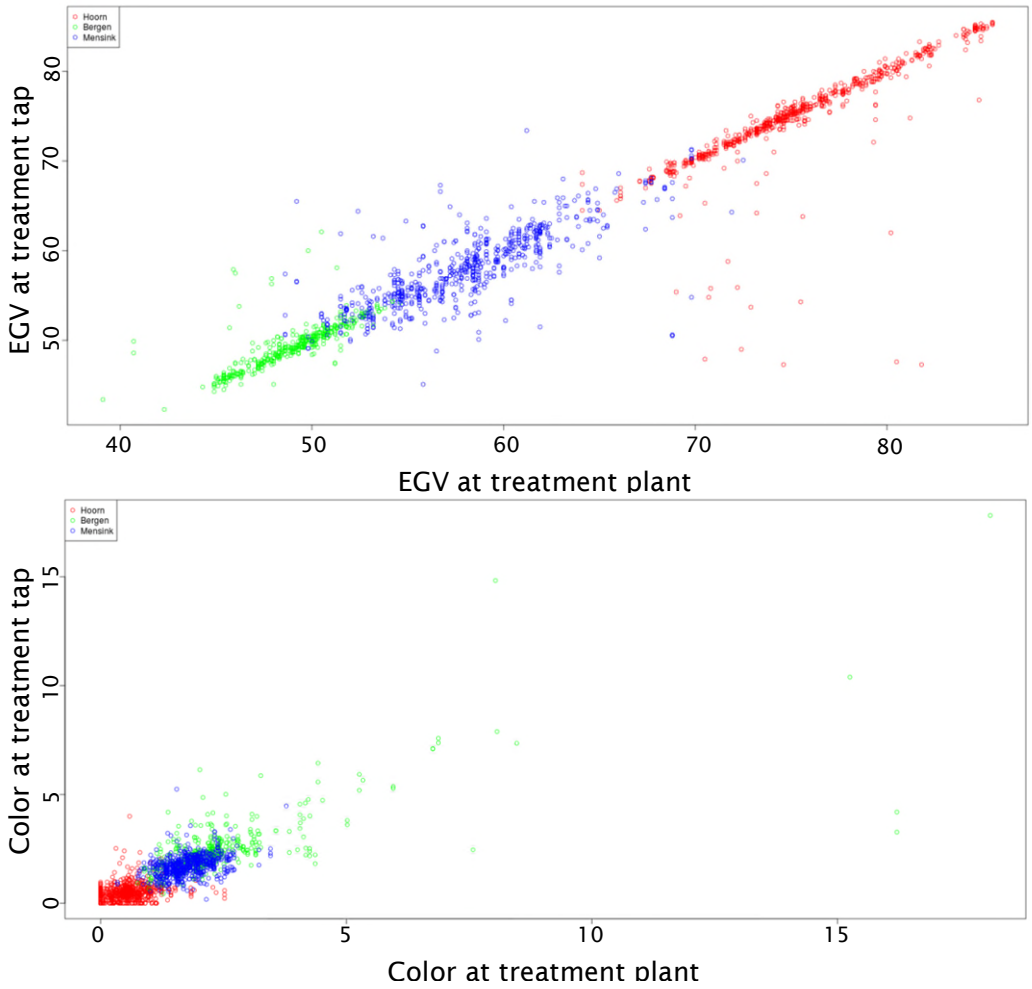


Figure 22: These figures show the values of EGV (top) and color (bottom) measured at the treatment plant (horizontal axis) and tap (vertical axis). The different colors indicate the treatment plant the water oirginate from according to the hydraulic model: Hoorn red, Wim Menksink blue, Bergen green.

It is tempting to compare the results of the clustering with the results shown here but this is not straightforward. The PDF's shown in the previous section (Figure 21) are based on *all* measurements, while here only a selection is used. Furthermore, here results try to illustrate the change of the quantities between treatment location and tap while in the previous section the *values* of the quantities are investigated.

Nevertheless, some interesting observations are made when comparing the DCT measurements. The PDF's in Figure 21 show that high DCT-L-LNA (>100,000) have highest likelihood to occur with Bergen water, while Figure 23 show that most sample have increased with less than 50,000. This suggests that the high values found in Bergen water (Figure 21) are caused by the water coming out of the treatment plant and do not develop in the network. Regarding the DCT-L-HNA of Bergen and Wim Mensink water Figure 21 shows their values are nearly all below 70,000 while Figure 23 shows they increase less than 10,000 (<15%) in the network. On the other hand, DCT-L-HNA of the water of Andijk is about half of the time above 70,000 while it increases up to 50,000 in the network. This suggests that higher values of DCT-L-HNA in Andijk water

are formed after it leaves the treatment plant. This is in-line with current knowledge and understanding at PWN; Andijk water is less biological stable.

