

Experimental design for evaluating WWTP data by linear mass balances

Le, Quan H.; Verheijen, Peter J.T.; van Loosdrecht, Mark C.M.; Volcke, Eveline I.P.

DOI

[10.1016/j.watres.2018.05.026](https://doi.org/10.1016/j.watres.2018.05.026)

Publication date

2018

Document Version

Accepted author manuscript

Published in

Water Research

Citation (APA)

Le, Q. H., Verheijen, P. J. T., van Loosdrecht, M. C. M., & Volcke, E. I. P. (2018). Experimental design for evaluating WWTP data by linear mass balances. *Water Research*, 142, 415-425. <https://doi.org/10.1016/j.watres.2018.05.026>

Important note

To cite this publication, please use the final published version (if applicable). Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

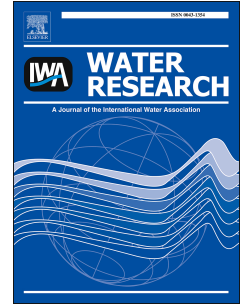
Takedown policy

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

Accepted Manuscript

Experimental design for evaluating WWTP data by linear mass balances

Quan H. Le, Peter J.T. Verheijen, Mark C.M. van Loosdrecht, Eveline I.P. Volcke



PII: S0043-1354(18)30394-4

DOI: [10.1016/j.watres.2018.05.026](https://doi.org/10.1016/j.watres.2018.05.026)

Reference: WR 13791

To appear in: *Water Research*

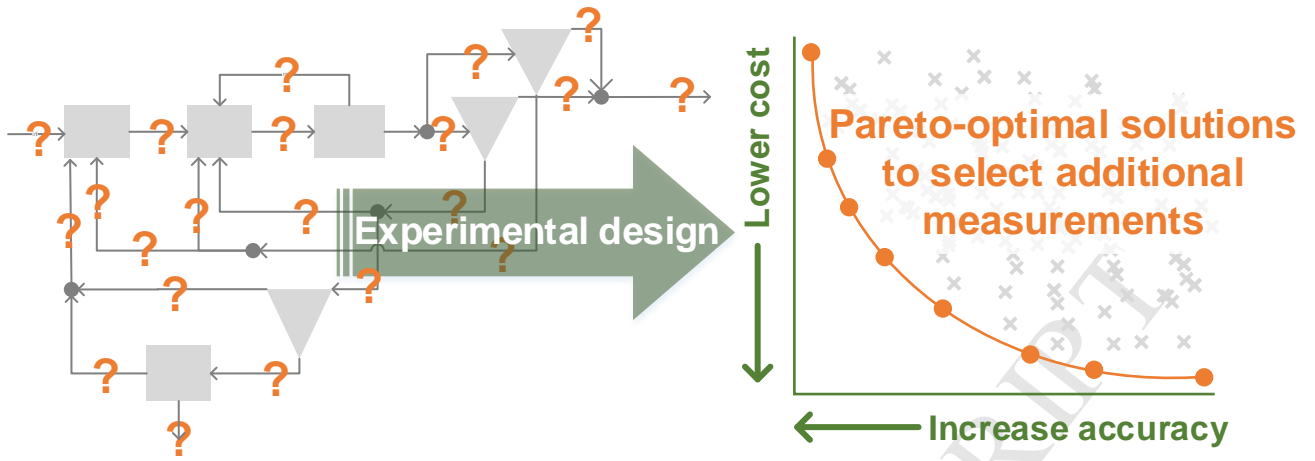
Received Date: 2 February 2018

Revised Date: 23 April 2018

Accepted Date: 14 May 2018

Please cite this article as: Le, Q.H., Verheijen, P.J.T., van Loosdrecht, M.C.M., Volcke, E.I.P., Experimental design for evaluating WWTP data by linear mass balances, *Water Research* (2018), doi: 10.1016/j.watres.2018.05.026.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Experimental design for evaluating WWTP data by linear mass balances

Quan H. Le^a, Peter J.T. Verheijen^b, Mark C.M. van Loosdrecht^b, Eveline I.P. Volcke^{a,*}

^a Department of Green Chemistry and Technology, Ghent University, Belgium

^b Department of Biotechnology, Delft University of Technology, The Netherlands

* Corresponding author: Eveline Volcke (Eveline.Volcke@UGent.be)

Abstract

A stepwise experimental design procedure to obtain reliable data from wastewater treatment plants (WWTPs) was developed. The proposed procedure aims at determining sets of additional measurements (besides available ones) that guarantee the identifiability of key process variables, which means that their value can be calculated from other, measured variables, based on available constraints in the form of linear mass balances. Among all solutions, i.e. all possible sets of additional measurements allowing the identifiability of all key process variables, the optimal solutions were found taking into account two objectives, namely the accuracy of the identified key variables and the cost of additional measurements. The results of this multi-objective optimization problem were represented in a Pareto-optimal front.

The presented procedure was applied to a full-scale WWTP. Detailed analysis of the relation between measurements allowed the determination of groups of overlapping mass balances. Adding measured variables could only serve in identifying key variables that appear in the same group of mass balances. Besides, the application of the experimental design procedure to these individual groups significantly reduced the computational effort in evaluating available measurements and planning additional monitoring campaigns. The proposed procedure is straightforward and can be applied to other WWTPs with or without prior data collection.

Keywords: experimental design; data validation; mass balances; data reconciliation; wastewater treatment plant;

27 **1 Introduction**

28 The importance of reliable data for wastewater treatment plant (WWTP) design, process optimization,
29 operator training, developing control strategies, benchmarking and simulation is commonly advocated
30 (Meijer et al., 2015, 2002; Puig et al., 2008; Rieger et al., 2010; Spindler, 2014; Villez et al., 2013a). Typical
31 data in this respect concern flows and concentrations of components. Depending on the objectives, available
32 historical data are complemented with additional data obtained through one or more intensive monitoring
33 campaigns using classic sampling followed by laboratory analyses and/or online sensors.

34 Data reconciliation is a proven technique to evaluate the consistency of collected data (Crowe, 1996; Ozyurt
35 and Pike, 2004). It involves a procedure of optimally adjusting estimates for variables such that these
36 estimates satisfy the conservation laws and other constraints (Crowe, 1996) and are therefore more accurate
37 than the original values. Data reconciliation is often accompanied by statistical tests for gross error detection
38 (measurement validation), which verify whether the deviation between each estimate and its measurement is
39 acceptable compared to the measurement error.

40 Even though data reconciliation has been widely applied in (bio)chemical engineering for decades (Madron
41 et al., 1977; Madron and Veverka, 1992; van der Heijden et al., 1994b), this concept so far has received
42 relatively little attention in wastewater treatment process engineering. Some studies applied the concept of
43 redundancy analysis and variable classification, which are closely related to the principles and objectives of
44 data reconciliation, for sensor fault detection (Villez et al., 2016, 2015, 2013b) or for describing redundancy
45 in the data set (Spindler, 2014). In other studies, data reconciliation was directly applied for the validation of
46 a WWTP process data set for modelling, process optimization or plant performance evaluation (Behnami et
47 al., 2016; Meijer et al., 2015, 2002; Puig et al., 2008). The effects of erroneous data on modelling errors was
48 investigated by Lee et al. (2015), applying gross error detection. Rieger et al. (2010) put the concept of data
49 validation by mass balancing in a general data collection framework, stressing the importance of
50 measurement planning to guarantee a successful subsequent data validation for WWTP. Besides full-scale
51 processes, data reconciliation was also applied to long-term data of a lab scale wastewater treatment reactor
52 to identify different anabolic reactions pathways (Lotti et al., 2014).

53 The abovementioned studies explicitly or implicitly pointed out that it is vital for a measurement plan to satisfy
54 the redundancy and steady-state conditions, as important prerequisites for successful data reconciliation.
55 While obtaining data fulfilling the steady-state condition was discussed in detail by Meijer et al. (2015), this
56 work focuses on the redundancy requirement.

