A. Spindler. Structural redundancy of data from wastewater treatment systems.
Determination of individual balance equations. Water Research, 57:193–201, June 2014.
https://doi.org/10.1016/j.watres.2014.03.042
PMID: 24721666
© 2014. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/

Structural redundancy of data from wastewater treatment systems. Determination of individual balance equations.

A. Spindler
Institute of Water Quality and Resource Management, Vienna University of Technology, Karlsplatz 13/226-1, 1040 Wien, Austria (E-mail: *a.spi@iwag.tuwien.ac.at*)

Abstract

Although data reconciliation is intensely applied in process engineering, almost none of its powerful methods are employed for validation of operational data from wastewater treatment plants. This is partly due to some prerequisites that are difficult to meet including steady state, known variances of process variables and absence of gross errors. However, an algorithm can be derived from the classical approaches to data reconciliation that allows to find a comprehensive set of equations describing redundancy in the data when measured and unmeasured variables (flows and concentrations) are defined. This is a precondition for methods of data validation based on individual mass balances such as CUSUM charts. The procedure can also be applied to verify the necessity of existing or additional measurements with respect to the improvement of the data's redundancy. Results are given for a large wastewater treatment plant. The introduction aims at establishing a link between methods known from data reconciliation in process engineering and their application in wastewater treatment.

Keywords
data validation; gross error detection; mass balancing; observability; redundancy

1 Introduction

This work discusses a fundamental approach to the validation of operational data from wastewater treatment plants through mass balancing. Historic records of plant data reflect the performance of a treatment plant and are regularly exploited for monitoring, benchmarking and simulation, to adjust control strategies and to plan for process redesign or plant extension. However, poor quality of historic data records is the main obstacle for these tasks. This has been agreed upon widely in literature (e.g. Rieger et al., 2010; Puig et al., 2008; Meijer et al., 2002; Barker and Dold, 1995) as well as different IWA workshops on this question (e.g. Mont Sainte-Anne 2010, Budapest 2011).

The type of operational data typically used for these tasks are daily flow volumes and concentrations measured in 24h-composite samples (where flow-proportionality is required for matching balances, especially in flows with strongly varying concentrations such as the influent). Higher frequency sensor data is more relevant in automated process control and therefore not of primary interest here. However, sensor readings are usually adjusted to the less frequent but more reliable laboratory measurements. Therefore, the validation of operational data from composite samples is also of considerable relevance for plant control.

Spindler and Vanrolleghem (2012) showed that the application of CUSUM charts is a suitable approach to continuous mass balancing[1] and detects off-balance periods more reliably than mass balances based on long term averages of data. Continuous mass balancing following this method requires individual balance equations which describe redundancy of the measured data.

This work will provide a procedure for the computational determination of the complete set of possible redundancy equations (also: balance equations) for a given plant layout. This aim is different from, but closely related to the principles and objectives of *data reconciliation*. With mass balancing as the key to data reconciliation and gross error detection, there appears to exist a gap between development and application of methods used in process engineering and wastewater treatment. Therefore a very short overview and comparison of the developments in both fields is given in the following parts of the introduction. After the presentation of the proposed method results will be given for its application to a large and complex wastewater treatment plant.

1.1 Data reconciliation in process engineering

Data reconciliation has developed mainly in the field of (chemical) process engineering. It allows to improve the measured values of process variables such as flows and concentrations based on the laws of conservation. Data reconciliation requires *redundancy* of the measured variables which means that they can also be calculated from other measured variables.

A vast amount of literature exists. Research began some 50 years ago when the concept of data reconciliation was introduced by Kuehn and Davidson (1961). Further research developed initially in two lines – the *topology oriented approach* first presented by Václavek (1969; Václavek and Louĉka, 1976) and the *equation oriented approach*, represented among others by Crowe (1986; Crowe et al., 1983). Some of the most recent progress in the field has been achieved by Kelly (e.g. 1998; 2004). Four comprehensive books have been written (Madron and Veverka, 1992; Narasimhan and Jordache, 2000; Romagnoli and Sánchez, 2000; Bagajewicz, 2010). Good overviews about research development are also provided in Crowe (1996) and Ponzoni et al. (1999).

A basic step in data reconciliation is the classification of the process variables. A process variable can either be directly *measured* (observed) or *unmeasured*. Unmeasured refers to variables that could be measured (at least theoretically) but are not for some reason. A process variable is *observable*, if it can be calculated from a subset of other measured variables. Measured observable process variables are called *redundant*. Crowe (1989) also classifies *barely observable* (unmeasured) variables which require at least one non-redundant measured variable to be calculated. *Structural* redundancy refers only to the theoretical calculability of a measured variable while *practical* redundancy also considers numerical and statistical accuracy of this calculation. The following short

[1] The application of CUSUM charts had originally been labelled “dynamic mass balancing” to differentiate from the established approaches. But because it does not actually target kinetic rates this naming will be avoided in the future.

example is given to illustrate the difference between structural and practical redundancy.

The volume of dewatered sludge is negligible compared to influent and effluent of a wastewater treatment plant. For structural redundancy of the overall flow it would, however, still be required to be measured. Obviously the amount of dewatered sludge cannot be reconciled from this balance as the propagation of errors would pose a very high uncertainty on this calculation. On the other hand, in- and effluent would still be practically balanceable without the amount of dewatered sludge being measured.

1.2 Data validation in wastewater treatment

So far the concept of data reconciliation has received little attention in wastewater treatment. This becomes obvious in the terminology. The term *mass balance* is prevalent, possibly inspired by the work of Nowak (1994; 1999). Rieger et al. (2010) actually refer to the *order of redundancy* as "overlapping balances". It reveals the practitioner's perspective where the individual mass balances receive higher attention than the reconciliation of the entire data set. This will be discussed further in the following section.