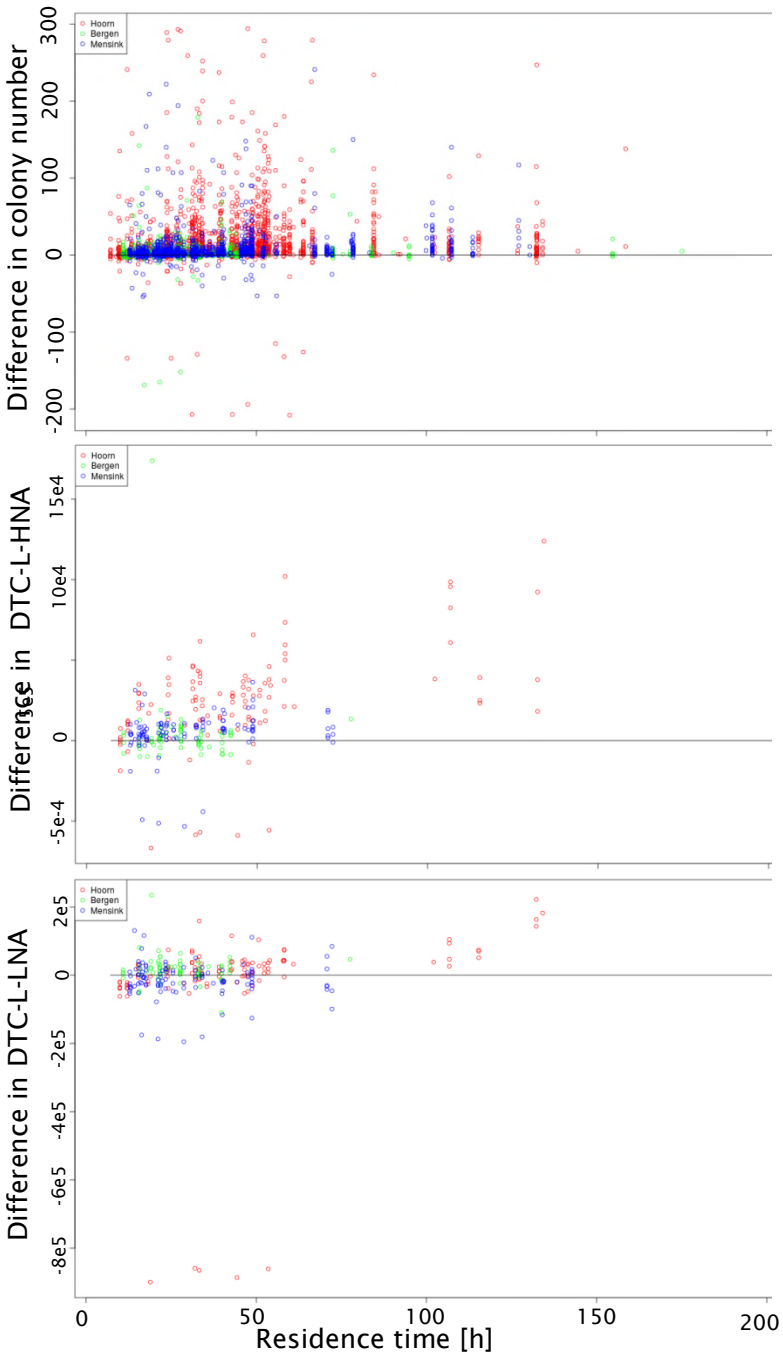


Figure 23: The difference between the values at the treatment plant and at the tap are shown as function of residence time for different bological quantities (top to bottom): colony number; cell counts of Living (L) High Nucleic Acid content (HNA) bacteria; and cell counts of living Low Nucleic Acid content (LNA) bacteria.

4.4 Conclusion on the results

In this pilot at PWN, two case studies are performed. In one the data of bag filters from the distribution network are analysed. In the other case study, water quality measurements at the tap of customers are analysed and combined with measurements at treatment plant as well as data of a hydraulic model.

The data of the bag filter case consists of measured content found in bag filters that are placed in the distribution network of PWN. It is divided in measurements of organisms as well as measured fractions of sediment. The aim of the study is to find out whether this data could tell something about food chain that supposedly exists in these bag filters. Based on univariate and multivariate regression no food chain can be distinguished. Some other relations are distinguished however, i.e., all organism become more abundant with temperature. As a consequence, the occurrence of all organisms correlate with each other. By using classification algorithm it is also revealed that for some organism their occurrence increase with flushing, while for others it decreases. Finally, seasonal effects are evaluated based on the cumulative sum of all species. For all seasons, it is shown that the organisms are more abundant at higher temperature. In spring the relation between temperature is quite strong, but in autumn where temperature is similar, much more variance is observed. This can be explained by chaos theory; every year, spring begins at more or less the same state because in winter all organisms die or sleep. Because every spring has the same starting point, abundance of the organisms has a predictable relation with temperature. However, autumn starts every year with a vastly different state caused by how both spring and summer have developed. For example one year organisms diminish in a cold summer leading to low organisms count in an autumn that may be warm. Next year, summer is hot and as a consequence, autumn has much higher organism counts while it is the same temperature as the year before.

It is concluded that the dataset of this case study is too low in quality to acquire detailed information like food chains or predict the occurrence of a species with a regression model. Possibly, the quality can be improved with higher sample frequencies. But it is more likely that viable information is missing, for example, there is no information on how much food is present in the system. It is also possible that the biology at the bag filters is mainly determined by upstream circumstances. In that case, such information should be included as well.

In the other case study, water quality measurements at the tap are successfully clustered based on EGV and colour. The clusters coincide with the supplied area of the different treatment plants: Andijk, Bergen and Wim Mensink. This is clearly shown for the distribution region that both Andijk and Bergen supplied. This region does not always receive water from Andijk and Bergen in the same ratio. Therefore, measurements are selected based on the supplying ratio of the treatment plants. When more of the water is supplied by Andijk, the measurements that fall into Andijk cluster lie closer to Bergen; at the same time, measurements in the Bergen cluster lie further away from Andijk. The clustering is also performed by using iron and pH measurements as well. However, using iron does not improve (or deteriorate) the clustering result. Using pH gives worse result for the clustering and should therefore never be used for clustering. This is to be expected because pH is probably not conservative in the system due to interaction between water and pipe material.

