UPPSALA
UNIVERSITET

Error detection in wastewater treatment plants using mass balances

Maja Karlsson

Abstract

Error detection in wastewater treatment plants using mass balances

Maja Karlsson

Process data from wastewater treatment plants are often corrupted by errors. These data provide a basis for operating the plant, therefore effort should be made to improve the data quality. Currently, Stockholm Vatten och Avfall uses a method where they quantitatively verify water flow measurement data by comparing it to water level measurements. In this thesis, an alternative approach based on mass balancing to detect errors was evaluated. The aim was to find, implement and evaluate a mass balance based method to detect and locate errors. The objective was to use this method to corroborate the flow verification method used by Stockholm Vatten och Avfall, and to improve flow data from Bromma Wastewater treatment plant. The chosen method consisted of two major steps, gross error detection and data reconciliation. A case study was performed where the method was tested on both simulated data with known added errors, real process data and finally a case where the suggested method was compared to the flow verification method. The results showed that this method was efficient in detecting a gross error when only one flow measurement was erroneous and that the estimation of the error magnitude was good. However, the suggested method was not useful for corroboration of the flow verification method. With the flow verification method, the flow in one filter basin at the time was examined. The suggested method required the combined flow in all 24 filter basins, which made it difficult to compare the two methods. The method has potential to be valuable for error detection in wastewater treatment plants, and to be used as a live tool to detect gross errors.

Key words: mass balance, wastewater treatment plant, error detection, data reconciliation, gross errors, random errors, bias

Referat

Feldetektion med hjälp av massbalanser i avloppsreningsverk

Maja Karlsson

Processdata från avloppsreningsverk innehåller ofta fel. Dessa data utgör basen för att driva reningsverket, därför bör resurser läggas på att kontrollera och förbättra datakvaliteten. I dagsläget använder Stockholm Vatten och Avfall en metod som kontrollerar vattenflöden genom att jämföra flödesmätningar med vattennivåmätningar. I det här examensarbetet testades en alternativ metod baserad på massbalansering för att detektera och lokalisera fel. Syftet var att hitta, implementera och utvärdera en massbalansbaserad feldetektion-smetod. Målet var att använda denna metod för att utvärdera flödesverifieringsmetoden som Stockholm Vatten och Avfall använder, samt att förbättra flödesdata från Bromma Reningsverk. Den valda metoden bestod av två steg: detektion av systematiska fel och databalansering. En fallstudie utfördes, där testades metoden på simulerad data med kända fel, på verklig flödesdata samt i ett sista fall där den föreslagna metoden jämfördes med flödesverifieringsmetoden. Resultaten visade att den föreslagna metoden effektivt kan detektera systematiska fel när endast en flödesmätning var felaktig samt att uppskattningen av felens magnitud var bra. Dock var den föreslagna metoden inte användbar för att verifiera Stockholm Vatten och Avfalls metod. Med flödesverifieringsmetoden undersöks flödet i ett filter åt gången, medan den föreslagna metoden kräver att flödet i alla 24 filter summeras. Detta gjorde att det var svårt att jämföra de två metoderna på ett bra sätt. Metoden har potential att vara värdefull för feldetektion i avloppsreningsverk samt att användas som ett realtidsverktyg för att detektera fel.

Nyckelord: massbalans, avloppsreningsverk, feldetektion, databalansering, systematiska fel, slumpmässiga fel, bias

Institutionen för informationsteknologi, Uppsala universitet (UU). Lägerhyddsvägen 2, SE 752-37 UPPSALA.

Populärvetenskaplig sammanfattning

För att underhålla och förbättra avloppsreningsverk är det viktigt att noga kontrollera processen. Drift, processmodellering och ekonomiska utvärderingar kräver att data insamlas. Exempelvis mäts halter av fosfor, kväve samt hur mycket vatten som flödar genom de olika delarna av anläggningen. Mätningarna sker kontinuerligt vilket genererar stora datamängder. Processdata som insamlats innehåller ofta fel vilket t.ex. kan bero på att ett mätinstrument täpps igen med smuts eller att mätaren är felkalibrerad. För att inte beslut ska baseras på felaktig information är det viktigt att försöka hitta dessa fel.

Processdata från reningsverk kan innehålla två olika typer av fel: systematiska- och slumpmässiga fel. Systematiska fel kan uppstå t.ex. när en mätare täpps och därigenom långsamt börjar visa ett lägre värde än det verkliga. Slumpmässiga fel är alltid närvarande i processdata och uppstår vanligen av oregelbundna störningar. Systematiska fel är generellt större än slumpmässiga, och svårare att detektera.

I dagsläget kontrollerar Stockholm Vatten och Avfall flödesdata i sandfilterbassängerna genom att jämföra dessa med vattennivåmätningar. Från denna jämförelse fås information om någon av mätinstrumenten ger felaktiga data, dock inte om felet ligger hos nivå- eller flödesmätaren. De önskar att verifiera sin metod. Syftet är därför att hitta, tillämpa och utvärdera en alternativ metod. Målet är att med hjälp av denna metod verifiera metoden som Stockholm Vatten och Avfall använder, samt att förbättra flödesmätningar från Bromma reningsverk.

Metoden som valdes är baserad på massbalansering. Massbalansering innebär att undersöka hur olika flöden i ett system förhåller sig till varandra. Ingen massa försvinner eller tillkommer i ett slutet system. Den föreslagna metoden bygger på algebra och statistisk analys, och består av två huvudsteg. Först detekteras systematiska fel, därefter balanseras data för att reducera slumpmässiga fel.

Ett system med fyra flöden på Bromma reningsverk undersöks, däribland flödet genom sandfilterbassängerna. Innan metoden testades på riktig flödesdata utvärderades den på simulerade data med tillagda fel som ökade med tiden. Resultaten visade att felet detekterades relativt snabbt, och uppskattningen av felets storlek stämde överens med det kända tillagda felet. Med hjälp av metoden kan fel detekteras i verkliga data om endast ett av de ingående flödena avviker från balansekvationerna, det vill säga innehåller fel. Om fler flöden innehåller fel var resultaten svåra att tolka. För att lösa det problemet bör större system med fler flöden undersökas. Det visade sig också att metoden inte var lämplig att använda som verifiering av Stockholm Vattens och Avfalls metod. Detta eftersom Stockholm Vatten och avfalls metod utnyttjar information från en flödesmätare åt gången, medan den föreslagna metoden kräver att flödet i alla 24 sandfilter summeras. Därigenom blev metoderna svåra att jämföra.

Den föreslagna metoden genererar ibland svårtolkade resultat, men har potential att vara värdefull för feldetektion i avloppsreningsverk och att användas som ett realtidsverktyg för att detektera fel. Metodiken är generell och tillämpbar på olika typer av system där energi- eller massbalanser kan ställas upp.

Preface

This master thesis is the final part of the Master Programme in Environmental and Water Engineering at Uppsala University and corresponds to 30 c. The thesis was performed at IVL Swedish Environmental Research Institute, Stockholm, and data was received from Stockholm Vatten och Avfall. Supervisor was Oscar Samuelsson, Postgraduate student at IVL Swedish Environmental Research Institute. Subject reviewer has been Bengt Carlsson, Professor at the Department of Information Technology, Division of System and Control Research at Uppsala University.

I would like to thank my supervisor Oscar Samuelsson for great support and guidance throughout the project. Also, I would like to thank Erik Lindblom and Anders Pålsson at Stockholm Vatten och Avfall for setting aside time to answer all my questions.

Contents

1 Introduction

1.1 Motivation

In most wastewater treatment plants (WWTPs) process data are collected, and normally a large amount of it. Common process parameters that are measured are e.g. water flow, total phosphorus and total nitrogen. These data provide a basis for operating the plant, process modelling and control, economic evaluations and so on. These actions rely heavily on the accuracy of data [Thomann, 2008].

The large amount of process data inevitably contain errors and efforts should be made to improve the data quality. These errors can originate from the harsh environment where the measuring equipment has to function. Process measurements can be corrupted by two types of errors, random- and systematic (gross) errors. Random errors are mainly measurement noise which is unavoidable and always present in process data. Gross errors are often caused by malfunctioning measuring instruments. By maintaining the measuring equipment, gross errors can be avoided to some extent [Narasimhan and Jordache, 2000].

To improve the accuracy of process data from WWTPs, several authors (e.g. [Meijer et al., 2002; Puig et al., 2008; Seungchul et al., 2015]) propose the use of data reconciliation (DR) combined with a statistical method to detect gross errors. In their suggested approaches, detection of errors are performed by checking mass balances of WWTPs. These methods provide valuable new information about the process [Seungchul et al., 2015], and errors where successfully detected in each study.

Error detection methods based on mass balancing are a well-known and commonly used technique to improve process data. These techniques have been applied in e.g. petrochemical plants, chemical plants, refineries and mineral processing industries [Narasimhan and Jordache, 2000]. Studies related to WWTPs are though rarely performed [Seungchul et al., 2015].

Only when a compound is completely recoverable from in- and outgoing flows, it can be balanced. In practice, many compounds, e.g. nitrogen and organic matter, converts into gas during the treatment process. Generally this is difficult to measure. Phosphorus do not convert to gas, and the amount of water that convert to gas is negligible. Therefore, all flows can be measured, which makes water and phosphorus suitable for mass balancing [Meijer et al., 2002].

In this thesis, focus will be set on finding a systematic method to detect errors using mass balancing in process data from Stockholm Vatten och Avfall. They wish to verify the current flow verification method used at Bromma WWTP, which will be one of the goals of this project. Flow measurements lay the foundation for process control, mass balancing and modelling in WWTPs, and are measured continuously. Therefore, flow measurements of water will be used for mass balancing and the term *flow balancing* will be used henceforth.

1.2　Problem formulation

Data from flow gauges are never perfect. There are always random errors present (random noise), which often is normally distributed [Exell, 2001]. Systematic gross errors include errors that are not normally distributed, e.g. when a gauge consistently measure too high or low flow. This type of errors are more complicated to detect. Gross errors are important to detect in order not to make decisions based on inaccurate information.

Stockholm Vatten och Avfall would like to corroborate their flow verification method, and to increase the reliability in flow measurements at Bromma WWTP. The current method to check flow data at Bromma WWTP is a comparison between measurements from flow gauges and level gauges in the sand filter section. The difference in water level and volume leaving the sand filter basins are compared during a certain period of time. From this comparison, it can be determined if the measuring devices are providing inaccurate data. However, it can not be determined in which gauge (flow or level) the error lies in.

An alternative approach to verify the accuracy of data is the use of flow balancing. Knowing how the different flows relate to one another enables detection, diagnosis and estimation of both random and gross errors. A flow balance based method could therefore provide valuable new information. To corroborate the flow verification method and to check data for errors, a method based on flow balancing will be used in this thesis.

1.2.1　Aim and objective

The aim of this project is to find, implement and evaluate a method based on flow balancing to detect and locate random and gross errors.

The objective is to use this method to corroborate a flow verification method currently used by Stockholm Vatten och Avfall, and to improve the accuracy of flow measurements at Bromma WWTP.

1.2.2　Limitations

The following limitations were made

- Only a section of Bromma WWTP was considered: influent, aeration basins, filter basins and effluent. Thus, four flows were analyzed. This could limit the method if more than one flow is erroneous.

- The choice of periods to study was narrowed due to limited access of data for the effluent flow and missing data for when the inlet hatches opens and closes. This will be further explained in the method description.

2 Background and theory

2.1 Bromma WWTP

Bromma WWTP has two facilities, Åkeshov and Nockeby. At Åkeshov, there is pre-treatment, pre-sedimentation, sludge treatment and biogas production. At Nockeby there is an activated sludge water treatment plant and a filter plant. Every day, 126 000 m^3 wastewater is treated, which is wastewater from over 300 000 people in Stockholm [Stockholm Vatten, ND]. An illustration of the section of Bromma WWTP which will be considered in this study can be seen in Fig. 2.1.
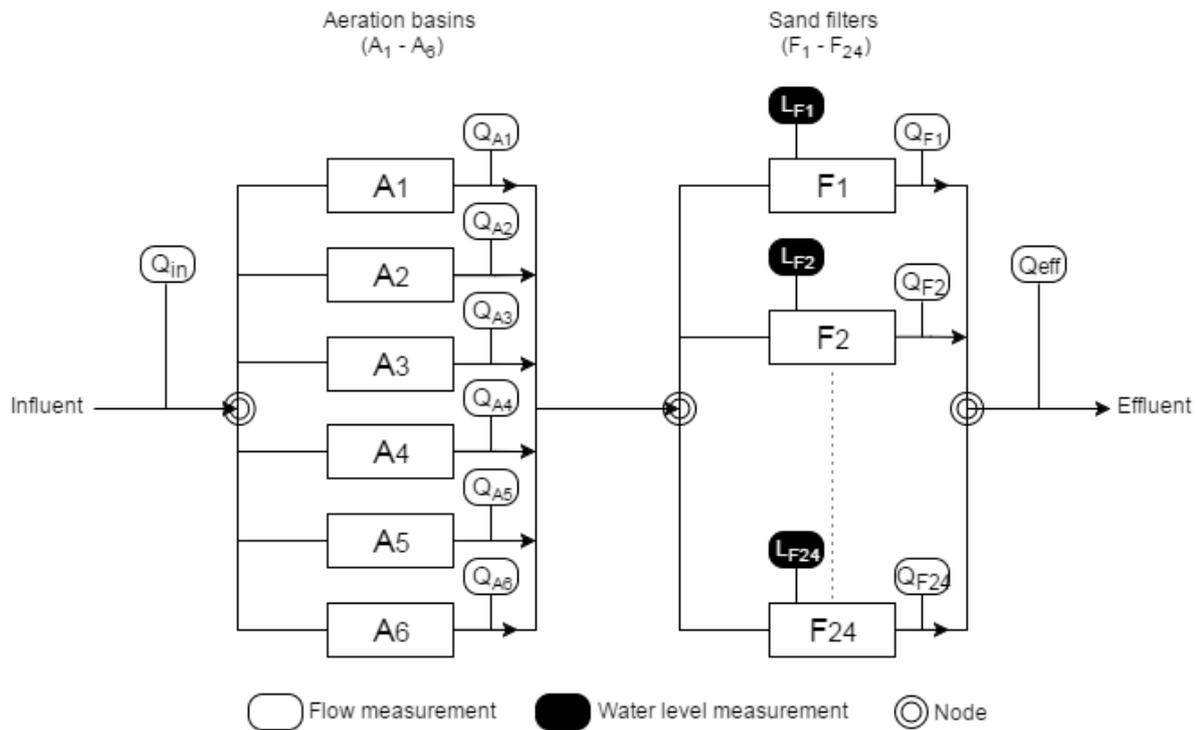

Figure 2.1. The section of Bromma WWTP which was considered in this study.