Literature in wastewater treatment focuses mainly on sensor fault detection and so far hardly regards redundancy of measurements. Until recently wastewater related literature cited only two works from the field of data reconciliation in process engineering (Meijer et al., 2002; Puig et al., 2008; Schraa et al., 2006).

Van der Heijden et al. (1994) adapt research from the field of chemical process engineering and apply it to elemental mass balances in fermentation processes. Following works in the field of wastewater treatment (Meijer et al., 2002; Puig et al., 2008) apply the methods of Van der Heijden et al. (1994) thus re-adapting them back into process oriented applications where they originally stem from. Meijer (2002) stress the importance of validation of operational data for use in simulation studies. Puig et al. (2008) point out that the dynamic nature of wastewater treatment makes mass balancing difficult. Both works rely exclusively on the method developed by Van der Heijden et al. (1994) which was implemented in the software Macrobal (Hellinga, 1992). However, when applying data reconciliation to elemental mass balances (Macrobal's purpose) the composition of substances is exactly known (fixed) which is not the case for the composition of wastewater treatment streams. Hence only in volumetric and mass flow rates the measurement variability was accounted for, but not in measured concentrations. Additionally, the high variability of flow measurements (around 50% relative standard deviation) includes process dynamics which is disputable given the fact the steady state is a prerequisite for the applied method of data reconciliation.

Schraa, et al. (2006) does mention data reconciliation citing Crowe (1996) but focuses on sensor fault detection. He did investigate data reconciliation in an earlier publication (Schraa and Crowe, 1998) when he was not yet involved with wastewater treatment.

Very recently two papers on redundancy classification and fault detection based on mass balances where published by Villez et al. (2013a; 2013b). In both papers the methods of data reconciliation are explicitly applied to (synthetic) data from wastewater treatment. The basic applicability of these methods is proven for the situation of sludge thickening in a settler. In the paper on redundancy classification (Villez et al., 2013a) influent TSS is concluded to be observable when measurements are taken only in the activated sludge tank, the wastage sludge and the effluent. The example obviously refers to inorganic TSS in a plant without chemical phosphorus precipitation.

1.3 Data reconciliation vs. individual mass balancing

In data reconciliation the aim is to adjust the entire data set to fit the constraints. To achieve this, the remaining random error (after removal of *gross errors*) is distributed over all variables according to an allowance that is defined by the variance of the single measurement errors. The variance of the measurement error needs to be known. *Steady state* is another frequent requirement for the established methods of data reconciliation. Even though approaches to integrated data reconciliation and gross error detection exist, considerable difficulties remain in dynamic systems (Narasimhan and Jordache, 2000).

In many industrial applications the preconditions for data reconciliation are met closely enough for its successful application. Substance influents to processes are usually controlled and set point changes of such controlled variables have rather low frequencies. In contrast, the influent is the main disturbance to the process of wastewater treatment and makes the dynamic adjustment of actuators such as pumps and blowers a constant challenge. Therefore wastewater treatment plants, especially those with combined sewer influent, are *dynamic systems*. This is also true if the measured data consists of daily means of the process variables (flow sums / composite samples). Another important difference to many industrial processes are the low concentrations and significant heterogeneity (dissolved/suspended) of the relevant compounds. The various sources of measurement random errors (representative sampling, interference from additional compounds, range of expected values, dynamic flows and concentrations) add up to comparatively larger uncertainty and make it complex and time-consuming, if not impossible, to determine the random measurement error variances.

Continuous balancing by means of CUSUM charts avoids these two main obstacles. The input variable to this method is the error vector of daily mass balances and therefore error distributions of the single measurements do not need to be known a priori. Continuous mass balancing has been proven suitable for gross error detection in dynamic systems (Spindler and Vanrolleghem, 2012). It requires individual balance (redundancy) equations, the determination of which is addressed in the following.

2 Methods

The single steps to determine individual redundancy equations which consist only of measured variables are provided below. While the setup of the incidence matrix and classification of redundancy and observability (steps 1a and 2) are typical for data reconciliation, steps 1b and 3 (incidence matrix expansion and elimination of observable variables) are characteristic for the algorithm described here. It follows the idea, that an observable (i.e. calculable) variable can be removed from an equation by expressing it in terms of other (measured) variables. If the observable variable can be calculated in various ways, several different redundancy equations are found.

Step 1: Incidence matrix setup and expansion

The description of a flow network is commonly given as *directed incidence matrix M*, where columns represent streams (edges in the network graph) and rows represent single subsystems (nodes in the network graph). The *environmental node* (Mah et al., 1976) is the source and sink of streams coming into and leaving the overall system, it represents the outside world. The values a_{ij} of matrix *M* are:

- 1, if stream *j* enters node *i*,
- -1, if stream *j* leaves node *i* and
- 0, if stream *j* is not incident with node *i*.

A complete incidence matrix *M* consists of *m* independent rows where *m* is equal to the number of nodes in the process network. In its most evident form the rows of the incidence matrix represent the single nodes themselves (or subsystems, e.g. an activated sludge tank). The representation of a single node in the incidence matrix can be directly transformed into linear and bilinear equations describing (mass) flow in and out of the corresponding subsystem.

Following its setup, the incidence matrix *M* is expanded to represent *all possible* combinations of single subsystems of the given process network. This is achieved by finding all XOR-combinations of the *m* linearly independent rows in *M*. The new resulting matrix is *M2*. It needs to be reduced to *M3* in an extra step because *M2* is likely to contain rows of zero, double entries and rows that represent combinations of subsystems which do not share any stream and thus are physically independent of one another. For example, thickening and dewatering facilities of a wastewater treatment plant usually do not share any input or output streams. When setting up redundancy equations, these types of balances should be avoided. The procedure to clean *M2* of the latter type of unnecessary rows is simply by stepwise comparison of each row with all other rows. If other rows have entries different from zero in exactly the same columns as the current row (and maybe more) they can be deleted. A graph theoretical approach to finding the relevant set of subsystem combinations might proof more efficient but was not investigated here.

