BTO 2019.011 | March 2019

BTO report

Data Quality Control

KWR Watercycle Research Institute

BTO

Data Quality Control

BTO 2019.011 | March 2019

Project number
402045-071-001

Project manager
Jos Frijns, MSc

Client
BTO - Thematical research - Drinking water technologies for the future

Quality Assurance
Christos Makropoulos, MEng DIC MSc PhD FRGS FHEA

Author(s)
Mario Castro-Gama, MSc; Claudia Agudelo-Vera, PhD MSc

Redaction(s)
Dimitrios Bouziontas, MSc

Sent to
Participants HI-Platform

More information
MSc, Mario Castro Gama
T +31 30 606 9644
E Mario.Castro.Gama@kwrwater.nl

KWR
PO Box 1072
3430 BB Nieuwegein
The Netherlands

KWR Watercycle Research Institute

BTO *Managementsamenvatting*

Verbeter de beheersing van datakwaliteit door kennis uit te wisselen en proefprojecten uit te voeren

Auteur(s) Mario Castro-Gama, Claudia Agudelo-Vera

Gegevens spelen een sleutelrol bij de besluitvorming en bij het ondersteunen van efficiënte systemen. Veel waterbedrijven erkennen data nu als een belangrijk aspect van de organisatie, dat goed moet worden beheerd. Tegelijkertijd neemt de complexiteit van drinkwatersystemen toe en ontstaan steeds meer data in en voor de operationele omgeving. Ook de datakwaliteit en het vermijden van fouten in datastromen worden steeds belangrijker, ook al wordt er nog niet altijd in de volle breedte naar dit inzicht gehandeld. Gegevensvalidatie is een belangrijk onderdeel van datakwaliteitscontrole. Ondanks de overvloed aan technieken en verschillende operationele cases, zijn er gemeenschappelijke patronen tussen organisaties te zien: waterbedrijven worden grotendeels geconfronteerd met vergelijkbare problemen voor gegevensvalidatie. Samenwerken aan deze gedeelde onderwerpen kan bijdragen aan de implementatie van consistente kaders voor gegevenskwaliteitscontrole die op meerdere niveaus werken. Er is niet één oplossing voor alle behoeften: afhankelijk van het gemonitorde gebeurtenis kunnen verschillende technieken met verschillende complexiteit worden toegepast, ook met betrekking tot de reikwijdte van de validatie.

Belang: gegevens essentieel voor alle beslissingen en modellen bij drinkwaterbedrijven

Gegevens zijn geen 'bijproduct', dit (onder)vinden ook waterbedrijven steeds meer. Gegevens van goede kwaliteit vormen een basis voor goede beslissingen en zijn essentieel voor het creëren van realistische modellen en voor het verbeteren van *evidence-based reporting*. Efficiënt gegevensbeheer is een organisatorische eis voor elke dienst, ook in verband met datagevoelige wetgeving als de EU INSPIRE-richtlijn (gegevensuitwisseling). Hierbij speelt hydro-informatica een belangrijke rol omdat er steeds meer datasets uit het veld komen (van sensoren tot slimme meters) en de rol van data in de beslissingsondersteuning groeit. Voor sommige doeleinden is de huidige kwaliteit van de gegevens niet altijd voldoende. Datakwaliteitscontrole is daarom een cruciale stap in de transformatie van data naar wijsheid. Onderdeel daarvan is gegevensvalidatie, waarbij drie verschillende stappen kunnen worden onderscheiden: voorbewerking, detectie van foutieve gegevens, en beslissing over eventuele correctie van de gegevens. Gegevensvalidatie is hier geen doel op zich, maar een stap op weg naar grote strategische doelstellingen, zoals een betrouwbaar en efficiënt drinkwatersysteem. Gegevensvalidatie reikt veel verder dan het informatierijk en moet worden gecombineerd met menselijke kennis. Voor een zinvolle analyse is het dus nodig om ruwe gegevens (van sensoren) te combineren met (technische) kennis, bijvoorbeeld over hoe een systeem is ontworpen, of kennis over de nauwkeurigheid en grenzen van elke meettechnologie.

Voor datavalidatie zijn twee uitvoeringsniveaus nodig: *strategisch* (top-down) door i) het ontwikkelen van kaders en normen voor de watersector die aansluiten bij de normen van andere sectoren en *operationeel* (bottom-up) door i) het uitvoeren van pilots, ii) het evalueren van casestudies en iii) het delen van ervaringen tussen waterbedrijven. *Dit* onderzoek heeft zich vooral gericht *op de operationele* implementatie voor gegevensvalidatie.

Aanpak: detectietechnieken voor datafouten getest op twee problemen met drie waterbedrijven.

In dit project lag de focus op het identificeren van foutieve gegevens. Om de aard van 'fouten' beter te onderscheiden, moet onderscheid worden gemaakt tussen systeemuitbijters (zoals extreme

gebeurtenissen) en data-anomalieën (zoals kwaliteit van geregistreerde getallen). Met een literatuurstudie zijn de huidige stand van de techniek en de beschikbare technieken voor gegevensvalidatie worden geïdentificeerd. Vervolgens zijn twee voor waterbedrijven belangrijke soorten problemen uitgewerkt in pilots: i) volumestroom en ii) waterkwaliteitsdatasets van temperatuur, troebelheid, pH en chloor. Voor deze problemen werkten vertegenwoordigers van drie waterbedrijven samen met datawetenschappers van KWR. Er is een stapsgewijs protocol ontwikkeld voor het toepassen van vier eenvoudige tests en toegepast op de datasets. Voor een benchmarknetwerk is een test uitgevoerd met een complexere analyse in combinatie met een hydraulisch model. De resultaten werden besproken met de respectievelijke waterbedrijven en er werden *best practices* en aanbevelingen gedefinieerd voor de verdere implementatie van detectie van foutieve data.

Resultaten: verschillende validatietechnieken beschikbaar: deels getest, nog geen finetuning

Alle waterbedrijven voeren gegevensvalidatie uit, maar op verschillende niveaus van complexiteit. Ook heeft elk bedrijf zijn eigen datamanagementtools en databasesystemenen gebruikt het verschillende specificaties voor tijdsstappen, eenheden, opslag, metadata, datamodel en gegevenscategorieën. In de drinkwatersector ontbreken specifieke richtlijnen (standaarden) voor de keuze van datasets en methodologieën voor validatie. Om deze redenen moeten voor vergelijkbare problemen op maat (per bedrijf) verschillende oplossingen worden ontwikkeld

Er zijn verschillende technieken voor detectie van foutieve gegevens, variërend van eenvoudige tests tot complexe analyses. Eén techniek die op alle problemen toepasbaar is, bestaat niet. Afhankelijk van de gemonitorde gebeurtenis of variabele moeten verschillende technieken met verschillende parameters worden toegepast, afhankelijk van het doel van de validatie.

Waterbedrijven worden geconfronteerd met vergelijkbare problemen op operationeel en strategisch niveau. Welke datareeksen moeten worden gevalideerd en op welk niveau? Welke technieken kunnen worden toegepast? Op dit moment zijn binnen de bedrijven tools op maat ontwikkeld. Rond datavalidatie bestaat een dynamische en continue verbetering en de ontwikkelingen per bedrijf hebben kennis opgeleverd. Toch bestaat de indruk dat waterbedrijven het wiel steeds opnieuw uitvinden. Er is een gemeenschappelijke behoefte aan *best practices* voor kwaliteitscontrole van gegevens en voor centrale validatie van procesgegevens.

Datakwaliteitscontrole vereist inzicht in welke variabelen worden gemonitord en hoe te werk te gaan bij de identificatie van foutieve gegevens. Welke data als foutief worden aangemerkt, hangt sterk af van de 'regels' die worden gebruikt voor de identificatie ervan. De implementatie van eenvoudige technieken heeft aangetoond dat eenvoudige tests het meest afwijkende gedrag kunnen identificeren. Maar zelfs eenvoudige foutieve-data-detectietechnieken moeten door een deskundige worden bijgesteld voor elke variabele bij elk waterbedrijf. Voor een volledige identificatie van anomalieën zijn operationele en onderhoudslogboeken nodig. Op dit punt is er ruimte voor verbetering bij de waterbedrijven. Een test met een complexere analyse (gebaseerd op een hydraulisch model) laat zien dat de mogelijkheden van dergelijke technieken uitgebreider zijn, omdat ze ook rekening houden met de fysica van watertransport in leidingen.

Implementatie: veel mogelijkheden voor betere datakwaliteit, kennisuitwisseling nodig

Specifieke richtlijnen (standaarden) voor de keuze van datasets en methodologieën voor validatie kunnen nuttig zijn voor eventuele toekomstige uitwisseling van gegevens. Ook kennisuitwisseling is nodig, binnen en buiten de drinkwatersector. Verschillende sectoren werken aan systemen om hun datasystemen te verbeteren. Er zijn bijvoorbeeld lessen te trekken uit ervaringen van Nederlandse overheidsinstellingen, zoals Rijkswaterstaat, waar de validatie van gegevens de afgelopen 10 jaar sterk is ontwikkeld.

Rapport

Dit onderzoek is beschreven in rapport *Data Quality Control* (BTO-2019-011).

More information
MSc, Mario Castro Gama
T +31 30 606 9644
E Mario.Castro.Gama@kwrwater.nl

KWR
PO Box 1072
3430 BB Nieuwegein
The Netherlands

KWR Watercycle Research Institute

BTO *Management summary*

Fostering Data Quality Control by exchanging knowledge and implementing pilot projects

Author(s) Mario Castro-Gama, Claudia Agudelo-Vera

Data play a key role in decision-making and in supporting efficient systems. A growing number of companies now view data as a key organizational aspect that has to be properly managed. At the same time, drinking water systems increase in complexity and feature smarter elements, which in turn leads to data-richer operation environments for water services. Given this challenging context, the often-overlooked factor of ensuring high data quality and preventing errors in data streams becomes increasingly important. Data validation is an important part of data quality control. Despite the plethora of techniques and different operational cases, common patterns can be seen across organizations. Water companies largely face similar issues for data validation, and thus working together in these common topics facilitates and speeds up the implementation of consistent data quality control frameworks that work across multiple levels. No single solution fits all needs. Depending on the monitored event, different techniques with varying complexity can be applied, also with regards to the scope of validation.

Importance: data essential for all decisions and models at water companies

Data should not be seen as a 'side product'. Water companies increasingly acknowledge that data of good quality provides a basis for good decisions and that data is essential for creating realistic models, as well as in improving evidence-based reporting. At the same time, efficient data management grows into an organizational requirement for any service, as data-sensitive legislation, such as, for example, the EU INSPIRE Directive, becomes implemented. In this process, Hydroinformatics plays an important role, due to the constant increase of datasets collected from the field (from sensors to smart meters) and due the growing role of data in decision support. However, for some of these purposes the present quality of the data may not suffice. Data quality control represents a key step of the transformation from data to wisdom. A key step of data quality control is the process of data validation, in which three different steps can be differentiated: pre-processing, detection of faulty data, and decision on data correction. Within this context, data validation is not a goal in itself, but should be seen as a step in the path to achieve large strategical objectives such as a reliable and efficient drinking water system. Data validation extends well beyond the information realm and needs to be combined with human knowledge. To provide a meaningful analysis, it is thus needed to combine raw data (from sensors) with (engineering) knowledge, e.g. knowledge of how a system is designed, or knowledge of the accuracy and limits of each measurement technology.

Two levels of implementation for data validation are required: Strategic (Top-down) by i) developing frameworks and standards for the water sector which are compatible with standards of other sectors and, Operational (Bottom up) by i) implementing pilots, ii) evaluating case studies and iii) sharing experiences among utilities. This report focuses mainly on the operational implementation for data validation.

Approach: Faulty data detection techniques were tested in two problems with three water companies.

In this project the focus was on the identification of faulty data. To better distinguish the nature of 'faults', a distinction has to be made between system outliers (e.g. extreme events) and data anomalies (i.e. quality of register entries). With a literature review,

the current state-of-the-art and available techniques for data validation are identified. Two types of problems are then identified as important to be worked out in two pilots: i) volume flow rate and ii) water quality datasets of temperature, turbidity, pH and chlorine. For these problems, representatives of three water companies worked together with data scientists of KWR. A step-by-step protocol of how to apply four simple tests was developed and applied to the data sets. A test using a more complex analysis together with a hydraulic model, was performed for a benchmark network. The results were discussed with the respective water companies and best practices and recommendations were defined for further implementation of faulty data detection.

Results: Several techniques are available, some are being tested, fine tuning is still missing.

All water companies implement data validation, however at different levels of complexity. Each company has its own data management tools and database systems, as well as different specifications regarding time steps, units, storage, metadata, data model and categories of data. The drinking water sector lacks specific guidelines (standards) for the choice of datasets and methodologies for validation. For these reasons, different solutions must be developed for similar problems, tailored to each company.

There are several techniques available for faulty data detection, varying from simple test to complex analysis. There is no such thing as a one size fits all technique.
Depending on the monitored event/variable, different techniques with different parameters should be applied also according to the objective of the validation.
Water companies face similar issues in the operational and strategic level. Which data series have to be validated and to which level? Which techniques can be applied? Customized tools have been developed within the companies, which is good because they learnt in each process leading to a dynamic and continuous improvement. Yet there is the impression that water companies keep reinventing the wheel. Current practices at water utilities show that there is a common need for best practices to built-up data quality control, and as a central process data validation. Data quality control requires an understanding of which variables are monitored and how, to proceed with the identification of faulty data. Faulty data is highly dependent on the 'rules' used for its identification.

Implementation of simple techniques showed that simple tests can identify the most anomalous behaviour. However, even simple faulty data detection techniques need fine tuning by an expert for each variable at each utility. For a complete identification of anomalies, operational and maintenance logs are required, and this is also a topic for improvement by utilities.
A test performed using a more complex (hydraulic model-based) analysis, shows the potentials of using such techniques for data validation going beyond data-based techniques to also take into account the physics of the water transport in pipes.

Implementation: a lot of potential to improve data quality, knowledge exchange is needed

Specific guidelines (standards) for the sector to define which datasets and methodologies are used for validation, can be useful for potential future exchange of data. Knowledge exchange is also necessary, both within and outside the drinking water sector. Different sectors are working on systems to improve their data systems. For example, lessons can be learned from their experiences, including experiences from government institutions in the Netherlands such as Rijkswaterstaat, where data validation has significantly grown as a process in the last 10 years.

Report

This research is a project report of *Data Quality Control* (BTO-2019-011).

More information
MSc, Mario Castro Gama
T +31 30 606 9644
E Mario.Castro.Gama@kwrwater.nl

KWR
PO Box 1072
3430 BB Nieuwegein
The Netherlands

Contents

1 Introduction

1.1 Background

During the last decades, the role of data as a vital resource that enhances decision-making and supports efficient systems operation has become evident, with a growing number of companies viewing data as a key organizational aspect that has to be properly managed, instead of an operational side-product (Tayi and Ballou 1998). At the same time, drinking water systems increase in complexity and feature smarter elements (Mutchek and Williams 2014, Mudumbe and Abu-Mahfouz 2015), which in turn leads to data-richer operation environments for the water services. Given this challenging context, the often-overlooked factor of ensuring high data quality and preventing errors in data streams becomes increasingly important.

Despite the emerging need for holistic, efficient data management policies, implementing a proper Data Quality Control (DQC) strategy is generally a non-trivial task, as the protocols and techniques used are process- and context-dependent. For the water sector, protocols to standardize data acquisition and analysis are being developed for different parts of the water cycle, targeting the data streams of specific processes. For example, management frameworks in the context of urban hydrology and sewer systems have been developed (Bertrand-Krajewski, et al. 2003) (Mourad and Bertrand-Krajewski 2002), as well as initiatives for European ocean and sea data management (EC 2010). In the Netherlands, KWR is developing a protocol for data quality aimed at the registration of groundwater levels and hydraulic heads (von Asmuth 2012) (von Asmuth 2015) (von Asmuth and van Geer 2015) together with the development of its own validation tools (von Asmuth, Maas, et al. 2012). In the drinking water sector, a uniform registration protocol of pipe burst data has been recently developed (Beuken and Moerman 2017).

Within these protocols, one of the core ways of improving data quality is by performing data validation. Data validation or, in other words, fault detection and isolation (FDI), refers to the identification and handling of anomalies and outliers in data that cannot be explained by the underlying physical rules of the measured system[1]. These anomalies, otherwise known as errors, can be further distinguished in three types (Lynggaard-Jensen, Hansen and Bertrand-Krajewski 2012):

i. measurement errors (e.g. failure of data registration, maintenance problems, drifts, bias, strong gradients, lack of redundancy, problems of coherence at both local and global scale, duplication of data),
ii. human errors (e.g. sensor placement, sensor settings, faulty/inadequate calibration, unit conversions, round-off and data conversion errors) and
iii. any occurrence of unexpected processes, modifications and events in the monitored urban water systems, either controlled or uncontrolled (i.e. pipe bursts, flooded pump station, maintenance of a filter at a treatment plant).

When untreated or mismanaged, e.g. due to the lack of a proper protocol for data validation, these errors cause a decrease in the reliability of measured data on the system. In turn, this decreases data quality. It strongly impacts the service operation, as it propagates deeper (Yoo, et al. 2006) into the decision-making process and leads to erroneous or ill-informed decisions

[1] Given this definition, any outliers or anomalies in data owing to natural rare and/or extreme events, including very low probability cases such as black swans (Paté-Cornell 2012), should not be considered as faulty data due to errors that have to be corrected.

on the system operation, organizational mistrust, reduced service efficiency and, eventually, customer dissatisfaction (Redman 1998). The detection and identification of the aforementioned errors can be done with a variety of methods that include threshold, data-driven and model-based approaches, further discussed in Chapter 3.

1.2 Objective

In light of the background provided in Section 1.1, the aim of this project is to find a suitable set of more or less generic data validation techniques, which can be applied and fine-tuned to two problem types related to drinking water services:

i. validation of flow meter data used for leak identification and
ii. validation of water quality sensor data in water treatment plants.

As part of this project, the applicability of the selected data validation techniques is demonstrated in three different pilots and an additional benchmark case. In addition to the application of specific data validation techniques, a protocol in the form of a flowchart is created, in order to help determine the applicability and need for ad hoc modification of generic data validation techniques to specific cases related to the drinking water industry. At the same time, the project extends beyond the limits of a simple data validation analysis, as it serves as a pilot for Hydroinformatics (HI) research in the context of BTO projects. More specifically, it can be used to demonstrate how researchers and practitioners interact in a HI research project and how its results can be implemented in practical applications, feeding analyses with better data than before. As such, this project serves as a basis to support the exploration of a need (or otherwise) for a Hydroinformatics research theme within the BTO.

1.3 Scope and approach

This project aims at providing a contribution towards better Data Quality Control (DQC) policies in the drinking water sector, by providing insights on (raw) data validation in two problem types, one in water quantity and one in water quality. The focus of this project is thus on a specific aspect of the overall DQC chain, which deals with faulty data detection and isolation (FDI). Furthermore, of interest to the project are errors in the measurement, sensing and human data editing process that lead to raw data distortion in the form of e.g. drift, bias, precision degradation or sensor failure (Alferes, et al. 2013). Mapping this focal point to the typology of errors seen in Section 1.1, it becomes evident that this project focuses only on errors of type (i) and type (ii), i.e. measurement and human errors. Moreover, the focus lies on data validation to determine faulty data and the identification techniques, without expanding further on the decision-making process of whether to accept or reject the faulty data. In fact, rejecting the faulty data will trigger a correction or rewrite of the faulty data, which is not covered in this report.

As a first step, in order to identify the needs of the industry and its current practice, an inventory of current applications regarding data quality control within the water companies was conducted. Visits or interviews with four water companies took place during the period January-May 2018, as well as surveys sent to the members of the Hydroinformatics Platform (HI-Platform) (Makropoulos, van Thienen and Agudelo-Vera 2018). Secondly, to gain insight on different approaches on data validation, a literature review on faulty data detection techniques was performed, resulting in an overview of available techniques that are relevant for the drinking water companies. This overview differentiates between simple and complex techniques and also includes an analysis on the range of applications of each one. The insight gained by the literature review allows a step-by-step protocol of data quality control for simple tests to be defined.

Based on the findings of the previous steps, an application in four cases focusing on two types of problems in drinking water follows (TABLE 1). The two problem types, also described in Section 1.2, comprise: (i.) the detection of anomalies in volume flow rate, as an example of data validation in water quantity, and (ii.) anomaly detection in datasets of temperature, turbidity, pH and chlorine, as an example of validation in water quality. The analysis of the case studies was performed in close cooperation with the water companies. Finally, using the information collected from all previous steps, best practices and issues regarding data quality control by the water utilities are identified, as well as recommendations for future application of faulty detection techniques, along with ideas for future research in the field of Data Quality Control.

TABLE 1. OVERVIEW OF THE CASES

	Case type 1	**Case type 2**
Company	Drinking water distribution	Water quality
Company A	X	X
Company B	X	
Company C	X	X
C-Town (hypothetical case)	X	

1.4 Outline of the report

A brief overview of the contents of the following Chapters is provided in this section. In Chapter 2, the necessary foundations and theoretical background in Data Quality Control is defined and an overview of current experiences of the drinking water companies is provided. Having set the foundations, Chapter 3 contains the literature review on faulty data detection techniques. This overview leads to a selection of techniques directly applicable to water utilities and introduces a protocol, in the form of a flowchart, to implement simple techniques for data quality control.

Chapter 4 describes the different studied cases and their results, derived from the application of simple techniques following the proposed protocol. In addition to the real cases, a theoretical model case is introduced to highlight additional possibilities, by using a hydraulic model to generate synthetic datasets of a hypothetical water distribution network. At the end of Chapter 4, the best practices and issues found during the implementation of the pilots is summarised. Following the analysis, Chapter 5 contains the discussion and recommendations that highlight the potential of future research. Chapter 6 then describes the main conclusions draws from each case study, as well as general conclusions drawn from the application of the methodology.

2 Data Quality Control – Theory and Current Practice

2.1 Principles of Data Quality Control

Data is now considered one of the fundamental pieces of the daily operation across many services. Over the recent decades, rapid technological changes have transformed multiple service fields into data-rich environments, where managers and decision-makers are increasingly called to handle, evaluate and decide based on data. Smarter and more frequent metering, along with advances in hardware, editing technologies and new data analysis techniques have reshaped decision-making from an empirical to an increasingly data-driven process (Donhost and Anfara 2010). Furthermore, the role of data is foreseen to grow, with the inclusion of technologies such as cloud-based systems and big data analytics in the systems analysis and, eventually, the decision-making culture, thus causing a paradigm shift in the value of information, the nature of expertise and, eventually, the practice of management and decision-making itself (McAfee, et al. 2012, Kitchin 2014). This paradigm shift is also occurring in the drinking water industry, as drinking water networks become smarter, more networked and more complex (Mudumbe and Abu-Mahfouz 2015, Mutchek and Williams 2014), thus providing increasingly data-rich inputs to the operators and the decision-makers.

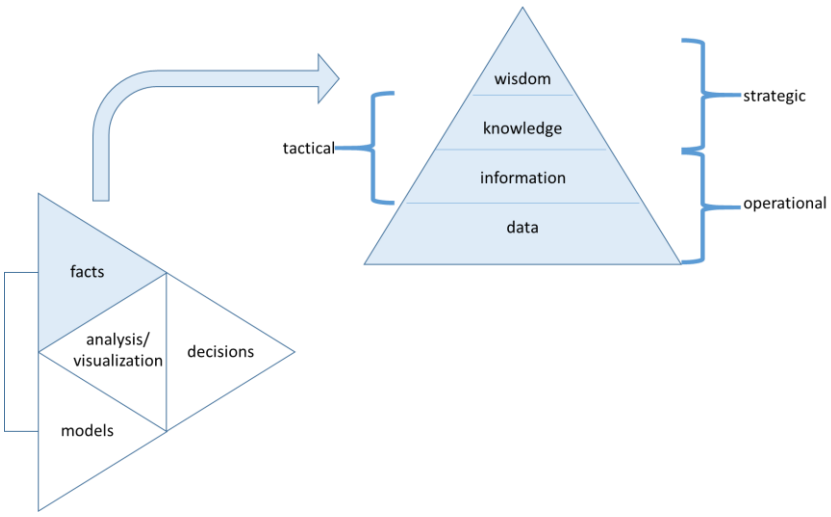


FIGURE 1. OVERVIEW OF THE COMPONENTS FEEDING THE DECISION MAKING PROCESS

The elevated role of data in decision-making leads to a pressing need for more efficient data quality services, as poor data quality leads to less knowledgeable operational decisions and, thus, less reliable systems and customer dissatisfaction (Redman 1998). This becomes evident when data is seen as part of the larger picture of decision-making (FIGURE 1), where data can be considered the foundation of knowledge creation that leads to information, knowledge and, eventually wisdom. In this structure, data acts as the founding component with which the organizational time scales are shifted from the operational collection of (raw) bytes to information analyses at tactical level and, finally, strategic interpretation of the analytical results that provides knowledge and wisdom to management groups. It follows, as a result, that the water companies have acknowledged that good data provides a basis for good

decision making[2]. As such, the policies to control data quality serve a fundamental function to the transformation from data to wisdom (Ackoff 1989) for water utilities, along with the broader frameworks that extend data applications for decision making provided by the concept of hydroinformatics (Makropoulos, van Thienen and Agudelo-Vera 2018).

As in other product, process and service cycles in organizations, ensuring data of good quality requires an encompassing framework of continuous quality improvement, which can be defined as a framework for Data Quality Control (DQC). To design such a framework, classic quality improvement methodologies can be employed, such as the Plan-Do-Check-Act (PDCA) approach (FIGURE 2) (Deming 1986, Ishikawa 1986, Shewhart 1931), which can be used to describe the continuous improvement of measurement systems and their data products as a cyclic process. In the context of data quality, the PDCA approach can be viewed as a proactive framework which continuously monitors and registers data, checks their integrity, acts upon the checked datasets to feed information-based decision-making and plans strategies, including proposition of improvements on the sensing/monitoring system which in turn closes the loop (English 2001, Stausberg, et al. 2006).

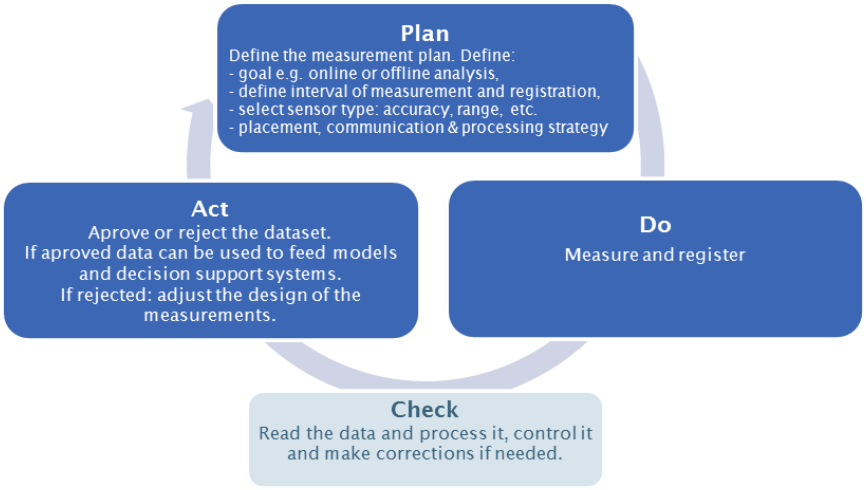


FIGURE 2. THE PDCA APPROACH FOR DATA QUALITY IMPROVEMENT, ADAPTED FROM (DEMING 1986).

2.2 The role of validation in Data Quality Control

Within the DQC context, validation techniques play a key role in connecting the wealth of information obtained by raw data acquisition with decision-making and planning. The acquired data (i.e. the result of a "Do" step in a PDCA cycle, see FIGURE 2) needs to be checked against errors and, in case faults are detected, needs to be corrected before feeding any decision-making process (i.e. the steps of "Act" and "Plan" in a PDCA cycle that complete the loop). To complete this transition, a "Check" step is needed, which is better known in information analysis as Data Validation (Di Zio, et al. 2016).

As seen in FIGURE 3, data validation can be further distinguished in three steps: Collection, Detection and Correction. Data collection refers to the process of gathering data through data streams from each sensing device to a central database, otherwise known as a data warehouse. The step that follows is the detection of a subset of data which could be deemed faulty. Detection techniques typically rely on mathematical modelling and are not trivial, as they have to ensure that they can safely distinguish between actual faulty data and data which appears

[2] Minutes, Hydroinformatics platform 12 October 2017

suspicious but its deviation could be attributed to something else than an error (FIGURE 4). As a last step, the data that is confirmed to be faulty need to be corrected (e.g. empty values filled, outliers corrected based on other close values etc.) before the data can be interpreted further and used as a basis for decision-making. This stepwise process of identifying and correcting faulty data is also known in literature as fault detection and isolation (FDI) (Khorasani 2009).

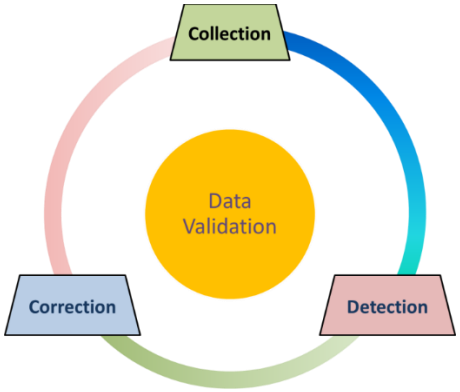


FIGURE 3. THE THREE STEPS COMPRISING DATA VALIDATION.

Primarily, the goal of data validation lies in identifying and extracting the subset of data which may be considered faulty (FIGURE 4), i.e. not representing a valid measurement of reality, due to a measurement or human error (Lynggaard-Jensen, Hansen and Bertrand-Krajewski 2012). From the likely faulty subset of data, some data represent occurrences of irregular/unexpected processes in the system (i.e. pipe bursts, catastrophes, maintenance downtime etc.). This data constitutes a third type of human error that lies beyond the scope of this study, as explained in Section 1.3. The focal point of this study lies, therefore, in techniques that can be used to detect the subset of faulty data whose faultiness can be explained and attributed to measurements, i.e. the first two types of errors seen in Section 1.1.

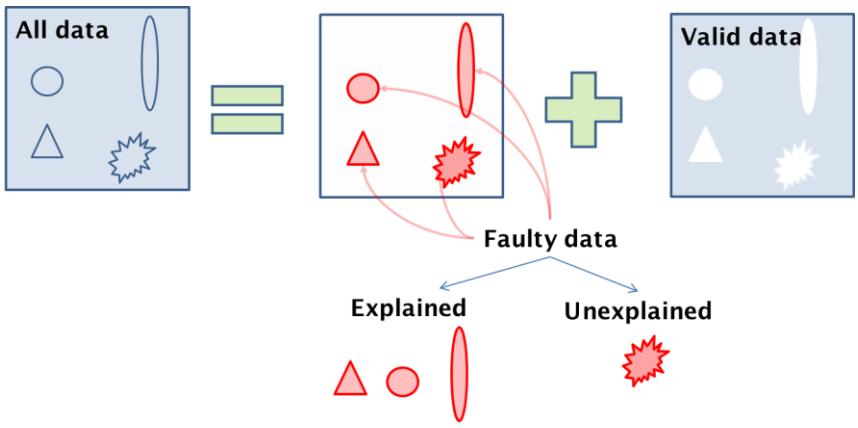


FIGURE 4. THE DATA TARGET GROUP OF VALIDATION.