The influent flow is a corrected flow; measured flows of excess sludge and flushing water is removed from the measured influent flow to obtain the actual influent flow. This will be refereed to as Q_{In}. The uncorrected influent flow and the effluent flow (Q_{Eff}) are measured using a venturi tube. The venturi tube has a constricted section that generates a pressure difference. This pressure difference is used to calculate the flow. At high water levels, over 0.7-0.8 meter, measurements are not correct regardless if the flow is high or low, since the tube gets jammed up. An other common problem with venturi tubes is that they get clogged slowly over time and therefore need to be continuously maintained (cleaned) to avoid erroneous measurements.

The flow in the aeration basins (Q_A) and the sand filter basins (Q_F) are measured using a straight constricted weir. These are constructed as an obstruction across an open channel to measure the flow rate. The water flows over the top of the weir, then falls down to a lower level. Since all of the water flows over the weir, and the geometry (rectangular) of the weir is known, the water depth behind the weir can be used to calculate the flow rate [OOF, 2017].

Like the venturi tubes, these weirs need to be well maintained to measure the flow accurately.

The current method used to verify the accuracy of measurements at Bromma WWTP, is a comparison between outgoing flow and water level in sand filter 1-24, located in the Nockeby facility. About once a day, the filter basins are emptied and backwashed (not simultaneously). When the inlet hatches closes and the water level decreases, the difference in water level during 2 minutes is compared to the water flow leaving the basin during the same period.

To enable comparison between water level and flow, the difference in water level is calculated to a flow using the basin area. The relation between measured flow (Q_F) and water level (L_F) can be expressed as

$$Q_F = \frac{\Delta L_F}{\Delta t} \cdot A \tag{1}$$

$$Q_{F_i} : measured\ flow\ [m^3/s]$$
$$L_{F_i} : measured\ water\ level\ [m]$$
$$A : basin\ area\ [m^2]$$

Ideally, the measured flow equals difference in water level multiplied with the basin area and Eq. 2 is valid. However, due to errors, they sometimes differ and an error term e is needed to describe the relation:

$$Q_F = \frac{\Delta L_F}{\Delta t} \cdot A + e \tag{2}$$

Fig. 2.2 is an illustration of how the method works. The comparison can only be made during depletion of water (at point of comparison plus a few minutes).
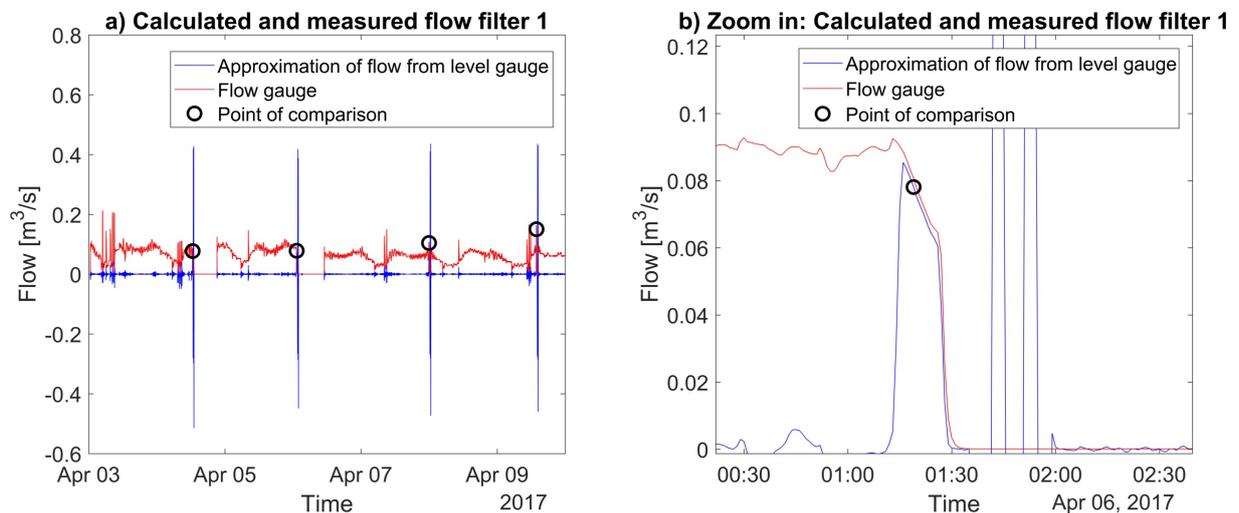

Figure 2.2. The flow verification method used by Stockholm Vatten och Avfall; an example of how if works. a) When the inlet hatches closes and the water level decreases, an approximation of flow calculated from the level gauge is compared to the measured flow (point of comparison), b) A zoom in of such a period. The difference in water level can only be recalculated as a flow during depletion of water (when the basins are emptied), thus, the blue curve is only an actual flow at the point of comparison plus a few minutes.

4

2.2 Errors

The total error in a measurement can be described by the sum of random and gross errors. Random errors can appear due to process or measurement noise. An average of several measurements containing random errors will generally be closer to the true measured value [Bell, 2001]. Gross errors are caused by non-random events such as malfunctioning instruments, fouling of sensors or faulty calibration [Narasimhan and Jordache, 2000]. The difference between these two types of errors is illustrated in Figure 2.3. Random errors increase the variation, but the average is not affected, whereas the average of a measurement containing gross errors is changed depending on the magnitude of the errors.
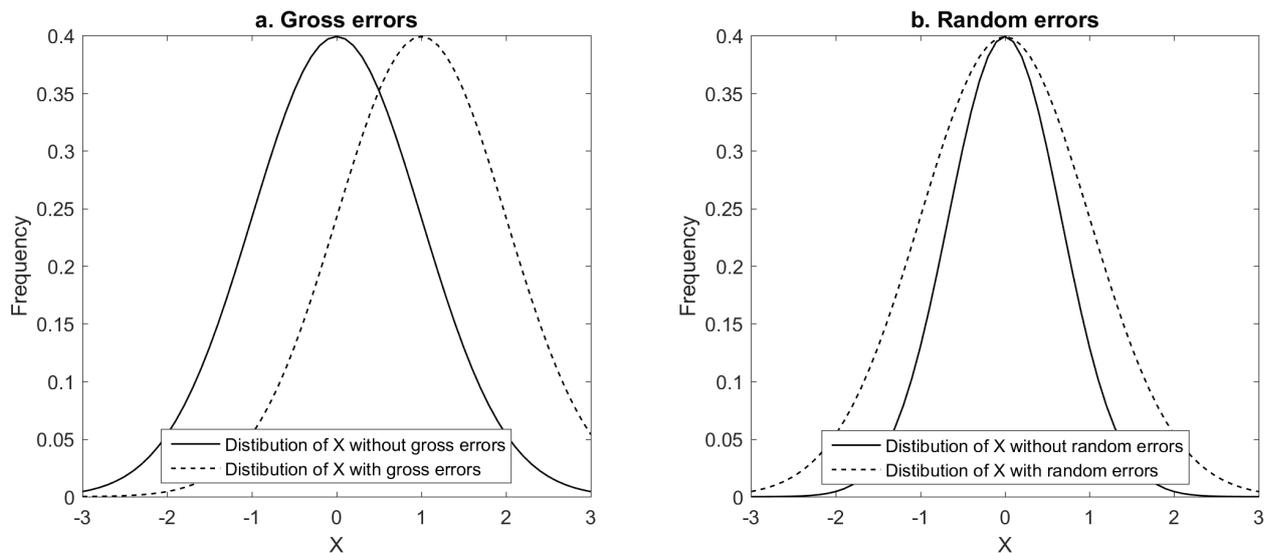

Figure 2.3. A simple example of how a measurement containing *a)* gross errors and *b)* random errors can deviate from the true measurement.

The relation between measured value (y), random error (ϵ), gross error (δ) and true value (x) can be expressed as

$$y = x + \epsilon + \delta \tag{3}$$

[Narasimhan and Jordache, 2000].

2.3 Error detection based on flow balancing

Knowing how different flows in a WWTP relate to each other (checking flow balances) enables detection, estimation and diagnosis of both random and gross errors. If the relationship between the flows is incorrect according to the balance equations, data is corrupted by an error. The basic principle in detecting gross errors is based on the detection of bias in statistical applications. Gross errors, or equivalently *significant errors*, are large relative the variable's variance [van der Heijden et al., 1994b]. Hypothesis testing is the most commonly used method to detect gross errors, where the null hypothesis H_0 is that there are no gross errors present in the data, and the alternative hypothesis H_A is that there are one or several gross errors present. The null hypothesis is accepted or rejected by a comparison with a test criterion [Narasimhan and

Jordache, 2000].

Data reconciliation (DR) is a method to reduce the effect of random errors in a dataset. DR uses process model constraints, adjusting measurements so that the constraints, i.e. mass balances, are satisfied [Narasimhan and Jordache, 2000]. A data set that has been reconciled contains less errors and satisfies the mass balances exactly [van der Heijden et al., 1994a].

Conveniently, gross error detection and data reconciliation require the same available information when processing the data [Narasimhan and Jordache, 2000]. A prerequisite for balancing- and gross error detection techniques to be applicable is that the measurements need to be redundant. This means that a measured flow also can be calculated from other measured flows [van der Heijden et al., 1994a].

2.3.1 Gross error detection

A method for detecting gross errors should be able to detect the presence of one or several errors in the data set, have the ability to identify type and location of the error and have the ability to estimate the magnitude of the errors.

For the case when the measurements satisfy the balance equations exactly, the linear constraint model is assumed to be given by

$$Ay = 0 \tag{4}$$

$$A : N \times m \text{ Process matrix}$$

$$y : m \times 1 \text{ vector of flow measurements}$$

[van der Heijden et al., 1994b].

The process matrix is a linear constraint vector that contain the elemental composition of the balances. It specifies how one flow relate to another. Every row represents a node, and the columns correspond to the flows, thus A can be called a $N \times m$ matrix where N is number of nodes and m number of flows. The elements in A are either negative, positive or 0 depending on whether the flow is an input, an output or if it is not associated with the balance equation [Narasimhan and Jordache, 2000]. Figure 2.4 shows an example of a process matrix constructed from a simple system with three flows denoted F1-F3 and two nodes.

Figure 2.4. Example of a system with three flows (F1-F3) and two nodes (the small circles).

From Figure 2.4 following balances are obtained

$$Node\ 1: \quad F1 - F2 = 0$$

$$Node\ 2: \quad F2 - F3 = 0$$

6

And the process matrix A becomes

$$A = \begin{array}{c} \begin{array}{ccc} F1 & F2 & F3 \end{array} \\ \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{array}{l} Node1 \\ Node2 \end{array} \end{array}$$

Due to measurement errors, Eq. 4 rarely add up to zero. For all variables measured, the $N \times 1$ vector of balance residuals r is given by

$$Ay = r \tag{5}$$

If there are no gross errors present, r has a normal distribution with a zero mean value and a covariance matrix Σ:

$$\Sigma = cov(r) = AVA^T \tag{6}$$

$\Sigma : m \times m$ residual covariance matrix

$V : m \times m$ measurement covariance matrix

Under $H_0, r \sim N(0, \Sigma)$, the elements in the residual vector r reflect if the process constraints (flow balances) are violated. Also, the covariance matrix Σ contains information of both the process matrix (A) and the measurement covariance matrix (V). Therefore, both r and Σ are useful in constructing statistical tests to detect the existence of gross errors [Narasimhan and Jordache, 2000].

2.3.2 Data reconciliation

DR is a technique to reduce the effect of random errors in data and improve the accuracy of measurements. A set of balance equations is needed to verify the consistency and to improve accuracy of measured data. Constraints can then be incorporated in that set of equations. After such relations between measured values are made, it is possible to adjust the measured values into estimate values consistent with the constraints.

To improve the accuracy of data using DR, certain assumptions must hold. The measured values must not contain gross errors, or large random errors. If this is not the case, DR can lead to incorrect adjustments of the measured values. Thus, it is crucial to check the data for gross errors prior to obtaining final estimates via DR. Also, no relevant components (e.g. flows) in the set of linear constraint equations can be omitted [van der Heijden et al., 1994b].

The general data reconciliation problem can be stated as

$$\min_{y-\hat{y}} (y - \hat{y})^T V^{-1} (y - \hat{y}) \tag{7}$$

$$\text{s.t } A\hat{y} = 0$$

7

$y : m \times 1$ vector of raw measurements

$\hat{y} : m \times 1$ vector of estimates

$V : m \times m$ measurement covariance matrix

$A : N \times m$ process matrix

2.4   Related research

Statistic techniques for detecting gross errors and to reconcile data from biochemical processes was developed by van der Heijden et al., [1994a]. The use of DR and methods for gross error detection in WWTP is an active research area, but has not received much attention so far [Le et al., 2016]. Only a limited number of studies related to WWTP are available at this time. Those found to be relevant to this study, ([van der Heijden et al., 1994a,b,c; Meijer et al., 2002; Puig et al., 2008; Seungchul et al., 2015; Le et al., 2016]) will be discussed below.

van der Heijden's method, described in van der Heijden et al., [1994a,b,c] is based on the fact that all systems can be described by a set of linear equations. It was implemented in "Macrobal", a free domain software originally developed to balance flows and compounds from the fermentation industry. The method can though easily be implemented in e.g. MATLAB[1] since it is based on matrix algebra and statistical analysis [van der Heijden et al., 1994a].

Meijer et al., [2002] was the first to apply Heijden's method on data from a WWTP [Brdjanovic et al., 2015]. They applied the method on annual average measurements from a full-scale WWTP, which revealed major errors in the process flows and data could be improved. This method has been tested on several WWTPs in the Netherlands, with successful results [Meijer et al., 2002]. Puig et al., [2008] continued in the same line by proposing a practical methodology to detect errors in historical data of a full-scale WWTP. They obtained useful and new information for evaluation, design and benchmarking purposes, and concluded that faulty historical data result in large errors when key operational conditions are calculated [Puig et al., 2008].

van der Heijden's method consists of four steps, and the step for detecting gross errors in turn consists of four additional steps (a-b):

1. Selection of measured and non measured components (e.g. flow)

2. Classification of components into four categories; balanceable, non-balanceable and calculable, non-calculable

3. Calculation of the optimal estimates for the components

4. Gross error detection, diagnosis and estimation

    (a) Detection: Testing if there are one or more significant error present

    (b) Classification of errors (1. One or several of the measurements has a significant error, 2. The definition of the system is incorrect or 3. The test is too sensitive due to small variances)

---

[1]© 2017 The MathWorks, Inc. MATLAB and Simulink

(c) Locating the error source

(d) Estimation of error magnitude

In Macrobal, gross errors are detected by evaluating the residuals for each balance equation using a statistical test. In an ideal case, the mass balances add up to zero, but due to errors the balances have residuals. This vector of residuals (r_n) is constructed for each set of measured values. For n number of massbalances and $j = m + u$ where m is measured flows and u unmeasured (calculated) flows (e.g. water flow, P, N etc.), the residuals are calculated as

$$r_n = \sum_{i=1}^{j} Q_i X_{i,n} \qquad (8)$$

Q_i : Vector of flows (measured or calculated)

$X_{i,n}$: Elements in each mass balance (elements in the process matrix)

n : Number of mass balances

When $n = u$, (thus when the number of mass balances equals the number of calculated flows), the system can be solved, and if $n > u$ the system is called over determined, which can be expressed as the degree of redundancy.

$$n - u = degree\ of\ redundancy \qquad (9)$$

If the degree of redundancy is equal to or larger than 1, the system can be balanced [Meijer et al., 2002].