From the clusters, a Probability Density Function (PDF) of some microbiological quantities is derived. They reveal a very distinct distribution between the different supplying treatment plants. Apparently, the different treatment plants and/or their network yield very different microbiology in the network. For most quantities, the tail of their PDF was longest for Andijk, i.e., water from Andijk is most likely to have high values. Except for cell counts of living Low Nucleic Acid content (LNA) bacteria. In that case Bergen has higher likelihood for the highest values. Andijk seems to have less biological stable water.

These PDF's are well described by a Poisson or log-normal distribution; they can therefore be used to describe the system. This can be utilized to detect suspicious measurements; measurements that fall in the minimal and maximal 5% (or another threshold) of the distribution do not comply to the normal behaviour of the system. They are possibly a measurement error or, more importantly, a consequence of something being out of the ordinary in the system, that is the treatment plant and/or its distribution network. Because the measurement can be qualified as *out of the ordinary* it can be utilized as an early warning mechanism for problems in the system. That is, measures can be taken before any thresholds are exceeded or problems have occurred.

These measurements are also coupled to a hydraulic network model of the PWN network; every measurement is linked to the closest node in the model. From the information of the model, the supplying treatment plant is determined for each measurement as well as its residence time. The model was solved for a certain distribution ratio of the supplying treatment plants; only measurements that have the same distribution ratio are used in this model. The selection is narrowed down further by selecting measurements at location that have a supply ratio of 0.9 or higher. For this selection there is high confidence the supplying treatment plant and the residence time are reliably coupled to each measurement.

From this data it is concluded that the EGV and colour are indeed very distinct between the treatment plants, especially between Bergen and Andijk. This supports the results of the clustering. Furthermore, the EGV has a direct proportional relation between treatment plant and tap; it is conservative in the network. Because this holds for a lesser extent for colour as well this explains why the above described clustering of measurements based on colour and EGV works very well.

Furthermore, it is shown that most microbiological quantities do not change more than 10% in the distribution network. However, the increase of colony number for Andijk water is sometimes 100 while most of its measured values are below 40 (not shown). This indicates that they grow significantly in the network. Although it should be noted that the sampling plays a role as well. Before taking the sample, the sample point should be flushed sufficiently long such that all water in the internal plumbing *and* in the service main are refreshed. Only then is the sample representative for the water in the distribution network in the street. If this is not done properly, cell counts are higher than those in the distribution net itself. Furthermore, DCT-L-HNA of Andijk increases up to 250,000 for residence times above 100 hours, while most of the measurements are below 100,000. This supports what is already known at PWN: Andijk water is less biological stable.

Finally, results show that residence time does not have much influence on the microbiological growth in the system for Bergen and Wim Mensink. For Andijk, the colony number is maximal around 50 hours of residence time; longer residence times yield lower colony numbers. Only the DCT-L-HNA of Andijk water increases more with residence time while its increase of DCT-L-LNA only seem to increase if the residence time is above 100 hours. Consequently, from the data is concluded that residence time can only for a small part explain microbiological growth in the network. E.g., the water quality of the supplying treatment plant is of much more importance.

It is concluded that the dataset of measurements at the tap combined with log-files of supplying ratios of the treatment plants to each area its supplies; water quality

measurements at the treatment plants; and a hydraulic model are of good quality. This made it possible that differences in water quality of different treatment plants have been revealed, as well as the change of quantities in the distribution net. In this study, these are analysed in a more qualitatively matter due to the explorative nature of the study. It is recommended that its findings are investigated further to obtain more rigorous prove. This could be used to make a real impact on the business. For example, one could use the data to prove only a certain percentage of the customers have a water quality above a certain threshold. Or this can be disproved which would urge to take measurement to still reach this threshold

4.5 Evaluation of the process

This evaluation is based on participation in progress meetings, spending time with the project team and two rounds of interviews (see Table 56Table 4) with the main participants in the project.

Table 12: Interviewees

Name	Organisation
Peter Schaap	PWN
Arjen Schimmel	PWN
Lianne van der Laan-Smit	PWN
Laurens van der Drift	Phinion
Martin Korevaar	KWR

The PWN case was selected after an inspiration session that took place in September 2018 at the PWN office. At the request of PWN, an independent data scientist was added to the team. As in the other projects, it was decided to do the work in situ at the office of PWN. The project team was located near the Business Intelligence Competence Center where also PWN's 'data lake' is managed.

Two cases were selected for the project: (1) analysis of bag filter sediments to gain insight into food chains of the organisms living inside the distribution network. (2) gaining insight in changes in water quality parameters potentially caused by different supply patterns and transit time of water in the distribution network.

In the first part of the project, the problem owner and main contact point for PWN was visiting the UK, so meetings took place by Skype. It took one day to gather the necessary data for both cases. The problem owner could inform the project team about the data and how it could be interpreted.

Initial results of the bag filter data proved not very promising, with weak correlations. In hindsight this could have been a reason to drop the case and proceed to the second case, but it was decided to continue with the univariate analysis to see if results could be obtained through different approaches. In the end, the results did not improve significantly, although some insight was gained in seasonal variation of population sizes of different species. It was noted by the respondents that the role of project leader, with a clear view on the project planning, was lacking in the project team, which leaded to more time being spent on the bag filter problem at the expense of the second case.

The data for the second case was readily available, but not in a suitable format for the analysis, so some time had to be invested in data preparation. Moreover, the spatial coordinates between two datasets was not consistent. It would have taken a lot of time to fix this, but through matching with a third dataset, this could be circumvented. The issue was pickup by the BICC and the source data (which was from an outside party) was adjusted based on this issue.