At the same time, the detection process has to ensure that irregular but non-faulty data are not classified as faulty. For instance, outliers owing to extreme events and even unprecedented events such as black swans (Paté-Cornell 2012) belong to the valid data subgroup and should not be classified as faulty data. As a core process in Data Quality Control, data validation is not a new concept and has been developed heavily in DQC platforms (Di Zio, et al. 2016), relying largely on algorithms and mathematical techniques of faulty data detection. However,

automating the entire process of data validation is not realistic (V. Venkatasubramanian, R. Rengaswamy and S. Kavuri, et al. 2003), and expert judgment is still required to cross-validate the results produced by mathematical methods.

2.3 Data quality control in the context of drinking water

The concepts on data quality control and validation described in Sections 2.1 and 2.2 can be applied to any data-fed production environment, including of course drinking water (DW). In that case, data describing the status of the DW network is acquired by sensing devices from multiple points within the production, transport and distribution chain and typically stored in a central repository called a data warehouse. The aim of Data Quality Control is thus to ensure that the data stored are accurate representations of reality (i.e. physical variables such as water quantity, quality, water level, pressure head etc.) and can be safely used to support decisions that concern the water system operation.

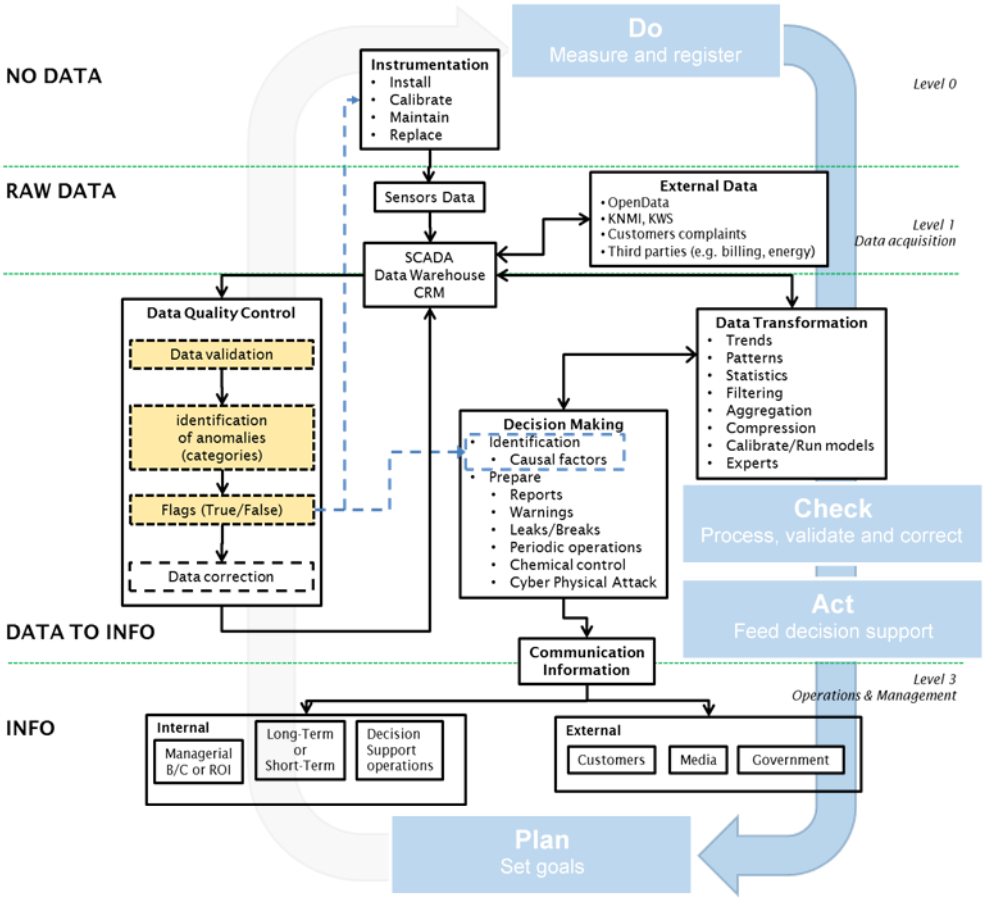


FIGURE 5. FROM DATA TO INFORMATION AT WATER UTILITIES. ADAPTED FROM (HARGESHEIMER, CONIO AND POPOVICOVA 2002)

To demonstrate this, Figure 5, presents the data-to-information workflow typically seen in the context of drinking water. Elements from the PDCA approach, as analysed in Section 2.1. have been mapped, focusing on the Do-Check-Act-Plan parts of the loop that described the pathway from data to information (and, eventually at the plan stage, knowledge). One may observe three distinct levels: acquisition of data (level 1), followed by transformation and quality control (level 2), and finally dissemination of the information produced by data (level 3). In the data acquisition level, data is coming from sensors (e.g. in real time) or can be fed from

periodical manual checks, such as local visits, regular sampling etc. The incoming data are then stored in repositories named data warehouses. Such data can be considered raw data, which means that they are stored as obtained by the sensors. After acquisition, a common workflow inside a data warehouse is to upscale fine-scaled data through aggregation or averaging, in order to produce metrics and time-series at intervals meaningful to management or to identify extreme or periodic events (Gaag and Volz 2008) and causal factors (Bertrand-Krajewski, et al. 2003).

If data coming from sensors and storage were perfect, then such an analysis would be possible almost in real-time. However, in reality raw data are prone to errors, for instance due to sensor failure (maintenance problems, bias, de-calibration, communication failure, physical damage due to catastrophes etc.), due to human mistakes (incorrect installation of measuring equipment e.g. sensor settings, unit conversions, not using the validation protocol issued by the manufacturer or forgetting registering information) and due to unexpected processes, phenomena and events in the monitored urban water system (electrical power outage, failure of a pump). When this happens, data cease to be accurate representations of reality and thus constitute a very poor – and potentially misleading - basis for decision-making.

This likelihood of errors in raw data makes data validation a necessary step of any data quality control protocol in drinking water. With a proper data validation scheme, faulty data is identified, isolated and corrected and can be then standardised to information (e.g. through proper transformation, formatting and metadata inclusion) and used for decision-making and further dissemination to customers, external parties, internal management for tactical/strategic analysis etc. (Figure 5). Due to the numerous processes involved, data validation is not trivial but depends on:

- the type of variable monitored,
- the overall measurement and sensor/monitoring network conditions and more specifically:
    - the degree of system complexity
        - Larger systems may require larger sensor networks in this way more variables are measured simultaneously.
        - Sensors located far away from each other may be correlated or measure completely different patterns with delays.
    - the operational age of the sensor/monitoring network, which is translated in the time length of available data
    - the type and technology sensors/equipment used
        - Precision, accuracy, type of measurement, uncertainty of measurement.

- the characteristics of the phenomenon being captured and more specifically:
    - the type of problem (leakage detection, water balance closure, water quality etc.)
    - the way data is represented (i.e. real numbers as pressures and flows, binary as pump switches, categorical as status of data provided by most systems)
    - data resolution (i.e. both temporal or spatial)
    - the method/technique used for validation (see Chapter 3)
        - Not a single technique can be used for all instances
        - The time spent between techniques can vary between pre- and post-processing
    - the metrics used (i.e. some variables such as pH, temperature and turbidity, are based on a sensor calibration made through laboratory tests)
- the user and objective
    - data may be used for Real Time Control (RTC) or offline historical analysis.

- some methods may be used for Data Warehouse administrators, while for the case of water accountant managers as end users only performance indicators or data aggregation as post-process are relevant.

2.4 Current implementation of data validation techniques by the drinking water companies

Drinking water companies own and manage extensive systems (with several facilities), which are continually monitored in different points, e.g. production, transport and distribution. WBG and WMD for example, have 18 drinking water production sites, 11 industrial water sites and 1 waste water treatment plant. For all these facilities WBG and WMD monitor approx. 27.000 different variables (tags). Meanwhile Company A has approximately 73.000 variables in total, measured every second. Currently every company is dealing with data quality issues. Due to the exponential growth of data and the specific characteristics of each variable, these cannot be easily manually validated.

Additionally, time series (TS) are becoming increasingly necessary for modelling such us hydraulic, risks and decision models.

Other emerging drivers are stricter laws and regulations for the definition of standardization of data models, protocols and congruent at the inside of the EC. For example, the European *INSPIRE* directive[3], defines the technical guidelines for data specification of Infrastructure and its spatial information, although such initiative is currently an invitation for standardization moving forward (which may facilitate exchange of information), rather than a mandatory application for future implementations for the drinking water utilities.

Despite these drivers, there are still several barriers to validate the data. The volume of real-time information has become so extensive that validation of all the variables by a human becomes unrealistic. To deal with it, in some cases data validation is limited to aggregated data (e.g. daily water use in a supply area). In other cases, software tools are built up to screen and flag the data which is identified as suspicious. In general, validation for process automation is sufficient for Programmable Logic Controllers (PLC), but this validation may still not be sufficient for sharing information within or outside an organization.

Currently, it is not possible to validate all the variables. In general, the most important data is validated. This is done by a mix of by hand and automatized routines. One of the companies introduced the concept of a '*data diet*', which implies a profound consideration of which data have to be measured, in which kind of time interval they have to be stored and which of them have to be validated, before starting to generate data. For some datasets, it is not really needed to develop a high level of validation. For those datasets where validation is essential, we should look into the possibility of correlation between different variables, measure all these variables and use correlation techniques (data science, statistics, models) for the validation.

The techniques used by the water companies to validate the data include:

- manual validation (expert judgment);
- visual comparison;
- control of measuring range, plausibility, data types;
- cross-correlation, statistical methods and models;

[3] Commission Regulation (EU) No 1312/2014 of 10 December 2014 amending Regulation (EU) No 1089/2010 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data services

- combining own data with validated external sources (e.g. *BAG - Basisregistratie Adressen en Gebouwen* – Basic registration of addresses and buildings);

Examples of current practices on data validation are:

i) Filling missing data in the records of produced water using registered energy use and relation between energy use and produced m³ of water, or

ii) determining missing year of installation of the pipes using the age of the buildings of the area. Although these methods are not exact, they help to improve the quality of the datasets. To the question regarding which platforms drinking water companies use to store and process the data, each company has its own (customized) systems, examples are shown in TABLE 2.

TABLE 2. OVERVIEW OF TOOLS PER UTILITY

Utility	Systems
1	FEWS, Aspen and Midas.
2	A MS SQL server database.
3	PI (real-time information Assets), SAP (context information of the Assets) Sample manager (information regarding water quality) and SCADA (events and all process information)
4	PGIM (database van ABB 800xa process automatization) and own data warehouse (Microsoft SQL).
5	PA (PIMS) and SAP.
6	GIS (ESRI) own information system (Accent) and SAP SharePoint.
7	Data warehouse and Infor PGIM.
8	Oracle Data warehouse, SQL, Excel, MS Power and BI ARCGIS.

2.5 Data validation - Experiences of a front runner: Company D

Drinking water company Company D is identified as one of the front runners in relation with automatization of routines for data validation. Company D collects a lot of measurement data in PI (from OsiSoft) but still it only validates just a small percentage of all the data.

Company D has a system to validate water volume flows. Company D validates the daily volume flow at measurement points on the boundaries of the DMAs (about 150 pieces per day). The validation consists of checking if the difference between meter readings at the beginning and end of the day is equal to the integral of the analogue readings during the day. If that does not turn out to be correct, the user of the system is assisted in correcting the daily quantity for instance by showing typical values/ranges for this type of day and the historical values of the last 7 or 14 days. Validating always consists of two steps: 1) Check whether the data to be validated is plausible, if not, 2) correct the data. Large customers (> 10k m³/y, approx. 600 units) are validated on a monthly basis. Meter readings of each month are compared with the previous month. It is also checked whether the difference between both meter readings is equal to the sum of hourly values (which are collected to determine peak rates).

Within the data validation process the responsibilities are well defined: An employee (from Company D' Control Center) validates the measurement points in the net and checks that the validation actions are carried out by production sites (if they appear to be necessary). An employee from the Industrial Water department validates large customers (>100k m³/y) and all the industrial water customers. An employee of the Customer Contact Centre validates customers with drinking water consumption between 10k-100k m³/y. The operators on site (*Dienstdoende Operators*) validate the outgoing flows of production sites, plus the waste water

flows and the incoming water flows. There is also a guideline with respect to the extension of the files e.g. xlsx or txt. for different types of validated data. The validation rules are also reported. Despite automation, data validation represents a lot of work and needs constant attention. It has become increasingly clear that not only the quality of the sensor and the data logger, but also a good interface to PI are very important links in the chain.

2.6 Experiences of standardisation towards better data quality

2.6.1 In the water sector

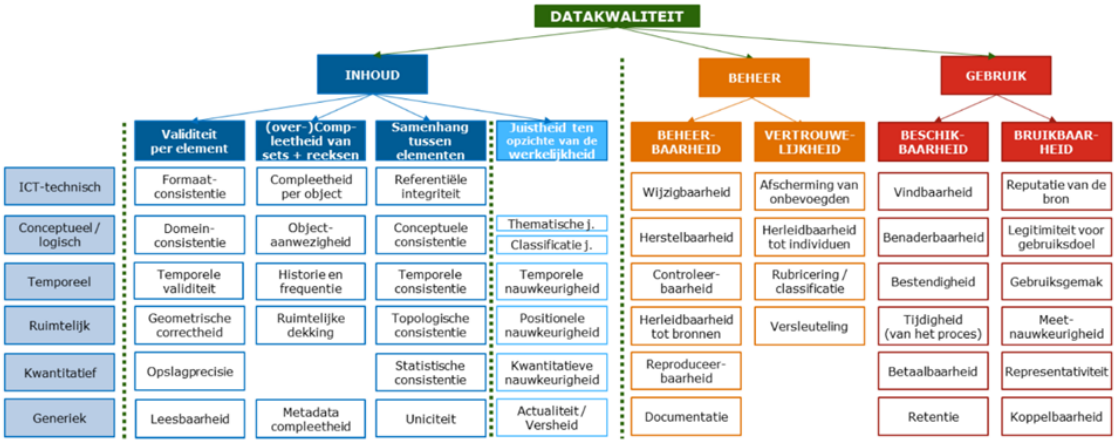
Standardisation is also taking place by developing protocols aiming to improve data quality. The acceptance and implementation of these protocols may take several years. For instance, the provinces in the Netherlands have a long history of validating data regarding monitored ground water levels. Since 2012 a protocol for data quality is being developed for registration of groundwater levels and hydraulic heads between KWR and TNO (von Asmuth 2011) (Post 2013) (von Asmuth 2012) (von Asmuth, Maas, et al. 2012), (von Asmuth and van Geer 2013) (Leunk 2014) (von Asmuth and van Geer 2015). Based on this protocol, tools are being developed to automatize data analysis. As a result, data quality labels can be added to the collected groundwater data classifying them into categories such as: Reliable; Questionable; Unreliable; Censored; Estimate and Missing. The process involving different provinces in that project has shown that data quality is an issue for all organizations, but some organization are less aware of its importance than others. The main lessons learnt in the process are: work on 1) understanding the problem, 2) speak the same language, 3) apply the same methods and 4) use the same tools.

Similar protocols have been developed for the uniform registration of pipe failure data USTORE (Uniforme STOringsRegistratie). This has been an on-going process of continuous improvement, with both the complexity of the registration and data requirements increasing over the years. The initiative started in 2001 and in 2017 it was included in the PCD (Praktijkcode Drinkwater) no. 9: 'Uniform failure registration. In 2018 this PCD is being implemented. This guideline differentiates between process and results oriented requirements (Beuken and Moerman 2017). Having the protocol is only one of the steps of data improvement. For the implementation of these protocols it is crucial that the person who registers the data, in this case the fitters, has to see the added value of good registration. The system requires continuous maintenance and evaluation and it is a continuous process, consistent with the process presented in FIGURE 2.

2.6.2 Outside the drinking water sector - Rijkswaterstaat

In the Netherlands, Rijkswaterstaat (RWS) is responsible for the design, construction, management and maintenance of the main infrastructure facilities in the Netherlands. This includes the main road network, the main waterway network and water systems. RWS is composed of 7 regions, which worked independently and are now centralized. Data quality issues cause time losses e.g. finding missing data, and affects different departments, e.g. Operational, F&C, HR, etc. Retrieving data afterwards, when the data system is not in order, is a costly and difficult (if not impossible) task. A data quality system helps to improve different processes. Improving data quality is not be considered at RWS a project, but an iterative continuous process. RWS developed a framework for data quality (FIGURE 6). It differentiates between data content, data management and data use. This framework is divided in 8 main dimensions and 47 sub-dimensions. In total there are 3 objectives which this framework covers: 1) common language, 2) Inspiration for drawing up requirements and 3) Comparable outcomes of measurements.

FIGURE 6. DATA QUALITY FRAMEWORK DEVELOPED BY RWS[4]



In 2016, RWS begun the inventory of data quality control practices as a pilot. As a result a dashboard supporting data quality process was envisioned. By the end 2016, a functional dashboard became available. On top of that once many data sources were available, it became an issue for RWS to program all rules required for data acquisition, for each specific data set. For that reason, a larger project was signed and it is still under development.

The framework of RWS was implemented as a 'validation factory'. This means that standard validation methods can be applied. A repository is used to store the data, while the dashboard is able to show the status of the 8 dimensions and 47 sub-dimensions. Such dashboard is useful for both managers and operators at different levels of the organization. Data content is process oriented, meanwhile data management and data use are focused on the intended product to be derived from data (product oriented).

For RWS, it is expected that the framework will help operators identify potential improvements. In fact, at different levels of the organization not all dimensions have to be implemented. The dashboard in itself helps with the prioritization of certain dimensions. Some dimensions are checked annually with a complete traceability (check list includes: check data boundary conditions, backup, user, etc).

[4] Adapted from presentation by Kasper Kisjes (Rijkswaterstat), 27 maart 2018 KWR, Nieuwegein.

3 Literature review on faulty data detection techniques for water utilities

3.1 Background

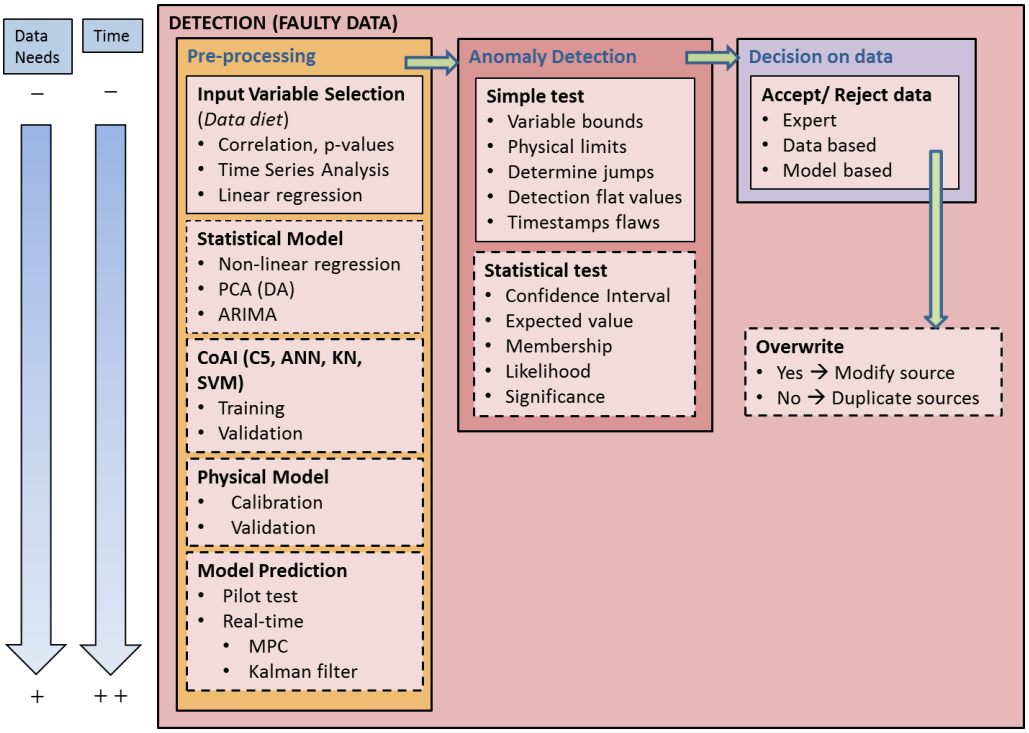
Detection of anomalies corresponds to the first line of defense against faulty data. When operations of the system require different types of measures to perform data quality control, detection allows for the identification of individual values or sets of values which are not properly measured or that indicate a deviation from the measured variables of a system (i.e. a water distribution network, a treatment plant).

This is done by creating a validation mechanism for the object that generates data (e.g. a sensor) at the time it generates the data, with various techniques having been developed for this purpose (Sun, et al. 2011, EPA 2006, EC 2010). In most cases data validation is carried out manually by expert judgment using both analytic and visualization tools. The issue is that with current data streams only a small amount of data can be validated by operatives from the utilities (Mourad and Bertrand-Krajewski 2002), and as evidenced by several authors there is always a human bias in the decision making of determination whether data corresponds to anomalous behavior (explained) or faulty data (unexplained) (V. Venkatasubramanian, R. Rengaswamy and K. Yin, et al. 2003).

Here a distinction must be made as anomalies can be obtained also when sensors measure leakages or pipe bursts, however in the context of data validation, if the proper records of such events are kept, then the data samples containing such anomalies are not considered faulty data, and will be considered valid data with the connotation that such data corresponds to specific events. In this research, faulty data corresponds to data which is anomalous and remains unexplained after performing data validation.

An inventory of data validation techniques is presented in FIGURE 7, based on a literature review on the subject during this project. In general, there are 3 relevant steps for data validation (FIGURE 7) corresponding to 0) data-diet, i) pre-processing, ii) anomaly or faulty data detection, and iii) decision on the data subset (Branisqvljevic, Kapelan and Prodanovic 2011). Currently due to the lack of tools to validate all the available data, it is recommended an extra preliminary step: ‘data-diet’, which consist of the selection of which data have to be validated. Input Variable Selection is a dynamic process, which means that it is not an static process done once, but and iterative process. Some variables may not be of interest as explanatory now, but in the future with improvements of the volume of data and capacity of computation they can become of interest. The diagram of FIGURE 7, also provides an indication of the amount of data required (Data vector) and the amount of time (Time vector) that such task may require for implementation at a drinking water company. In pre-processing techniques, the use of models or meta-models may drastically increase the data and computational requirements, however the level of uncertainty could be reduced significantly.

FIGURE 7. INVENTORY OF FAULTY DATA DETECTION TECHNIQUES. DOTTED LINES REPRESENT TECHNIQUES IDENTIFIED BUT NOT IMPLEMENTED IN THIS PROJECT, WHILE SOLID LINES REPRESENT TECHNIQUES IMPLEMENTED ON CASE STUDIES.



3.2 Faulty data detection techniques

Faulty data detection techniques generally classify the data into two classes:

- class of correct data and
- class of faulty or doubtful data.

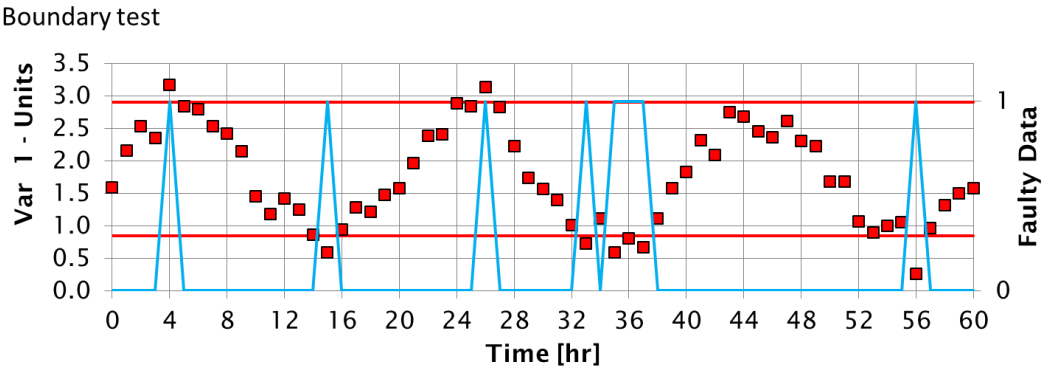
Some techniques for faulty data detection have been developed for generic problems (Di Zio, et al. 2016) (Waal 2013) (Wilson 1993). In the water domain, such techniques have been developed for sewer systems (Branisqvljevic, Kapelan and Prodanovic 2011), geo-hydrological systems (Sun, et al. 2011), (von Asmuth, Maas, et al. 2012) (von Asmuth 2011) (von Asmuth 2012) (von Asmuth and van Geer 2015), analysis of water quality data (McKenna 2007), automatic or real-time data validation in urban systems (sewers mainly) (Mourad and Bertrand-Krajewski 2002), and for specific problems such as the determination of leakages as anomalous data in water supply systems (Mounce, et al. 2014).

In the following sections we will present an overview of anomaly detection techniques in increasing order of complexity and data requirements.

3.2.1 Boundary or range detection

There are several techniques which may be used for determining if data is contained inside a certain boundary or boundaries (FIGURE 8).

FIGURE 8. TYPICAL BOUNDARY TEST. MEASURED SAMPLES IN RED, BOUNDARIES AS RED LINES. VALUES OUTSIDE THE BOUNDARIES ARE MARKED AS FAULTY DATA.



- **Zero value detection**

In some cases, it is necessary to identify whether zero is a valid category or value of sensor data, otherwise data is identified as faulty data. The determination depends on the variable and the sensor type and characteristics.

- **Minimum and maximum detection values**

This analysis is based on geometric, hydraulic and data quality constraints. For instance, tanks have a limited capacity, as such, levels above the maximum value or negative values will represent faulty data. These types of errors can be easily identified by setting proper thresholds on data before storage to the database or data warehouses. Such test requires a coordinated effort between the field operatives and data managers. By using this data detection technique, it can also happen that the sensor when failing (e.g. due to calibration, power failure or maintenance) triggers data such as “Null”, “-9999”. Such data is a representation of sensor downtime and can be easily identified by this simple test as faulty data.

- **Minimum and maximum based on historical values**

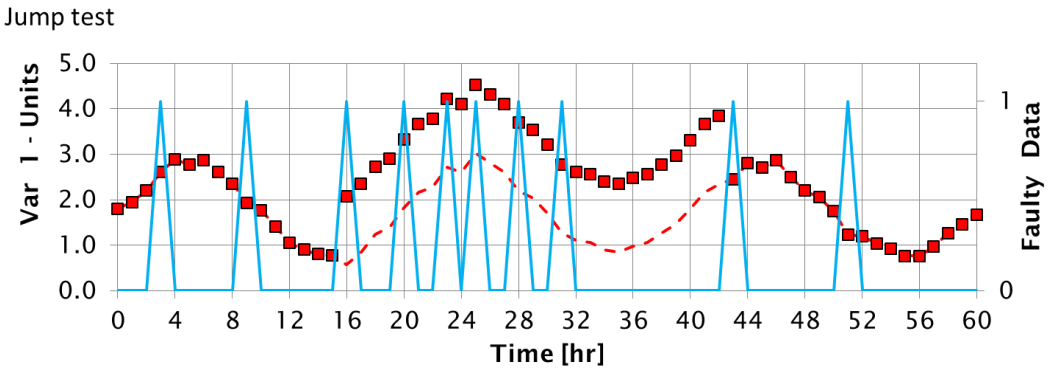
Another case, is when some water quality variables contain limited thresholds. In a treatment plant, temperatures will fall in certain ranges, and extremely high or low temperature measurements can be considered as faulty data. It is also possible, for example, that the temperature of certain reactors in a treatment plant require a range of variability. Such threshold can be easily set for detection of faulty data. Once some portion of the data crosses a threshold it can be easily identified as faulty or suspicious data. It should be noticed that in this case the range of variability of a sensor can be dramatically reduced and the number of issuing warning due to faulty data will increase. For that reason, knowledge-based thresholds must be provided by the operators of the system.

- **Jump or leap detection**

In this case, it is also plausible to determine leaps or jumps in signal data. If one knows that pressure can’t vary in a second 30m, or that a tank can’t become full in consecutive time steps due to the physical limitations of capacity of the inlet and outlet pipes/pumps, then it is possible to determine with a certain precision whether or not data is outside the physical rate changes (FIGURE 9). In the case of data quality, there are several aspects which need to be

taken into account. For example, water quality data tends to be more expensive than pressure of flow data (e.g. the cost of both sensors and its periodic maintenance). This limits the number of data samples. If errors are present in water quality data, when aggregation is performed at hourly, daily or monthly scales is performed, errors may become additive. One possible way of identifying faulty data due to jump/leap is by obtaining the difference among successive timestamps.

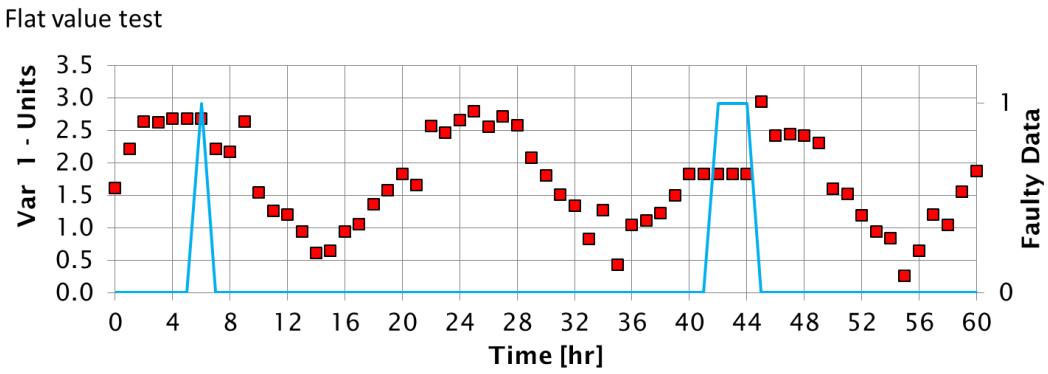
FIGURE 9. TYPICAL JUMP TEST. MEASURED SAMPLES IN RED, EXPECTED TREND AS DOTTED LINE. VALUES WHICH ARE DEVIATING FROM EXPECTATION ARE MARKED AS FAULTY DATA.



- **Flat line detection**

Flat line detection can be used for multiple purposes, as a tool for the detection of gaps in the data or for determining constant values which tend to be very rare inside a Water Distribution Network (WDN). Contrary to the previous cases, a trend or pattern inside the data could be identified in which the sensor measures the same value in two, three or more time-steps. If that is the case faulty data can be identified, or at least flagged, as the dynamics of a complex system such as a WDN usually do not allow for such 'stable' behavior.

FIGURE 10. FLAT VALUE TEST. SAMPLES ARE RED DOTS. WHEN SEVERAL CONSECUTIVE SAMPLES DISPLAY THE SAME VALUE, SUCH SAMPLES ARE MARKED AS FAULTY DATA.



This type of verification can also indicate that for some particular time windows, a subset of values is measured corresponding to the maximum feasible value for a particular sensor (saturation). This type of anomaly shows that the capacity/selection of the sensor is insufficient for the variable of interest. It could also indicate that a re-calibration of the sensor is required.

Exceptions of variables for which flat line detection is not required, are pump status where a flat line can indicate only that the pump is switched on or off for a particular period of time, such that faulty data is not being identified by the system but rather a normal operation.