The residual vector r_n is compared to certain *compare vectors* which corresponds to a specific source of an error. For a measurement k, the compare vector c_k is the corresponding column in the reduced process matrix A', which is obtained by removing all linearly dependent rows in A, so that A becomes linearly independent.

$$c = A' = RA \qquad (10)$$

c : Compare matrix

A' : Reduced process matrix

R : Reducing matrix

The residual vector and the compare vector are compared in a statistical test. Gross errors present in the flow vector can be detected by systematically redefining the measured flow to an unmeasured flow until the statistical test is passed. Gross errors caused by incorrectly defined mass balances of faulty measurements can be detected by systematically removing balances until the statistical test is passed [van der Heijden et al., 1994b].

Another approach to find gross errors after the detection step, is the serial elimination strategy. This is performed by removing one measured component at a time until the statistical test is

9

passed, used by e.g. Xiaolong et al., [2014]. Their aim was to identify gross errors in power plants via DR [Xiaolong et al., 2014]. This strategy is often used in combination with balancing methods, but has according to van der Heijden [1994b] a few shortcomings compared to their method. The serial elimination strategy provide more unambiguous results, but consideration is not taken to all possible events that can result in inconsistency; identification of errors in the system descriptions is not explicit. Another advantage with van der Heijden's method is that the magnitude of the error can be estimated.

Seungchul et al., [2015] uses a similar approach as previously mentioned authors [van der Heijden et al., 1994a,b,c; Meijer et al., 2002; Puig et al., 2008] but proposes a new DR scheme using a closed loop mass balance and the Lagrange multiplier method. Influent data were generated using a Monte Carlo simulation, thus, the study was based on a simulated WWTP model. Three case studies were performed. In the first case, no gross errors were considered, in the second case, a gross error was set in the process data and in the third case, the first two cases were compared. Seungchul's method consists of two steps:

1. Mass balancing and the Lagrange multiplier method

2. Data verification and data reconciliation

To solve the constraint optimization problem 7, Seungchul propose the use of a Lagrange multiplier (λ). The Lagrange multiplier method is a straight forward strategy for finding the local maxima and minima of a function, subject to certain constraints.

Le et al,. [2016] demonstrates an application of a bilinear steady-state DR and gross error detection, in contrast to the work of van der Heijden et al., [1994a,b,c] where the mass balances were expressed in linear terms. The aim was to evaluate consistency of measured data and to estimate unknown parameters. Implementation was made in MATLAB, and the algorithm was tested on measured data from a full-scale partial nitration reactor (SHARON), which is used for treating wastewater with high levels of ammonium. This approach allowed to reducing the number of unknown variables, and increase the number of variables that could be balanced and estimated [Le et al., 2016].

3   Method

3.1   Choice of method

The system examined in this thesis was quite simple relative to the systems that were balanced in Meijer et al., [2002] and Puig et al., [2008]. It was unlikely that errors would be present due to an inaccurate system description, thus, the major drawback with the serial elimination method was indifferent. However, the magnitude of the errors could not be estimated after the serial elimination method, which was desirable if the erroneous data was to be corrected.

The approach by Le et al., [2016] would be interesting for a process with many unmeasured flows, which was not the case in this study. Therefore, the basis of van der Heijden's method was used. This method had generated successful results and had been well cited in published articles on the subject.

A few modifications were made. The initial steps were somewhat simplified since all variables were measured, and the Lagrange multiplier method used in [Seungchul et al., 2015] was used since it was a convenient approach to solve an constrained optimization problem. Initially the method was evaluated using simulated data where known errors was added.

The following methodology was chosen for error detection in this thesis.

   Step 1:   Formulation of flow balances and the process matrix

   Step 2:   Gross error detection, diagnosis and estimation

   Step 2a: Detection; testing if there are one or more significant gross error present

   Step 2b: Locating the error source (diagnosis)

   Step 2c: Estimation of error magnitude

   Step 2d: Correcting data for gross errors

   Step 3:   Calculation of the optimal estimates for the flows via data reconciliation using the Lagrange multiplier method

The suggested method was tested in three case studies. Before describing the method in detail, the conditions in each case were described.

3.2   The case studies

In all three cases, series of averages were created for subsets of the full datasets (not a moving average). The range was set to 60 data points, e.g. if the entire dataset contained 3000 data points, it was divided into 50 subsets for which an average was calculated, resulting in 50 data points. All test parameters were calculated for each average series, consequently 60 time steps at the time. For Case 2 and 3 where real process data was analyzed, this corresponded to 1 hour since the data resolution was in minutes. For Case 1 the time was given in time steps.

3.2.1 Case 1

In Case 1, the suggested method was evaluated on three different simulated WWTP datasets with known added errors.

Initially, four static signals (amplitude = 1) were created. To resemble a dataset only containing random errors, normally distributed random noise (mean value = 1, variance = 0.0065-0.0067) were added (simulated dataset 1). To investigate how the method would perform with systematic errors present, a bias (slope = 0.1) was added on one of the four signals (Simulated dataset 2). In Simulated dataset 3, the same bias was added, but the variance of the random noise was increased (mean value = 1, variance = 0.013-0.015).

3.2.2 Case 2

In Case 2, the method was evaluated on real WWTP flow data from Bromma WWTP. Stockholm Vatten och Avfall provided flow data for the influent, aeration basins and sand filter basins. Data for the effluent flow was provided by Norrenergi and limited to four datasets over totally 22 days during 2014 and 2015. To investigate all four flows, these periods were used.

In Fig. 3.1 a), one of the available datasets were shown (all datasets are given in Appendix A.). It could be seen that there was a time shift in Q_{Eff}. Also, Q_A and Q_F were considerably lower than Q_{In} and Q_{Eff} when the flow quickly increased and leveled out after October 8th. This likely occurred because the flow suddenly increased fast and there was a flow restriction in these two basins; the excess water was side-stepped. The time shift on Q_{Eff} probably occurred because the water travels a distance before it was measured. In Fig. 3.1 b), the time shift in Q_{Eff} was corrected by displacing the flow data vector two hours back in time. Thereof there were missing data for Q_{Eff} the last two hours in this graph. Also, the bypass flow was added to Q_A and Q_F.
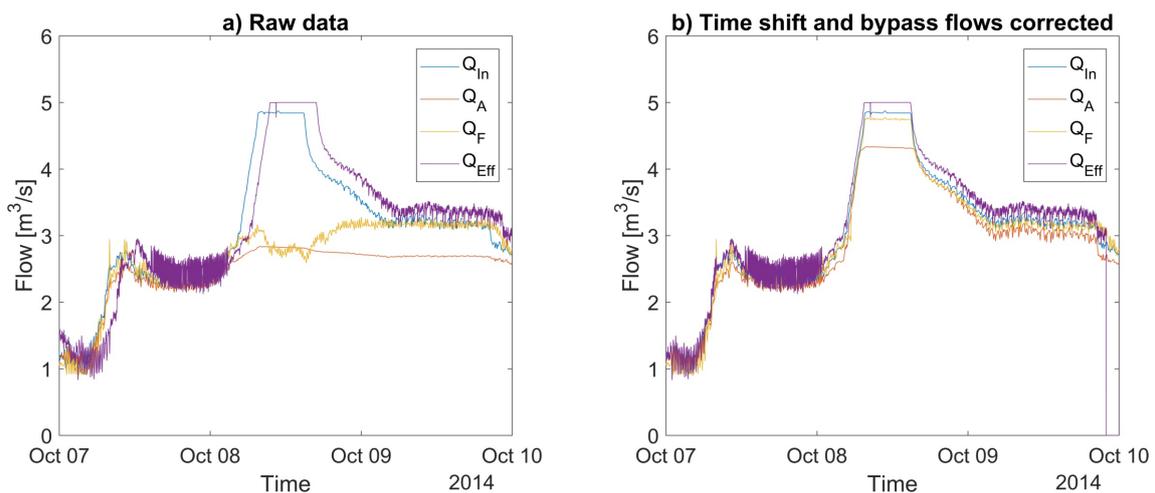

Figure 3.1. One of the available datasets. a) The raw data set, b) Q_{Eff} has been time-shifted 2 hours back in time. For Q_A and Q_F there was a limit to how much water that could enter the basins. When the flow increased, some of the water bypassed. This bypass flow was added.

Between October 8th to 9th in Fig. 3.1, the flow gauges leveled out, which probably occurred due to high water levels or high flows; the measuring equipment could only measure up to a certain level. These periods were not considered in this study since the suggested method would misinterpret the difference between the flows as gross errors.

The suggested method was evaluated on two shorter periods (one or two days at the time) on both uncorrected data and data corrected for time shift and bypass flow (see Fig. 3.1). The first period was during October 7th and the second during September 8-10th.

3.2.3 Case 3

In case 3, the method was compared to the flow verification method used by Stockholm Vatten och Avfall. Step 1 and 2(a-c) was performed on flow data from April 3-10th 2017 and May 15-21th 2017. A description of the flow verification method is given in Background and Theory. The initial plan was to perform the comparison between the two methods on the same periods as in Case 2. However, due to absence of data for the inlet hatches during 2014 and 2015 (which was necessary to enable implementation of the flow verification method), the comparison to the flow verification method was performed on more recent data from 2017. Data for Q_{Eff} was not available for these periods, thus, there were only three flows to include in the suggested method.

The flow verification method was used to calculate the difference between measured and calculated flow in the 24 sand filters (Q_{F_i}, $i = 1, ..., 24$) during these periods. Differences between measured and calculated flow in the filters could thereby be compared the results from the suggested method during the same periods. If a difference between measured and calculated flow could be connected to a significant error found in Q_F using the suggested method, this would imply that the measured flow was incorrect. If differences between measured and calculated flow could not be connected to the results from the suggested method, this would imply either that the calculated flow (thus level measurement) was incorrect or that the two methods were unsuitable to compare.

3.3 Step 1: Formulation of flow balances and the process matrix

As a first step, flow balances and the process matrix A were constructed.

3.3.1 Case 1

For all three tests in Case 1, the system was described as follows:

$$\text{Node } 1 \; : Q_1 - Q_2 = 0 \tag{11}$$

$$\text{Node } 2 \; : Q_2 - Q_3 = 0 \tag{12}$$

$$\text{Node } 3 \; : Q_3 - Q_4 = 0 \tag{13}$$

with the raw measurement vector

$$y = [Q_1, Q_2, Q_3, Q_4] \tag{14}$$

13

and the process matrix

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{15}$$

Since the rows in A were linearly independent, A did not have to be reduced according to Eq. 10 to obtain the compare vectors c_i, hence $A_i = c_i$ in this case.

3.3.2 Case 2

From the system in Figure 2.1 following flow balances were constructed over each node:

$$\text{Node 1} : Q_{in} - \sum_{i=1}^{6} Q_{A(i)} = 0 \tag{16}$$

$$\text{Node 2} : \sum_{i=1}^{6} Q_{A(i)} - \sum_{i=1}^{24} Q_{F(i)} = 0 \tag{17}$$

$$\text{Node 3} : \sum_{i=1}^{24} Q_{F(i)} - Q_{eff} = 0 \tag{18}$$

with the raw measurement vector

$$y = [Q_{in}, \sum_{i=1}^{6} Q_{A(i)}, \sum_{i=1}^{24} Q_{F(i)}, Q_{eff}] \tag{19}$$

and the process matrix

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{20}$$

The same conditions as in case 1 applied in this case: A did not have to be reduced to obtain the compare matrix c.

3.3.3 Case 3

Since there were three flows and two nodes in this case (Q_{Eff} was not included), following flow balances describe the system:

$$\text{Node 1} : Q_{in} - \sum_{i=1}^{6} Q_{A(i)} = 0 \tag{21}$$

$$\text{Node 2} : \sum_{i=1}^{6} Q_{A(i)} - \sum_{i=1}^{24} Q_{F(i)} = 0 \tag{22}$$

with the raw measurement vector

$$y = [Q_{in}, \sum_{i=1}^{6} Q_{A(i)}, \sum_{i=1}^{24} Q_{F(i)}] \tag{23}$$

the process matrix became

$$A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \tag{24}$$

Since the rows in A were linearly independent, A did not have to be reduced to obtain the compare matrix c.

3.4   Step 2: Gross error detection, diagnosis and estimation

As a final step, found gross errors were removed from the initial measurements, resulting in an improved set of data.

3.4.1   Step 2a: Detection

The step for detecting gross errors were based on hypothesis testing, where:

$$H_0 : No \; gross \; errors \; present$$
$$H_A : One \; or \; several \; gross \; errors \; present$$

H_0 was accepted or rejected by comparing a test statistic γ with a threshold value [Narasimhan and Jordache, 2000].

$$\gamma = r^T \Sigma^{-1} r \tag{25}$$

γ : Test statistic

r : $1 \times N$ vector of balance residuals

Σ : Covariance matrix of the balance residuals (see Eq. 6)

It could be proven that the test statistic γ follows a χ^2-distribution with υ degrees of freedom equal to the rank of Σ. For the chosen level of significance α, the test criterion became $\chi^2_{1-\alpha,\upsilon}$. If $\gamma \geq \chi^2$, an error was significant [van der Heijden et al., 1994b].

In all performed gross error detection tests, the test criterion was set to $\chi^2_{1-0,005,DF}$, which means that the level of significance was 0,005. The χ^2-value (the test criterion) was read from a χ^2-distribution table and differed depending on the numbers of degrees of freedom, which was equal to the rank of Σ.

15

3.4.2 Step 2b: Locating the error source

If a significant gross error was detected in Step 2a, its source could be found by performing an error diagnosis.

Given the residual vector r with a covariance matrix Σ, it must be determined if the compare vector c or a multiple of c was a sample of the residual vector r. In van der Heijden et al., [1994b] this was denoted " if r and c point in the same direction". As explained in Background and theory (see Eq. 10), the compare matrix c was the reduced process matrix A. Each column in c were compared to the corresponding measurement.

