BTO 2018.085 | October 2018

BTO report

Explorations in Data Mining for the Water Sector

KWR Watercycle Research Institute

BTO

Explorations in Data Mining for the Water Sector

BTO 2018.085 | October 2018

Project number
402045/023

Project manager
G. (Geertje) Pronk PhD

Client
BTO - Exploratory Research

Quality Assurance
C. (Christos) Makropoulos PhD

Author(s)
P. (Peter) van Thienen PhD, H.-J. (Henk-Jan) van Alphen MSc, A. (Andrea) Brunner PhD, Y. (Yuki) Fujita PhD, B. (Bram) Hillebrand MSc, R. (Rosa) Sjerps MSc, J. (Joost) van Summeren PhD, A. (Anthony) Verschoor PhD, B. (Bart) Wullings MSc.

Sent to
The report was distributed among BTO participants and will become public one year after its initial publication.

Year of publishing
2018

More information
Dr. P. (Peter) van Thienen
T 030 6069602
E peter.van.thienen@kwrwater.nl

Keywords: data mining

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

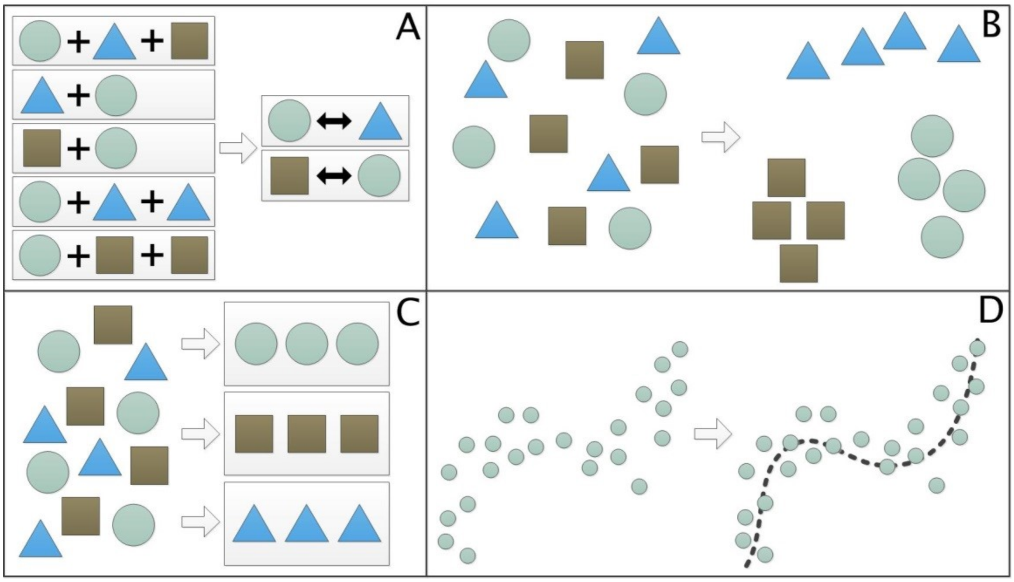
KWR Watercycle Research Institute

BTO *Executive summary*

Data mining techniques ready for water utility applications

Author(s) Peter van Thienen PhD, Henk-Jan van Alphen MSc, Andrea Brunner PhD, Yuki Fujita PhD, Bram Hillebrand MSc, Rosa Sjerps MSc, Joost van Summeren PhD, Anthony Verschoor PhD, Bart Wullings MSc.

Techniques for extracting knowledge from (combinations of) databases, often presented under the flags of data mining and big data, have shown significant development over recent years. However, attempts at their implementation in the water sector have produced interesting but not yet 'revolutionary' results. Identifying and resolving barriers towards deploying already mature data mining methods and tools for the benefit of the water sector, related to data ownership and access; availability and quality of data; organizational dynamics and culture, will allow the sector to take advantage of past work and recent developments in the field – capitalizing on a vast array of mathematical methods to address both current and emerging challenges.



Four main objectives of data mining: association rules, clustering, classification, and regression. From Vonk and Vries (2016)

Relevance: data mining techniques provide opportunities for insight

Techniques for extracting knowledge from (combinations of) databases, often presented under the flags of data mining and big data, have shown significant development over recent years. Many 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 (developments in sensors, 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. The aim of this research is 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.

Approach: current state of affairs from the literature and practitioners

Existing approaches and applications have been scouted in the literature and practitioners have been interviewed. From the information we gathered, we have identified opportunities for the application of data mining techniques in the water sector, both from a domain perspective and from a data perspective.

Results: methods and data are there, but some barriers need to be addressed

In the methodological sense, data mining or more precisely knowledge discovery from databases is a mature field which offers many fully developed methods with a plethora of reference applications. In the specific water cycle management domain, numerous applications in both an academic and operational context are available internationally. From this perspective, there is no immediate need for KWR and the BTO utilities to put more effort in the development of (completely) new methods, but rather in the implementation, or customization of existing methods. Both the datasets and the applications are readily identifiable, presenting opportunities. A number of successful applications have been reported also by Dutch utilities, such as the prediction of pipe failures by Oasen. However, practitioners indicate a number of obstacles, including data ownership and access, availability of good data analysts, availability and quality of data, and organizational dynamics/culture.

Implementation: collaborative effort to set up the data to decision chain

This report has been written as a deliverable of the first phase of the BTO exploratory research project *VO datamining*. Based on the conclusions of the first phase, as described above, we recommend that the following phases of the project focus on the actual implementation of a number of data mining cases with BTO utilities. In doing so, we no longer aim for methodological exploration and innovation, but rather for innovation in the application. Important research questions to be answered include practical issues related to the streamlining of the complete chain from data acquisition through quality assurance and data mining to decision(s) (support). At a higher abstraction level, they also include questions on how to organize a successful implementation of data mining techniques – ideally a template approach can be defined. We have seen that the methods, the data and the potentially fruitful applications are there. For the water utilities, our recommendations focus on resolving the barriers which have been identified and which are within their sphere of influence. These include data ownership and access, availability and quality of data, and organizational dynamics/culture. In this study, the root causes of these barriers have not been considered, but this would be a first step in resolving them. Organizations outside the water sector have taken steps and set up frameworks that address all of these issues. A good example is Rijkswaterstaat, which presented its framework in one of the meeting of the Hydroinformatics Platform. We recommend that their approach be considered as a starting point.

Report

This research is described in the report *Explorations in Data Mining for the Water Sector* (BTO 2018.085).

More information
Dr. P. (Peter) van Thienen
T 030 6069602
E peter.van.thienen@kwrwater.nl

KWR
PO Box 1072
3430 BB Nieuwegein
The Netherlands

Watercycle Research Institute

Contents

1 Introduction

1.1 Context

Techniques for extracting knowledge from (combinations of) databases, often presented under the flags of data mining and big data, have shown significant development over recent years. Many 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 (developments in sensors, 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.

This exploratory study makes a wide inventory 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, the project identifies which current and emerging applications (combination of techniques and one or more data sources) are promising, on technical grounds, but also on the basis of the expected information yield to support decision processes at water companies. This report covers the results of both of these activities.

1.2 Aim and approach

The aim of the exploratory research *VO datamining*, for which this report is the first deliverable, is 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.

A priori, the latter point was envisioned to involve the development and deployment of an innovative data mining technique. However, progressing insights gained during the execution of the project, both from the first phases of the project itself and from the knowledge exchange meetings of the Hydroinformatics Platform, have led to a shift in the approach. Innovation is sought in the application rather than the technique, and attention is given to the entire data chain from gathering to decision making in an explicit collaboration between KWR and water utilities.

This report describes the results of the first two activities and provides and overview of available data mining techniques and applications within and outside the water sector. It also identifies promising applications for the water sector.

Because of the enormous number of fields in which data mining techniques can be and are applied, the overview provided here cannot be complete. It is not meant to be. It is meant to provide an impression of the current state of affairs, as a starting point for concrete new developments in a water sector context.

1.3 Scope

For this report, we have elected to use a broad definition of data mining. In fact, when we say *data mining*, we mean *knowledge discovery from databases* (Vonk and Vries 2016). This includes all generic statistical and heuristic techniques which are included in a narrower, more formal definition, but we also include less generic, more ad hoc or case specific approaches which serve the same purpose of discovering knowledge in datasets (such as QSAR and QMRA). A more elaborate overview of the considered methods is given in Chapter 2.

The structure and classification of problem types applied in this report reflects the organization of research fields within KWR. An overview is given in Table 1, which also describes the scope in terms of topics.

1.4 Guide to this report

We start by giving an overview of data mining methods in Chapter 2. Chapter 3 gives an overview of data mining applications, both within the water sector and outside. Also, interviews on the topic with practitioners, again within and outside the water sector, are presented. The potential for new applications in the water sector is discussed in Chapter 5. Finally, we present our conclusions and recommendations from the material in the previous chapters in Chapter 6.

TABLE 1: CLASSIFICATION OF PROBLEM TYPES. SUBFIELDS CORRESPOND TO KWR TEAMS, WHICH ARE ORGANIZATIONAL CORES OF EXPERTISE.

Field	subfield	topics for application of data mining techniques
Hydrology	Ecohydrology	- Advanced and sustainable water management for nature conservation; - freshwater provision to agriculture, industry, drinking water; - blue-green solutions (i.e. smart combinations of water technology and ecology) to address the consequences of climate change in urban environments.
	Geohydrology	- Quality and supply of groundwater; - subsurface techniques for the provision of water and heat; - well technology and management.
(Waste) water treatment and distribution Water	Water treatment	- Assessment of water treatment performance; - real-time prediction of coagulant type and dosage.
	Water distribution	- Water demand forecasting; - leak localization; - water quality event detection; - drinking water discoloration; - pipe integrity risk assessment; - customer behavior analysis; - decision support to manage water quality incidents.
	Wastewater treatment	- Prediction of influent flow rate; - prediction of solids in effluent; - energy optimization of pumps; - optimization of treatment plants.
Water Quality	Chemical water quality[1]	- Design of risk based monitoring programs for drinking water (sources); - wastewater based epidemiology; - interpretation of high-resolution mass spectrometry (HRMS) data based non-target screening (NTS) analyses; - interpretation of micro- and nanoplastics analysis coupled to Fourier-transform infrared (FTIR) spectra, thermogravimetric (TGA) spectra; - prediction of relevant transformation products; - evaluation and prediction of a chemical's human health effects, using toxicological *in vivo* and *in vitro* data, *in silico* methods, adverse outcome pathways (AOPs) and bioassays; - evaluation, prediction and modelling of a chemical's exposure in drinking water sources; - evaluation and prediction of a chemicals removal in treatment systems.
	Biological water quality[2]	- biological water quality analysis

[1] The methods by which these topics are addressed are often grouped under the term *cheminformatics*.

[2] In a similar vein, methods by which these topics are addressed are often grouped under the term *bioinformatics*.

2 Big data and data mining

2.1 Introduction

Data mining, in the sense of knowledge discovery from databases, is often linked (and applied) to big data. Big data is characterized by the 5 V's (Zhai, Ong et al. 2014):

- *Volume*: large amounts of data;
- *Velocity*: high rate of generation;
- *Variety*: combinations of different data types and sources;
- *Veracity*: varying levels of quality;
- *Value*: to be extracted from the data.

It is clear that this definition applies, at least to some degree, to datasets gathered by water utilities. Take for example the ensemble of flow and pressure data gathered at all water production locations in a supply area. The amount of data generated and the rate of data generation are large by human standards, but not really when compared to many other datasets (e.g. radio astronomy). Flow and pressure data are different in nature, and different types of measuring devices or sensors (brands, models, measuring principles, dimensioning) may be combined, which will also result in different levels of data quality. These data contain information on leakage and the state of the network, which can be converted to financial value.

Data mining is sometimes considered a subfield of machine learning:

Definition of machine learning
An application of artificial intelligence aiming at (i) learning from historical datasets and (ii) generalizing this knowledge to analyse new datasets. Machine learning procedures typically consists of 3 steps using independent subsets of data: training (the step in which the model parameters are fitted), validation (the step to evaluate the fit of the trained model to a validation dataset while tuning the model's hyperparameters), and testing (the step that evaluates the final model's performance against an independent test dataset).

However, in the broad definition of data mining which is applied here (see §1.3), there are also data mining techniques that do not qualify as *machine learning*. We start our treatment of the subject in this chapter using the machine learning perspective, and then broaden the scope to include other approaches.

2.2 Basics of data mining from a machine learning perspective

With data mining, different goals can be aimed for. Common goals are association rules creation (i.e. identifying cause and effect), clustering, classification, and regression (Figure 1). To achieve those goals, there are thousands of machine learning algorithms available. Each machine learning problem consists of three basic components (Domingos 2012): representation model (a model which represents the data in a certain way), evaluation criteria (a measure to evaluate the representation model), and an optimization algorithm (a method to improve the score of the representation model against the evaluation criteria). Depending on the data mining goals and type of dataset, specific combinations of algorithms of the three elements should be chosen.

Note that the choice of the representation model, which is often overrated in the whole process of data mining, is not the decisive element for successful data mining. Many other elements (e.g. data preparation such as feature engineering, algorithms of evaluation criteria and optimization) are equally or sometime more important (Zhang et al., 2003, Domingos 2012). In this light, the goal of the data mining task and identification of available datasets should be the starting point, not the available techniques of data mining.

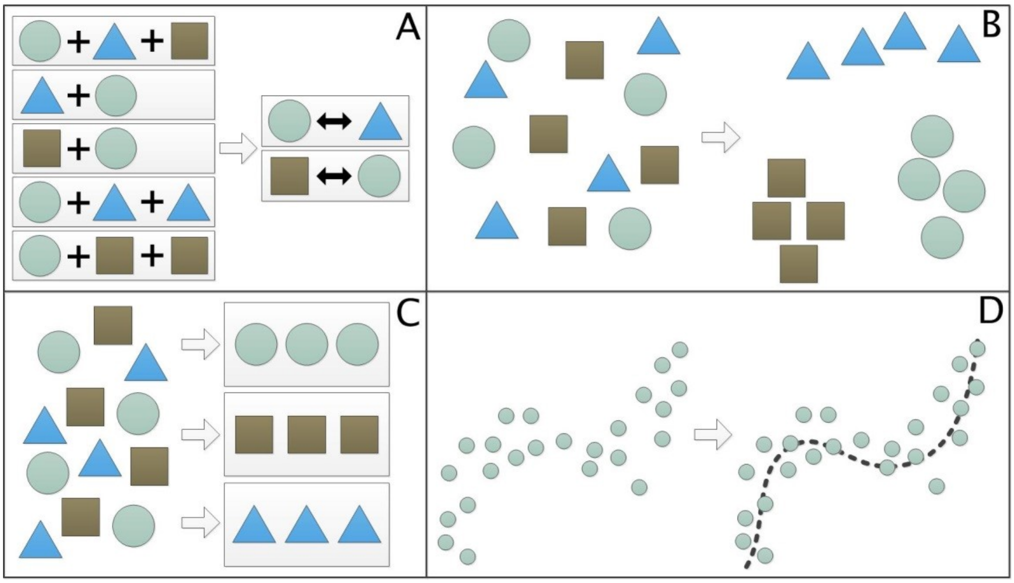


Figure 1. Graphical representation of main goals of data mining. (A) association rules, (B) clustering, (C) classification, (D) regression. Source: Vonk and Vries (2016).

2.3 Overview of formal generic approaches

In Table 1, commonly used algorithms of representation models are listed. For more details about different processes and types of machine learning, see Vonk and Vries (2016). Furthermore, we also listed some of the data mining techniques which are mainly used in the phase of data preparation (e.g. dimensionality reduction) rather than to achieve the ultimate goal of the data mining.

TABLE 1. LIST OF COMMONLY-USED ALGORITHMS OF REPRESENTATION MODELS, GROUPED PER GOAL OF DATA MINING (OR DATA PREPARATION). THE REPRESENTATION MODEL IS CATEGORIZED INTO ONE OF THE CLASSES LOGICAL/GEOMETRIC/PROBABILISTIC/ENSEMBLE (SENSU FLACH 2012), WHICH ARE DESCRIBED IN TABLE 2. THE *INTERPRETABILITY* COLUMN GIVES AN EXPERT JUDGMENT BY THE AUTHORS OF HOW EASILY/STRAIGHTFORWARDLY THE OBTAINED MODEL CAN BE INTERPRETED.