From a computational perspective, data is stored in a database and it is not exempt of numerical rounding errors. If the verification is based only on equality constraints between consecutive timestamps, then the identification may be impossible to perform. The numerical precision may be too low. For this a threshold must be defined as a difference between consecutive measurements. This helps to avoid issuing a warning of faulty data, when in reality there could be a steady increase/decrease with low magnitude for a particular time window. This indicates, that proper knowledge of the system is extremely valuable. If the threshold is "coarse" many faulty data will be identified, if the threshold is "fine" it is possible that no faulty data is identified. It should be noted that the duration or time window for a flat value test requires a calibration for each variable and for each sensor (McKenna 2007).

In this category it is also possible to use slope/gradient tests, to determine sensor drift, however such techniques are not considered in this report.

3.2.2 Statistical test of variables that follow certain distributions

Not all data from every sensor can be simultaneously checked, for that reason, statistical analyses are relevant for faulty data detection.

- **Comparison of Flow Pattern Distributions (CFPD)**

As an example, inflows and consumption patterns into a Water Distribution Network (WDN) will follow a very limited number of patterns throughout the day. The water demand is a stochastic process. However, at certain hours of the day, the consumption tends to be lower such as during Minimum Night Flow (MNF). There will be also a variation of consumption due to season change (i.e. winter, summer) (van Thienen, et al. 2012) or due to specific events in a short window duration (Bakker 2014). In this case, a comparison of flow pattern distributions (CFPD) method for the identification, quantification and interpretation of anomalies in district metered areas (DMAs) or supply area flow time series relies, for practical applications, on visual identification and interpretation of features in CFPD block diagrams. Such analysis of features and automated screening of data with seasonal statistics can be used to measure deviations for the expected value and infer whether anomalous data has been detected or not (Thienen and Vertommen 2015).

- **Extreme value check using statistics**

When performing detection of certain variables, there are statistical tests which may be applied to the data. One of them is to perform as statistical analysis of the probability distributions of known observed values which are considered as good values, and representative of the behavior of the sensor. In this way, the quartiles of the data can be obtained and a hypothesis test can be performed to identify certain samples which are outside the statistical boundaries. If there is significance of it, then the value can be registered as an outlier, implying that its consideration as faulty data.

- **Principal Component Analysis (PCA)**

This technique will be presented as a selection of explanatory variables using regression.

- **Spatial deviation**

Sometimes, variables are correlated in space. For this a complete set of techniques for faulty data identification can be explored. Kriging and other geo-statistical techniques can be applied to estimate deviation between the measured value and the estimate (Clarke 2013). However this techniques have been mostly applied in hydrology and fell outside of the scope of the pilot.

3.2.2.1 Regression

In general terms, regression analysis can be used to determine which data points contain faulty data. In this case, a pre-processing is required in which a regression model based on "good" data is prepared. Such a regression model contains 2 components. *Predictors* or *explanatory variables* and *response* variables. For example, one may be interested in the relation between the monthly water production [m^3] of a utility and the total energy consumption [kWh] used for treatment, transmission and distribution. If there is a missing or suspect faulty data in water production, then this value can be estimated based on the total energy consumption of the utility. In statistical terms, energy consumption becomes the explanatory variable of the water production. Notice also, that it is possible to have several explanatory variables for the water production. In that case, the regression is not linear but multivariate. In some cases, one may estimate non-linear regressions among explanatory and response variables, however the number of possible transformations is very large and a selection of the best available regression model needs to be determined (Furnival 1971) (Hocking and Leslie 1967) (Schatzoff, Fienberg and Tsao 1968). Another approach, is the use of weighted regression (Hirsch, Moyer and Archfield 2010) where some subsets of data can be given more importance during the regression. In that case decisions regarding anomalies are made stirred by an analysts' decision. Some techniques exist, even when data is sampled in unevenly distributed intervals (Lomb 1976), as it is mostly the case of water utilities, where most of the records are stored at specific events and subsequently interpolated prior to storage in databases and data warehouses.

One may estimate the regression between the 2 (or more) variables. If the adjustment of the regression (Root Mean Square Error - RMSE, Coefficient of Determination - R^2, Nash-Sutcliffe Efficiency - NSE, Kling Gupta Efficiency - KGE) is good (for RMSE must be of low magnitude, while for efficiencies near 1.0 is optimal) then the model represents the response variable as a function of the explanatory variables. Then to determine whether a new data point (as in the example of water production vs energy use) is categorized either as good or faulty, the regression will determine if by comparison there is a large error in the estimation. If that is the case, the new data point should be marked as faulty data.

There are two issues which are of relevance for regression for data validation:

- How to select the explanatory variables for a certain response variable?
- How to decide whether or not measured value is far from expectation?

As posed during the first workshop of the Hydroinformatics Platform in 2018 (February 15, 2018), the first question is related with the concept of Data Diet. Data Diet in the sense of looking for a correlation between different variables by: 1) deciding which correlations are possible by using domain knowledge (expertise of drinking water) and 2) selection of the correlation technique used by data scientists.

As an example, one may have data from a thousand variables ($n_v \sim 1000$) within a SCADA, Data Warehouse or database. If an operator wants to perform the correlation analysis of each pair of variables with the goal of identifying all possible relationships among variables (e.g. correlated not correlated). The total number of such combinations of variables (n_{dd}) as response variables of another one is estimated as $n_{dd} = \frac{n_v(n_v-1)}{2} \sim 500.000$, or half a million different analysis. An operative may not be able to analyze all possible combinations in its own laptop or desktop computer, so here domain knowledge is of relevance. Such task is cumbersome, although it needs to be done only once. This opens another issue, the necessary development of tools for handling such large datasets and number of variables. Specifically, the datasets grow every minute in the data warehouses due to the constant feeding of data from sensors.

For that reason, there are several techniques which can be used for the selection of explanatory variables. Such techniques are known as Input Variable Selection (IVS) and have broader applications in engineering and environmental sciences. We refer to Galelli (2014), Castro-Gama et al. (2014) and (Hocking and Leslie 1967) which have developed methodologies for this subject in diverse water resources and environmental systems. Such techniques apply correlation and stepwise selection of explanatory variables to perform the identification of significant dependencies. Two main techniques used for IVS are:

- Correlation analysis. Based on time series analysis (Box, Jenkins and Reinsel 2008), the selection of the variables can be performed based on a preprocessing of data. A division here must be made between stochastic and deterministic time series analysis (Fatichi, et al. 2009) depending on the consideration of errors and memory of the variable. Some authors have taken into account long term or long range dependency (Beran 2010) (Beran, Feng, et al. 2013), long term persistence (Ehsanzadeh and Adamowski 2010) (Lennartz and Bunde 2009) (Rea, et al. 2009), although with applications to broader fields than water resources. In water quality analysis correlation has been extensively applied for open and pressurized flow when data validation is required (Hirsch, Alexander and Smith 1991) (Darken, et al. 2002).
- Principal Component Analysis (PCA), as its name states had been applied mainly for the determination of variables which may project the data into its principal components. This means that data is transformed using an orthogonal transformation, and converted into a set of variables which are uncorrelated among themselves, such set is denominated principal components. One particular application is consumption patterns and leakage detection (Palau, Arregui and Carlos 2012). In a sense, the values which deviate from the PCA's, pre-identified for a specific WDN or a DMA, are considered as faulty data of water use.

An additional important question is related to the need to determine what is the measure to be used once a screening of PCA or correlation has been performed. We discussed RMSE, R^2, NSE and KGE as possible metrics. However, several other metrics can be also used to evaluate the performance of the techniques used, depending on the frequency of registration and the variable type to be analyzed (Castelletti, Galelli and Ratto, et al. 2012), or with the aid of meta-models (black box models) of the system (Castelletti, Galelli and Restelli, et al. 2012) (Ratto, Castelletti and Pagano 2012).

- **ARIMA**

Auto-Regressive Integrated Moving Average (ARIMA) models are a particular type of regression based on previous values from the time series itself and of estimations of averages from previous time steps. It could be seen as a regression of a variable on itself and many other variables (Oriani, et al. 2016) (Montanari, Rosso and Taqqu 2000) (Montanari, Taqqu and Teverovsky 1999). This technique can be used to identify anomalies by comparing the estimated ARIMA regression for a particular variable to the measured data. However, when data is complex and the variable of interest non-stationary in nature (e.g. when statistical properties of the system change in time) ARIMA's can't account for it. For this reason, an extension to ARIMA has been developed denominated ARIMAX.

3.2.2.2 CoAI and Classification

A higher level of data abstraction for faulty data detection is originated from the field of Computational- or Artificial- Intelligence (CoAI) as Data Driven Modeling (DDM) (Vries, et al. 2016) (Hill and Minsker 2010). Recently, a report on Explorations of Data Mining has addressed many different methodologies (Thienen, et al. 2018).

When several variables are available, with a large number of samples, with known status as good values, it is possible to generate multiple data subsets or clusters among the variables. Implicitly, it is possible to obtain information or detect faulty data when data measured by sensors is outside of the typical cluster to which a certain variable is expected to belong.

The added value by splitting available data into clusters also allows operatives to understand typical patterns of behavior of their system, and may reduce the burden of decision making of the faulty detection.

One advantage of this set of techniques is its efficiency, as after training is performed on the classifier model, it becomes easy to use it for data analysis purposes. The drawback is that for training purposes, a large amount of data is required which does not contain faulty data for the model to learn the current system status. Among the techniques used for this type of analysis there are:

- Decision Trees (DecT)

Probably the simplest technique to perform classification of large datasets is DecT (Quinlan 1992). This technique allows to create rules or boundaries among datasets. It is available in several software packages such as Weka[5] and has been used for example for online prediction of leaks and breaks for WDN models (Allen, et al. 2011). Once the DecT have been trained there is no way to perform a correction without regenerating the trees.

- One-class Support Vector Machine classification

Support Vector Machines (SVM) are a specific type of non-linear regression. SVM correspond to specific classification tools in multidimensional spaces. If a SVM is built based on a large collection of data from a water utility, then it is possible to perform comparisons and indicate whether it contains faulty data or not. In principle, SVM's require large datasets for training.

[5] https://www.cs.waikato.ac.nz/ml/weka/downloading.html

Previously, SVM's have been used for leakages and demand pattern identification (Mashford, et al. 2009) (Mounce, Mounce and Boxall 2011) (Candelieri, Soldi and Archetti 2014), however to our knowledge not directly for DQC of WDN.

- Artificial neural networks (ANN)

Use of ANNs as classifiers

A particular subgroup Artificial Neural Networks (ANN), is **Kohonen networks** (or Self-Organizing Maps) used mainly for (unsupervised learning-based) classification. This technique has been applied for water balance in small DMA's for the identification of leakages (Aksela, Aksela and Vahala 2009). Values identified with a deviation from the cluster predicted by the model, correspond to faulty data.

Use of ANNs as surrogate models of WDNs

The use of a physically based model for the simulation of a real system can be extremely complex and its use can be computationally cumbersome. For that reason, ANN has been used as a surrogate or a meta-model to capture the main features which drive a system.

Although ANN are not new, for problems involving WDN their applications are relatively recent. Mainly, because data collection for WDNs has become available mostly in the last 30 years. Even with more data being available though the number of possible scenarios of operation of a WSS is unbounded and only a subset may be used for ANN training.

To create an ANN which simulates the behavior of a system, some authors have created/collected a large collection of pressures and flows in a WSS. The goal is to create an accurate, but fast-simulation surrogate or meta-model of the system. After training with the data of the ANN is performed, the ANN can then be used to simulate the behavior of a WDN and reduce computational time with respect to a physically based model, to be able to compare with measured variables, in near real-time. This technique has been applied for example to the case of the WDN of Haifa, Israel (Preis 2011) (Perelman and Ostfeld 2011).

The use of ANN has been implemented for detection of anomalies, specifically for water losses and leakages in WSS (Mounce, et al. 2014). In other cases, ANNs have been used to simulate the behavior of a treatment/purification plant, given that the latter systems although complex tend to be more constrained in their operational variability, and as such more suitable for ANN. In all these cases, the ANN was used as to simulate the expected performance of a water system and then compare this to measured data. In that way an ANN could be a reliable, computationally inexpensive tool for faulty data detection.

On the downside, extreme events or situations (i.e. combination of operational variables and treatment conditions) which have not been used as data for ANN training will be easily (but erroneously) detected as faulty data. For that reason, a long time series containing multiple scenarios of operation is required for training and validation purposes of ANN.

3.2.2.3 Physical models

Physical models such as Synergy/EPANET/Infoworks/WaterGEMS are probably the most reliable way to identify faulty data. If a WDN model of the system is available (and properly calibrated), it is possible to simulate the behavior of the network and then extract the data of pressures and flows if given the proper drivers (e.g. current levels in tanks and reservoirs, demand forecast) are known.

Once the measured data have been collected, it is possible to use simulation output (model) to compare with filed data and extract valuable insights: In most of the cases, there will be a similarity between the two. It is also possible, that for some locations there could be a difference or deviation between modelled and measured flows and pressures. This could indicate an increase of losses due to pipe breaks or background leakages (Mesman and van Thienen 2015), or in some cases even the detection of faulty sensors (i.e. low battery, power line downtime, communications network collapse or even scheduled maintenance).

Running an ensemble of models with different parameters can even allow utilities to obtain a probability distribution of system behavior, under certain conditions (Poulakis, Valougeorgis and Papadimitriou 2003). The trade-off for utilities is one between model reliability and computational cost. The more reliable the model is, the larger the effort to keep models up-to-date and properly calibrated. This requires extensive field- and office- work, which in the end becomes cost for the utility.

It should be note that physically based models have the limitation of typically requiring longer computational time. The growth of GIS based models gave practitioners the possibility to increase the size of models, and these tend to become larger every day. Its computational times become a burden as compared with meta-models such as ANN, and tend to require update and continuous calibration as WDN are continuously evolving.

An alternative to curb the computational time of physically based models is hydraulic model reduction. For this, several algorithms exist. Among such Skeletonization (Martínez-Solano 2017), (Ulanicki 1996), hydraulic simplification (Anderson and Al-Jamal 1995) and Topological aggregation of serial pipes (Giustolisi, et al. 2012) can be used to increase computational speed of a network model. Skeletonization and hydraulic simplification result in the sacrifice of energy balance in each simulation, while topological aggregation is impossible to use with commercial packages, as it is only implemented inside research tools.

Another interesting development in this context comes from a promising strategy recently developed (Tsoukalas, et al. 2016) to address these shortcomings using of surrogate modeling techniques. This entailed the development of the Surrogate-Enhanced Evolutionary Annealing-Simplex (SEEAS) algorithm that couples the strengths of surrogate modeling with the effectiveness and efficiency of the evolutionary annealing-simplex method. SEEAS combines three different optimization approaches (evolutionary search, simulated annealing, downhill simplex). Its performance is benchmarked against other surrogate-assisted algorithms in several test functions and two water management applications (model calibration, multi-reservoir management) with promising results revealing the significant potential of using SEEAS in challenging optimization problems on a (computational) budget.

At the intersection between data and models, Bragalli et al. (2016), developed a 3-level data assimilation technique in WDN. The technique is based in Ensemble Kalman filter (EnKF), where the updates of the system states (i.e. pressures, flows, demands) are based on the sensor data and the estimation in a hydraulic model built in EPANET. EnKF has shown better convergence and stability than other filters when applied to non-linear systems. The algorithm proposed by Bragalli et al., progressively assimilates data from known pressures (e.g. sensors inside the WDN or tank levels), then flows (e.g. from pumps or flow meters) and finally (if available) demands (e.g. from AMR). It can be used to reduce the uncertainty of models, perform model predictive control and also to identify faulty data. However, the disadvantage is that in the absence of such data, the reduction of uncertainty becomes limited (Bragalli, Fortini and Todini 2017). It can be expected however that with the increase of computational capacity (Hadka 2013), in the near future online optimization of asynchronous data streams, with possible

application for data validation of WDN, will be feasible for multiple sensors, sensor types, resolutions and data sources.

3.3 Knowledge based techniques or tools

3.3.1 Check of status of sensor or asset

Most SCADA and data warehouses contain large amounts of log files from the different sensors which are used for monitoring purposes. In many cases, information related to maintenance and battery replacement are available. Some variables related to water quality such as pH and turbidity require a calibration of the signal of the sensor with respect to samples which are analyzed in the laboratory. Then a regression is performed between the measured signal and the laboratory sample. Sometimes the logbooks contain metadata and information such as: which laboratory realized the analysis, and also which technique or laboratory test was used for the estimation of the regression "sample data vs signal".

In other cases, Pipe Failure Data (PFD) is of relevance as it helps to update physically based models, and clarify certain anomalies which otherwise would be identified as faulty data (Vreeburg, et al. 2013). For water companies, this is still a challenge as information of logbooks has become more available recently, but there is still a need to integrate PFD and replacement techniques in the same repository as it is the case of USTORE project for 8 water utilities of The Netherlands (Kwakkel, et al. 2015) (Moerman, Beuken and Wols 2017).

3.3.2 Check the duration between sensor maintenances

This information is extremely valuable as it determines the lifetime of data. In the era of Big Data, the reliability of sensor data becomes a burden, more sensors require constant maintenance and more data is constantly stored in the data warehouses. Performing a proper scheduling of maintenance of sensors will become a relevant task in the coming years for all water utilities.

3.3.3 Other tools available

From a scientific perspective there are plenty of available tools to perform DQC in diverse programming languages. A summary of a short survey of such is presented in TABLE 3. These tools are mainly statistical packages which contain different libraries for Data Validation or DQC.

TABLE 3. A SHORT SUMMARY OF SOFTWARE FOR DATA QUALITY CONTROL AND DATA VALIDATION

Tool Name	Programming Language	Reference
Menyanthes	MATLAB	(von Asmuth, Maas, et al. 2012)
PIMS	Low level languages such as C/C++	Developed by OSIsoft
AURA	C/C++	(Mounce, et al. 2014)
CANARY	MATLAB, although it is sealed as p-codes	(McKenna 2007), and its application by (Housh and Ohar 2017)
CAPTAIN	MATLAB and Java, both as p-codes	(Young, Tych and Taylor 2009)
FRACTAL	R language, open	(Constantine and Percival 2014)
FITDISTRPLUS	R language, open	(Delignette-Muller and Dutang 2015)
CTS	R language, open	(Wang 2013)
ZOO	R language, open	(Zeileis and Grothendieck 2005)

Tool Name	Programming Language	Reference
GAUSS18	Python library	(Aptech, 2018)
SLICE	Python library	(Waal 2001)
CHERRYPI	Python	(Waal 1996)
EMS	Python library	(Scheel 2000)
FEAR		(Wilson 1993) and (Wilson 2008)
LOWESS		(Cleveland 1981)
Minitab	Itself	
SPSS	Itself	
STATA	Itself	

Other tools used for specific tasks of DQC by the utilities can be found in Section 2.4

3.3.4 Periodic calibration of sensors and measuring systems

It is important for utilities to constantly identify and calibrate sensors which may be not complying with standards for accuracy. Through time, flow and pressure meters tend to show a reduction in their accuracy and it is necessary to perform analysis of the measurements which are made by the sensors (Aisopou, Stoianov and Graham 2012). In a sense, this preserves both the lifetime of the sensor, and avoids its replacement. Even more serious is the case when (almost) no maintenance is performed and the data provided by a sensor must be considered non-valid data.

3.4 Protocol for data quality control

In the framework of this study, a protocol has been developed, which is meant to describe the process from data acquisition (raw data) until validated information (used mainly for internal and external communication) is generated. In general, we depart from other methodologies used by the *Central Bureau voor de Statistiek* (CBS) (Pannekoek, Scholtus and van der Loo 2013), or for water resources such as: groundwater DQC (von Asmuth and van Geer 2015) (von Asmuth 2015) (Leunk 2014) (von Asmuth 2012), asset management registration (Beuken and Moerman 2017) or environmental monitoring (EPA 2006), currently available for a complete data quality control for two reasons:

1) because data, administered by drinking water utilities, contain characteristics different to other data types. In social sciences, for example variables change in time but with a low dynamic, while in drinking water data changes happen at every point in time. In other fields such as groundwater, there is the possibility to link data to a physical model of aquifer(s). This is also possible for WDN, where the physics are known. However, the type of variables of the two water systems under study are dissimilar in magnitude and time resolution. Processes in groundwater tend to develop during years, while in WDN extreme events can happen within minutes. In general, as it will be presented, the case studies belong to only two categories water quantity and quality, so many of the techniques contained in social sciences and other water resources systems may not be applicable in short time.

2) Because there are large differences in the number of available data samples from the different water utilities and it is not possible to customize the protocol for each of them. A generic protocol was therefore developed and applied across the cases.

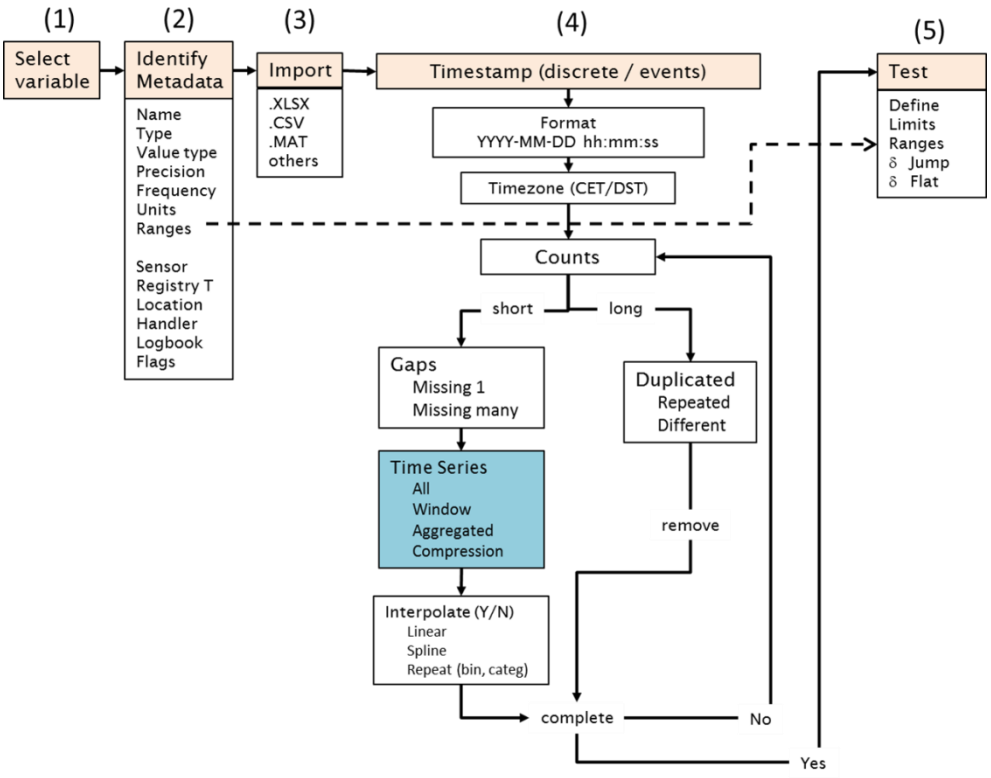
We consider here anomaly or faulty data detection using simple tests. Such tests are known as: i) variable bounds, ii) physical limits, iii) determine jumps (shift or sudden change in trend), iv) flat values and v) errors in the timestamps.

The step-by-step protocol for data validation is presented in FIGURE 11.

Step (1) corresponds to the selection of a variable. It should be noted here that the data validation is performed individually for data point of each time series.

After variable selection, some tasks are to be done by the operator, by gathering or collecting other data related to the variable of interest. This corresponds to step (2) Identification of metadata. This may become a cumbersome task in case that many variables need to be analyzed, however this needs to be performed only once. Also, some of the values tend to be available by the utilities, such as the sensor type, registration time step, location of the sensor, handler of the sensor at the water utility and its associated logbook. One important feature of the metadata collection is the determination of flags. For some utilities, the system automatically identifies anomalous data and provides different *Flags*, once the data is fetched from the data warehouse or SCADA system. For other utilities, this is currently under development. Flags represent a quality metric of data samples, not a quantitative metric.

FIGURE 11. STEP BY STEP PROTOCOL OF DATA QUALITY CONTROL APPLYING SIMPLE TESTS



Step (3), once the data has been inventoried and metadata has been added it is necessary for the utilities to decide the format in which the data will be presented. This has the advantage that once an agreement has been made for the data format at the water company, exchange of data across different levels of the informatics chain (see Figure 5) will be easier to overcome.

Step (4), requires an in-depth analysis of the timestamps and time format with which the selected variable is presented. In some cases data is missing for some periods of time, so a decision to whether or not to interpolate must be made (Graham 2009) (Helsel and Hirsch 2002). In other cases, due to sensor malfunction duplicates are observed. If that is the case a decision must be made in terms of which data to remove. This is not a trivial task and to our

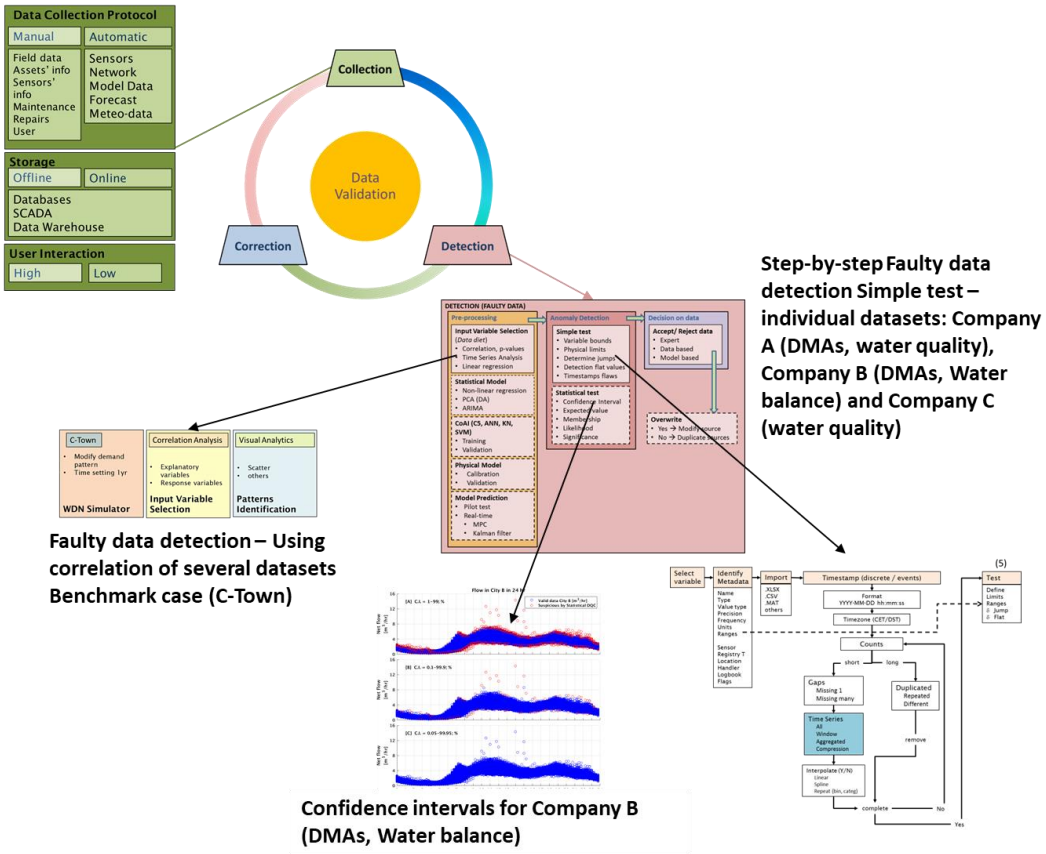
knowledge with the data presented by the utilities, many errors of this kind are still present in the raw data delivered. Of notice here, is also that data aggregation and visualization are throughout the utilities still necessary for the identification of faulty data.

Step (5), once the time series is complete, then it is possible to proceed to the simple tests of data validation (see FIGURE 11).

Although we do not consider data correction, we have performed additional queries to the water companies with the idea of using the proposed protocol for data validation and to be able to evaluate its possible application or additional requirements.

Finally, FIGURE 12, shows the framework for data validation applied to different case studies in this research, see chapter 4.

FIGURE 12. FRAMEWORK OF APPLICATIONS OF DATA VALIDATION IN CASE STUDIES.



4 Case studies

4.1 Overview of the cases and selection of the techniques

To test a number of data validation techniques, two problems were identified for their application. In the field of drinking water distribution, the applications are: anomaly detection in volume flow rate, and in the field of water quality anomaly detection in datasets of temperature, turbidity, pH and chlorine. In TABLE 4, the overview of the cases and the data validation techniques are presented, the available datasets, locations, which analysis were performed for each utility.

TABLE 4. OVERVIEW OF DATA AND TECHNIQUES USED FOR EACH CASE STUDY.

Company	Datasets	Remarks	Action	Current validation	Techniques
Company A	Water quality WWTP I Production	One location: Temperature, pH, Turbidity. Company A has a built up system that labels the data with different Flags.	Compare identification of anomalies with own system and with registration of maintenance activities or reported incidents	Data is validated automatically by the system	Simple test Confusion matrix
	Flow City A Pump Stations	Large scale City A. 5 pumping stations. Company A has a built up system that labels the data with different Flags.		Data is validated automatically by the system	Data aggregation Regression Knowledge based
	Energy City A Energy provider	Large scale City A. 5 pumping stations. Data comes from a third party. No Flags available for this data.		Data form a third party not validated by Company A.	Data aggregation Regression Knowledge based
Company B	Flow (DMA) City B Residential	Contains information of Flow meters in the boundaries of the DMA. Positive and negative flows. Also data from customers inside the DMA. No Flags for faulty data are available.	Estimate an analysis of water balance with the recently installed flow meters	Data is part of pilot and it is not yet validated	Data aggregation Simple test Knowledge based
	Flow (DMA) City H Industrial	Contains information of Flow meters in the boundaries of the DMA. Also data from customers inside the DMA. No Flags for faulty data are available.	Estimate an analysis of water balance with the recently installed flow meters	Data is part of pilot and it is not yet validated	Data aggregation Knowledge based
Company C	Water quality City Bk	Short data sets of one location for Temperature, pH, Turbidity and Chlorine.	Test approach proposed for Company A on Chlorine	Data is part of pilot and it is not yet validated	Simple test on Chlorine data.

Company	Datasets	Remarks	Action	Current validation	Techniques
	Flow (DMA) City Db	Large scale - Complex DMA	Make inventory of the preconditions to perform data validation	Data is part of pilot and it is not yet validated	N/A. The dataset has many
C-Town (hypothetical benchmark system)	Simulation run of an hypothetical system	Synthetic small scale WDN with a single source, and multiple boosters. System split into 5 different DMA's. Combination of hydraulic	Elaborate analysis of input variable selection	Not applicable. Synthetic case.	Correlation and input variable selection.

The data provided by the water utilities contains several differences. In terms of variables water quality data tends to be more homogeneous, with the particularity that Chlorine data is available for Company C. Data resolution was also variable and depends on the type of registration for each utility. It can vary in minutes, quarters (15min), events (as a significant change occurs), pulses. Time stamp registration is also very heterogeneous across utilities, dates can contain summer and winter time as number or other indicator (+1.00 or +0.00), some information is mechanistically absent on the same timestamps, most likely due to data communication. In this regard, at 00:00, some utilities contain no data, indicating that the data transfer is most likely to occur at this time at night. Data from one utility was provided as a Last In First Out (LIFO) format (reverse dates) for some variables, while the same utility provided a more common First In First Out (FIFO) format. Length of time series was also very variable as it was not possible more than a few months of data for some utility, where some sensors have recently started to send live data. Due to the high variability in data types and content it was not possible to perform a *one size fits all* analysis of data validation and more specifically of faulty data detection, so the focus given to specific datasets is variable to present a broader set of analysis within this BTO. In each specific case study, the data used and the test applied are presented. Taking advantage of feedback sessions, it was possible to implement expert knowledge in the determination of faulty data of 2 utilities.