To examine how well c_i compared with r, the residual fit Ψ of the vector c was calculated, which was the following probability

$$\Psi = p(\Delta^2 \leq \chi^2) \tag{26}$$

Δ^2 : Test statistic

As any probability, Ψ lied within 0 and 1. A small residual fit implied that c did not resemble r, hence, the corresponding source of error was not a suspect and vice-versa with a high value of Ψ. It was not a definitive measure but an indication to which flow that differed the most from the others.

For a measurement l, the test statistic Δ_l^2 was calculated as

$$\Delta_l^2 = r^T \Sigma^{-1} r - \frac{(r^T \Sigma^{-1} c_l)^2}{c_l^T \Sigma^{-1} c_l} \tag{27}$$

Δ_i^2 : Test statistic for measurement i

r : $1 \times N$ vector of balance residuals

Σ : Covariance matrix of the balance residuals (see Eq. 6)

c_l : $1 \times N$ Compare vector for measurement l

Δ^2 had a χ^2-distribution with a number of degrees of freedom equal to the rank of Σ minus 1 [van der Heijden et al., 1994b].

3.4.3 Step 2c: Estimation of magnitude

When a source of a significant error had been located, its magnitude s could be estimated by [van der Heijden et al., 1994a]

$$\hat{s}_l = \frac{c_l^T \Sigma^{-1} r}{c_l^T \Sigma^{-1} c_l} \tag{28}$$

$\hat{s}_i$: Estimate of the error magnitude for measurement l

$c_l : 1 \times N$ Compare vector for measurement l

Σ : Covariance matrix of the balance residuals (see Eq. 6)

$r : 1 \times N$ Vector of balance residuals

3.4.4 Step 2d: Correcting data for gross errors

If an error had been detected, located and its magnitude was estimated, the corresponding measurement should be corrected. This could be performed by simply subtracting the estimated error ($\hat{s}$) from the erroneous flow.

3.5 Step 3: Calculation of optimal estimates

In this step, the optimal estimates were calculated via DR using the Lagrange multiplier method. Eq. 7 was minimized using a Lagrange multiplier λ. The estimates ($\hat{y}$) were calculated from

$$\lambda = I - V A^T (A V A^T)^{-1} A \tag{29}$$

$$\hat{y} = \lambda y \tag{30}$$

λ : Lagrange multiplier

$I : m \times m$ identity matrix

$V : m \times m$ covariance matrix

$y : m \times 1$ vector of raw measurements

This step should only be performed if the gross error detection steps showed that there were no gross errors present, or if detected errors were successfully removed[Plasma Processing Laboratory, 2013].

4 Results

4.1 Case 1: Simulated data

The rank of Σ was 3, which gave a test criterion $\chi^2 = 12.84$.

4.1.1 Gross error detection

Simulated dataset 1
The dataset (Fig. 4.1 a) consisted of four generated signals with added random noise, which resembled a dataset containing only random errors. The noise had a standard deviation of 0.08 and was normally distributed with a variance between 0.0065 - 0.0067, see Appendix B.

The test statistic (Fig. 4.1 c) was not exceeded at any time, thus there were no gross errors or large random errors found.
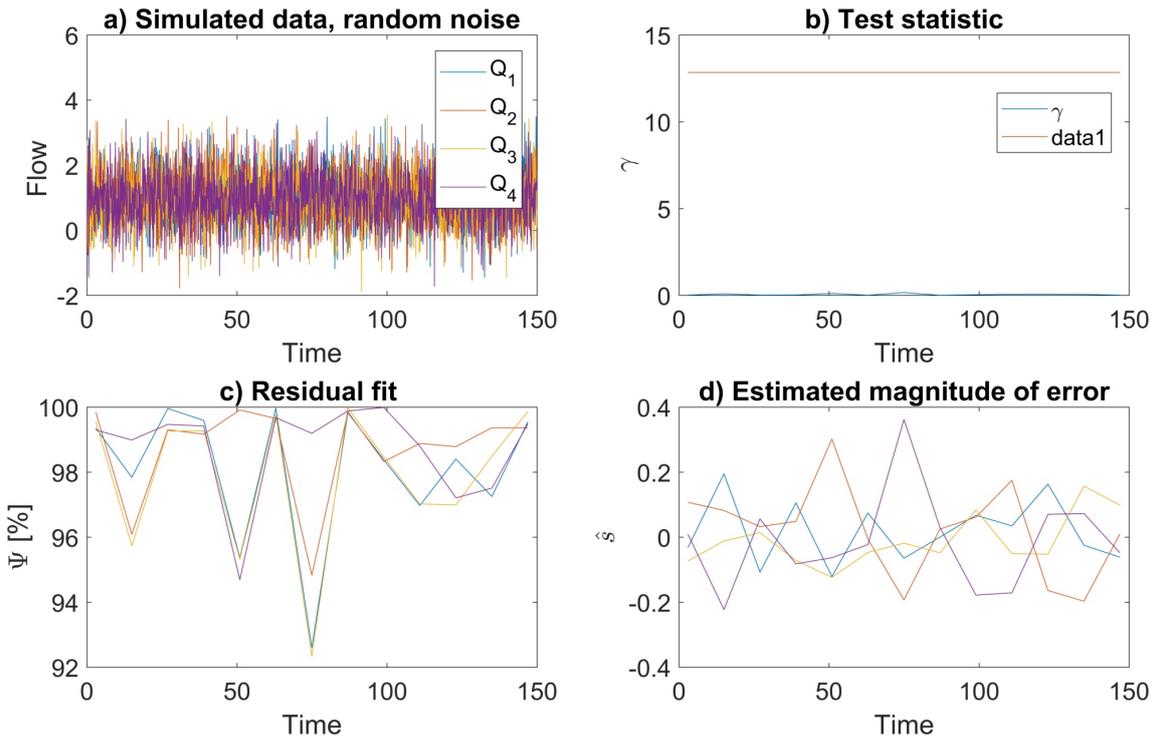
Figure 4.1. Gross error detection performed on simulated dataset 1 where there were no gross errors present, only random noise. a) The simulated dataset, b) Step 2a, testing if there were any significant errors present, c) Step 2b, locating the error source, d) Step 2c, estimating the error magnitude.

Simulated dataset 2

A gradually increasing bias (slope 0.1) was added Q_2 (Fig. 4.2 a), which means that the dataset contained both random- and gross errors. The noise had the same properties as in Simulated dataset 1. As could be seen in Figure 4.2 b) a bias was detected around time step 30. To draw conclusions about what signal that was biased, the residual fit (Fig. 4.2 c) was examined. The residual fit was the probability that the error was in that specific flow. It was about equal for all flows at t=0. As the bias increased, the probability that the error was in Q_1, Q_3 or Q_4 decreased and the residual fit for Q_2 remained $> 90\%$. This indicated that the bias lied in Q_2. The estimated error magnitude (Fig. 4.2 c) for Q_2 seemed to correlate well with the added bias. At t=100 the magnitude was approximately 1, which looked likely when comparing to the dataset.
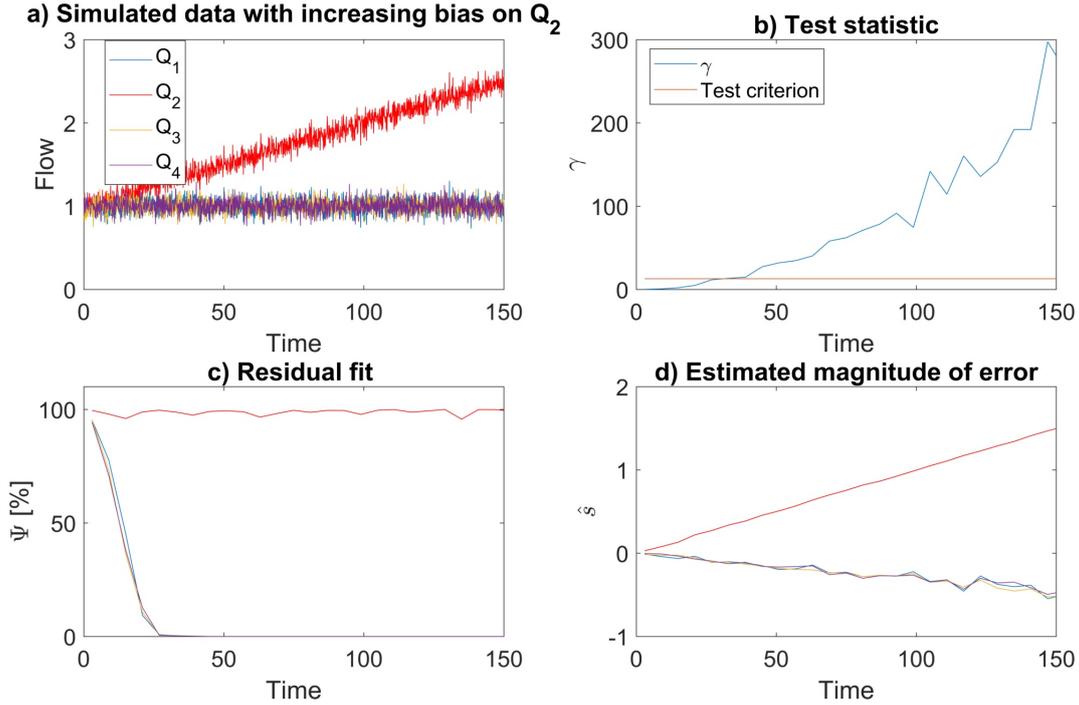
18

Figure 4.2. Gross error detection performed on simulated dataset 2 where there were an increasing bias added to Q_2 and random noise in each signal. a) The simulated dataset, b) Step 2a, testing if there were any significant errors present, c) Step 2b, locating the error source, d) Step 2c, estimating the error magnitude.

To investigate how well the method detected this type of bias, and to remove the bias that was found in the signal Q_2, a linear regression model was created for the estimated magnitude of the error found in Q_2. A linear model was created instead of simply subtracting the errors since the added bias was linear. This enabled comparison between the added slope to the estimated one. If the estimated errors were assumed to be normally distributed, a linear regression model describing $\hat{s}$ in Fig. 4.2 d) was given by

$$\hat{\hat{s}} = 0.00999t + 0.0004 \tag{31}$$

This model had a r^2-value of 0.899, which indicated that the model described $\hat{s}$ well. Also, the estimated slope was very similar to the added one, which indicated that the estimation of the error was good. In Fig.4.3 a), the bias was corrected by subtracting the coefficient of slope, $0.00999t$, from the raw data from time step 30. When re-doing the error detection test, no significant errors were found.
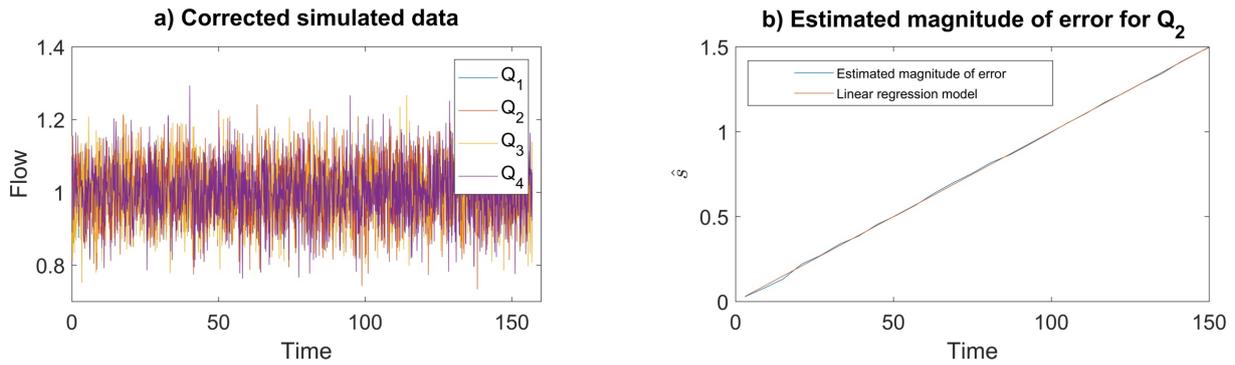
Figure 4.3. A linear regression model was created for the estimated magnitude of the error found in Q_2. a) The simulated data in Figure 4.2 had been corrected by subtracting the estimated coefficient of slope, b) The estimated error on Q_2 and a linear model to describe it.

Simulated dataset 3

In Figure 4.4 the same bias was added to Q_2, but the noise standard deviation on all signals was increased to 0.12. The noise was normally distributed with a variance between 0.013 - 0.015, see Appendix B. Here, the bias was not detected until around time step 88, see Fig. 4.4 b). This was expected since the variance was larger.
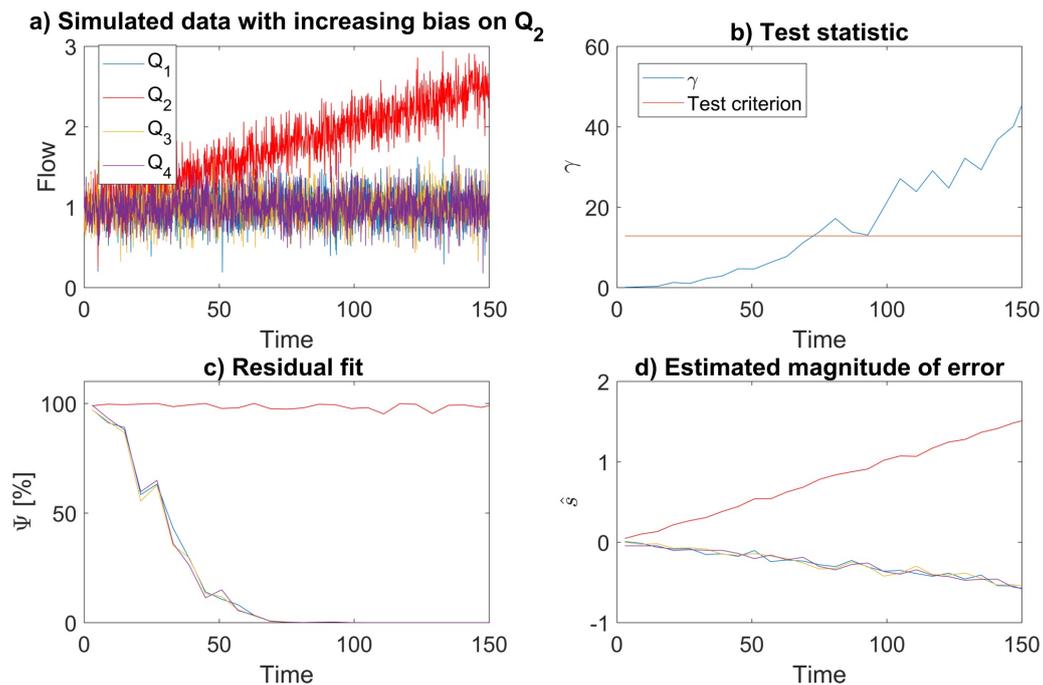
Figure 4.4. Gross error detection performed on simulated dataset 3 where there was an increasing bias added to Q_2 and random noise (increased compared to simulated dataset 2) in each signal. A) The simulated dataset, b) Step 2a, testing if there were any significant errors present, c) Step 2b, locating the error source, d) Step 2c, estimating the error magnitude.

