Pergamon

Wat. Sci. Tech. Vol. 36, No. 5, pp. 19–26, 1997.
© 1997 IAWQ. Published by Elsevier Science Ltd
Printed in Great Britain.
0273–1223/97 $17·00 + 0·00

PII: S0273-1223(97)00496-4

A STOCHASTIC MODEL FOR TWO-STATION HYDRAULICS EXHIBITING TRANSIENT IMPACT

J. L. Jacobsen*, H. Madsen** and P. Harremoës***

* *Institute for Mathemtical Modelling, Building 321 DTU, DK-2800 Lyngby, Denmark and PH-Consult, Ordruphøjvej 4, DK-2920 Charlottenlund, Denamrk*
** *Institute of Mathematical Modelling, DTU*
*** *Institute of Environmental Science and Engineering, DTU and PH-Consult*

ABSTRACT

The objective of the paper is to interpret data on water level variation in a river affected by overflow from a sewer system during rain. The simplest possible, hydraulic description is combined with stochastic methods for data analysis and model parameter estimation. This combination of deterministic and stochastic interpretation is called grey box modelling.

As a deterministic description the linear reservoir approximation is used. A series of linear reservoirs in sufficient number will approximate a plug flow reactor. The choice of number is an empirical expression of the longitudinal dispersion in the river. This approximation is expected to be a sufficiently good approximation as a tool for the ultimate aim: the description of pollutant transport in the river.

The grey box modelling involves a statistical tool for estimation of the parameters in the deterministic model. The advantage is that the parameters have physical meaning, as opposed to many other statistically estimated, empirical parameters. The identifiability of each parameter, the uncertainty of the parameter estimation and the overall uncertainty of the simulation are determined. © 1997 IAWQ. Published by Elsevier Science Ltd

KEYWORDS

Grey box modelling; Maximum likelihood estimation; Stochastic hydraulic model.

INTRODUCTION

It is well known from models of reactor hydraulics, usually pertaining to waste water treatment plants, that modelling the reactor as a series of totally mixed reactors will mathematically merge towards a description for plug flow when the number of reactors in the series goes towards infinity. In hydrology this is called a linear reservoir model (see e.g. Chow et al. (1988)).

This description is proposed for a simple hydraulic model of an urban river, influenced by transient impact during rain events. Whitehead and Young (1975) used a similar concept in a deterministic streamflow forecasting model, where the variation in flow was modelled by a set of first order differential equations and combined with a stochastic black box model for rainfall runoff. This essentially comprises what we call a grey box model. Due to uncertainty in the model, as well as uncertainty in the measurements, the stochastic variation must be included as an integrated part of the model as proposed in this paper. Furthermore Young and Beck (1974), Beck and Young (1975) also used a linear reservoir type model with one reservoir and a transportation delay, while for the approach considered in this paper a series of reservoirs are used.

In this paper a grey box modelling approach of a two-station hydraulics is outlined. For simplicity a very simple system with only three reservoirs is considered. However, this simple model turns out to give a rather good description of the dynamics of the hydraulics under the influence from rain events. Furthermore, the number of reservoirs is supported by statistical analyses.

In the last section it is demonstrated that the modelling approach makes it very easy to suggest extensions of the model; for instance to cope with rivers, where the characteristics of the hydraulics change along the river stretch.

A simple two-station hydraulic model

For practical purposes an infinite number of reservoirs is not feasible. An adequate number must be determined by testing the model on real river data. The data used in this investigation has been sampled for about a week with a resolution of fifteen minute intervals from an urban river, which is influenced by transient impact from rain events. Water levels have been measured at two measuring stations about a kilometer and a half apart and rain has been measured a short distance upstream from these.

For each reservoir, the linearized change in volume over time is:

$$\frac{dV}{dt} = Q_{in} - Q_{out} + Q_{add} \tag{1}$$

where Q_{in} and Q_{out} is the input and output water flow to the reservoir, respectively. The state variable V is the volume and t is the time. Q_{add}, is a term which describes the transient impact from rain events. For simplicity, we assume that $Q_{add} = A_c\ \Phi\ P_r$, where A_c is the catchment area, Φ the runoff coefficient and P_r is the rain used as a proportional factor. It is assumed that the additional water flow is proportional to the intensity of the rain. A system with three reservoirs is shown in Figure 1.