4.2 Company A

During the HI Pilot, Company A has contributed with data from the treatment plants L1 and L2. It corresponds to data form the filtration processes. A total of 4 time series and 3 different variables (i.e. Temperature (1), pH (1) and Turbidity (2)) was delivered by the utility, although it currently monitors 73,000 variables (simultaneously) for drinking water and 23,000 for wastewater. Data collection corresponds to time series in the period between 1st January 2016 and 31st December 2017. According to our visit to the facilities of Company A, it is also possible to fetch directly from their data warehouse values interpolated at different resolutions. However for the time being a resolution of 1 min for all variables was selected for further analysis. Data is presented in its raw form as an event-driven collection, this means that once there is a variation in the measurement in any of the sensors above a certain threshold, a new sample (row) was stored for a particular sensor variable. Due to this, the total length and number of timestamps collected vary among variables. Data contains 4 fields (columns), (A) the sensor ID, (B) the time stamp, (C) the measured value and (D) a status of signal's health, established by the system. Such DQC is split in four different categories as flags: (1) Good data - "Goed", (2) Faulty data - "Slecht", (3) Dubious data - "Doubtful" and (4) Out of range - "Buiten". It was not possible to determine the specific rules which drive the definition of different categories as they are automatically triggered by the system of Company A.

To this particular data set only simple tests were performed in order to estimate the accuracy of the identification of the DQC proposed here. It was possible to identify most of the faulty data, without previous knowledge of the system rules. However, in some cases with the simple tests, some additional "likely" faulty data was identified among the time series. This does not mean that the statuses provided by Company A are not reliable enough, rather than one of the detection rules presented here (i.e. flat value detection) may not be currently implemented inside their data warehouse.

4.2.1 Company A data

Data was collected from 2 sources the Pumping Stations (PS) for the whole system and Water Quality (WQ) data at Waste Water Treatment Plant (WWTP) I and WWTP II. Data from the pumping stations was analyzed only for the identification of faulty data of timestamps and consistency of status identification by Company A. Data from WQ corresponds to only 4 out of thousands of sensors in the treatment plants, but represent examples of a comparison between current Company A data status and our proposed protocol.

Water quality data at WWTP I & II

Given the existence of a verification system at Company A, only 4 different data validation tests were performed:

- Verification of boundaries
- Verification of timestamps
- Verification of flat values
- Verification of jumps

The confusion matrices (TABLE 6) present the comparison between the observed faulty data in the raw data collection and the ones identified by applying the 4 verifications.

TABLE 5. NUMBER OF FLAGS PRESENT IN WATER QUALITY DATA FROM COMPANY A

		pH		Temperature		Turbidity	
Treatment plant	Acronym	(-)	%	(C)	%	(%)	%
WWTP I	L01	25	0.00	33	0.00	30	0.00
WWTP II	L02	N/A	-	N/A	-	67	0.01

N/A: Not available

For each time series the corresponding number of flags identified in the data is presented in TABLE 5. It is evident that there is a limited number of flags identified by the system. For that reason, to this particular data set, simple tests were performed in order to estimate the accuracy of the identification of Company A flag system. This is indeed a cumbersome task for Company A as to our knowledge more than 73,000 variables are updated every minute by their system only for drinking water.

TABLE 6, presents the confusion matrix for the different variables. *This report* must be understood the data quality control performed with simple rules, while Company A implies the flag status obtained from reading raw data which was submitted. If any timestamp sample is identified as faulty data by any of the simple tests of this report, then data is considered faulty. When both DQC schemes agree in the identification a Yes-Yes coincidence is identified. This can be understood as a validation of Company A's flag identification.

For a perfect coincidence among the two analyses a Yes-Yes cell must contain all faulty data flags for both DQC schemes. In the case that Company A's flag data system is unable to identify a feasible faulty data compared to ours, a No-Yes coincidence is identified. There is the possibility that certain rules are not implemented inside Company A flag data system, but this can only be hypothesized. Absence of knowledge of the algorithmic scheme for identification of faulty data at Company A is then a constraint.

It is also of interest, that for all variables the number of feasible faulty data points identified is larger for this report analysis than the number of flags in the raw data from Company A.

This could be seen in two ways. First as this report data validation is more compact and sensitive to faulty data, and second there could be more tests which can be automated as flag data for water quality by Company A's system.

In the first case, this is a disadvantage for the operatives, as this will translate to a large number of verifications required. For example, Temperature data in WWTP I, confirms more than 1200 flags require verification during a 2-year period, or almost 2 flags per day. As it is now data validation for Company A is already cumbersome, so the possibility of performing such task for more than 73.000 variables seems impossible to come to reality. A calibration process of the parameters for each of the variables requires to be performed in individual basis.

In the second case, it is possible that this research project data quality control has identified additional faulty data. Although this may sound controversial, some examples are presented to discuss the reliability of DQC of Company A.

TABLE 6. CONFUSION MATRICES OF WATER QUALITY DATA FOR COMPANY A'S AND THIS REPORT on RAW DATA

pH		This Report		
		Yes	No	Total
Company A	Yes	25	9	34
	No	254	-	
	Total	279		

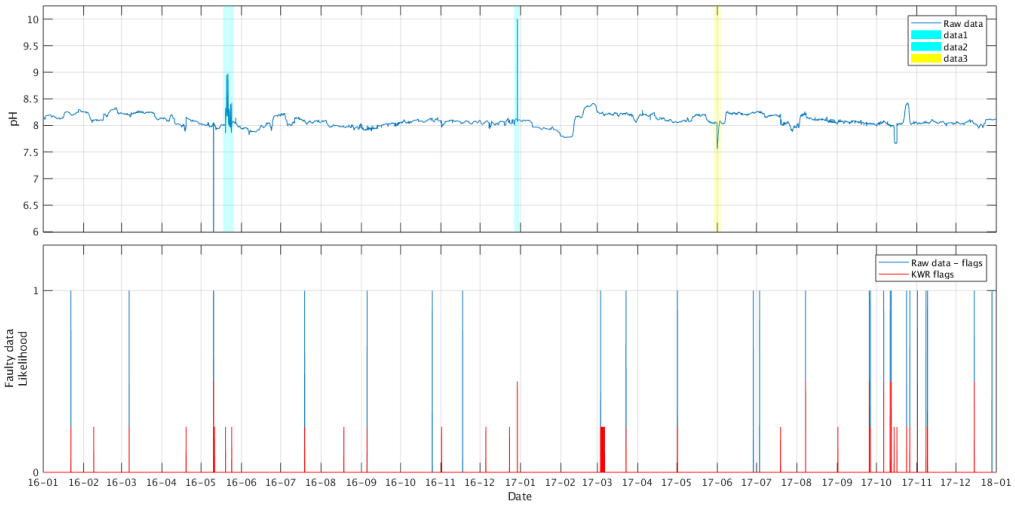
Temperature		This Report		
		Yes	No	Total
Company A	Yes	29	4	33
	No	1250	-	
	Total	1279		

Turbidity II		This Report		
		Yes	No	Total
Company A	Yes	61	6	67
	No	187	-	
	Total	248		

Turbidity I		This Report		
		Yes	No	Total
Company A	Yes	20	10	30
	No	39	-	
	Total	59		

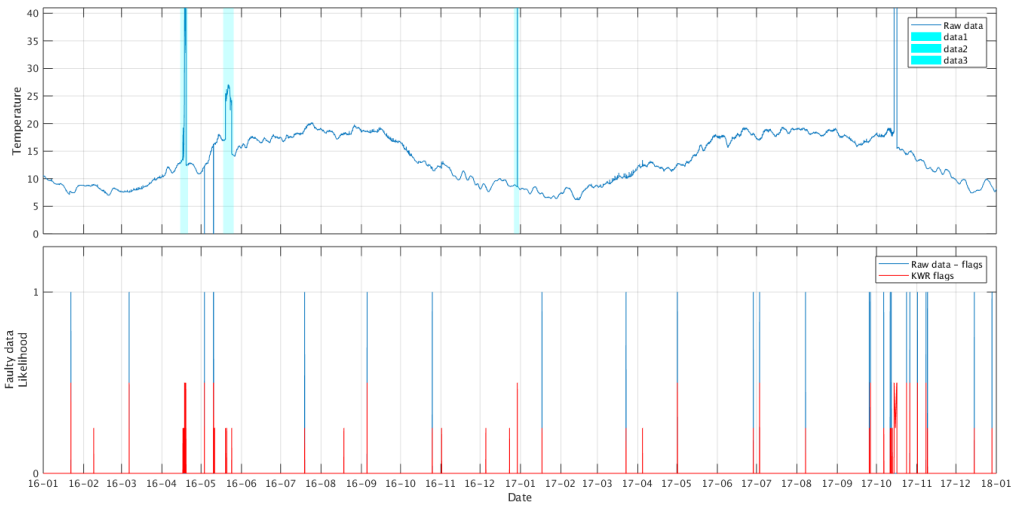
FIGURE 13, presents the time series for pH at WWTP I. The range of variability of pH is quite small due to the need by Company A to keep its magnitude within a narrow band. However, there are some spikes present in the data which are identified both by Company A's system and KWR's. Also present in this figure, in *cyan*, time windows in which KWR's DQC identified feasible faulty data (No-Yes), as May 2016 and January 2017. In *yellow*, time windows in which there was no identification by any system (No-No) but visual inspection shows that there is feasible faulty data as in June 2017.

FIGURE 13. TIME SERIES OF PH IN WWTP I. SHOWING ALSO SYSTEM FLAGS AND THIS REPORT DQC.



In FIGURE 14, time series of temperature in WWTP I is presented, in this case most of Company A's system flags are captured by the simple analysis by KWR, and indeed three time windows in which the possibility of faulty data was identified are presented. Such time windows are: centered in 2016 around April 18th , May 22nd and December 29th. Although there are other faulty data windows in a sense both systems identify the faulty data as in October 13 of 2017.

FIGURE 14. TIME SERIES OF TEMPERATURE IN WWTP I. SHOWING ALSO SYSTEM FLAGS AND THIS REPORT DQC.



In the case of turbidity in is presented in FIGURE 15 (in logarithmic scale). In this case the most relevant feature for DQC is that the time series presents jumps at different values. Such jumps (drifts or changes in variance), occur during 2016, around July 1st, and in 2017 around 3rd of April and 9th of August. This can be due to a modification in the operational conditions of the treatment plant or more locally in a filter, however without further information it was not possible to elaborate a robust hypothesis on this change of behaviour at this location.

A duplicate analysis for the same variable, this time at WWTP II, (see FIGURE 16, vertical axis in logarithmic scale) shows that Company A's flag system tends to allow higher values of turbidity as normal events. An example is the spike in 2016, during April 1st. This could have an operational reasoning, however it is not identified in the logbooks provided by the utility.

FIGURE 15. TIME SERIES OF TURBIDITY IN WWTP I. SHOWING ALSO SYSTEM FLAGS AND THIS REPORT DQC.

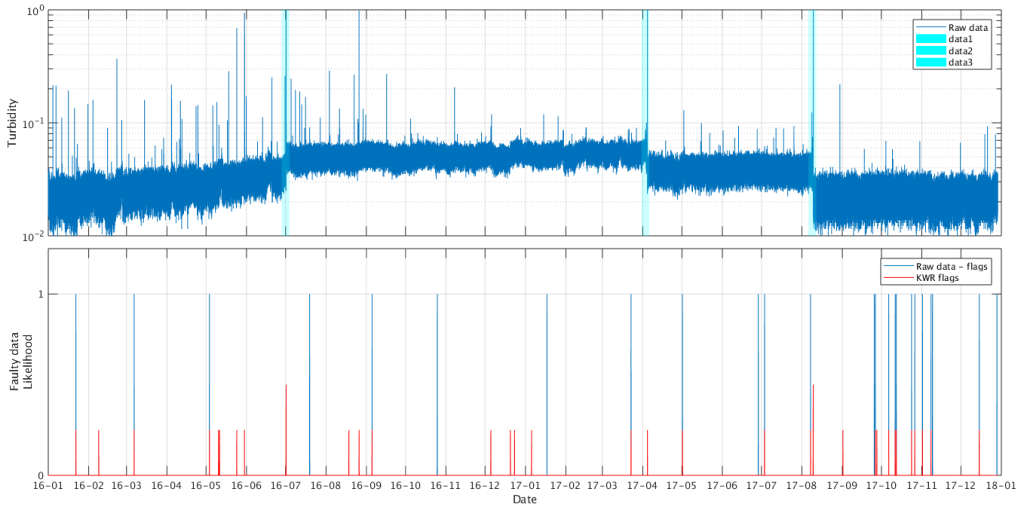
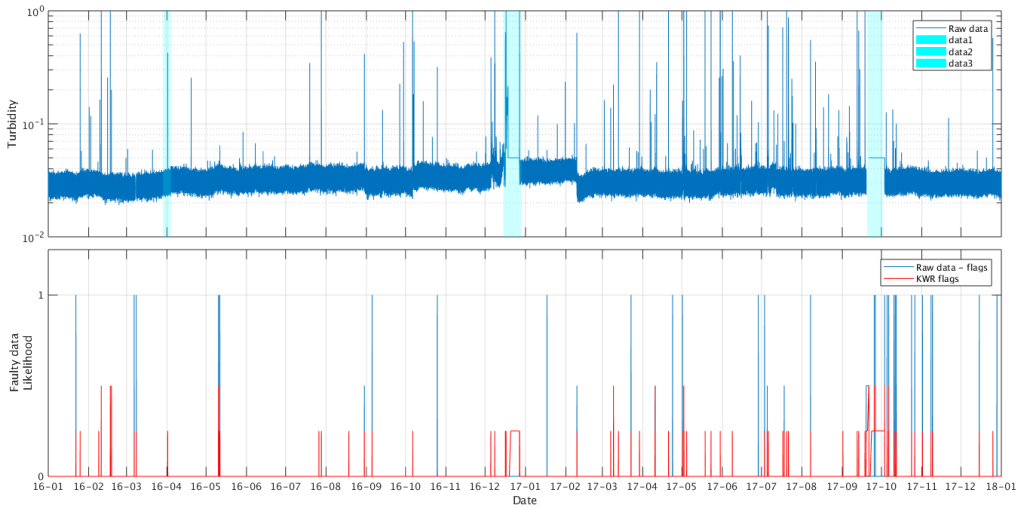


FIGURE 16. TIME SERIES OF TURBIDITY IN WWTP II. SHOWING ALSO SYSTEM FLAGS AND THIS REPORT DQC.

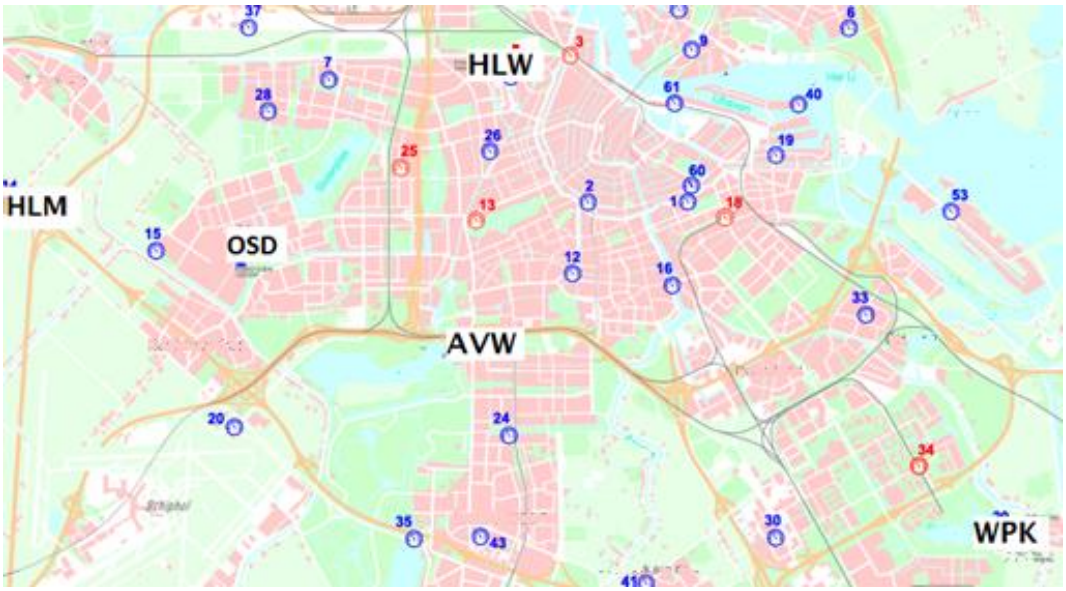


On the other hand there are some time windows in which the simple tests may simply identify plateau values registered in the raw data, while Company A's system was not able to do so (see FIGURE 16). The time windows are in 2016 around December 22nd and in 2017 around September 26th.

Pumping stations data

There are a total of 5 Pumping Stations (PS) in City A system: WPK, AVW, HLW, OSD and HLM (see FIGURE 17).

FIGURE 17. GENERAL LOCATION OF PUMPING STATIONS IN CITY A (SOURCE: COMPANY A)



For each one of the them data of pressure, flow and energy use was available. Pressure (kPa) and flows (m^3/h) have a time resolution of approximately 1 minute while energy use (kWh) contains data with a time resolution of approximately 15 minutes. The available data, covers the period between 01 January 2016 and 31 December 2017 (included). Data delivered by the utility contains the system's flag for DQC. The flags are categorized from 0 to 4. According to the contact person, flags of 0 and 4 can be considered as *Good* data, while flags of 1, 2 or 3 are considered as *Faulty* data.

The time series of the flows are presented in FIGURE 18, the pressures are presented in FIGURE 19, and the energy use in FIGURE 20, discriminated for each pumping station. In such figures, the flag status of the utility (Company A) is presented as red lines in the corresponding time stamp. On the contrary, If data is considered as valid or *Good*, the red line has a value of zero.

TABLE 7. NUMBER OF FLAGS AND PERCENTAGE FROM TOTAL OF TIME STAMPS FROM RAW DATA PROVIDED BY COMPANY A

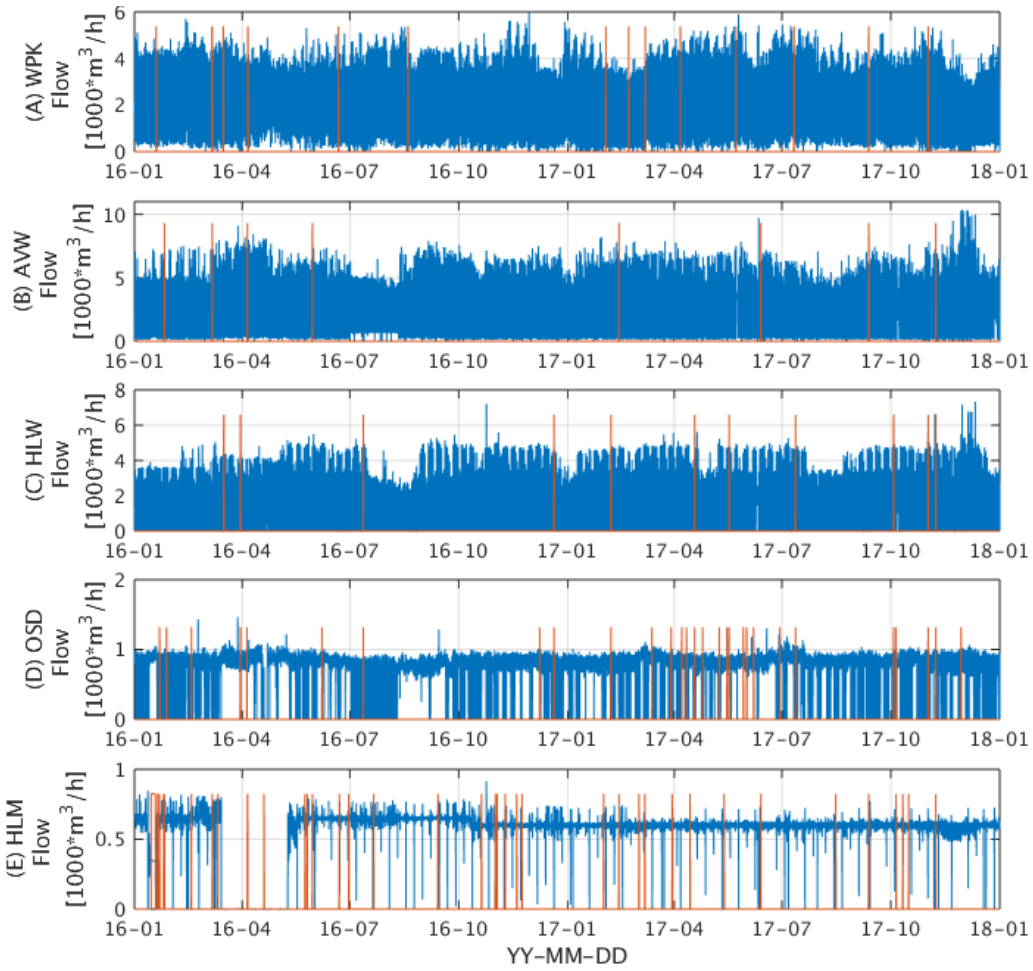
	Flow		Pressure		Energy*	
Pump Station	**(m^3/hr)**	**%**	**(kPa)**	**%**	**(kWh)**	**%**
WPK	475	0.05	474	0.05	0	0.00
AVW	266	0.03	68	↓0.01	0	0.00
HLW	41	↓0.00	311654	↑29.63	0	0.00
OSD	299	0.03	228	0.02	0	0.00
HLM	6769	↑0.64	6769	0.64	0	0.00

* Energy does not contain reported anomalies in the data. ↑ indicates highest percentage of anomalies. ↓ indicates lowest percentage of anomalies.

The total number of flags identified in the raw data for each time series and variable are presented in TABLE 7. It is of notice that no anomalies are identified by the system or provided

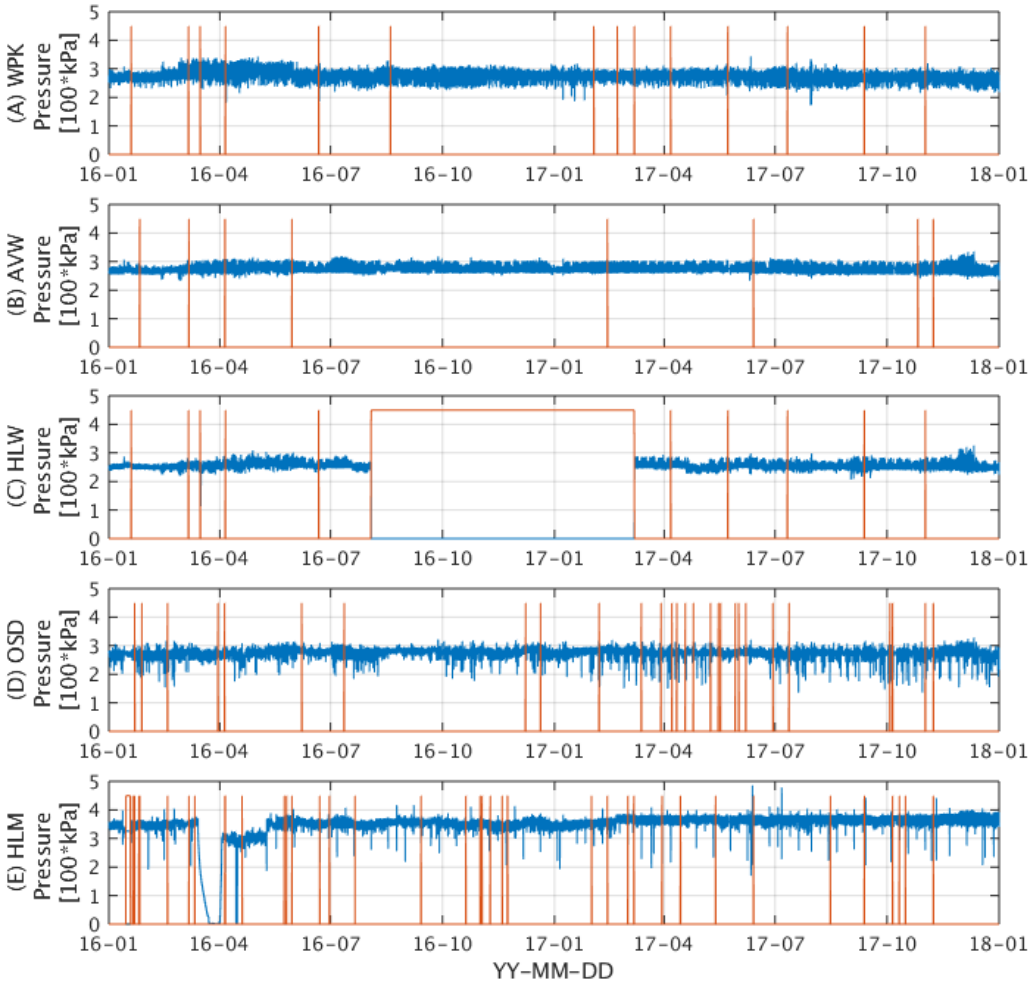
by the utility for energy use of any pump station as such data is provided by a third party handling the energy consumption of the utility.

FIGURE 18. TIME SERIES OF FLOW AT 5 PUMPING STATIONS OF COMPANY A. FAULTY DATA REPORTED BY COMPANY A DISPLAYED WITH RED LINES. VERTICAL AXES ARE DIFFERENT TO ALLOW VISIBILITY OF TIME SERIES.



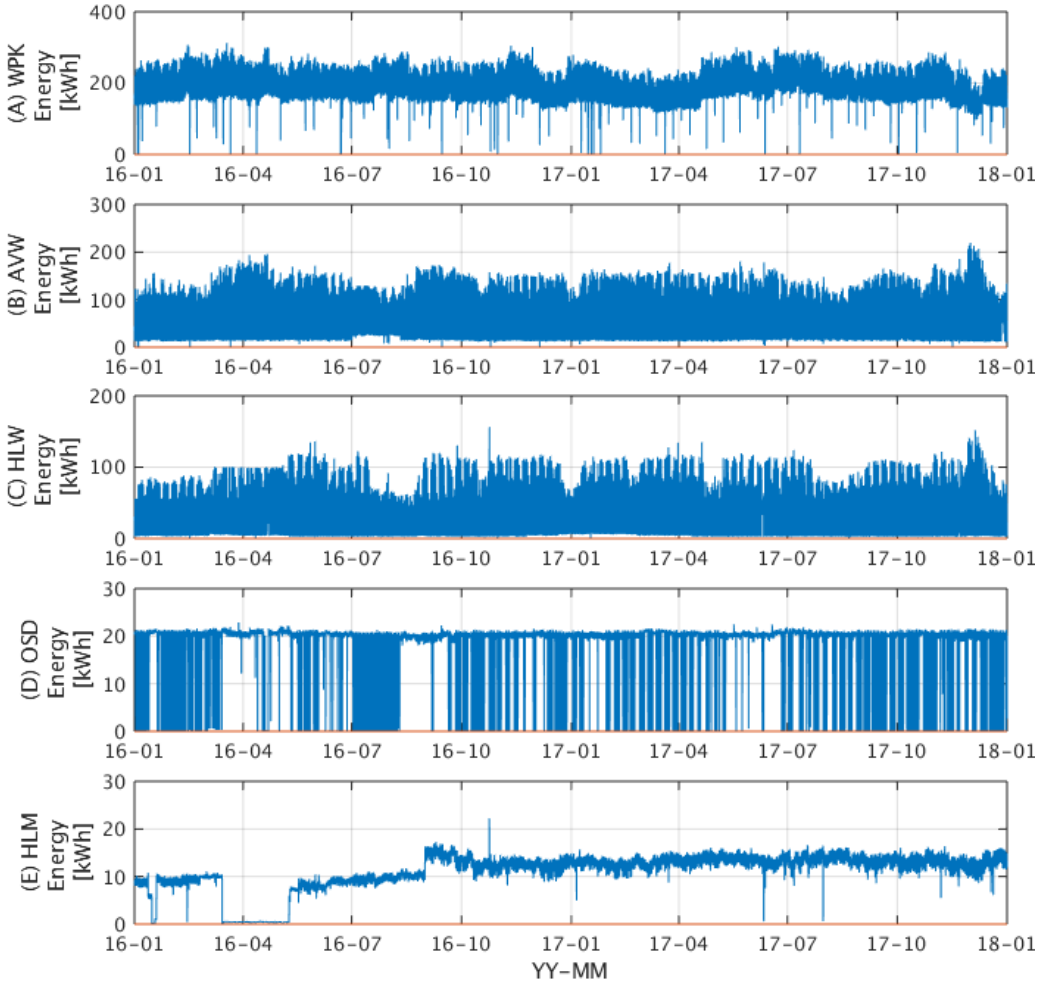
For flow time series (see FIGURE 18), the largest amount of flags is found at PS OSD (6769 or 0,64% of the TS). The lowest number of flags is observed in the PS HLW (41 or 0.004%). In the case of PS HLM (see FIGURE 18E), in the period between March 2016 and May 2016, there is no data of flows and this drop is not identified by the system flags. This is similar to what may be observed for the pressures at the same PS.

FIGURE 19. TIME SERIES OF PRESSURES AT 5 PUMPING STATIONS OF COMPANY A. FAULTY DATA REPORTED BY COMPANY A DISPLAYED WITH RED LINES.



For pressure time series, the largest amount of flags is found at PS HLW (311654 or 29,6% of the TS), in the period comprehended between August 2016 and March 2017 (see FIGURE 19C) where data is not completely absent, but presents data with values close to zero.This flags are correctly picked by the utility's system. The lowest number of flags is observed in the PS AVW (68 or 0.01%). As a side note, for PS HLM, in the period comprehended between March 2016 and April 2016, there is a drop in pressure which is not flagged. It is a steady process, with no sudden jumps on data, or variance changes. This occurred simultaneously for flow and pressure, given that there was a renovation taking place in this pump station. This demonstrates the relevance of performing data validation with respect to logbooks of operational activities and maintenances in the utility.

FIGURE 20. TIME SERIES OF ENERGY USE AT 5 PUMPING STATIONS OF COMPANY A. NO FAULTY DATA REPORTED BY THIRD PARTY.