A linear regression model for the estimated magnitude of the error was created for the data set in Figure 4.4 as well.

$$\hat{\hat{s}} = 0.01t - 0.0089 \tag{32}$$

This model had a r^2-value of 0.901, which indicated a good fit to the estimated error. The estimated slope equaled the one added to Q_2, which indicated that the estimation of the error was good. In Figure 4.5 a), the bias was corrected by subtracting the coefficient of slope $0.01t$ from the raw data from time step 88. When re-doing the error detection test, no significant errors were found.
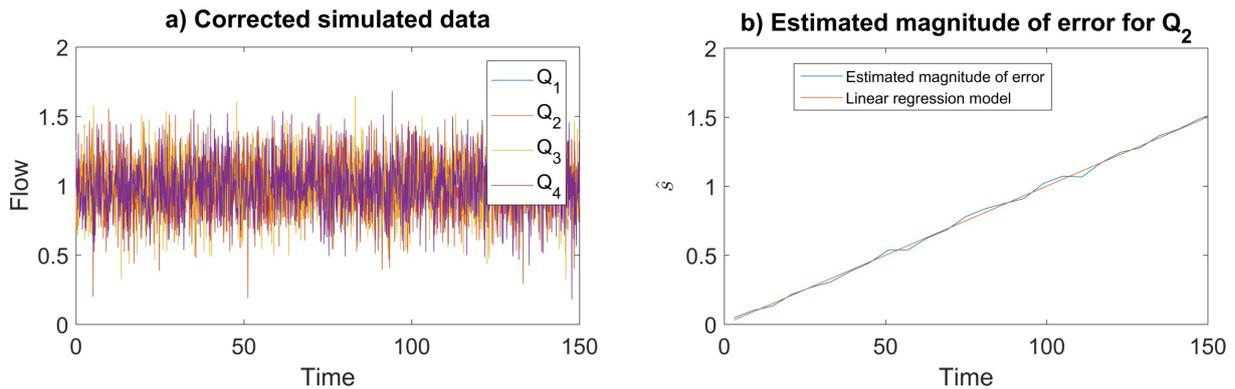

Figure 4.5. A linear regression model was created for the estimated magnitude of the error found in Q_2. a) The simulated data in Figure 4.4 has been corrected by subtracting the estimated coefficient of slope, b) The estimated error in Q_2 and a linear model to describe it.

4.1.2 Data reconciliation

Since the gross errors were detected and removed, the DR step could be performed to obtain the final estimates. The DR was performed per 60 time steps. In Figure 4.6, simulated dataset 2 and 3 had been reconciled. The flows $(Q_1, ..Q_4)$ in each dataset (simulated dataset 2 and 3) were now identical and free from systematic errors. They satisfied the balance equations exactly.
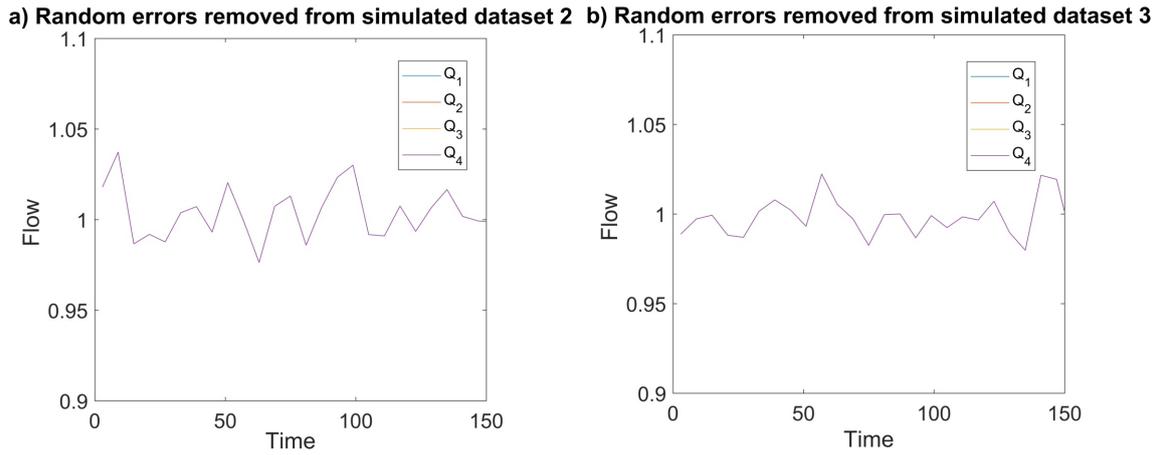
a) Random errors removed from simulated dataset 2   b) Random errors removed from simulated dataset 3

Figure 4.6. DR was performed on simulated dataset 2 (a) and 3 (b). The flows $(Q_1, ..Q_4)$ in each dataset (simulated dataset 2 and 3) were now identical and the balance equations were satisfied exactly.

In Table 4.1, mean values of the raw datasets, the gross error free datasets and the reconciled datasets were shown. As expected, the mean value $\bar{y}$ for Q_2 were higher than for the other flows. After the removal of gross errors, the mean values for the flows in dataset 2 were identical, and for dataset 3 they were almost identical. The standard deviation for Q_2 were lowered in both cases. After the DR, the mean values were identical i both cases, and the standard deviations were decreased.

Table 4.1. Mean values $\pm$ standard deviation of the raw data ($\bar{y}$), the data where gross errors were removed ($\bar{\bar{y}}$) and the reconciled data where random errors were removed ($\bar{\hat{y}}$) for simulated dataset 2 and 3.

Simulated dataset 2	$\bar{y}$	$\bar{\bar{y}}$	$\bar{\hat{y}}$
Q_1	1.00 ± 0.08	1.00 ± 0.08	1.00 ± 0.01
Q_2	1.79 ± 0.46	1.00 ± 0.08	1.00 ± 0.01
Q_3	1.00 ± 0.08	1.00 ± 0.08	1.00 ± 0.01
Q_4	1.00 ± 0.08	1.00 ± 0.08	1.00 ± 0.01
Simulated dataset 3	$\bar{y}$	$\bar{\bar{y}}$	$\hat{y}$
Q_1	1.00 ± 0.19	1.00 ± 0.19	1.00 ± 0.01
Q_2	1.79 ± 0.49	1.01 ± 0.20	1.00 ± 0.01
Q_3	1.00 ± 0.20	1.00 ± 0.20	1.00 ± 0.01
Q_4	1.00 ± 0.19	1.00 ± 0.19	1.00 ± 0.01

4.2   Case 2: Real WWTP data

As in Case 1, the rank of Σ was 3, which gave a test criterion $\chi^2 = 12.84$. The method was tested on both the raw and the corrected data sets, see Figure 3.1, mainly to see if the method was able to detect the time shift.

4.2.1 Gross error detection: Period 1

In Fig. 4.7 a) there was a distinct time shift on Q_{Eff} from 06:00 plus a few hours. This deviation seemed to be detected when viewing the test statistic graph (Fig. 4.7 b). The residual fit (Fig. 4.7 c) implied that the error lied in Q_{Eff} at this time. The estimated magnitude or the error (Fig. 4.7 d) showed that Q_{Eff} was lower than the other flows during this time. The second top in the test statistic graph, at 12:00, was difficult to interpret when looking at the residual fit.
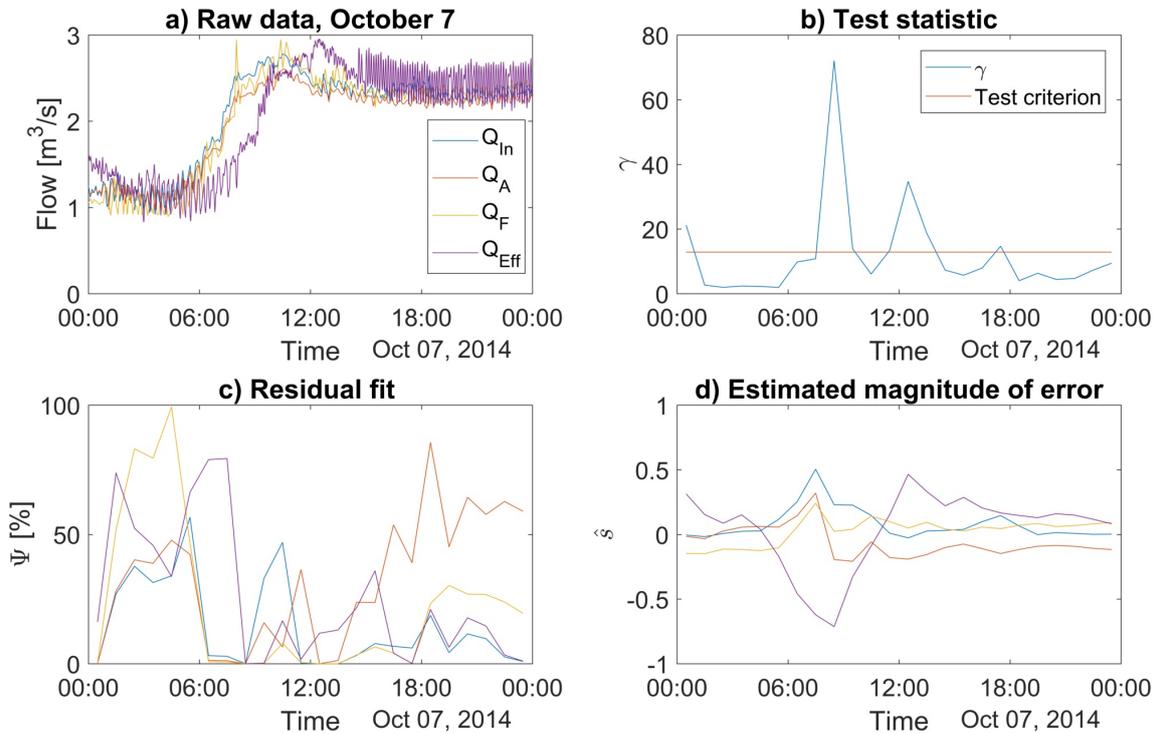
Figure 4.7. Step 1-2abc performed on data from 7 October 2014. a) Raw data, 24 h, b) Calculated test statistic, c) Calculated residual fit, d) The estimated magnitude of the error.

In Fig. 4.8 a), the time shift was corrected (shifted 2 hours back in time). A significant error was still found between 06:00 and 12:00, see Fig. 4.8 b), only now the residual fit graph (Fig. 4.8 c) implied that the error most likely lied in Q_A at this time. Viewing the corrected raw data in Fig. 4.8 a), this looked likely since Q_A lied lower than the other flows at this time.
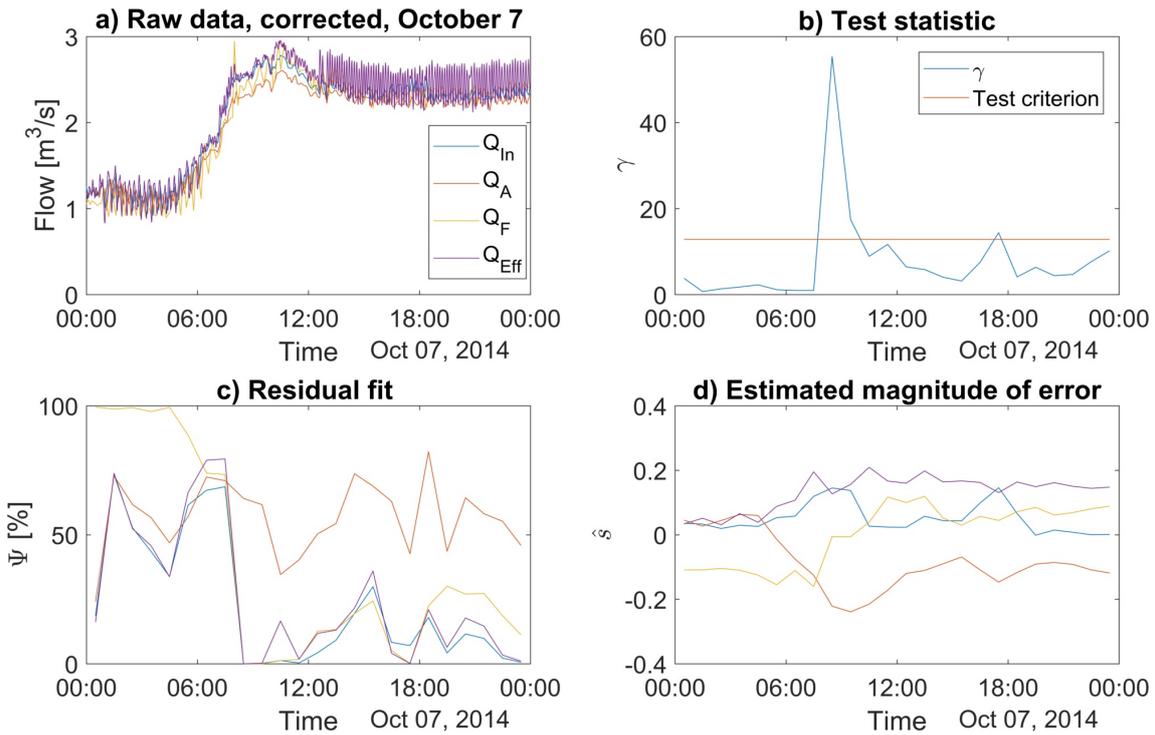
23

Figure 4.8. Step 1-2abc performed on data from 7 October 2014. a) Corrected raw data, 24 h, b) Calculated test statistic, c) Calculated residual fit, d) The estimated magnitude of the error.

