Comparison of two wavelet-based tools for data mining of urban water networks time series

K. Villez*, G. Pelletier, C. Rosén***, F. Anctil**, C. Duchesne**** and P.A. Vanrolleghem*,****

*BIOMATH, Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure Links 653, B-9000 Gent, Belgium (E-mail: *Kris.Villez@biomath.ugent.be*; *Peter.Vanrolleghem@gci.ulaval.ca*)

** gEAU: Groupe de recherche en génie des eaux, Département de génie civil, Pavillon Adrien-Pouliot, Université Laval, Québec G1K 7P4, Canada (E-mail: *Genevieve.Pelletier@gci.ulaval.ca*; *Francois.Anctil@gci.ulaval.ca*; *Peter.Vanrolleghem@gci.ulaval.ca*)

*** IEA: Department of Industrial Electrical Engineering and Automation, Lund University, LTH, Box 118, SE-221 00, Lund, Sweden (E-mail: *Christian.Rosen@iea.lth.se*)

**** LOOP: Laboratoire d'observation et d'optimisation des procédés, Département de génie chimique, Pavillon Adrien-Pouliot, Université Laval, Québec G1K 7P4, Canada (E-mail: *Carl.Duchesne@gch.ulaval.ca*)

Abstract In this paper, two approaches to data mining of time series have been tested and compared. Both methods are based on the wavelet decomposition of data series and allow the localization of important characteristics of a time series in both the time and frequency domain. The first method is a common method based on the analysis of wavelet power spectra. The second approach is new to the applied field of urban water networks and provides a qualitative description of the data series based on the cubic spline wavelet decomposition of the data. It is shown that wavelet power spectra indicate important and basic characteristics of the data but fail to provide detailed information of the underlying phenomena. In contrast, the second method allows the extraction of more and more detailed information that is important in a context of process monitoring and diagnosis

Keywords B-splines; qualitative representation of trends (QRT); urban water networks; wavelet analysis

Water Science & Technology **Vol 56 No 6 pp 57–64** © IWA Publishing 2007

Introduction

In the past decade wavelet analysis has become a mature approach to localize important characteristics of time series at once in both the time and frequency domain. Applications to process monitoring and diagnosis include fault detection (e.g. Luo *et al.*, 1998, Rosén and Lennox, 2001) and data reconciliation (e.g. Tona *et al.*, 2005). For a complete overview of applications of wavelet-based methods for process monitoring we refer to Ganesan *et al.* (2004). While most wavelet applications deliver a numeric output (a reconciled or filtered signal, statistics), Bakshi and Stephanopoulos (1994) provide a method to obtain a qualitative description of time series following wavelet decomposition, which is improved for robust inflection point detection in Villez *et al.* (in preparation). In Flehmig *et al.* (1998) and Akbaryan and Bishnoi (2000), other wavelet-based approaches are presented for qualitative interpretation of data. As operators typically spend a large proportion of their time to the monitoring of trends in process measurements (Yamanaka and Nishya, 1997) while their reasoning or knowledge is of a qualitative nature, automated extraction of qualitatie information has large potential in the context of fault detection and isolation. Indeed, information about trends may be addressed to the operator only in case of abnormal behaviour so to reduce the time spent by operators on the proofing of normal data. As a result, anomalies can be focused on which will likely result in a faster reaction to and analysis of abnormal situations. Reported applications of wavelet-based

doi: 10.2166/wst.2007.590

methods aim at process monitoring and diagnosis (Akbaryan and Bishnoi, 2001, Rubio *et al.*, 2004 and Flehmig and Marquardt, 2006) and process data mining (Stephanopoulos *et al.*, 1997). Alternative methods to obtain qualitative descriptions of trends are based on piece-wise polynomial fitting (Dash *et al.*, 2004, Charbonnier *et al.*, 2005), PCA-based clustering (Wang and Li, 1999) or neural networks (Rengaswamy and Venkatsubramanian, 1995, Maurya *et al.*, 2005). In view of model structure discrimination, polynomial fits are used by Vanrolleghem and Van Daele (1994). B-spline smoothing is used by Schaich *et al.* (2001) for the same purpose.