The description of the series of linear reservoirs, depicted in Figure 1 involves a resistance, R_h, which is a function of the specific characteristics of the river, such as slope, roughness and the geometry. The term Q_{add} includes direct overflow from the sewer system, when the sewers and the storage basins cannot keep up with the combined amount of water from runoff and regular sewage, as well as increased masses of water being discharged to the recipient from a waste water treatment plant. Thus, rain data are used as a proportionality factor. Assuming that

$$Q_{in} = \frac{A\,h_{i-1}}{R_h} \quad \text{and} \quad Q_{out} = \frac{A\,h_i}{R_h} \tag{2}$$

We obtain the following equation for one reservoir:

$$\frac{dh_i}{dt} = \frac{h_{i-1} - h_i}{R_h} + \frac{A_c\,\Phi}{A} P_r \tag{3}$$

R_h is the resistance against water flow between the reservoirs, A is the surface area of the stretch of river, and P_r is the proportional factor from the intensity of rain.

For simplicity the resistances are assumed equal in the following. This reflects the belief that the resistance, as well as the river in question is homogeneous. For the proposed method in genereeral, it is however, possible to let the resistance vary along the river. This would reflect stretches of rivers with different riverbed characteristic (e.g. different plants) or varying physical shape.

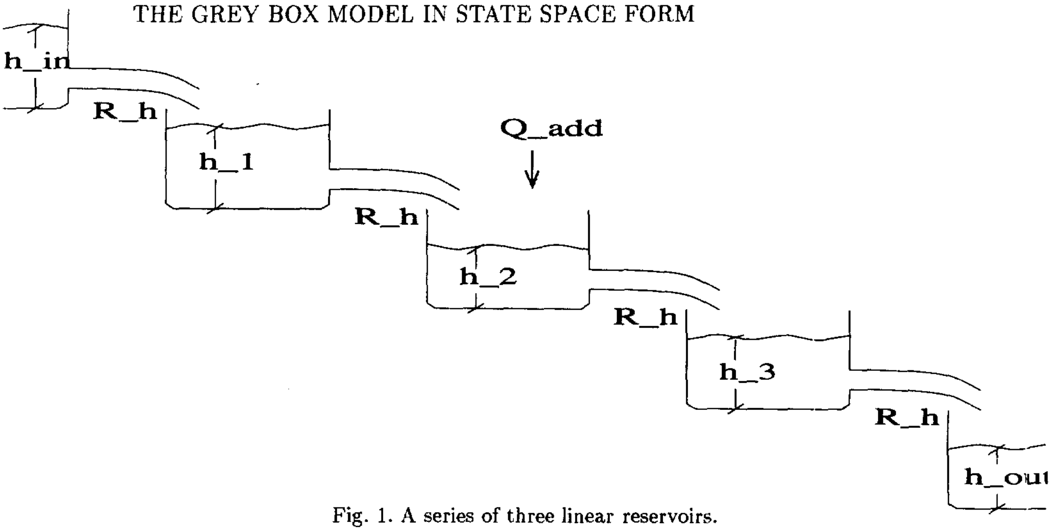

Fig. 1. A series of three linear reservoirs.

A combination of physical, biological and chemical knowledge from e.g. existing oand well-known deterministic models with the information provided by data is a modelling method which uses all available information. This modelling principle was first labeled "grey box modelling by Bohlin (1984), Bohlin (1994). Since traditional time series models based on data only, are called black box modelling and well-known models based on physical, biological and chemical information are sometimes called white box models, we suggest the label: Grey box modelling for the combined principle as is also done in Tulleken (1993). Young and Whitehead (1975) uses the term "internally descriptive" and "databased mechanistic" for essentially the same type of model. A similar linear reservoir model, as the one proposed here, is reported in Whitehead et al. (1979). In both cases, however, they consider models in discrete time, whereas this paper considers continuous time models.

The suggested model is expected to consist of a number of reservoirs and the waterlevels, h_i in reservoir no i, are unmeasured state variables of the model. The known levels are h_{in} and h_{out} at the two measuring stations, respectively.

For the actual river the influx from rain is not considered to be diffuse along the whole stretch of the river section in question, but the influx is rather a point attribution. For the illustrated case the input from rain is only added for one reservoir.