4.2.2 Gross error detection: Period 2

In Fig. 4.9 a), neither the time shift or the bypass flows were corrected. It could be seen that Q_A distinctly differed from the other flows during the entire period, but the time shift in Q_{Eff} was not so apparent. The test statistic exceeded the test criterion almost the entire period (see Fig. 4.9 b), and the residual fit implied that the significant error lied in Q_A (see Fig. 4.9 c). The estimated error magnitude decreased with time (see Fig. 4.9 d), as did the difference between Q_A and the other flows in the raw data figure.
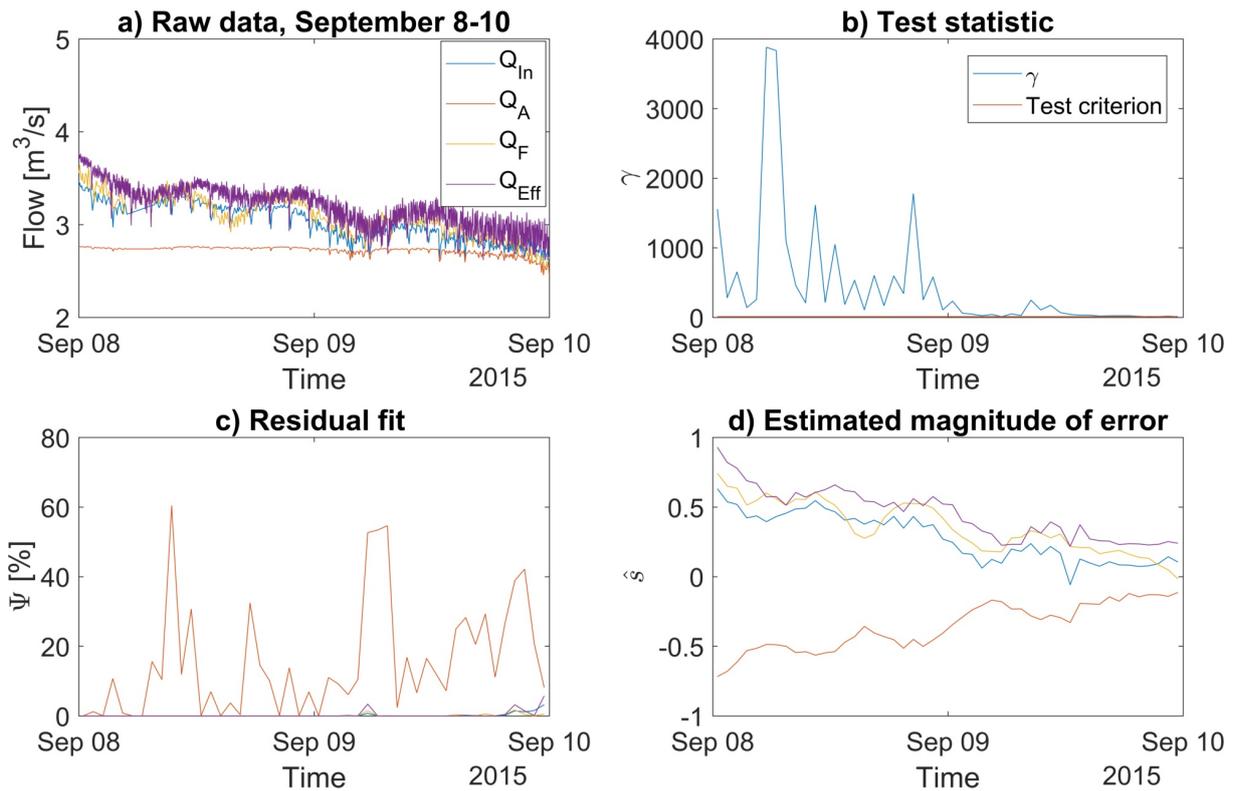
Figure 4.9. Step 1-2abc performed on data from 8-10 September 2015. a) Raw data, 24 h, b) Calculated test statistic, c) Calculated residual fit, d) The estimated magnitude of the error.

Since the test statistic exceeded the test criterion almost the entire period (Fig. 4.9 b), and Q_A dominated in the residual fit graph (Fig. 4.9 c), it was very likely that the error lied in Q_A. Therefore, the estimated error in Q_A was removed to investigate how good the estimation was. In Figure 4.10, the estimated error in Q_A was subtracted from the raw data. The error seemed to be well estimated since Q_A now matched the other flows well. To investigate this further, the time shift in Q_{Eff} also was removed from this dataset and the method was tested once again.
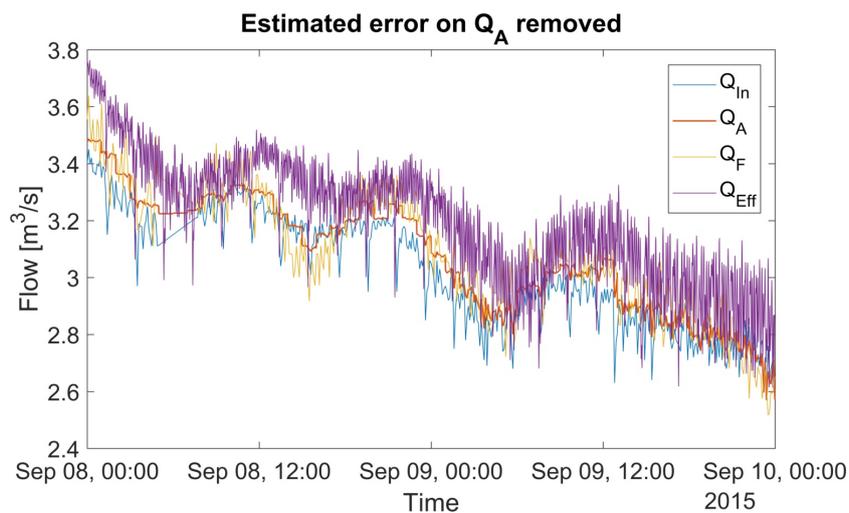

Figure 4.10. The estimated error found in Q_A had been subtracted from the raw data for Q_A.

25

In Fig. 4.11 a), both Q_A and the time shift in Q_{Eff} was corrected. Compared to Figure 4.9, data was less erroneous. The test criterion was exceeded between 04:00 and 8:00 on the 8th (Fig. 4.11 b), but it was difficult to connect this error to any of the flows since the residual fit was zero for all flows 06:00. However, data was improved compared to the original dataset in Figure 4.9.
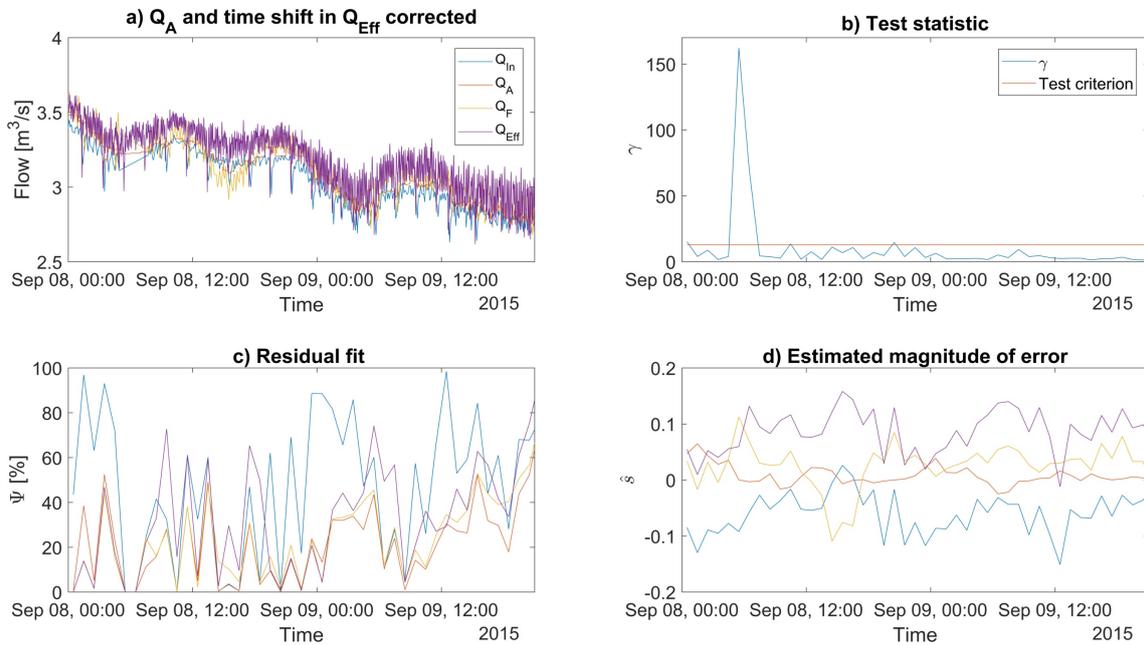

Figure 4.11. Step 1-2abc performed on data from 8-10 September 2015. a) Corrected raw data, 24 h, b) Calculated test statistic, c) Calculated residual fit, d) The estimated magnitude of the error.

4.2.3   Data reconciliation: Period 2

Since the most of the gross errors had been removed from the dataset in Figure 4.11, DR could be performed. In Figure 4.12, DR had been performed. The flows were now identical and free from random errors (random noise).
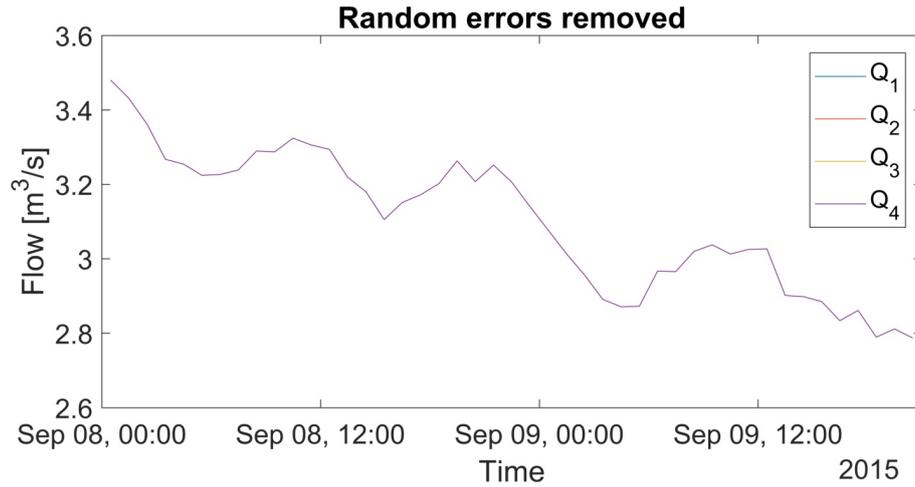
Figure 4.12. Random errors removed from the gross error free dataset in Fig. 4.11.

In Table 4.2, mean values of the raw dataset, the gross error free dataset and the reconciled dataset were shown. The mean values for Q_A in the raw dataset (Figure 4.9) was lower compared to the other flows. After the gross errors were removed, the mean for Q_A was closer to the other flows. After the DR the mean values for all four flows were identical.

Table 4.2. Mean values $\pm$ standard deviation of the raw data ($\bar{y}$), the data where gross errors were removed ($\bar{\bar{y}}$) and the reconciled data where random errors were removed ($\bar{\hat{y}}$).

	$\bar{y}$	$\bar{\bar{y}}$	$\bar{\hat{y}}$
Q_{In}	3.04 ± 0.19	3.04 ± 0.19	3.10 ± 0.18
Q_A	2.72 ± 0.05	3.10 ± 0.18	3.10 ± 0.18
Q_F	3.13 ± 0.20	3.13 ± 0.20	3.10 ± 0.18
Q_{Eff}	3.19 ± 0.20	3.19 ± 0.20	3.10 ± 0.18

4.3 Case 3: Comparison to the flow verification method

The comparison was performed on two periods: April 3-10th 2017 (Period 1) and May 15-21th 2017 (Period 2). The rank of Σ was 2, which gave a test criterion $\chi^2 = 10.6$.

4.3.1 Period 1

In Fig. 4.13 b), the test statistic did not exceed the test criterion at any time, thus no gross errors were found during this period.
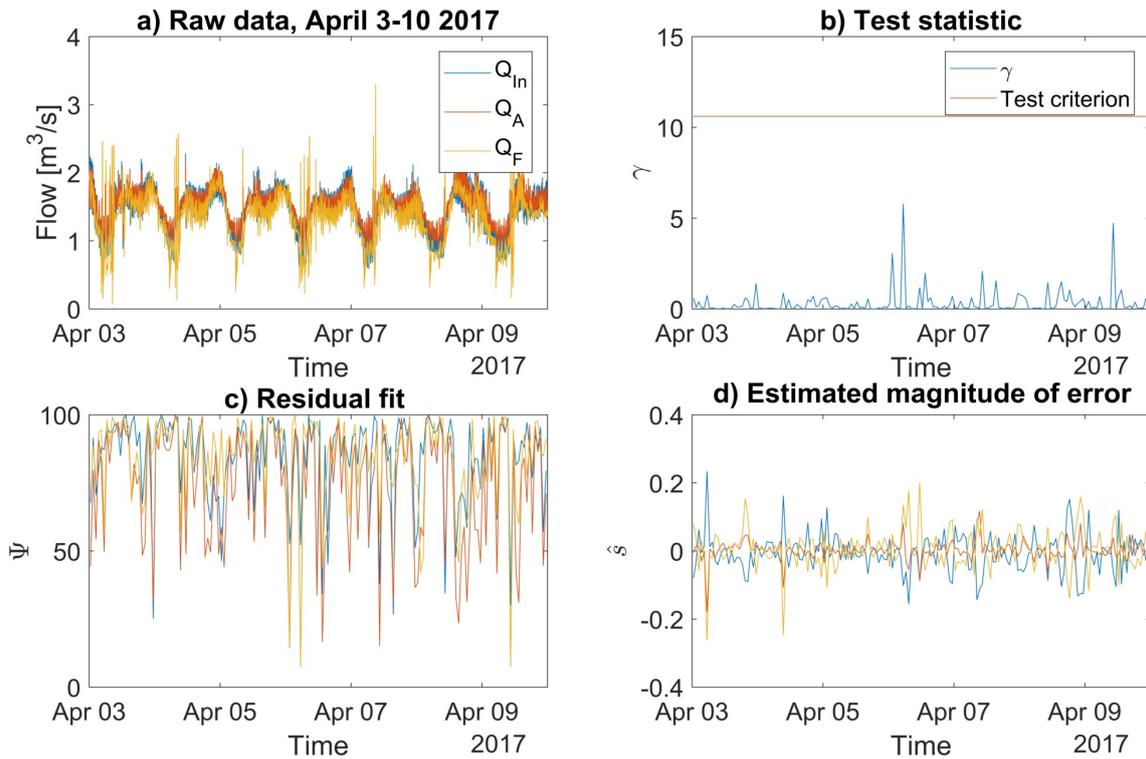

Figure 4.13. The method was tested on a dataset with three flows. a) The dataset on which the method was tested, b) Step 2a, testing if there were significant errors present, c) Step 2b, Locating the error source, d) Step 2c, Estimating the error magnitude.

The results from the flow verification method reveled no major differences between calculated and measured flow in Q_F in any of the filters during this period. The largest difference was about 2.9 % in filter 4 on April 8th 13:00.