Step 2: Classification of redundancy and observability

In wastewater treatment, the equations that describe a balance around a node can be of two types. *Flow balances* contain only measured (volumetric) flows and are therefore *linear*. *Mass balances* of a specific compound are calculated from the products of flows and the compound's concentration in each stream and are therefore *bilinear* (a linear structure composed of simple products). Mass flows that are not bound to a water flow such as methane, nitrogen and oxygen uptake (digested COD) are linear parts of otherwise bilinear balance equations. Mass flow or concentration of a compound can also be zero in certain streams such as phosphorus in the gas phase. This can be relevant when equations are actually set up in step 3.

The linear and bilinear nature of the equations that describe flow and mass flow in wastewater treatment simplifies the classification of observability and redundancy (as opposed to sometimes nonlinear equations in process engineering). A straightforward classification method for the bilinear case has been described by Ragot et al. (1990). They base their classification of observability on a simple analysis of measured and unmeasured flows and concentrations around the single nodes. It yields that only one unmeasured concentration of a compound can be calculated from a single balance equation and only if all flows of that balance are observable. Flows on the other hand might be calculated from known concentrations, too. As proposed in Ragot et al. (1990) the procedure is iterative and stops when no further observable variables are found. Here, the algorithm is adapted to determine both redundancy and observability in each node (row of *M3*). The necessity of iteration is met in step 3.

For a single node it might be possible to directly set up a flow or mass balance equation, to set up a balance equation through elimination of (an) unmeasured flow(s), or to calculate a flow, concentration or mass flow. The rules are:

(1) If all flows Q are measured, a redundancy equation can be set up.
   (a) If additionally all concentrations or mass flows of one compound are measured, another redundancy equation can be set up.
   (b) If only one concentration or mass flow of a compound is unmeasured, it can be calculated in this node (for later elimination in another node).
(2) If only one flow Q is unmeasured it can be calculated from the other flows in this node (for later elimination in another node).
(3) If one or more flows Q are unmeasured and
   (a) there are as many or more compounds with all concentrations / mass flows measured than missing flows, the missing flows can be eliminated and a redundancy equation for this node be set up.
   (b) the number of compounds with all concentrations / mass flows measured is one less than the number of unmeasured flows Q , the missing flows can still be calculated in this node (for later elimination in another node).
   (c) only one concentration or mass flow of a compound is unmeasured and the number of other compounds with all concentrations / mass flows measured is not less than the number of unmeasured flows Q, still all unmeasured values can be calculated in this node (for later elimination in another node).

Some additional attention has to be paid to nodes such as splitters, where a compounds concentration is equal in all streams. Therefore unmeasured flows cannot be calculated from known concentrations in a splitter. The only meaningful redundancy equations for these nodes are those for flow Q. Splitters have to be indicated seperately. In a system with only 3 streams, no storage and just one compound X the classification can be illustrated easily (Figure 1).

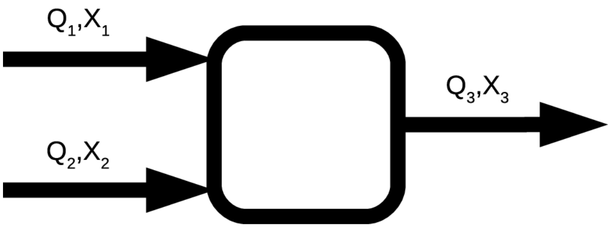

Figure 1: Single system with 2 input streams and 1 output stream, carrying 1 component (X)

The balance equations are:

$Q_1 + Q_2 + Q_3 = e_{1a}$ (1a) flow balance

$Q_1 X_1 + Q_2 X_2 + Q_3 X_3 = e_{1b}$ (1b) mass balance

Each balance equation yields an error *e* with an expected value of zero.

When Q_1, Q_2, Q_3, X_1 are measured and X_2, X_3 unmeasured, only the flows *Q* are redundant (eq. 1a). When another concentration, e.g. X_2 is measured, the remaining concentration X_3 becomes observable but all concentrations are still not redundant.

When X_1, X_2, X_3, Q_3 are measured, none of them are redundant and all flows are (barely) observable. When another flow, e.g. Q_2 is measured, the concentrations become redundant, Q_2 and Q_3 are redundant and Q_1 is observable. The redundancy equation becomes:

$Q_1 X_1 + Q_2 X_2 - (Q_1 + Q_2) X_3 = e_{1c}, \quad X_1 \neq X_2 \neq X_3$ (1c)

The method this classification of observability and redundancy is based on (Ragot et al. 1990) is especially obvious and simple to follow. Other methods – which give the same classification results – often require more involved mathematics and/or are part of the iterative reconciliation process thus depending on measurement data. A good overview is given in Bagajewicz (2010).

Step 3: Elimination of observable variables and setup of redundancy equations

For the computerized setup of the actual redundancy equations software capable of symbolic calculations is beneficially applied. First, for each row in *M3* that contains variables observable in that node, one or several equations that solve for this variable can be written. After one cycle through the rows of *M3* there exists an incomplete set of equations that can be used to calculate observable variables. With the observable variables assumed to be measured, another cycle starts after the repetition of step 2. This is repeated until no further equations to solve for observable variables are found.