Class	***Goal of data mining (or data preparation)***	***Type of data (for each observation i)***	***Algorithms of representation model***	***Type of model***	***Machine learning ?***	***Interpre-tability***
Unsupervised	Association rules	A set of categorical variables	A priori	Logical	Y	moderate
	Clustering	A set of numerical variables	K-means	Geometric	Y	moderate
	(Dimensionality reduction)	A set of numerical variables	Principal Component Analysis	Geometric	N	good
		Two sets of numerical variables	Canonical correspondence analysis	Geometric	N	moderate
	(Anomaly detection)	A set of numerical variables	Local outlier factor	Geometric	Y	moderate
	(Density Estimation)	A numerical variable	kernel density estimation	Probabilistic	Y	moderate
Supervised	Classification	A categorical variable + A set of numerical variables	Decision trees	Logical	Y	good
			ANN (Artificial neural networks)	Geometric	Y	poor
			Support vector machine	Geometric	Y	moderate
			Naïve Bayes	Probabilistic	Y	poor
			Random forests	Ensemble	Y	moderate
	Regression/classification	A categorical/numerical variable + A set of numerical variables	Feedforward ANN	Logical/ geometric	Y	poor
	Regression	A set of categorical variables	Linear regression	Geometric	N	good
			generalized linear (mixed) model	Geometric	N	good
			Regression trees	Logical	Y	good
			Support vector regression	Geometric	Y	moderate
			Gradient boosting regression	Ensemble	Y	moderate

TABLE 2: DIFFERENT TYPES OF REPRESENTATION MODELS OF MACHINE LEARNING (SENSU FLACH 2012).

Type of representation model	Description
Logical	Logical models describe the features with logical formulas. They are most commonly represented in the form of rules or decision trees.
Geometric	Geometric models are based on concepts from geometry (such as distance, lines, or planes) and interpret data as points in a multidimensional space. The advantage of geometric models is that the results can be (relatively) easily visualized. The model fitting is often strongly influenced by data distribution, and therefore data standardization and outlier removal is often recommended.
Probabilistic	Probabilistic models assume a certain probability that attributes in data are related to each other. Probabilistic models are commonly based on the Bayes theorem.

2.4 Deep learning

Some cases require a more complex approach, such as deep learning. This approach is defined as follows (Deng and Yu, 2014):

Definition of deep learning
A class of machine learning techniques, where many layers of information processing stages in hierarchical supervised architectures are exploited for unsupervised feature learning and for pattern analysis/classification. The essence of deep learning is to compute hierarchical features or representations of the observational data, where the higher-level features or factors are defined from lower-level ones. The family of deep learning methods have been growing increasingly richer, encompassing those of neural networks, hierarchical probabilistic models, and a variety of unsupervised and supervised feature learning algorithms.

This means that data are transformed in a number of steps into successively more abstract representations for classification or pattern recognition. A good example of a concrete algorithm is an artificial neural network with multiple hidden layers.

2.5 Ad hoc approaches in cheminformatics and bioinformatics

An overview of more domain specific approaches is given for the subfield of chemical and microbiological water quality in Table 3. Note that many of these methods (contain steps that) are based on methods listed in Table 1. These methods are commonly grouped under the terms chem(o)informatics, which was defined by Brown (1998), and bioinformatics, as defined by Luscombe et al. (2001). We mix, modify and broaden their definitions here to better cover the applications of both fields in the water sector:

Definition of cheminformatics
Chem(o)informatics is the application of "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to the field of chemistry, for inter alia the identification of chemical compounds, understanding of their interactions, and organization of information regarding these, in order to transform (large volume, multi-source) data into information and information into knowledge.

Definition of bioinformatics
Bioinformatics is the application of "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to the field of biology, for inter alia the identification of organisms, their genetic information, and/or organic compounds, understanding of their interactions, and organization of information regarding these, in order to transform (large volume, multi-source) data into information and information into knowledge.

TABLE 3: SPECIFIC APPROACHES FOR CHEMICAL AND MICROBIOLOGICAL APPLICATIONS

label	*technique/ description*	*method class*	*type of data*	*state of development*	*reference*
QSAR=quantitative structure-activity relationship for effect prediction	Statistical link between chemical structure and chemical activity	Correlation and classification	Chemical structure, HRMS fragmentation spectra, toxicity data	Different models developed	(Altenburger, Nendza et al. 2003, Bhatia, Schultz et al. 2015, Dimitrov, Diderich et al. 2016)
QSAR=quantitative structure-activity relationship for removal efficiency prediction	Statistical link between chemical structure and removal efficiency with different treatment technologies	Correlation and classification	Chemical structure, removal rates	Different models developed	(Wols and Vries 2012, Vries, Wols et al. 2013, Vries, Bertelkamp et al. 2017)
QSAR=quantitative structure-activity relationship for environmental fate prediction	Statistical link between chemical structure and environmental fate	Correlation and classification	Chemical structure and chemical fate properties	Different models developed	(Mackay and Paterson 1991, Mackay, Shiu et al. 1992, Sabljić, Güsten et al. 1995, Scheringer 2009, Scheringer, Jones et al. 2009, Zarfl, Scheringer et al. 2011)
Read across	Relevant information from analogous substances to predict	Correlation and classification	Chemical structure and chemical properties	Standard method for substances registration (REACH)	(Escher and Fenner 2011, Shah, Liu et al. 2016)

label	***technique/ description***	***method class***	***type of data***	***state of development***	***reference***
	chemical properties				
Chemical graph theory	Molecular representation by a graph: atoms as vertices and molecular bonds as edges.	Graph theory	Chemical structure	Developed since 1980. Used in cheminformatics	(Bonchev and Rouvray 1992, Ivanciuc 2013)
QMRA: quantitative microbial risk assessment of exposure	Statistical link between reference pathogens, exposure pathways and hazardous events	Regression & classification	Quantitative data or assumptions on pathogen occurrence and exposure, data on frequency of hazardous events and severity of hazards, data on pathogen removal rates by relevant treatment processes	Different models developed	WHO, 2016
QMRA: quantitative microbial risk assessment of health effects	Statistical link between reference pathogens and health effects	Regression & classification	Quantitative data on dose of reference pathogens and infection response, risk of illness per infection, disease burden and susceptible population	Different models developed	WHO, 2016
QMRA: quantitative microbial risk assessment: risk characterization	Statistical link between exposure and health effects within relevant scenarios	Regression & classification	Exposure & health effects, scenarios accounting for variability and uncertainty	Different models developed	WHO, 2016
Large-scale genomic data mining	Combination of different algorithms	Association, classification & regression	Sequence data (genomic), microarray	Different models developed	Huttenhower & Hofmann (2010); Ju & Zhang

label	***technique/ description***	***method class***	***type of data***	***state of development***	***reference***
			data (DNA expression), interaction variable data (chemical/ physical, regulatory, protein modifications, etc.)		(2015a,b); Kennedy et al. (2010); Lee et al. (2008)
Coliform monitoring	Machine learning	Gradient tree boosting, decision trees, distance weight	Water quality	Several models developen	Dawsey & Minsker (2007)
Cyanotoxin monitoring	Multivariate adaptive splines	Classification, regression	Water quality	Model developed	Garcia Nieto et al. (2010)
Public response analysis	Mining social media	Association rules	Social media (keywords)	Developed (social science)	Cha & Stow (2015)
Modelling eutrophication	Deep learning	PCA, self-organizing feature map (SOFM), fuzzy logic modelling	Water quality	Several models developed	Chen & Mynett (2003)
Assessing microcystin concentrations	Deep learning	Association, path analysis, clustering, classification, forecasting (genetic algorithms)	Water quality (incl. microcystin), remote sensing (satellite)	Developed (geoscience)	Chang et al. (2014)

3 Current applications of data mining: the state of play

3.1 Introduction

This chapter provides an overview of current applications of data mining. The first part of this chapter is based on a literature study and focusses on the water sector and water applications. The second part describes a number of interviews of practitioners from inside and outside the water sector.

3.2 Overview from the literature

Table 4 provides an overview of data mining applications in the water sector that have been realized and presented in the scientific literature. The table clearly illustrates that a vast number of applications have already been realized, using many of the methods described in the previous chapter on a wide range of datasets. Because of the wide range of applications described in the table, it is difficult to extract a common denominator or conclusion, other than the following:

- the applications of data mining themselves also conform to the 5 V's of data mining (see §2.1);
- a focus on application rather than method development seems appropriate for the Dutch water sector.

TABLE 4: OVERVIEW OF DATA MINING APPLICATIONS IN THE WATER SECTOR FROM THE LITERATURE.

Field	***label***	***application + purpose***	***Temporal/ spatial scale***	***type of data***	***Goal of data mining[1]***	***quality/usability of results***	***reference***
hydrology	Classification of drone/satellite images into vegetation type	Converting satellite image into a vegetation map based on ground survey data, both with pixel-based and Patched-based approach	8 x 10 km image from 5 different timings of a year	Raster data of satellite images (Sentinel 2A) Ground survey data of vegetation class	Classification (Random forest, with texture statistics to account for spatial information)	Accuracy of prediction ca. 74 % The method can be applied for drone images, and for prediction of other categories (e.g. soil pH)	On-going VO research (400695/060, 2017-2018)
	Predicting vegetation types from environmental variables (PROBE)	Predict vegetation association types from habitat variables via indicator values of species	Training dataset from whole Netherlands	35 000 vegetation relevé data[2] (species identification and abundance) with known vegetation type Indicator values for environmental gradients (soil acidity, soil moisture, and nutrient availability) for all vascular and moss species in the Netherlands Habitat variables (e.g. soil type, (modeled) soil moisture)	Classification (Bayesian classification) Density estimation (Gaussian Mixture Model) Regression (Structural equation modeling, linear mixed model, variation partitioning)	Accuracy of prediction 85 % Software PROBE-2 is readily available, enabling area-covering prediction of vegetation types The PROBE model is applied in a number of projects concerning nature restoration and climate adaptation	BTO-2010.024 (Witte, Bartholomeus et al. 2010), BTO-2016.011 (Cirkel, Fujita et al. 2016), BTO-2016.071 (Fujita, Bartholomeus et al. 2016) (Witte, Wójcik et al. 2007, Ordoñez, Van Bodegom et al. 2010, Ordonez, van Bodegom et al. 2010, Fujita, van Bodegom et al. 2013,

[1] algorithms or representation models
[2] A relevé is a quadrat that encloses the minimal area that can be expected to contain e.g. 95% of all species present in a community.

						Witte, Bartholomeus et al. 2015)
Translating Vegetation relevé data to ecological information (ESTER)	Translation of vegetation relevé data into ecological information (soil acidity, soil moisture, nutrient availability, salinity)	Input point data from whole Netherlands	Vegetation relevé data (i.e. species identity and abundance) Indicator values of species (IV) Habitat variables (soil pH, nutrient mineralization, groundwater level)	Regression (linear regression)	Explained variance of IV-habitat variable relations: 45 – 79 % Software ESTER v.01 is readily available Prediction can be used for calibration/validation of (hydrological) models in a cost-efficient manner	KWR 2014.054 (Witte, Bartholomeus et al. 2014)
Exploring causes of well clogging	Find factors which influence clogging of wells and quantify their influence	For each study, ca. 15 wells x time series of height (30 seconds interval) for multiple years	Aggregated data (to a monthly interval) of groundwater quality (e.g. Cl), utilization factors, frequency of switching on/off, geometry of wells	Regression (Gradient boosting regression)	Explained variance ca. 65% Main factors causing clogging were identified, which gives knowledge base to build decision support system for well field operation	3 on-going BTO research (401826/001, 400554/191, 401827/001)
virus removal during soil passage	Quantifying importance of soil (hydro)geochemical factors on efficiency of virus removal during soil passage	Meta-dataset of 4 regions in the Netherlands	Dataset of virus stacking efficiency (SE) and a number of hydrogeochemical variables from 4 existing studies	Dimensionality reduction (PCA) Regression (linear mixed model)	Explained variance of SE: 56%	On-going BTO project (BTO 2018.014)

	Evaluation of restoration measures	Evaluation of effectiveness of restoration measures via analysis of soil and vegetation variables	Meta-dataset of 4 Dutch coastal dunes	Dataset of soil variables (e.g. pH, organic matter content) and vegetation composition from 4 Dutch coastal dunes	Dimensionality reduction (Canonical Correspondence Analysis) Regression (ANOVA, linear regression, log-response ratio)	Effectiveness of different restoration measures was quantitatively evaluated, which helps to make robust suggestions for future management strategies	DPWE grey dunes (Fujita and Aggenbach 2015), TKI ijzerslib (Dorland, Fujita et al. 2017)
	Water demand analysis	Prediction of water demand based on meteorological and holiday statistics	8 water supply areas, 20 years of time-series data	Holiday statistics, daily data of meteorological measurement, water daily volume	Regression (Support vector regression)	Explained variance ca. 60-95% Prediction of water demand under different scenarios (including climate change)	BTO 2017.043 (Vonk, Cirkel et al. 2017)
	Flood prediction	Detection of flooded area (1), prediction of flooding event (2) using machine learning techniques	Temporal and spatial data	e.g. Satellite image, topography, distance to rivers, dike morphology, weather data, river water levels	1 Classification (e.g. Support Vector Machine, Random forest) 2) Regression (e.g. Bayesian Linear, Boosted Decision Tree) 3) Classification / Anomaly detection (Neural Cloud)	1) Accuracy >80% 2) RMSE up to ca.0.1-0.2	1) Lamovec P., Matjaž M. et al. (2013) 2) Noymanee, Nikitin et al. (2017) 3) Pyayt, Mokhov et al. (2011)
water treatment and distribution	Demand Forecasting	Urban Water demand forecasting on monthly, weekly, daily and hourly scales	Time data different time scales	Time series data of water demand and weather information (AMR data) or DMA inlet data	Prediction: 1) DAN2, dynamic artificial neural network 2) support vector machine regression	1) >90% accuracy for monthly, weekly, daily and hourly demands 2) MAPE (Mean Absolute Percentage Error) < 30% for 50% of AMRs	1) (Ghiassi, Zimbra et al. 2008) 2) (Candelieri, Soldi et al. 2015) 3) (Adamowski and Karapataki 2010)

				3) linear regression and artificial neural network 4) Deep learning modeling: DBN, deep belief network	3) Best results with the Levenberg-Marquardt Artificial Neural Network 4) Prediction errors of 2% for daily demand, 5.5% for 15 minute demand	4) (Wu and Rahman 2017) (Wu and Rahman 2017)
Leak Localization	Leak localization in water distribution networks	Spatial data	Residuals, obtained by comparing pressure measurements with estimations provided by models, pressure data from sensors	Classification: 1) kNN, 2) Support vector machine Regression: 2) support vector machine	1) Depends on network size / requires epanet model of the investigated network 2) 100% of the cases was the predicted leak node within 500m of the actual leaking node and 35% exact location success rate.	1) (Soldevila, Blesa et al. 2016) 2) (Mashford, Silva et al. 2009)
Anomaly/ contamination detection	Contamination (intrusion) location and timing determination	Time and spatial data	Sensor data (every 5 minutes) total of 30 minutes	Anomaly detection (Maximum Likelihood method)	7 out of 285 nodes found as possible with the correct node as the most likely (SIE) 18 out of 285 nodes found as possible with the correct 2 among the 6th most likely (clustered)	(Huang and McBean 2009)
Asset management (Risk models)	Risk assessment models of pipe networks	Asset data and repair data	Asset data and repair/broken asset data	1) Classification, Scoring model 2) Regression, Evolutionary Polynomial Regression	2) Correctly describes 95.7% of all bursts	1) (Babovic, Drécourt et al. 2002) 2) (Giustolisi, Savic et al. 2004) (Berardi, Kapelan et al. 2008)

