Dynamic mass balancing for wastewater treatment data quality control using CUSUM charts

A. Spindler and P. A. Vanrolleghem

ABSTRACT

Mass balancing is a widely used tool for data quality control in wastewater treatment. It can effectively detect systematic errors in data. To overcome the limitations of the mean balancing error as a measure of data quality, a well established method for statistical process control (the CUSUM chart) is adopted for application on the error vector of balancing data. Two examples show how time periods with stable low mass balancing errors can be detected by the method. The detectability of such time periods depends on the variability of the balancing error which is an important measure for the precision of the data.

Key words | data quality control, fault detection, mass balancing, statistical process control

A. Spindler (corresponding author)
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*

P. A. Vanrolleghem
model*EAU*, Dép. de génie civil et de génie des eaux,
Université Laval, Québec,
QC G1V 0A6,
Canada

INTRODUCTION

On wastewater treatment plants (WWTPs) data are routinely collected for reasons of treatment performance evaluation as well as process monitoring and control. The collected data can be a valuable source of information for process redesign, treatment plant extension or simulation. It usually provides a long-term record of the plant performance and is readily available to the engineer. Typically, concentrations of influents and effluents are measured in 24 h composite samples and flows are recorded as daily sums. The advantage of routine data is their availability for long time periods at no extra cost. In contrast, dedicated measurement campaigns might provide a higher sampling frequency but are costly in terms of time and labor and can only cover a comparably short period of time.

To serve as a basis for further engineering tasks, the quality of the routine collected data has to be controlled. Simple or advanced plausibility tests as well as mass balancing are generally applied to meet this requirement (Rieger *et al.* 2010). Plausibility testing is necessary but not sufficient in terms of redundancy. Plausible values can still be (systematically) wrong and sometimes right values might not be plausible. Redundant verification is therefore necessary. Mass balancing can often effectively detect systematic errors in data. Thomann Haller (2002) showed a possibility of testing the significance of the mean balancing error.

Basics of mass balancing

Typical compounds for mass balancing include water H_2O (as flow), and elemental fluxes such as chemical oxygen demand (COD), total phosphorus (P), total nitrogen (N) and iron (Fe). Other compounds can be balanced over systems in which they are not subject to reactions, e.g. total suspended solids (TSS) in dewatering stages.

The mass balance over a system for one compound and for a time period of n days is calculated from all mean fluxes $\bar{F}$ entering (positive) or leaving (negative) the system (Figure 1). It yields the mean balancing error $\bar{e}$ for the particular time period. If accumulation (storage ΔS) of the compound occurs in the system, it also has to be considered (Equations (1a, b)).

It is easily understood that the mean balancing error $\bar{e}$ can be calculated in two distinct ways due to the distributive property of the mean:

(i) as sum of vector means

$$\bar{e} = \sum_{i=1}^{x}\left(\frac{1}{n}\sum_{t=1}^{n} F_{i,\%\text{in},t}\right) + \sum_{j=1}^{y}\left(\frac{1}{n}\sum_{t=1}^{n} F_{j,\text{out},t}\right) + \frac{\Delta S_{n,1}}{n} \quad (1a)$$

(ii) as mean of a vector of sums

$$\bar{e} = \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{i=1}^{x} F_{i,\%\text{in},t} + \sum_{j=1}^{y} F_{j,\text{out},t} + \Delta S_{t,t-1}\right) \quad (1b)$$

doi: 10.2166/wst.2012.125

Figure 1 | Simple balancing layout. Several fluxes may enter or leave a system, accumulation (ΔS) is possible.

In Equation (1a) the means of all single time series of fluxes F in and out of the system as well as the mean accumulation are computed and than added. In Equation (1b) however, balances are calculated for each time step (usually 1 day) thus giving a vector e of (daily) balancing errors of length n (the *error vector*), the mean of which is calculated at the end to give $\bar{e}$.