For a model with three reservoirs, the differential equations become:

$$\begin{aligned}
\frac{dh_{out}}{dt} &= \frac{h_3 - h_{out}}{R_h} \\
\frac{dh_3}{dt} &= \frac{h_2 - h_3}{R_h} \\
\frac{dh_2}{dt} &= \frac{h_1 - h_2}{R_h} + \frac{A_c\,\Phi}{A} P_r \\
\frac{dh_1}{dt} &= \frac{h_{in} - h_1}{R_h}
\end{aligned} \tag{4}$$

The water levels (h_1, h_2 and h_3) are not measured, but due to the state space formulation this is not necessary, since they can be estimated using a Kalman filter (see Jacobsen and Madsen (1996) or Jacobsen et al. (1996)).

For n reservoirs, the differential equations can be written as the state space model:

$$\begin{bmatrix} \frac{dh_{out}}{dt} \\ \frac{dh_n}{dt} \\ . \\ . \\ \frac{dh_1}{dt} \end{bmatrix} = \begin{bmatrix} \frac{-1}{R_h} & \frac{1}{R_h} & 0 & . & 0 \\ 0 & \frac{-1}{R_h} & \frac{1}{R_h} & . & 0 \\ 0 & . & . & . & 0 \\ . & . & . & . & . \\ 0 & . & . & . & \frac{-1}{R_h} \end{bmatrix} \begin{bmatrix} h_{out} \\ h_n \\ . \\ . \\ h_1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ . & . \\ 0 & \frac{A_c \Phi}{A} \\ \frac{1}{R_h} & 0 \end{bmatrix} \begin{bmatrix} h_{in} \\ P_r \end{bmatrix} \tag{5}$$

For simplicity, it is assumed that the contribution from rain is affecting only the second last reservoir. The observed water level, at the downstream measuring station, is:

$$Y = \begin{bmatrix} 1 \; 0 \; . \; . \; 0 \end{bmatrix} \begin{bmatrix} h_{out} \\ h_n \\ . \\ . \\ h_1 \end{bmatrix} \tag{6}$$

To include the stochastic variation, an additive noise term, which represents the uncertainty in the input plus the uncertainty in the model, is added to Eq. (5). This is assumed to be a Wiener process (also called a Brownian motion). Also a term for describing the measurement error is introduced. Thus, the stochastic model, when written in matrix notation, becomes:

$$d\boldsymbol{X} = \boldsymbol{A}\boldsymbol{X}dt + \boldsymbol{B}\boldsymbol{U}dt + d\boldsymbol{w}(t) \tag{7}$$

$$\boldsymbol{Y}(t) = \boldsymbol{C}\boldsymbol{X}(t) + \boldsymbol{e}(t) \tag{8}$$

where $\boldsymbol{X}$ is the state vector containing the states of water depths, $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices which characterise the dynamical behaviour of the system and specify how the input signals enter the system, respectively. The input vector $\boldsymbol{U}$ contains the measured input water level and the rain.

In order to be able to find maximum likelihood estimates of the parameters based on Gaussian distributions, the process $\boldsymbol{w}(t)$ is assumed to be a Wiener-process. The measurement error $\boldsymbol{e}(t)$ is assumed to be Gaussian distributed white noise with zero mean. Furthermore, it is assumed that $\boldsymbol{w}(t)$ and $\boldsymbol{e}(t)$ are mutually independent.

In the next section it is briefly outlined how the parameters of the model (state space model (5)), are estimated by a maximum likelihood method. The method accommodates missing observations (for a further description of this feature, see Jacobsen and Madsen (1996) and Jacobsen et al. (1996)).

MAXIMUM LIKELIHOOD ESTIMATES

All observations are equidistantly spaced at fifteen minute intervals. However, in order to simplify the notation, we shall assume that the time index t belongs to the set $\{0, 1, 2, ..., N\}$, where N is the number of observations. Introducing

$$\mathcal{Y}(t) = [\boldsymbol{Y}(t), \boldsymbol{Y}(t-1), \ldots, \boldsymbol{Y}(1), \boldsymbol{Y}(0)]' \tag{9}$$

i.e. $\mathcal{Y}(t)$ is a vector containing all the observations up to and including time t.