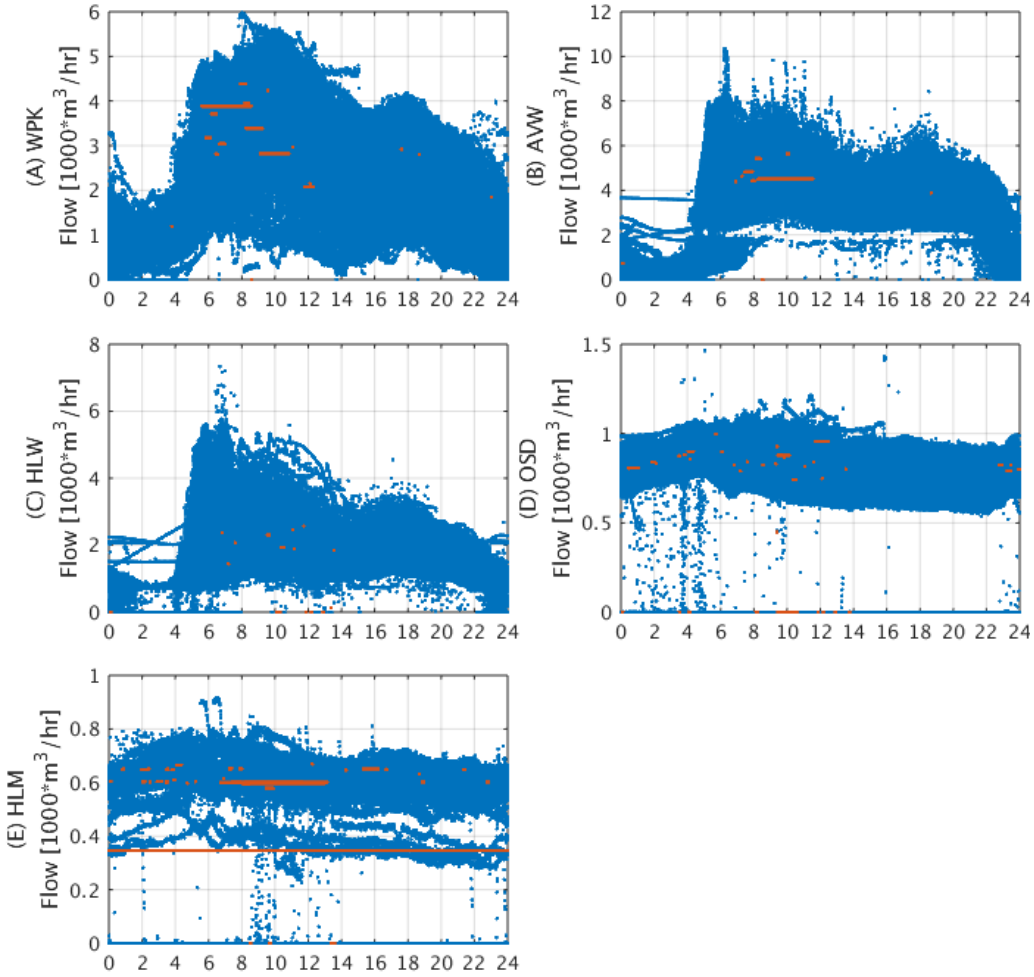
For energy time series, there are no flags identified by the system. The data is provided by a third party. Energy use at PS HLM (FIGURE 20E) displays drops during extended period between February, March, April and May 2016. After feedback with the utility this period corresponds to a maintenance of the pumping station.

In addition for OSD PS (FIGURE 20D), raw data contains a large number of pump switches (values = 0). This can indicate no register or that indeed the pumps are shut off. However, this is unlikely as the data represents an integrated value during a 15min interval.

Subsequently, data from each time series has been processed to identify if there are particular periods, throughout a daily operation (in 24 hr), when faulty data is more likely to occur. For this reason, data has been rearranged and categorized as hourly data, disregarding the dates. Such analysis is presented in FIGURE 21 for flows, FIGURE 22 for pressures, and FIGURE 23 for energy use.

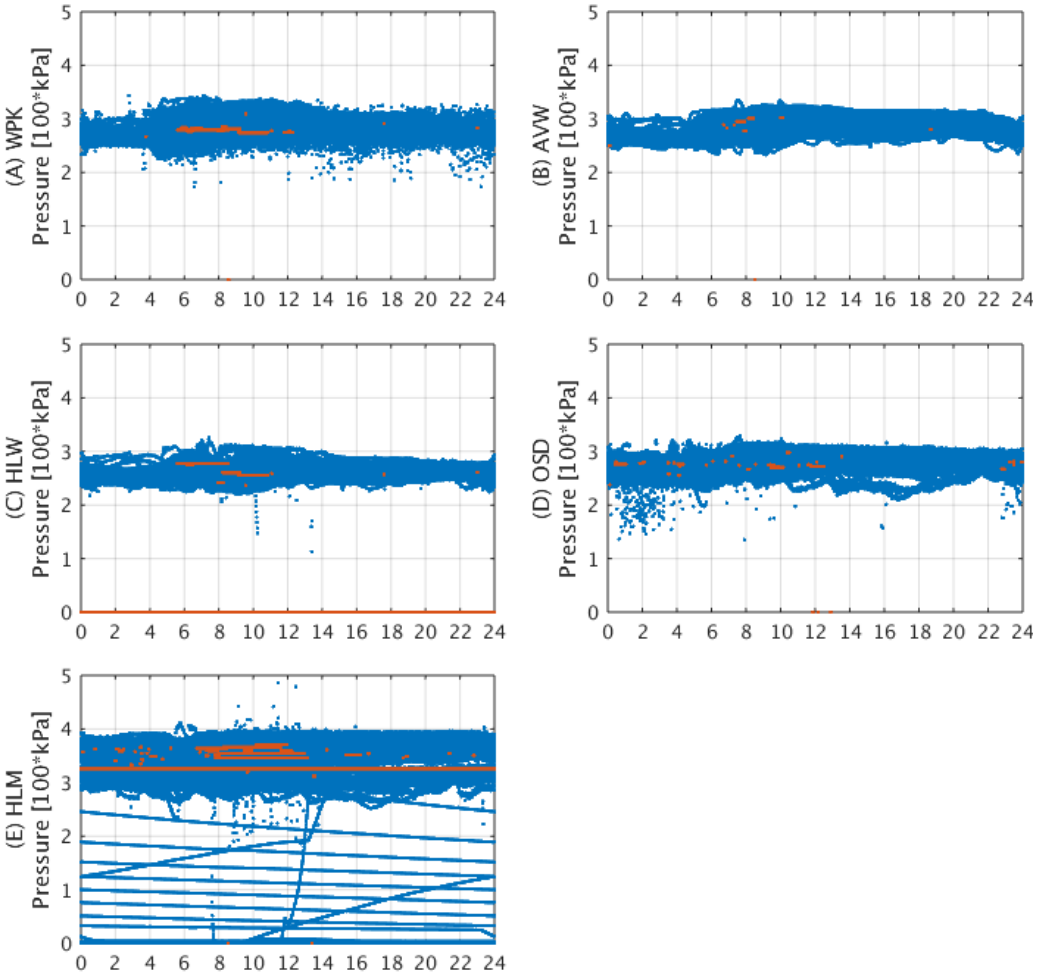
Due to these figures being a similar analysis of time series, in all figures red dots represent the same faulty data flags on raw data by Company A (TABLE 7).

FIGURE 21. SCATTER OF FLOWS AT EACH PUMPING STATION DURING A DAY. INCLUDES ANOMALIES FROM COMPANY A (RED DOTS).



In the case of flows, most of the flags are present during the peak consumption hours. This behavior is similar to what is observed for pressures. However, the faulty data detected does not correspond to high or low values either, but to intermediate ones. In the case of PS WPK (see FIGURE 21A), there are a 3 values constantly picked by the system as faulty data near 4000, 3390 and 2800 m^3/hr. In the case of pumping station AVW (see FIGURE 21B), there is a value constantly picked by the system as faulty data near 4510 m^3/hr. In the case of pumping station OSD (see FIGURE 21D), the predominant faulty detection value is 0 m^3/hr between 9 am and 11 am. In the case of PS HLM (see FIGURE 21E), there is a value constantly picked by the system as faulty data near 346 m^3/hr (all day) and near 600 m^3/hr (between 6:30 am and 13:30 am). Once again, this is consistent with a disruption of the system. After a feedback session with the contact person from Company A, it was discovered that this specific values for which data is identified as faulty corresponds to the period when data is fetched to the data warehouse, and it is flagged as faulty by their system. The explanation is due because there is always a delay for the data transmission, triggering the flag in the system.

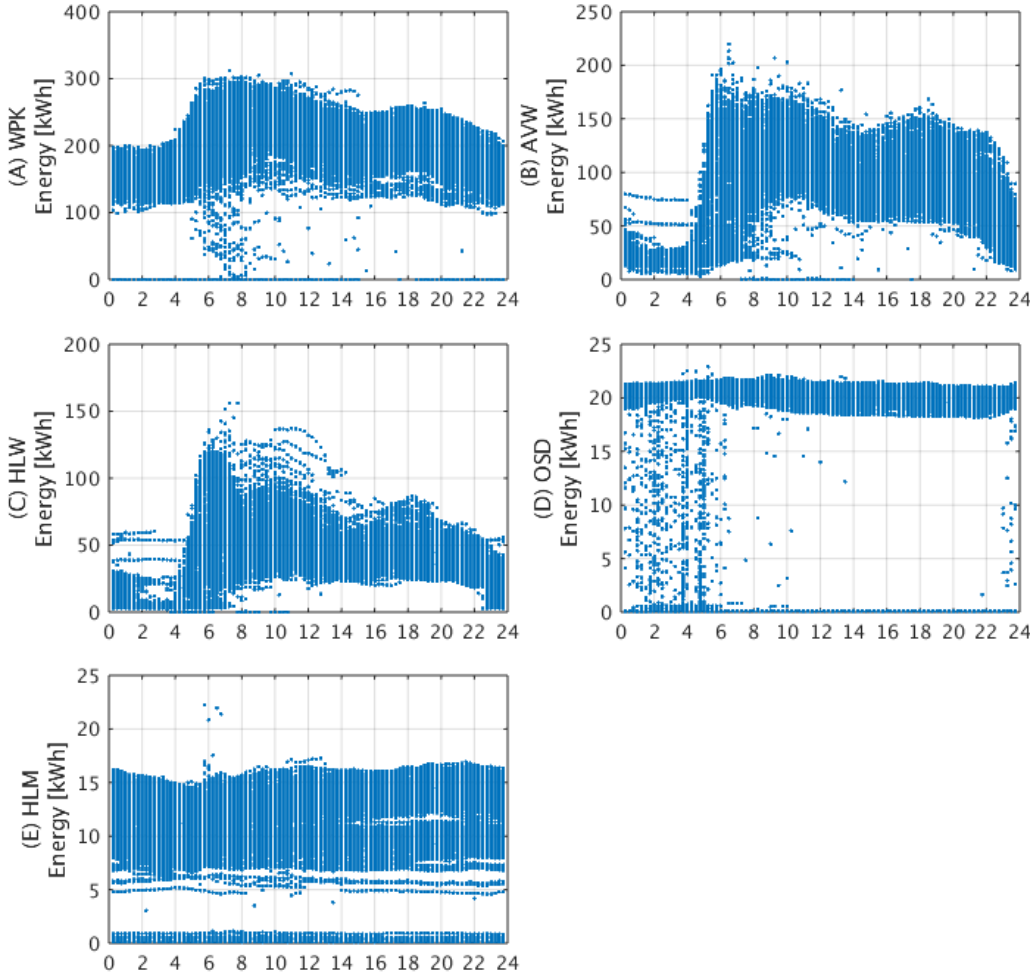
FIGURE 22. SCATTER OF PRESSURES AT EACH PUMPING STATIONS DURING A DAY. INCLUDES ANOMALIES FROM COMPANY A (RED DOTS).



In the case of pressures, the system tends to pick faulty data more often during the peak consumption hours of the morning, with the exception of HLW where there is a huge gap of data which is picked as near zero values (see FIGURE 22C). In the case of HLM (see FIGURE 22E), there is a value near 325 kPa which is constantly being identified as faulty data throughout the day. This may be consistent with a disruption of the system or with a known operational setup. Indeed, there was no possible explanation to be found on the identification of such data as faulty by the system.

The values below the average trend in HLM (see FIGURE 22E) show a steady increase (or decrease) in the pressure for some days. The possible explanation obtained for this behavior is that such data samples represent an specific event HLM refurbishment which is visible in the period of March to May 2016 in the time series of flows and pressure for the same station (see FIGURE 18E and FIGURE 19E)

FIGURE 23. SCATTER OF ENERGY USE AT EACH PUMPING STATIONS DURING A DAY. NO ANOMALIES REPORTED BY THIRD PARTY.



The variation of energy consumption during the day shows that the 3 larger pumping stations (WPK, AVW and HLW) have a similar pattern to the one of flows. The highest range of variability of energy use during the day is present for WPK between 5:00-9:00 am (see FIGURE 23A), midnight to 5:00 am for AVW and HLW (FIGURE 23B and FIGURE 23C). For OSD (FIGURE 23D), there is a large variability of energy consumption from 11:00 pm and during the following 6 hours of the day. In the case of HLM (FIGURE 23E), the existence of gaps in the data (previously discussed) creates two different levels of energy consumption. A near zero level and what can be called a regular or central pattern.

In the case of flows, once the anomalies have been identified for each time series, a tag has been assigned to the data if any of the pump stations for a particular timestamp contains an anomaly. Subsequently the data from all the flows has been aggregated to estimate the total flow delivered by Company A. The time series of accumulated flows itself will not provide additional information, as the anomalies have been removed. For that reason a bivariate probability distribution of the data in time every 15 min (horizontal axis) and with a flow resolution of 250 m^3/h is elaborated (vertical axis). The obtained result of the bivariate probability analysis is presented in FIGURE 24A, where the color represents the probability of total flow in the city of City A (and surroundings) at a certain 15 min interval during the day. Red values represent a higher probability while gray values correspond to low probability. In fact FIGURE 24A, is a representation of the flow pattern of the city.

FIGURE 24. BIVARIATE PROBABILITY OF (A) TOTAL DEMAND CONSUMPTION AND (B) TOTAL ENERGY USE IN CITY A. ANOMALIES HAVE BEEN REMOVED.

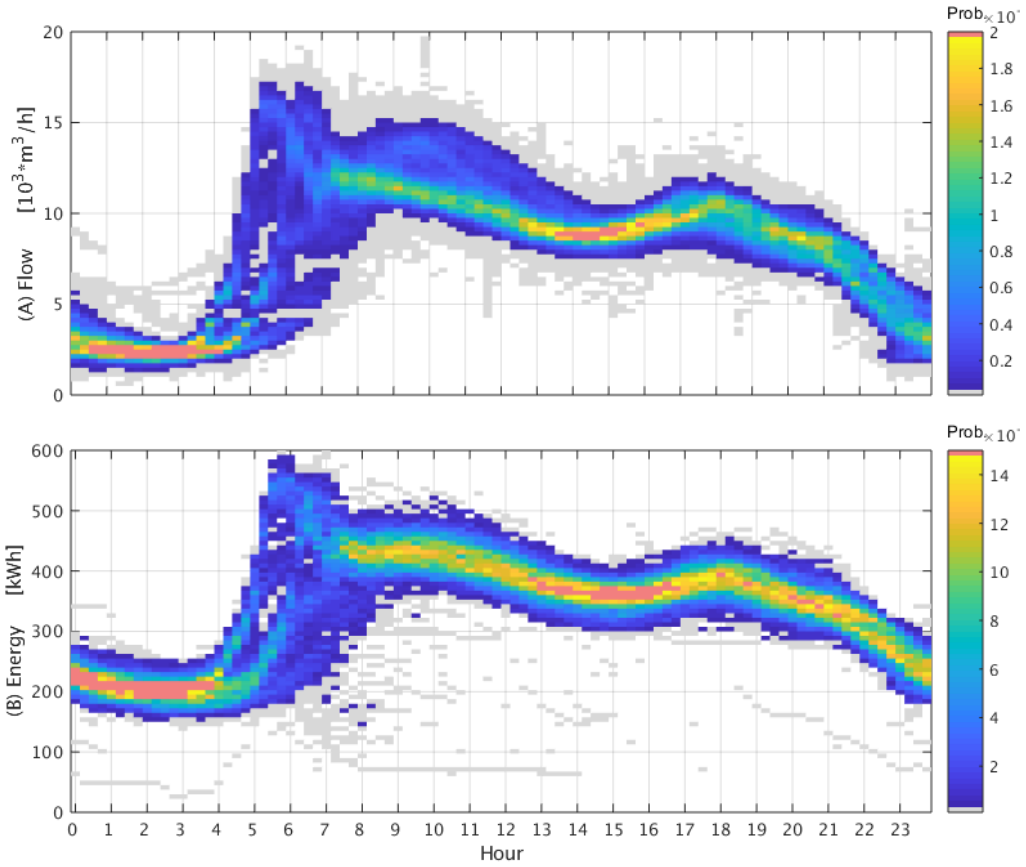


FIGURE 24A, shows a double peak consumption (as expected) during the early hours of the morning (05:00-09:00) and during the dinner time (17:00-19:00). On the other hand the Minimum Night Flow (MNF) occurs after midnight (01:00-04:00) with a high probability.

Another output which can be obtained from a data validation perspective is the fact that some samples correspond to a large demand consumption around 10:00. Although its probability is low, it is evident that such events have occurred in City A during the last 2 years. In fact, the highest registered values for the entire system occur for this particular time of the day.

As a contrast some samples show that there has been a lower demand consumption than the average trend at 11:30. There was no possible explanation identified by the utility, on such atypical pattern variation.

One drawback of this analysis, is that the demand pattern may vary by season, and the number of rules required to identify faulty data may become untraceable. Second issue with this analysis is the limitation of scale in the data, meaning that it is not feasible to easily identify anomalies at HLM as in is for AVW, given that there is 1 order of magnitude between them. A faulty data value in HLM can be easily covered by atypical or extreme event occurring at AVW.

Another possible use of this analysis for data validation, is to perform a correlation analysis of flows with respect to the energy use obtained from the utility. In fact, as presented in FIGURE 24B, the energy use pattern follow obviously a similar trend than that of the flows. As a matter of fact, this is not distant from the current operation of the system as a single District Metered Area (DMA), or fully interconnected WDN.

Similar to what was done in the case of flows of the 5 pumping stations, an aggregation of energy use is performed for the whole system (see FIGURE 24B). Given that the energy consumption is already in a 15 min interval, it is not a burden to perform additional data and time stamp conversions for data validation. This may result in a different scenario for other water companies where energy may be collected at different time resolutions.

The obtained histogram of the bivariate probability distribution of energy use in City A is presented in FIGURE 24B. As it was the case of flows there is a double peak in the energy consumption (this is expected) and a higher probability during the MNF.

From a data validation perspective, it is of particular interest that in this case a larger number of timestamps occur with low probability for energy consumptions under the average pattern (gray values), particularly during the hours of 06:00-10:00. However, given that there is no indication of status or flag data for any of this time series, it is not possible to conclude whether this is due to data anomalies itself or to a regular operation of the system. As an hypothesis most of this low values below the trend of energy pattern, can be attributed in part to the fact that there is a huge number of pump switches for all PS' as it is presented in FIGURE 23. Of notice is also that such pump switches identified in the energy consumption are not represented all the time in the flow and pressure data, mainly because pump switches occur at a 1 minute resolution, while energy is a cumulative variable stored every 15 minutes.

From a feedback session with personnel from the utility, it was possible to determine that such behavior of low energy registrations in the morning hours is due to the accounting of energy by the utility. Sometimes during the morning the energy for the system is provided from diverse sources, and the third party which delivers the energy data is not aware of such energy for accounting. In this case, expert knowledge of the daily operations and workings of the utility became far more relevant, otherwise all such data would be flagged as faulty by a data validation system.

Aggregated data and obtaining volume from energy

Given the available data from all the pumping stations, it was decided to perform a regression analysis between the aggregated values of volume, which were not considered as faulty data, as provided by Company A, and the energy use (from a third party) in City A the whole WDN system. There are 3 things that need to be done in order to obtain a linear regression which can reflect the majority of the data without including faulty data:

i) the flow data is verified and faulty data is removed for all pumping stations if at least one of them has identified the data point as faulty. For example, the extended periods where faulty data was identified in HLW and in HLM are removed only if flow data contains samples with flags indicating faulty data.

ii) Once this is done the data from flow in m^3/h is converted into volume m^3. The current flow resolution for Company A is 1min, so the data is divided by 60. This data is Volume every 1min. Before aggregation is performed, If there was a 15 min window with more than 3 missing values in WPK, AVW or HLW, any of them, that complete window was removed for regression purposes, given that it would imply a reduction of the total volume accounted for in City A as a whole of more than 2%.

iii) The data of Volume is aggregated at 15min intervals, to obtain the total volume used at the same time intervals of the energy use data available. The energy data is pruned also of the same timestamps for volume where faulty data is identified.

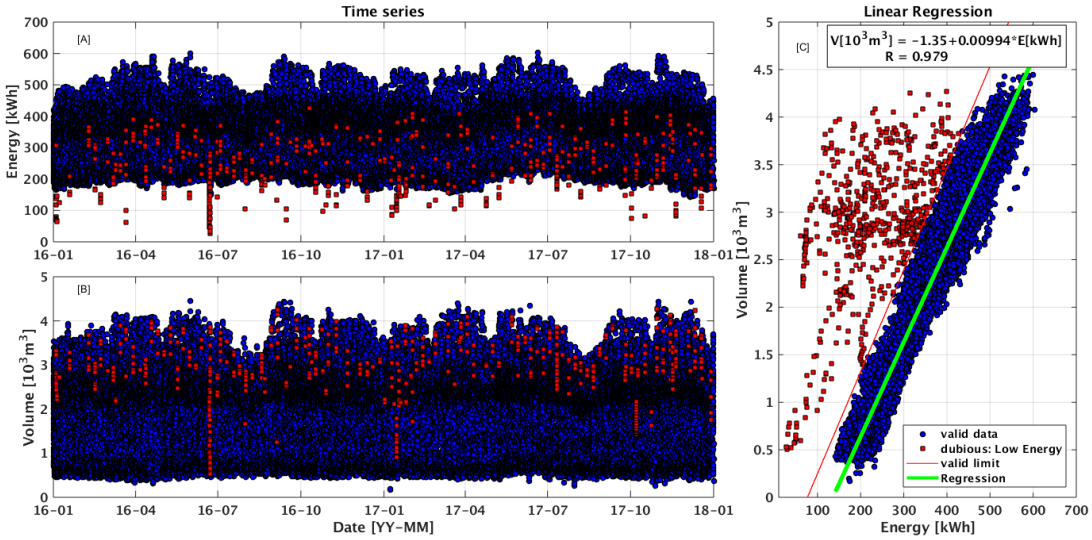
In FIGURE 25, the time series (A, B), and linear regression analysis between Volume and Energy (25C) are presented. Even after removing the faulty data samples, some data can still be classified as faulty data, as will be presented. It is shown that even after limiting the number of points which may create anomalies in this regression, there are still a number of points

(shown in red) which indicate different behavior from the rest of samples. The best regression in this case indicates that there is a relationship with a high correlation R = 0.979 between volume and energy use. The regression in itself indicates that:

$$Volume\ [10^3 m^3] = -1.35 + 0.0094 * E\ [kWh]$$

Such regression rose additional questions about the need of proper data validation for energy use. In this case, when the volume tends to be close to zero there would be a negative energy use, which is not possible. Although filling missing volume data based on energy use seems a good alternative, it is required to verify the energy data, which in this case, coming from a third party, is difficult to do. So it is recommended to install own energy/electricity sensors at crucial places in the drinking water infrastructure.

FIGURE 25. VOLUME VS ENERGY FOR CITY A. RED DOTS INDICATE ANOMALOUS DATA.



After a feedback session with the personnel from Company A, and further research by the BTO contact person at Company A, it was concluded that there is a reduced energy use for a portion of the data in FIGURE 25, above the central regression data. Such anomaly is caused by the way that energy is accounted for by Company A. In fact, a total balance of energy and its disaggregation into different components (e.g. treatment, distribution, and produced by Company A) requires of other additional variables not analyzed in this report. Being that the case, there could be different sources of uncertainty in a simple regression analysis as the one presented in this report for the utility.

Energy data thus becomes a source of yet another data validation process in itself, and demonstrates how important it is to perform data quality control across the different organizational levels, and for different purposes inside every utility. By looking blindly at the aggregated data of both volumes (production) and energy (consumption) it is not be possible to understand the registration of low energy values without the support of experts from the utility.

4.3 Company B

From Company B, most of the data retrieved corresponds to the DMA's of City B and City H. The two case studies are completely different and require a different analysis from a data validation perspective. The focus in this report is only in the analysis of data for water balance estimation at City B, which corresponds to a small DMA with residential consumption.

TABLE 8. DATA COLLECTED FROM COMPANY B FOR CITY B

Locations	Data Source	Variable	Units	Start Date			End Date			Time Resolution
Bal	DMA	Pos flow	m^3/h	2016	6	4	2018	3	18	15 min
		Neg flow	m^3/h	2016	6	4	2018	3	18	15 min
Hoo	DMA	Pos flow	m^3/h	2016	11	23	2018	3	18	15 min
		Neg flow	m^3/h	2016	11	23	2018	3	18	15 min
Koel	DMA	Pos flow	m^3/h	2016	6	4	2018	3	18	15 min
		Neg flow	m3/h	2016	6	4	2018	3	18	15 min
Koer	DMA	Pos flow	m3/h	2016	6	4	2018	3	18	15 min
		Neg flow	m3/h	2016	6	4	2018	3	18	15 min

For this there are 4 flow meters located in the boundaries of the DMA (FIGURE 26). The data collected for City B is presented in TABLE 8.

FIGURE 26. COMPANY B, CITY B. LOCATION OF FLOW METERS INSIDE THE DMA.

Meters are known as Bal, Hoo, Koel and Koer. For each of them, the time series corresponds to the last 2 years in 15 minutes intervals. In all cases, data is presented for flow (positive: going into the DMA) and backflow (negative: going out of the DMA). The corresponding time series of raw data are presented in FIGURE 27. It is evident that the data available is limited, given that the sensors have been installed recently the time series are short, but constitute a big step into the understanding of behavior in the DMA.

There are two issues that the utility must face for the estimation of water balance, from a data quality control perspective:

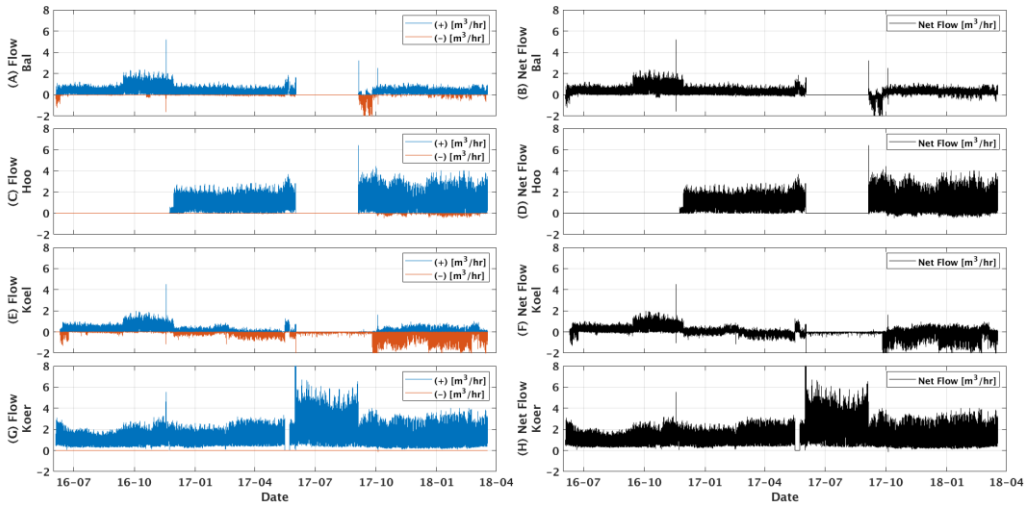
- *Lack of measurements inside the DMA's at a similar resolution*. Currently, only total customer consumption from households is known at a yearly basis, for more customers in the DMA. Only an average daily demand may be estimated for the DMA. The accuracy of such surrogate analysis contributes to the uncertainty of the water balance. In this case, the identification of anomalies poses a hurdle. It is not possible to differentiate faulty data in the water balance from a systematic increase of background losses. Although the Non-

Revenue Water of this utility is low (less than 5% obtained by communication with contact person), this lack of internal measurements shows that there is room for improvement in term of data redundancy.

- *Extended periods of no records*. The data from the flow meter sensors has extended periods of no availability. In fact, it is possible to establish that the area was not fully isolated, given that there is still this interaction with surrounding areas for the period of analysis. That is the reason behind a registration of positive and negative flows in the boundaries of the DMA.

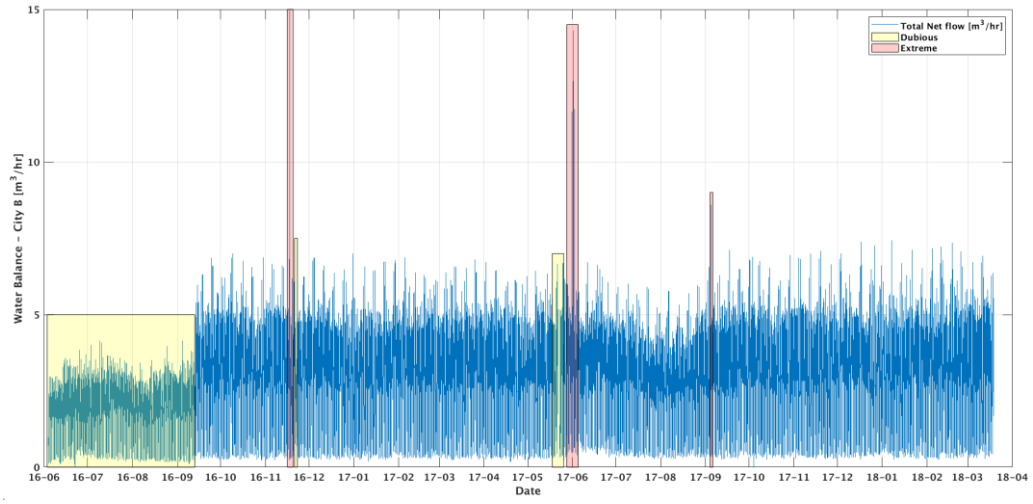
As an example of this behavior the net flow in Hoo (FIGURE 27D), shows that there is no data before December 2016. After verification with the utility, the sensor was installed during November 2016. This demonstrates that the utility is moving towards better understanding of the consumption in their DMA, but the identification of anomalies and faulty data is hard to assess with such a limited time series.

FIGURE 27. LEFT: TIME SERIES OF DATA IN CITY B. POSITIVE FLOW (+) AND BACKFLOW (-), RIGHT: NET FLOW AT EACH LOCATION.



The total net flow into the DMA is presented in FIGURE 28. Net flow data has been added to obtain a total flow. In this case, it is evident that some anomalies can be easily identified. In red some windows were extreme anomalous data were present, and in yellow a window in which dubious data is present.

FIGURE 28. TOTAL WATER BALANCE IN CITY B (APRIL 2016 – MARCH 2018)



Regarding the first window of dubious data (June-September 2016), it is evident that there is a lower magnitude, it is easily identified that during this window the flow meter in Hoo was not in service, and as such this anomaly corresponds to lack of sensor data in the period. In the other two dubious cases, of short duration, it was possible to identify that there was an absence of data in other sensors.

Regarding the extreme data identified for the total water balance in City B DMA, some records coming from the log books were obtained indicating that there was an operation taking place in the system, forcing almost all water to enter the DMA by Koer. The extreme values 2nd and 3rd correspond to the indication of dates in which this occurred in other parts of the system, while the 1st extreme value corresponds exactly to the date in which Hoo entered in operation. Such analysis may not be feasible if the log books of internal operation in the DMA were not available, demonstrating once again the importance to preserve such records in a practical and standardized to identify or get rid of anomalies.