	Estimation drinking water discoloration	Use data mining to find a relation between regeneration and pipe factors	Field data during flushing events (U.K. and Dutch DWDNs)	Turbidity time series and flushing flow rates, pipe specs., hydraulic andiron content, treatment type, etc.	Self-Organising Maps, SOMs + Evolutionary Polynomial Regression (EPR). (EPR combines Genetic Algorithms with numerical regression)	Confirms key factors to higher material accumulation rate. Predicts daily regeneration rates (R^2=0.930 (for cases with regular and repeated flushing).	(Mounce, Husband et al. 2014, Mounce, Blokker et al. 2016)
	Assessment of Water Treatment performance	The assessment of water treatment performance (OM removal)		Water samples and fluorescence excitation-emission wavelengths	Regression (stepwise regression, partial least squares, multiple linear regression and neural network with back-propagation algorithm) Decomposition (principal components analysis, parallel factor analysis and self-organizing map)	Best results with combining parallel factor analysis with partial least squares or self organizing maps with neural network with back propagation algorithm	(Bieroza, Baker et al. 2012)
	Decision support	Decision support to manage water quality incidents in distribution systems	Incident reports of 3-year period.	Incident reports of water companies (2009-2011); expert subjective scoring	Knowledge-based problem-solving (memory based): Case Based Reasoning (CBR).	Supports decision –making by providing guidance for water utilities in managing drinking water incidents.	(Mounce, Mounce et al. 2015)
	Customer Behavior Analysis	Assess public response to environmental event (shutdown of drinking water supply)	Social media data of ~500k residents	Web-based data (Twitter, Google Trends) in response to supply		Offers perspective on public response, the associated collective knowledge, and public perception.	(Cha and Stow 2015)

	Water treatment (coagulant type)	Determining the coagulant type	Real time data	Water parameters (pH, temperature, alkalinity) which change over time	Classification (Decision tree for type of coagulant)	Overall good but prone to over fitting (50/50 train/test better results than 70/30)	(Bae, Kim et al. 2006)
	Water treatment (real time) coagulant dosage)	Prediction model for real time PAC dosage especially for extreme cases (such as storms)	Real time data	Water parameters (pH, temperature color…)	Prediction 1) Neural network 2) ANN and ANFIS	ANFIS outperforms ANN	1) (Bae, Kim et al. 2006) 2) (Wu and Lo 2008)
		Prediction of alum dosage in drinking water treatment processes	9 year period	Coagulation-related parameters, collected from water works authority in Thailand, period 2006-2015.	Comparison of WEKA classifiers (M5P, M5Rules and REPTree) to a feedforward ANN (multilayer perceptron, MLP)	Highest accuracy yielded by M5Rules method.	(Chawakitchareon, Boonnao et al. 2017)
	Optimization of water distribution systems	Analyzing WDS solutions (obtained by evolutionary algorithms) to generate rules or relations between variables to obtain better solutions	Solution space	WDS solutions obtained from evolutionary algorithms	(SOM) and Bayesian Network	Significant reduction in search space	(Izquierdo, Montalvo et al. 2014)
	Optimization of wastewater treatment pumps	Energy optimization of wastewater pumps in a treatment plant to conserve energy	Real time data	Speed of pumps elevation of wet well	Regression 1) Several among which kNN, decision (regression) trees, MLP	Incorporation of time series is important for accuracy	1) (Kusiak, Zeng et al. 2013) 2) (Torregrossa, Hansen et al. 2017)

					2) Signal decomposition		
	Water treatment plant optimization	Water treatment plant optimization by developing software sensors	Real time data	Data from the SCADA of the treatment plant	Regression 1) generalized least squares regression, artificial neural networks, self-organizing maps and random forests Classification 2) various methods	Expert knowledge inclusion is still important	1) (Dürrenmatt and Gujer 2012) 2) (Comas, Dzeroski et al. 2001)
	Prediction of solids in wastewater	Using CBOD and influent flow rate time series to create a day ahead prediction of TSS (Total suspended Solids)	Time series data	Influent flow rate and influent CBOD (carbonaceous bio-chemical oxygen demand)	Prediction: Multi-layered perceptron, k nearest neighbor, multivariate adaptive regression spline, support vector machine and random forest	Around 70% accuracy for a 7 day ahead prediction	(Verma, Wei et al. 2013)
	Prediction of influent flow rate	Short term prediction of influent flow rate in a wastewater treatment plant	Time series data	Influent flow rate, rainfall data radar reflectivity data	Prediction: Multilayer perceptron neural network	Well up to 150 min, after that a lag develops	(Wei, Kusiak et al. 2013)

chemical water quality	Risk based monitoring	Design of risk-based monitoring program to efficiently monitor water quality	Spatial time series data	Target and non-target monitoring data, effect studies, chemical property and toxicity data	Prioritization, correlation, pattern recognition, PCA, hierarchical clustering, k-means clustering	The design of a risk-based monitoring program is requested by drinking water law	(von der Ohe, Dulio et al. 2011, Sjerps, ter Laak et al. 2014, Sjerps, Vughs et al. 2016) Sjerps et al, in prep.
	Occurrence surveys	Temporal and spatial analysis of available monitoring data	Spatial time series data	Target monitoring data	Temporal or spatial profiles, prioritization and ranking techniques	Large scale reliable exposure data	(Loos, Gawlik et al. 2009, Loos, Locoro et al. 2010, Loos, Carvalho et al. 2013, van Loon, Sjerps et al. 2017)
	Non-target data interpretation (including suspect screening)	Identification of formerly unknown features detected in non-target screening analyses	Lab-scale, environmental data	Non-target screening data	Cheminformatics, PCA, clustering techniques, temporal or spatial profiles, chemical patterns: isotopic patterns, mass defects, homologous series, functional groups	Identification of formerly unknown or non-detectable chemicals in environmental samples	(Hoh, Dodder et al. 2012, Moschet, Piazzoli et al. 2013, Chiaia-Hernandez, Schymanski et al. 2014, Schymanski, Singer et al. 2014, Thurman, Ferrer et al. 2014, Gago-Ferrero, Schymanski et al. 2015, Ruff, Mueller et al. 2015, Zonja, Delgado et al. 2015, Sjerps, Vughs et al. 2016, Hollender, Schymanski et al. 2017, Merel, Lege et al. 2017, Muz, Ost et al. 2017)
	Identification of nano- and microplastics	Identification of nano- and microplastics is time-consuming and	Lab-scale, environmental data from	Fourier-transform infrared (FTIR) spectra and Thermogravimetric (TGA) spectra	Comparison, correlation and clustering of IR	Nano- and microfibers are identified manually and has	(Mintenig, Int-Veen et al. 2017)

	not yet high throughput	spatial time series		spectra, distance and similarity analyses	the potential to be mined with data mining techniques	
Identification of transformation products (TPs)	TPs present in the environment often unknown, can be detected in non-target screening, but rarely identified. In silico prediction can facilitate identification	Lab-scale, environmental data from spatial time series	Non-target screening data, chemical structure data and known biotransformation rules.	Volcano plots, statistical tests, fold changes, PCA, clustering, similarity searches. Prediction tools: enviPATH	Database of TPs that are formed in drinking sources and during drinking water treatment	(Fenner 2016, Schollée, Schymanski et al. 2016, Wicker, Lorsbach et al. 2016, Schollee, Schymanski et al. 2017)
EDA=Effect Directed Analysis	Identifying toxicants from environmental samples by linking effects to hazardous chemicals with fractionation	Lab-scale, environmental data	In vitro effect tests of samples combined with chemical analysis	Correlation	EDA is a promising tool for identifying predominant toxicants in complex environmental mixtures	(Brack, Ait-Aissa et al. 2016)
Environmental fate and exposure modeling	Predictive modeling for chemicals in water systems	Spatial model scale	Chemical structure and properties, customer data, climate data, catchment-based land use activities and soil characteristic. Validation with water quality data	QSAR model, correlation, SML, GIS model	Chemical exposure prediction supports decision making	(Mackay and Paterson 1991, Mackay, Shiu et al. 1992, Scheringer 2009, Scheringer, Jones et al. 2009, Zarfl, Scheringer et al. 2011, Wambaugh, Setzer et al. 2013, Judson, Houck et al. 2014, Zijp, Posthuma et al. 2014, Comber, Smith et al. 2018, Schulze, Sättler et al. 2018). TRANSATOMIC.xlsx

	Removal prediction	Study and predict chemicals removal in treatment systems	Model scale	Chemical structures and available removal data	QSAR model, SML	Predicted removal efficiencies supports decision making	(Wols and Vries 2012, Vries, Wols et al. 2013, Vries, Bertelkamp et al. 2017)
	Human health effects prediction	Study and predict chemical human health effects	Model scale	Toxicological and effect data, in vitro toxicity data, chemical structures	Toxicity databases, bioassays results, QSAR models, read across, AOPs - informed computational models	Reliable toxicological experimental data is sparse, high throughput human health risk evaluation is valuable	(Judson, Richard et al. 2009, Wambaugh, Setzer et al. 2013, Judson, Houck et al. 2014, Blackwell, Ankley et al. 2017, Wittwehr, Aladjov et al. 2017, Zang, Mansouri et al. 2017, Baken 2018, Brunner, Dingemans et al. submitted)
	Waste-water based epidemiology	Population characterization: deriving public health and (illicit) drug use from wastewater analyses	Spatial time series combined with real-time data	Target, and non-target screening data, pH, conductivity, rain fall, flow, mobile phone data	Correlation, PCA, temporal and spatial trend profiles	Wastewater contains information of human health characteristics that are not yet fully explored	(Bade, Causanilles et al. 2016, Causanilles, Baz-Lomba et al. 2017, Causanilles, Kinyua et al. 2017, Causanilles, Ruepert et al. 2017, Causanilles, Nordmann et al. 2018)
biological water quality	Large-scale genomic data mining	Species identification, community characterisation, assessing genomic diversity, finding patterns of gene expression, gene	global databases	Sequence data (genomic), microarray data (DNA expression), interaction variable data (chemical/ physical, regulatory, protein modifications, etc.)	Combination of different algorithms	Moderate-good (fair amount of user expertise required for correct data treatment and interpretation)	Huttenhower & Hofmann (2010); Ju & Zhang (2015a,b); Kennedy et al. (2010); Lee et al. (2008); McPerson (2009), Romano et al. (2017); Segata et al.

		discovery, drug discovery, enzyme discovery, community functioning, toxic potential, disease potential					(2011); Wassenaar (2004)
	Risk management						
	Quantitative microbial risk assessment (QMRA)	Water safety management (water safety plans, sanitation safety plans)	Several years, water system-dependent	Quantitative data or assumptions on pathogen occurrence and exposure, frequency of hazardous events and severity of hazards, pathogen removal rates by relevant treatment processes, dose of reference pathogens and infection response, risk of illness per infection, disease burden and susceptible population, scenarios accounting for variability and uncertainty	Classification, regression	Good	WHO (2016)
	Predicting cyanobacteria toxins	Remote monitoring (predicting) of cyanotoxins, early warning system	Near real-time (daily), global (1); Multi-year, reservoir (2)	Remote sensing (satellite) (1); water quality (including microcystins) (1,2)	Deep learning (1); multivariate adaptive splines (2)	Good	1: Chang et al. (2014); 2: Garcia Nieto et al. (2010)
	Assessing risk of faecal infection	Exposure risk assessment, tracking faecal contamination	Years, water (distribution) system	Hydrology, hydrometric (1); water quality and faecal indicator species (1,2)	Classification trees (1); machine learning (2)	Poor-moderate (1); good (2)	1: Bichler et al. (2014); 2: Dawsey & Minsker (2007)
	Infection and disease management	Discovery of infections and antimicrobial resistance patterns	Months, hospital	Public health surveillance data, hospital infection control data	Association rules	Moderate, dependent on expert evaluation	Brossette et al. (1998)

	Infection and disease management	Assessment of climate change effects on disease occurrence (decision support system)	12 years, global literature	Literature data (key facts), environmental data, data on food and waterborne disease occurrence	Network maps (classification)	Good	Semenza et al. (2012)
	Management of species invasions	Assessment of factors contributing to species invasions	Decades, global shipping network	Vessel movement, ballast discharge, various ecological & environmental data	Graph clustering	Good	Xu et al. (2014)
	Ecosystem response monitoring						
	Coastal management	Discovery of ecological thresholds	Weeks, coastal area	Flow cytometry data, water quality	Artificial neural networks	Good	Pereira et al. (2009)
	Multiple stressor effects on community functioning	Predict response of macro-invertebrate community to multiple stressors	Years, river basin	Water quality, hydrology, hydromorphology, macrofauna & macrofauna traits	A priori association, boosted regression tree	Good	Mondy et al. (2016)
	Predicting fish communities	Predicting fish community composition and species richness	River sections	Land cover data (GIS), hydrologic data, topographic data, data on fish communities	Classification and regression tree (CART), random forests (RF)	Moderate-good	He et al (2010)
	Modelling eutrophication	Predicting (harmful) algal biomass	Multi-year, lake	Water quality	Deep learning	Moderate-good	Chen & Mynett (2003)

4 Current applications of data mining: interviews with practitioners

4.1 Interviewees

A number of practitioners of data mining within and outside the water sector (see Table 5) have been interviewed by Henk-Jan van Alphen (KWR) with respect to current applications, technologies, drivers, opportunities and challenges. This chapter provides an integrated summary of these interviews.

TABLE 5: OVERVIEW OF INTERVIEWEES.

Name	Organization
Rob van Putten	Waternet
Jurjen den Besten	Oasen
Jan Urbanus	Evides
Stijn Heemskerk	ABN AMRO
Dumky de Wilde	Professional on data-analytics
Laurens Koppenol	ProRail

4.2 Regarding specific applications

The interviewees do not have an exhaustive list of applications, but gave some examples in the interviews for which data analytics have been used so far at their respective companies.

4.2.1 Waternet

Waternet has used gradient boosting to determine the best location of monitoring wells. Mr Van Putten was involved in the application of Deep Learning (DL). With this method, he gained knowledge in *IJkdijk* project 4 years ago. They equipped a dike with sensors, to measure its state. Mr Van Putten then applied DL to a case of industrial water use. Data was fed from images from video taken from water meters, and the goal was to read numbers. Waternet has also used DL to identify deer on video images of drones.

4.2.2 Oasen

The issues for application of DM at Oasen are related to asset management. Examples of applications are pipe break forecasts and optimizing the replacement of water meters. The core principle is that applications contribute to an improved Oasen's customer satisfaction. No studies are done based on correlations in the data (without prior hypothesis). These correlations are not so easy to find, because the data is complex and often not so well structured.

4.2.3 Evides

Data analysis at Evides is relatively new. There is currently a data team of about 13 people (not fte) that focuses on the subject. Evides wants to focus more on statistical analysis, because internally there is a sense that a better performance can be achieved. As a development standard the team uses a SCRUM Agile method, which involves a SCRUM master and a *product owner*.

Evides, together with external data analysts, has made an analysis of water quality data from the Biesbosch storage areas (time series of 30-40 years length). In it, four different trends in water quality were identified, three of which were already known by Evides. Although the analysis did not yield any real new insights, it did confirm existing insights.

Currently, data analysis is also used to detect leakages and background losses by comparing the inflow and outflow of a district metered area (DMA). Such analysis is carried out by the Israeli company Takadu. In the current projects, most attention is paid to data validation.

Evides used data visualization to investigate the relatively high percentage of Non-Revenue Water (NRW). One hypothesis is that this is partly due to invoicing. The invoice data was combined with the municipal administration (BAG) and visualized with ArcGIS. In that way, it was clear which addresses do and do not receive an invoice. GIS visualizations provided immediate insights and lead to follow-up questions. Individual questions are then addressed by a combination of (in-house) specialists with support from ESRI specialists, which Evides hires for this purpose.

4.2.4 ABN

The retail department of ABN mainly uses customer data, with the purpose of enhancing its marketing strategy. Most data mining issues involve optimizing the communication with existing customers. For example, which messages are sent to which customers at which time. This can be done from a commercial purpose, such that customers will increase the volume of savings or investments, or from a customer satisfaction perspective, in which the goal is to facilitate procedures such as the application for a debit/credit card. Specific applications are chosen on the basis of the strategic objectives of the bank and strongly driven by the perspective of the retail marketing professionals.

A good example is attracting new investors. The bank has a lot of data about the current investors: when they signed in with the bank; what they invested in, which quantities are involved per transaction and all other transaction data and personal data. By means of data analytics, it can be determined which customers should receive notifications about new forms of investment and at which time. This can be done, for example, with the use of Decision Trees (DT). The analyst chooses (i.e. on the basis of correlations) a number of variables that relate to the likelihood that someone will invest or not. A computer can generate a decision tree of customer data, which classifies the customers into groups with an higher likelihood to invest in a certain number of iterations. The system then issues a group with the highest likelihood to invest, which is reported to the bank analysts. Then based on this information, a marketing or communication action is executed. A feedback loop is generated to the system, such that this action is subsequently analyzed in order to improve the model.

