Model-order selection of output-error models - BSM1 as case study

Christian Wallin[1,2,*] Jesús Zambrano[2]

[1]ABB AB, Power Generation, Västerås, Sweden.
[2]School of Business, Society and Engineering, Mälardalen University, Västerås, Sweden.
[*]Corresponding author e-mail: christian.wallin@se.abb.com

Abstract

Output-Error (OE) System Identification is used to estimate the nonlinear behavior of an activated sludge process (ASP) in a Wastewater Treatment Plant (WWTP). The aim is to identify dynamic models to reproduce the effect of different plant dynamics. How the dissolved oxygen concentration of the aerobic tank affect the effluent ammonia concentration and how the internal recirculation affect the nitrate concentration of the anoxic tank is studied. The best fit of the model is estimated by varying the model order through a trial-and-error approach. Three different scenarios are investigated: one Single-Input-Single-Output (SISO) and two Multiple-Input-Multiple-Output (MIMO) structures. In the SISO scenario only the oxygen to the effluent ammonia dynamics is investigated. Then for both the MIMO scenarios the internal recirculation to nitrate concentration dynamics in the anoxic tank is included and in the last scenario the influent flow rate is also included. The approach is evaluated using the Benchmark Simulation Model no.1 (BSM1).

*Keywords: Benchmark Simulation Model No. 1, Model Predictive Control, Output Error Model, System Identification.*strict

1 Introduction

The activated sludge process (ASP) in a wastewater treatment plant (WWTP) are large non-linear systems subject to perturbations and uncertainty in the influent composition. However, these process should operate continuously and following strict effluent regulations.

From the point of view of control, a system identification of the process is important, mainly because it will improve the control performance of the process, which is typically formed by PI controllers. Another reason is that the system identification can be used to carry out stability analysis of the closed-loop system (Chistiakova et al., 2017).

The system identification involves defining a model structure, mainly a black-box model, where the model parameters are adjusted to fit the data and do not reflect physical consideration Ljung (1999). These models involve a model-order definition with adjustable parameters. The definition of such a model order is still empirical. Typically, a certain model order is assumed, see for example Chistiakova et al. (2017) where an Output-Error (OE) model and nonlinear models were estimated, Ekman (2008) where a bilinear model is estimated and Vrečko et al. (2004) where a state-space model is estimated, those cases used an ASP as case study.

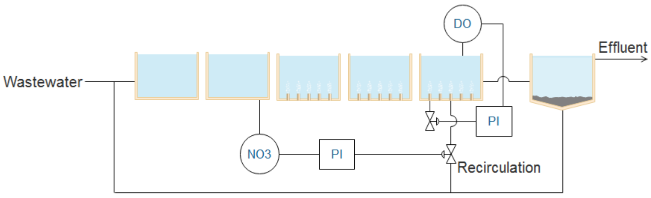

Figure 1. 1 basic layout where the influent water first passes 2 anoxic tanks followed by 3 aerated tanks and then passes a settler before being released. Two control loops are shown: One measures the nitrate concentration of the 2nd anoxic tank to control the internal recirculation and the other measures the oxygen concentration in the last aerated tank to control the air-flow rate.

The aim of this work is to present a way to get an appropriate model-order in a system identification of the process. For the system identification, an OE model is used. An ASP was used case study using data from the Benchmark Simulation Model no. 1 (BSM1) (Alex et al., 2008).

2 The Benchmark Simulation Model

2.1 Description

Data from BSM1 is used for system identification, see the model layout in Figure 1. The BSM1 is a platform that defines a conventional ASP, and includes a simulation model, plant layout, default control systems, performance criteria and test procedures. The plant layout of the BSM1 is formed by a five-compartments ASP, consisting of two anoxic tanks followed by three aerobic tanks and a settler.

The process model is based on the Activated Sludge Model No. 1 (ASM1) (Henze et al., 1987) for the ASP compartments and the Takács model (Takács et al., 1991) for the secondary settler. The BSM1 kinetics and stoichiometric parameter values were kept as default.

The BSM1 includes a constant and a dynamic influent. The constant influent (150 days) was used for the system identification of the SISO case and the MIMO case without influent flow rate, whereas the dynamic influent (14 days) was used for the case MIMO* where influent flow rate is also considered.

https://doi.org/10.3384/ecp18153243

The BSM1 includes two default control loops. One control loop is formed by a PI controller that controls the dissolved oxygen (DO) concentration in the last aerobic tank via air flow rate regulation which is shown in Figure 1 as a valve controlling the incoming air to the blowers. The other control loop is also formed by a PI controller which deals with the control of the nitrate concentration in the last anoxic tank via the internal recirculation flow rate. The sampling time was 15 minutes. The dry weather scenario was used as dynamic influent.