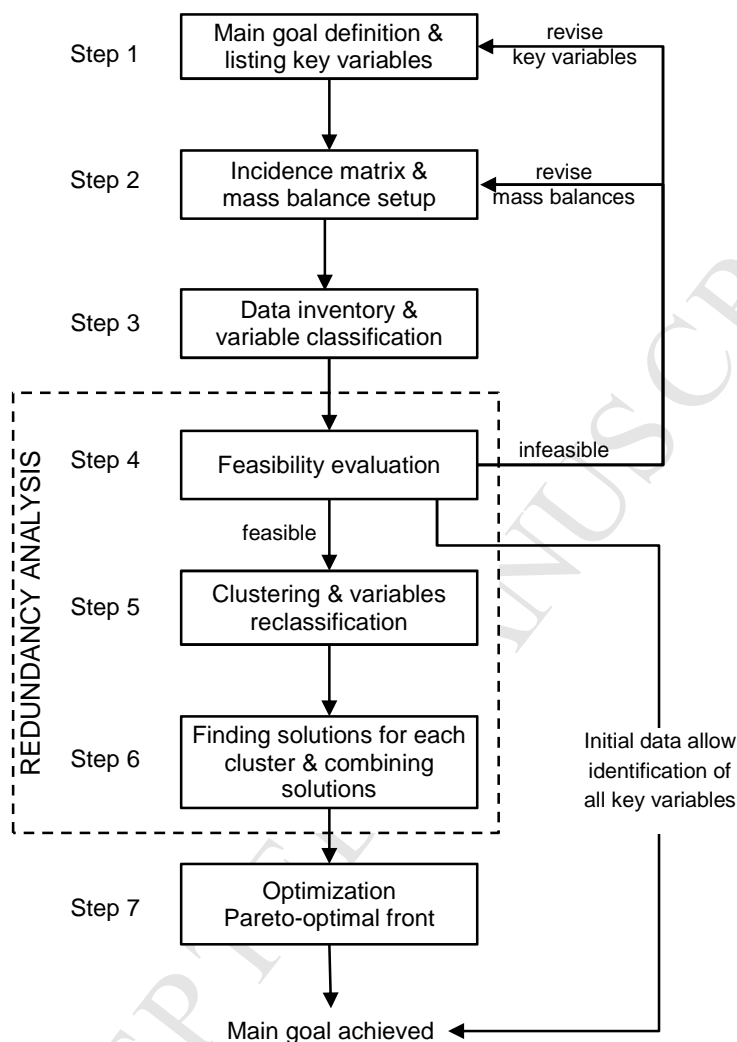
57 Redundancy of variables means that their measured values can also be calculated from other (measured)
58 variables. However, for many WWTPs, there are often not sufficient initially measured data available and
59 additional measurements typically need to be carried out (in a monitoring campaign) to ensure the required
60 degree of redundancy for data reconciliation. In this respect, “overlapping mass balances” and “closed mass
61 balances” are typically aimed at. The term “overlapping mass balances” refers to mass balances over single
62 or combined subsystems that share one or more mass flows or have at least one variable in common. The
63 term “closed mass balances” refers to mass balances in which all variables are measured and which can
64 typically be set up for conserved quantities such as total mass flows or total phosphorus mass. This practice
65 of adding overlapping and closed mass balances increases the overall system redundancy and has therefore
66 been commonly applied for data reconciliation in wastewater process engineering (Lee et al., 2015; Meijer et
67 al., 2015, 2002; Puig et al., 2008).

68 However, increasing the *overall* system redundancy does not guarantee the possible identification of
69 specified key variables (van der Heijden et al., 1994a). The approach of Meijer et al. (2015, 2002) and Puig
70 et al. (2008), aiming at increasing redundancy by adding measurements to set up overlapping and closed
71 mass balances, therefore, involved the risk of adding trivial mass balances and associated unnecessary
72 additional measurements. For WWTP data reconciliation, the question remains in what manner and to which
73 extent additional measurements, entailing additional overlapping and closed mass balances, effectively lead
74 to reliable and improved estimates of the key variables under concern.

75 This work provides a practical stepwise procedure to determine sets of additional measurements that
76 guarantee the possible identification of key process variables, which means that their value can be
77 calculated from other, measured variables. More specifically, these sets of additional measurements satisfy
78 the required degree of redundancy for data reconciliation considering constraints in the form of linear mass
79 balances. The focus of this work is on the experimental design, i.e. the determination of additional
80 measurements allowing the identification of key variables. The actual application of data reconciliation to
81 obtain reliable and improved estimates for key variables is the topic of a follow-up paper. The redundancy of
82 measurements was analysed to gain insight in the way measured variables are related through linear mass
83 balances. Particular attention was paid to the contribution of additional overlapping and closed mass
84 balances. Through this comprehensive redundancy analysis, shortcomings of previous studies in selecting
85 meaningful additional measurements were overcome. Moreover, the accuracy of the reconciled results and
86 the cost of additional measurements were considered in finding optimal sets of additional measurements.
87 The procedure was demonstrated for a full-scale WWTP.

88 **2 Experimental design procedure**

89 An experimental design procedure for practical application to wastewater treatment processes was derived
 90 (Figure 1).



91

92 **Figure 1.** Experimental design procedure for the selection of sets of additionally measured variables that
 93 allow the identification of key variables.

94 The key variables are defined first (Step 1), followed by the set-up of an incidence matrix and mass balances
 95 based on the process flow diagram (Step 2) and the inventory of available data (Step 3). Even though these
 96 3 steps have been addressed previously in an intuitive approach for data collection (Meijer et al., 2015), they
 97 were now included in a more formal experimental design procedure, focusing on key variables, simplifying
 98 the mass balance set-up and reducing associated efforts. Moreover, a comprehensive redundancy analysis
 99 has been added in this study (Steps 4-6), to overcome the shortcoming of previous studies. It is now

100 checked up-front that the list of key variables and/or the set of set up mass balances are relevant in the
101 sense that key variables are identifiable (Step 4). Mass balances and their corresponding variables are
102 clustered (Step 5), which greatly improves the efficiency of finding all solutions, i.e. sets of additionally
103 measured variables that satisfy the defined main goal (Step 6). Finally, a procedure to select the optimal
104 solution in terms of additional measurement costs and accuracy of identified key variables has now been
105 provided as well (Step 7). Step 4 to Step 7 were implemented in MATLAB 2014b (MathWorks®). More
106 details on the individual steps are given below. Details on the applied procedures and on the theoretical
107 background are provided in Supplementary Material A and B, respectively.

108 **Step 1. Main goal definition - listing key variables**

109 Data reconciliation can be applied to identify key process variables and at the same time detect possible
110 gross errors. Key variables may be measured or not; their identification means that improved estimates of
111 their values are obtained. These new estimates meet all the constraints (i.e., fit all mass balances) and are
112 therefore considered more reliable and accurate (have a smaller standard deviation or error) than the original
113 values. In case a key variable is measured, the new estimate is considered improved compared to the
114 original measurements. In case the key variable is not measured, the new estimate is considered improved
115 compared to the value directly calculated from original measurements (using the available set of mass
116 balances).

117 In this step, all key process variables are listed. Typical examples of key process variables in a WWTP that
118 need to be known with high accuracy concern influent and effluent mass flow rates of the activated sludge
119 process (biological reactor) as well as the waste activated sludge mass flow rate. The oxygen requirements
120 for chemical oxygen demand (COD) and nitrogen removal are also important process variables and
121 therefore typically need to be calculated – they are typical unmeasured key variables.

122 The constraints which the new estimates of key variables need to meet, are in the form of linear mass
123 balances, consisting of mass flow terms. For this reason, key variables, denoted as K^* are expressed in
124 terms of total mass flows and mass flows of individual components (as indicated by the superscript *). It is
125 important to note that the mass flow of a certain component at a certain place is only considered measured if
126 both the corresponding flow rate and component concentration are measured.

127 The experimental design procedure aims at determining one or more sets of additional measurements that
128 guarantee the identification of all key variables, while minimizing the cost of additional measurements and
129 maximizing the accuracy of the identified key variables.

130 Step 2. Incidence matrix and mass balance setup

131 The process flow diagram of WWTP is translated into a so-called incidence matrix, which is a mathematical
132 description of the flow network. The columns of the incidence matrix represent process streams and the rows
133 represent individual or combined unit processes. The elements of this matrix are:

- 134 • 1, if stream enters a unit process,
- 135 • -1, if stream leaves a unit process,
- 136 • 0, if stream is not incident with a unit process.

137 To visualize the spatial distribution of the interrelated subsystems, it is advised to number and arrange the
138 flows and unit processes in the matrix following the water line, starting from the influent and primary tanks
139 and ending towards the dewatered sludge. In this way, the matrix diagonal represents the water flow through
140 the WWTP (Meijer et al., 2015).

141 Following the setup of the incidence matrix, linear mass balances of total mass flows ($\rho \times Q$, or Q when
142 assuming the same density ρ for all streams in that mass balance) and individual mass flows, e.g. total
143 phosphorus (mTP), COD ($mCOD$) and total nitrogen (mTN), are set up for all subsystems considered. These
144 subsystems could either be individual or combined unit processes. The resulting mass balances need to
145 contain all key variables listed in Step 1. More detailed practical guidance on the selection of conservative
146 quantities is provided by Meijer et al. (2015).