4.2.5 De Wilde

Much of what Mr. De Wilde did is related to data visualization. He has worked making scientific insights accessible to policy staff at the immigration policy department of the IND (Dutch immigration and naturalization service). In his case, data mining for migration data was the main task. To make data accessible to stakeholders, in this case IND, the essential part is that it has to be visible at a glance. But to give meaning to data, many choices have to be made in which dimensionality issues appear. Every choice for a selection of a scope on a data subset is also a choice to make it more relevant than the other. If the scope is unclear, the more likely that information is lost with the stakeholder. This in fact is related to the use of color and representation.

Mr. De Wilde has also done research into the frequency of elevator confinements for the fire brigade. After collection, a number of peaks were visible in the data. Analysis of the data from the fire brigade itself did not produce anything relevant. However, a combination of data with cross analysis of news reports showed that it was correlated with power failures.

4.2.6 ProRail

The initial scope of DM activities at ProRail was predicting failure of infrastructure. An assessment has been done of the most common causes of disruption of the trains and they were analyzed using data analytics. One notable success was identifying the likelihood and location of people walking on or along the train tracks, one of the main causes of disruptions in train traffic. They also looked into the decision being made in guiding trains through the network. The question there is whether it is possible to simulate human decision making.

ProRail has a datalab of 10 fte which was founded by the innovation department and financed by the IT-department. Formally it is not part of a specific department which gives more freedom to experiment. Data scientist are being 'lent' by other department and around the 10 fte core is a layer of trainees and external analysts.

4.3 Regarding tools and methods used

During the interviews a number of tools were mentioned which are widely used for data analytics:

- SQL
- SAS
- R
- Python
- ArcGIS (data visualization)
- Microsoft Azure Machine Learning Studio
- Microsoft Excel
- Google Search

During the interviews a number of machine learning methods were mentioned which are widely used for data analytics:

- Gradient Boosting
- Decision Trees / Random Forest
- Artificial Neural Networks (ANN)
- Support Vector Machines (SVM)

The interviewees agreed that the application of analytical methods is not the biggest challenge during the whole process. Obtaining and adjusting the data is considerably more complex and time-consuming than the actual analysis. Mr. De Wilde estimates that 75% of the time is spent preparing the data, 10% on the analysis and visualizing 15%.

According to Mr. Den Besten, the majority of scientific research in data analysis focuses on the development of analytical methods. It is relatively easy to apply that knowledge to your own issues, for example with packages such as Microsoft Azure Machine Learning Studio.

For low-level issues, Mr. Van Putten uses Google to search for suitable methods or existing solutions. As an example, he mentions recognizing *deer* on videos of drones. For more complex methods such as ANN and DL, he follows the (very) recent (and dynamic) scientific publications.

Since Mr. Heemskerk predominantly deals with very large, well-structured databases, SQL is the basic tool for him. Knowledge of the specific way in which the databases are organized is also necessary in order to find the right fields. Excel, R, SAS, or Python is usually used for further analysis. Mr. Heemskerk has also experimented with DL, but in his specific field, it performs with a lower performance than Random Forest, which is now the benchmark at ABN.

Mr Koppenol notes that there is a sense of urgency in applying a model to the data. He distinguishes roughly two issues: classification and regression, the latter just providing you with a number that needs to be interpreted. He also uses ML.

Both Mr. De Wilde and Mr. Urbanus underlined the importance of data visualization. The clear visualization of data often provides many insights without actual analysis or transformation being applied to data.

4.4 On the relationship with the need to domain knowledge for data mining

All interviewees indicate that data analysis without specific subject knowledge does not lead to meaningful results. Mr. Urbanus experienced this when a number of data analysts from outside the water sector started using the quality data from the Biesbosch storage areas. In doing so, they were not able to produce any new real insight.

According to Mr. Den Besten, there are three types of skills required for a data analyst in an organization such as Oasen.

(1) Hacking skills, such as modeling and programming;
(2) statistics skills and
(3) domain knowledge.

Mr. De Wilde also indicated that if those three types of skills can’t be united in a person, then there must be a bridge closing the gap between the domain knowledge and the other two areas.

Mr. Den Besten adds that the relationship with domain knowledge at Oasen is 'unconsciously skilled'. In their case, data analysts work (and are often trained) as asset managers.

Subject knowledge is also leading in the field where Mr. Heemskerk is active. The marketeers make a proposition and must also be able to provide evidence that the proposition has been delivered. This is also reflected in the type of data analysts that ABN recruits. They must have good social and communication skills, while also being able to translate their results into the practice of the marketing professionals. Mr. Heemskerk explains that he thinks marketeers will not be able to work without elementary data skills in the future, whereas nowadays data analysts are the ones who bridge the gap.

Mr. Van Putten also agrees that the combination between data analysis and specific knowledge is particularly interesting. Only data analysis focused on methods can lead to strange results, if the subject knowledge is not taken into account. On the other hand, data knowledge may make current practices obsolete. The combination between data and specific knowledge is the most powerful.

Mr. Koppenol also does not use data analytics without subject knowledge, but mostly out of data and time constraints. With enough data and enough time, pure data analytics may lead to useful results.

4.5 Regarding data management

The interviewees agree that data management is the weakest link in the chain of data analytics. According to Mr. Den Besten, in the water sector this is the bottleneck of data analytics. The quality of the results can’t be better than the quality of the data you use. As such the largest gain to be made by his water company is in data management. He also notes that the development of methods and tools is mainly taking place outside of the water sector and that water companies can profit from that instead of developing their own tools. Data generation and the collection of good data is instead a very specific problem for water companies in which water companies can innovate themselves.

At the moment most of Oasen's data comes from sensors and observations from operators. The latter category is crucial because it often concerns data about anomalies. At this moment it depends on the expertise and assumptions of the operator which data is collected in the event of a pipe break. All employees should be aware that data could be the key to the solution of many issues, and that properly recording data is therefore a high priority of the company.

Mr. De Wilde indicates that if you want to use data for problem solving, you also have to look at how the organization is structured to deal with data. In fact, in most cases everything revolves around digitalization/automatization of existing processes, so that a constant data flow is available. Then it is also a concern who has access to the data, which external data you need and which external tools are useful to use.

According to Mr. Urbanus, the use of external tools is a problem at Evides. The IT department is struggling to keep up with the rapidly developing needs of the data analysts. It is already difficult for IT to meet the needs of the business practices and needs. Data analysts often want to use new programs, libraries, software which in many cases must then be adapted to existing security protocols by its IT department. That has shown to be a demotivating force for data analysts.

Errors in the data, according to Mr. van Putten, can be an important factor in the quality of the results. Especially for anomaly detection, uncertainty and reliability of forecasts, data validation plays a major role. His advice: be open about the methods you use, use metadata and be very open about the error margins used. Mr. Van Putten often adds a column with a reliability indicator, based on his own criteria for his track. The recording of metadata is also of great importance for him.

Mr. Urbanus wonders how clean the data should be before reliable analyses can be made. And second, whether you can keep a data analysis team happy for a long time if you steadily put them in charge of long data validation processes. Mr. Koppenol estimates that about 50% of the data preparation process can be skipped while still getting reliable results from the data.

In the past, Mr. Heemskerk has experienced few obstacles on data management and data quality. A lot of data (especially transactions data) is available at his organization and due to its nature (personal financial data) it is well maintained and updated. In contrast, in his case, the vast amount of data makes it somewhat difficult to focus on the research questions of interest and not the data itself.

Mr. Koppenol also considers data quality the bottleneck of the process. He gets most of his data from the system logs from infrastructure, but also there it is not always clear what the data actually represents. ProRails data is often matched with public data sets, such as geo data or weather data. Some sensors have been put in place but according to Mr. Koppenol they need to be in place for a longer period (2 to 3 years) to be useful.

4.6 Regarding current data infrastructure

Waternet has a *data point* where data is collected and made available (about 18 FTE) and a *datalab* for analysis (as of now only Mr. Van Putten). There is cooperation with universities on the subject. The datalab focuses in particular (and at this moment) on sensoring and automatization (robots). Waternet's datalab is now only intended for Waternet’s people, but cooperation is sought with the municipality of Amsterdam.

For the data warehousing, Waternet is currently under negotiation with Microsoft for the use of their Azure cloud computing service. Data storage and availability is a big challenge for the company. Due to the network restrictions it is difficult for employees to store large amounts of data or to access different data sets directly from Waternet’s intranet. The result is that people often use their own laptops, USB sticks or other forms of storage (less reliable). This makes it difficult to manage and share data. The data from the treatment plants (77.000 sensors) is distributed over two databases (i.e. drinking water and sewer), but these are difficult to access due to additional protocols for cyber security. No use is currently made of data compression.

Mr. De Wilde notes that if you really want to work with Big Data (> 10 million documents) organizations like drinking water companies do not want to store and manage that themselves. The data centers of Microsoft, Google and Amazon are cheaper and more reliable for such tasks.

ProRail's data is stored in a private cloud hosted by KPN, but negotiations about a new environment are taking place, Mr. Koppenol uses a Hadoop environment to do the actual analytics. Storing data with a third party poses no issues for ProRail and neither does using third party systems to run analyses. The data is generally not accessible for the public, although there are a few API in place with map layers of the infrastructure.

4.7 Regarding obstacles and driving forces

4.7.1 Obstacles

Mr. Van Putten mentioned issues of ownership and opening up data as a major obstacle. Being able to produce data could yield a lot, but it can also lead to conflicts about the interpretation during, for example, lawsuits. Everyone wants access to data, but no one wants to be responsible for managing and cleaning data.

Another obstacle that has been mentioned by all entries was the fact of bringing to the organizations sufficiently qualified data analysts. That is mainly a matter of money. The price for good data analysts is so high that for example, a bank like ABN with their existing salary structure can’t afford to hire good people. This also applies to water companies. Additionally, to be of use as a data analyst in the water sector, you also need some knowledge of the relevant processes in the water cycle.

The question is whether public organizations can let the right people know what their digital ambitions are. Think of interim data scientists in companies such as: *VODW*, *Xomnia* and *Anchormen*. In most cases, data analyst rates are higher than 100 €/hr. The advantage in the case of public organizations is that data scientists can publish about their results and that the issues that they are working on have much relevance to society. Mr. Koppenol argues that public organizations can also be attractive because the wide application of data analytics is still in its infancy and as a young data analyst you can really make a difference, instead of being one of the many at a tech company.

A third obstacle is the availability and quality of data, which has been described before.

Organizational dynamics can also work as an obstacle. Mr. De Wilde mentions the Immigration and Naturalization Service (IND) as an example, where due to a culture of '*security concern*' within the organization, it is very difficult to use new digital tools. Something that Mr. Urbanus also signals at Evides. Mr. De Wilde also found out that when there is not a high ranking manager(of IND and other cases) in charge of the development of data analytics, the implementation of such practices becomes unsuccessful or unlikely to develop in short time.

According to Mr. De Wilde, the value of a data team is determined either by their peers within the organization or the service which they provide to customers. This is only possible if people within an organization start asking the questions for which data analysis is needed. You have to train people to know that they can ask certain questions and ask the proper/right ones. Only then, data analysts teams can show what you can do with the data.

According to Mr. Van Putten, the islands that exist in the water sector are frequently seen as obstacle. People are afraid to lose their jobs to people with other competences and skills. Traditional knowledge sometimes conflicts with (results from) data science. This was evidenced from the IJkdijk project (previously mentioned).

4.7.2 Strengths and driving forces

Mr. Van Putten cites better cooperation between water boards and water companies as a main driving force in making data, methods and results public.

Mr. De Wilde also sees the collection and availability of data as an important driving force. Especially, when it means that all kinds of data can be combined. For example, the municipality of Amsterdam has shared a lot of geographical data via web services such as *maps.amsterdam.nl*.

Mr. Urbanus sees a growing awareness of the value of data at Evides. The importance of collecting more and new types of data is also seen. This certainly also applies to Evides Industrial Water (IW), where data has a lot of economic value for the company.

The fact that at Waternet the datalab is so highly regarded on the strategic agenda has given momentum to the awareness of employees of data. Data science is just beyond the hype. Many employees from various departments have taken or are taking Python courses. Internships on subjects such as robotics and drones are being executed and the possibilities for permanent employment are being looked into. In a way, self-managing teams reduce the number of middle management positions available and all kinds of employees search for ways to do something with data.

Mr Koppenol sees a shift in how organizations deal with ICT. What would really boost results is when data is freely available in the cloud with a lot of freedom for analysts to do their analyses. There is some cooperation with the Dutch Railways in equipping trains with sensors. This also stimulates cooperation and integration between the two organizations.

4.8 Regarding data and decision making

The decision-makers at Oasen's asset management are positive about the results obtained so far, but they keep some reserve in making decisions on the results. Usually when the system generates a prediction (of a pipe break), one does not immediately make a decision, but rather first looks at whether the prediction becomes true. There is a fundamental problem with predictive/forecast algorithms. If you act proactively on the basis of a forecast algorithm (for asset management), you will never be able to evaluate whether the system estimate was right or wrong, and then it is impossible to validate the forecast algorithm.

At Evides, the increased use of data means that it is now also expected that certain statements are substantiated with data or statistics. The processing of data into information (for example through visualization) also leads to new insights that can support decision making. Mr. Urbanus recognizes that much in the water sector is done out of habit. Results from data analysis can offer a different perspective for different issues. As an example, at Evides, there is the issue of condition-dependent maintenance. Until recently, fire hydrants and valves were checked annually or biennially. That was a matter of habit. Perhaps data analysis can substantiate this choice or suggest a different maintenance regime. Perhaps these checks are not necessary at all, or one/some fire hydrant(s) must be checked more often than the other(s).

According to Mr. Van Putten, it is not true that decision-makers need to judge data science and traditional knowledge separately. Policy recommendations are being formulated bottom-up, i.e. consisting of insights from both areas. Ideally, subject knowledge and experiential knowledge will be integrated with data science models in the future.

Mr. Koppenol stresses the importance of showing the added value of data analytics and finding the right person in the organization to show it to. It also helps to let decision makers join in the process and show which choices are being made and why. There is sufficient goodwill in the organization to try this out, but it needs to prove itself.

4.9 Regarding opportunities for the future

Mr. van Putten does not mention any specific application for data mining in the water sector, but considers all engineering issues to have a relevant link with data analysis. He suggests that each engineer within a water company should have a minimum knowledge of data science competences. He gives a Python course every year, and he sees that more people from diverse backgrounds register.

Mr. Den Besten sees that in the future there are plenty of opportunities for applications which may improve customer satisfaction. For example, by improving the service level to customers by reducing the contact and delivery times or better planning of system maintenance. Also, simple information can be transmitted to customers through the use of chatbots. In another organization, the chatbots work so well that even the own employees use them to find data in their systems.

According to Mr. Den Besten, the water sector should start focusing now on the collection of data which may be of relevance for future issues. Business cases have to be formulated for data collection or a valorization model that shows the potential revenue of systematic data collection and management. Based on current methods of data analysis and available information, an analysis can be made of the gap between the analytical potential and the availability of data.

Mr. Urbanus expects more applications of machine learning in the coming years in the water sector. Vitens and Evides currently perform their billing process through *Facturatie BV*. In general, data which is available from that process is insufficiently used at the moment. That offers interesting possibilities for the future, to make the customer satisfaction more efficient and effective.

Mr. Urbanus indicates that there is still little understanding of what exactly happens in the water distribution network. The combination of hydraulic data and water quality data can lead to more insights. The need to collect this data from their system is great, but implementing new sensor locations is expensive in terms of time and money. Currently, there is a great need for cheap, easy-to-install sensors (for both quantity and quality). The autonomous inspection robot AIR can also yield a lot of data which will be of relevance. The challenge is to turn that data into useful information.

At Evides IW, there is a need for good forecast of incidents. At IW, the redundancy of the pump system is much smaller than that of the drinking water. A lot of demineralized water is also supplied, which means a heavy load on infrastructure. Interruption of supply can lead to contractual fines with IW customers. Therefore, there is a great need for good forecast to avoid such issues. The software used for the pumps already has all kinds of data available and the additional measurements could be done with sensors. These sensors are considerably cheaper and easier to install than those in the distribution network.