Confidence intervals for anomaly detection

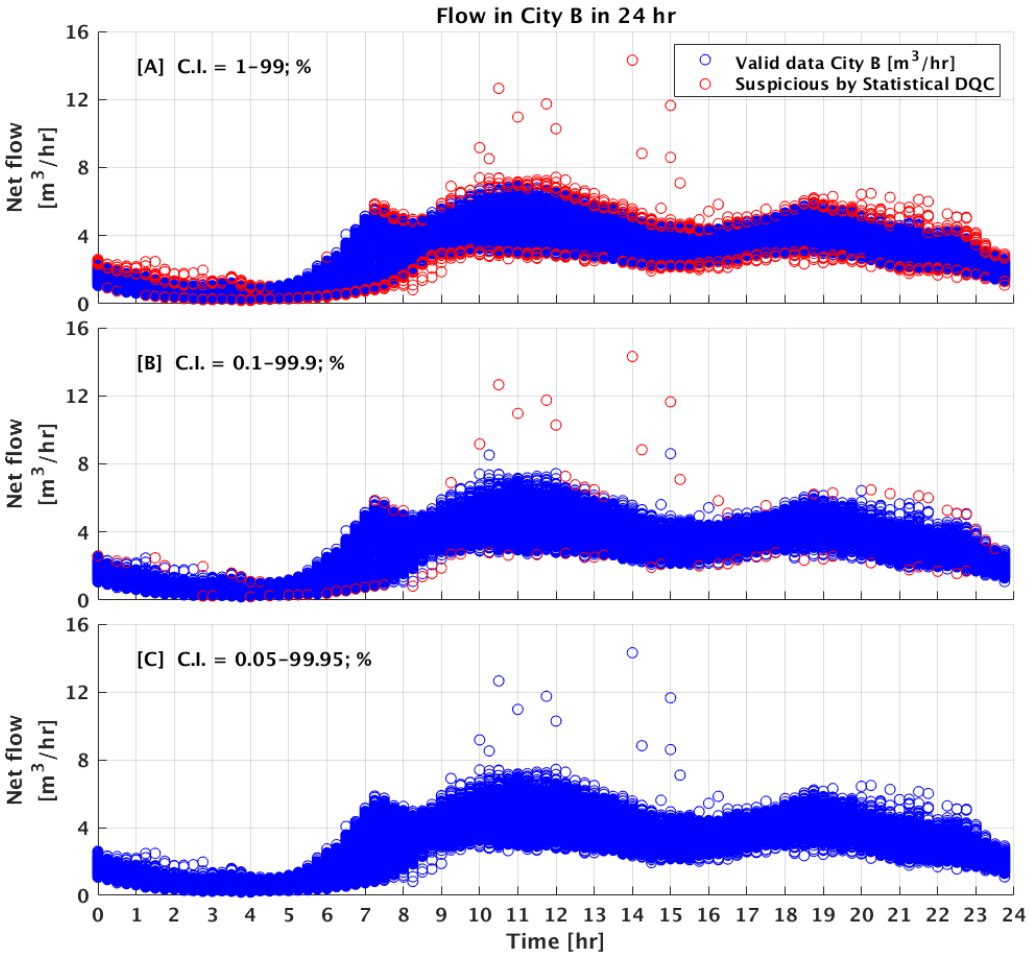
Finally, in this DMA, it is not known directly what is the customer pattern and how much flow can be considered as an anomaly or deviation throughout the day. For that reason a confidence interval for each 15 minutes of daily operation was elaborated. The goal is to identify the minimum magnitude of anomalies. In this case, focus is not made in the anomalies of the patterns as it was done for other utility. In this case, what is relevant is how sensitive an statistical analysis such as Confidence Intervals (CI) may be in a small DMA as City B.

For the confidence intervals, data of the total net flow was rearranged in a scatter in 24 hours. A confidence interval is defined for each time frame (in this case each 15 min), by using normal distributions based on the sample statistics (mean and standard deviation) of the samples for that specific hourly time stamp. By doing so, the dependence of the confidence interval is only based on previous days and not dependent of daily operation at different hours of the day.

The confidence interval depends on a parameter α (in %), which indicates how far a data point is from the mean value. The lower α is, more data will be included in the confidence interval, such data may be considered an outlier. If a sample value is far away from the central trend it is said to be outside of the confidence intervals (CI) at a α significance. Although such analysis

of confidence intervals may be also performed to each individual net flow estimation in the boundaries of City B (FIGURE 27B-D-F-H), here it is only done for the total water balance of the complete DMA (FIGURE 28).

FIGURE 29. CONFIDENCE INTERVALS FOR WATER BALANCE IN CITY B. THREE DIFFERENT CASES. .



The analysis of confidence intervals is additionally performed for 3 different values of α = {1.00, 0.10, 0.05} in FIGURE 29. The selection of α is made just to indicate the sensitivity of identification of outliers as anomalies in the water balance, by specifying fixed α values throughout the day. Once again, it is emphasized that such anomalies can be either explained or unexplained. In the absence of flags issued by Company B data system, an assumption is made such that each anomaly is in fact faulty data.

In FIGURE 29, it is possible to see that there is a trade-off in the identification of outliers given different values of α. For a CI of 1-99% many outliers are detected (FIGURE 29A). Generally speaking outliers are removed from the sample data. This may not be the case as most values of low consumption at any time during the day are considered are not outliers, but are considered as such by this technique. This creates an additional issue for the utility, as the number of anomalies needs to be verified individually making the process cumbersome.

As a work around, a confidence interval with a higher significance between 0.1 and 99.9 % (FIGURE 29B), reduces the number of anomalies detected. Some anomalies are identified as values closer to the central trend of the demand pattern of the DMA, but can be easily verified

as good data by simple visual inspection. Also, it is of notice that some values around 15:00 hours, have a high deviation from the central trend, however these are not identified as anomalies by the CI technique.

In an extreme case, in FIGURE 29C, a confidence interval is selected with a low value of α=0.05%, making the identification of anomalies impossible for this particular case, at any time during the day. This demonstrates, that confidence intervals used for data quality control and anomaly detection must be fine-tuned for each specific case (e.g. water quality, water balance, energy balance), and that this approach may not be the best way to perform validation of data from water balance.

The use of a specific technique or tool for water balance anomaly identification without the correct understanding of the system becomes a technical challenge at each utility managing diverse DMA sizes as it is the case of The Netherlands.

4.4 Company C

Company C supplies water through a network of over 32,000 km to 180 municipalities, serving around 3 million customers and hundreds of companies. In the past it was provincial, therefore geographically oriented. The technical management (*Technisch beheer*) is done by an operation service (*explotatiedienst*). Company C has its own SCADA system to identify trends and to perform control in real time. Data from approximately 60.000 variables is collected at least each 15 minutes as FIFO (First-In-First-Out). Data is daily saved in 'flat files'[6]. Everything is stored in Microsoft SQL Server. Data validation is made in the SQL server every 1 day.

By the installation of new sensors, a lot of attention is given to defining physical boundaries, calibration, but afterwards that occurs less often. Sensors that monitor Chlorine, Turbidity and pH are once per month are regularly calibrated. Pressure and flow meters are checked on a request basis.

Just before the starting of the pilot, there was a server crash and a bulk of information (not quantified by the utility) stored in the server was lost. Therefore, only a limited number of datasets were available for this pilot project. The server crash evidences the vulnerability of systems and highlights the need to use robust systems for data storage, because data is the crucial component of basic statistical and data driven models/applications. Currently, Company C is working on recovering the data, in this report we address the DMA case in a qualitative way and the water quality case similar to the case of Company A. To avoid repeating results and add value in the case of water quality, only the analysis for Chlorine is presented as this is a variable not used in the Netherlands.

4.4.1 DMA - City Db

The DMA of City Db was selected to be analyzed in the pilot project. City Db is a complex DMA, it is connected to the supply areas of Ev, JE and Mz (see FIGURE 30A).

[6] A **flat file database** is a database stored as an ordinary unstructured file called a "flat file". To access the structure of the data and manipulate it on a computer system, the file must be read in its entirety into the computer's memory. Upon completion of the database operations, the file is again written out in its entirety to the host's file system. In this stored mode the database is said to be "flat", meaning that it has no structure for indexing and there are usually no structural relationships between the records. A flat file can be a plain text file or a binary file (Source wikipedia).

FIGURE 30. (A) WATER PRODUCTION CENTRES, (B) DETAILED OF THE WATER BALANCE IN AND AROUND CITY DB

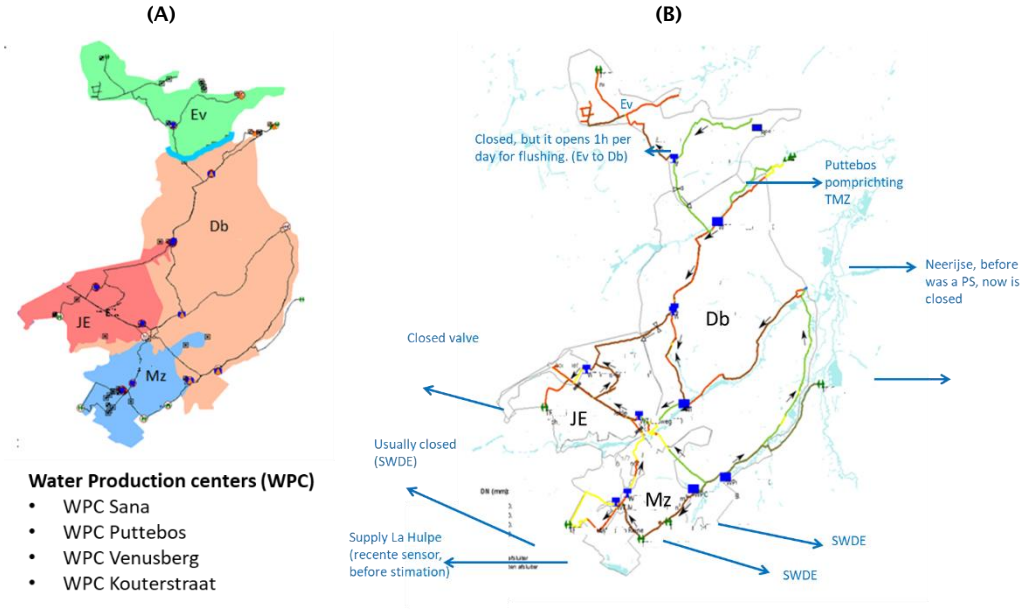


TABLE 9 shows a generic description of the water balance meters of the supply area City Db. Often the stored water in the water towers is neglected, which can be a source of inaccuracies and uncertainty in the water balance.

TABLE 9. OVERVIEW OF THE FLOW METERS OF THE CITY DB SUPPLY AREA (WPC: WATER PRODUCTION CENTER, WT: WATER TOWER)

Direction	Supply Area Ov	Type	Unit
+	WPC OvK	counter	m³
+	WPC Puttebos City Db	counter	m³
+	WPC Sana City Db	counter	m³
+	WPC Venusberg	counter	m³
+	WPC Hoeilaart	counter	m³
+	MTK WT. Ev richting City Db	counter	m³
-	MTK WT. Ev richting Ev	counter	m³
	WT JE	Analog signal	m
	WT Mz 1	Analog signal	m
	WT Mz 2	Analog signal	m
	WT City Db 1	Analog signal	m
	WT City Db 2	Analog signal	m
	WT Losweg	Analog signal	m

Looking in detail to the operations described in FIGURE 30B it can be seen that the system is regularly re-configured, e.g. as the valves are dynamically operated. This means that the water balance formulation changes continually, as input and outputs are constantly adapted. Furthermore, in the last period one pumping station was closed, forcing other pumping stations to provide more water to the supply area. As such, at this stage and within this project is not possible to perform an analysis of water balance in this DMA.

To successfully describe and validate data of water balance in City Db, additional data is required. Specifically:

- As water is being continually exchanged among the different supply areas, exact dates and times of pump switches, or changes to valves openings is needed to be able to determine the water use in the DMA.
- Expert knowledge is highly needed to successfully validate the water balance data. This system is more complex than others analyzed in this report. More information about the daily operations is required to understand the changes in data provided by the utility.
- Sensor installations are also of relevance. For instance, recently, one sensor has been installed to measure the flow going to a nearby DMA, while old data collections of such flow were generated in the past as an estimation by the utility. In fact, this sensor was installed because it became evident (by the utility operators) that the infrastructure (sensors and software) did not guarantee a complete set of data relevant for water balance of this system.

Furthermore, this is a complex case, for that reason it is also recommended for Company C to collect also:

- Log-books of maintenance and operations performed in their system. Currently, not collected, and required to be able to differentiate between events and normal operation. This is the basis for the determination of faulty data detection.
- A calibrated hydraulic model for further evaluation of (changes in) operation of the system will increase the capacity to understand the changes in the system. Currently, this DMA is not isolated from other areas, as such a larger model available by the utility can be used to perform data validation by means of deviations between observed data and simulation results. The data validation methods proposed so far on this report can't be easily adapted at the current stage for this DMA.

The exploration of data validation based on a hydraulic model will be further explored in this report with an application on a benchmark water distribution network case.

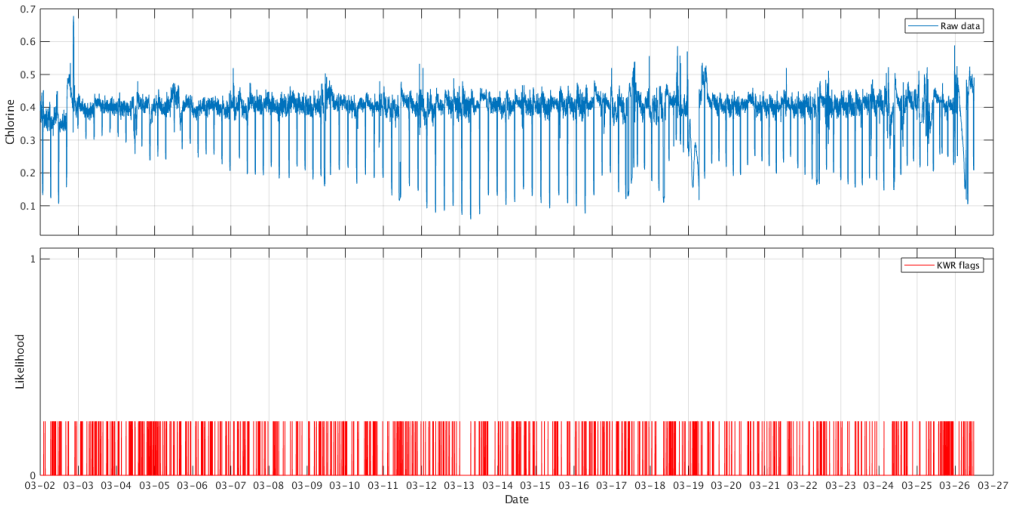
4.4.2 Case study water quality - Chlorine

For the water quality datasets: chlorine, temperature, turbidity and pH. The analysis performed for Company A was replicated. However, the parameters that were adjusted for the datasets of Company A do not match the datasets of Company C. Therefore, these parameters should be again calibrated. For that reason, a selection was made to analyze and present here the time series corresponding to chlorine as this is not a common sensor variable in The Netherlands.

The flags obtained from Company C are present in the data with a numerical representation with long integers. Values of 262336 and 268697792 are present in data from turbidity and pH. Although the data has a label for data quality control, after a feedback session with Company C it was clear that from the user perspective, the utility is not able to describe clearly the meaning of these flag numbers from the software provider. A simple hypothesis is that

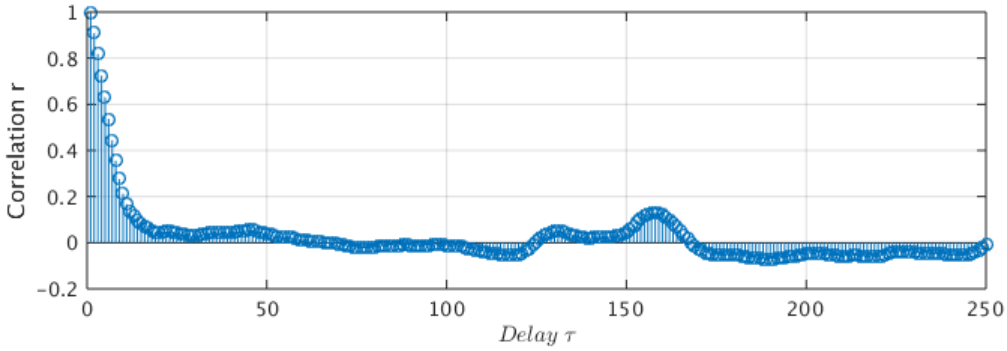
these numbers correspond to a translation from a bit string tag, as 262336 ~ 2^{18} and 268697792~2^{28}, however this does not clarify what each flag represents or its meaning. Data submitted by the utility of chlorine did not contain flags. Here, only the simple data validation tests proposed on this report were applied and the corresponding results presented.

FIGURE 31. TIME SERIES AND DATA VALIDATION OF CHLORINE DATA COMPANY C.



The number of flags identified are 609, which corresponds to 3.7% of the data samples. It was not possible to obtain a lower number of flags as there are limitations of the data regarding its numerical precision. Many values are repeated in consecutive timestamps, triggering a flat value flag. At the same time, the jumps of the time series (change of variance), appear to be a systematic behavior. For that reason, the jump test was eliminated from the analysis to this particular time series. In order to estimate the frequency in which such jumps or drifts occur an autocorrelation analysis of the time series was performed.

FIGURE 32. AUTOCORRELATION OF CHLORINE TIME SERIES.

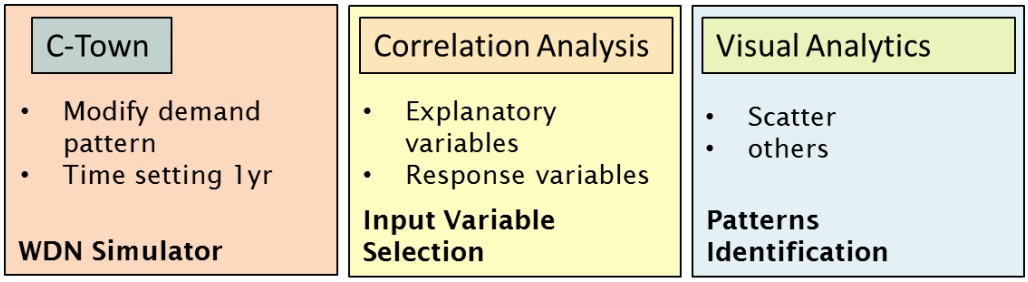


Such analysis is presented in FIGURE 32, where it is seen that the time series is highly correlated up to a delay of 15 minutes. This initial correlation pattern indicates that the time series of chlorine is linearly dependent with the time series of the last quarter of hour. A second peak shows that when the delay is 160 minutes chlorine is linearly dependent. This indicates that during the operation of the plant a mechanistic process occurs which triggers the sudden drop of Chlorine every ~3hr.

4.5 C-Town, benchmark Network.

Some techniques for data validation require additional input which may be not available in a ready-to-process form by the utilities. One of such inputs, is the use of water distribution network simulation models to support data validation. For that purpose, and with the goal of test other techniques for data validation with controlled data, an artificial system is used. By controlled is meant that its behavior is known in advance and easily modified.

FIGURE 33. PROCESS OF ANALYSIS OF DATA IN A BENCHMARK WATER DISTRIBUTION SYSTEM.

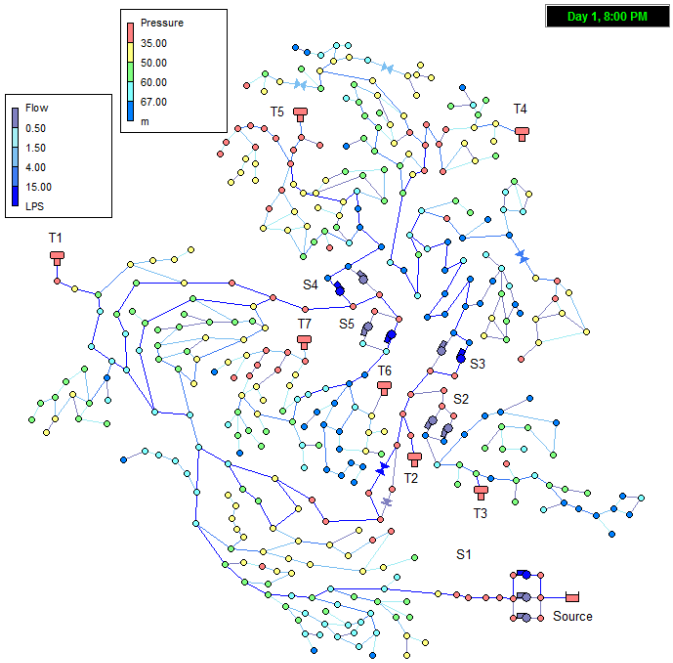


Initially, the model patterns will be modified, and the system will be run in an extended period simulation for 1 year. In a sense, the simulation results are considered here the data which will be used for data validation. Then a correlation analysis among selected variables considered as simulation results will be performed. Anomalies are applied to data and visual analytics is used to analyze whether the anomalies correspond to faulty data or not.

4.5.1 Model set up

In this case, C-Town (FIGURE 34) water distribution system, which is based on a real-world medium-sized network, first introduced for the Battle of the Water Calibration Network (Ostfeld, et al. 2011) was selected as a test bed for data validation analysis. The WDN consists of 429 pipes, 388 junctions, 7 storage tanks, 11 pumps (located as 5 pumping stations), 5 valves (1 of them check valve, and one of them Flow Control Valve “FCV”), and a single reservoir or source located in the southeast. Information regarding the distribution system was incorporated into an EPANET2 (Rossman 2000) input file (see FIGURE 34).

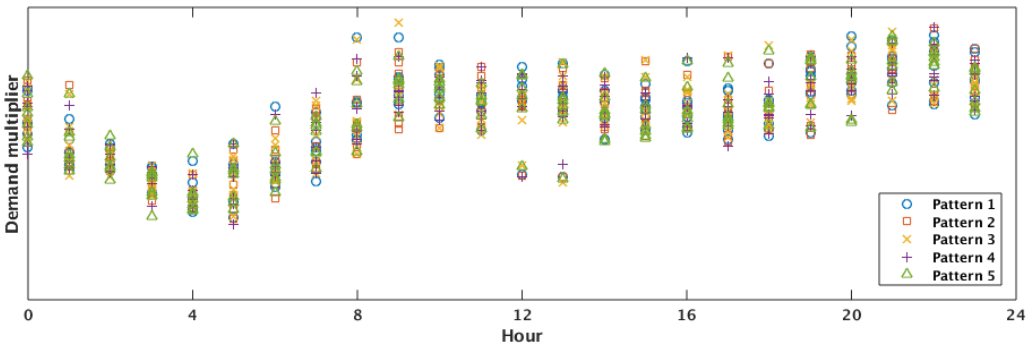
FIGURE 34. SNAPSHOT OF THE OPERATION OF C-TOWN AT 20 HOURS OF SIMULATION. FLOW ON PIPES AND PRESSURE ON NODES.



4.5.2 Added perturbations to demands

Water consumption is regular throughout the day with different consumption patterns in different parts of the network. The model contains patterns (5 of them) of a 24 hours duration. This means the system is split in 5 DMA's.

FIGURE 35. PATTERNS OF C-TOWN AFTER APPLYING A NORMALLY DISTRIBUTED PERTURBATION

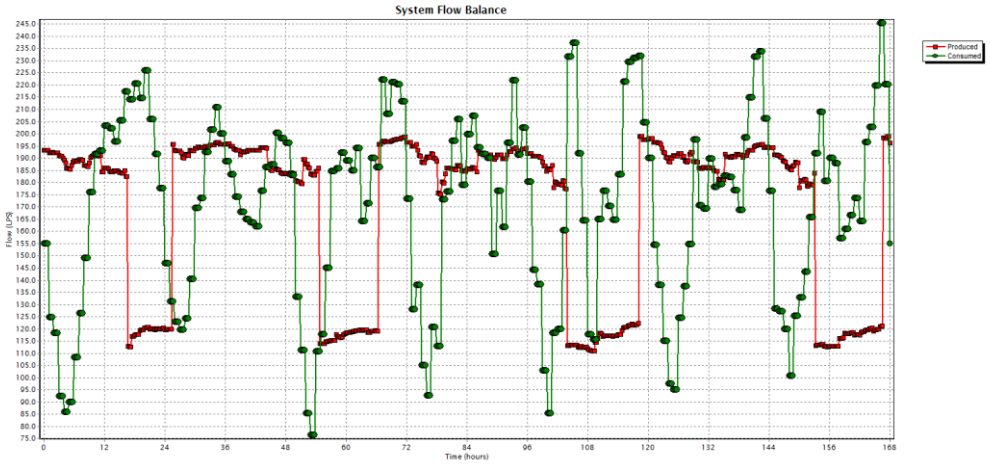


A perturbation is applied on demand patterns to simulate what will happen with real-life measurements where uncertainties (independent of its source) are present in data. Such random perturbation helps to simulate the stochastic nature of user demands. It is desirable to have a variable demand pattern throughout an extended period simulation, rather than to have the same daily demand pattern in all demand locations. Subsequently, the demand patterns were altered by adding a random perturbation.

The named perturbation is normally distributed, with a zero mean and a standard deviation equal to 10% of the demand multiplier at the particular hour. The perturbation is then added (or subtracted) to the demand multiplier (Kapelan, Savic and Walters 2005). In this way, the

demand multipliers are close to its original values, and the probability of having negative (unrealistic) demand multipliers is minimized. FIGURE 35 shows the hourly variability of the 5 patterns. The total flow balance during a 1-week period (168 hours) is presented in FIGURE 36.

FIGURE 36. TOTAL WATER BALANCE FOR C-TOWN DURING 1 WEEK SIMULATION. (IN RED PRODUCTION AND IN GREEN CONSUMPTION).



4.5.3 Selection of measured variables, a SCADA analogy.

From the system it is possible to obtain all variables of flows and pressures at every pipe and node, however this is not realistic in a real system. Not every pressure or flow is monitored in real systems. Consequently, the model was run for a whole year at hourly time intervals (8760 timestamps). Being that the case a selection is made for certain variables which are of interest, similar to the ones which may be obtained from a SCADA system. Only 43 variables were collected from C-Town for the 1-year simulation, corresponding to:

- the tank levels (7 measured in m),
- the flows of the pumps (11 in l/s),
- the flow of the flow control valve – FCV (1 in l/s),
- the status of the pumps (11 as ON/OFF),
- the status of the FCV (1 as ON/OFF), and
- pressure observations downstream of pumping stations and the FCV in the WDN (12 of them).

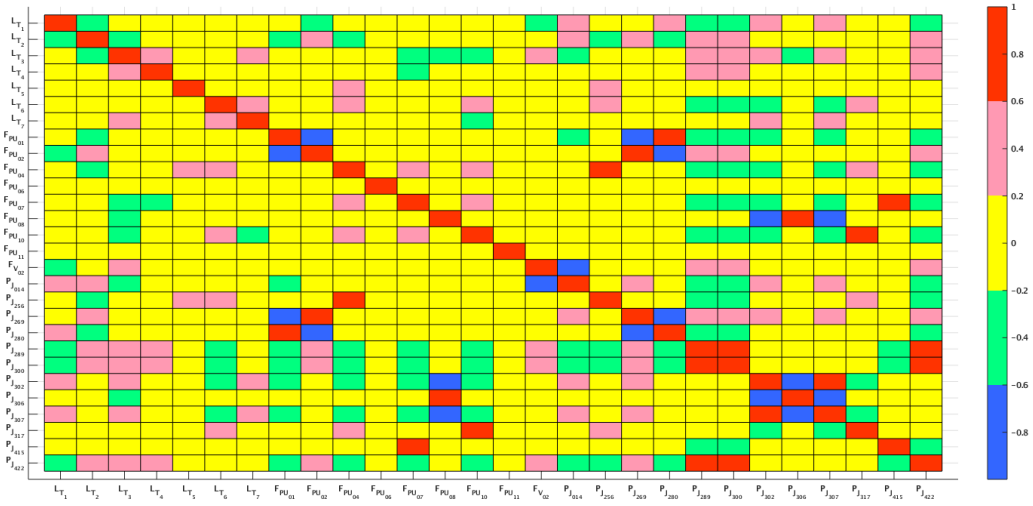
The interaction of pumps and valves creates a nonlinear variability in the internal nodes and elements. The main drawback, is that there is no water quality data to test data validation.

Cross-correlation among variables

It was decided to present a correlation analysis which resembles Input Variable Selection (IVS) (Galelli, et al. 2014). This means that the correlation among time series is performed and the most significant explanatory variables are selected for each variable. For simplicity, the status of the pumps (binary variable 0 or 1) was taken out of this analysis. The reasons for the dismissal of pump status are: (i) it requires a different approach for correlation analysis due to the fact that is a binary variable, and (ii) because the status of pumps is directly correlated to the flows obtained after simulation, so it is redundant. In a way, pump switches in a pumping station can be easily retrieved from the time series of flows, while the opposite is not possible. After removal of pump status, a total of 28 variables presented a behavior which

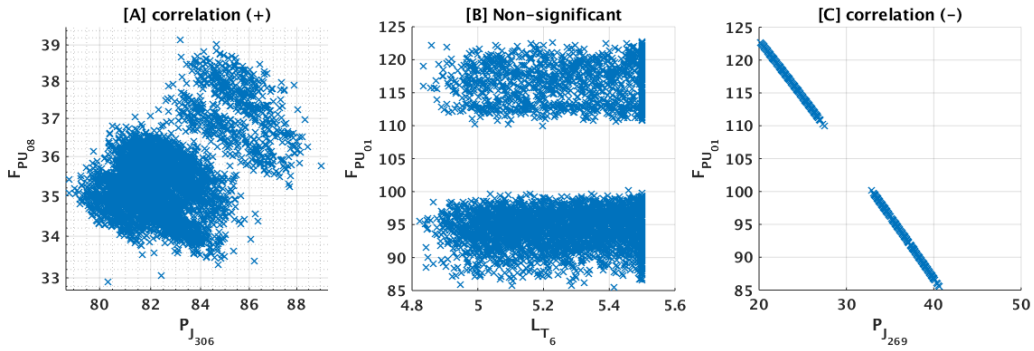
could imply a correlation (see FIGURE 37). The levels in the 7 tanks (L_{Tx}), flows in 8 pumps (a total of 4 pumps do not operate during the whole simulation), flow in 1 valve (F_{PUx}, F_{Vy}) and 12 pressure 'sensors' (P_{Jzzz}). By sensors it is meant that observations are retrieved from the simulation or as virtual sensor data.

FIGURE 37. CORRELATION ANALYSIS OF TIME SERIES FROM C-TOWN DURING A 1 YEAR SIMULATION AT 1 HOUR TIME PATTERN. NO DELAY AMONG TIME SERIES.



It is possible to observe that some time series have a significant correlation (red values) or proportional behavior (almost redundancy between time series). In general, correlation of variables with themselves must be dismissed as this results in a correlation of 1.0 by definition, that is why the diagonal of FIGURE 37 is completely red. Other time series may present a behavior with no statistical significance (yellow, close to zero correlation). While other time series, may present a negative correlation near -1.0 (blue values) or inversely proportional behavior. To demonstrate what it is meant, in FIGURE 38, it is presented an example of the scatter of the 3 different cases.

FIGURE 38. SCATTER OF DATA FOR 3 DIFFERENT OBTAINED CORRELATIONS (A) POSITIVE, (B) NON-SIGNIFICANT, AND (C) NEGATIVE.