147 Step 3. Data inventory and variable classification

148 Once the mass balances are set up, an inventory is made of initially measured and initially unmeasured
149 process variables that appear in the mass balances. The values of measured variables are obtained from
150 routine lab analyses or through online monitoring. These are typically flows (Q) and concentrations of
151 individual components such as COD, total nitrogen (TN) and total phosphorus (TP).

152 For optimization purposes, the expected measurement costs of all unmeasured variables (in the form of flow
153 and concentration) and the measurement errors (standard deviation of the mean of the measurements in the
154 form of mass flow) of all variables are also inventoried. In case the measurement error of a variable is not
155 known or cannot be realistically estimated from expert knowledge, one could use a small error compared to
156 those of other variables, essentially assuming a relatively good measurement, which still allows to track the

157 error propagation to the identifiable variables. Note that the relative magnitude of the measurement errors is
158 of importance, rather than their exact values.

159 Let M be the set of initially measured variables and U be the set of initially unmeasured variables resulting
160 from the data inventory. Part of the initially unmeasured variables are unmeasurable; they constitute the
161 subset X of U . The remaining initially unmeasured variables could potentially be measured and constitute a
162 complementary subset P_a of U ($P_a = U - X$). The aforementioned variables are typically expressed in terms of
163 (volumetric) flows and concentrations.

164 **Step 4. Feasibility evaluation**

165 The feasibility of satisfying the main goal, i.e. of identifying all listed key variables, is evaluated for two
166 extreme cases of measurement availability:

167 (i) All potential additionally measured variables P_a are measured additionally. It is thus checked whether
168 all key variables are identifiable for the largest set of potential (additional) measurements and for the
169 given set of mass balances. In case one or more key variables are not identifiable, it is recommended
170 to first review the set of mass balances. The mass balances need to contain all key variables.
171 Besides, non-identifiability could also result from mistakenly neglected flows or because of an
172 oversimplified plant layout. If revising mass balances does not result in the identifiability of all key
173 variables, there is insufficient redundancy in the system and it is advised to remove unidentifiable key
174 variables, i.e. to return to Step 1. Once all key variables are identifiable considering the largest
175 possible set of additional measurements, possibly after revising mass balances and/or key variables,
176 the second extreme case of measurement availability is evaluated.

177 (ii) Only initially measured data are available. If all key variables are identifiable from the set of initial
178 measurements, the main goal is fulfilled *a priori* and there is no need for additional measurements. If
179 this is not the case, the procedure proceeds to Step 5 and Step 6 to determine sets of additional
180 measurements resulting in the identifiability of all key variables. The existence of such sets of
181 additional measurements is guaranteed by (i), which ensures the best possible definition of mass
182 balances and removed key variables that are not identifiable *a priori*.

183 The identifiability of key variables is checked through redundancy analysis, based on the procedures of van
184 der Heijden et al. (1994) and Klamt et al. (2002), as detailed in the Supplementary Material (section B1 for
185 the theoretical background and section A1 for the practical implementation).

186 **Step 5. Clustering and variables reclassification**

187 Once the identification of key variables has been evaluated feasible, it will be investigated which set(s) of
188 additional measurements are required to this end. This procedure is simplified by clustering the mass
189 balances in groups of overlapping mass balances, i.e. mass balances that have at least one variable (total or
190 individual mass flow rate) in common.

191 Clustering is based on redundancy analysis, involving the set-up of redundancy equations (see
192 Supplementary Material B1). The redundancy equations are obtained from the original set of mass balances
193 by eliminating all unmeasured variables, such that only measured variables remain. Variables that appear in
194 a single redundancy equation will be used in data reconciliation to identify each other. When redundancy
195 equations are interrelated by one or more variables, they will also be used to identify the variables in the
196 related equations. The identifiability of variables in a group of interdependent redundancy equations is
197 independent from the identifiability and measurement availability of variables in the other groups.

198 In order to cluster the mass balances in groups of overlapping mass balances, the redundancy equations are
199 derived assuming all variables are measured. In this way, the maximum number of relations between
200 (measured) variables can be identified, allowing subsequent variable reclassification clustering in groups of
201 interdependent variables. First, the redundancy equations are clustered in groups of redundancy equations
202 that are related by one or several variables. Second, groups of variables that belong to the corresponding
203 groups of redundancy equations are formed (variable reclassification). Finally, based on groups of variables,
204 the mass balances are clustered in group of overlapping mass balances that only contain variables of the
205 corresponding groups.

206 After clustering the mass balances in groups of overlapping mass balances, variable classification was
207 retaken for each group. Each group has its own measured variables (M), unmeasured variables (U),
208 potential additionally measured variables ($P_a = U - X$), unmeasurable key variables (X), and key variables
209 (K^*) that contribute to mass flow terms in the overlapping mass balances of that group. It is important to
210 realize that flow variables (Q) are implicitly taken up in the individual mass flows (mTP , $mCOD$, mTM). For
211 this reason, concentration variables (TP , COD or TM) always appear together with the flow rate (Q) of the
212 corresponding stream while clustering. It is thus possible that a single (flow) variable appears in multiple
213 groups.

214 The routine of clustering and variable reclassification is provided in Supplementary Material A2.

215 **Step 6. Finding solutions**

216 Clustering mass balances into groups of overlapping mass balances significantly simplifies the procedure of
 217 finding solutions. Indeed, the identifiability of key variables in one group of overlapping mass balances is
 218 independent from the measurement availability of variables in other groups; the measurement availability of
 219 a variable in a group of overlapping mass balances only helps identifying other variables in that group. The
 220 solutions for each group of overlapping mass balances can thus be derived separately and then combined.

221 Solutions are found by checking for all potential sets of additional measurements (per group) whether they
 222 guarantee the key variables (of that group) to be identifiable. The identifiability of key variables is checked
 223 through redundancy analysis, based on the procedures of van der Heijden et al. (1994) and Klamt et al.
 224 (2002), analogously as in Step 4 (see Supplementary Material B1). Since the key variables of all groups
 225 need to be identified simultaneously, the overall solutions are derived by combining the solutions for the
 226 individual groups of overlapping mass balances, while discarding duplicates. Step 6 is detailed in
 227 Supplementary Material A3.

228 **Step 7. Optimization**

229 In Step 7, the costs and accuracy are calculated for all solutions. Each set of additional measurements that
 230 guarantees all key variables K^* to be identifiable, is referred to as a solution and is characterized by a $1 \times$
 231 n_p row vector $A = (a_1 \dots a_{n_p})$ consisting of binary decision variables a_j that indicate whether the
 232 corresponding potential additionally measured variables in P_a were selected to measure additionally ($a_j = 1$)
 233 or not ($a_j = 0$).

234 For every solution, the corresponding cost of additional measurements is calculated as the sum of the
 235 individual costs w_{aj} of additional measurements a_j , similar to the approach of Villez et al. (2016):

$$236 \quad f_c(A) = \sum_{j=1}^{n_p} w_{aj} \cdot a_j = W_a \cdot A' \quad (\text{Eq.1})$$

237 $W_a = (w_1 \dots w_{n_p})$ is a $1 \times n_p$ weighing vector, in which each element is w_{aj} .

238 The average variance of new estimates of key process variables (inversely related to accuracy), is calculated
 239 relative to the variance for the so-called reference solution, according to Eq. 2, and is termed $f_v(A)$. The
 240 reference solution, expressed as a $1 \times n_p$ vector $A_r = (1)$, is the solution obtained when all possible
 241 additional measurements P_a are measured additionally.

$$242 \quad f_V(A) = \frac{1}{n_k} \sum_{i=1}^{n_k} v_i / v_i^r = \frac{1}{n_k} \sum_{i=1}^{n_k} w_{vi} \cdot v_i = \frac{1}{n_k} W_v \cdot V' \quad (\text{Eq.2})$$

243 $V = (v_1 \dots v_{n_k})$ denotes a $1 \times n_k$ vector of variances of new estimates ($v_i \geq 0$) of the key variables (hereafter
 244 referred to as variance of key variables) when the solution A is implemented. The calculation of V is detailed
 245 in the Supplementary Material (B2 for theoretical background and A4 for practical implementation). $W_v =$
 246 $(w_{v1} \dots w_{v_{n_k}})$ is a $1 \times n_k$ vector of non-negative weights, $w_{vi} = 1/v_i^r$, in which v_i^r represents the variance of the
 247 key variables i when the reference solution is implemented.