Within ABN, much thought is given to the role of banks in the long term. The expectation is that ICT companies will take on many tasks from banks. In that regard, as a result of the Payment Service Directive 2 (PSD2), banks have to share their transaction data with third parties at the request of their customers (under certain conditions). This means that other parties (such as Google or Amazon) can also use this data. The strongest position of the banks (in relation to their customers) is then in the field of automated financial advice.

Within ABN, many ideas are being currently developed at the moment, such as using data from sensors of ATMs, converting speech to text at call centers or using DL for risk (security of asset positions).

Mr Koppenol sees a future in which much more is being measured. Not by the data scientist but by everybody in society. This will lead to much more data and to much more impact of data analytics.

5 Potential for new applications

5.1 From a domain perspective

An overview of potential new applications, driven by emerging interest in different domains within the water sector is given in Table 6. This list was compiled by the authors of this report from their reading of the literature, knowledge of their fields and current initiatives in their professional networks.

5.2 From a data perspective

An overview of potential new applications in the water sector driven by the emerging availability of new and upcoming data sources is given in Table 7. This list was compiled also by the authors of this report from their reading of the literature, knowledge of their fields and current initiatives in their professional networks.

5.3 Discussion and outlook

5.3.1 Surface and subsurface water

For many fields of ecohydrology and geohydrology, the underlying processes are well understood and therefore physical models are predominantly used to understand/assess/predict the phenomena of interest. Examples of such fields are groundwater dynamics, soil water dynamics in unsaturated zones and evapotranspiration, transport of heat and water regarding ASR (Aquifer storage and recovery) and ATES (aquifer thermal energy storage).

However, there are opportunities for data mining techniques to be of help by adding new knowledge, especially where ‘fuzzy’ variables are involved and therefore a physical model alone is not sufficient to reveal causal relationships. These ‘fuzzy’ variables, which can also be described as imprecise or linguistic variables, include those related to human behavior (e.g. operational conditions of wells, management types of nature restoration) or variables which represent multiple (functional) properties that cannot be attributed to a simple physical metric (e.g. spectrum information of satellite images, species composition).

One of the difficulties in applying data mining techniques in the field of Ecohydrology/Geohydrology lies in the fact that the existing datasets usually cover a certain limited spatial extent, whereas the area of interest is typically large (e.g. the area where infiltration water passes through, the area nature restoration measures are implemented). The mismatch of spatial scale probably continues to be an issue in future due to the labor-intensive nature of data acquisition methods in this field. A promising way forward is to combine small-scale datasets with area-covering databases (which are available from other data sources, e.g. RIVM, KNMI, RIWA), or with process-based models, in order to extrapolate the insight gained from the small scale dataset. Analysis of satellite or drone image, which combines area-covering images and intensive ground survey points, is a good example of such an approach.

TABLE 6: OVERVIEW OF DATA MINING OPPORTUNITIES FROM A DOMAIN PERSPECTIVE.

	label	technique/ description	method class	type of data	state of development	reference
hydrology	Manure control for minimizing leaching & maximizing yield	Seeking for optimal manure application to agricultural land to maximize yield and minimize environmental impacts (e.g. nutrient leaching to ground water)	Regression	Soil map, land use map, hydrological data (soil water flows), meteorological data, manure policy, behavior of farmers, groundwater quality data	'Precision agriculture / smart farming' (mainly for the purpose of maximizing yield) is already taking place in many regions in the world (Wolfert, Ge et al. 2017) . Dutch government invested in making satellite image available for promotion of precision agriculture (www.spaceoffice.nl/nl/satellietdataportaal).	
	Precision nature restoration	Taking spatial and temporal information into account to make effective planning of nature restoration measures	Regression?	Soil map, climate data, hydrological (modeled) data, topographical data, plant characteristics (e.g. N content) predicted by drones	'Precision conservation' has been implemented in the States since early 2000's, mainly for reducing erosion risks (Berry, Detgado et al. 2003)	
	Automatic quality control of groundwater chemistry data	Testing quality of groundwater chemistry data for chemical consistency and outliers, as pre-treatment needed for obligatory data delivery to BRO	Anomaly detection Classification	Chemical variables in groundwater (on each filter of wells, with different sampling time)	Manual quality control has been done for Province Zeeland	
	Predicting forest fire	Predicting fire hazard level of a day or burned area	Regression, Classification	Meteorological data Soil and vegetation data	Prediction method of forest fire using machine learning has been developed for arid forest systems (Cortez and Morais 2007, Sakr, Elhajj et al. 2010). Process-based model might serve as a	

	label	technique/ description	method class	type of data	state of development	reference
					better method for the ecosystems in the Netherlands.	
water treatment and distribution	Leak Localization	Test EPANET trained algorithms for leak localization with real leakage events and sensor data from around its occurrence to asses validity for real world examples		Sensor data and leak location	Possible implementation of multiple leak localization algorithms	(Soldevila, Blesa et al. 2016) (Mashford, Silva et al. 2009)
	Water treatment plants optimization	There is much data being recorded in a water treatment plant which can be analyzed using data mining techniques for optimal performance and asset management purposes		SCADA data	Several researches have been performed on parts of the water treatment process with a main focus on energy efficiency	(Comas, Dzeroski et al. 2001) (Kusiak, Zeng et al. 2013, Torregrossa, Hansen et al. 2017) (Dürrenmatt and Gujer 2012)
chemical water quality	Risk based monitoring	Temporal and spatial trend profiles, correlation and clustering	Combination of methods	Target and non-target monitoring data, land use data, hydrological data	Pilot study performed at Vitens. Potentials for other drinking water companies	(Sjerps, Brunner et al. in preparation)
	Wastewater based epidemiology	Correlation, PCA, temporal and spatial trend profiles	Combination of methods	Target, and non-target screening data, pH, conductivity, rain fall, flow, mobile phone data	Pilot study TKI and WATCH. Potential to expand the broad range of health indicators. Data mining needed for integration of different data (formats), not yet implemented	(Causanilles, Baz-Lomba et al. 2017), TKI, WATCH

	label	***technique/ description***	***method class***	***type of data***	***state of development***	***reference***
	Identification of unknowns, including transformation-products	Correlation, PCA, temporal and spatial trend profiles, Cheminformatics - QSARs, machine learning Prediction tools	Combination of methods	HRMS target and NTS data Chemical structure data, biotransformation rules	Current analysis includes manual steps, however potential to mine data with data mining techniques Ongoing projects target automation/data mining - BTO 2018 high throughput identification - BTO 2018 VO Integration - DPWE robustness treatment trains - AquaNES - UVPD However: machine learning not part of these efforts yet, yet highly relevant!	(Krauss, Singer et al. 2010, Hug, Ulrich et al. 2014, Schymanski, Jeon et al. 2014, Bletsou, Jeon et al. 2015)
	Identification of nano- and microplastics	Spatial comparison of detected spectra	Spatial comparison	Fourier-transform infrared (FTIR) spectra and Thermogravimetric (TGA) spectra	Currently manual analysis, potential to mine data with data mining techniques	(Mintenig, Int-Veen et al. 2017)
	Chemical exposure modeling	QSAR model, GIS model, correlation	Combination of methods	Chemical use and property data, spatial data (rainfall, soil type, hydrology)	TRANSATOMIC.xlsx Pilot VO Waterkwaliteitskaart	(Mackay and Paterson 1991, Mackay, Shiu et al. 1992, Scheringer 2009, Scheringer, Jones et al. 2009, Zarfl, Scheringer et al. 2011, Wambaugh, Setzer et al. 2013, Judson, Houck et al. 2014, Zijp, Posthuma et al. 2014, Schulze, Sättler et al. 2018). TRANSATOMIC.xlsx

	label	technique/ description	method class	type of data	state of development	reference
biological water quality	Geohydrology	Predicting hydrology from DNA material	Machine learning	Genomic sequences, hydrology	Early	Good et al. (2018)
	Infection and disease management	Assessment of climate change effects on disease occurrence (decision support system)	Network maps (classification)	Literature data (key facts), environmental data, data on food and waterborne disease occurrence	Intermediate	Semenza et al. (2012)
	Management of species invasions	Assessment of factors contributing to species invasions	Graph clustering	Vessel movement, ballast discharge, various ecological & environmental data	Early	Xu et al. (2014)
	Predicting fish communities	Predicting fish community composition and species richness	Classification and regression tree (CART), random forests (RF)	Land cover data (GIS), hydrologic data, topographic data, data on fish communities	Early	He et al (2010)
	Predicting cyanobacteria toxins	Remote monitoring (predicting) of cyanotoxins, early warning system	Deep learning	Remote sensing (satellite); water quality (including microcystins)	Intermediate	Chang et al. (2014)
	Understanding drinking water advisories (DWAs)	Identification of factors contributing to DWAs	Decision tree	Data on likelihood of DWAs, water system characteristics	Early	Harvey et al. (2015)
	Public response analysis	Assessment societal relevance of environmental events	Association rules	Social media mentions of specific terms (Twitter, Google Trends)	Early	Cha & Stow (2015)

	label	technique/ description	method class	type of data	state of development	reference
	Disease and infection management	Connecting disease and infection data on pathogen occurrence outside hospitals		Disease surveillance data, infection management data, genomic data	Idea	
	Quantitative ecological risk assessment (QERA)	Stepwise approach to quantify ecological risks and evaluate suitable management scenarios		Environmental data, data on management effectivity, data on disturbance events	Idea	

TABLE 7: OVERVIEW OF DATA MINING OPPORTUNITIES FROM A DATA PERSPECTIVE.

	label	Description	scale of data	state of development	Possible application
hydrology	Nitrate sensor in soils	Nitrate concentrations in soil water and in leaching water, measured directly or indirectly with sensors	Point data Real-time, continuous measurement is possible (when equipped with wireless data-logger)	Direct and continuous measurements of nitrates in soil are at this moment not operational. Indirect measurement via EC has the highest potential to be applied in field conditions. Methods of nitrate sensors were reviewed upon request of Vitens (SPO bodemsensor (Cirkel, Fujita et al. 2017))	Monitoring/prediction of nitrate leaching Optimizing manure application (on both temporal and spatial scales)
	Distributed Temperature Sensing	Distributed sensoring with glass fiber cable for temperature measurement	Continuous, real-time measurements along the entire length	The sensor has been used in on-goind TKI project Koppert Cress to monitor temperature around geothermal energy storage.	Real-time monitoring of geothermal energy
	Drone images	Drone image with different types of sensor data	Area-covering, up to ca. 1x1 km / flight	KWR purchased a drone (with multispectral sensor, thermal sensor). A certified pilot will be available soon. TKI proposal (on change detection in vegetation using drone) is in preparation	Mapping of vegetation patterns, assessment of evapotranspiration, detection of drought damage to agricultural crops, signaling of pipe leakages and illegal discharges, counting of animals in nature areas.

	label	Description	scale of data	state of development	Possible application
water treatment + distribution	Detailed soil map	Soil maps with detailed categories (i.e. much more than current	Area-covering for whole Netherlands	Provinces are taking initiatives to aggregate detailed maps	Replace current rough soil maps (which is used as model input for PROBE and hydrological models), leading to better model predictions
	Hydro-geo-eco database	Integrated database of hydro, geochemical, and ecological variables obtained in the past projects of KWR.	Point data, scattered all over the Netherlands	No initiative (yet). Data and knowledge is scattered in a number of colleagues of team GEO/ECO. Need lots of efforts to integrate and standardize. LIMS (Laboratory information management system) may help promoting integrated use of existing data within KWR.	Analysis of factors influencing removal efficiency of virus/nitrate/other unwanted substances in infiltration water. Identifying factors to determine successful establishment of specific vegetation groups
	Standardized groundwater quality data	Standardized monitoring data of groundwater quality by BRO (Basis Registratie voor de Ondergrond), which will be ready by 2020	Point data of well (with multiple filters on different depths) from all over the Netherlands Yearly measurement	Provinces (which are responsible for the monitoring) have delivered groundwater data to IHW as a temporary database (which proceeds to the BRO database).	Assessment of groundwater quality change (in relation to, e.g. well clogging, manure policy, etc.)
	Sensor data in treatment plants	Continuous, real-time measurements of water quality & quantity parameters	Measurements inside treatment plants	- Electrical conductivity sensor data and software sensors have been used to predict the chemical water quality in a small-scale groundwater treatment plant (Hoenderloo, Netherlands) - KWR & WLN are discussing ideas to optimize treatment processes (DW, industrial, and WW) with data mining of sensor data and meta-data.	- Optimization of drinking water treatment - Understand and optimize treatment processes and extract operational rules
	Sensor data in water	Continuous, real-time measurements of water	Throughout a drinking water distribution network - for now mostly at supply area inflows, but	- anomaly detection on flow and pressure data is becoming common practice;	- anomaly detection in infrastructure

label	Description	scale of data	state of development	Possible application
distribution networks	quality & quantity parameters	increasingly also at other locations in the networks	- water quality anomaly detection is still in its infancy; - combined analysis of multiple signals in early research stage	- anomaly detection in water quality - event identification - improved understanding of the functioning of the system
Smart meter data in distribution systems	Continuous (real-time) measurements of water quantity and quality at households or in the distribution network	Throughout a drinking water distribution network	Implemented at several demonstration sites, mostly as pilot projects (e.g. SmartWater4Europe; PUB, Singapore; Aguas de Valencia)	- Leak detection - Water quality and temperature monitoring - Automated invoicing - Improve hydraulic models - Social alarm system (e.g. in case of absence of demand at elderly homes)
Asset data	Registration of asset properties, measurements of asset condition and environment data.	At the level of individual assets or components.	In the Netherlands: Ongoing uniform registration of pipe failures (USTORE); KWR develops an automated inspection robot in collaboration with tech partners and water companies. A new initiative has been piloted and is being discusses for a data platform to register assets at the level of components (Citadel).	- Optimize asset management (replacement of pipes).
Social media & customer data	Customer information (statements, opinions, alerts) on Twitter, Facebook, Google Trends, etc.	Data from customers in a distribution network or for a certain customer profile.	The use of social media platforms is widespread and increasing. Data is partly free & accessible. In a KWR study (BTO 2015.024) a correlation between customer data, water quality incidents and causes has been demonstrated.	- Improve customer demand profiles - Improve customer information service - Early warning for water quality and quantity anomalies

	label	Description	scale of data	state of development	Possible application
water quality	Substance structures and properties databases	Databases of compounds including physicochemical properties and/or toxicological information	Chemical space. Meta data increasingly included.	Steadily growing with the goal to cover the entire chemical space. Integration of various data sources and datatypes. Inclusion of transformation products. U.S EPA CompTox Chemistry dashboard: well curated, 761000 chemicals (Williams, Grulke et al. 2017); PubChem 94 million chemicals; ChemSpider 64 million compounds, 243 data sources; StoffIDENT: database of water-relevant chemicals	Use of cheminformatics and QSARs to predict chemical and toxicological properties. Use of databases as suspect lists in suspects screening. Identification of chemicals, including transformation products in non-target screenings. Use of toxicological information for prioritization and risk based monitoring.
	Mass spectral library databases	Databases of MS2 fragmentation spectra, required for structural identification of compounds in HRMS screenings.	MS2 fragmentation spectra. Meta data.	Commercial (mzcloud) and non-commercial databases (Massbank.eu, Mass Bank of North America (MoNA)). Varying curation levels. Need for more high quality spectra, acquired with different parameters and instruments.	MS2 based identification of chemicals in non-target screening data. Substructure searches for transformation products. Contribute own spectra to existing databases. Predict fragmentation patterns through mining of existing databases.
	Chemical water quality monitoring databases	Chemical water quality databases including REWAB, RIWA and BRO (basisregistratie ondergrond) and bestrijdingsmiddelen atlas	Monitoring data in source and drinking water	Monitoring data is collected for authorization and extra-legal purposes. Chemical water quality databases are present and under construction, quality assurance is often under development	Development of risk based water quality monitoring programs, decision support for drinking water companies to produce safe drinking water, decision support for the authorization and/or licensing of chemicals

5.3.2 Treatment and transport infrastructure

Looking at infrastructure, including treatment, distribution and wastewater infrastructure, many utilities and water boards are moving from a reactive towards a pro-active and, ultimately, a forecasting style of water system management. This perspective requires real time monitoring, complemented with real-time forecasting and real time control – which suggests a progressive focus on available data resources and the development of data-analysis tools. Another important ingredient of effective data management is data visualization (algorithms, analytics and platforms) to improve the ability of operators to understand multi-dimensional data.