Finally, for each balance in *M3* that contains no unobservable variables the redundancy equations are set up with the observable variables being replaced by their solving equations. Each balance equation then consists only of redundant variables. In case several equations are available for the calculation of an observable variable, multiple redundancy equations will be set up for this balance. It is advisable to limit the number of replaced observable variables in the redundancy equations to control complexity of the resulting equations.

The procedure was implemented using the Sage Mathematics Software (Stein et al., 2012). Sage itself relies on a number of other computational programs out of which use has primarily been made of The R Project for Statistical Computing (R Core Team, 2013) as well as Singular (Decker et al., 2011) and Maxima (2013) for symbolic calculations.

Simpel sensor placement

The expansion of *M* into *M3* can also be applied to determine useful additional measurements. Assuming that in order to establish redundancy of a measured variable the new redundancy equation should be simple and contain only few variables, it follows that it will be taken directly from a row in *M3*. Therefore, the linear and bilinear redundancy equations resulting from *M3* simply need to be scanned for those that contain both the variable that should become redundant and at the same time the minimum number of unmeasured variables, preferably only one. This unmeasured variable(s) need to be measured additionally. While this approach to sensor placement is utile due to its simplicity, it is also limited. It does not guaranty the smallest possible number of additional measurements in order to establish overall redundancy of a given variable but does provide for a simple redundancy equation. It does not aim at data reconciliation either.

3 Results

Results are presented for the application of the above method to a large two-stage wastewater treatment plant (160.000 p.e.). The numbering of the subsections is in accordance with the single steps in the methods section.

The plant layout of the application example is given in Figure 2. The plant treats wastewater from various municipal (M1-M4) and industrial (I1-I4) sources. For a full analysis, all flows regardless of their size are included with only the polymer and precipitant dosage being neglected. For example, the main industrial source (I3) is sampled in a side stream and for that reason a splitter can be found in the plant layout. The mass flows leaving AST1 and AST2 and labelled "gas" refer to oxygen uptake and elementary nitrogen. Because each activated sludge tank and its clarifier are one functional unit they are not separated.

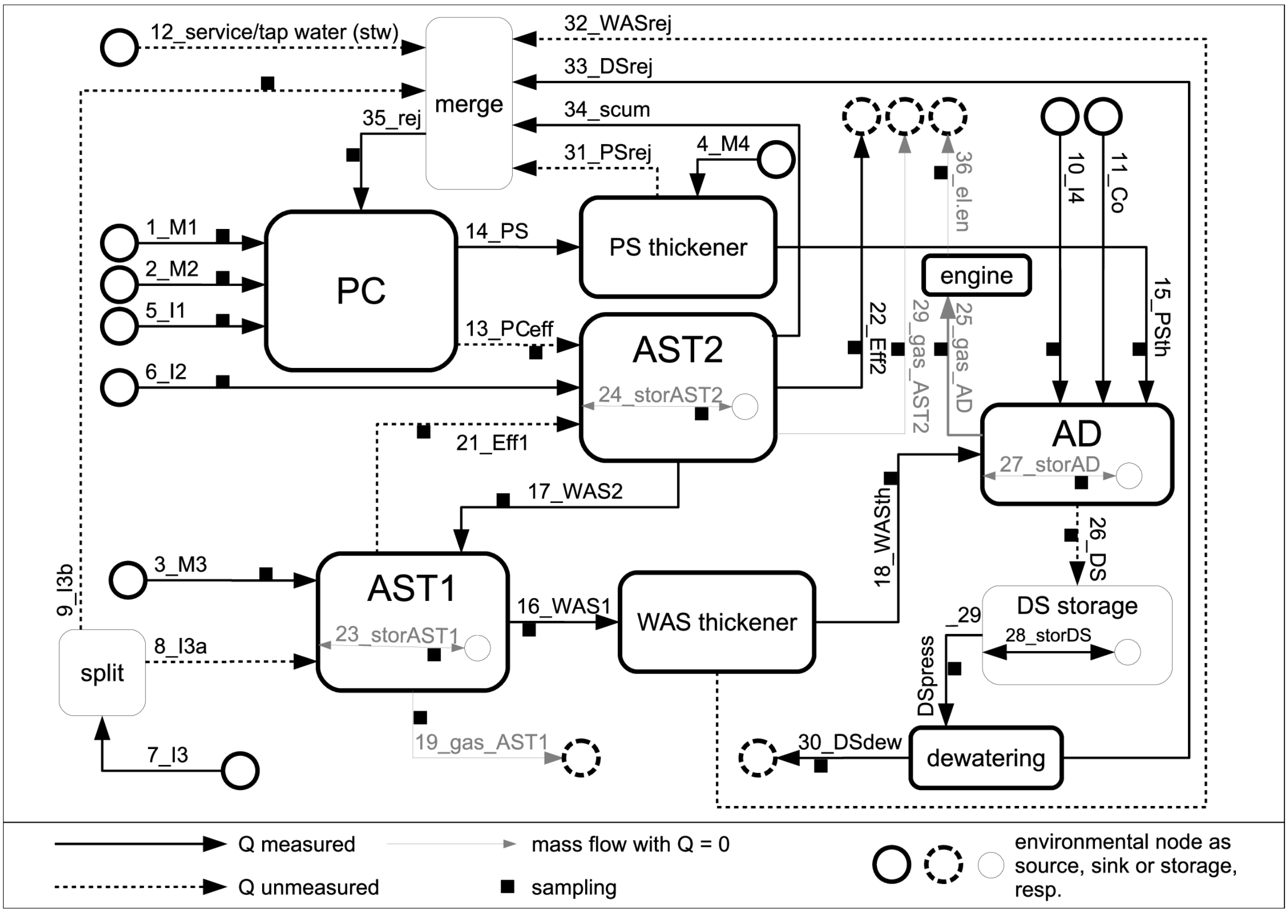

Figure 2: Plant layout of the application example.