248 Adding measurements to an existing set of measurements results in a smaller variance of new estimates
 249 obtained through data reconciliation (van der Heijden et al., 1994). Therefore, the reference solution A_r is a
 250 best known solution, which results in the smallest variance v_i^r (highest accuracy) of new estimates of the
 251 key variables ($\forall i \in \{1, 2, \dots, n_k\}: v_i^r \leq v_i$). The objective function $f_V(A)$ is a variation on the V-optimality choice
 252 in the experimental design theory (Pukelsheim, 2006). Essentially, the use of relative variances of a solution
 253 to a best known solution is a relevant choice to circumvent the problems due to the different units in which
 254 different key variables are expressed. The division by number of key variables (n_k) makes this objective
 255 such that in the best case the objective function $f_V(A)$ equals unity.

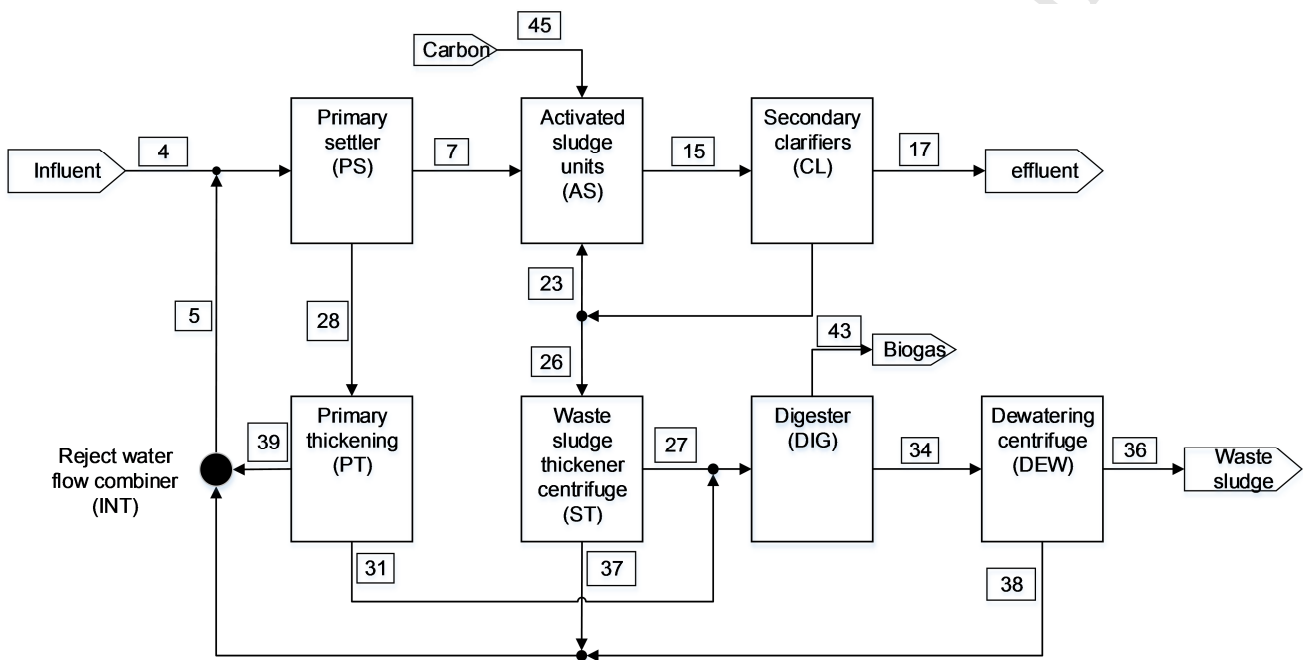
256 Finding an optimal solution is a multi-objective optimization problem consisting of finding the solution that
 257 minimizes both $f_C(A)$ and $f_V(A)$. In this study, the Pareto-optimal solutions were determined, for which a
 258 lower cost can only be obtained at the expense of a lower accuracy and vice versa. The implementation of
 259 this step is detailed in Supplementary Material A4.

260

261 **3 Application to a full-scale WWTP**262 **WWTP under study**

263 The proposed experimental design procedure was applied to WWTP Houtrust, The Hague, The Netherlands.
 264 Figure 2 displays a simplified configuration of this plant including all important streams; comprising a “three
 265 stage Phoredox process” or A2/O design. The full configuration of this plant is given in Supplementary
 266 Material C1; a more extensive plant description can be found in Meijer et al. (2015).

267



268

269

Figure 2. Simplified process flow diagram of WWTP Houtrust.

270

271 **Step 1. Main goal definition - listing key variables**

272 The experimental design procedure aims at determining one or more sets of additional measurements that
 273 guarantee the identifiability of all key variables, while minimizing the cost of additional measurements and
 274 maximizing the accuracy of the identified key variables. More specifically, key variables in the form of total
 275 mass flow and individual mass flows (COD, total nitrogen and total phosphorus) of the following streams had
 276 to be identifiable:

- 277 – Settled influent, i.e. influent of the activated sludge process (stream 7),
- 278 – WWTP influent (stream 4) and effluent (stream 17),
- 279 – Waste activated sludge (stream 26),

- 280 – WWTP waste sludge (stream 36),
 281 – Reject water (stream 5),
 282 – Biogas (stream 43, in this case only the mass flow of COD had to be identified).

283 Besides, the following unmeasurable key variables related to process performance had to be identifiable:

- 284 – Required oxygen for the oxidation of COD (OC_{cod} , $kg.day^{-1}$),
 285 – Amount of denitrified nitrogen ($DENI$, $kg.day^{-1}$),
 286 – Primary sludge flow and associated mass flows of COD, total nitrogen and total phosphorus (stream
 287 28).

288 Step 2. Incidence matrix and mass balance setup

289 The incidence matrix of WWTP Houtrust was set up based on the simplified process flow diagram (Figure 2),
 290 representing the WWTP layout by the minimum numbers of subsystems and streams but still contained all
 291 the variables of interest. The resulting matrix (Table 1) contained 8 rows (or subsystems) and 17 columns (or
 292 streams).

293 **Table 1.** Incidence matrix of the WWTP Houtrust. Ingoing and outgoing streams are denoted by '1' and '-1',
 294 respectively.

	Stream number in process flow diagram																
	4	5	7	15	17	23	26	27	28	31	34	36	37	38	39	43	45
Subsystems ↓	Total Influent	Reject water	Settled influent	Inflow secondary clarifiers (CL)	WWTP effluent	Return activated sludge	Waste activated sludge (WAS)	Thickened WAS to digester	Primary sludge	Thickened primary sludge	Digested sludge	Evacuated sludge	Centrate WAS thickening	Centrate dewatering	Overflow primary thickener	Biogas	External carbon source
Primary settler (PS)	1	1	-1						-1								
Activated sludge units (AS)			1	-1		1											1
Secondary clarifiers (CL)				1	-1	-1	-1										
Waste sludge thickener (ST)							1	-1					-1				
Primary sludge thickener (PT)									1	-1					-1		
Digester (DIG)								1		1	-1					-1	
Dewatering centrifuge (DEW)											1	-1		-1			
Reject water flow combiner (INT)		-1											1	1	1		

295

296 In simplifying the full process flow diagram (Supplementary Material C1), the activated sludge unit processes
297 (selector, predenitrification, anaerobic, anoxic, aeration and de-aeration tanks) were grouped into a
298 combined unit (AS, Figure 2), since they involve the unmeasured loss and supply of components through the
299 gas phase (N_2 , CO_2 and O_2), which do not need to (and cannot) be distinguished among them. Buffer units
300 were not explicitly considered, reasonably neglecting accumulation, separation and/or conversion of
301 components in these units. The small streams, such as clean water stream (stream 40), ferric chloride
302 sulfate ($FeClSO_4$) added for phosphorus removal (stream 44 to selector and 46 to digester) and grit removed
303 from the primary sludge (stream 41), were neglected. Bypass streams not used during normal operation
304 (Q18, Q19 and Q20) were not considered either.

305 Based on the incidence matrix, 32 linear mass balances were set up (Supplementary Material C2). Four
306 main types of mass balances were accounted for, describing the conservation of total flow (Q) and individual
307 mass flows of total phosphorus (mTP), chemical oxygen demand ($mCOD$) and total nitrogen (mTN) around
308 individual subsystems. The external carbon source (stream 45) and the biogas (stream 43) were reasonably
309 assumed to represent only COD; their total mass flow rates were neglected (in mass balances #2 and #6,
310 respectively). The oxygen required for COD removal (OC_{cod}) and the amount of denitrified nitrogen (DENI)
311 were taken into account in the COD balance of the activated sludge unit (mass balance #18). Note that, the
312 resulting set of mass balances contains all key variables, as required.