The first results of the second case were available shortly before the final project meeting and proved more promising than the results of the first case. It was agreed in the final meeting that PWN and Phinion would spent some additional time to elaborate on these initial results.

4.6 Success and fail factors

4.6.1 Success factors

Working in situ, close contact with problem owner. As in the other cases, working at the location of the problem owner proved beneficial to the project. The collaboration between three different organisations (PWN, Phinion, KWR) went well, with a good division of roles and competencies, with Phinion focusing more on the statistical part and KWR on data preparation and modelling.

During the first part of the project, the main contact for the project team was visiting the UK, so meetings had to be organized through Skype, which was not ideal. Generally, working at PWN allowed for quick resolving of issues with data or sharing results and discussing next steps. However, travel time to the PWN office and adjusting to a new working environment can also put a strain on the researchers.

Involving an outside data scientist. Involving an independent data scientist, with no in-depth knowledge of the water sector can generate fresh insights into existing issues. People working in the water sector are likely to be influenced by the 'dominant culture' including existing knowledge about cause and effect. An outsider with a strong focus on data, may be able to falsify existing claims or see connections that insiders will easily overlook.

Iterative process, regular updates, possibility to adjust. Respondents noted that sharing results often and adjusting the research process accordingly can be beneficial for the outcome these kinds of projects. In this case, initial results were interpreted and matched with additional domain knowledge to identify additional analytical steps. Although there is always uncertainty whether these additional steps will yield better results, the opportunity to do so is helpful. An iterative process has the downside of delaying the overall process, resulting in time limitations at the end of the project.

Quick access to data. It took the project team only one day to gather the necessary data for both cases. The availability of a 'data lake' and the close connection to the BICC as well as their flexibility and responsiveness proved essential.

4.6.2 Fail factors

Some issues with the data. Although the data was quickly available, it was not always well suited for the purpose of the analysis. The bag filter data came in six separate spreadsheets that had to be manually merged and the water quality data had inconsistencies in the spatial coordinates. Also the bag filter data was not detailed enough to do the kind of analysis required for the case.

According to the project team, the data management is in transition and at this point sufficient for a research project. For actual implementation a more structural approach is needed. Typically this would involve first sitting together with the IT department to develop a data interface.

Focus on finding positive and 'surprising' results. Data analysis can have the tendency to apply increasingly complicated statistical methods to find patterns or causality in the data. This conflicts with the traditional scientific approach of formulating hypotheses and testing them on a data sample. In this case, a relation between variables is identified in the data and a post hoc hypothesis is formulated of what the relationship means.

Second, there is an implicit (and sometimes explicit) expectation that data mining needs to yield surprising and new results, instead of confirming existing knowledge. This would be a surprise in itself, because when data mining would indeed lead to surprising results, our current state of knowledge would be seriously flawed. Substantiating or nuancing existing knowledge with data should be regarded a positive result in itself.

Results difficult to follow for non-technical audience. Results of data analysis can be complicated. Usually there is an underlying process that is complicated in itself (such as evolutionary development of microbial species) which is then analyzed using complicated statistical methods. This became apparent in the progress meetings and final meeting of this project. Although the results were nicely visualized with graphs and figures, these were difficult to interpret for a non-technical audience. The elaborate discussion that followed was mainly between the problem owner and the project team about the interpretation of the results. When results are presented to a broader audience, more emphasis should lie on the broader conclusions and applicability and less on the technical and scientific details.

Too little focus on overall planning. Too much time was invested in the bag filter case, as in the beginning of the project there was too little focus on the overall planning. This can also be seen as a potential pitfall of an iterative approach (and perhaps of any research project). It can also be the result of a focus on positive results (e.g. strong correlations), while weak or no correlations can also be considered a result but are not particularly helpful for decision support. Finally, once time and effort is invested, it can be hard to drop a case and 'lose' the invested effort.

4.7 Conclusions on the process

For all organisations involved, the project has been a good learning experience. PWN gained experience in data science projects as well as working more closely with outside organisations (KWR and Phinion). It was suggested in the interviews that these cooperative models are the way forward for PWN on the issue of data science, since it employs no data scientists of its own.

The project was also a good opportunity for a better connection between ICT and data at PWN and the water expertise. Through the progress meetings people from BICC and information management did learn more details about the core business of PWN. This was also the case for the data scientists involved in the project, who had the opportunity to learn about the different water quality and water quantity issues in the PWN distribution network.

The results of the project (in both cases) has emphasized the need for more (and more detailed) data collection and for more centralized and standardized data management.

There is a data infrastructure (data lake) though which data is accessible, but data ownership is still fragmented and the way data is collected, validated and maintained is more or less up to the individual data owner.

This project was set up both as a research project and an implementation project, but in the end leaned more towards the research side. The goals was to see what can be learned from two categories of data that PWN has available and how this can influence decision making. In the case of the bag filters, it was hypothesized that food chains could be identified and that the results could direct future research. In the second case, the goal was to explain differences in water quality by comparing the share of supply from two pumping stations. If this relation could be established, this would justify the addition of meta data to the monstering locations and results. So, in short, one of the primary aims of doing data science is to justify systematic data collection.

5 Discussion

5.1 Comparison between cases

The three cases that are part of this study have a very similar set up. The goal in all cases is to establish a chain in which data is gathered, prepared, analysed, interpreted with the goal of supporting decision making. Ideally, this would lead to a concrete and applicable tool.