4.3.2 Period 2

In Fig. 4.14 b), the test statistic exceeded the test criterion for a short period during the 18th of May.
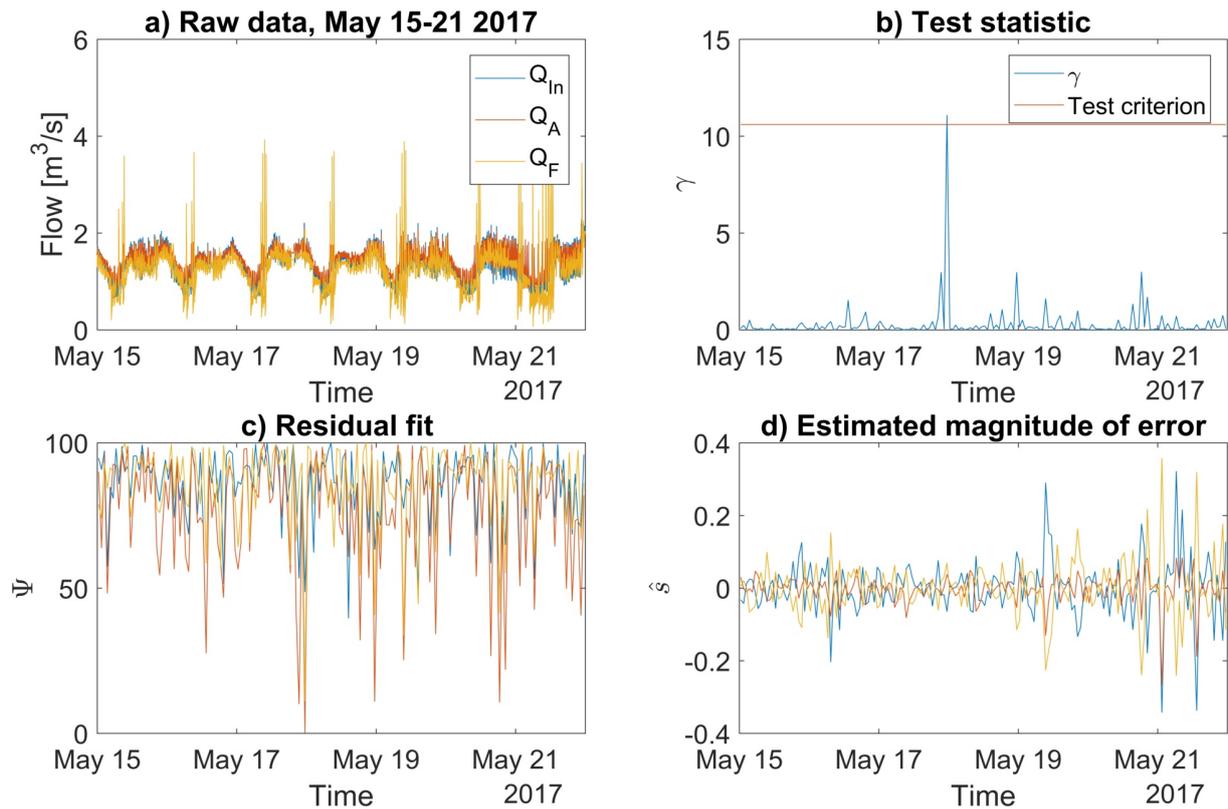
Figure 4.14. The method was tested on a dataset with three flows. a) The dataset on which the method was tested, b) Step 2a, testing if there were significant errors present, c) Step 2b, Locating the error source, d) Step 2c, Estimating the error magnitude.

When zooming in at the period where the error was found (Fig. 4.15), it was difficult to interpret the results. The error was found approximately May 17th 23:30, but from the residual fit (Fig. 4.15 c) it was difficult to determine in which flow the error was in.
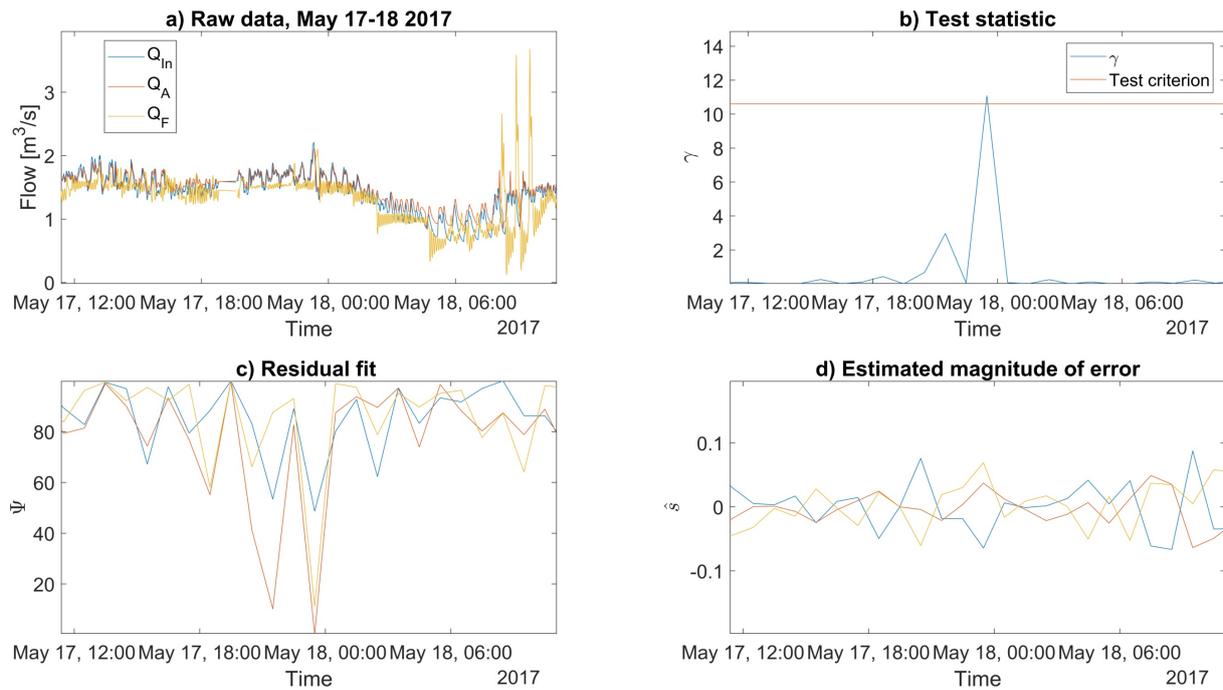
Figure 4.15. Testing the method on a dataset with three flows. a) The dataset on which the method was tested, b) Step 2a, testing if there were significant errors present, c) Step 2b, Locating the error source, d) Step 2c, Estimating the error magnitude.

The results from the flow verification method reveled a few differences between the measured flow and the flow calculated from the level measurement. The largest differences were presented in Table 4.3. The biggest difference was 9,3 % in filter 16 on May 16th. However, non of the differences presented in Table 4.3 could be related to the error found in Fig. 4.15, partly because it could not be concluded if the error was in Q_F but also because the date and time did not match.

Table 4.3. Results from the flow verification method; the largest differences found when the measured flow was compared to the flow calculated from the level measurement between 15-21 May 2017.

Filter no.	Difference [%]	Date
3	8,5	17-May-2017 06:00
3	5,5	18-May-2017 18:42
6	6,5	17-May-2017 18:33
9	5,6	18-May-2017 05:51
16	9,3	16-May-2017 02:15

5 Discussion

The aim with this thesis was to find, implement and evaluate a method based on flow balancing to detect and locate random and gross errors. By compiling relevant literature, a suitable method was found and successfully implemented in MATLAB. The evaluation was performed on simulated data and showed that the method was efficient in detecting an increasing bias and that the estimation of the error magnitude corresponded well to the added bias.

One of the objectives was to improve the accuracy of flow measurements. For that purpose, the method was tested using real WWTP data. The method was able to detect distinct time shifts, which meant that time shifts needed to be corrected prior to using the gross error detection method to avoid false positives. Gross errors were detected as well, although, sometimes it was difficult to interpret the residual fit to determine which flow that contained an error. When a gross error had been detected, and the results were explicit, this error could be corrected using the estimated error magnitude.

In Figure 4.9 where the method was tested on raw data, there was a distinct difference between Q_A and the other flows. Viewing the entire dataset in Figure 3.1 it looked as if this difference occurred because the bypass flow was not included. Thus, the difference probably did not concern a systematic error. However, it was an interesting period to test the method since only one flow differed from the others.

The objective was also to use the chosen method to corroborate the flow verification method used by Stockholm Vatten och Avfall. Nothing explicit could be concluded from this comparison. When constructing the flow balances which were necessary for the chosen method, the 24 measured flows in the sand filter section were combined. Using the flow verification method, each of the 24 flows were examined. This made it difficult to compare the two methods. The method used in this thesis was probably not sensitive enough to detect deviations in one filter basin at the time.

The gross error detection method was not only able to detect and estimate gross errors, but large random errors as well (Fig. 4.11). If the test statistic only exceeded the test criterion for a short period of time, this was most likely evidence of a large random error since gross errors often increased or decreased over time. However, it was important to detect large random errors as well.

The method was implemented in MATLAB, and the script was written so that series of subsets, 60 time steps at the time, was created of the full dataset. Each parameter (residual fit, test statistic etc.) was calculated per subset. The range of 60 time steps was chosen after testing different ranges. When using a smaller range, e.g. 30 data points, smaller deviations and random errors were detected but the results were difficult to interpret. When using a wider range, e.g. 120 time steps, only major errors were detected. Since gross errors often are quite major and increasing/decreasing over time, e.g if a measuring device slowly gets clogged, a wider range is probably preferable. Depending on what type of errors that one wished to detect, the choice of time range was important. If re-doing the tests, a time range of 120 data point would be preferable. This would make the results easier to interpret, and the method would only find major errors. The script was also tested on the entire datasets (the mean value of the entire

period). This approach was not chosen since the method got to insensitive to deviations in the included flows. Also, dividing the datasets into subsets and performing the calculations on these subsets produced vectors with several data points of the test parameters which gave more detailed results.

The interpretability of the results were good when analyzing simulated data. The test statistic steadily increased over time as the difference between the biased flow and the other flows increased. Simultaneously, the residual fit was close to 100 % for the biased flow while it for the other flows quickly decreased towards zero. In case 2 where real WWTP data was analyzed, it was not as easy to interpret the results. Real data had more noise and variation between the flows. As described in van der Heijden et al., [1994b], it could be difficult to conclude which measurement the error descended from (interpreting the residual fit). There were often many possibilities since the residual fit was a probability calculated for each flow and these could be similar. For that reason, the method was tested on short periods, 1-3 days at the time in case 2. However, since all parameters were calculated for determined subsets of the full dataset, it would be possible to analyze longer periods.

The data reconciliation step reduced random noise and yielded datasets that fitted the balance equations exactly. This proved to be an easy and efficient approach to reduce the effect of random errors.

In Fig. 4.2 (d) and 4.4 (d), there was a negative slope on the estimated magnitude of the error for Q_1, Q_3 and Q_4. This was unexpected since the residual fit in both cases implied that only Q_2 was erroneous, thus, the estimated magnitude of the error for Q_1, Q_3 and Q_4 should be zero. However, since the estimated magnitude of the error ($\hat{s}$) was a calculation of how large the difference between the flows were with respect to the balance equations, this was not very strange. When calculating $\hat{s}$, the other test parameters (test statistic and residual fit) were not considered. Basically, in this case, $\hat{s}$ told us that there either was a negative error in Q_1, Q_3 and Q_4 or a positive error in Q_2, and to find out which statement that were correct, one had to observe the residual fit and test statistic first.

The flow verification method had a few disadvantages compared to the method used in this thesis. The information that could be retrieved from the flow verification method was whether or not there was a difference between measured and calculated flow from level measurements in the 24 sand filters, thus if one of these measurements were erroneous. However, it could not be determined in which measurement the error was in. With the suggested method used in this thesis, a lot more information could be gained: did data contain errors, what kind of error (random or gross) was present, which flow measurement the error descend from and the magnitude of the error. Also, the suggested method could be used to analyze different kinds of process data and large systems. The flow verification method demanded manual monitoring, whereas the suggested method easily could be automatized.

In comparison to previous research where the same or similar methods were used, the systems examined in this thesis were very simple. The constructed mass balances basically said that the influent flow equaled the flow in the aeration basins which equaled the flow in the sand filters and so on. This resulted in few constraints and few balance equations. Only analyzing a few

flows, as in this case three or four flows, was probably not optional when using this method. If more than one flow are erroneous, the method was not able to distinguish which flows that were erroneous and which were not. An example of this problem can be seen in Figure 4.11. The test statistic was exceeded but the residual fit was zero for all flows at this time. This was probably because none of the flows correlated well with each other; the method could not point out which flow that did not fit the balance equations. Therefore it would be preferable to analyze more flows, since it is more likely that the method would have several correct flows to compare with.

The serial elimination method was another approach to locate the error source, used by e.g. Xiaolong et al., [2014]. Since the system examined in this thesis were simple, and it was unlikely that errors descended from an incorrect system description, this method could have been used and probably resulted in a decent identification of the error source. However, according to van der Heijden et al., [1994b], the serial elimination method provided ambiguous results. Using van der Heijden's method enabled estimation of the error magnitude, which were one major reason to why it was chosen in this thesis.

In future studies it would be interesting to analyze more flows and more complicated systems to increase the reliability and to see how the method performed. The time range should be increased, mainly to simplify interpretation of the results. There was potential in developing this method to function as a live tool to detect errors, to run the detection script continuously and get a notification whenever a significant error was found, thus get a warning when the flows did not fit the balance equations and something was wrong. The number of flows (or other process data) to be included and the structure of mass balances were straight forward to alter.

It was difficult to compare the results retrieved in this thesis with results from the previous studies. van der Heijden et al., [1994b] examined a fermentation process, and the few authors that examined process data from WWTPs balances, analyzed sludge production systems or other processes, not only water flow as in this study. However, the conclusions in their studies had a common denominator: the method provided valuable information regarding measurement accuracy and errors were detected, which also applies on the results from this study.

6 Conclusions

In this thesis, a suitable method for detecting gross and random errors was found and successfully implemented. The evaluation showed that the method was efficient in detecting gross errors when only one out of the analyzed flows were biased, and when that bias was large. The estimation of the magnitude of such an error was good, which made it straight forward to correct data and to obtain a dataset without gross errors.

Furthermore, the following conclusions could be made

- The gross error detection method was not only able to detect gross errors, but large random errors as well

- Time shifts needed to be corrected prior performing the gross error detection, otherwise

this deviation would be assessed as a gross error

- A larger number of flows should be analyzed to increase the credibility of the results

- A wider time range should be chosen to reduce the detection of random errors and to simplify the interpretation of the results

- The DR step reduced the variation and yielded datasets that fitted the balance equations exactly

- The method had potential to be used as a live tool

References

S. Bell. A beginner's guide to uncertainty of measurement. *Measurement Good Practice Guide*, 2:11, 2001.

D. Brdjanovic, C. F. Meijer, S, M. H. C. M. Lopez-Vazquez, C, and M. van Loosdrecht, Mark C. *Applications of Activated Sludge Models*. IWA Publishing, London, UK, 2015.