313 The question may arise whether adding mass balances containing off-gas measurements would lead to
314 additional solutions. This will be the case when the added mass balances contain key variables or stay in the
315 same group with other mass balances that contain the key variables. Sampling in the gas phase, however, is
316 typically difficult and associated with a large uncertainty (all the reactors are open and off-gas is dispersed
317 over a large surface area) and significant costs. For these reasons and to limit the complexity of the given
318 example, it was therefore decided not to consider mass balances containing off-gas measurements for
319 demonstrating the experimental design procedure in this study.

320 **Step 3. Data inventory and variable classification**

321 An overview of the initially measured and initially unmeasured data of WWTP Houtrust in terms of flows and
322 concentrations is given in Table 2.

323 **Table 2.** Data inventory in terms of flows (Q) and concentrations of total phosphorus (TP), chemical oxygen
 324 demand (COD) and total nitrogen (TN) for WWTP Houtrust. mTP , $mCOD$ and mTN present mass
 325 flow terms.

PFD ^(*)	Short Name	Q			TP			mTP	COD		$mCOD$	TN		mTN
		m	c	σ	m	c	σ	m	p	σ	m	p	σ	
4	WWTP influent	1	11	2,000	1	75	20	1	35	1500	1	75	40	
5	Reject water	1	11	100	1	75	35	1	35	900	1	75	60	
7	Settled influent		11	2,000		75	35		35	100		75	70	
15	Inflow secondary clarifiers	1	11	3,000		75	15	1	35	2000	1	75	35	
17	WWTP effluent		11	2,000	1	75	10	1	35	150	1	75	50	
23	Return activated sludge	1	11	2,000		75	15	1	35	1500		75	50	
26	Waste activated sludge	1	11	15		75	15	1	35	550		75	30	
27	Thickened WAS		11	50		75	65		35	1500		75	100	
28	Primary sludge		11	100		75	40		35	1500		75	45	
31	Thickened primary sludge		11	15		75	20	1	35	600		75	50	
34	Digested sludge		11	50		75	80	1	35	1200		75	35	
36	WWTP waste sludge	1	11	15	1	75	80	1	35	2000	1	75	40	
37	Centrate WAS thickening		11	50		75	35		35	600		75	25	
38	Centrate dewatering		11	50		75	55	1	35	400		75	30	
39	Overflow primary thickener	1	11	100		75	90		35	300		75	25	
43	Biogas	1	11	100	1	75	2	1	35	200	1	75	2	
45	External carbon source	1	11	2	1	75	2	1	35	150	1	75	2	

326 (*) Stream number in process flow diagram (Figure 2).
 327 m = indicating whether this flow/concentration variable is initially measured (1) or not (empty);
 328 σ = estimated error of the corresponding mass flow of the measurements (standard deviation of the mean, used in Step
 329 7);
 330 c = weighing factor represents the cost of a single measurement.
 331 Unit: flow and concentration = $m^3 \cdot day^{-1}$ and $g \cdot m^{-3}$; mass flow: $kg \cdot day^{-1}$

332 Errors of the measurement or the standard deviations of the mean measurements of all variables (in terms of
 333 total and individual mass flow) were estimated based on previous monitoring campaign (Meijer et al., 2015).
 334 From initial data, variables were classified into 4 groups: initial measured variables (M), initial unmeasured
 335 variables (U), unmeasurable variables (X) and potential additionally measured variables (P_a) (Table 3).

336 **Table 3.** Variable classification

Description	Corresponding variables
M Initially measured variables ($n_m = 34$)	Q4, Q5, Q15, Q23, Q26, Q36, Q39, Q43, Q45, TP4, TP5, TP17, TP36, TP43, TP45, COD4, COD5, COD15, COD17, COD23, COD26, COD31, COD34, COD36, COD38, COD43, COD45, TN4, TN5, TN15, TN17, TN36, TN43, TN45
U Initially unmeasured variables ($n_u = 34$)	Q7, Q17, Q27, Q28, Q31, Q34, Q37, Q38, TP7, TP15, TP23, TP26, TP27, TP28, TP31, TP34, TP37, TP38, TP39, COD7, COD27, COD28, COD37, COD39, TN7, TN23, TN26, TN27, TN28, TN31, TN34, TN37, TN38, TN39
X Unmeasurable variables ($n_x = 4$)	Q28 TP28 COD28 TN28
P_a Potential additionally measured variables ($n_p = 30$)	Q7, Q17, Q27, Q31, Q34, Q37, Q38, TP7, TP15, TP23, TP26, TP27, TP31, TP34, TP37, TP38, TP39, COD7, COD27, COD37, COD39, TN7, TN23, TN26, TN27, TN31, TN34, TN37, TN38, TN39
K^* Key variables ($n_k = 31$)	Q4, Q5, Q7, Q17, Q26, Q28, Q36 $mTP4, mTP5, mTP7, mTP17, mTP26, mTP28, mTP36$

337 $OCcod =$ required oxygen for COD removal; $DENI =$ denitrified nitrogen

338 While the classification of variables and the measurement cost quantification are rather straightforward, the
339 estimation of the measurement accuracy may be more difficult. Any expert knowledge and/or information
340 from previous monitoring campaigns is most useful in this respect. Keeping in mind that the relative
341 magnitude of the error terms is more important than their absolute values, it is interesting to note that, e.g.,
342 the error term on the volumetric mass flow of the influent (Q4) is of the same magnitude as the error term on
343 its COD mass flow (mCOD4), on its turn being one magnitudes higher then COD mass flow in the effluent
344 (mCOD17).

345 **Step 4. Feasibility evaluation**

346 The feasibility evaluation for the WWTP Houtrust confirmed that the identification of key variables is feasible,
347 at least in the case that all potential additionally measured variables (all variables in P_a) are measured
348 additionally. However, the initial data were not sufficient to identify all key variables. Therefore, the procedure
349 is continued to find all sets of additional measurements that allow the identification of key variables and
350 select the optimal solutions in terms of cost and accuracy.

351 **Step 5. Clustering and variable reclassification**

352 The redundancy equations were set up and analysed in view of clustering (Supplementary Material C3). A
353 first group of redundancy equations contains only variables in terms of flows (equations #1-8 in), a second
354 group express the relations between total phosphorus loads (equations #9-16). A third group of redundancy
355 equations (equation #17-32) contains variables from both the COD and nitrogen balances; they can be used
356 to identify both $mCOD$ and mTN variables. The COD and total nitrogen balances need to be considered
357 together because they are related through the amount of denitrified nitrogen, $DENI$. Consequently, the mass
358 balances were also clustered into three corresponding groups.

359 Variable classification was retaken for each group (Table 4). Each group has its own measured variables (M),
360 unmeasured variables (U), potential additionally measured variables (P_a), unmeasurable key variables (X),
361 and key variables (K^*) that appear in the set of (overlapping) mass balances of that group. Consider, for
362 example, the group of overlapping mass balances of flow Q (Supplementary Material C3, mass balances #1-
363 8). In this group, seven key variables K^* need to be identifiable are flow measurements: WWTP influent (Q4),

364 reject water (Q5), settled influent (Q7), WWTP effluent (Q17), waste activated sludge (Q26), primary sludge
 365 (Q28) and waste sludge (Q36). Their identifiability needs to be checked for all subsets of potential
 366 additionally measured variables $P_a = (Q7, Q17, Q27, Q31, Q34, Q37, Q38)$ in this case being $2^7 = 128$ (with
 367 7 the number of elements in P_a of this group).

368 For the group of total phosphorus mass balances (Supplementary Material C3, mass balances #9-16), there
 369 are seven key variables K^* , namely, the total phosphorus mass flow in the influent ($mTP4$), reject water
 370 ($mTP5$), settled influent ($mTP7$), WWTP effluent ($mTP17$), waste activated sludge ($mTP26$), primary sludge
 371 ($mTP28$) and waste sludge ($mTP36$). Their identifiability needs to be checked for all subsets of potential
 372 additionally measured variables $P_a = (Q7, Q17, Q27, Q31, Q34, Q37, Q38, TP7, TP15, TP23, TP26, TP27,$
 373 $TP31, TP34, TP37, TP38, TP39)$ in this case being $2^{17} = 131,072$ (with 17 the number of elements in P_a).

374 Analogously, variable classification was applied to the group of chemical oxygen demand and total nitrogen
 375 balances. Note that, as the volumetric flows Q contribute to all individual mass flow terms, they are part of
 376 potential additionally measured variables of each group (Table 4).