From $\bar{e}$, the relative mean balancing error $\bar{e}_{\mathrm{rel}}$ is computed by normalization with the mean flux through the system. As a matter of common agreement, the mean influent flux is chosen.

$$\bar{e}_{\mathrm{rel}} = \frac{\bar{e}}{\sum_{i=1}^{x}\left(\frac{1}{n}\sum_{t=1}^{n} F_{i,\mathrm{in},t}\right)} \tag{2}$$

Measures for data quality

Accuracy and precision are the quality criteria for good data. They correspond to systematic and random errors, respectively. Although mass balancing has been accepted as a method of choice for redundant data quality control in the field of wastewater treatment (with a focus on accuracy), little has been said about decision criteria.

The mean balancing error $\bar{e}$ is mainly perceived as the most important decision variable. Thomann Haller (2002) also focused on this measure and showed how to find a confidence interval for $\bar{e}$ to test its significance. However, an insignificant difference between $\bar{e}$ and zero does not determine high data quality alone. A small (relative) mean balancing error can still be significantly different from zero if the precision of the single measurements is high. Low precision might accordingly yield a large confidence interval for $\bar{e}$ thus leading to the misinterpretation of a large $\bar{e}_{\mathrm{rel}}$ as not significantly different from zero. Acceptability of a certain mean relative error therefore seems to be more important than significance. The level of acceptability depends on the task that is addressed using the data.

Another aspect is dynamic variability. While a large $\bar{e}_{\mathrm{rel}}$ certainly signals low data quality (or poor system description), a low $\bar{e}_{\mathrm{rel}}$ could still have been calculated from an error vector e that drifts in time from unacceptably high to unacceptably low values. If data quality is checked relying only on the mean, not much can be said about the data quality in the time series. This is of special importance, when historic data are to be used as input for simulation.

The CUSUM method is suggested to approach the dynamic behavior of the error vector. In the literature, only Zaher & Vanrolleghem (2003) are known to have used this method in the same context, however without explicitly investigating it. Among other control charts, CUSUM is one of the more sensitive. Exponentially weighted moving average (EWMA) charts, another sensitive type of control chart, had also been investigated, but did not yield results of comparable quality. The detectability of changes of the balancing error by the CUSUM method depends on the variability of the error vector and therefore on the precision of the data. This will become clear in the course of this paper.

THE CUSUM CHART

CUSUM charts, introduced by Page (1954), are used widely in statistical process control to detect small changes (e.g. shifts or drifts) in the mean μ (the target value) of a monitored process variable (Montgomery 2009). Small in this context means changes of less than one standard deviation.

CUSUM charts are designed to detect one-sided changes (increase or decrease) of the monitored variable X. For the two-sided case (increase and decrease), one upper (positive) and one lower (negative) CUSUM chart have to be combined. For convenience, data are normalized to a mean value of 0 and a standard deviation of 1. The CUSUM is a modified cumulative sum of a process variable X, consecutively adding up the values x_t, $t = 1,\ldots,n$ where n is the length of vector X. The two modifications are:

(i) The upper (positive) CUSUM may not drop below zero, the lower (negative) CUSUM may not rise above zero.
(ii) A smoothing parameter (reference value k) restricts the sensitivity of the method by constantly drawing the CUSUM series towards the target value (zero for normalized data).

The two-sided CUSUM for normalized data may be defined as:

$$\begin{aligned} C_t^+ &= \max\left(0, C_{t-1}^+ - k + x_t\right) \\ C_t^- &= \min\left(0, C_{t-1}^- + k + x_t\right) \end{aligned} \quad \text{with} \quad C_0 = 0 \tag{3}$$

The CUSUM series signals an undesired shift $\Delta\mu$ of the process mean by exceeding a chosen control limit ($+h$ or $-h$). Thus, the reference value k and the control limit h are the two parameters which determine the behavior of the CUSUM chart. The optimal value of k is $\Delta\mu/2$, half the size of the shift to detect (Lucas & Crosier 1982). The control limit h may then be chosen according to the desired average run length ARL_0 of the CUSUM series (Montgomery 2009).

The average run length ARL_0 is the average number of time steps (i.e. data points) after which the CUSUM series will give a signal even though the true shift of the mean is zero (false alarm). Indeed, due to the probabilistic nature of the data (random errors), a long enough CUSUM series will eventually exceed any control limit. This corresponds to the type I error (false positive) in statistical tests. Therefore, a compromise has to be made. In the past, ARL_0 was chosen as 370 which is equivalent to a 3σ control limit on a Shewart control chart (Montgomery 2009).