This process will give rise to smart water distribution grids, with data collected through automated meter reading (AMR) and sensors for water quantity and quality placed throughout networks. The present focus in the application of the data which is collected from these sensors is on event or anomaly detection. Future potential includes a more acute comprehension of water demand patterns, water quality, customer profiles, and the functioning and deterioration of the system through a combination of data sources, e.g. to enhance network efficiency, improve water planning, manage billing and propose new customer services (Cheifetz, Noumir et al. 2017, March, Morote et al. 2017, Stewart et al. 2018).

Apart from these direct and (near) real time measurements, additional data sources are helping to improve asset management of water transport and distribution and wastewater transport infrastructures. Up till now, detailed information of components of assets is lacking, but this turns out to be an essential component of life predictions with an acceptable accuracy. A new collaborative initiative in the Dutch sector –the data platform *Citadel*– could help to improve subsurface asset information. *Citadel* is aimed at the registration of batch and serial numbers at component level will be registered. The structured and comprehensive registration holds a potential to increase the quality of the entire chain, improve delivery security and reduce costs over the entire lifespan. The collaborative pipe failure registration platform USTORE (Moerman and Beuken 2015) and information architecture UKNOW (Moerman, Van Vossen et al. 2016) are examples of registering pipe failures and asset information that allow for detailed statistical analysis with the aim of optimized pipe replacement. It can be expected that environmental data availability and quality will continue to grow and become more fine-grained, for example through high-resolution data from satellites and drones. The use of environmental data could further improve asset management strategies, e.g. by using soil, weather and traffic data for improved pipe stress calculations and degradation predictions. (The use of environmental data is however not limited to asset management.)

Analysis of social media content (such as Twitter, Facebook, and Google Trends) aggregated through mining approaches can broaden the types of information available to water utilities and water boards, including sentiments among their customers that indicate issues that are important to them.

Finally, graph databases may be useful for water distribution networks because of the similarity of the spatial nature of the data with the setup of such databases. Topological relations between sensor measurements can be easily reflected in the structure of such databases. Graph databases excel in many-to-many, network-type of problems and large volume, large variety data sets. With the addition of smart meters and the subsequent amount of data, graph databases need to be considered in more detail as a basis for combining data from multiple sensors in a meaningful way (Rose 2015, Creaco et al., 2016).

5.3.3 Chemical water quality

Through the continuous exploration of existing and new datasets and their integration, chemical water quality data can be used not only to analyze the status quo but also predict the future impact of chemicals on drinking water quality, human and environmental health. This will help identify vulnerabilities and data gaps that need additional attention or protection, and ultimately be more predictive and responsive. In the field of chemical water quality, data mining should focus on its application to the fields of risk-based monitoring, wastewater based epidemiology and identification of unknowns from non-target screening data, including transformation products, identification of nano- and microplastics, and the modeling of chemical exposure.

5.3.4 Biological water quality

Biological water quality management is increasingly benefiting from the opportunities offered by data mining. A rapidly expanding field is that of genomics (analysis of genome structure and expression), where high-throughput technologies such as Next Generation Sequencing (NGS) have accelerated the availability of DNA sequence information and methods to process and analyse this genetic information. In biotechnology it has become common practice to apply data mining to relate information present in genomic databases worldwide to functional properties (Huttenhower et al., 2008, Kennedy et al., 2010, Ju and Zhang, 2015a,b) such as the production of desirable chemicals (pharmaceuticals, Lee et al., 2008), to find novel pollutant degradation pathways or to relate identified genes to desired or undesired functions (e.g., antibiotic resistance genes, Liu). Current applications within biological water quality management are mainly on species identification based on small (amplified) DNA sequences (amplicon sequencing/metabarcoding) and on the genes and their expression (metagenomics and metatranscriptomics). For a review of NGS applications for biological water quality assessment we refer to Tan et al. (2015), with examples on identification of specific micro-organisms and their genes for e.g. disease risk estimation, presence of potentially toxic micro-organisms or antibiotic resistance. In genomics, the increase in analytical capacity exceeds that of Moore’s law, meaning that the genome sequencing capacity increases at a faster rate than that of microprocessor computing capacity. This has already been addressed as a challenge for the coming years, otherwise the costs of data analysis will significantly exceed the costs of analysis. New (decentralized) computing solutions also have their drawbacks, and in the end the users should always be able to make informed decisions on analysis tools and understanding the analytical results (McPherson, 2009). This urges for novel approaches to handle these data in terms of hardware, but also in data mining tools that are able to combine and extract relevant information from multiple sources in an efficient manner.

Biological water quality has traditionally relied on data mining approaches for risk management. One of the most widely applied data mining frameworks is that of quantitative microbial risk assessment (QMRA), a preventive, risk-based approach to water quality management. QMRA uses a single assessment of the risk of waterborne infectious disease transmission and has developed as a scientific discipline over the last two decades. QMRA has become embedded in the water-related guidelines of the World Health Organization (WHO, 2016), and is at the basis of many water safety plans (WSPs) and sanitation safety plans (SSPs) worldwide. Specific biological risks are also being managed using data mining, for example to predict the occurrence of cyanobacterial toxins or the risk of faecal infection.

Another class of applications is where data mining is used to monitor or predict the response of the ecosystem or parts thereof to environmental changes. This application is not widely applied yet, but has good potential since it gives insight in the typically complex and nonlinear response of these systems.

Novel applications often arise at the intersection of scientific disciplines, and this also holds for biological water quality. In a multidisciplinary approach, often novel data mining techniques are adopted or combined, and/or different types of data are being related. This may lead to unexpected results, requiring careful interpretation and judgement. Nevertheless, this approach may identify patterns otherwise undetected. Examples of these are the prediction of river flow rates from genomics data (Good et al., 2018), understanding climate change effects by combining measured data and literature data (Semenza et al., 2012), management of invasions by combining traffic (ship movements) with species data (Xu et al., 2014), predicting fish communities from combining species data with data on land use and hydrology (He et al., 2010) or predicting cyanotoxin risks from near-real time remote sensing data and biological data (Chang et al., 2014). From social sciences, another type of analysis is possible, namely monitoring the response of the human system to changes in biological water quality, which can give more insight into the response of both water managers (Harvey et al., 2015) as well as the end user/general pubic (Cha & Stow, 2015). Examples of such cross-disciplinary approaches are rare, but have great potential. One obvious application is infection and disease management outside of hospitals: data from disease surveillance systems and hospital infection control data can be linked to (genomic) data on pathogen occurrence in drinking water distribution and sewage systems, and analyzed using spatiotemporal analysis techniques. This might give better insight in infection dynamics, disease spreading or hotspots of antimicrobial resistance. Another area is in ecological risks, where managers often have to deal with the response of a very complex ecosystem to disturbance events. Here, a stepwise approach analogous to the QMRA framework (termed QERA, see e.g. Bayliss et al., 2012) might help in managing these risks.

6 Conclusions and recommendations

6.1 Conclusions

In the methodological sense, data mining or more precisely knowledge discovery from databases is a mature field which offers many fully developed methods with a plethora of reference applications. In the specific water cycle management domain, numerous applications in both an academic and operational context are available internationally. From this perspective, there is no immediate need for KWR and the BTO utilities to put more effort in the development of (completely) new methods, but rather in the implementation and customization of existing methods. Both the datasets and the applications are readily identifiable, presenting opportunities.

A number of successful applications have been reported also by Dutch utilities. However, practitioners indicate a number of obstacles:

- data ownership and access;
- availability of good data analysts;
- availability and quality of data;
- organizational dynamics/culture.

They also stress the importance of knowledge both on data analytics methods and the application domain. No meaningful additional insights are to be expected if one is lacking.

6.2 Recommendations to the water sector

We have seen that methods, data and domain questions are there – with more emerging constantly. For the water utilities, our recommendations focus on resolving the barriers which have been identified and which are within their sphere of influence. These include data ownership and access, availability and quality of data, and organizational dynamics/culture. In this study, the root causes of these barriers have not been considered, but this would be a first step in resolving them. Organizations outside the water sector have taken steps and set up frameworks that address similar issues. A good example is Rijkswaterstaat, which presented its framework in one of the meetings of the Hydroinformatics Platform(e.g. Kisjes, 2016). We recommend that their approach be considered as a starting point for data consolidation, quality control and data sharing across the water distribution industry.

6.3 Recommendations to KWR

This report has been written as a deliverable of the first phase of the BTO exploratory research project *VO datamining*. Based on the conclusions of the first phase, as described above, we recommend that the following phases of the project focus on the actual implementation of a number of data mining cases with BTO utilities. In doing so, we no longer aim for methodological exploration and innovation, but rather for innovation in the application. Important research questions to be answered include practical issues related to streamlining of the complete chain from data acquisition through quality assurance and data mining to decision (support). At a higher abstraction level, they also include questions on how to organize a successful implementation of data mining techniques – ideally also defining a ‘template’ approach. In order to identify the barriers and success factors, we recommend that social scientists also be included in the second phase of the project to identify broader organizational root causes for barriers to and success factors in the implementation of data mining techniques within water utilities.

7 References

Adamowski, J. and C. Karapataki (2010). "Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms." Journal of Hydrologic Engineering **15**(10): 729-743.

Altenburger, R., M. Nendza and G. Schüürmann (2003). "Mixture toxicity and its modeling by quantitative structure-activity relationships." Environmental Toxicology and Chemistry **22**(8): 1900-1915.

Babovic, V., J.-P. Drécourt, M. Keijzer and P. Friss Hansen (2002). "A data mining approach to modelling of water supply assets." Urban Water **4**(4): 401-414.

Bade, R., A. Causanilles, E. Emke, L. Bijlsma, J. V. Sancho, F. Hernandez and P. de Voogt (2016). "Facilitating high resolution mass spectrometry data processing for screening of environmental water samples: An evaluation of two deconvolution tools." Sci Total Environ **569-570**: 434-441.

Bae, H., S. Kim and Y. J. Kim (2006). "Decision algorithm based on data mining for coagulant type and dosage in water treatment systems." Water Science and Technology **53**(4-5): 321.

Baken (2018). Tools for human health risk evaluation of emerging chemicals. Nieuwegein, The Netherlands, KWR Watercycle Research Institute.

Bayliss, P., R.A. van Dam and R. E. Bartolo (2012) "Quantitative Ecological Risk Assessment of the Magela Creek Floodplain in Kakadu National Park, Australia: Comparing Point Source Risks from the Ranger Uranium Mine to Diffuse Landscape-Scale Risks." Human and Ecological Risk Assessment: An International Journal 18:115-151.

Berardi, L., Z. Kapelan, O. Giustolisi and D. A. Savic (2008). "Development of pipe deterioration models for water distribution systems using EPR." Journal of Hydroinformatics **10**(3): 265.

Bernhardt, E. S., E. J. Rosi and M. O. Gessner (2017). "Synthetic chemicals as agents of global change." Frontiers in Ecology and the Environment **15**(2): 84-90.

Berry, J. K., J. A. Detgado, R. Khosla and F. J. Pierce (2003). "Precision conservation for environmental sustainability." Journal of Soil and Water Conservation **58**(6): 332-339.

Bhatia, S., T. Schultz, D. Roberts, J. Shen, L. Kromidas and A. Marie Api (2015). "Comparison of Cramer classification between Toxtree, the OECD QSAR Toolbox and expert judgment." Regulatory Toxicology and Pharmacology **71**(1): 52-62.

Bichler, A., Neumaier, A., & Hofmann, T. (2014). A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. *Journal of Hydrology, 519*(PA), 909-917. doi:10.1016/j.jhydrol.2014.08.013

Bieroza, M., A. Baker and J. Bridgeman (2012). "New data mining and calibration approaches to the assessment of water treatment efficiency." Advances in Engineering Software **44**(1): 126-135.

Blackwell, B. R., G. T. Ankley, S. R. Corsi, L. A. DeCicco, K. A. Houck, R. S. Judson, S. Li, M. T. Martin, E. Murphy, A. L. Schroeder, E. R. Smith, J. Swintek and D. L. Villeneuve (2017). "An "EAR" on Environmental Surveillance and Monitoring: A Case Study on the Use of Exposure-Activity Ratios (EARs) to Prioritize Sites, Chemicals, and Bioactivities of Concern in Great Lakes Waters." Environ Sci Technol **51**(15): 8713-8724.

Bletsou, A. A., J. Jeon, J. Hollender, E. Archontaki and N. S. Thomaidis (2015). "Targeted and non-targeted liquid chromatography-mass spectrometric workflows for identification of transformation products of emerging pollutants in the aquatic environment." TrAC - Trends in Analytical Chemistry **66**: 32-44.

Bonchev, D. and D. H. Rouvray (1992). Chemical Graph Theory: Introduction and Fundamentals, Gordon and Breach Science Publishers.

Brack, W., S. Ait-Aissa, R. M. Burgess, W. Busch, N. Creusot, C. Di Paolo, B. I. Escher, L. Mark Hewitt, K. Hilscherova, J. Hollender, H. Hollert, W. Jonker, J. Kool, M. Lamoree, M. Muschket, S. Neumann, P. Rostkowski, C. Ruttkies, J. Schollee, E. L. Schymanski, T. Schulze, T. B. Seiler, A. J. Tindall, G. De Aragao Umbuzeiro, B. Vrana and M. Krauss (2016). "Effect-directed analysis supporting monitoring of aquatic environments--An in-depth overview." Sci Total Environ **544**: 1073-1118.

Brown, F.K. (1998). Chemoinformatics: What is it and How does it Impact Drug Discovery. Annual Reports in Med. Chem. Annual Reports in Medicinal Chemistry. 33: 375. doi:10.1016/S0065-7743(08)61100-8.

Brunner, A. M., M. L. Dingemans, K. A. Baken and A. P. van Wezel (submitted). "Prioritizing anthropogenic chemicals in drinking water and sources through combined use of mass spectrometry and ToxCast toxicity data."

Candelieri, A., D. Soldi and F. Archetti (2015). "Short-term forecasting of hourly water consumption by using automatic metering readers data." Procedia Engineering **119**: 844-853.

Causanilles, A., J. A. Baz-Lomba, D. A. Burgard, E. Emke, I. Gonzalez-Marino, I. Krizman-Matasic, A. Li, A. S. C. Love, A. K. McCall, R. Montes, A. L. N. van Nuijs, C. Ort, J. B. Quintana, I. Senta, S. Terzic, F. Hernandez, P. de Voogt and L. Bijlsma (2017). "Improving wastewater-based epidemiology to estimate cannabis use: focus on the initial aspects of the analytical procedure." Anal Chim Acta **988**: 27-33.

Causanilles, A., J. Kinyua, C. Ruttkies, A. L. N. van Nuijs, E. Emke, A. Covaci and P. de Voogt (2017). "Qualitative screening for new psychoactive substances in wastewater collected during a city festival using liquid chromatography coupled to high-resolution mass spectrometry." Chemosphere **184**: 1186-1193.

Causanilles, A., V. Nordmann, D. Vughs, E. Emke, O. de Hon, F. Hernandez and P. de Voogt (2018). "Wastewater-based tracing of doping use by the general population and amateur athletes." Anal Bioanal Chem **410**(6): 1793-1803.

Causanilles, A., C. Ruepert, M. Ibanez, E. Emke, F. Hernandez and P. de Voogt (2017). "Occurrence and fate of illicit drugs and pharmaceuticals in wastewater from two wastewater treatment plants in Costa Rica." Sci Total Environ **599-600**: 98-107.

Cha, Y. and C. A. Stow (2015). "Mining web-based data to assess public response to environmental events." Environmental Pollution **198**: 97-99.

Chang, N. -., Vannah, B., & Jeffrey Yang, Y. (2014). Comparative sensor fusion between hyperspectral and multispectral satellite sensors for monitoring microcystin distribution in lake erie. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7*(6), 2426-2442. doi:10.1109/JSTARS.2014.2329913

Chawakitchareon, P., N. Boonnao and P. Charytragulchai (2017). Prediction of Alum Dosage in Water Supply by WEKA Data Mining Software. Frontiers in Artificial Intelligence and Applications. **Volume 292: Information Modelling and Knowledge Bases XXVIII:** 83-93.

Cheifetz, N., Z. Noumir, A. Samé, A.-C. Sandraz, C. Féliers and V. Heim (2017). "Modeling and clustering water demand patterns fom real-world smart mete data." Drinking Water Engineering and Science **10**: 75-82.