Step 1: Incidence matrix setup and expansion

The incidence matrix *M* resulting from the plant layout is given in Table 1. It has 11 independent rows (subsystems) and 36 columns (streams).

Table 1: Incidence matrix M describing the example plant layout.

	1 M1	2 M2	3 M3	4 M4	5 I1	6 I2	7 I3	8 I3a	9 I3b	10 I4	11 Co	12 stw	PCeff	14 PS	PSth	16 WAS1	17 WAS2	WASth
PC	1	1	·	·	1	·	·	·	·	·	·	·	-1	-1	·	·	·	·
PS thick.	·	·	·	1	·	·	·	·	·	·	·	·	·	1	-1	·	·	·
AST1	·	·	1	·	·	·	·	1	·	·	·	·	·	·	·	-1	1	·
AST2	·	·	·	·	·	1	·	·	·	·	·	·	1	·	·	·	-1	·
WAS thick.	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	1	·	-1
AD	·	·	·	·	·	·	1	·	·	·	1	·	·	·	1	·	·	1
DS storage	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
Dewatering	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
split	·	·	·	·	·	·	·	-1	-1	1	·	·	·	·	·	·	·	·
merge	·	·	·	·	·	·	·	·	1	·	·	1	·	·	·	·	·	·
Gas engine	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
env. node	-1	-1	-1	-1	-1	-1	-1	·	·	-1	-1	-1	·	·	·	·	·	·

	19 gas AST1	20 gas AST2	21 Eff1	22 Eff2	23 stor AST1	24 stor AST2	25 gas AD	26 DS	27 stor AD	stor DS	29 DS press	30 DS dew	PS rej	WAS rej	DS rej	34 scum	35 rej	36 el.en
PC	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	1	·
PS thick.	·	·	·	·	·	·	·	·	·	·	·	·	-1	·	·	·	·	·
AST1	-1	·	-1	·	1	·	·	·	·	·	·	·	·	·	·	·	·	·
AST2	·	-1	1	-1	·	1	·	·	·	·	·	·	·	·	·	-1	·	·
WAS thick.	·	·	·	·	·	·	·	·	·	·	·	·	·	-1	·	·	·	·
AD	·	·	·	·	·	·	-1	-1	1	·	·	·	·	·	·	·	·	·
DS storage	·	·	·	·	·	·	·	1	·	-1	·	-1	·	·	·	·	·	·
Dewatering	·	·	·	·	·	·	·	·	·	·	-1	1	·	·	-1	·	·	·
split	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
merge	·	·	·	·	·	·	·	·	·	·	·	·	1	1	1	1	-1	·
Gas engine	·	·	·	·	·	·	1	·	·	·	·	·	·	·	·	·	·	-1
env. node	1	1	·	1	-1	-1	1	·	-1	1	1	·	·	·	·	·	·	1

The variables' division into measured and unmeasured flows and concentrations is indicated in Figure 2 and explicitly given in Table 2. Note that for the splitter, all concentrations are known (measured) despite only one sampled.

The expansion *M2* of Matrix *M* yields a total of 2047 different combinations of subsystems (2^m*-1, m=11*). The number of subsystem combinations increases exponentially with the number of independent subsystems. When reduced to *M3*, 688 combinations of subsystems remain for which linear and bilinear balance equations could be set up.

Table 2: Classification of measured and unmeasured flows, concentrations and mass flows.

Flow	**COD**
zero:19,20,23,24,25,27,36	zero:---
	concentration
measured:1,2,3,4,5,6,7,10,11,14,15,16, 17,18,22,28,29,30,33,34,35	measured:1,2,3,5,6,7,8,9,10,13,15,16, 17,18,21,22,26,28,29,30,35
unmeasured:8,9,12,13,21,26,31,32	unmeasured:4,11,12,14,31,32,33,34
	mass flow
	measured:19,20,23,24,25,27,36
	unmeasured:---
Phosphorus	**Nitrogen**
zero:19,20,25,36	zero:25,36
concentration	***concentration***
measured:1,2,3,5,6,7,8,9,13,15,16, 17,18,21,22,26,28,29,30,35	measured:1,2,3,5,6,7,8,9,10,13,15,16, 17,18,21,22,26,28,29,30,35
unmeasured:4,10,11,12,14,31,32,33,34	unmeasured:4,11,12,14,31,32,33,34
mass flow	***mass flow***
measured:23,24,27	measured:23,24,27
unmeasured:---	unmeasured:19,20

Step 2: Classification of observability and redundancy

Overall, there are 96 measured and 35 unmeasured variables in the example. Of the measured variables 81 are redundant and 22 unmeasured variables are observable. There are 21 measured flows (all but one redundant) and 75 measured concentrations and mass flows out of which 14 remain structurally not redundant. The 8 unmeasured flows in the example are all observable and out of the 27 unmeasured concentrations or mass flows 14 can still be calculated from other variables. Because the classification of observability and redundancy is not the primary aim of this work, the detailed results for each individual variable are not included here.