The positive correlation (FIGURE 38A) is explained as junction 306 (J_{306}) is located nearby upstream of pump 8 (PU_{08}), the negative correlation corresponds to the typical case in which the pressure and flow are measured for the same pump (FIGURE 38C). The case of non-significant correlation (flow in pump 1 vs level in tank 6 in FIGURE 38B), is interesting because

it demonstrates how complex behavior among variables may indicate no statistical significance between them.

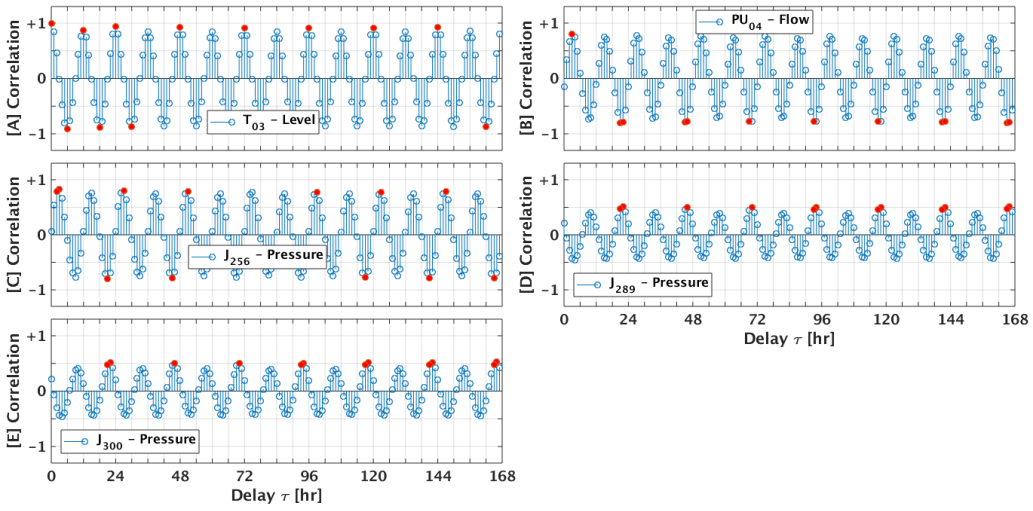
If the simple tests which were implemented during this project were used for the time series in FIGURE 38B, most likely the upper or lower set of samples would be identified as faulty data. In fact, it is evident that the tank 6 reaches its maximum level several times during an annual operation, that said the simple boundary test may be defined based on this measurements.

4.5.4 Correlation with lags or previous time steps.

After analyzing all the time series correlations, an aspect which has not been discussed is the use of correlation at different lag times, meaning, that there can exist hidden correlations among variables with respect to previous time steps or that there can exist 'memory' among time series. This has been discussed, previously when ARIMA models were introduced as tools for data validation, and for the analysis of Chlorine data on itself. Here, the analysis is performed for multiple variables.

First of all, we will focus in a specific response variable. The response variable selected is the water level in tank 3 of C-Town (T_{03}). The correlation with respect to all other explanatory variables with a delay (memory among series) up to 168 hours (1 week) is presented in FIGURE 39. Only the most significant 5 explanatory variables are presented: A) itself (T_{03}), B) flow of pump 4 (PU_{04}), C) pressure at junction 256 (J_{256}), D) pressure at junction 289 (J_{289}), and E), pressure at junction 300 (J_{300}).

FIGURE 39. CORRELATION OF EXPLANATORY VARIABLES WITH RESPECT TO WATER LEVEL IN TANK 3. THE MOST SIGNIFICANT DELAYS FOR EACH VARIABLE ARE HIGHLIGHTED IN RED.



For each explanatory variable, the most significant delays are presented as red dots in FIGURE 39, and are listed in TABLE 10. For example, T_{03} shows a significant correlation with respect to data of PU_{04} from 3, 21, 22, 45, 46,...165, 166 hours ago. Also it is of notice that J_{289} and J_{300} present exactly the same significant delays.

TABLE 10. LIST OF EXPLANATORY VARIABLES AND ITS SIGNIFICANT DELAYS FOR CORRELATION WITH RESPECT TO WATER LEVEL IN TANK 3.

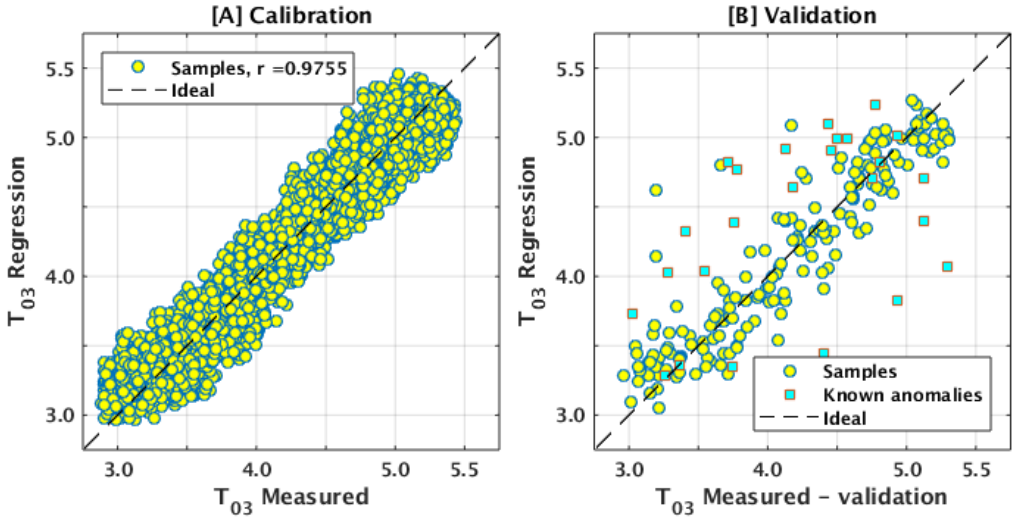
Name	Type	Units	Delays of significance for each variable
T_{03}	Tank level	m	0*, 6, 12, 18, 24, 30, 48, 72, 96, 120, 144, 162
PU_{04}	Pump Flow	l/s	3, 21, 22, 45, 46, 69, 93, 117, 141, 142, 165, 166
J_{256}	Pressure	m	2, 3, 21, 27, 45, 51, 99, 117, 123, 141, 147, 165
J_{289}	Pressure	m	21, 22, 46, 70, 93, 94, 117, 118, 141, 142, 165, 166
J_{300}	Pressure	m	21, 22, 46, 70, 93, 94, 117, 118, 141, 142, 165, 166

**A delay of zero on the same variable corresponds to autocorrelation r = 1.0.*

4.5.5 A regression model for Tank 3

Based on the most significant delays for all 5 explanatory variables a multivariate linear regression model is built for the water level in tank 3 (T_{03}). A total of 59 explanatory variables (see TABLE 10, excluding delay zero for T_{03}), are used to estimate a time series for tank 3. The scatter of estimated vs measured tank levels is presented in FIGURE 40A.

FIGURE 40. MULTIVARIATE LINEAR REGRESSION SCATTER PLOT OF WATER LEVEL IN TANK 3. A) CALIBRATION BASED ON MEASURED DATA (FROM EPANET MODEL), B) VALIDATION OF REGRESSION WITH ANOMALIES ON NEW DATA SET.



As performance indicator, the correlation between estimated tank levels obtained by regression and measured levels (originally from the C-Town's EPANET model) is r = 0.9755. This means that the regression is reliable. The *ideal*, or perfect regression would correspond to having all dots aligned in the diagonal. To measure how far are the estimations form the measurement, two metrics i) Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are obtained. RMSE is 0.14m (14 cm), while MAE is 0.11m (11cm). As an additional control for the regression model it is verified that no values are estimated outside the physical boundaries of the tank (0.0 m to 5.5 m). Knowing in advance this regression model based for Tank 3 water level, offers the opportunity to use it for data validation.

4.5.6 Anomalies as faulty data

Once a regression model for Tank 3 is available, anomalies (i.e. single or multiple time steps) were introduced in different timestamps of data of two weeks at 1-hour resolution. In a sense, by knowing in advance where, and when the anomalies were injected it was possible to verify the additional gains in using a regression to identify anomalies for data validation.

Such anomalies correspond to repeated time stamps, sudden jumps in the data, hiding values and changing some values to zero (as when data is not retrieved by the SCADA system). A total of 25 faulty data samples were inserted (7% of time series length) in a 2-week data set. The regression model is applied to the time series of Tank 3, based on the same explanatory variables used during calibration (TABLE 10). Results are presented in FIGURE 40B, where the inserted anomalies are marked as squares. It is evident that most anomalies lie outside the central trend of the regression, so by using deviation or MAE as a boundary of data validation most faulty data is identified. As a drawback, false positives appear or values which lie far away from the central trend but that it is known that do not correspond to the faulty data inserted. In addition, three faulty data samples are close to the central trend of the validation and would not be identified by such data validation technique.

Subsequently, the protocol for data validation was applied, to compare the basic tests with the regression analysis. The anomalies identified by the data validation protocol of this report with simple tests is able to track only 5 samples. This is a drawback of the basic methods as it entails simple methods. It also highlights the need to use additional more complex methods for data validation, which may facilitate the identification of faulty data in more complex datasets as the one obtained with this simple benchmark network.

4.5.7 Other non-linear models

It would be interesting to test to what degree synthetic data errors (which conform to real data errors) can be detected and filtered out using several techniques. That would show the potential of each technique to the utilities. However, this not realistic in the context of this project (at least at this stage).

4.6 Lessons learnt from the case studies

From the visits to the water companies and the pilots a number of best practices and issues can be identified. TABLE 11 contains the best practices and TABLE 12 contains the issues identified during the pilot.

TABLE 11. BEST PRACTICES IDENTIFIED DURING THE PILOT

Best Practices	Company			
	A	B	C	D
Data scientist works together with a domain expert				X
Clear responsibilities	x			X
Validation rules reported				X
Implementation of automatized routines which allows continuo validation of some datasets	x		x	X
Validation of aggregated data				X
Implementing pilot projects to learn from it		x		

TABLE 12. ISSUES FOUND DURING THIS PILOT PROJECT

Issues	Company			
	A	B	C	D
Data storage integration. Different databases or a single database. If multiple DB, then sometimes data is not linked one to one.	x	x		
Lack of overview of metadata: difficulties to track additional information e.g. log books of maintenance, data is still in different databases stored	x	x	x	
Lack of priorities/time to (at least) tag the large number of signals and known events		x	x	
Problems related to the interfaces				X
Current approaches are often somewhat ad hoc.		x	x	
A lot of techniques, but still data validation/correction is largely based on expert opinions	x			X
Own customize system, data models and tools (not compatible with other companies)	x		x	X
Only a small percentage of the data is validated	x	x	x	X
Vulnerability of failure of servers, data from third parties	x	x	x	
Black box in the built-in tools. Automatic filtering of suspicious data and not clear which rules they use to validate the data.	x		x	

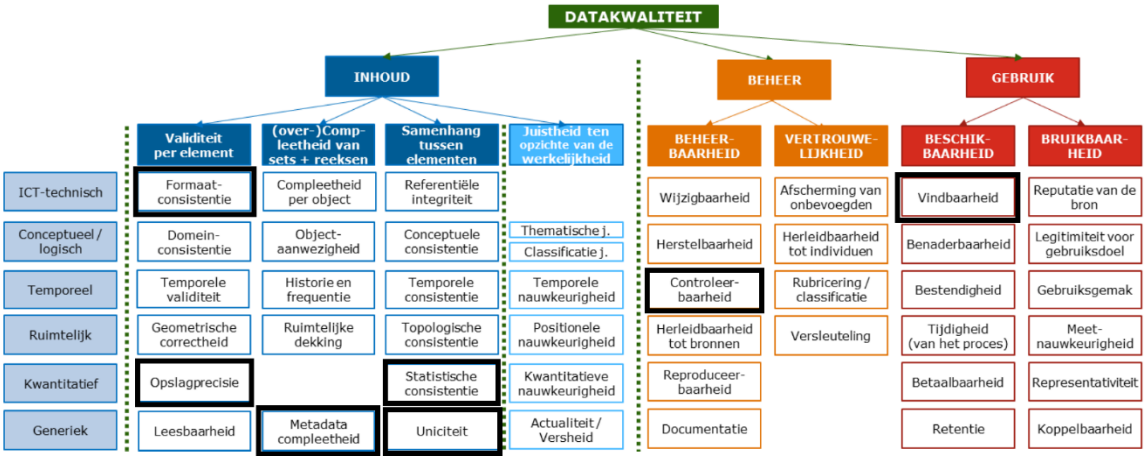
A general issue that arises from the comparison of the four companies is the lack of standards. There are different types of registrations within and between the companies. Additionally, there is a lack of knowledge about how data is validated by third parties e.g. energy companies.

Among the utilities the purpose of validation is very diverse e.g. water flow, billing, determining the water balance, identifying leaks, changes in turbidity due to new filter installation and of course, depending on the objective the requirements of the validation change. The following issues have been identified that hinder the potential for implementation of more complex/advanced data validation techniques:

- Known deviations and operations are usually poorly logged by the utilities, or when this is done it is very limited to some variables across utilities,
- lack of specialized manpower to perform this task on a regular basis: a team consisting of both data scientists and hydraulic engineers is required,
- specific techniques for data validation are still hard to adopted because there is no overview regarding which data are needed for each of them.

Some of these issues identified in the case studies can also be found in the framework developed by *Rijkswaterstaat* (Directorate-General for Public Works and Water Management), see black boxes in FIGURE 41. The black boxes are topics to which at least some attention is paid (or drawn by this research); the non-boxed topics may not even be on the radar.

FIGURE 41. POSITIONING THE ISSUES FOUND IN THE PILOTS WITHIN THE FRAMEWORK DEVELOPED BY RIJKSWATERSTAAT



Although only a small fraction of the data is validated, all of these data are potentially still used as input for different types of models e.g. reliability analysis of predictions of systems performance under certain scenarios.

Significant differences in the resolution of different parameters was also observed defined by the frequency with which they are stored in utility databases. For example, the year of installation of pipes, usually has a resolution of 1 year. On the other hand, water quality data for turbidity, can be stored with timestamps every 5s in a database.

There is still a lack of knowledge about how to, first select the appropriate technique for validating a given dataset, and second how to *fine tune* the parameters of each validation technique, to minimize the tradeoff between identification of possible anomaly and extreme events in the system. Utilities face the same challenges when defining flags for data validation:

- Too many flags, (false positives) operatives become reluctant to verification and data validation as the trust on the data validation diminishes.
- Too few flags and robustness is lost as water companies would not be able to differentiate between a regular event, and extreme event and a real anomaly.
- Changes in the system of historic data is not always recorded or archived. This is a challenge when the system configurations change dynamically (e.g. in the case of Company C)

The theoretical case of C-town shows the potential to implement more complex techniques in the near future by using hydraulic models to understand the plausible behavior of components of the system and to find out correlations between variables.

5 Discussion, recommendations and future work

5.1 Introduction

Data quality is a key consideration for the reliable functioning of drinking water systems, as data is used to monitor and operate systems, to bill customers, to report the performance of the company, to feed different types of models. Improving the quality of the data and making it more accessible will benefit every departments of a company.

The following sections discuss a number of recommendations and future work in the field of data quality control for drinking water utilities.

5.2 Recommendation regarding future work

During the project, it has been evident that most utilities apply diverse methods of Data Quality Control (DQC). One of the features which is lacking across is a degree of standardization of DQC. A proposal for a Dutch initiative for standardization of DQC for drinking water utilities is suggested as a possible follow up of this work. Such a task might require the development of a tool which could help operatives with the task of daily time series analysis, IVS, regression, interpolation, data smoothing, data aggregation and data correction.

5.2.1 Standardization

Defining when the quality of the data is good enough, is case specific. Standardization can help to 1) Understand common problems among utilities, 2) speak the same language so that similar issues can be addressed across utilities, 3) apply the same methods and 4) use the same tools. Challenges for which data standardization can provide a solution to, include:

- Allowing integration of data that come from different sources, origins and formats.
- Automatizing data control and correction (support to automate processes) – less by hand and subjective handling of data.
- Allowing faster and better analysis and understanding of processes (more objective, reproducible and comparable results).
- Improving and facilitating reporting and compliance.
- Reducing cost (and time) by providing certainty of units, protocols, event types, etc.
- Allowing data sharing and implementation of hydroinformatics tools
- Facilitating interoperability of (IT) tools within a company and between companies
- Allowing comparison within departments or companies e.g. benchmarking
- Enhancing transparency and clarity about what can and cannot be done with data

Water companies can use as a starting point, relevant experiences of other sectors. There are several standards which are relevant to the water sector, developed for instance within Internet of Things (IoT) initiatives (using smart appliances) or smart cities initiatives. Standardization within the Water Sector is something that the European Commission (Makropoulos, van Thienen and Agudelo-Vera 2018) is very interested to achieve with different actions are taking place, but bottom-up action is also needed from utilities.

5.2.1.1 Data model

It is recommended to adopt and adapt a framework such as the one developed by *Rijkswaterstaat* where different dimensions and categories are clearly identified not only for the data content, data management, but also for diverse user are considered. Currently data from the utilities are highly variable in volume and resolution, format, metadata, and shape. However, they all measure the same types of variables. It would be very helpful a standard data model for DQC control can be agreed at a national level in due course. Given that the customers are relatively the same in the country, this implies that the water utilities can develop an intercompany data model for water accounting, with the possibility to extend it in the future as new variables are incorporated in their data warehouses.

5.2.1.2 Data collection redundancy

As presented in this report, the analysis of the *data diet* of large data collections may help utilities identify which sensors are more reliable and how much data storage is required. It is not clearly known by the utilities which variables are correlated with statistical significance to other sensors in a definite way. An effort should be made to develop such analysis of redundancy with the utilities using their data mining techniques discussed in this project.

5.2.1.3 Aggregation of data

Aggregation of data is performed by all companies, for specific purposes. Additional data analysis which is of interest for all utilities can be mainly allocated to obtain regular water balance calculations. All utilities require to identify and account for the water produced and commercialized. However, it was identified that the utilities have serious concerns about the limits or boundaries of the anomalies in the water balance. Given that Non-Revenue Water (NRW) is not a serious concern in the country, the main issue is to be able to identify when the water supply system presents a deviation from its regular pattern. As demonstrated in this project, for some utilities, the need to establish such boundaries for water balance is a current issue. Even with advanced tools for water accounting there are deviations present in the data of water balance for all utilities. Therefore, it is recommended to implement advanced techniques such as model based validation to tackle this issue. Here a pilot for a DMA configuration, with an optimal time step and additional info: (e.g. pressure, flows) and expert knowledge (e.g. operators' knowledge, and logbooks of operations and maintenance), can be compared with a simulation model to determine anomalies.

5.2.1.4 Data correction techniques

Most methods applied for incomplete time series have been developed for surface and groundwater hydrology (E. Veling 2010) (von Asmuth 2011) (von Asmuth and van Geer 2013), and some applications have been made for WDN (Oriani, et al. 2016). This project did not address the issue of data correction techniques. A proposal is made to continue the process of use Hydroinformatics tools, but with the goal of identifying specific techniques for data correction which may be suitable for a subset of water distribution network variables. If for the case of data validation, the spectrum of techniques was broad, a similar content is expected in the case of data correction techniques.

5.2.1.5 Data reconciliation

The decision which must be made by utilities about which data to trust and which data not to trust is a constant struggle. Often during the selection of data and case studies of this BTO, a common concern was voiced by operatives of utilities about not trusting the data of a region or DMA. However, no quantification or metric was made available to express this in each case, and to our knowledge such metrics are not available in the Dutch context. This means that expert knowledge has been applied by utility's experts, with prior experience on the management of their system. However, such expert knowledge is currently only encapsulated

in the minds of experts. There are several methodologies available to transform such qualitative decision making into quantifiable rules for the determination of likelihood of data as being faulty. This can determine improvements in data collections and reduce dependence of utilities of specific experts due to absence (i.e. holidays, leave, sickness, pension, death).

5.2.2 Selection of faulty detection techniques

Regarding data quality control, there is no such as one technique fits all. Depending on the monitored event/variable, different techniques with different parameters should be applied also according the objective of the validation.

Software tools cannot validate data by themselves. Expert knowledge is always needed for a good determination of the parameters to perform the data validation, especially in complex systems, such as drinking water systems.

Given the variability of datasets, number of records, timestamps, time resolution and variables only simple techniques for data validation have been applied for the water utilities data, and a simple regression model for the benchmark case.

So it is recommended that in a future project, similar techniques are specifically calibrated for subsets of data from the same variables. For example, in this report a short analysis of data validation for water balance is presented, but an extensive literature on the matter is also available. This means that the possibility to increase the identification of faulty data can be explored with many more techniques than the ones presented here on that subject. Once this is done, then a proper selection of best practice techniques for water balance can be established for the context of Dutch water companies.

5.2.3 Modeling for anomaly identification

During this project, only a short portion of techniques for data validation was explored and only simple tests were implemented. One of the biggest obstacles in applying data validation techniques (see FIGURE 7), corresponds to the use of modeling tools to identify anomalies.

Several methodologies are available which have not been implemented in this work, ranging from statistical, surrogate, and physically based models with the aim to properly define when a sample of data corresponds or not to an anomaly. Having a calibrated model of a WDN, presents the possibility to estimate deviations between measured data and simulated data and as such detect anomalies.

6 Conclusions

6.1 Specific conclusions based on the pilots regarding faulty data detection

6.1.1 Company A

Data validation applied to water quality demonstrated the validity of the utility's flag system, as it is able to identify most anomalies. In general, data of Company A is of good quality with a low percentage of flags issued by the system as faulty data.

In comparison with the simple data validation rules proposed in this report, the implementation of a flat value test can be a possibility to increase validation criteria for the utility. With the protocol proposed in this work, it was possible to demonstrate that the turbidity time series tend to present jumps (i.e. drifts or changes in average). This is another area which could be of future improvement for data quality control.

From the validation of data for water balance, it can be concluded that the temporal data resolution is sufficient, however the process of data collection and aggregation is quite demanding. If a water balance needs to be performed at different intervals (e.g. 1 day, 1 week, 1 month, 1 year), the tasks of data validation would require extensive searches from diverse sources to confirm information from logbooks, installation and maintenance records. This is highly time consumptive and requires improvement as it is not fully integrated.

Expert knowledge of the utility helped to clarify the validity of large collections of data from the last 2 years, i.e. the pumping station maintenance in Haarlemmermeer. The integration of such expert knowledge in the system has yet to be implemented. This is evidenced in the fact that most queries of additional data validation were solved by internal communication via-via, and not through a complete record of operations in any database.

An analysis of regression of volume vs energy consumption also demonstrates that there is a significant correlation between the two variables. There exists the possibility to fill gaps of volume using energy. A data validation of energy data is required, because: i) the data is provided without flags by a third party, and ii) the data contains samples with low energy consumptions for regular water consumption. A hypothesis was made regarding the source of such anomalies in third-party data, however this must be explored in detail. This proves that even with the known correlation between volume and energy, there is a need to increase the number of data validation techniques applied to avoid misrepresentation of the system by the utility.

6.1.2 Company B

The process of data validation at Company B has just started, no flags on data are reported. The main concern from a data validation is to be able to identify faulty data in water balances.

A data aggregation was performed for 4 sensor meters in City B as a single DMA, and a time series of net inflow was estimated. It was possible to easily identify periods where data validation can identify doubtful data, however the lengths of the time series are rather short and demonstrate how the process of data validation is a current effort for this utility.

The use of confidence intervals was demonstrated for the time series of net flow in City B. Analysis presented in this report points out that confidence intervals can be easily identified for varying hours of the day in the DMA. In the current report, only a time resolution of 15 minutes was used, because data was provided at that resolution.

A sensitivity analysis of data validation for the identification of faulty data (assuming all data outside the confidence interval is faulty) with a single parameter was explored. Such analysis demonstrates that: 1) additional fine tuning is required at each location and each utility, 2) that expert knowledge should be actively integrated in data validation models.

In the DMA of City B, customer consumption is obtained from billing records, and due to that, the time resolution is 1 year. Such resolution may not be enough to perform additional data validation of net consumption in the DMA by comparing net inflow vs net outflow.

6.1.3 Company C

Data of this utility was analyzed for water quality in a treatment plant, and for water balance in the DMA of City Db. Both datasets have a short duration due to an issue with the historical records in one of the databases for this area.

An analysis of a time series of chlorine was performed, given that such variable is not existent in the Netherlands. The results show that the chlorine has a significant correlation of less than 10 minutes and a change of average almost every 160 minutes. The second one must be a dosing event in the treatment plant.

For the water balance, it was not possible to perform an estimation of net flow in the DMA, as the system is currently interconnected with other close areas to City Db, for which no data was available.

It is concluded that a new data must be obtained in order to understand the operation of the DMA.

6.1.4 C-Town, benchmark case

For this system we show how to create a multivariate linear model based on Input Variable Selection (IVS). The response variable Tank 3, has a significant correlation with respect to the flow of pump 4, and of 3 different nodes in the system (not necessarily from the same DMA). A regression model is built based on these explanatory variables. Results show that the regression has a high correlation with the actual measured data.

With this statistical regression model, a subsequent validation is then performed by applying anomalies to a second dataset. Results show that a regression model, can reduce the number of unidentified anomalies. The protocol of simple tests proposed in this report is able to identify only 20% of anomalies, while the regression technique validates close to 70%.

It was therefore demonstrated that that the use of simulation models, and statistical models can be useful for the identification of faulty data in a DMA or water distribution system. Although the regression model seems to have better performance, than the protocol of validation, only a small data set was used for validation (2 weeks). Additional analysis and regression techniques can be applied for data validation.

6.2 General conclusions regarding data quality control

Data quality control is a continuous process instrumental in achieving large strategic objectives such as reliable and efficient drinking water system operations. Data validation for

water companies is not an exclusive task of a data scientist. Data validation is a collaboration between operators who understand the system and data analysts who must validate large proportions of data to improve models and, as a consequence, decisions. Currently only a small percentage of the data is validated but we suggest that in the near future there is a need to become even more active in data management at every level of a water company.

All involved water utilities implement individually data validation at different levels of complexity. Although water companies face similar issues, several customized tools/software are being developed per company. This is because each company has its own registration and database system, as well as different specifications regarding time steps, units, storage, etc. Working together on specific guidelines (standards) for the sector to define which datasets and methodologies are used for validation, can facilitate and speed up implementation of data quality control systems and be useful for potential future exchange of data, and may facilitate auditing operations as a nationwide goal in the near future.

From the pilots it was concluded that there is a need to exchange data and develop proper data models (i.e. metadata + format + platform).

There are several techniques to validate the data. Regarding data quality control, there is no such thing as one technique that fits all. Depending on the monitored event/variable, different techniques with different parameters should be applied also according the objective of the validation. Best practices identified during the pilots are: 1) encourage cooperation between data scientist and expert, 2) Define responsibilities e.g. who, when and how is the data validated and report validation rules, 3) Implement automatized routines, 4) In early stages, validate aggregated data first to speed up implementation, 5) Implement pilot projects to identify potentials for improvements.

Barriers identified during the pilot include:

- different types of data compression or interpolation and data storage are used even within one organization;
- only a small percentage of the data is validated;
- there is a lack of overview of metadata: difficulties to track additional information e.g. log books of maintenance, data is still stored in different databases;
- there is a lack of priorities/time to (at least) tag the large number of signals and known events,
- current approaches are often somewhat ad hoc. There are a lot of customized systems, some of them ‘Black boxes’, data models and tools. It is not always clear which rules are used to validate the data. The systems are not compatible with the ones of other companies even in some cases from the same vendor;
- a lot of techniques are applied, but still data validation/correction is (by hand) largely based on expert opinions which are never transformed into concepts or applications;

These barriers are further complicated by the following vulnerabilities:

- failure of servers, or
- changes in the system/models, such vulnerabilities cannot be traced because files are overwritten or deleted,
- data dependency from third parties (partly or not validated) and iv) issues related to the interfaces which result in data errors.

We conclude that a two-way implementation of DQC procedures is needed:

i. Strategic (Top-down) by developing frameworks and standards for the water sector which are compatible with standards of other sectors, and
ii. Operational (Bottom up) by implementing pilots, evaluating case studies, and sharing experiences across utilities.

To progress both, it is envisioned that utilities can start with simple cases and techniques such as the ones presented in this report and steadily scale-up as the needs and goals of the utilities are met.

7 Glossary

This section presents an overview of fundamental, recurring data-related terms that are used throughout this report.

Data Quality Control (DQC), as a term, encompasses all elements of the data quality improvement strategy in a company, extending beyond data validation and reaching into the sensing system itself, as well as the decision-making process based on the validated data. Management frameworks for Data Quality Control are based on the classic Plan-Do-Check-Act approach for production systems (Deming 1986, Ishikawa 1986), which refers to the design of a holistic, cyclic policy to ensure data quality.

Data stream is a term describing the transmission of data from a single source. According to (NITS 1996), a data stream is a sequence of digitally encoded signals used to represent information in transmission. In the water industry, such streams are composed of data coming from sensors in situ (Hill and Minsker 2010).

Data Validation refers to the process of determining the subset of faulty data and deciding whether it is required to perform an action on it (e.g. correct data) or flag it as potentially faulty data that can't be explained. Data validation is a key process within the adopted, general Data Quality Control framework that is responsible for identifying faulty data.

Data warehouse refers to a central repository of data integrated from one or more distinct sources (Ponniah 2010). A data warehouse can be considered synonymous to a **database** that stores and manages data from multiple sources. Sources which can be integrated include, among others: sensors, SCADA, smaller databases, business platforms, customer-relationship management systems. The advantage of data warehouses lies in the possibility to perform integrated analysis and reporting as a by-product of the integration of data into information. The goal of data warehouses is to transform information to make it ready for use by decision makers.

Faulty Data Detection is a term introduced by certain authors to characterize the identification of sensor data faults which may cause substantial performance degradation of all decision-making systems or processes that depends on data integrity for making decisions (Khorasani 2009). Within the larger framework of **Data Validation**, faulty data detection is considered the second step in the Collection-Detection-Correction three-step process, seen in FIGURE 3. Fault detection methods, such as built-in tests, typically log the time that the error occurred and either trigger alarms for manual intervention or initiate automatic recovery of information. The faulty detection techniques can be classified into qualitative and quantitative.

Fault Detection and Isolation is a synonymous term to **Data Validation** and corresponds to the ability to isolate anomalies which can be analyzed separately (V. Venkatasubramanian, R. Rengaswamy and K. Yin, et al. 2003).

8 References

Ackoff, R. L. 1989. "From data to wisdom." *Journal of Applied Systems Analysis* 15: 3-9.

Aisopou, A, I Stoianov, and N Graham. 2012. "In-pipe water quality monitoring in water supply systems under steady and unsteady state flow conditions: a quantitative assessment." *Water Research* 46: 235–246.

Aksela, K, M Aksela, and R Vahala. 2009. "Leakage detection in a real distribution network using a Self-Organizing Maps." *Urban Water* 279-289.

Alferes, J, S Tik, J Copp, and P Vanrolleghem. 2013. "Advanced monitoring of water systems using in situ measurement stations: data validation and fault detection." *Water Science & Technology* 1022-1030.

Allen, M, A Preis, M Iqbal, S Srirangarajan, H Lim, L Girod, and A Whittle. 2011. "The application of real-time in-network monitoring of the water distribution system to improve operational efficiency." *Journal of American Water Works Association (JAWWA)* 103: 63-75.

Anderson, E, and K Al-Jamal. 1995. "Hydraulic-network simplification." *J. Water Resour. Plann. Manage.* 121: 235–240.