While the interpretation of wavelet spectra has become common for time series analysis, the use of qualitative methods for time series analysis has a limited coverage in the literature. It is therefore unclear which of these methods is best for time series analysis. In this paper, two state-of-the-art methods will be compared, being wavelet power spectrum analysis (see e.g. Torrence and Compo, 1998) and wavelet-based qualitative description of trends as described originally by Bakshi and Stephanopoulos (1994), following the modifications of Villez *et al.* (submitted). While classical wavelet spectrum analysis indicates where important features in a series are situated in the time and frequency domain, this does not provide information on the type or shape of the identified features. It will be shown that explicit information regarding the first and second order behaviour of a series can be extracted and leads to additional relevant information about the studied series.

Following this introduction, the data set and applied methods are described in materials and methods. Results and discussion are given in separate sections, where after conclusions are drawn.

Materials and methods

Data description

Drinking water is supplied to a residential neighbourhood of 20.500 inhabitants in the Quebec City area from five groups of wells distributing water to three pressure zones with average water use of 1,050, 4,050 and 600 m^3/d in the lower, intermediate and high pressure zone respectively. Of these groups, four are located in the lower pressure zone, while one is located in the intermediate zone. All groups of wells have a proper local distribution network, but are all connected to a booster station which fills a storage tank (6,800 m^3) located in the high pressure zone. During periods of low water demand, all excess well discharges are pumped to the latter tank. During periods of high water demand, the water tank supplies peak demand to all three zones. The data used in this study are flow measurements from the outlet of this storage tank from November 15th, 2002 to February 1st, 2003 at one minute intervals. One expects low flow rates during the filling of the tank, since it should only supply the high pressure zone but, unfortunately, a flaw in design permits water released from the tank to be pumped back, leading to high energy costs. The daily water demand pattern from such a residential neighbourhood is expected to show two peaks: one in the morning (breakfast, showers) and one in the evening (supper, dishwashing, clothes washing, baths/showers) altered with low flow rate periods in-between, the night time flow rates being the lowest. Weekday patterns should differ from weekends and seasonal patterns would be expected for longer data sets.

Method 1: Wavelet power spectra

The first method used in this paper is based on the wavelet power spectra as described by Torrence and Compo (1998). In this method, the original signal is decomposed by means of a continuous wavelet transform. Practically, this means that the (finite) discrete time series, x_n, is convoluted with a daughter wavelet, $\psi(j,s)$, being a scaled and dilated version of the mother wavelet, ψ_0, for a determined set of scales, $\boldsymbol{s}$, and all discrete time

instants, k (1 to N):

$$W(k,s) = \sum_{j=0}^{N-1} x_j \cdot \psi(j,s)^* \left[\frac{(j-k)\cdot \delta t}{s}\right] \quad (1)$$

where $*$ indicates the complex conjugate and δt the sampling period.

In this work, the Morlet wavelet is used as the mother wavelet for the first method, as in the authors' experiences (e.g. Parent *et al.*, 2006), this wavelet allows a crisp discrimination in the frequency domain. This wavelet is defined in the frequency domain as:

$$\psi_0(\omega) = \pi^{-1/4} \cdot H(\omega) \cdot e^{-(\mathbf{s}\cdot\boldsymbol{\omega}-\boldsymbol{\omega}_0)^2/2} \quad (2)$$

where $H(\omega)$ presents a heaviside step function ($H(\omega) = 1$ if $\omega > 0$, $H(\omega) = 0$ otherwise) and ω_0 is the non-dimensional frequency parameter of the wavelet, set to 6 so that the wavelet function has zero mean and is localized in both time and frequency space (Farge, 1992). To obtain the daughter wavelet in Equation (1), the mother wavelet is dilated, translated and normalized:

$$\psi(j,s)^* \left[\frac{(j-k)\cdot \delta t}{s}\right] = \left(\frac{\delta t}{s}\right)^{1/2} \cdot \psi_0 \left(\frac{(j-k)\cdot \delta t}{s}\right) \quad (3)$$