Step 3: Elimination of observable variables and setup of redundancy equations

Based on the division into measured and unmeasured variables, only 4 linear balance equations out of the 688 different subsystem combinations can be readily calculated with all their components being measured. Three of those are the equations describing flow balances around the anaerobic digester and the dewatering facilities (see Figure 2 for comparison). The validation of flows 10, 11, 15, 18, 28, 29, 30, 33 is possible from these equations. Only one directly available redundancy equation for a compound can be found. It is the simple balance around the gas engine, where the methane content of the gas and the electrical efficiency of the engine are needed to calculate the COD mass flows (superscript *mf* refers to “mass flow”). The respective equations are:

$$\begin{aligned}
&Q_{\mathrm{29DSpress}} - Q_{\mathrm{30DSdew}} - Q_{\mathrm{33DSrej}} &&= \mathrm{e}_{\mathrm{2a}} \\
&Q_{\mathrm{10I4}} + Q_{11Co} + Q_{15PSth} + Q_{18WASth} - Q_{\mathrm{28storDS}} - Q_{\mathrm{29DSpress}} &&= \mathrm{e}_{\mathrm{2b}} \\
&Q_{\mathrm{10I4}} + Q_{11Co} + Q_{15PSth} + Q_{18WASth} - Q_{\mathrm{28storDS}} - Q_{\mathrm{30DSdew}} - Q_{\mathrm{33DSrej}} &&= \mathrm{e}_{\mathrm{2c}} \qquad \text{(2a-d)} \\
&COD_{25gasAD}^{mf} - COD_{36el.en}^{mf} &&= \mathrm{e}_{\mathrm{2d}}
\end{aligned}$$

It can be verified from Figure 2 that for the linear flow balance equations (2a-c) the corresponding bilinear balance equations describing mass flow cannot be set up due only to missing values for the co-substrate and the reject from dewatering.

Two valid redundancy equations can be found directly through the elimination of unmeasured flows from within the same node. They describe balances of the high load activated sludge tank (AST1) and its combination with the splitter (eq. 3a-b). In these two cases there are 2 flows unmeasured but 2 concentrations (COD and P) fully measured in all streams giving 3 equations with 2 unknowns which combine to 1 redundancy equation. Owing to the splitter, equation 5b contains the term *(COD₇-COD₉)·Q₇* that effectively yields zero.

$$\begin{aligned}
&(COD_{21}-COD_{8})\cdot(Q_3\cdot P_3-Q_{16}\cdot P_{16}+Q_{17}\cdot P_{17}+P_{23}^{mf})\\
&-\left[(COD_{21}-COD_{3})\cdot Q_3-(COD_{17}-COD_{21})\cdot Q_{17}+(COD_{16}-COD_{21})\cdot Q_{16}+COD_{19}^{mf}-COD_{23}^{mf}\right]\cdot P_8\\
&-\left[(COD_{3}-COD_{8})\cdot Q_3+(COD_{17}-COD_{8})\cdot Q_{17}-(COD_{16}-COD_{8})\cdot Q_{16}-COD_{19}^{mf}+COD_{23}^{mf}\right]\cdot P_{21} &&=e_{3a}\\
&(COD_{21}-COD_{9})\cdot(Q_3\cdot P_3+Q_7\cdot P_7-Q_{16}\cdot P_{16}+Q_{17}\cdot P_{17}+P_{23}^{mf})\\
&-\left[(COD_{21}-COD_{7})\cdot Q_7+(COD_{21}-COD_{3})\cdot Q_3-(COD_{17}-COD_{21})\cdot Q_{17}+(COD_{16}-COD_{21})\cdot Q_{16}+COD_{19}^{mf}-COD_{23}^{mf}\right]\cdot P_9\\
&-\left[(COD_{7}-COD_{9})\cdot Q_7+(COD_{3}-COD_{9})\cdot Q_3+(COD_{17}-COD_{9})\cdot Q_{17}-(COD_{16}-COD_{9})\cdot Q_{16}-COD_{19}^{mf}+COD_{23}^{mf}\right]\cdot P_{21} &&=e_{3b}
\end{aligned}$$

(3a-b)Equations 4a and 4b are again mass balances around the storage tank, but the missing flow rate from the anaerobic digester, *Q₂₆*, is calculated from the flow balances around other neighboring subsystems. In the same way, equation 4c balances flows around the system PC-merge-AST1-AST2 where flow *Q₈* is missing and can be calculated from the combination of flow and COD balances around AST1.

$$\begin{aligned}
&Q_{28}\cdot COD_{28}+Q_{29}\cdot COD_{29}-(Q_{28}+Q_{30}+Q_{33})\cdot COD_{26} &&=e_{4a}\\
&Q_{28}\cdot COD_{28}+Q_{29}\cdot COD_{29}-(Q_{10}+Q_{11}+Q_{15}+Q_{18})\cdot COD_{26} &&=e_{4b}\\
&Q_1+Q_2+Q_3+Q_5+Q_{35}+Q_6-Q_{14}-Q_{16}-Q_{22}-Q_{34}\\
&-\left[(COD_{21}-COD_{3})\cdot Q_3-(COD_{17}-COD_{21})\cdot Q_{17}+(COD_{16}-COD_{21})\cdot Q_{16}+COD_{19}^{mf}-COD_{23}^{mf}\right]\div(COD_{21}-COD_{8}) &&=e_{4c}
\end{aligned}$$

(4a-c)Equations 5a-b show examples, where two observable variables had to be replaced in order to set up redundancy equations:

$$\begin{aligned}
&(COD_{13}-COD_{5})\cdot Q_5+(COD_{13}-COD_{35})\cdot Q_{35}+(COD_{13}-COD_{2})\cdot Q_2\\
&-(COD_{1}-COD_{13})\cdot Q_1-COD_{13}\cdot(Q_1-Q_{14}+Q_2+Q_{35}+Q_5)+COD_1\cdot Q_1\\
&-COD_{13}\cdot Q_{14}+COD_2\cdot Q_2+COD_{35}\cdot Q_{35}+COD_5\cdot Q_5 &&=e_{5a}
\end{aligned}$$