Also in each of the three cases a collaboration was set up between a drinking water company, an independent data scientist and KWR. The work was carried out mostly at the drinking water company in close collaboration with the problem owner and internal experts. The cases followed roughly the same approach with a sequence of activities from data gathering and preparation, modelling and statistical analysis and interpreting the results to the development of a concrete tool.

The cases differ in the extent to which the outcomes are readily operationally (or strategically) applicable. In the Oasen case, a concrete tool was developed (EDWARD) which can be applied to estimate the future peak factor, based on climate and behavioural data. In the PWN case, an exploration was done of water quality data from bag filters and, in a second case, water quality was linked to supply from two different production facilities. The results of these analyses did not lead to operational tools, but underlined the need for systematic data gathering and management and consistent labelling of measurements with meta data. In the WBG case, machine learning models were developed to predict the failure of water meters. These models are still in an early stage of development, but have shed light on the importance of systematic gathering of water meter data and consistent labelling with meta data. It has also resulted in a closer inspection of the data gathering process from water meters by WBG.

The difference in outcomes (the extent to which applicable results have been obtained) can be largely attributed to the availability and quality of the data, as well as the level of understanding of the issue at hand. On the one hand there is the Oasen case, were all necessary data was available, in very good quality, and there was a thorough understanding of the issue due to the fact that it had been modelled in previous research already. On the other hand, there is the WBG-case, were crucial data was not available and there were multiple issues with labelling and data preparation. Also, the issue at hand was complicated and yet unsolved (the causal mechanisms behind meter stagnations were not understood) and the data analysis was aimed at providing more insight into the issue. The PWN case took the middle ground, with data readily available, but not always sufficient to get strong results.

5.2 Success and fail factors

In all three cases, collaboration between the different organisations at the location of the problem owner were mentioned as the primary success factor for this kind of project. There are a number of practical issues with data project, such as finding the right data and getting access to internal servers and networks which are much easier to solve on site than through e-mail. Also, in the research process, it is much easier to share results with the problem owner and to discuss next steps when this can be done face to face than through e-mail or Skype. If additional information is needed from

experts, this can be done in a couple of minutes if you can drop by someone's desk, instead of multiple days if you send an e-mail.

Having a data science team working on location also creates awareness among colleagues that a data science project is going on. This awareness is very important in the current stage of development of data science competencies in water companies. The fact that water professionals know that their data is actively being used for analyses as well as their interaction with the data scientists can be important motivations for professionalising data management.

Another success factor, which is related to working on site, is the interaction between domain experts and data scientists. It is hard to imagine how these cases could have been successful if only data was shared with the data scientists who would then do their data mining without feedback and interpretation from domain experts. In all cases, domain expertise was present in the project teams (through the KWR researchers) and through interaction with the problem owner and experts from the water companies. Also the fact that KWR researchers can involve their own colleagues and their broad expertise was considered an important asset. The research process was going back and forth between results from analysis, reflection by data scientists and domain experts and consecutive analyses. While domain experts can provide meaning to the data (and the results of the analyses), independent data scientists can challenge 'dominant thinking' or unfounded assumptions with their exclusive focus on the data. This can generate important new insights into existing problems and practices.

These success factors were present in all the three cases, so we cannot definitely conclude that these are necessary factors. We can, however, draw that conclusion from the third success factor: data availability and data quality. As has been described in the previous chapter, this has been crucial for the success of the cases. Data quality strongly determines the quality and validity of the results, as a popular saying goes: "garbage in, garbage out". But even if there is sufficient data of sufficiently good quality, a poor structure of the data sets, or an inconsistent labelling of data or a lack of metadata, can mean that a lot of time has to be spent on data preparation. Typically, this is about 80% of the time[2], but in this project it ranged from around 30% (Oasen) to 95% (WBG). This means the time left for all other steps ranges from 70% to 5%. This is, of course, a crucial difference, particularly when time is limited.

All three project teams followed an iterative approach, sharing and evaluating results regularly and determining next steps based on those results. This can be iterations of data validation and preparation, or doing different statistical analyses exploring the data, or running consecutive versions of a machine learning algorithm, adjusting it according to the results. Going through these iterations clearly improves the results, but it can be difficult to decide how many iterations are enough and when to proceed to the next phase. This became apparent in both the WBG case and the PWN case. In the WBG case, too much time was spent on data preparation leaving little time for developing the machine learning algorithm. In the PWN case, too much time was spent on the bag filter analyses, leaving less time for the second case. When taking an iterative approach, it is crucial to keep an eye on the overall planning and not to end up in endless iterations that only lead to marginal improvements in the results.

[2] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#22a55f186f63

One of the objectives of this study was to establish a chain from data analysis to decision support. A crucial part of that chain is translating the results from the analysis (or the data itself) into information relevant for decision-making. This was particularly successful in the Oasen case, where the information was presented in an interactive dashboard with graphs that are easy to interpret. In the PWN case, the results were also nicely visualized, but the graphs were difficult to interpret for an audience not trained in statistics or hydrological modelling.

The role of the ICT-department is also important for the success of data science projects. It is often required that external data scientists have access to internal servers and data sets or that links to external data sets (sometimes permanent) have to be established. This is in conflict with the focus on security of internal systems and keeping inside and outside firmly separated with firewalls and comparable measures. Also, data science needs virtual space to experiment, link, combine and alter datasets and run models. Often, the required platforms are not available, so outside platforms have to be used (such as Dataiku in the WBG case) or just offline laptops (which have limited capacity for data storage and processing capacity). In the case of Oasen a separate server was made available, but access to that server for outside data scientist took a long time to provide. The same goes for the data infrastructure. PWN has a data lake in place, at Oasen and WBG the data warehouse is in development. ICT has an important role to play in this development. It was remarked that this requires active participation of the ICT department up to the highest level, instead of a more reactive approach in trying to facilitate incidental demands by individual data project teams.