When k and h have been chosen, the average run length $ARL_{\Delta\mu}$ (for detection of a true shift $\Delta\mu$ of the mean) can be calculated (Knoth 2009). $ARL_{\Delta\mu}$ increases with decreasing values of k (when h is adjusted to keep a constant ARL_0) and therefore with smaller shifts $\Delta\mu$. In statistical process control a fast response, i.e. low $ARL_{\Delta\mu}$, is desirable.

Synthetic example

Figure 2 (left) depicts data of a synthetic example. The time series has length 200. At intervals [1:40] and [91:140] the random data are $N(0,1)$ distributed. In the interval [41:90] the target value (mean) was changed to $+0.5$. From data point 141 to the end of the series, the mean drifts from 0 to -1. In Figure 2 (right), the results of a CUSUM chart applied to the data are shown. The reference value k was chosen to 0.25 for optimal detection of a shift of ± 0.5. ARL_0 is kept at 370 with a control limit h of ± 8.01 The crucial parts of the CUSUM series are those where it moves away from zero crossing the control limit. In the example, the faulty periods would be interpreted as occurring in intervals [45:100] and [165:200].

Application of the CUSUM method to the error vector of a mass balance

When applying the CUSUM method for the analysis of the error vector of a mass balance, several special characteristics have to be considered:

(i) Historic data are being used. The fastest possible detection of a change of the target is therefore not crucial. This allows for a trial-and-error approach at specifying the design parameters k and h and for more sensitive detection.
(ii) The length of the CUSUM series is determined by data availability. This influences the possible average run length before detection of a true change.
(iii) The CUSUM series does not stop or cause corrective action upon a signal. Therefore, the behavior of the series after a signal is also of interest (as in the synthetic example).
(iv) The process mean (target) is known *a priori*. The expected value of the error vector of a mass balance is always zero.

The ratio between the standard deviation s_e of the error vector before normalization and the total mean input into the system will be shown to be an important indicator for the setup of the CUSUM chart. If the standard deviation of the error vector is relatively high, the data lacks precision.

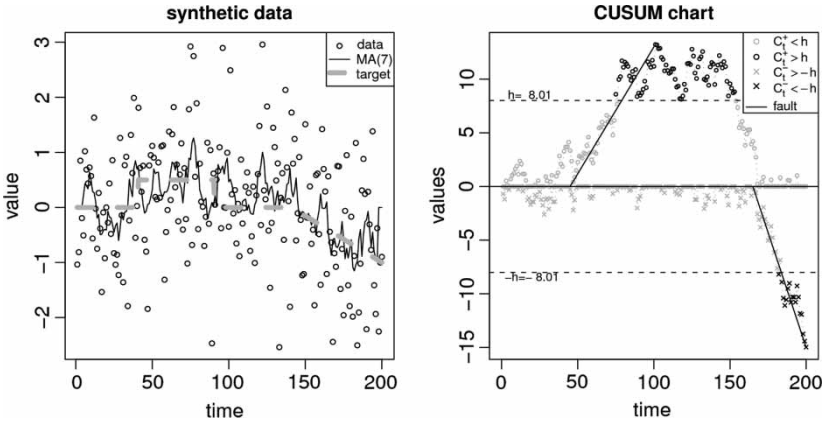

Figure 2 | Left: Synthetic $N(0,1)$ data including a shift and a drift and its 7-day moving average. Right: CUSUM chart of the data. Plotted slopes indicate interpreted faulty periods.

A small shift in the mean of the error vector of less than 0.5 s_e (which is hard to detect) might then already mean a considerable change in one of the fluxes associated with the balance. Therefore, a small reference value k has to be selected. A smaller reference value at constant ARL_0 causes a higher $ARL_{\Delta\mu}$.

The CUSUM method can be applied quite straightforwardly to flow data. The application becomes more challenging, when daily changes in storage also have to be considered. This is the case with all other measured variables, i.e. elemental flux balances. As storage is strongly coupled with TSS concentrations, reliable and representative measurement of this variable is important.