(5a-b)

$$\begin{aligned}
&Q_3-Q_{16}+Q_{17}\\
&-\left[(P_{21}-P_3)\cdot Q_3-(P_{16}-P_{21})\cdot Q_{17}+(P_{16}-P_{21})\cdot Q_{16}-P_{23}^{mf}\right]\div(P_{21}-P_8)\\
&-\left[\begin{array}{l}(COD_7-COD_9)\cdot Q_7+(COD_3-COD_9)\cdot Q_3+(COD_{17}-COD_9)\cdot Q_{17}\\ -(COD_{16}-COD_9)\cdot Q_{16}-COD_{19}^{mf}+COD_{23}^{mf}\end{array}\right]\div(COD_{21}-COD_9) &&=e_{5b}
\end{aligned}$$

Obviously, the number and complexity of equations increases with the number of replaced observable variables allowed per equation. At the same time, practical usability is likely to deteriorate. While only 10 redundant variables can be put in four balance equations when solutions of observable variables where not allowed, this number increases to 31 with those two additional equations where observable variables are eliminated within the same node. With one observable variable calculated from another node, there are 21 distinct equations expressing redundancy of 57 variables. Equations including solutions for two observable variables yield 199 distinct equations for 74 redundant variables.

Sensor placement

When the incidence matrix expansion into *M3* is scanned to improve overall redundancy, it turns out that the additional measurement of the reject flow from primary sludge thickening (Q_{31}) and sampling of the scum (COD_{34}, P_{34}, N_{34}) would have the greatest effect on overall structural redundancy. While before the introduction of these additional measurements there were 81 variables redundant out of 96 measured, this ratio increases to 96 redundant variables out of 100 measured. For this structural analysis, reasonability of the suggested additional measurements was not regarded.

4 Discussion

The computational determination of bilinear redundancy equations has been shown for the case of structural redundancy. It allows to set up suitable mass balances for data validation procedures that require individual balance equations such as CUSUM charts. This is of particular interest when the dynamic nature of wastewater treatment is considered where reliable gross error detection is still a challenge. The computational approach also provides equations that might not be obviously visible to the expert's eye, particularly for large and complex wastewater treatment systems. This way, substantially more process variables become accessible to the data validation procedure. In the example only 10 out of 81 redundant variables could be expressed in simple balance equations whereas 74 redundant variables became accessible when the calculation of 2 observable variables per equation was allowed. Additionally, the approach of incidence matrix expansion allows for a simple investigation about the placement of additional measurements to provide redundancy of chosen variables.

The expansion of the incidence matrix M is possible even for large and complex wastewater treatment plants. However, the number of subsystems even in those wastewater treatment plants is rather limited compared to some chemical industries. Due to the exponentially growing computational effort, the approach of incidence matrix expansion might not be feasible in other fields.

For practical applicability of the method further research is necessary. As most of the resulting redundancy equations (such as eq. 5b) are very complex and include many variables, some criteria will be needed to select equations that are actually useful for data validation. A sensitivity analysis could reveal which variables in such equations can be validated and for which variables in such equations no conclusions can be drawn. Much alike, many redundancy equations cannot be set up because they include variables that are in fact negligible. In the example, neglecting Q_8 (the flow of the sampling side stream of the industrial influent *7_I3*) would allow the setup of a flow balance around the primary clarifier and the activated sludge tanks AST1 and AST2. However, flow Q_8

might not be negligible with respect to the merging of the various reject waters. These questions address *practical redundancy* in addition to *structural redundancy* of the variables. An extension of the above described algorithm should be possible to find *approximate redundancy equations.* This would be based on an estimation of all variables, where possible by the classical methods of data reconciliation. Following an analysis of sensitivity, for each equation in *M2* the negligible terms would be eliminated before solutions for observable variables and redundancy equations are calculated. Investigations in this direction shall be the objective of a subsequent paper.

5 Conclusions

An algorithm is presented that allows the determination of all structurally possible redundancy equations for a given plant layout and classification of measured and unmeasured variables. Due to the separate treatment of flows and concentrations not only linear redundancy equations can be found. The algorithm is derived from data reconciliation methods which are applied extensively in the field of (chemical) process engineering but so far hardly present in wastewater treatment. Because of a possibly large number and high complexity of the resulting redundancy equations, the investigation of practical redundancy appears necessary. The underlying concept of incidence matrix expansion also allows a simple investigation on the effect of additional measurements.

It has been shown, that the setup of individual redundancy equations for data validation based on mass balancing can be fully computerized. This is an important step in the development of automated data validation in wastewater treatment systems.

Acknowledgements

The basic ideas of this work were developed during the first author's stay at model*EAU* in Québec, Canada, which co-funded the exchange. Peter Vanrolleghem holds the Canada Research Chair in water quality modeling. During the 1st IWA YWP Publication Workshop sponsored by and held at UTM Johor Baru, Malaysia, helpful remarks and intense discussions from and with Gustav Olsson and Helmut Kroiss contributed significantly to improve the paper.

References

Bagajewicz, M. J. (1998) Gross error modeling and detection in plant linear dynamic reconciliation. Computers & Chemical Engineering, 22, 1789–1809.

Bagajewicz, M. J. (2010) *Smart process plants software and hardware solutions for accurate data and profitable operations*, New York :, McGraw-Hill.

Barker, P. S. and Dold, P. L. (1995) Cod and Nitrogen Mass Balances in Activated-Sludge Systems. Water Research, 29(2), 633–643.

Crowe, C. M. (1996) Data reconciliation — Progress and challenges. Journal of Process Control, 6(2-3), 89–98.

Crowe, C. M. (1989) Observability and redundancy of process data for steady state reconciliation. Chemical Engineering Science, 44(12), 2909–2917.

Crowe, C. M. (1986) Reconciliation of process flow rates by matrix projection. Part II: The nonlinear case. AIChE Journal, 32(4), 616–623.