For practical details and an efficient computation of the described convolution in the Fourier domain, we refer to Torrence and Compo (1998). Of special interest are the studied scales, given as follows:

$$s_p = s_0 \cdot 2^{p\cdot \delta p}, \quad p = 0, 1, ...P \quad (4)$$

In our study, s_0, P and δp were set to 2, 14 and 0.125 respectively so that the studied scales ranged from 2 times the measuring interval (period $= 2$ minutes or approximately 0.0014 days) to 2^{14} times the measuring interval (period $= 32{,}768$ minutes or approximately 23 days) with intervals of 0.125 on a $\log_2$ scale. As the wavelet scale is not necessarily the same as its equivalent Fourier period, the results have been analytically adjusted for this discrepancy to allow a correct interpretation, following the discussion of Torrence and Compo (1998). Future references in this paper to the term "period" indicate the equivalent Fourier period, while scale remains the term for the wavelet period or scale. Torrence and Compo (1998) also provide the cone of influence which defines the region in the obtained spectrum where edge effects distort the wavelet power spectra in such a way that interpretation becomes ambiguous.

Following the derivation of the wavelet coefficients, $W(k,s)$, the powers are calculated as the squared real part of the coefficients, $|W(k,s)|^2$, and normalized by the overall variance of the time series, $|W(k,s)|^2/\sigma^2$. The given values are then a relative measure of the power of a white noise process with the same overall variance. As such, wavelet spectra are straightforward tools to assess non-stationarity, amplitude changes and dominant frequencies of a studied process in time.

Method 2: Analysis of qualitative trends at different scales in wavelet decomposition

The second method, originally described by Bakshi and Stephanopoulos (1994) and modified in Villez *et al.* (in preparation), essentially aims at the description of a time series as a set of contiguous periods in which the series exhibits a constant sign of the first and second derivatives, called triangular episodes. As the human eye cannot detect discontinuities in the third derivative, such a presentation often confirms descriptions given by human operators. Given that the presence of noise prevents the assessment of such a description on the basis of the (discrete) first and second derivatives, the method

can be regarded as a non-parametric adaptive smoother. The method consists of three steps, being wavelet decomposition, qualitative trend identification at each scale and assessment of relevant qualitative features. Each of these steps is described here below.

Step 1: wavelet decomposition. The wavelet decomposition is performed as in the first method (Equations (1) and (2)). The mother wavelet is however replaced with the cubic spline wavelet. Following the work of Mallat and Zhong (1992) the mother wavelet is defined in the Fourier domain as follows:

$$\psi_0(\omega) = \left(\frac{\sin(\omega/2)}{\omega/4}\right)^4 \tag{5}$$

Quite interestingly for trend representation is that (1) the coefficients resulting from the cubic spline wavelet filtering (= band pass filtering) indicate the extrema and inflection points in the original signal and (2) the corresponding low pass filter, defined as the scaling function, does not add extrema or inflection points to an analysed signal because of its smoothing properties. These two properties allow the construction of a filter bank of cubic spline band pass filters and their corresponding low pass filters in such a way that, at each scale, the qualitative trend representation can be determined straightforwardly (see next paragraph). In addition, (3) the constructed filter bank is computationally efficient. The former characteristics allow the use of the cubic spline wavelet for accurate assessments of qualitative features of signals at different frequency scales.

Step 2: qualitative trend identification at each scale. After wavelet decomposition, the qualitative trend representation is constructed. For this purpose, the extrema and inflection points in a signal are identified. Following this identification, 7 types of qualitative behaviour (triangular primitives) are identified (see Table 1). The qualitative representation of a trend is then defined by the maximal time windows in which the qualitative behaviour is the same (maximal time windows with a unique sign of 1st and 2nd derivatives), called triangular episodes (Cheung and Stephanopoulos 1990a,b). Note that the episodes with zero second derivatives (triangular primitives E, F and G) are not occurring often since filtering often distorts the shape of linear parts in a series so that the second derivative in the filtered series is non-zero.