6 Conclusions and recommendations

6.1 Conclusions

- The three cases followed a similar approach establishing a chain from data gathering upon reporting analyses for decision support.
- The cases differ in the extent to which the outcomes are readily applicable in strategy or operation.
- In the Oasen case, a concrete tool was developed (EDWARD) which can be applied and re-applied to estimate the future peak factor, based on climate and behavioural data.
- In the PWN case, the results underlined the need for systematic data gathering and management and consistent labelling of measurements with meta data.
- In the WBG case the results have shed light on the importance of systematic gathering of water meter data and consistent labelling with meta data. It has also resulted in a closer inspection of the data gathering process from water meters by WBG.
- The difference in outcomes can be largely attributed to the availability and quality of the data, as well as the level of understanding of the issue at hand.
- In all three cases, collaboration between the different organisations at the location of the problem owner were mentioned as the primary success factor for this kind of project.
- Having a data science team working on location also creates awareness among colleagues that a data science project is going on. This awareness is very important in the current stage of development of data science competencies in water companies.
- Interaction between domain experts and data scientists is important. It is hard to imagine how these cases could have been successful if only data was shared with the data scientists who would then do their data mining without feedback and interpretation from domain experts.
- Data quality strongly determines the quality and validity of the results. But even if there is sufficient data and the data is of good quality (let say the data contains few mistakes), then a poor structure of the data sets, or inconsistent labelling of data among different data sets, or too little meta data, can mean that a large share of the time has to be spent on data preparation.
- An iterative approach clearly improves the results, but it can be difficult to decide how many iterations are enough and when to proceed to the next phase.
- A crucial part of that chain from data gathering to decision support is translating the results from the analysis (or the data itself) into decision relevant information.
- The role of the ICT-department is important for the success of data science projects. The ICT department can contribute in providing access to internal servers and data sets and a virtual space to experiment, link combine and alter datasets and run models.

6.2 Recommendations

Based on the analysis of the three cases and the success and fail factors, the following recommendations were formulated:

- It is recommended that water utilities improve their data management so that data quality is consistent; sufficient meta data is available (and consistently labelled) and it is known throughout the organisation which data is available and where. Currently, data is often being managed for specific applications, by the owner of the application. Ideally, data should not only be suitable for the immediate purpose it was collected for, but for as wide a range of purposes as possible.

- Water utilities should (continue to) develop a companywide vision on data. All parts of the organisation that are somehow involved with data (either in using it, collecting it, or facilitating both) should also be involved in developing this vision. This vision can address issues such as data ownership, data management, data infrastructure, privacy and security, the role of data science in decision support and the skills and competencies that are required in the organisation or should be sourced elsewhere.

- When pursuing data science projects, start with creating and evaluating a list of available and usable data and build-in several go/no-go points to refine the selection, based on data availability, data quality and initial results. This implies taking an iterative approach in data science projects, in which after each step results are shared and new steps are formulated. This should be combined with clear decision making on whether pursuing the case is still worthwhile. While taking an iterative approach, a broader planning should be made assigning time and resources to the different steps in the project to avoid time constraints later in the project.

- In data science projects, it is recommended to work on location, in close cooperation with problem owner, data owner and end user. This makes it easier to solve challenges, obtain information, share results and create awareness of the fact that data science is being pursued. Make sure that outside researchers have sufficient access to internal systems and datasets and that a platform is available on which data can be prepared and analysis can be conducted.

- Set up multi-disciplinary project teams with data scientists and domain experts, closely connected to the problem owner, data owners and ICT support. All skills and knowledge should be present in the project team, but not necessarily in the same person. Having an independent data scientist (not from the client organisation) on board can provide a fresh look at existing issues. KWR researchers should link with their colleagues for additional domain expertise, which can be of added value to data science projects.

- Make a distinction between research projects and implementation projects. A research project necessarily presupposes a level of uncertainty about the outcomes. Will the data be helpful in establishing causal relations or predicting certain events? Implementation is focused on developing practical applications out of research results that already have been established. Implementation is a broad term, which can range from minor policy changes based in scientific results to setting up automated decision support tools. In the latter case, the

broader scope and scale could require involvement of outside ICT experts/companies.

- When implementation is the primary goal, start with the end state in mind. Determine beforehand where the tool should be located and which connections to internal and external data sets are necessary. If this is done *ad hoc*, by downloading data to a temporary server and making impromptu adjustments, this can be difficult to replicate and implement in existing systems.

7 Literature

Van Thiel, L. (2016) Watergebruik Thuis 2016. Vewin rapport C8732 (http://www.vewin.nl/SiteCollectionDocuments/Publicaties/Cijfers/Watergebruik-Thuis-2016.pdf, visited Jan 7, 2019).

Van Thienen, P., H.-J. van Alphen, A. Brunner, Y. Fujita, B. Hillebrand, R. Sjerps, J. van Summeren, A. Verschoor, and B. Wullings (2018) Explorations in Data Mining for the Water Sector. BTO 2018.085

Van der Marel, P. and Van der Woerdt, D. (2014)/ Eén eenvoudige parameter (EGV) geef t inzicht in drinkwaterdistributie. H2O-Online, 3 september 2014.