H. B. Exell, R. Error analysis. http://www.jgsee.kmutt.ac.th/exell/PracMath/ErrorAn.htm, 2001.

H. Le, Q, J. T. Verheijen, P, E. Mampaey, K, C. M. van Loosdrecht, M, and I. P. Volcke, E. Non-linear data reconciliation for a partial nitritation (sharon) reactor. *IFAC-PapersOnLine*, 49-7:1139–1144, 2016.

C. F. Meijer, S, H. Van der Spoel, S. Susanti, J. Heijnen, J, and C. M. Van Loosdrecht, M. Error diagnostics and data reconciliation for activated sludge modelling using mass balances. *Water Science and Technology*, 45:145–156, 2002.

S. Narasimhan and C. Jordache. *Data Reconciliation and Gross Error Detection, An Intelligent Use of Process Data*. Gulf Publishing Company, Houston, Texas, 2000.

OOF. Weirs. https://www.openchannelflow.com/weirs, 2017.

Plasma Processing Laboratory. Bbl: Matlab data reconciliation. http://ppl.research.mcgill.ca/matlab-105/, 2013.

S. Puig, M. van Loosdrecht, J. Colprim, and C. F. Meijer, S. Data evaluation of full-scale wastewater treatment plants by mass balance. *water research*, 42:4645–4655, 2008.

L. Seungchul, R. Sankara, K. MinJeong, E. Iman, Janghorban, and Y. ChangKyoo. Assessment of environmental data quality and its effect on modelling error of full-scale plants with a closed-loop mass balancing. *Environmental Technology*, 36 No. 24:3253–3261, 2015.

Stockholm Vatten. Bromma avloppsreningsverk - för stockholmarnas och miljöns bästa. http://www.stockholmvattenochavfall.se/globalassets/pdf1/informationsmaterial/broschyrer/avloppsverk/bromma_webb.pdf, ND.

M. Thomann. Quality evaluation methods for wastewater treatment plant data. *Water Science and Technology*, 57:1601–1609, 2008.

T. J. M. van der Heijden, R, J. Heijnen, J, C. Hellinga, B. Romein, and C. A. M. Luyben, K. Linear constraint relations in biochemical reaction systems: 1. classification of the calculability and the balanceability of conversion rates. *Biotechnology and Bioengineering*, 43: 3–10, 1994a.

T. J. M. van der Heijden, R, J. Heijnen, J, C. Hellinga, B. Romein, and C. A. M. Luyben, K. Linear constraint relations in biochemical reaction systems: 2. diagnosis and estimation of gross errors. *Biotechnology and Bioengineering*, 43:11–20, 1994b.

T. J. M. van der Heijden, R, J. Heijnen, J, C. Hellinga, B. Romein, and C. A. M. Luyben, K. Linear constraint relations in biochemical reaction systems: 3. sequential application of data reconciliation for sensitive detection of systematic errors. *Biotechnology and Bioengineering*, 44:781–791, 1994c.

J. Xiaolong, L. Pei, and L. Zheng. Data reconciliation and gross error detection for operational data in power plants. *Energy*, 75:14–23, 2014.
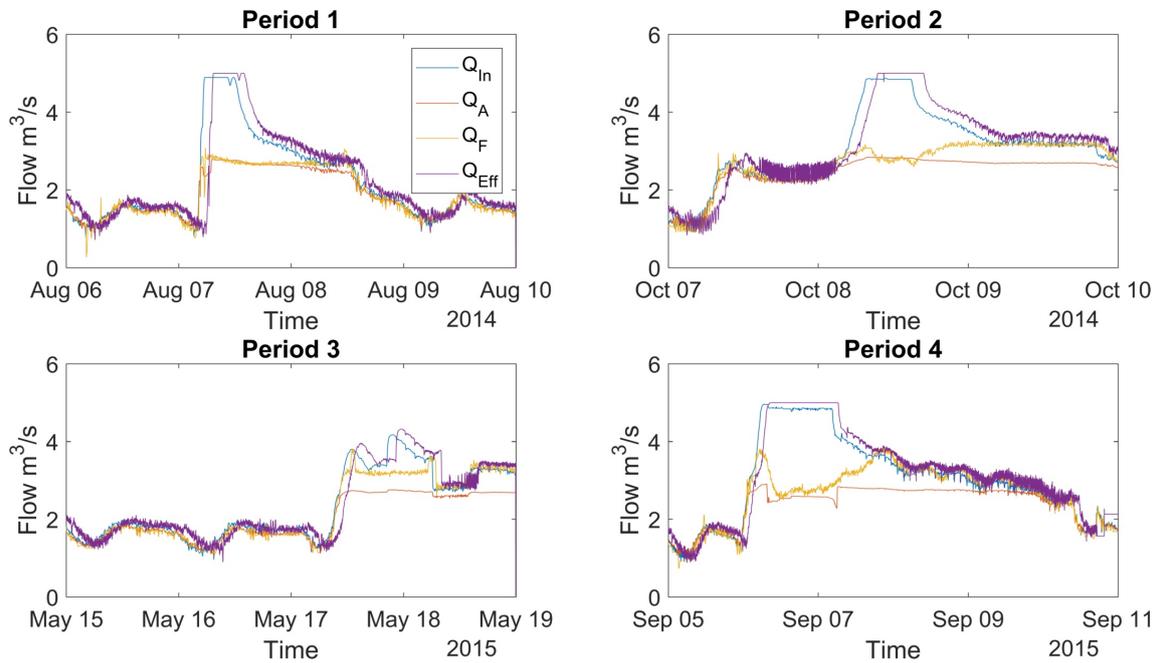
Appendices

A

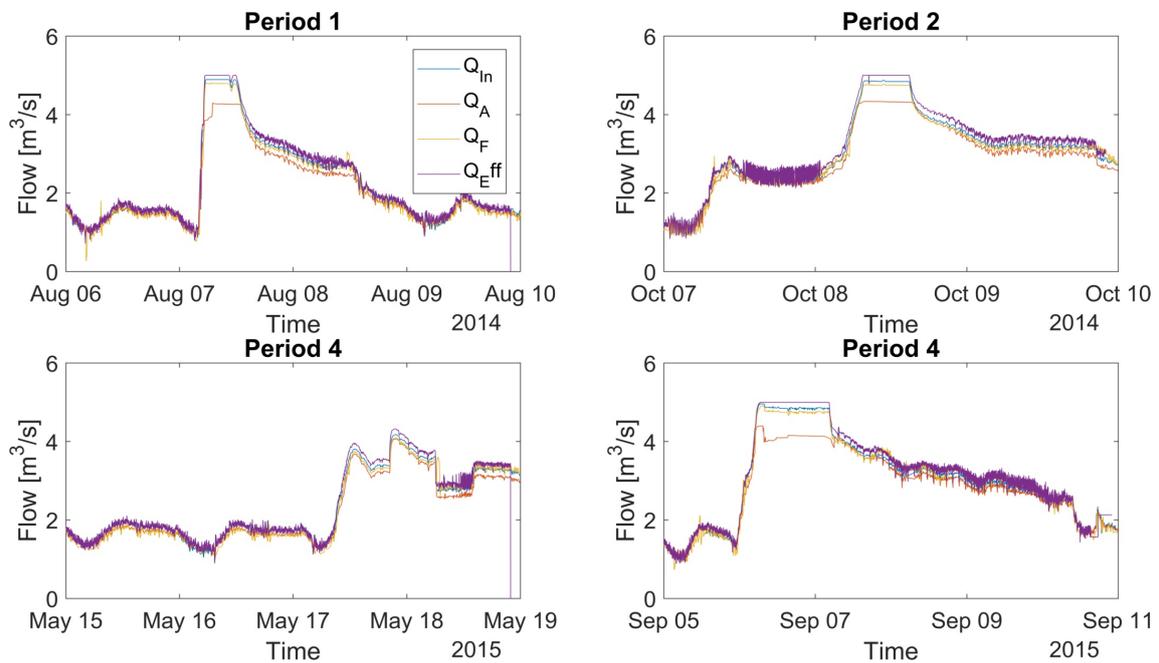Fig. A .1. Available datasets in Case 2.

Fig. A .2. Available datasets corrected for time shift and bypass flows in Case 2.
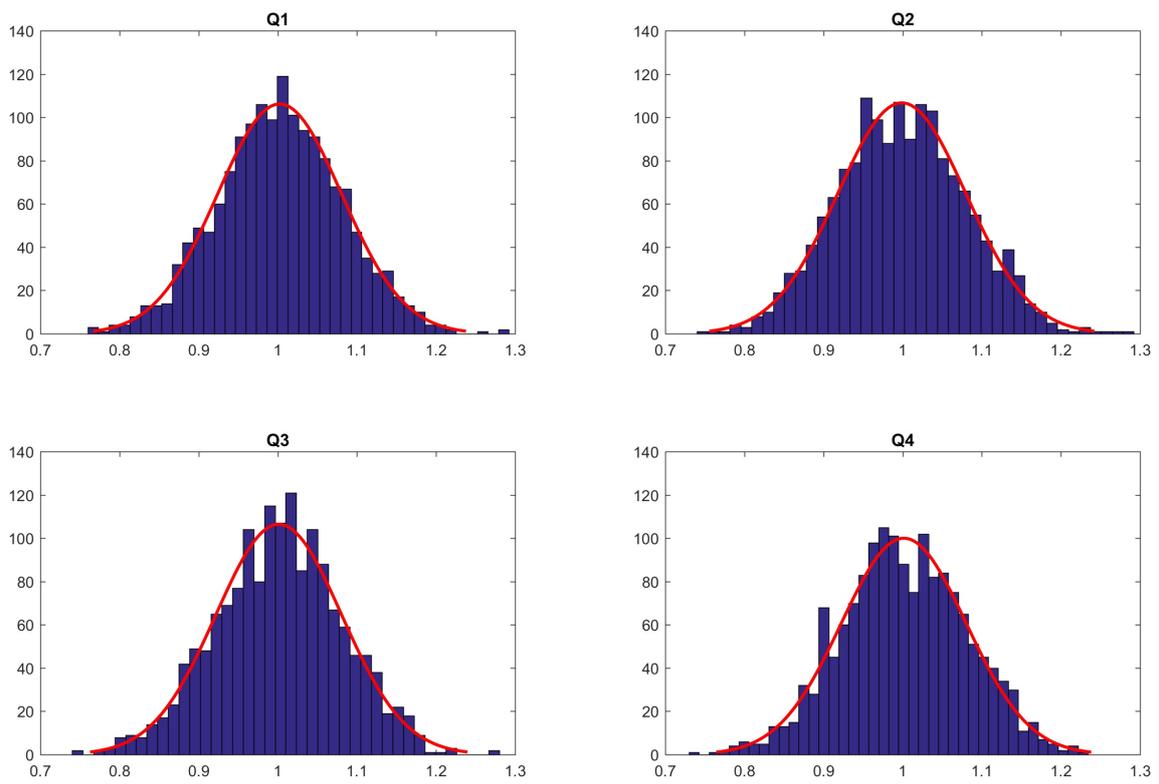
B

Fig. B .3. Histogram with a distribution fit of the simulated signals $Q_1 - Q_4$ in Simulated dataset 1 and 2 (the histogram showed the distribution of Q_2 without the increasing bias). The signals had a variance between 0.0065 - 0.0067.
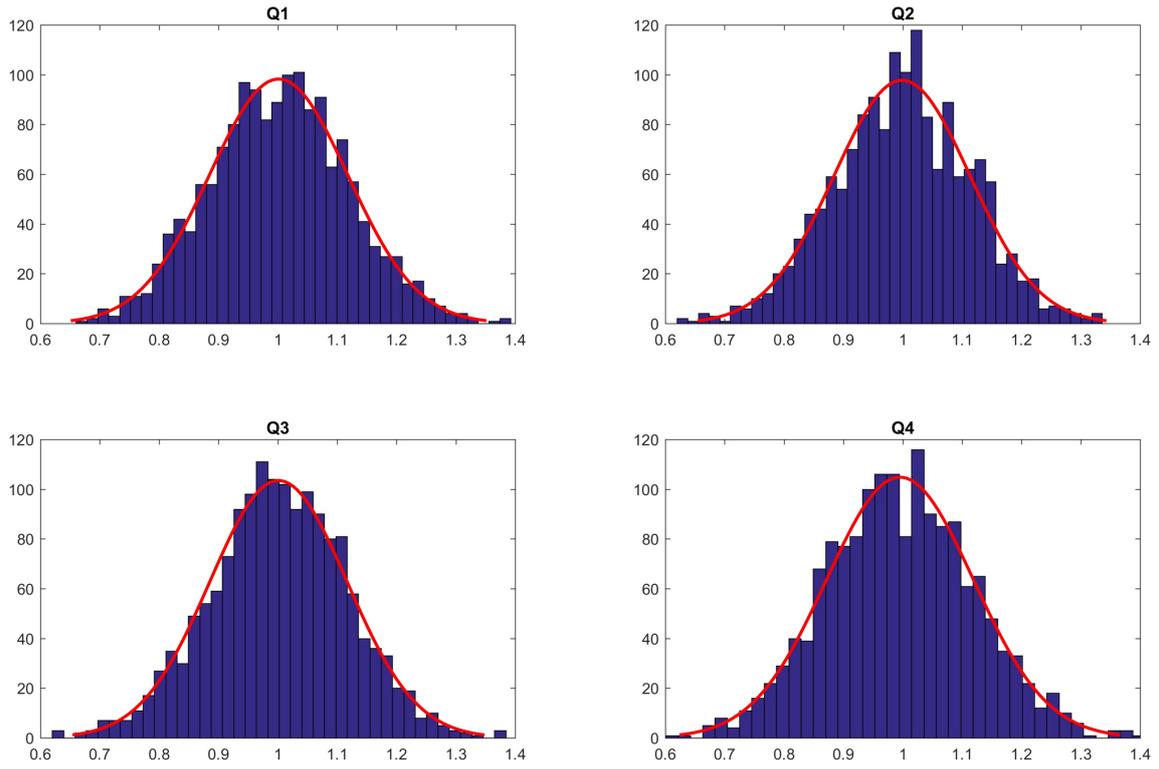
Fig. B .4. Histogram with a distribution fit of the simulated signals $Q_1 - Q_4$ in Simulated dataset 3 (the histogram showed the distribution of Q_2 without the increasing bias). The signals had a variance between 0.013 - 0.015.