Step 3: assessment of relevant qualitative features. Given the qualitative representations of a series at all scales, one needs to assess which features in these representations are relevant. To do so, the corresponding essential points (maxima, minima and inflection points) are connected over all scales. Then, going from coarser to finer scales, episodes will be split into new episodes with more details. To assess whether such a split is relevant, one uses Witkin's stability criterion, a heuristic criterion originally developed for Gaussian scale-space filtering (Witkin, 1983). This criterion defines that a split into

Table 1 Overview of primitives used to characterize signals on the basis of 1st and 2nd order derivative (U = upward, D = downward)

derivatives		primitives		derivatives		primitives	
1st	2nd	monotonic	triangular	1st	2nd	monotonic	triangular
+	−	U	A	+	0	U	E
−	−	D	B	−	0	D	F
−	+	D	C	0	0	G	G
+	+	U	D				

new episodes is relevant if the mean range of scales over which the more detailed episodes exist is larger than the range over which the coarser episode exist. For more details and an extensive example we refer to Bakshi and Stephanopoulus (1994). In the original method this criterion is applied to monotonic episodes only, i.e. time windows with maximal extent in time in which no extrema lie. In this work, Witkin's stability criterion is first applied for monotonic episodes. Then, the criterion is again applied for candidate triangular episodes. Villez *et al.* (submitted) show that this leads to a more robust assessment of the relevant inflection points.

Results

Method 1: wavelet power spectrum analysis

In Figure 1, a contour plot of the relative power spectra is shown for the data collected between November 15th, 2002 and February 1st, 2003. Over the whole period, high powers are observed in bands at periods of 1 day and $\frac{1}{2}$ day respectively, suggesting regular cyclic behaviour. This concurs with the apparition of two peaks in water demand during each day. The fact that the peaks do not occur at a distance in time of 12 hours leads to dominant powers at two distinct frequency bands. The wavelet power is generally lower at frequencies lower than $\frac{1}{2}$ days, but remarkably, daily peaks in the wavelet powers are seen for a couple of hours during the day. A closer look at the data (not shown) revealed that a sharp increase of the water flow typically occurs during the morning, indeed being a highly dynamic event during a short period during the day. At longer periods, non-regular patterns are observed over the whole period, suggesting non-stationary behaviour in these scales.

Method 2: qualitative description of trends

With the first method, a dominant cycle of 1 day is observed in the data. Based on this observation, the data series is now split into sections of 1 day, each starting and ending at midnight. A separate qualitative representation is assessed for each of these sections. To reduce edge effects during filtering, the data was padded with anti-symmetric data at the

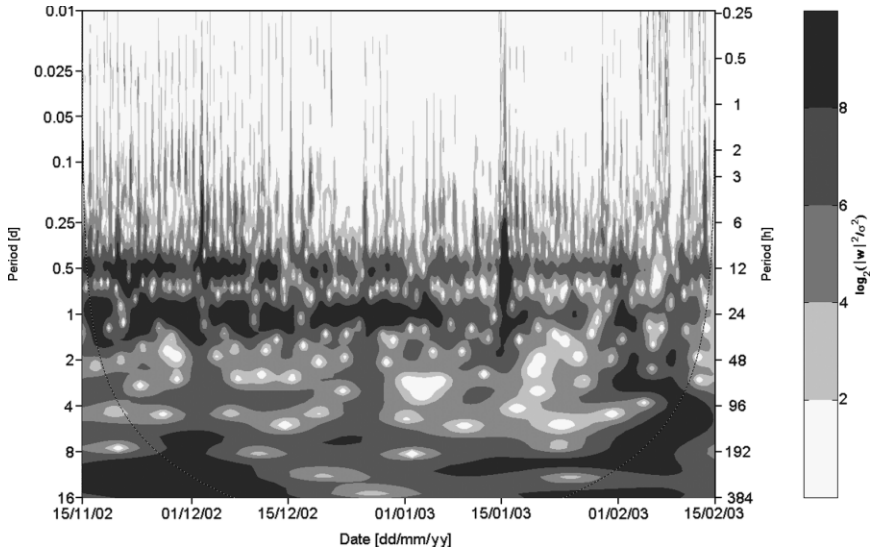

Figure 1 Contour plot of the wavelet power spectra from November 15th, 2005 to February 1st, 2006 and for periods from 0.01 to 16 days. The shading legend is shown at the right hand side. The dashed line shows the cone of influence