All the unknown parameters, denoted by the vector $\boldsymbol{\theta}$, are embedded in the continuous time state space model (eq.s (7) and (8)). The observations are, however, given in discrete time. Hence, the state space model has to be sampled in order to calculate the likelihood function. Under the assumption that the input variables are constant through the sampling time, the discrete-time state-space model can be written (Madsen and Melgaard (1991), Melgaard and Madsen (1993), Jacobsen and Madsen (1996)):

$$\boldsymbol{X}(t+1) = \boldsymbol{\Phi}\,\boldsymbol{X}(t) + \boldsymbol{\Gamma}\,\boldsymbol{U}(t) + \boldsymbol{v}(t) \tag{10}$$

$$\boldsymbol{Y}(t) = \boldsymbol{C}\,\boldsymbol{X}(t) + \boldsymbol{e}(t) \tag{11}$$

where $\boldsymbol{\Phi} = e^{\boldsymbol{A}\tau}, \boldsymbol{\Gamma} = \int_0^\tau e^{\boldsymbol{A}s}\boldsymbol{B}ds$ and $\boldsymbol{v}(t)$ is a white noise sequence.

The likelihood function is the joint probability density of all the observations assuming that the parameters are known, i.e.

$$\begin{aligned} L'(\boldsymbol{\theta};\mathcal{Y}(N)) &= p(\mathcal{Y}(N)|\boldsymbol{\theta}) \\ &= p(\boldsymbol{Y}(N)|\mathcal{Y}(N-1),\boldsymbol{\theta})p(\mathcal{Y}(N-1)|\boldsymbol{\theta}) \\ &= \left(\prod_{t=1}^{N} p(\boldsymbol{Y}(t)|\mathcal{Y}(t-1),\boldsymbol{\theta})\right)p(\boldsymbol{Y}(0)|\boldsymbol{\theta}) \end{aligned} \tag{12}$$

where successive applications of the rule $P(A \cap B) = P(A|B)P(B)$ is used to express the likelihood function as a product of conditional densities.

Since both $\boldsymbol{v}(t)$ and $\boldsymbol{e}(t)$ are normally distributed the conditional density is also normal. The normal distribution is completely characterized by the mean and the variance. Hence, in order to parameterize the conditional distribution, we introduce the conditional mean and the conditional variance as

$$\hat{\boldsymbol{Y}}(t|t-1) = E[\boldsymbol{Y}(t)|\mathcal{Y}(t-1),\boldsymbol{\theta}] \qquad \text{and} \qquad \boldsymbol{R}(t|t-1) = V[\boldsymbol{Y}(t)|\mathcal{Y}(t-1),\boldsymbol{\theta}] \tag{13}$$

respectively. It may be noticed that these correspond to the one-step prediction and the associated variance, respectively. Furthermore, it is convenient to introduce the one-step prediction error (or innovation)

$$\epsilon(t) = \boldsymbol{Y}(t) - \hat{\boldsymbol{Y}}(t|t-1) \tag{14}$$

Using (12) - (14) the conditional likelihood function (conditioned on $\boldsymbol{Y}(0)$) becomes

$$L(\boldsymbol{\theta};\mathcal{Y}(N)) = \prod_{t=1}^{N}\left((2\pi)^{-m/2}\det\boldsymbol{R}(t|t-1)^{-1/2}\exp(-\tfrac{1}{2}\epsilon(t)'\boldsymbol{R}(t|t-1)^{-1}\epsilon(t))\right) \tag{15}$$

where m is the dimension of the $\boldsymbol{Y}$ vector (in the present case m= 1). Traditionally the logarithm of the conditional likelihood function is considered

$$\log L(\boldsymbol{\theta};\mathcal{Y}(N)) = -\tfrac{1}{2}\sum_{t=1}^{N}\left(\log\,\det\boldsymbol{R}(t|t-1) + \epsilon(t)'\boldsymbol{R}(t|t-1)^{-1}\epsilon(t)\right) + \text{const} \tag{16}$$

In the linear case, the conditional mean and variance in Eq. (13) can be calculated recursively by using a Kalman filter. In the non-linear case, an extended Kalman filter is used for calculating the conditional mean and variance. The numerical details of the algorithms can be found in Madsen and Melgaard (1991).

The maximum likelihood estimates are found by maximization of the log likelihood function (16), and the uncertainties of the parameter estimates are found using the observed curvature of the log likelihood function, evaluated at the final estimates (Madsen and Melgaard (1991), Melgaard and Madsen (1993)).