Chemlist. (2018). "Regulated Chemicals - CHEMLIST." Retrieved February 21, 2018, from http://www.cas.org/content/regulated-chemicals.

Chen, Q., & Mynett, A. E. (2003). Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu lake. Ecological Modelling, 162(1-2), 55-67. doi:10.1016/S0304-3800(02)00389-7

Chiaia-Hernandez, A. C., E. L. Schymanski, P. Kumar, H. P. Singer and J. Hollender (2014). "Suspect and nontarget screening approaches to identify organic contaminant records in lake sediments." Anal Bioanal Chem **406**(28): 7323-7335.

Cirkel, D. G., Y. Fujita, R. P. Bartholomeus and J. P. M. Witte (2016). Inbouw van bodemnutriënten en zuurgraad in PROBE, KWR Watercycle Research Institute**:** 35.

Cirkel, D. G., Y. Fujita and J. Rozemeijer (2017). Kwantificeren van nutriëntenuitspoeling van sensoren**:** 33.

Comas, J., S. Dzeroski, K. Gibert, I. R.-Roda and M. Sanchez-Marre (2001). "Knowledge discovery by means of inductive methods in wastewater treatment plant data." AI Communications **14**(1): 45-62.

Comber, S. D. W., R. Smith, P. Daldorph, M. J. Gardner, C. Constantino and B. Ellor (2018). "Development of a chemical source apportionment decision support framework for lake catchment management." Science of The Total Environment **622-623**: 96-105.

Cortez, P. and A. Morais (2007). A Data Mining Approach to Predict Forest Fires using Meteorological Data. Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, Guimaraes, Portugal.

Creaco, E. P. Kossieris, L. Vamvakeridou-Lyroudia, C. Makropoulos, Z. Kapelan, D. Savic (2016). "Parameterizing residential water demand pulse models through smart meter readings." Environmental Modelling and Software. Environmental Modelling & Software 80: 33-40

Dawsey, W. J., & Minsker, B. S. (2007). Data mining to inform total coliform monitoring plan design. *8th Annual Water Distribution Systems Analysis Symposium 2006,* , 158. doi:10.1061/40941(247)158

Deng, L. and Yu, D. (2014) Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing. 7 (3–4): 1–199. doi:10.1561/2000000039.

De Corte, A. and K. Sörensen (2014). "HydroGen: an Artificial Water Distribution Network Generator." Water Resources Management **28**(2): 333-350.

Dimitrov, S. D., R. Diderich, T. Sobanski, T. S. Pavlov, G. V. Chankov, A. S. Chapkanov, Y. H. Karakolev, S. G. Temelkov, R. A. Vasilev, K. D. Gerova, C. D. Kuseva, N. D. Todorova, A. M. Mehmed, M. Rasenberg and O. G. Mekenyan (2016). "QSAR Toolbox – workflow and major functionalities." SAR and QSAR in Environmental Research **27**(3): 203-219.

Domingos, P. (2012). "A few useful things to know about machine learning." Commun. ACM **55**(10): 78-87.

Dorland, E., Y. Fujita, W. Chardon and A. K. de Jong (2017). Toepassing van drinkwaterslib op fosfaatrijke gronden t.b.v. natuurontwikkeling**:** 110.

Dürrenmatt, D. J. and W. Gujer (2012). "Data-driven modeling approaches to support wastewater treatment plant operation." Environmental Modelling & Software **30**: 47-56.

Engel, T. (2006). "Basic overview of chemoinformatics." J Chem Inf Model **46**(6): 2267-2277.

Ertl, P., J. Muhlbacher, B. Rohde and P. Selzer (2003). "Web-based cheminformatics and molecular property prediction tools supporting drug design and development at Novartis." SAR QSAR Environ Res **14**(5-6): 321-328.

Escher, B. I. and K. Fenner (2011). "Recent advances in environmental risk assessment of transformation products." Environmental Science and Technology **45**(9): 3835-3847.

Fenner, K. (2016). Transformation product analysis: Ready to go beyond suspect screening? Non-target screening of organic chemicals for a comprehensive environmental risk assessment, Conference Centre Monte Verità, Ascona, Switzerland.

Flach, P. (2012). Machine Learning - The art and science of algorithms that make sense of data. Cambridge, Cambridge University Press.

Fujita, Y. and C. Aggenbach (2015). Effects of mowing, sod-cutting, and drift sand on development of soil and vegetation in Grey Dunes, KWR**:** 126.

Fujita, Y., R. P. Bartholomeus and J. P. M. Witte (2016). PROBE-3: A succession model for ecosystem services, KWR Watercycle Research Institute**:** 43.

Fujita, Y., P. M. van Bodegom and J.-P. M. Witte (2013). "Relationships between Nutrient-Related Plant Traits and Combinations of Soil N and P Fertility Measures." PLoS ONE **8**(12): e83735.

Gago-Ferrero, P., E. L. Schymanski, A. A. Bletsou, R. Aalizadeh, J. Hollender and N. S. Thomaidis (2015). "Extended Suspect and Non-Target Strategies to Characterize Emerging Polar Organic Contaminants in Raw Wastewater with LC-HRMS/MS." Environ Sci Technol **49**(20): 12333-12341.

Garcia Nieto, P. J., Sánchez Lasheras, F., de Cos Juez, F. J., & Alonso Fernández, J. R. (2011). Study of cyanotoxins presence from experimental cyanobacteria concentrations using a new data mining methodology based on multivariate adaptive regression splines in Trasona reservoir (Northern Spain). *Journal of Hazardous Materials, 195*, 414-421. doi:10.1016/j.jhazmat.2011.08.061

Ghiassi, M., D. K. Zimbra and H. Saidane (2008). "Urban Water Demand Forecasting with a Dynamic Artificial Neural Network Model." Journal of Water Resources Planning and Management **134**(2): 138-146.

Giustolisi, O., D. A. Savic and D. Laucelli (2004). "Data Mining for Management and Rehabilitation of Water Systems: The Evolutionary Polynomial Regression Approach." Wasserbauliche Mitteilungen **27**.

Good, S. P., URycki, D. R., & Crump, B. C. (2018). Predicting hydrologic function with aquatic gene fragments. Water Resources Research, 54, 2424–2435. doi:10.1002/2017WR021974

Grimme, S. and P. R. Schreiner (2017). "Computational Chemistry: The Fate of Current Methods and Future Challenges." Angew Chem Int Ed Engl.

Guha, N., K. Z. Guyton, D. Loomis and D. K. Barupal (2016). "Prioritizing chemicals for risk assessment using chemoinformatics: Examples from the IARC monographs on pesticides." Environmental Health Perspectives **124**(12): 1823-1829.

Harvey, R., Murphy, H.M., McBean, E.A. & Gharabaghi, B. (2015) Using Data Mining to Understand Drinking Water Advisories in Small Water Systems: a Case Study of Ontario First Nations Drinking Water Supplies. Water Resources Management 29(14), 5129–5139. doi:10.1007/s11269-015-1108-6

He, Y., Wang, J., Lek-Ang, S., & Lek, S. (2010). Predicting assemblages and species richness of endemic fish in the upper Yangtze river. Science of the Total Environment, 408(19), 4211-4220. doi:10.1016/j.scitotenv.2010.04.052

Hogenboom, A. C., J. A. van Leerdam and P. de Voogt (2009). "Accurate mass screening and identification of emerging contaminants in environmental samples by liquid chromatography-hybrid linear ion trap Orbitrap mass spectrometry." Journal of Chromatography A **1216**(3): 510-519.

Hoh, E., N. G. Dodder, S. J. Lehotay, K. C. Pangallo, C. M. Reddy and K. A. Maruya (2012). "Nontargeted comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry method and software for inventorying persistent and bioaccumulative contaminants in marine environments." Environ Sci Technol **46**(15): 8001-8008.

Hollender, J., E. L. Schymanski, H. P. Singer and P. L. Ferguson (2017). "Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go?" Environmental Science & Technology **51**(20): 11505-11512.

Huang, J. J. and E. A. McBean (2009). "Data Mining to Identify Contaminant Event Locations in Water Distribution Systems." Journal of Water Resources Planning and Management **135**(6): 466-474.

Hug, C., N. Ulrich, T. Schulze, W. Brack and M. Krauss (2014). "Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening." Environmental Pollution **184**: 25-32.

Huttenhower C, Hofmann O (2010) A Quick Guide to Large-Scale Genomic Data Mining. PLoS Comput Biol 6(5): e1000779. doi:10.1371/journal.pcbi.1000779

Ivanciuc, O. (2013). "Chemical graphs, molecular matrices and topological indices in chemoinformatics and quantitative structure-activity relationships." Curr Comput Aided Drug Des **9**(2): 153-163.

Izquierdo, J., I. Montalvo, R. Pérez-García and E. Campbell (2014). "Mining Solution Spaces for Decision Making in Water Distribution Systems." Procedia Engineering **70**: 864-871.

Jolly Matthew, D., D. Lothes Amanda, L. Sebastian Bryson and L. Ormsbee (2014). "Research Database of Water Distribution System Models." Journal of Water Resources Planning and Management **140**(4): 410-416.

Ju, F., & Zhang, T. (2015a). 16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions. *Applied Microbiology and Biotechnology, 99*(10), 4119-4129. doi:10.1007/s00253-015-6536-y

Ju, F., & Zhang, T. (2015b). Experimental design and bioinformatics analysis for the application of metagenomics in environmental sciences and biotechnology. *Environmental Science and Technology, 49*(21), 12628-12640. doi:10.1021/acs.est.5b03719

Judson, R., K. Houck, M. Martin, T. Knudsen, R. S. Thomas, N. Sipes, I. Shah, J. Wambaugh and K. Crofton (2014). "In vitro and modelling approaches to risk assessment from the U.S. environmental protection agency ToxCast programme." Basic and Clinical Pharmacology and Toxicology **115**(1): 69-76.

Judson, R., A. Richard, D. J. Dix, K. Houck, M. Martin, R. Kavlock, V. Dellarco, T. Henry, T. Holderman, P. Sayre, S. Tan, T. Carpenter and E. Smith (2009). "The toxicity data landscape for environmental chemicals." Environ Health Perspect **117**(5): 685-695.

Kar, S. and K. Roy (2010). "QSAR modeling of toxicity of diverse organic chemicals to Daphnia magna using 2D and 3D descriptors." J Hazard Mater **177**(1-3): 344-351.

Kennedy, J., Flemer, B., Jackson, S. A., Lejon, D. P. H., Morrissey, J. P., O'Gara, F., & Dobson, A. D. W. (2010). Marine metagenomics: New tools for the study and exploitation of marine microbial metabolism. Marine Drugs, 8(3), 608-628. doi:10.3390/md8030608.

Khan, K. and K. Roy (2017). "Ecotoxicological modelling of cosmetics for aquatic organisms: A QSTR approach." SAR QSAR Environ Res **28**(7): 567-594.

Kisjes, K. (2016) "Grip op datakwaliteit van AAT tot Z", presentation file: https://waterinfodag.nl/wp-content/uploads/2016/01/20180329-Handout-Waterinfodag-Kasper-Kisjes.pdf, retrieved October 3, 2018.

Krauss, M., H. Singer and J. Hollender (2010). "LC-high resolution MS in environmental analysis: From target screening to the identification of unknowns." Analytical and Bioanalytical Chemistry **397**(3): 943-951.

Kusiak, A., Y. Zeng and Z. Zhang (2013). "Modeling and analysis of pumps in a wastewater treatment plant: A data-mining approach." Engineering Applications of Artificial Intelligence **26**(7): 1643-1651.

Lamovec P., Matjaž M. and O. K. (2013). "Detection of flooded areas using machine learning techniques : case study of the Ljubljana moor floods in 2010." Disaster Advances **6(7)**(July).

Lee, J. K., Williams, P. D., & Cheon, S. (2008). Data Mining in Genomics. *Clinics in Laboratory Medicine*, *28*(1), 145–viii. Doi:10.1016/j.cll.2007.10.010

Liu, B., & Pop, M. (2009). ARDB - antibiotic resistance genes database. Nucleic Acids Research, 37(SUPPL. 1), D443-D447. doi:10.1093/nar/gkn656

Loos, R., R. Carvalho, D. C. António, S. Comero, G. Locoro, S. Tavazzi, B. Paracchini, M. Ghiani, T. Lettieri, L. Blaha, B. Jarosova, S. Voorspoels, K. Servaes, P. Haglund, J. Fick, R. H. Lindberg, D. Schwesig and B. M. Gawlik (2013). "EU-wide monitoring survey on emerging polar organic contaminants in wastewater treatment plant effluents." Water Research **47**(17): 6475-6487.

Loos, R., B. M. Gawlik, G. Locoro, E. Rimaviciute, S. Contini and G. Bidoglio (2009). "EU-wide survey of polar organic persistent pollutants in European river waters." Environmental Pollution **157**(2): 561-568.

Loos, R., G. Locoro, S. Comero, S. Contini, D. Schwesig, F. Werres, P. Balsaa, O. Gans, S. Weiss, L. Blaha, M. Bolchi and B. M. Gawlik (2010). "Pan-European survey on the occurrence of selected polar organic persistent pollutants in ground water." Water Research **44**(14): 4115-4126.

Luscombe, N.M., D. Greenbaum, M. Gerstein (2001) "What is bioinformatics? A proposed definition and overview of the field." Methods Inf Med, 40(4): 346-58.

Mackay, D. and S. Paterson (1991). "Evaluating the multimedia fate of organic chemicals: A level III fugacity model." Environmental Science and Technology **25**(3): 427-436.

Mackay, D., W. Y. Shiu and K. Ma (1992). Illustrated Handbook of physical-chemical properties and environmental fate for organic chemicals. Boca Raton, FL, USA, Lewis Publishers.

March, H., Á.-F. Morote, A.-M. Rico and D. Sauri (2017). "Household smart water metering in Spain: Insights from the experience of remote meter reading in Alicant." sustainability **9**(582).

Mashford, J., D. D. Silva, D. Marney and S. Burn (2009). An Approach to Leak Detection in Pipe Networks Using Analysis of Monitored Pressure Values by Support Vector Machine. 2009 Third International Conference on Network and System Security.

McPherson, J.D. (2009) Next-generation gap. Nature Methods 6: S2–S5. doi:10.1038/nmeth.f.268

Merel, S., S. Lege, J. E. Yanez Heras and C. Zwiener (2017). "Assessment of N-Oxide Formation during Wastewater Ozonation." Environ Sci Technol **51**(1): 410-417.

Mintenig, S. M., I. Int-Veen, M. G. J. Löder, S. Primpke and G. Gerdts (2017). "Identification of microplastic in effluents of waste water treatment plants using focal plane array-based micro-Fourier-transform infrared imaging." Water Research **108**: 365-372.

Moerman, A. and R. Beuken (2015). "USTORE, hét kennisinstrument voor het onderbouwen van vervangingsbeslissingen van waterleidingen." H2O

Moerman, A., J. Van Vossen and R. Beuken (2016). UKNOW; zicht op leidingdegradatie door samenhang in informatiesystemen, KWR Watercycle Research Institute.

Mondy, C.P., Muñoz, I. & Dolédec, S. (2016). Life-history strategies constrain invertebrate community tolerance to multiple stressors: A case study in the Ebro basin. *Science of The Total Environment* 572, 196-206. doi:10.1016/j.scitotenv.2016.07.227

Moschet, C., A. Piazzoli, H. Singer and J. Hollender (2013). "Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry." Anal Chem **85**(21): 10312-10320.

Mounce, S. R., E. J. M. Blokker, S. P. Husband, W. R. Furnass, P. G. Schaap and J. B. Boxall (2016). "Multivariate data mining for estimating the rate of discolouration material accumulation in drinking water distribution systems." IWA Journal of Hydroinformatics **18**(1): 96-114.

Mounce, S. R., S. P. Husband, W. R. Furnass and J. B. Boxall (2014). "Multivariate Data Mining for Estimating the Rate of Discoloration Material Accumulation in Drinking Water Systems." Procedia Engineering **89**: 173-180.

Mounce, S. R., R. B. Mounce and J. B. Boxall (2015). "Case-based reasoning to suport decision making for managing drinking water quality events in distribution systems." Urban Water Journal **13**(7): 727-738.