left ($y(start\text{-}k) = -y(start + k)$) and right side ($y(end + k) = -y(end\text{-}k)$) of each section. In Figure 2a the monotonic presentations (episodes with constant sign of the first derivative) are displayed for each day in the studied period. Each horizontal bar represents the triangular episodes in a single day. The left and right ends of each rectangle in this bar are the start and end points of the monotonic episodes. It is observed that a larger part of the days have the following pattern: D-U-D-U-D, which means two minima and maxima are observed for these days. The maxima correspond with a peaking water demand in the morning and evening while minima lie in between. For a few days, the second maximum does not occur or is too weak to be accepted into the presentation. Interestingly, these days occur between 21/12 and 6/1, corresponding to Christmas holidays. A significant part of the studied time series exhibit frequent changes in the qualitative behaviour (e.g. 17/12). Visual exploration (not shown) proved that this behaviour is correctly identified (they are not stemming from an erroneous acceptance of noise as relevant features) and showed that these patterns exhibited many step changes. The causes of these (abnormal) step changes could however not be unambiguously linked to sensor failure (incorrect measurement) or control failure (incorrect action). In addition to the qualitative information (chronology of up/down episodes), the location in time of the observed maxima and minima was shown to be relevant as well. It can be seen for instance that the first maximum occurs later on 1/12 and 2/12 when compared to the days just before and after. Remarkably, these two days are a Saturday and a Sunday. The same delay of the first peak occurs for all the other weekend days in the studied period. Such a delay is not observed for the second peak (evening). In addition to the weekend-related delay of the first peak in the day, a similar effect is seen for all the days between 21/12 and 6/1, corresponding to Christmas Holidays. The observed weekend-effect thus also applies to these holidays. It can thus be concluded that the water flow data exhibits a distinct pattern in weekend days and holidays that reflects the behaviour of the city's population (Campos and Van Sperling, 1996).

In Figure 2b, the triangular presentations (episodes with constant sign of first and second derivatives) are given. In addition to the extrema, inflection points are thus shown

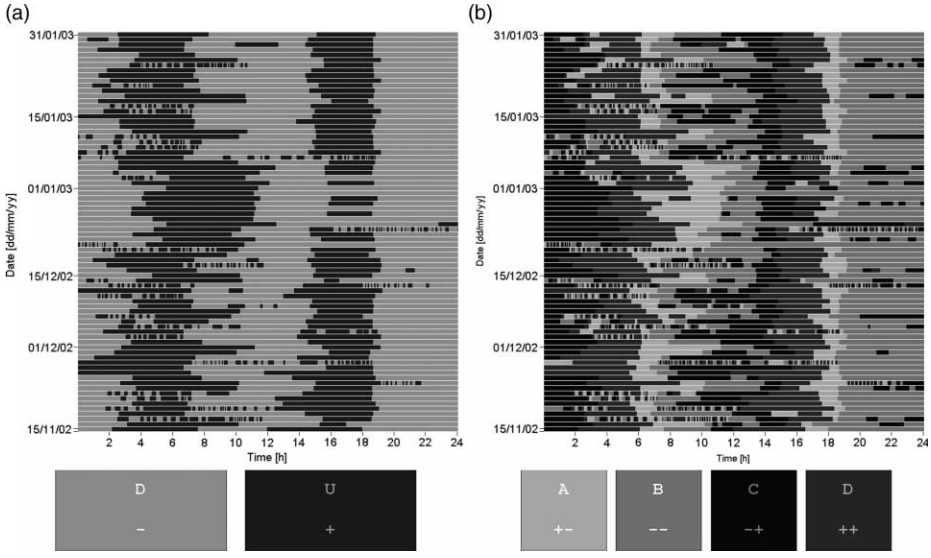

Figure 2 Qualitative representations of time series by means of (a) monotonic primitives and (b) triangular primitives. Each horizontal bar shows the representations of the data of a single day. Shading legends are shown below the respective graphs

as well in this graph. A single inflection point is observed between two extrema during the larger part of the data set, indicating that the acceleration of the flow is typically monotonically increasing or decreasing within each monotonic episode. In a few cases, multiple inflection points occur within one monotonic episode. This happens almost exclusively at night, after the second maximum (e.g. see 01/12). In other words, after the second peak demand, the speed at which the demand decreases shows maxima and minima (inflection points are the extrema of the 1st derivative) for some days.