RESULTS OF APPLICATION TO REAL DATA

The CUSUM method was applied to existing routine data of a large WWTP (170.000 PE). The plant has six influents. The two major influents are one municipal and one industrial (refinery). Another two influents stem from the nearby airport (wastewater and surface water). The industrial wastewater (about half of the influent flow) is pretreated in a high-load aerobic stage before joining the aerobic/anoxic treatment for nutrient removal. Because flow Q is the basis for the calculation of fluxes the examples given are: (1) a flow balance over the entire treatment plant; and (2) a flow balance over the anaerobic digester. Unfortunately, it was not possible to include a phosphorus balance as well due to missing data in some fluxes.

The error vectors were calculated from daily flow balances over the two systems for a time period of $n = 366$ days. Table 1 gives the absolute and relative mean flow balance errors and the standard deviation of the error vectors. Figure 3 illustrates the error vectors themselves.

Table 1 | Influent and effluent flow sums for the two examples, absolute and relative mean balancing error and standard deviation of the balancing error

	Whole plant flow balance	Anaerobic digester flow balance
Mean influent flow $\Sigma F_{i,\text{in}}$ m^3/d	24,648	139.6
Mean effluent flow $\Sigma F_{j,\text{out}}$ m^3/d	−25,237	146.9
Mean balancing error $\bar{e} = \Sigma F_{i,\text{in}} + \Sigma F_{j,\text{out}}$ m^3/d	−589	−7.3
Relative mean balancing error $\bar{e}_{\text{rel}} = \bar{e}/\Sigma F_{i,\text{in}}$	−2.4%	−5.3%
Standard deviation s_e m^3/d	848	74.2

Both balances have relatively small mean errors of 2.4 and 5.3%, respectively. The ratio of standard deviation s_e to total mean influent flow, however, is relatively small for the flow balance over the whole WWTP (3.4%) but large for the flow balance over the anaerobic digester (53%). Therefore, the reference value k was chosen differently for each of the two examples. Table 2 illustrates the steps for the setup of the CUSUM chart.

For the whole plant flow balance k was chosen for optimal detection of a shift in the mean of $\Delta\mu = \pm 1.0\ s_e$ ($k = 0.5$). For the flow balance over the anaerobic digester a more sensitive choice was necessary. The reference value was chosen as $k = 0.15$ in order to optimally detect shifts in the mean of $\Delta\mu = \pm 0.3\ s_e$. Note that the detectable relative mass balance errors (i.e. optimally detectable shifts, step 5 in Table 2) are very different. Even though the example of the anaerobic digester was set up for more sensitive detection only balancing errors of about 16% can be optimally detected.

The control limits h were chosen to give an ARL_0 of 370. The resulting $ARL_{\Delta\mu}$ are $ARL_{1.0} = 9.2$ and $ARL_{0.3} = 51$ (Knoth 2009). For the flow balance over the anaerobic digester, a 'design shift' would be detected approximately 51 data points after its occurrence. Given the length of the error

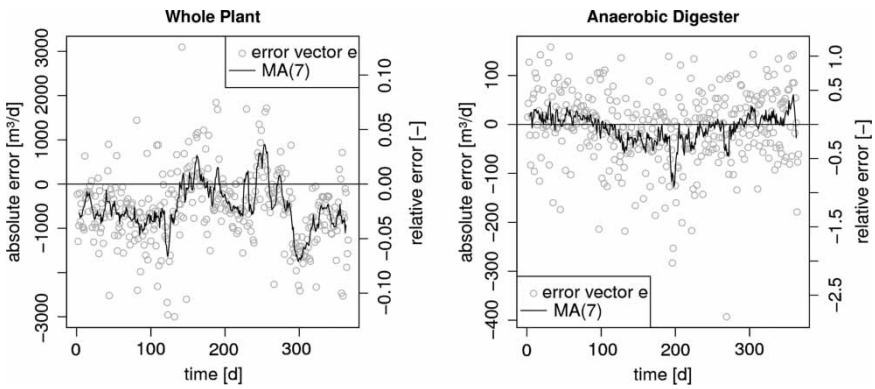

Figure 3 | Error vector e and its 7-day moving average for the two examples.