Muz, M., N. Ost, R. Kuhne, G. Schuurmann, W. Brack and M. Krauss (2017). "Nontargeted detection and identification of (aromatic) amines in environmental samples based on diagnostic derivatization and LC-high resolution mass spectrometry." Chemosphere **166**: 300-310.

Noymanee, J., N. O. Nikitin and A. V. Kalyuzhnaya (2017). "Urban Pluvial Flood Forecasting using Open Data with Machine Learning Techniques in Pattani Basin." Procedia Computer Science **119**: 288-297.

Ordoñez, J. C., P. M. Van Bodegom, J. P. M. Witte, R. P. Bartholomeus, H. F. Van Dobben and R. Aerts (2010). "Leaf habit and woodiness regulate different leaf economy traits at a given nutrient supply." Ecology **91**(11): 3218-3228.

Ordonez, J. C., P. M. van Bodegom, J. P. M. Witte, R. P. Bartholomeus, J. R. van Hal and R. Aerts (2010). "Plant Strategies in Relation to Resource Supply in Mesic to Wet Environments: Does Theory Mirror Nature?" American Naturalist **175**(2): 225-239.

Pereira, G. C., Figueiredo, A. R., & Ebecken, N. F. F. (2009). Mining for ecological thresholds and associations in cytometric data: A coastal management perspective. Paper presented at the WIT Transactions on Information and Communication Technologies, , 42 85-92. doi:10.2495/DATA090091

Ponce Romero, J., S. Hallett and S. Jude (2017). "Leveraging Big Data Tools and Technologies: Addressing the Challenges of the Water Quality Sector." Sustainability **9**(12): 2160.

Pramanik, S. and K. Roy (2013). "Environmental toxicological fate prediction of diverse organic chemicals based on steady-state compartmental chemical mass ratio using quantitative structure-fate relationship (QSFR) models." Chemosphere **92**(5): 600-607.

Pramanik, S. and K. Roy (2014). "Modeling bioconcentration factor (BCF) using mechanistically interpretable descriptors computed from open source tool "PaDEL-Descriptor"." Environ Sci Pollut Res Int **21**(4): 2955-2965.

Pyayt, A. L., I. I. Mokhov, B. Lang, V. V. Krzhizhanovskaya and R. J. Meijer (2011). "Machine learning methods for environmental monitoring and flood protection." World Academy of Science, Engineering and Technology **78**: 118-123.

Romano, G., Costantini, M., Sansone, C., Lauritano, C., Ruocco, N., & Ianora, A. (2017). Marine microorganisms as a promising and sustainable source of bioactive molecules. Marine Environmental Research, 128, 58-69. doi:10.1016/j.marenvres.2016.05.002

Rose, A. (2015). "On the uses of graph databases in water distribution systems." Retrieved 14 march, 2018, from https://www.icwmm.org/Archive/2015-C024-20/on-the-uses-of-graph-databases-in-water-distribution-systems.

Ruff, M., M. S. Mueller, M. Loos and H. P. Singer (2015). "Quantitative target and systematic non-target analysis of polar organic micro-pollutants along the river Rhine using high-resolution mass-spectrometry--Identification of unknown sources and compounds." Water Res **87**: 145-154.

Sabljić , A., H. Güsten, H. Verhaar and J. Hermens (1995). "QSAR modelling of soil sorption. Improvements and systematics of log KOC vs. log KOW correlations." Chemosphere **31**(11-12): 4489-4514.

Sakr, G. E., I. H. Elhajj, G. Mitri and U. C. Wejinya (2010). Artificial intelligence for forest fire prediction. 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics.

Scheringer, M. (2009). "Long-range transport of organic chemicals in the environment." Environmental Toxicology and Chemistry **28**(4): 677-690.

Scheringer, M., K. C. Jones, M. Matthies, S. Simonich and D. Van De Meent (2009). "Multimedia partitioning, overall persistence, and long-range transport potential in the context of pops and pbt chemical assessments." Integrated Environmental Assessment and Management **5**(4): 557-576.

Schollée, J. E., E. L. Schymanski and J. Hollender (2016). Statistical Approaches for LC-HRMS Data To Characterize, Prioritize, and Identify Transformation Products from Water Treatment Processes. Assessing Transformation Products of Chemicals by Non-Target and Suspect Screening − Strategies and Workflows Volume 1, American Chemical Society. **1241:** 45-65.

Schollee, J. E., E. L. Schymanski, M. A. Stravs, R. Gulde, N. S. Thomaidis and J. Hollender (2017). "Similarity of High-Resolution Tandem Mass Spectrometry Spectra of Structurally Related Micropollutants and Transformation Products." J Am Soc Mass Spectrom **28**(12): 2692-2704.

Schulze, S., D. Sättler, M. Neumann, H. P. H. Arp, T. Reemtsma and U. Berger (2018). "Using REACH registration data to rank the environmental emission potential of persistent and mobile organic chemicals." Science of the Total Environment **625**: 1122-1128.

Schymanski, E. L., J. Jeon, R. Gulde, K. Fenner, M. Ruff, H. P. Singer and J. Hollender (2014). "Identifying small molecules via high resolution mass spectrometry: Communicating confidence." Environmental Science and Technology **48**(4): 2097-2098.

Schymanski, E. L., H. P. Singer, P. Longree, M. Loos, M. Ruff and M. A. Stravs (2014). "Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry." Environ Sci Technol **48**.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology, 12*(6) doi:10.1186/gb-2011-12-6-r60

Semenza, J. C., Herbst, S., Rechenburg, A., Suk, J. E., Höser, C., Schreiber, C., & Kistemann, T. (2012). Climate change impact assessment of food- and waterborne diseases. Critical Reviews in Environmental Science and Technology, 42(8), 857-890. doi:10.1080/10643389.2010.534706.

Shah, I., J. Liu, R. S. Judson, R. S. Thomas and G. Patlewicz (2016). "Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information." Regulatory Toxicology and Pharmacology **79**: 12-24.

Singh, P. and P. D. Kaur (2017). "Review on Data Mining Techniques for Prediction of Water Quality." International Journal of Advanced Research in Computer Science **8**(5).

Sjerps, R. M., D. Vughs, J. A. van Leerdam, T. L. ter Laak and A. P. van Wezel (2016). "Data-driven prioritization of chemicals for various water types using suspect screening LC-HRMS." Water Res **93**: 254-264.

Sjerps, R. M. A., A. Brunner, Y. Fujita, B. Bajema, M. de Jonge, P. Bauerlein, J. de Munk and A. P. van Wezel (in preparation). "Target and suspect chemical screening combined with clustering and prioritisation techniques to design a risk based monitoring program in groundwater supply zones for drinking water.".

Sjerps, R. M. A., T. ter Laak and A. van Wezel (2014). Prioriteren van stoffen voor de (drink)waterketen. Nieuwegein KWR**:** 65.

Sjerps, R. M. A., D. Vughs, J. A. van Leerdam, T. L. ter Laak and A. P. van Wezel (2016). "Data-driven prioritization of chemicals for various water types using suspect screening LC-HRMS." Water Research **93**: 254-264.

Soldevila, A., J. Blesa, S. Tornil-Sin, E. Duviella, R. M. Fernandez-Canti and V. Puig (2016). "Leak localization in water distribution networks using a mixed model-based/data-driven approach." Control Engineering Practice **55**: 162-173.

Stewart, R.A., Nguyen, K., Beal, C., Zhang, H., Sahin, O., Bertone, E., Silva Vieira, A., Castelletti, A., Cominola, A., Giuliani, M., Giurco, D., Blumenstein, M., Turner, A., Liu, A., Kenway, S., Savić,, D.A., Makropoulos, C., and Kossieris, P. (2018) "Integrated intelligent water-energy metering systems and informatics: Visioning a digital multi-utility service provider." Environmental Modelling & Software105: 94-117.

Tan B, Ng C, Nshimyimana JP, Loh LL, Gin KY-H and Thompson JR (2015) Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. Front. Microbiol. 6:1027. doi: 10.3389/fmicb.2015.01027

Tebes-Stevens, C., J. M. Patel, M. Koopmans, J. Olmstead, S. H. Hilal, N. Pope, E. J. Weber and K. Wolfe (2018). "Demonstration of a consensus approach for the calculation of physicochemical properties required for environmental fate assessments." Chemosphere **194**: 94-106.

ter Laak, T. L., L. M. Puijker, J. A. van Leerdam, K. J. Raat, A. Kolkman, P. de Voogt and A. P. van Wezel (2012). "Broad target chemical screening approach used as tool for rapid assessment of groundwater quality." Science of the Total Environment **427-428**: 308-313.

Thurman, E. M., I. Ferrer, J. Blotevogel and T. Borch (2014). "Analysis of hydraulic fracturing flowback and produced waters using accurate mass: identification of ethoxylated surfactants." Anal Chem **86**(19): 9653-9661.

Torregrossa, D., J. Hansen, F. Hernández-Sancho, A. Cornelissen, G. Schutz and U. Leopold (2017). Pump Efficiency Analysis of Waste Water Treatment Plants: A Data Mining Approach Using Signal Decomposition for Decision Making.

Torres, J. (2006). MIcropolis: A virtual city for water distribution system research applications. Undergraduate, Texas A&M University.

van Loon, A., R. M. A. Sjerps and K. J. Raat (2017). Gewasbeschrmingsmiddelen en hun afbraakproducten in Nederlandse drinkwaterbronnen. Nieuwegein, KWR**:** 60.

Verma, A., X. Wei and A. Kusiak (2013). "Predicting the total suspended solids in wastewater: A data-mining approach." Engineering Applications of Artificial Intelligence **26**(4): 1366-1372.

von der Ohe, P. C., V. Dulio, J. Slobodnik, E. De Deckere, R. Kühne, R. U. Ebert, A. Ginebreda, W. De Cooman, G. Schüürmann and W. Brack (2011). "A new risk assessment approach for the prioritization of 500 classical and emerging organic microcontaminants as potential river basin specific pollutants under the European Water Framework Directive." Science of the Total Environment **409**(11): 2064-2077.

Vonk, E., D. G. Cirkel and I. Leunk (2017). De gevolgen van klimaatverandering en vakantiespreiding voor de drinkwatervraag. Nieuwegin, KWR Watercycle Research Institute**:** 54.

Vonk, E. and D. Vries (2016). Datamining voor assetmanagement - inventarisatie en voorbeelden uit de watersector, KWR**:** 49.

Vries, D., C. Bertelkamp, F. Schoonenberg Kegel, B. Hofs, J. Dusseldorp, J. H. Bruins, W. de Vet and B. van den Akker (2017). "Iron and manganese removal: Recent advances in modelling treatment efficiency by rapid sand filtration." Water Research **109**: 35-45.

Vries, D., B. A. Wols and P. de Voogt (2013). "Removal efficiency calculated beforehand: QSAR enabled predictions for nanofiltration and advanced oxidation." Water Science & Technology: Water Supply **13**(6): 1425-1436.

Wambaugh, J. F., R. W. Setzer, D. M. Reif, S. Gangwal, J. Mitchell-Blackwood, J. A. Arnot, O. Joliet, A. Frame, J. Rabinowitz, T. B. Knudsen, R. S. Judson, P. Egeghy, D. Vallero and E. A. Cohen Hubal (2013). "High-throughput models for exposure-based chemical prioritization in the ExpoCast project." Environmental Science and Technology **47**(15): 8479-8488.

Wassenaar, T. M. (2004). Risk assessment prediction from genome sequences: Promises and dreams. Journal of Food Protection, 67(9), 2053-2057. doi:10.4315/0362-028X-67.9.2053

Wei, X., A. Kusiak and R. Sadat Hosseini (2013). "Prediction of Influent Flow Rate: Data-Mining Approach." Journal of Energy Engineering **139**(2): 118-123.

Wicker, J., T. Lorsbach, M. Gutlein, E. Schmid, D. Latino, S. Kramer and K. Fenner (2016). "enviPath--The environmental contaminant biotransformation pathway resource." Nucleic Acids Res **44**(D1): D502-508.

Williams, A. J., C. M. Grulke, J. Edwards, A. D. McEachran, K. Mansouri, N. C. Baker, G. Patlewicz, I. Shah, J. F. Wambaugh, R. S. Judson and A. M. Richard (2017). "The CompTox Chemistry Dashboard: A community data resource for environmental chemistry." Journal of Cheminformatics **9**(1).

Witte, J. P. M., R. P. Bartholomeus, D. G. Cirkel, E. Doomernik, Y. Fujita and J. Runhaar (2014). Manual and description of ESTAR, version 01: a software tool to analyse vegetation plots, KWR**:** 27.

Witte, J. P. M., R. P. Bartholomeus, J. C. Douma, H. Runhaar and P. M. Van Bodegom (2010). De vegetatiemodule van Probe-2**:** 49.

Witte, J. P. M., R. P. Bartholomeus, P. M. van Bodegom, D. G. Cirkel, R. van Ek, Y. Fujita, G. M. C. M. Janssen, T. J. Spek and H. Runhaar (2015). "A probabilistic eco-hydrological model to predict the effects of climate change on natural vegetation at a regional scale." Landscape Ecology **30**: 835–854.

Witte, J. P. M., R. B. Wójcik, P. J. J. F. Torfs, M. W. H. De Haan and S. Hennekens (2007). "Bayesian classification of vegetation types with Gaussian mixture density fitting to indicator values." Journal of Vegetation science **18**(4): 605-612.

Wittwehr, C., H. Aladjov, G. Ankley, H. J. Byrne, J. de Knecht, E. Heinzle, G. Klambauer, B. Landesmann, M. Luijten, C. MacKay, G. Maxwell, M. E. Meek, A. Paini, E. Perkins, T. Sobanski, D. Villeneuve, K. M. Waters and M. Whelan (2017). "How Adverse Outcome Pathways Can Aid the Development and Use of Computational Prediction Models for Regulatory Toxicology." Toxicological Sciences **155**(2): 326-336.

Wolfert, S., L. Ge, C. Verdouw and M.-J. Bogaardt (2017). "Big Data in Smart Farming – A review." Agricultural Systems **153**: 69-80.

Wols, B. A. and D. Vries (2012). "On a QSAR approach for the prediction of priority compound degradation by water treatment processes." Water Science and Technology **66**(7): 1446-1453.

World Health Organization (WHO) (2016). Quantitative Microbial Risk Assessment: Application for Water Safety Management. World Health Organization, xiv + 187 pp.

Wu, G.-D. and S.-L. Lo (2008). "Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system." Engineering Applications of Artificial Intelligence **21**(8): 1189-1195.

Wu, Z. Y. and A. Rahman (2017). "Optimized Deep Learning Framework for Water Distribution Data-Driven Modeling." Procedia Engineering **186**: 261-268.

Xu, J., Wickramarathne, T. L., Chawla, N. V., Grey, E. K., Steinhaeuser, K., Keller, R. P., .Drake, J.M. & Lodge, D. M. (2014). Improving management of aquatic invasions by integrating shipping network, ecological, and environmental data: Data mining for social good. Paper presented at the Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1699-1708. doi:10.1145/2623330.2623364.

Zang, Q., K. Mansouri, A. J. Williams, R. S. Judson, D. G. Allen, W. M. Casey and N. C. Kleinstreuer (2017). "In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning." J Chem Inf Model **57**(1): 36-49.

Zanzi, A. and C. Wittwehr (2017). "Searching Online Chemical Data Repositories via the ChemAgora Portal." Journal of Chemical Information and Modeling **57**(12): 2905-2910.

Zarfl, C., M. Scheringer and M. Matthies (2011). "Screening criteria for long-range transport potential of organic substances in water." Environmental Science and Technology **45**(23): 10075-10081.

Zhai, Y., Y. S. Ong and I. W. Tsang (2014). "The Emerging "Big Dimensionality"." IEEE Computational Intelligence Magazine **9**(3): 14-26.

Zhang, S., C. Zhang and Q. Yang (2003). "Data preparation for data mining" Applied Artificial Intelligence 17: 375-381.

Zijp, M. C., L. Posthuma and D. Van De Meent (2014). "Definition and applications of a versatile chemical pollution footprint methodology." Environmental Science and Technology **48**(18): 10588-10597.

Zonja, B., A. Delgado, S. Perez and D. Barcelo (2015). "LC-HRMS suspect screening for detection-based prioritization of iodinated contrast media photodegradates in surface waters." Environ Sci Technol **49**(6): 3464-3472.