Discussion

Two wavelet-based methods for mining of time series have been applied to a time series of hydraulic data. In the first method, the signal is transformed into a power measure over time and frequency. As such, relevant dynamics were observed primarily in the scale of days and $\frac{1}{2}$ days. Coarser scales seemed to be characterized by non-cyclic behaviour, while in more detailed scales, regular peaks in power were observed, suggesting highly dynamic events during a limited time-window during most days. In the second method, the resulting wavelet decomposition is further processed to obtain a qualitative representation of the data. As such, relevant maxima, minima and inflection points are identified and define the qualitative representation. By means of this method, typical maxima in water demand in the morning and evening were detected. In addition, the location in time was shown to be dependant on the type of day (working day, weekend day and holiday). Also, inflection points were observed during some nights, indicating a minimal decrease of the water demand during some days. Clearly, the qualitative presentation of the data delivers interesting information regarding the behaviour of a city's population, which is not available from wavelet decomposition only. Conversely, the first method showed that a daily cycle prevails in the studied time series, hereby leading to a window definition for the qualitative representation. Wavelet power spectrum analysis thus functions as an excellent pre-analysis step.

Conclusions

Two state-of-the-art approaches in time series analysis were applied to a time series for which little knowledge was available. It was shown that the first method, wavelet power analysis, indicates the location of major features in frequency and time, but not their type or shape. Since the second method does not aim solely at the analysis of the amplitude of the wavelet coefficients but also of their sign (First order behaviour) and changes over time (Second order behaviour), more detailed information can be extracted by means of qualitative description of trends. In the case presented, such information may be a helpful tool for the design measurement campaigns, modelling of the system and on-line detection of system failures (e.g. leaks).

Acknowledgements

This work was supported by the Institute for Encouragement of Innovation by means of Science and Technology in Flanders (IWT). Peter Vanrolleghem holds the Canada Research Chair in Water Quality Modelling.

References

Akbaryan, F. and Bishnoi, P.R. (2000). Smooth representation of trends by a wavelet-based technique. *Comput. Chem. Eng.*, **24**, 1913–1943.

Akbaryan, F. and Bishnoi, P.R. (2001). Fault diagnosis of multivariate systems using pattern recognition and multisensor data analysis technique. *Comput. Chem. Eng.*, **25**, 1313–1339.

Bakshi, B.R. and Stephanopoulos, G. (1994). Representation of process trends - part III. Multiscale extraction of trends from process data. *Comput. Chem. Eng.*, **18**, 267–302.

Campos, H.M. and von Sperling, M. (1996). Estimation of domestic wastewater characteristics in a developing country based on socio-economic variables. *Wat. Sci. Technol.*, **34**(3-4), 71–77.

Charbonnier, S., Garcia-Beltan, C., Cadet, C. and Gentil, S. (2005). Trends extraction and analysis for complex system monitoring and decision support. *Eng. Appl. Artif. Intell.*, **18**, 21–36.

Cheung, J.T.-Y. and Stephanopoulos, G. (1990a). Representation of process trends - part I. a formal representation framework. *Comput. Chem. Eng.*, **14**, 495–510.

Cheung, J.T.-Y. and Stephanopoulos, G. (1990b). Representation of process trends - part II. The problem of scale and qualitative scaling. *Comput. Chem. Eng.*, **14**, 511–539.

Dash, S., Maurya, M.R. and Venkatasubramanian, V. (2004). A novel interval-halving framework for automated identification of process trends. *AIChE J.*, **50**, 149–162.