Crowe, C. M., Campos, Y. A. G., and Hrymak, A. (1983) Reconciliation of process flow rates by matrix projection. Part I: Linear case. AIChE Journal, 29(6), 881–888.

Decker, W., Greuel, G.-M., Pfister, G., and Schönemann, H. (2011) *Singular — A computer algebra system for polynomial computations. Version 3-1-5*, [online] http://www.singular.uni-kl.de/Manual/3-1-1/sing_1.htm

Hellinga, C. (1992). *Macrobal 2.02*. Delft University of Technology. [online] http://www.tnw.tudelft.nl/en/about-faculty/departments/biotechnology/data-software/macrobal/

Kelly, J. D. (2004) Formulating large-scale quantity-quality bilinear data reconciliation problems. Computers & Chemical Engineering, 28(3), 357–362.

Kelly, J. D. (1998) On finding the matrix projection in the data reconciliation solution. Computers & Chemical Engineering, 22(11), 1553–1557.

Kuehn, D. R. and Davidson, H. (1961) Computer Control II. Mathematics of Control. Chemical Engineering Progress, 57(6), 44–47.

Madron, F. and Veverka, V. (1992) Optimal selection of measuring points in complex plants by linear models. AIChE Journal, 38(2), 227–236.

Mah, R. S., Stanley, G. M., and Downing, D. M. (1976) Reconcillation and Rectification of Process Flow and Inventory Data. Industrial & Engineering Chemistry Process Design and Development, 15(1), 175–183.

Maxima (2013) *Maxima, a Computer Algebra System. Version 5.30.0*, [online] http://maxima.sourceforge.net/

Meijer, S. C., van der Spoel, H., Susanti, S., Heijnen, J. J., and van Loosdrecht, M. C. M. (2002) Error diagnostics and data reconciliation for activated sludge modelling using mass balances. Water Science and Technology, 45(6), 145–156.

Narasimhan, S. and Jordache, C. (2000) *Data reconciliation & gross error detection: an intelligent use of process data*, Gulf Professional Publishing.

Nowak, O., Franz, A., Svardal, K., Muller, V., and Kühn, V. (1999) Parameter estimation for activated sludge models with the help of mass balances. Water Science and Technology, 39(4), 113–120.

Nowak, O., Schweighofer, P., and Svardal, K. (1994) Nitrification Inhibition - A method for the estimation of actual maximum autotrophic growth rates in activated sludge szstems. Water Science and Technology, 30(6), 9–19.

Ponzoni, I., Sánchez, M. C., and Brignole, N. B. (1999) A New Structural Algorithm for Observability Classification. Industrial & Engineering Chemistry Research, 38(8), 3027–3035.

Puig, S., van Loosdrecht, M. C. M., Colprim, J., and Meijer, S. C. . (2008) Data evaluation of full-scale wastewater treatment plants by mass balance. Water Research, 42, 4645–4655.

Ragot, J., Maquin, D., Bloch, G., and Gomolka, W. (1990) Observability and variables classification in bilinear processes. Journal A, 17–23.

R Core Team (2013) *R: A Language and Environment for Statistical Computing. Version 2.15.3*, Vienna, Austria, R Foundation for Statistical Computing. [online] http://www.R-project.org/

Rieger, L., Takács, I., Villez, K., Siegrist, H., Lessard, P., Vanrolleghem, P. A., and Comeau, Y. (2010) Data Reconciliation for Wastewater Treatment Plant Simulation Studies—Planning for High-Quality Data and Typical Sources of Errors. Water Environment Research, 82(5), 426–433.

Romagnoli, J. A. and Sánchez, M. C. (2000) *Data processing and reconciliation for chemical process operations*, Academic Press.

Schraa, O. J. and Crowe, C. M. (1998) The numerical solution of bilinear data reconciliation problems using unconstrained optimization methods. Computers & Chemical Engineering, 22(9), 1215–1228.

Schraa, O. J., Tole, B., and Copp, J. B. (2006) Fault detection for control of wastewater treatment plants. Water Science & Technology, 53(4-5), 375.

Spindler, A. and Vanrolleghem, P. A. (2012) Dynamic mass balancing for wastewater treatment data quality control using CUSUM charts. Water science and technology: a journal of the International Association on Water Pollution Research, 65(12), 2148–2153.

Stein et al., W. A. (2012) *Sage Mathematics Software. Version 5.5*, The Sage Development Team. [online] http://www.sagemath.org

Václavek, V. (1969) Studies on system engineering—III optimal choice of the balance measurements in complicated chemical engineering systems. Chemical Engineering Science, 24(6), 947–955.

Václavek, V. and Louĉka, M. (1976) Selection of measurements necessary to achieve multicomponent mass balances in chemical plant. Chemical Engineering Science, 31(12), 1199–1205.

Van der Heijden, R. T. J. ., Heijnen, J. J., Hellinga, C., Romein, B., and Luyben, K. C. A. . (1994) Linear constraint relations in biochemical reaction systems: I. Classification of the calculability and the balanceability of conversion rates. Biotechnology and Bioengineering, 43, 3–10.

Villez, K., Corominas, L., Vanrolleghem, P. A. (2013a). Structural observability and redundancy classification for sensor networks in wastewater systems. Proceedings of the 11th IWA conference on instrumentation control and automation (ICA2013), Narbonne, FR, Sept. 18-20, 2013, Appeared on USB-stick (IWA-12741).

Villez, K., Corominas, L., Vanrolleghem, P. A. (2013b). Sensor fault detection and diagnosis based on bilinear mass balances in wastewater treatment systems. Proceedings of the 11th IWA conference on instrumentation control and automation (ICA2013), Narbonne, FR, Sept. 18-20, 2013, Appeared on USB-stick (IWA-12744).