RESULTS AND DISCUSSION

Table 1 show the obtained maximum likelihood estimates for the parameters, and their standard deviations.

TABLE 1: MAXIMUM LIKELIHOOD PARAMETER ESTIMATES AND THEIR STANDARD DEVIATIONS

Symbol	Parameter	Estimate	Unit
R_h	Resistance between reservoirs	0.340	[h]
		(0.744 10^{-2})	
$A_c\,\Phi/A$	Additional water flow	29.750	[-]
		(1.305)	

A positive estimate for the resistance, R_h was to be expected. Since neither the surface area of the river section, nor the catchment area, is known, it is not possible to identify the runoff coefficient, Φ and fully assess this parameter. The ratio between these two areas was estimated to 29.75. With a width of the river about 6 - 7 m and the length of the section about a kilometer and a half, this corresponds roughly to a reduced catchment of about 25 hectars. The standard deviation of both parameters is reasonably small.

The measured, as well as the simulated water level, is seen in Figure 2. The simulation was performed using the estimated parameters. It is obvious that some of the variation is explained by the input (h_{in}); but a simulation without the input from rain has shown that the flow during rain has to be described by the rain input also. Otherwise, the model is not able to simulate the high peaks in the water level satisfactory.

The model was evaluated statistically and several tests for white noise behavior of the residuals have indicated that the model described the dynamics sufficiently. Hence, no more than three reservoirs are needed. On the other hand, less than three reservoirs leads to a model with autocorrelated errors. Hence, exactly three reservoirs are needed for the considered case.

It is seen from Figure 2 that the model slightly underestimates the water level. This could be interpreted as some sort of additional diffuse water supply regardless of whether it rains or not. Or it could be that the water level, h, alone cannot adequately describe the flow.

A constant diffuse water supply, Q_c, could easily be included in the description by a small modification of (5) to the following extended model:

$$\begin{bmatrix} dh_{out} \\ dh_n \\ . \\ . \\ dh_1 \end{bmatrix} = \begin{bmatrix} \frac{-1}{R_h} & \frac{1}{R_h} & 0 & . & 0 \\ 0 & \frac{-1}{R_h} & \frac{1}{R_h} & . & 0 \\ 0 & . & . & . & 0 \\ . & . & . & . & . \\ 0 & . & . & . & \frac{-1}{R_h} \end{bmatrix} \begin{bmatrix} h_{out} \\ h_n \\ . \\ . \\ h_1 \end{bmatrix} dt + \begin{bmatrix} 0 & 0 & Q_c \\ 0 & 0 & Q_c \\ . & . & . \\ 0 & \frac{A_c\Phi}{A} & Q_c \\ \frac{1}{R_h} & 0 & Q_c \end{bmatrix} \begin{bmatrix} h_{in} \\ P_r \\ 1 \end{bmatrix} \tag{17}$$

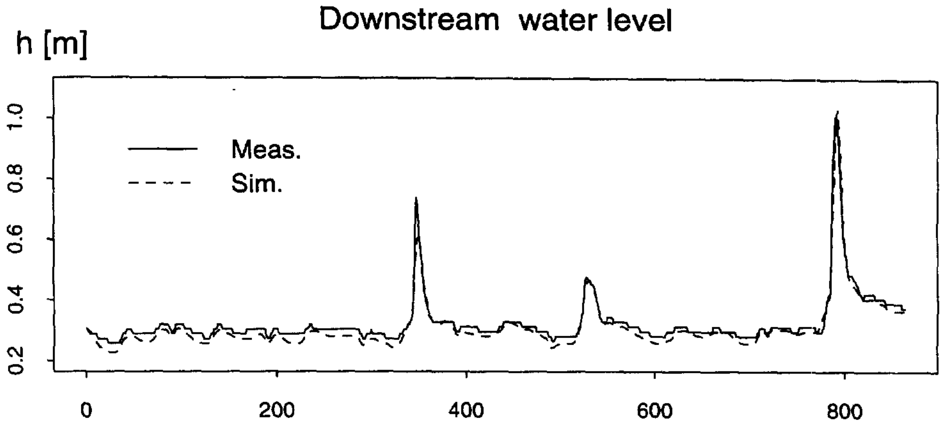

Fig. 2. Measured and simulated water level.