Farge, M. (1992). Wavelet transforms and their applications to turbulence. *Annu. Rev. Fluid Mech.*, **24**, 395–457.

Flehmig, F. and Marquardt, W. (2006). Detection of multivariable trends in measured process quantities. *J. Process Control*, **16**, 947–957.

Flehmig, F., Watzdorf, R.V. and Marquardt, W. (1998). Identification of trends in process measurements using the wavelet transform. *Comput. Chem. Eng.*, **22**(suppl), S491–S496.

Ganesan, R., Das, T.K. and Venkataraman, V. (2004). Wavelet-based multiscale statistical process monitoring: A literature review. *IIE Trans.*, **36**, 787–806.

Luo, R., Misra, M., Qin, S.J., Barton, R. and Himmelblau, D.M. (1998). Sensor fault detection via multiscale analysis and nonparametric statistical reference. *Ind. Eng. Chem. Res.*, **37**, 1024–1032.

Mallat, S. and Zhong, S. (1992). Characterization of signals from multiscale edges. *IEEE Trans. Pattern Anal. Mach. Intell.*, **14**, 710–721.

Maurya, M.R., Rengaswamy, R. and Venkatasubramanian, V. (2005). Fault diagnosis by qualitative trend analysis of the principal components. *Chem. Eng. Res. Des.*, **83**, 1122–1132.

Parent, A.-C., Anctil, F. and Parent, L.-É. (2006). Characterization of temporal variability in near-surface soil moisture at scales from 1 h to 2 weeks. *J. Hydrol.*, **325**, 56–66.

Rengaswamy, R. and Venkatasubramanian, V. (1995). A syntactic pattern-recognition approach for process monitoring and fault diagnosis. *Eng. Appl. Artif. Intell.*, **8**, 35–51.

Rosén, C. and Lennox, J.A. (2001). Multivariate and multiscale monitoring of wastewater treatment operation. *Wat. Res.*, **35**(14), 3402–3410.

Rubio, M., Colomer, J., Ruiz, M.L., Colprim, J. and Mélendez, J. (2004). Qualitative trends for situations assessment in SBR wastewater treatment process. In: *Proceedings of the 4th ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence (BESAI)*, Valencia, Spain, August 2004.

Schaich, D., Becker, R. and King, R. (2001). Qualitative modelling for automatic identification of mathematic models of chemical reaction systems. *Control Eng. Practice*, **9**, 1373–1381.

Stephanopoulos, G., Locher, G., Duff, M.J., Kamimura, R. and Stephanopoulos, G. (1997). Fermentation database mining by pattern recognition. *Biotechnol. Bioeng.*, **53**, 443–452.

Tona, R.V., Benqlilou, C., Espuña, A. and Puigjaner, L. (2005). Dynamic data reconciliation based on wavelet trend analysis. *Ind. Eng. Chem. Res.*, **44**, 4324–4335.

Torrence, C. and Compo, G.P. (1998). A practical guide to wavelet analysis. *Bull. Amer. Meteorol. Soc.*, **79**, 61–78.

Vanrolleghem, P. and Van Daele, M. (1994). Optimal experimental design for structure characterization of biodegradation models: on-line implementation in a respirographic bionsensor. *Wat. Sci. Technol.*, **30**(4), 243–253.

Villez, K., Rosén, C., Anctil, F., Duschesne, C. and Vanrolleghem, P.A. (submitted) Qualitative representation of trends: an improved method for multiscale extraction of trends from process data. *Submitted to IEEE Transactions on Signal Processing*.

Wang, X.Z. and Li, R.F. (1999). Combining conceptual clustering and principal component analysis for state space based process monitoring. *Ind. Eng. Chem. Res.*, **38**, 4345–4358.

Witkin, A.P. (1983). Scale space filtering: a new approach to multi-scale description. In: *Image Understanding*, Ullman, S. and Richards, W. (eds.), Ablex, Norwood, NJ, pp. 79–95.

Yamanaka, F. and Nishiya, T. (1997). Application of the intelligent alarm system for the plant operation. *Comput. Chem. Eng.*, **21**, S625–S630.