377 **Table 4.** Variable classification for each group of overlapping mass balances. n_m , n_u , n_x , n_p and n_K represent
 378 the number of measured variables M , unmeasured variables U , unmeasurable key variables,
 379 potential additionally measured variables P_a , key variables K^*

		Group of overlapping mass balances		
Description	Flow (Q)	Total phosphorus (TP)	Chemical oxygen demand and total nitrogen (COD & TN)	
M Set of measured variables ($n_m = 34$)	Q4, Q5, Q15, Q23, Q26, Q36, Q39, Q43, Q45 ($n_m = 9$)	Q4, Q5, Q36, Q43, Q45 TP4, TP5, TP17, TP36, TP43, TP45 ($n_m = 11$)	Q4, Q5, Q15, Q23, Q26, Q36, Q39, Q43, Q45 COD4, COD5, COD15, COD17, COD23, COD26, COD31, COD34, COD36, COD38, COD43, COD45, TN4, TN5, TN15, TN17, TN36, TN43, TN45 ($n_m = 28$)	
U Set of unmeasured variables ($n_u = 34$)	Q7, Q17, Q27, Q28, Q31, Q34, Q37, Q38 ($n_u = 8$)	Q7, Q17, Q27, Q28, Q31, Q34, Q37, Q38 TP7, TP15, TP23, TP26, TP27, TP28, TP31, TP34, TP37, TP38, TP39, ($n_u = 19$)	Q7, Q17, Q27, Q28, Q31, Q34, Q37, Q38 COD7, COD27, COD28, COD37, COD39, TN7, TN23, TN26, TN27, TN28, TN31, TN34, TN37, TN38, TN39 ($n_u = 23$)	
X Set of unmeasurable key variables ($n_x = 4$)	Q28 ($n_x = 1$)	Q28 TP28 ($n_x = 2$)	Q28, COD28, TN28 ($n_x = 3$)	
P_a Set of potential additionally measured variables ($n_p = 30$)	Q7, Q17, Q27, Q31, Q34, Q37, Q38 ($n_p = 7$)	Q7, Q17, Q27, Q31, Q34, Q37, Q38, TP7, TP15, TP23, TP26, TP27, TP31, TP34, TP37, TP38, TP39, ($n_p = 17$)	Q7, Q17, Q27, Q31, Q34, Q37, Q38 COD7, COD27, COD37, COD39, TN7, TN23, TN26, TN27, TN31, TN34, TN37, TN38, TN39 ($n_p = 20$)	

Group of overlapping mass balances

Description	Flow (Q)	Total phosphorus (TP)	Chemical oxygen demand and total nitrogen (COD & TN)
Set of key variables K^* $(n_K = 31)$	Q4, Q5, Q7, Q17, Q28, Q26, Q36 (*) $(n_K = 7)$	$mTP4, mTP5, mTP7,$ $mTP17, mTP26, mTP28,$ $mTP36$ $(n_K = 7)$	$mCOD4, mCOD5, mCOD7, mCOD17, mCOD26,$ $mCOD28, mCOD36, mCOD43, OCcod,$ $mTN4, mTN5, mTN7, mTN17, mTN26, mTN28,$ $mTN36, DENI$ $(n_K = 17)$

380 (*) Key variables expressed in volumetric flows are directly equivalent to key variables in total mass flows as the same
381 density is assumed for all streams.

382 Overall, three distinct groups of overlapping mass balances and associated groups of variables were
383 determined: the flow (Q), the mass of total phosphorus (mTP) and the combined group of mass of chemical
384 oxygen demand (mCOD) and mass of total nitrogen (mTN). Each group of mass balances can be effectively
385 used to identify variables that appear in that group – only those and no other ones.

386 Step 6. Finding solutions

387 The determination of sets of additional measurements that guarantee the identification of key variables was
388 performed separately for each group of overlapping mass balances and the obtained results were merged
389 subsequently.

390 For instance, the set of overlapping mass balances for total phosphorus contains seventeen potential
391 additionally measured variables ($n_p = 17$, Table 4), corresponding to $2^{17} = 131,072$ subsets (combinations of
392 variables) of P_a to be analysed. By applying the algorithm (Supplement Material A3), 337 out of 131,072
393 subsets of P_a were found as the solutions A allowing the identification of key variables K^* (Table 4) of this
394 group. Similar interpretation can be done for other groups.

395 Since the key variables of all groups need to be identifiable simultaneously, 80,004 overall solutions A were
396 derived by combining the solution vectors of one group to the ones of others, considering all possible
397 combinations.

398 A non-clustering approach, analysing all possible combinations of initially unmeasured variables and the
399 complete set of mass balances, without distinguishing between groups – essentially skipping Step 5 - was
400 also performed for comparison. The results are summarized in Table 5.

401 **Table 5.** Summary of solution of clustering and non-clustering approach.

Group	$n_p^{(1)}$	possible subsets	Number of solution ⁽²⁾	Execution
-------	-------------	------------------	-----------------------------------	-----------

		of P_a (2^{n_p})		A_z	A	time ⁽³⁾
Clustering	Q	7	128	100		
	TP	17	131,072	337	80,004	47 s
	COD & TN	20	1,048,576	200		
Non-clustering		30	1,073,741,824		80,004	7486 s

- 402 (1) n_p is number of potential additionally variables in terms of flow and concentration
403 (2) A_z is solutions for each group of overlapping mass balances
404 A is final solution after combining solutions of individual groups (duplicates were removed)
405 (3) Procedures were implemented by using Matlab 2014a on desktop CPU i7-4770, RAM 8GB.

406 The total number of subsets to be analysed (total number of P_a of each group) in the clustering approach
407 amounted to 1,179,776 ($= 128 + 131,072 + 1,048,576$), compared to all $2^{30} = 1,073,741,824$ subset of P_a in
408 non-clustering approach (Table 5). It is clear that clustering significantly reduced computational effort, which
409 enables the finding solutions to perform much faster, in this case by a factor of about 150 (47s versus 7486s).

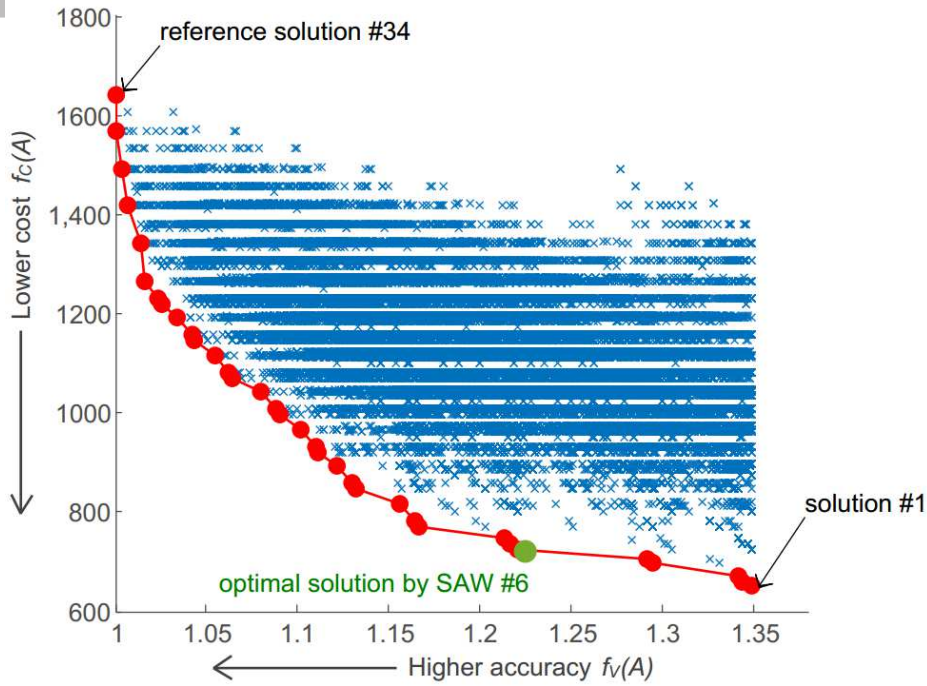
410 The more potential additionally measured variables the system has, the greater advantage of clustering will
411 be. For example, in case of 20 initially measured variables and 40 potential additionally measured variables
412 (compared to 30 initially measured variables and 30 potential additionally measured variables in the
413 presented case study), the number subsets of P_a to be checked in the non-clustering approach would be 2^{40}
414 (about 1×10^{12}). With an average speed of analysing of 150,000 subsets/s with available computational
415 resources, it would take about 80 days for non-clustering approach to solve the problem, while the clustering
416 approach took about 2 hours to complete. The execution time for finding solution greatly depends on the
417 number of initial measurements and the number of key variables.