Bakker, M. 2014. "Optimised control and pipe burst detection by water demand forecastingOptimised control and pipe burst detection by water demand forecasting." Ph.D. dissertation, Tu Delft. doi:10.4233/uuid:8d030ba9-da22-46d3-a018-f5d502e8d1d1.

Beran, J. 2010. "Long-range dependence." *Wiley Interdisciplinary Reviews: Computational Statistics* 2: 26-35. www.scopus.com.

Beran, J, Y Feng, S Ghosh, and R Kulik. 2013. *Long-memory processes: Probabilistic properties and statistical methods.* www.scopus.com.

Bertrand-Krajewski, J, J Bardin, M. Mourad, and Y Beranger. 2003. "Accounting for sensor calibration, data validation, measurement and sampling uncertainties in monitoring urban drainage systems." *Water Science and Technology* 47: 95-102.

Beuken, R, and A Moerman. 2017. *Uniforme storingsregistratie (USTORE). Praktijkcode voor het beheer van storingsregistratie van leidingnetten.* Nieuwegein, The Netherlands: KWR. PCD 9 2017.

Boutahar, M., V. Marimoutou, and L. Nouira. 2007. "Estimation methods of the long memory parameter: Monte Carlo analysis and application." *Journal of Applied Statistics* 34: 261-301. www.scopus.com.

Box, G, G Jenkins, and G Reinsel. 2008. *Time Series Analysis: Forecasting and Control.* www.scopus.com.

Bragalli, C, M Fortini, and E Todini. 2017. "Data Assimilation in Water Distribution Systems." *Procedia Engineering* 186: 506-513. doi:https://doi.org/10.1016/j.proeng.2017.03.263.

Bragalli, C, M Fortini, and E Todini. 2016. "Enhancing Knowledge in Water Distribution Networks via Data Assimilation." *Water Resources Management* 186: 3689–3706. doi:10.1007/s11269-016-1372-0.

Branisqvljevic, N, Z Kapelan, and D Prodanovic. 2011. "Improved real-time data anomaly detection using context classification." *Hydroinformatics.*

Candelieri, A, D Soldi, and F Archetti. 2014. "Short-term forecasting of hourly water consumption by using automatic metering readers data." *Procedia Engineering* 844-853.

Castelletti, A., S. Galelli, M. Ratto, R. Soncini-Sessa, and P. C. Young. 2012. "A general framework for Dynamic Emulation Modelling in environmental problems." *Environmental Modelling \& Software* 34: 5-18. doi:https://doi.org/10.1016/j.envsoft.2012.01.002.

Castelletti, A., S. Galelli, M. Restelli, and R. Soncini-Sessa. 2012. "Data-driven dynamic emulation modelling for the optimal management of environmental systems." *Environmental Modelling \& Software* 34: 30-43.

Castro-Gama, M., I. Popescu, S. Li, A. Mynett, and A. van Dam. 2014. "Flood inference simulation using surrogate modelling for the Yellow River multiple reservoir system." *Environmental Modelling & Software* 250-265.

Clarke, R.T. 2013. "Calculating uncertainty in regional estimates of trend in streamflow with both serial and spatial correlations." *Water Resources Research* 49: 7120-7125. www.scopus.com.

Cleveland, W. S. 1981. "Lowess: A program for smoothing scatterplots by robust locally weighted regression." *American Statistician* 35: 54-55. www.scopus.com.

Cohn, T. A., and H. F. Lins. 2005. "Nature's style: Naturally trendy." *Geophysical Research Letters* 32: 1-5. www.scopus.com.

Constantine, W., and D. Percival. 2014. "Fractal: Fractal time series modeling and analysis. R package version 2.0-0." *Fractal Time Series Modeling and Analysis.* www.scopus.com.

Darken, P. F., C. E. Zipper, G. I. Holtzman, and E. P. Smith. 2002. "Serial correlation in water quality variables: Estimation and implications for trend analysis." *Water Resources Research* 38: 221-227. www.scopus.com.

De Witte, K., en R. C. Marques. 2010. „Incorporating heterogeneity in non-parametric models: A methodological comparison." *International Journal of Operational Research* 9: 188-204. www.scopus.com.

Delignette-Muller, M. L., and C. Dutang. 2015. "fitdistrplus: An R package for fitting distributions." *Journal of Statistical Software* 64: 1-34. www.scopus.com.

Deming, W. E. 1986. *Out of the Crisis.* MIT Press. Cambridge, MA, page 88.

Di Zio, M., N. Fursova, T. Gelsema, S. Gießing, U. Guarnera, J. Petrauskienė, L. Quensel-von Kalben, et al. 2016. "Methodology for data validation 1.0." Tech. rep., Essnet Validat Foundation .

Donhost, Michael J., and Vincent A. Anfara. 2010. "Data-Driven Decision Making." *Middle School Journal* 42 (2): 56-63. doi:10.1080/00940771.2010.11461758.

EC. 2010. "SeadataNet Data Quality Control procedures." Tech. rep., 6ThFramework of EC DG Research.

Ehsanzadeh, E, and K Adamowski. 2010. "Trends in timing of low stream flows in Canada: Impact of autocorrelation and long-term persistence." *Hydrological Processes* 24: 970-980. www.scopus.com.

English, Larry P. 2001. "Information quality management: The next frontier." *ASQ World Conference on Quality and Improvement Proceedings. American Society for Quality.*

EPA. 2006. *Data Quality Assessment, A Reviewers Guide. EPA QA/G-9R.* Washington D.C.: US-Environmental Protection Agency.

Fatichi, S., S.M. Barbosa, E. Caporali, and M.E. Silva. 2009. "Deterministic versus stochastic trends: Detection and challenges." *Journal of Geophysical Research Atmospheres* 114. www.scopus.com.

Foster, G. 1996. "Wavelets for period analysis of unevenly sampled time series." *Astronomical Journal* 112: 1709-1729. www.scopus.com.

Furnival, G. M. 1971. "All Possible Regressions with Less Computation." *Technometrics* (Taylor & Francis) 13: 403-408. doi:10.1080/00401706.1971.10488794.

Gaag, B, and J Volz. 2008. "Real-time on-line monitoring of contaminants in water. Developing a research strategy from utility experiences and needs." Tech. rep., Kiwa Water Research.

Gaag, B., and J. Volz. 2007. "Realtime on-line monitoring of contaminants in water." Tech. rep., Kiwa Water Research.

Galelli, S., G.B. Humphrey, H.R. Maier, A. Castelletti, G.C. Dandy, and M.S. Gibbs. 2014. "An evaluation framework for input variable selection algorithms for environmental data-driven models." *Environmental Modelling and Software* 33–51.

Garside, M. J. 1965. "The best subset in multiple regression analysis." *Applied Statistics* 14: 196.

Giustolisi, O., D. Laucelli, L Berardi, and D. Savic. 2012. "Computationally Efficient Modeling Method for Large Water Network Analysis." *Journal of Hydraulic Engineering.*

Godsey, S. E., W. Aas, T. A. Clair, H. A. Wit, I. J. Fernandez, J. S. Kahl, I. A. Malcolm, et al. 2010. "Generality of fractal 1/f scaling in catchment tracer time series, and its implications

for catchment travel time distributions." *Hydrological Processes* 24: 1660-1671. www.scopus.com.

Graham, J. W. 2009. *Missing data analysis: Making it work in the real world.* Vol. 60. www.scopus.com.

Hadka, D. 2013. *Robust, adaptable many-objective optimization: the foundations, parallelization and application of the Borg MOEA.* Peen S: The Graduate School College of Engineering.

Hargesheimer, Erika, Osvaldo Conio, and Jarka Popovicova. 2002. "Online Monitoring for Drinking Water Utilities." Tech. rep., AWWA Research Foundation CRS PROAQUA.

Helsel, D. R., and R. M. Hirsch. 2002. "Statistical Methods in Water Resources." www.scopus.com.

Hill, and B. S. Minsker. 2010. "Anomaly detection in streaming environmental sensor data: A data-driven modeling approach." *Environmental Modelling and Software* 25: 1014–1022.

Hirsch, R, R Alexander, and R Smith. 1991. "Selection of methods for the detection and estimation of trends in water quality." *Water Resources Research* 27: 803-813. www.scopus.com.

Hirsch, R. M., D. L. Moyer, and S. A. Archfield. 2010. "Weighted regressions on time, discharge, and season (WRTDS), with an application to chesapeake bay river inputs." *Journal of the American Water Resources Association* 46: 857-880. www.scopus.com.

Hocking, R, and N Leslie. 1967. "Selection of the best subset in regression analysis." *Technometrics* 9: 531.

Hokkanen, J., and P. Salminen. 1997. "ELECTRE III and IV decision aids in an environmental problem." *Journal of Multi-Criteria Decision Analysis* 6: 215-226. www.scopus.com.

Hokkanen, J., and P. Salminen. 1997. "Locating a waste treatment facility by multicriteria analysis." *Journal of Multi-Criteria Decision Analysis* 6: 175-184. www.scopus.com.

Housh, and Z. Ohar. 2017. "Multiobjective Calibration of Event-Detection Systems." *Journal of Water Resources Planning and Management* 143. doi:10.1061/(ASCE)WR.1943-5452.0000808.

Hurst, H. E. 1951. "Long-term storage capacity of reservoirs." *Transactions of the American Society of Civil Engineers* 116: 770-808. www.scopus.com.

Ishikawa, Kaoru. 1986. *Guide to Quality Control.* Asian Productivity Organization. https://www.amazon.com/Guide-Quality-Control-Kaoru-Ishikawa/dp/9283310365?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=9283310365.

Kapelan, ZS, DA Savic, and GA Walters. 2005. "Multiobjective Design of Water Distribution Systems Under Uncertainty." *Water Resources Research.*

Khorasani, Ehsan Sobhani-Tehrani and Khashayar. 2009. "Fault Detection and Diagnosis." In *Fault Diagnosis of Nonlinear Systems Using a Hybrid Approach*, by Ehsan Sobhani-Tehrani and Khashayar Khorasani. Springer US.

Kirchner, J. W. 2005. "Aliasing in 1fα noise spectra: Origins, consequences, and remedies." *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 71. www.scopus.com.

Kitchin, Rob. 2014. "Big Data, new epistemologies and paradigm shifts." *Big Data & Society* 1 (1). doi:10.1177/2053951714528481.

Kwakkel, Jan H., Marjolijn Haasnoot, and Warren E. Walker. 2016. "Comparing Robust Decision-Making and Dynamic Adaptive Policy Pathways for model-based decision support under deep uncertainty." *Environmental Modelling \& Software* 86: 168-183. doi:https://doi.org/10.1016/j.envsoft.2016.09.017.

Kwakkel, M., I. Vloerbergh, P. van Thienen, R. Beuken, B. Wols, and K. van Daal. 2015. "2015, Uniform failure registration: from data to knowledge,." *Water asset management international* 10 (4): 18-22.

Lennartz, S, and A Bunde. 2009. "Trend evaluation in records with long-term memory: Application to global warming." *Geophysical Research Letters* 36. www.scopus.com.

Leunk, I. 2014. *Kwaliteitsborging grondwaterstands- en stijghoogtegegevens. Validatiepilot; analyse van bestaande data.; Rapportnr. KWR 2014.059.* Nieuwegein: KWR Watercycle Research Institute.

Leunk, I. 2014. *Kwaliteitsborging grondwaterstands- en stijghoogtegegevens: Validatiepilot, analyse van bestaande data KWR 2014.059.* Nieuwegein: KWR Watercycle Research Institute.

Lomb, N. R. 1976. "Least-squares frequency analysis of unequally spaced data." *Astrophysics and Space Science* 39: 447-462. www.scopus.com.

Lynggaard-Jensen, A., H.P. Hansen, and J.L. Bertrand-Krajewski. 2012. *Real time integrated monitoring system supporting new data validation methods - Methodology guidelines and examples of application.*

Makropoulos, Christos, Peter van Thienen, and Claudia Agudelo-Vera. 2018. "Towards a roadmap for hydroinformatics research BTO 2018.077." Tech. rep., KWR Watercycle Research Institute.

Marques, R. C., and K. De Witte. 2011. "Is big better? On scale and scope economies in the Portuguese water sector." *Economic Modelling* 28: 1009-1016. www.scopus.com.

Martínez-Solano, F., Iglesias-Rey, P., Mora-Meliá, D., Fuertes-Miquel, V.S. 2017. "Exact skeletonization method in water distribution systems for hydraulic and quality models." *Procedia Engineering* 186: 286 – 293. https://ac.els-cdn.com/S1877705817313991/1-s2.0-S1877705817313991-main.pdf?_tid=spdf-c8508d91-8b94-475c-bdb9-256f8b2c469b&acdnat=1519906589_45a30437ee48e8623570630a9b10ac20.

Mashford, John, Dhammika De Silva, Donavan Marney, and Stewart Burn. 2009. "An Approach To Leak Detection In Pipe Networks Using Analysis Of Monitored Pressure Values By Support Vector Machine." *Proceedings of the 3rd International Conference on Network and System Security* 1-6.

McAfee, Andrew, Erik Brynjolfsson, Thomas H Davenport, D J Patil, and Dominic Barton. 2012. "Big data: the management revolution." *Harvard business review* 90 (10): 60-68.

McKenna, H.D., Klise, K., Cruz, V., Wilson, M. 2007. "Event detection from water quality time series." *Proceedings of World Environmental and Water Resources Congress, ASCE, Reston, VA.*

Mesman, G., and P. van Thienen. 2015. *Lekzoeken met hydraulische modellen. BTO 2015.064.* Nieuwegein, NL: KWR Watercycle Research Institute.

Moerman, A., R.H.S. Beuken, and B.A. Wols. 2017. "Review on the development of uniform failure registration (USTORE) in the netherlands,." *LESAM conference.* Trondheim.

Montanari, A., M. S. Taqqu, and V. Teverovsky. 1999. "Estimating long-range dependence in the presence of periodicity: An empirical study." *Mathematical and Computer Modelling* 29: 217-228. www.scopus.com.

Montanari, A., R. Rosso, and M. S. Taqqu. 2000. "A seasonal fractional ARIMA model applied to the Nile River monthly flows at Aswan." *Water Resources Research* 36: 1249-1259. www.scopus.com.

Montanari, A., R. Rosso, and M. S. Taqqu. 1997. "Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation." *Water Resources Research* 33: 1035-1044. www.scopus.com.

Morrison, D. F. 1976. *Multivariate Statistical Methods.* www.scopus.com.

Mounce, R. Mounce, T. Jackson, J. Austin, and J.B. Boxall. 2014. "Pattern matching and associative artificial neural networks for water distribution system time series data analysis." *Journal of Hydroinformatics* (IWA Publishing) 16: 617-632. doi:10.2166/hydro.2013.057.

Mounce, S.R., R.B. Mounce, and J.B. Boxall. 2011. "Novelty detection for time series data analysis in water distribution systems using Support Vector Machines." *Journal of Hydroinformatics* 672–686.

Mourad, M., and J.L. Bertrand-Krajewski. 2002. "A method for automatic validation of long time series of data in urban hydrology." *Wat. Sci. Tech.* 45: 263–270.

Mudumbe, Mduduzi John, and Adnan M. Abu-Mahfouz. 2015. "Smart water meter system for user-centric consumption measurement." *2015 IEEE 13th International Conference on Industrial Informatics (INDIN).* 993-998. doi:10.1109/INDIN.2015.7281870.

Mutchek, Michele, and Eric Williams. 2014. "Moving Towards Sustainable and Resilient Smart Water Grids." *Challenges* 5 (1): 123-137. doi:10.3390/challe5010123.

Neter, J., W. Wasserman, and M. H. Kutner. 1990. *Applied Linear Statistical Models.* www.scopus.com.

NITS. 1996. "Federal Standard 1037C data stream." *ITS.* Aug 23. Accessed 10 20, 2018. https://www.its.bldrdoc.gov/fs-1037/dir-010/_1451.htm.

Oriani, Fabio, Andrea Borghi, Julien Straubhaar, Gregoire Mariethoz, and Philippe Renard. 2016. "Missing data simulation inside flow rate time-series using multiple-point statistics." *Environmental Modelling & Software* 264-276.

Ostfeld, A., E. Salomons, L. Ormsbee, J.G. Uber, C. M. Bros, P. Kalungi, R. Burd, B. Zazula-Coetzee, T. Belrain, and D., Kang. 2011. "Battle of the water calibration networks." *Journal of Water Resources Planning and Management.*

Palau, C, F Arregui, and M Carlos. 2012. "Burst detection in water networks using principal component analysis." *Journal of Water Resources Planning and Management* 138 (1): 47-54.

Pannekoek, J., S. Scholtus, and M. van der Loo. 2013. "Automated and Manual Data Editing: A View on Process Design and Methodology." *Journal of Official Statistics* 29: 511-537. doi:10.2478/jos-2013-0038.

Paté-Cornell, Elisabeth. 2012. "On “Black Swans” and “Perfect Storms”: Risk Analysis and Management When Statistics Are Not Enough." *Risk Analysis* 32 (11): 1823-1833.

Perelman, L., and A. Ostfeld. 2011. "Water Distribution Systems Simplifications through Clustering." *J. Water Resour. Plan. Management* 138 (6): 218–229.

Ponniah, Paulraj. 2010. *Data warehousing fundamentals for IT professionals.* Second. John Wiley & Sons.

Post, V.E.A., and Von Asmuth, J.R. 2013. "Hydraulic head measurements: New technologies, classic pitfalls." *Hydrogeology journal* 10.1007/s10040-013-0969-0.

Poulakis, Z., D. Valougeorgis, and C. Papadimitriou. 2003. "Leakage detection in water pipe networks using a Bayesian probabilistic framework." *Probab. Eng. Mech.* 315-327.

Preis, A., Whittle A.J., Ostfield A., Perelman L. 2011. "Efficient hydraulic state estimation technique using reduced models of urban water networks." *Journal of Water Resources Planning \& Management* 137: 343-351.

Quinlan, J. R. 1992. "Learning with continuous classes." *Proc. 5th Australian Joint Conf. on Artificial Intelligence, World Scientific, Singapore.* 343–348.

Ragsdale, C. T. 2001. "Spreadsheet Modeling and Decision Analysis." www.scopus.com.

Ratto, M., A. Castelletti, and A. Pagano. 2012. "Emulation techniques for the reduction and sensitivity analysis of complex environmental models." *Environmental Modeling and Software* 1-4. doi:10.1016/j.envsoft.2011.11.003.

Rea, W., L. Oxley, M. Reale, and J. Brown. 2009. "Estimators for long range dependence: An empirical study." *Electron.J.Stat.* www.scopus.com.

Redman, Thomas C. 1998. "The impact of poor data quality on the typical enterprise." *Communications of the ACM* 41 (2): 79-82.

Rossman, L.A. 2000. "EPANET 2: Users Manual. EPA/600/R-00/057." *U.S. Environmental Protection Agency, Washington, D.C.*

Salminen, P., J. Hokkanen, and R. Lahdelma. 1998. "Comparing multicriteria methods in the context of environmental problems." *European Journal of Operational Research* 104: 485-496. www.scopus.com.

Sang, Y., Z. Wang, and C. Liu. 2014. "Comparison of the MK test and EMD method for trend identification in hydrological time series." *Journal of Hydrology* 510: 293-298. www.scopus.com.

Schatzoff, M, S Fienberg, and R Tsao. 1968. "Efficient calculations of all possible regressions." *Technometrics* 10: 768.

Scheel, H. 2000. "EMS: Efficiency Measurement System User's Manual." www.scopus.com.

Sengupta, J. K. 1990. "Tests of efficiency in data envelopment analysis." *Computers and Operations Research* 17: 123-132. www.scopus.com.

Shamir, U., and E. Salomons. 2008. "Optimal real-time operation of urban water distribution systems using reduced models." *J. Water Resour. Plann. Manage.* 134: 181-185.

Shewhart, Walter A. 1931. *Economic Control of Quality Of Manufactured Product.* Martino Fine Books.
https://www.ebook.de/de/product/24077831/walter_a_shewhart_economic_control_of_quality_of_manufactured_product.html.

Shi, H., J. Gong, A. C. Zecchin, M. F. Lambert, and A. R. Simpson. 2017. "Hydraulic transient wave separation algorithm using a dual-sensor with applications to pipeline condition assessment." *Hydroiinformatics* 19: 752-765. doi:10.2166/hydro.2017.146.

Simon, A. 2013. "ESS.VIP.BUS Common data validation policy, Deliverable 2-5 Definition of “validation levels” and other related concepts." Tech. rep., European Commission – Eurostat/B1, Eurostat/E1, Eurostat/E6 .

Stausberg, Jürgen, Michael Nonnemacher, Dorothea Weiland, Gisela Antony, and Markus Neuhäuser. 2006. "Management of data quality - development of a computer-mediated guideline." *Studies in health technology and informatics* 124: 477-482.

Sun, S., J. Bertrand-krajewski, A. Lynggaard-Jensen, J. Broeke, F. Edthofer, M. Céu Almeida, M. Silva Ribeiro, and J. Menaia. 2011. "Literature Review of Data Validation Methods." Tech. rep.

Taormina, Galelli S. Ole Tippenhauer N. et al. 2018. "The BATtle of the Attack Detection ALgorithms: disclosing cyber attacks on water distribution networks." *Journal of Water Resources Planning and Management.* http://www.batadal.net/data.html.

Taqqu, M. S., V. Teverovsky, and W. Willinger. 1995. "Estimators for long-range dependence: An empirical study." *Fractals* 3: 785-798. www.scopus.com.

Tavallali, P., M. Razavi, and S. Brady. 2017. "A non-linear data mining parameter selection algorithm for continuous variables." *PLoS ONE* 12: e0187676. doi:https://doi.org/10.1371/journal.pone.0187676.

Tayi, Giri Kumar, and Donald P Ballou. 1998. "Examining Data Quality." *Communications of the ACM* 41 (2): 54-57. doi:10.1145/269012.269021.

Team, Core. 2013. "R: A language and environment for statistical computing." *R: A Language and Environment for Statistical Computing*. www.scopus.com.

Thienen, P., H.-J. Alphen, A. Brunner, Y. Fujita, B. Hillebrand, R. Sjerps, J. Summeren, A. Verschoor, and B. Wullings. 2018. "Explorations in Data Mining for the Water Sector BTO 2018.085." Tech. rep., KWR Watercycle Research Institute.

Thienen, P.V., I. Pieterse-Quirijnse, H.D. Kater, and J. Duifhuizen. 2012. "Nieuwe lekverliesbepalingsmethoden voor het drinkwaterdistributienet." *H2O.*

Thienen, Peter van, and Ina Vertommen. 2015. "Automated feature recognition in CFPD analyses of DMA or supply area flow data." *Journal of Hydroinformatics* 18 (3): 514-530.

Todini, and L. Rossman. 2013. "Unified Framework for Deriving Simultaneous Equation Algorithms for Water Distribution Networks." *Journal of Hydraulic Engineering* 139: 511-526.

Topcu, O. 2003. "Review of Verification and Validation Methods in Simulation: Literature Survey, Concepts, and Definitions." Tech. rep., DRDC Atlantic TM 2003-055 Defence R&D Canada – Atlantic. http://cradpdf.drdc-rddc.gc.ca/PDFS/unc13/p520183.pdf.

Tsoukalas, I, P Kossieris, A Efstratiadis, and C. Makropoulos. 2016. "Surrogate-enhanced evolutionary annealing simplex algorithm for effective and efficient optimization of water resources problems on a budget." *Environmental Modelling & Software 77* 122-42.

Ueda, T., and Y. Hoshiai. 1997. "Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs." *Journal of the Operations Research Society of Japan* 40: 477-478. www.scopus.com.

Ulanicki, B., Zehnpfund, A., Martinez, F. 1996. "Simplification of water distribution network models." *Proc., 2nd Int. Conf. on Hydroinformatics, Zurich, Switzerland,*. 493–500.

van den Broek, B., M. Loo, and J. Pannekoek. 2014. "Kwaliteitsmaten voor het datacorrectieproces." *Statistics Netherlands.*

van Selst, M., and P. Jolicoeur. 1994. "A Solution to the Effect of Sample Size on Outlier Elimination." *The Quarterly Journal of Experimental Psychology Section A* 47: 631-650. www.scopus.com.

van Thienen, Peter, I. Pieterse-Quirijnse, H.D. Kater, and J. Duifhuizen. 2012. *Nieuwe lekverliesbepalingsmethoden voor het drinkwaterdistributienet.* H2O.

Veling, E. 2010. "Approximations of impulse response curves based on the generalized moving Gaussian distribution function." *Advances in Water Resources* 33: 546-561. https://ac.els-cdn.com/S0309170810000412/1-s2.0-S0309170810000412-main.pdf?_tid=bc69be80-8a64-47e2-ab81-69e97488e2e5&acdnat=1520408993_b68c1dd8ae1b0bc7c56f47364e5a8dc2.

Veling, E.J.M. 2010. "Approximations of Impulse Response Curves based on the Generalized Moving Gaussian Distribution Function." Ph.D. dissertation, Faculty of Civil Engineering and Geosciences, Department of Water Management at TU Delft. https://repository.tudelft.nl/islandora/object/uuid:12512821-5563-4244-aa46-6c460b745148?collection=research.

Venkatasubramanian, V, R Rengaswamy, K Yin, and S Kavuri. 2003. "A review of process fault detection and diagnosis, Part I: Quantitative methods." *Computers & Chemical Engineering* (Elsevier BV) 27: 293-311. doi:10.1016/s0098-1354(02)00160-6.

Venkatasubramanian, V, R Rengaswamy, S Kavuri, and K Yin. 2003. "A review of process fault detection and diagnosis, Part III: Process history based methods." *Computers & Chemical Engineering* (Elsevier BV) 27: 327-346. doi:10.1016/s0098-1354(02)00162-x.

Venkatasubramanian, Venkat, Raghunathan Rengaswamy, and Surya N. Kavuri. 2003. "A review of process fault detection and diagnosis, Part II: Qualitative models and search strategies." *Computers & Chemical Engineering* (Elsevier BV) 27: 313-326. doi:10.1016/s0098-1354(02)00161-8.

von Asmuth, J. 2012. "Groundwater System Identification through Time Series Analysis." Ph.D. dissertation, Faculty of Civil Engineering and Geosciences, Department of Water Management at TU Delft. https://repository.tudelft.nl/islandora/object/uuid:b6ccd472-9b9d-4810-aa19-3a0b046017e0?collection=research.

von Asmuth, J. 2015. *Kwaliteitsborging grondwaterstands- en stijghoogtegegevens: Protocol voor datakwaliteitscontrole (QC) KWR 2015.013.* Nieuwegein: KWR Watecycle Research Insititute {\&} TNO.

von Asmuth, J. 2011. *Over de kwaliteit, frequentie en validatie van druksensorreeksen. KWR 2010.001.* Nieuwegein: KWR Watercycle Research Institute.

von Asmuth, J, and F van Geer. 2013. *Kwaliteitsborging grondwaterstands- en stijghoogtegegevens: op weg naar een landelijke standaard. KWR 2013.027.* Nieuwegein / Utrecht.: KWR Watercycle Research Institute / TNO.

von Asmuth, J, and F van Geer. 2015. *Kwaliteitsborging grondwaterstands- en stijghoogtegegevens: Systematiek en methodiek voor datakwaliteitscontrole (QC).* Nieuwegein: KWR Watecycle Research Insititute {\&} TNO.

von Asmuth, J, and F. van Geer. 2015. *Kwaliteitsborging grondwaterstands- en stijghoogtegegevens: Systematiek en methodiek voor datakwaliteitscontrole (QC). KWR 2015.004.* Nieuwegein / Utrecht: KWR Watercycle Research Institute / TNO.

von Asmuth, J, K Maas, M Knotters, M Bierkens, M Bakker, T Olsthoorn, D Cirkel, I Leunk, F Schaars, and D Asmuth. 2012. "Menyanthes software for hydrogeologic time series analysis, interfacing data with physical insight." *Environmental Modelling \& Software.* http://waterware.kwrwater.nl/uploadedFiles/Website_KWR_Waterware/Menyanthes/asmuth_menyanthes.pdf.

Vreeburg, J.H.G., I.N. Vloerbergh, P. Van Thienen, and R. De Bont. 2013. "Shared failure data for strategic asset management,." *Water Science & Technology: Water Supply 13(4):* 1154-1160.

Vries, van den Akker, Vonk, de Jong, and van Summeren. 2016. "Application of machine learning techniques to predict anomalies in water supply networks." *Water Science and Technology, Water Suppy* 16 (6): 1528-1535.

Waal, T. 1996. "CherryPi: A Computer Program for Automatic Edit and Imputation." Tech. rep., UN/ECE Work Session on Statistical Data Editing, Voorburg.

Waal, T. 2013. "Selective Editing: A Quest for Efficiency and Data Quality." *Journal of Official Statistics* 29: 473-488. doi:10.2478/jos-2013-0036.

—. 2001. "SLICE: Generalised Software for Statistical Data Editing."

Wang, Z. 2013. "cts: An R package for continuous time autoregressive models via Kalman filter." *Journal of Statistical Software* 53: 1-19. https://cran.r-project.org/web/packages/cts/vignettes/kf.pdf.

Wilson, P. W. 1993. "Detecting outliers in deterministic nonparametric frontier models with multiple outputs." *Journal of Business and Economic Statistics* 11: 319-323. www.scopus.com.

Wilson, P. W. 2008. "FEAR: A software package for frontier efficiency analysis with R." *Socio-economic planning sciences* 42: 247-254. www.scopus.com.

Yoo, C.K., K. Villez, I.B. Lee, S. Van Hulle, and P.A. Vanrolleghem. 2006. "Sensor validation and reconciliation for a partial nitrification process." *Water Science and Technology* 53 (4-5): 513-521.

Young, Peter C., Wlodzimierz Tych, and C. James Taylor. 2009. "The Captain Toolbox for Matlab." *IFAC Proceedings Volumes* 42: 758-763. doi:https://doi.org/10.3182/20090706-3-FR-2004.00126.

Yue, S., P. Pilon, B. Phinney, en G. Cavadias. 2002. „The influence of autocorrelation on the ability to detect trend in hydrological series." *Hydrological Processes* 16: 1807-1829. www.scopus.com.

Zeileis, A., and G. Grothendieck. 2005. "Zoo: S3 infrastructure for regular and irregular time series." *Journal of Statistical Software* 14. www.scopus.com.

Zetterqvist, L. 1991. "Statistical estimation and interpretation of trends in water quality time series." *Water Resources Research* 27: 1637-1648. www.scopus.com.

Zhang, Q., and W. P. Ball. 2017. "Improving riverine constituent concentration and flux estimation by accounting for antecedent discharge conditions." *Journal of Hydrology* 547: 387-402. www.scopus.com.

Zhang, Q., C. J. Harman, and J. W. \& Kirchner. 2018. "Evaluation of statistical methods for quantifying fractal scaling in water-quality time series with irregular sampling." *Hydrology and Earth System Sciences* 22: 1175-1192. doi:10.5194/hess-22-1175-2018.

Zhang, Q., D. C. Brady, W. R. Boynton, and W. P. Ball. 2015. "Long-Term Trends of Nutrients and Sediment from the Nontidal Chesapeake Watershed: An Assessment of Progress by River and Season." *Journal of the American Water Resources Association* 51: 1534-1555. www.scopus.com.

Zhu, J. 2003. "Quantitative models for performance evaluation and benchmarking: Data envelopment analysis with spreadsheets." *Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets.* www.scopus.com.

Appendix 1

LIST OF TABLES

LIST OF FIGURES