Vonk, E., D.G. Cirkel, I. Leunk (2017) De gevolgen van klimaatverandering en vakantiespreiding voor de drinkwatervraag. BTO 2017.043

Attachment I Pilot Waterbedrijf Groningen

Pilot Waterbedrijf Groningen

I.1 Approach 1

The goal of approach 1 is to automate the process of distinguishing between suspicious and non-suspicious meters, using only water consumption data. The steps that were necessary to reach this goal are described below and are represented in the infographic in Figure 24:

1. **Prepare a reference data set of consumption history per water meter**
   The first step was to aggregate meter readings on the level of individual water meters. We calculated water consumptions from successive meter readings. Since the period between two successive meter readings does not necessarily equal one year, we recalculated the consumption based on 365 days. Periods smaller than ¾ year (272 days) are excluded in the process to prevent statistical bias (i.e. random variations in consumption cancel out for longer periods). Negative and extremely high consumptions >1000 m^3/year are also excluded, because these likely include misreadings (or small businesses).
   We also removed anomalous meter readings. We only kept 90% of the meter readings that fell inside the ranges set up by Waterbedrijf Groningen.
   This gives us a prepared data set of meter-specific consumption histories.

2. **Calculate the average yearly consumption for each water meter**
   For each series of consumption history per water meter the average yearly consumption is calculated. For each meter reading the difference between the consumption and the meter-specific average is computed ('delta_Q').

3. **Assign household size categories**
   We assigned household sizes to individual water meters, based on the average of the consumption history of those meters. The mean metered consumption was compared to VEWIN data for 2016 of household size-dependent average consumption (Van Thiel, 2016). The class bounds are calculated as linear averages of the consumption of the two adjacent bins. For the purpose of this research we ignored small long-term trends in the consumption, as well as a possible region-specific dependency. Large industrial consumption is not explicitly accounted for. Consumption larger than 1000 m3/year is accounted as ≥11 persons per household. In the Results section, we focused only on household sizes from 1 to 6 persons, which includes the bulk of the data.

4. **Identify thresholds per household class for outlier identification.**
   Per household class a distribution of consumption deviations ('delta_Q's') is translated to a boxplot, so that outlier thresholds can be determined. This is done by using the interquartile range (IQR) multiplied by 1.5. The IQR is a measure used for statistical dispersion. The multiplication factor of 1.5 is

arbitrary but often used in outlier detection. Identifying these thresholds needs to be done only once, assuming that the consumption per household size doesn't change much over the years.

5. **Identify consumption outliers**
   The outlier thresholds are applied to the entire consumption data set (adding back the anomalous data we removed in step 1) to distinguish suspicious and non-suspicious consumption records. For each evaluation year, this results in a set of suspicious water meters and non-suspicious water meters for each household size.

6. **Compare outliers to water meters labelled by Waterbedrijf Groningen as 'suspected of stagnation'**
   The set of suspicious water meters for the evaluation year 2017 is compared to two sets of water meters labelled by Waterbedrijf Groningen:

   - the 10% anomalous water meters (S1)
   - the 3% water meters labelled as 'suspected of stagnation' (S2)

   Ideally, the set of suspicious water meters

   (i) fully includes the set S2, because this would exclude false negatives, i.e. water meters identified as normal that Waterbedrijf Groningen identified as suspicious, and
   (ii) is smaller than the set S1, because this would reduce the current number of false positives and suggests that application of the algorithm can reduce the manual effort by Waterbedrijf Groningen.