418 From the $2^{30} = 1,073,741,824$ combinations (subsets) of additional measurement analysed, 80,004 of them,
419 i.e. a fraction of less than 10^{-4} , were found to be solutions that will allow key variables to be identified.

420 Step 7. Optimization

421 The cost and accuracy objective functions were calculated for all 80,004 solutions and are displayed in
422 Figure 3. The Pareto-optimal front is also visualized, containing thirty-four (34) optimal solutions. For these
423 Pareto-optimal solutions, a lower cost can only be obtained at the expense of a lower accuracy and vice
424 versa, a higher accuracy can only be obtained at the expense of a higher cost. The specifications of the
425 Pareto-optimal solutions are listed in Supplementary Material C4.

426



427

428 **Figure 3.** Solutions A are expressed in terms of cost $f_c(A)$ (the lower, the better) and accuracy $f_v(A)$ (the
 429 lower the value, the higher accuracy of the solution or smaller variance of new estimate of key variables).
 430 Each x represents a solution; the line with filled circles (red) represents the Pareto-optimal front, containing
 431 all optimal solutions. The green filled circle denotes the optimal solution #6 selected by the simple additive
 432 weighting method (SAW), see Supplementary Material C4 (for interpretation of the references to colour in
 433 this figure, the reader is referred to the web version of this article).

434 The most accurate (but also most expensive) Pareto-optimal solution is the reference solution #34, for which
 435 all 30 potential additionally measured variables P_a are measured additionally. The reference solution is
 436 characterized by an accuracy $f_v(A) = 1.00$ and cost $f_c(A) = 1642$. The cheapest and least accurate Pareto-
 437 optimal solution is solution #1 with $a = 14$, $f_v(A) = 1.35$ and $f_c(A) = 650$. An accuracy of 1.35 means that the
 438 average variance ($f_v(A)$, see Eq. 2) identified through this solution is 35% higher than the lowest possible
 439 variance, i.e. that of the reference solution and the cost of 650 is the total cost of 14 additional
 440 measurements. An analogous interpretation holds for the other solutions.

441 From the 34 Pareto-optimal solutions, the user can select a favourite one. For instance, applying the additive
 442 weighting method (SAW) results in optimal solution #6 (green-filled circle, Figure 3), requiring $a = 15$
 443 additional measurements and characterized by an accuracy $f_v(A) = 1.22$ and cost $f_c(A) = 725$. While a
 444 minimum number of 14 additionally measured variables is required to have enough redundancy to identify all

445 key process variables, the SAW optimal solution only requires one more additional measurement to offer a
446 better accuracy.

ACCEPTED MANUSCRIPT

448 **Experimental design procedure in view of data reconciliation for wastewater treatment plants.**

449 This contribution presents an experimental design procedure to determine set(s) of additional measurements,
450 which should be carried out to guarantee the identifiability of key variables, meaning that their value can be
451 calculated from other variables based on available constraints – in this case linear mass balances. The
452 identifiability of key variables is a prerequisite for subsequent data reconciliation, through which the reliable
453 and improved estimates for key variables are obtained. The focus on a predefined (limited) number of key
454 variables is very relevant for monitoring campaigns at WWTPs since typically only a few volumetric flow rates
455 and/or components mass flows should be estimated with high accuracy and high reliability while others are
456 not of interest.

457 Experimental design for WWTP data collection has been addressed previously, e.g. by Meijer et al. (2015),
458 Puig et al. (2008) and Rieger et al. (2010). In these studies, measurements and/or mass balances were
459 added such that the number of constraints (independent mass balances) was higher than the number of
460 unknown variables, i.e. aiming at an overdetermined system. In this way, redundancy was considered as a
461 “global property” of the system. This approach, however, does not guarantee the identifiability of all specified
462 key process variables, which is required for the subsequent improvement of their estimates through data
463 reconciliation. It also involves the risk of adding measurements without added value in planned monitoring
464 campaigns. Redundancy is indeed not a “global property” but rather is a property of individual variables (van
465 der Heijden et al., 1994a).

466 In this study, the shortcomings of previous studies (Meijer et al., 2015, 2002; Puig et al., 2008; Rieger et al.,
467 2010) are overcome by unambiguously checking the identifiability of all key variables through the application
468 of redundancy analysis. The feasibility of identifying key variables for the given set of mass balances is
469 checked upfront; mass balances and/or key variables are redefined if needed. The proposed procedure also
470 simplified the set-up of mass balances. In previous studies, it was not always clear to which extent additional
471 mass balances actually provided additional information, i.e. whether they were linearly independent from the
472 previous ones. By applying a feasibility evaluation through redundancy analysis as proposed in this study,
473 one can be confident that the key variables are identifiable for the given set of mass balances.

474 In this work, redundancy analysis was performed following the method of van der Heijden et al. (1994a) and
475 Klamt et al. (2002). This analysis comprises the set-up of redundancy equations, which are derived by

476 eliminating unmeasured variables and linear dependencies from the set of mass balances. Graph-based
477 methods (Kretsovalis and Mah, 1988), as applied by Villez et al. (2016) to determine the optimal layout of
478 flow sensors, constitute an alternative way to analyse redundancy. Graph-based method is intuitive (directly
479 related to topology) and may avoid numerical problems in matrix inversion (particularly when dealing with
480 larger and sparse matrices). Nevertheless, the set-up of redundancy equations and mass balances will still
481 be required as they make up a fundamental part of the data reconciliation procedure. In addition, setting up
482 redundancy equations (redundancy matrix R) allows the identification of groups of overlapping mass
483 balances (clustering) and allows quantifying the accuracy by which key variables can be identified (variance
484 matrix V). For all of these reasons, equation-based redundancy analysis is preferred in this work.

485 **Clustering mass balances in groups of overlapping mass balances**

486 In this work, clustering mass balances in groups of overlapping mass balances was proposed for the first
487 time as an essential part of the experimental design procedure. Clustering significantly reduces the
488 computational effort in finding sets of additional measurements that allow the identification of key variable.
489 Solutions are determined independently for each group and the results for individual groups are
490 subsequently combined. This decomposition makes that a much smaller number of sets of potential
491 additionally measured variables need to be analysed. The advantages of clustering are more pronounced as
492 the number of potential additionally measured variables increases. The number of additional measurement
493 layouts to be analysed exponentially increases (2^n) with the increasing number of potential additionally
494 measured variables (n).

495 In addition, clustering reveals dependencies between variables. The identifiability of variables in one group of
496 overlapping mass balances is independent from the measurement availability of variables in other groups.
497 Therefore, increasing the number of measured variables in one group only helps identifying other variables
498 in the same group. There was not always full awareness of this in previous studies. Moreover, additional
499 measurements of conservative quantities are not always as useful as they were thought to be. For instance,
500 mass flow measurements of total phosphorus, combining measured flow and concentration, are often added
501 to increase system redundancy (Meijer et al., 2015, 2002; Puig et al., 2008). While those measurements
502 increased the number of total phosphorus mass flow variables that could be identified, however, they do not
503 have a direct influence on the identifiability of COD and total nitrogen mass flow variables. An additional
504 measurement of total phosphorus mass flow (flow rate and concentration) could, however, help identifying
505 the key variables in other groups in the coincidental case that the (volumetric) flow rate of the corresponding

506 stream was not initially measured and corresponds to key variables in other groups (mass flows of COD and
507 total nitrogen) of which the concentrations were already measured. Flow measurements contribute more to
508 the identifiability of key variables than concentration measurements in the sense that they contribute to all
509 mass flows of individual components and thus appear in more groups of overlapping mass balances.

510 **Selecting the optimal solutions among alternatives**

511 Among all solutions, the optimal solutions were found considering two objectives, namely the costs of
512 additional measurements and the accuracy of identified key variables. The results of this multi-objective
513 optimization problem were represented in a Pareto front. It is interesting to note that number of Pareto-
514 optimal solutions is very small compared to total number of solutions (fraction of less than 10^{-3}) and
515 represents an even smaller fraction of the total number of possible combinations of additional measurements
516 (less than 10^{-7}). The Pareto-front is a valuable decision tool from which the user can simply select the
517 preferred optimal solution based on expected accuracy and/or monitoring campaign budget. Alternatively,
518 the trade-off between cost and accuracy could be made based on mathematical methods such as simple
519 additive weighting (SAW), multiplicative exponent weighing (MEW), grey relational analysis (GRA), technique
520 for order of preference by similarity to ideal solution (TOPSIS), etc. (Wang and Rangaiah, 2016).