Table 2 | Steps for setup of CUSUM charts for the two examples (for $N(0,1)$ normalized data $s_e = 1$)

Step	Whole plant flow balance	Anaerobic digester flow balance
0. Consideration of ratio $s_e/\Sigma\bar{F}_{i,in}$	$s_{e,rel} = 3.4\%$	$s_{e,rel} = 53\%$
1. Choice of optimally detectable shift $\Delta\mu$	$\Delta\mu = 1.0\ s_e$	$\Delta\mu = 0.30\ s_e$
2. Reference value $k = \Delta\mu/2$	$k = 0.5\ s_e$	$k = 0.15\ s_e$
3. Calculation of control limit h to give desired ARL_0	$h = 4.77\ s_e$	$h = 11.0\ s_e$
4. Verification of $ARL_{\Delta\mu}$	$ARL_{1.0} = 9.2$ d	$ARL_{0.3} = 51$ d
5. Calculation of relative optimally detectable mass balance error	$\Delta\mu/\Sigma\bar{F}_i = \pm 3.4\%$	$\Delta\mu/\Sigma\bar{F}_i = \pm 16\%$

vector (366 data points) this result seems to be a reasonable compromise between detectability and run length for detection.

Figure 4 shows the CUSUM graphs for both balances. For the whole WWTP two time periods of worse than average balancing performance can be distinguished. Those are the intervals [20:135] and [280:366]. In these time periods the relative mean balancing errors are −3.0 and −4.1%, respectively. Between these two time periods, the mean balancing error drops to −0.3%.

As shown in the synthetic example, the faulty time periods were approximated by following back the slopes of the CUSUM chart. For the anaerobic digester the relative mean balancing error is largest in the time period [120:225] amounting to −28%. At data point 269 the CUSUM series shows a considerable jump, suggesting a major single erroneous data point. Excluding data point 269, the mean relative error for the anaerobic digester in the time period [226:366] is +2.3%.

DISCUSSION

The flow balance over the anaerobic digester obviously contains an error that cannot be neglected. Following the analysis, it was possible to diagnose the source of this error. Interviews with staff pointed to a faulty flow meter in the effluent of the digester. Data from an alternative flow meter were available. Its analysis showed considerably less systematic error (Figure 5). While the standard deviation of the error vector stays at 74.7 m^3/d, the relative mean balancing error drops to as little as +0.2% and is constant throughout the entire time period. For the balance over the whole plant, the error apparently stays small enough to be neglected in any practical application of the data. It might for example be due to minor miscalibration of the flow sensors.

From the two examples it becomes obvious that the calculation of the mean balancing error is not sufficient for determining the quality of routine data from WWTP. In both examples the overall mean balancing error seems relatively small and therefore acceptable at first sight. The application of the CUSUM method clearly showed time periods of varying performance of the error vector. In example 2 (anaerobic digester) a relative mean error of −28% over almost one-third of the entire time series was disguised by the rest of the data.

A 7-day moving average (Figure 3) may already give a good idea about intervals of different performance of the error vector. The CUSUM method however has the advantage of freely selectable control limits and gives a clearer picture. Additionally, the selection of the parameters

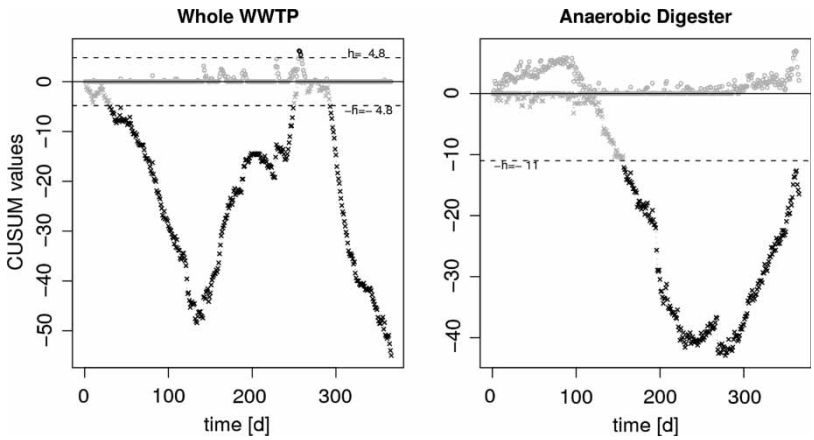

Figure 4 | Two-sided (positive and negative) CUSUM charts for the two examples.