Since the extended model is also a state space model of the general form described by (7) – (8), the parameters of the model (17) could easily be estimated by the suggested maximum likelihood method.

CONCLUSIONS

A description of the hydraulic system for a river receiving water from transient rain events has been proposed. This description is based on a series of linear reservoirs as an idealisation and uses measurements from two stations. The model is formulated as a grey box model in state space form.

A maximum likelihood method for estimating the parameters of state space models in continuous time for hydraulic dynamics is described. An ordinary Kalman filter is used to evaluate the likelihood function. The maximum likelihood approach makes it very easy to evaluate the model by statistical tools as indicated in the paper. Furthermore, the state space formulation is very flexible, and extensions like the incorporation of the diffuse water supply is easy without changing the statistical procedure.

The parameters of a very simple model, with three reservoirs, have been identified. Each of the parameter estimates is within reasonable bounds statistically. The model is based on physical laws for the system, as well as the actual data, making the model a grey box model. The continuous time formulation of the grey box approach has the advantage that a direct physical interpretation of the estimated parameters is possible. For model building, the grey box approach based on continuous time state space models is very attractive and flexible. As described in the paper, it is very easy to introduce and estimate a constant diffuse ground water infiltration, as well as hydraulic models where the resistance against flow varies along the stretch of the river.

The model will subsequently be used as the hydraulic part of a more complex model involving transient loads of water and pollutants and the effects on the oxygen concentration. Measured as dissolved oxygen, this quantity is one of the most universal environmental measures of water quality. As urban rivers and streams are often considered a recreational asset in towns and cities, it is important to be able to model the dynamics, such that an assessment of proposed regulations may be carried out (see e.g. Jacobsen et al. (1996)).

References

Beck, M. and Young, P. (1975). A dynamic model for DO-BOD relationships in a non-tidal stream. *Water Research*, **9**, 769–776.

Bohlin, T. (1984). Computer-aided grey-box validation. Tech. rep. TRITA-REG-8403, Department of Automatic Control, Royal Institute of Technology, Stockholm, Sweden.

Bohlin, T. (1994). A case study of grey box identification. *Automatica*, **30**(2), 307–318.

Chow, V. T., Maidment, D. R., and Mays, L. W. (1988). *Applied Hydrology*. McGraw-Hill International Editions.

Jacobsen, J. L. and Madsen, H. (1996). Grey box modelling of oxygen levels in a small stream. *EnvironMetrics*, **7**, 109–21.

Jacobsen, J. L., Madsen, H., and Harremoës, P. (1996). Modelling the transient impact of rain events on the oxygen content of a small creek. *Water, Science and Technology*, **33**(2), 177–187. (In: Uncertainty, Risk and Transient Pollution Events. Selected proceedings of the IAWQ Interdisciplinary Symposium).

Madsen, H. and Melgaard, H. (1991). The mathematical and numerical methods used in CTLSM - a program for ML-estimation in stochastic, continuous time dynamical models. Tech. rep. no. 7/1991, Institute of Mathematical Statistics and Operations Research, Technical University of Denmark, Lyngby, Denmark.

Melgaard, H. and Madsen, H. (1993). CTLSM continuous time linear stochastic modelling. In Bloem, J. (Ed.), *In: Workshop on Application of System Identification in Energy Savings in Buildings*, pp. 41–60. Institute for Systems Engineering and Informatics, Joint Research Centre.

Tulleken, H. J. A. F. (1993). Grey-box modelling and identification using physical knowledge and bayesian techniques. *Automatica*, **29**(2), 285–308.

Whitehead, P. and Young, P. (1975). *A Dynamic-Stochastic Model for Water Quality in part of the bedford-Ouse River System. In: Computer Simulation of Water Resources Systems, Ed. G.C. Vansteenkiste.* North Holland, Amsterdam.

Whitehead, P., Young, P., and Hornberger, G. M. (1979). A systems model of stream flow and water quality in the Bedford-Ouse river – I. Stream flow modelling. *Water Research*, **13**, 1155–1169.

Young, P. and Beck, B. (1974). The modelling and control of water quality in a river system. *Automatica*, **10**, 455–468.

Young, P. and Whitehead, P. (1975). *A Recursive Approach to Time-series Analysis for Multivariable Systems.* North Holland, Amsterdam.