521 The Pareto-optimal solutions are guaranteed to be globally optimal because an exhaustive search was
522 applied: (1) all possible combinations (2^{30} in total) of additional measurements were analysed (through
523 redundancy analysis) to find the solutions for the given set of mass balances and given data inventory, and
524 (2) an accuracy $f_V(A)$ and a cost $f_C(A)$ were calculated for every possible solutions (80,004) found under (1)
525 to find the Pareto-front (i.e., a discrete optimization problem).

526 To maximize the accuracy, this work aims to minimize the average variance of key process variables relative
527 to those of the reference solution (i.e., the solution for which all possible additional measurements are
528 measured additionally, leading to the smallest variance). Other options to maximize accuracy could be to
529 maximize the determinant of the covariance matrix of key variables (D-optimality) or to maximize its minimum
530 eigenvalue (E-optimality). This objective function then needs to be reformulated accordingly.

531 **Application to other WWTPs**

532 The proposed experimental design procedure is simple to apply to other similar WWTPs since it consists of a
533 fixed sequence of steps, all of which are fully explained and documented. Step 1 to step 3 require inputs
534 from the user (for listing key variables, setting up mass balances and inventorying data) following the

535 guidelines. Step 4 to step 7 are fully automated for any problem that can be formulated in the first 3 steps;
536 these steps do not require user intervention except in case there is one or more key variables that cannot be
537 identified for the given set of mass balances and key variables following the indication of Step 4.

538 The procedure was described as a retrofitting problem, in which initial measurements are already available
539 and standard error of variables could be estimated/collected easily. The proposed experimental design
540 procedure remains applicable in case no initial measurements are available, e.g. in case of a WWTP in the
541 design phase. In this case, the standard error of the variables need to be estimated relying on expert
542 knowledge, keeping in mind that their relative values are more important than the absolute values.

ACCEPTED MANUSCRIPT

543 **5 Conclusions**

- 544 – An experimental design procedure for WWTP is proposed to determine sets of additional
545 measurements, which guarantee that key variables can be identified in the sense that they can be
546 calculated from other measurements and therefore, more reliable and improved estimates of these
547 variables can be found through reconciliation.
- 548 – The comprehensive redundancy analysis takes advantage of independent groups of overlapping
549 mass balances to decompose a large system to smaller independent sub-systems, which then
550 significantly reduces computational effort for finding sets of additional measurements that allow the
551 identification of key variables.
- 552 – The search for optimal sets of additional measurements is solved as a multi-objective optimization
553 problem involving cost of additional measurements and accuracy of the improved estimates of key
554 variables. The final result is the enumerated Pareto-optimal front of additional measurements, which
555 is valuable for monitoring planning.
- 556 – The proposed procedure is straightforward and demonstrated for a case study and can easily be
557 applied to other WWTPs, even if no initial measured data are available.

558

559 **Acknowledgements**

560 The authors thank Sebastiaan Meijer for providing data files with the detailed configuration of the WWTP
561 Houtrust. This research did not receive any specific grant from funding agencies in the public, commercial, or
562 not-for-profit sectors.

ACCEPTED MANUSCRIPT

References

- 563
564
565 Behnami, A., Shakerkhatibi, M., Dehghanzadeh, R., Benis, K.Z., Derafshi, S., Fatehifar, E., 2016. The
566 implementation of data reconciliation for evaluating a full-scale petrochemical wastewater treatment
567 plant. *Environ. Sci. Pollut. Res.* 23, 22586–22595.
- 568 Crowe, C.M., 1996. Data reconciliation - Progress and challenges. *J. Process Control* 6, 89–98.
- 569 Klamt, S., Schuster, S., Gilles, E.D., 2002. Calculability analysis in underdetermined metabolic networks
570 illustrated by a model of the central metabolism in purple nonsulfur bacteria. *Biotechnol. Bioeng.* 77,
571 734–751.
- 572 Kretsovalis, A., Mah, R.S.H., 1988. Observability and redundancy classification in generalized process
573 networks—I. Theorems. *Comput. Chem. Eng.* 12, 671–687.
- 574 Lee, S., Rao, S., Kim, M.J., Esfahani, I.J., Yoo, C.K., 2015. Assessment of environmental data quality and its
575 effect on modelling error of full-scale plants with a closed-loop mass balancing. *Environ. Technol.*
576 (United Kingdom) 36, 3253–3261.
- 577 Lotti, T., Kleerebezem, R., Lubello, C., van Loosdrecht, M.C.M., 2014. Physiological and kinetic
578 characterization of a suspended cell anammox culture. *Water Res.* 60, 1–14.
- 579 Madron, F., Veverka, V., 1992. Optimal selection of measuring points in complex plants by linear models.
580 *AIChE J.* 38, 227–236.
- 581 Madron, F., Veverka, V., Vanecek, V., 1977. Statistical-Analysis of Material Balance of a Chemical Reactor.
582 *AICHE J.* 23, 482–486.
- 583 Meijer, S.C.F., van der Spoel, H., Susanti, S., Heijnen, J.J., van Loosdrecht, M.C.M., 2002. Error diagnostics
584 and data reconciliation for activated sludge modelling using mass balances. *Water Sci. Technol.* 45,
585 145–156.
- 586 Meijer, S.C.F., van Kempen, R.N.A., Appeldoorn, K.J., 2015. Plant upgrade using big-data and reconciliation
587 techniques, in: *Applications of Activated Sludge Models*. IWA publishing, p. 500.
- 588 Ozyurt, D.B., Pike, R.W., 2004. Theory and practice of simultaneous data reconciliation and gross error
589 detection for chemical processes. *Comput. Chem. Eng.* 28, 381–402.
- 590 Puig, S., van Loosdrecht, M.C.M., Colprim, J., Meijer, S.C.F., 2008. Data evaluation of full-scale wastewater
591 treatment plants by mass balance. *Water Res.* 42, 4645–4655.

- 592 Pukelsheim, F., 2006. Optimal Design of Experiments. Society for Industrial and Applied Mathematics.
- 593 Rieger, L., Takacs, I., Villez, K., Siegrist, H., Lessard, P., Vanrolleghem, P.A., Comeau, Y., 2010. Data
594 reconciliation for wastewater treatment plant simulation studies-planning for high-quality data and
595 typical sources of errors. *Water Environ. Res.* 82, 426–433.
- 596 Spindler, A., 2014. Structural redundancy of data from wastewater treatment systems. Determination of
597 individual balance equations. *Water Res.* 57, 193–201.
- 598 van der Heijden, R.T.J.M., Heijnen, J.J., Hellinga, C., Romein, B., Luyben, K.C.A.M., 1994a. Linear
599 Constraint Relations in Biochemical Reaction Systems .1. Classification of the Calculability and the
600 Balanceability of Conversion Rates. *Biotechnol. Bioeng.* 43, 3–10.
- 601 van der Heijden, R.T.J.M., Romein, B., Heijnen, J.J., Hellinga, C., Luyben, K.C.A.M., 1994b. Linear
602 Constraint Relations in Biochemical Reaction Systems .3. Sequential Application of Data Reconciliation
603 for Sensitive Detection of Systematic-Errors. *Biotechnol. Bioeng.* 44, 781–791.
- 604 Villez, K., Vanrolleghem, P.A., Corominas, L., 2016. Optimal flow sensor placement on wastewater treatment
605 plants. *Water Res.* 101, 75–83.
- 606 Villez, K., Vanrolleghem, P.A., Corominas, L., 2015. Sensor placement by means of deterministic global
607 optimization, in: *New Development in IT & Water Conference*. Rotterdam, The Netherlands.
- 608 Villez, K., Vanrolleghem, P.A., Corominas, L.I., 2013a. Sensor fault detection and diagnosis based on
609 bilinear mass balances in wastewater treatment systems , in: *11th IWA Conference on Instrumentation
610 Control and Automation (ICA2013)*. Narbonne, France.
- 611 Villez, K., Vanrolleghem, P., Corominas, L., 2013b. Structural observability and redundancy classification for
612 sensor networks in wastewater systems, in: *11th IWA Conference on Instrumentation Control and
613 Automation*. Narbonne, France.
- 614 Wang, Z., Rangaiah, G.P., 2016. Application and Analysis of Methods for Selecting an Optimal Solution from
615 the Pareto-Optimal Front obtained by Multi-Objective Optimization. *Ind. Eng. Chem. Res.*

616

Highlight

- Step-wise measurement planning procedure to obtain reliable data from wastewater treatment plants (WWTPs)
- Right combinations of measurements guarantee improvement of key variables
- Clustering in groups of overlapping mass balances speeds up calculation
- Optimal solutions are trade-off between measurement cost and variable accuracy
- Procedure is demonstrated for case study and straightforward to apply to other WWTPs