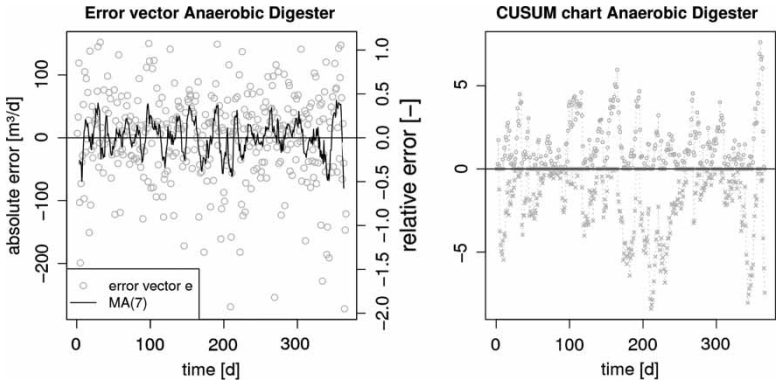

Figure 5 | Error vector and two-sided CUSUM chart for the corrected *Q* balance over the anaerobic digester. Control limits *h* for the CUSUM chart are outside the visible range of the *y*-axis at ±11.

for the CUSUM method allows for the calculation of the optimally detectable mass balance error.

The actually detected mass balance error can still be smaller than the optimally detectable mass balance error. This is the case in the first faulty period in example 1 (whole WWTP). The optimally detectable mass balance error is not a strict limit for detectability but does give a good idea to the user. This reflects the probabilistic nature of random errors which do have a certain unpredictable influence on the performance of the CUSUM method.

When applying the CUSUM method to elemental flux balances, it becomes necessary to also consider storage in the balances. This will mostly be done using daily TSS data and known ratios between the balanced element and TSS. However, representative measurement of TSS is not easily achieved and the resulting error vector might show too high variability. Smoothing of TSS data, i.e. by means of a moving average might solve this problem. Research in this respect is still going on.

CONCLUSIONS

When mass balances are used to determine the quality of routine data from WWTP and to search for systematic errors it is also necessary to consider the error vector of the balance rather than the mean balancing error alone. It has been shown that the CUSUM method can be applied to determine time periods of good balancing performance and to calculate the detectability limits for errors. The variability of the balancing error vector, preferably expressed as a ratio between standard deviation and total mean input load into a system, is an important indicator for these detectability limits.

ACKNOWLEDGEMENTS

The central parts 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.

REFERENCES

Knoth, S. 2009 *spc: Statistical process control*. R package version 0.3. Available from: http://CRAN.R-project.org/package=spc.

Lucas, J. M. & Crosier, R. B. 1982 Fast initial response for CUSUM quality-control schemes: give your CUSUM a head start. *Technometrics* **24** (3), 199–205.

Montgomery, D. 2009 *Introduction to Statistical Quality Control*. Hoboken N.J., Wiley.

Page, E. S. 1954 Continuous inspection schemes. *Biometrika* **41** (1–2), 100–115.

Rieger, L., Takács, I., Villez, K., Siegrist, H., Lessard, P., Vanrolleghem, P. A. & 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.

Thomann Haller, M. P. 2002 *Datenkontrolle von Abwasserreinigungsanlagen mit Massenbilanzen, Experimenten und statistischen Methoden*, PhD thesis, Swiss Federal Institute of Technology Zurich.

Zaher, U. & Vanrolleghem, P.A. 2003 *Data validation, deliverable 2.2*, research report – TELEMAC EU project no. 28156, European Research Framework – Information Society Technologies (IST), pp. 67.

First received 17 August 2011; accepted in revised form 31 January 2012