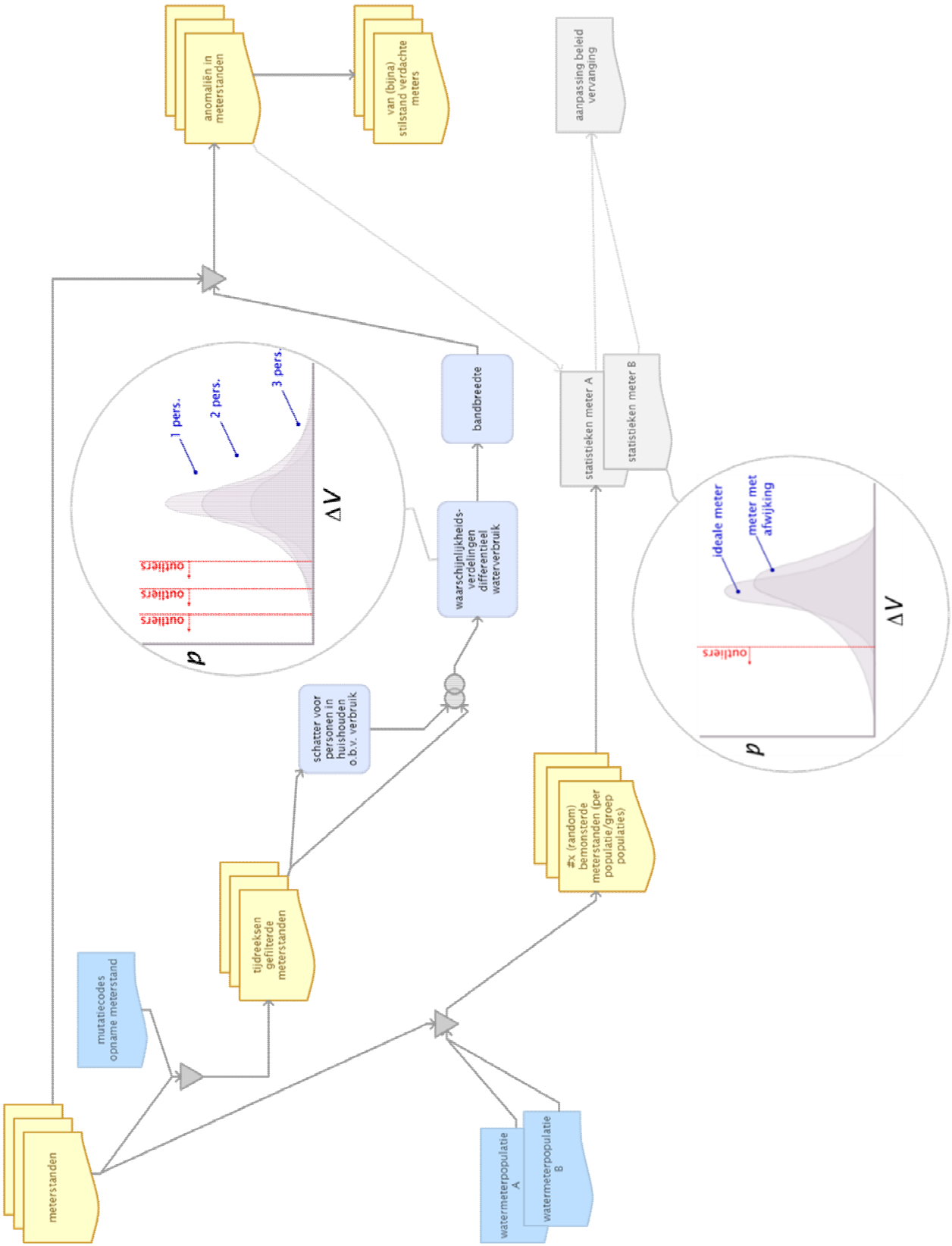
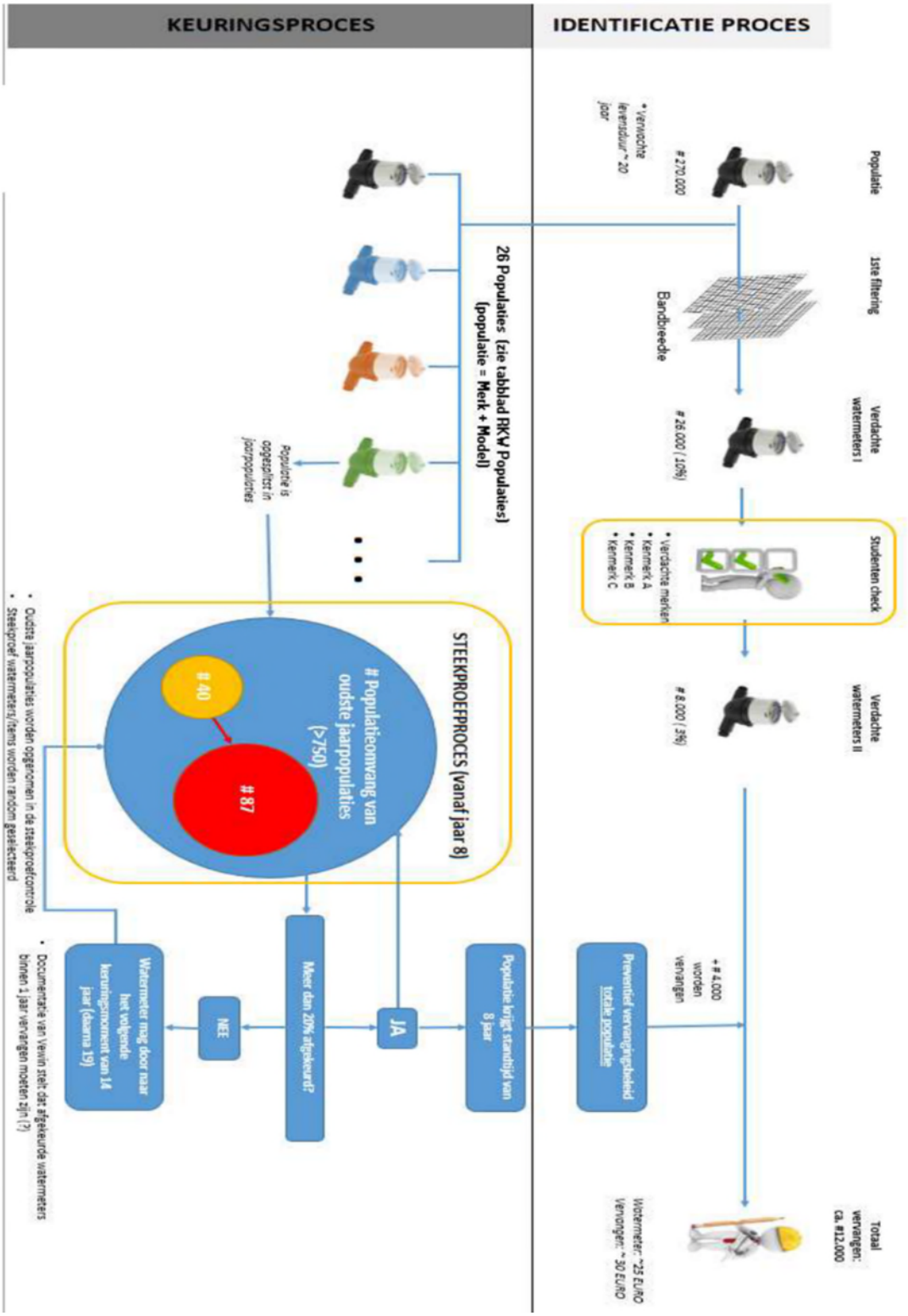


Figure 24: Infographic showing approach 1



I.2 Infographic containing information about the identification process and the quality assurance process