2.2 Case studies

Three different structures were studied for system identification. One structure (referred as SISO case) considers only the DO set-point in the last aerated tank ($S_{O,5}^{sp}$) as input signal, whereas the effluent ammonia concentration ($S_{NH,eff}$) was used as output signal. The DO was modified from the default constant value of 2 mg/l to a range between 0.8 and 2.4 mg/l with a minimum step-interval of the change of set-point set to 100 time-steps.

The other structure (referred as MIMO case) includes the input/output signals of SISO with the addition of the effect of the internal recirculation (Q_{int}) in the nitrate concentration of the second anoxic tank ($S_{NO,2}$).

The last structure (referred as MIMO* case) is the MIMO case including the effect of the influent flow rate (Q_{in}). Here the minimum step-interval is also changed to 10 time-steps because of the lower total simulation time.

3 Method

3.1 System Identification

The structure of the model includes a linear OE model (Ljung, 1999), given as follows

$$\hat{y}(t) = \frac{B(q^{-1})}{F(q^{-1})} u(t - nk) + e(t), \tag{1}$$

where $\hat{y}(t)$ is the output signal of the OE model, $u(t)$ is the input signal, $e(t)$ is the error, nk is a time delay. B and F have the form

$$B(q^{-1}) = b_1 + b_2 q^{-1} + \cdots + b_{nb} q^{-nb+1}, \tag{2}$$

$$F(q^{-1}) = 1 + f_1 q^{-1} + \cdots + f_{nf} q^{-nf}, \tag{3}$$

which are polynomials in the backward shift operator q^{-1} (i.e. $q^{-i}x(k) = x(k-i)$), where $b_i (i = 1,...,nb)$ and $f_i (i = 1,...,nf)$ are unknown parameters, nb and nf are the orders of the OE model.

Part of the system identification involves generating input signals and measurable outputs. The input signals were generated by multiplying a pseudo-random binary sequence (PRBS) with a uniformly distributed random factor. This gives a sequence where each constant value is multiplied with a uniformly distributed amplitude (Wigren, 2003). Figure 2 shows an example of data used for system identification in the MIMO case.

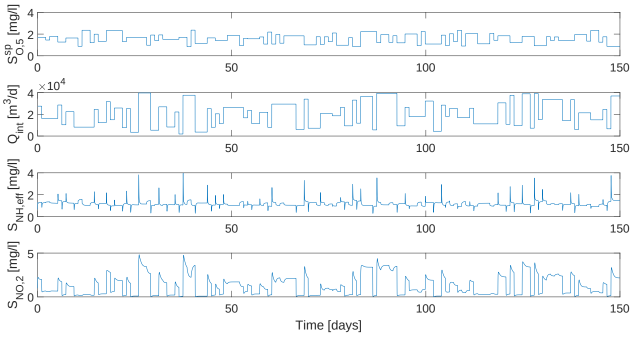

Figure 2. Input signals ($S_{O,5}^{sp}, Q_{int}$) and output signals ($S_{NH,eff}, S_{NO,2}$) used for system identification in MIMO case.

Multiple simulations are done and for each simulation multiple linear models are estimated by varying the order of the B and F polynomials (cf. (2)-(3)).

3.2 Model selection

The Akaike Information Criterion (AIC) (Akaike, 1974) is a way to compare and obtain a good model order. The AIC takes into account the number of parameters and the size of the data set in the following way

$$AIC = \log\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2\right) + \frac{2n_p}{N}, \tag{4}$$

where ε is the error between the estimated model and the BSM1 model, n_p is the number of estimated parameters, and N is the size of the estimation data set.

In this work, the small sample-size corrected Akaike's Information Criterion ($AICc$) has been used, since it was especially developed for regression and autoregressive time series models (Hurvich and Tsai, 1989). The $AICc$ is defined as follows

$$AICc = AIC + \frac{2n_p(n_p+1)}{N - n_p - 1}. \tag{5}$$

See in (5) that the number of estimated parameters is more relevant than the size of the data set. A model with the lowest $AICc$ is expected to be the model that better describes the data with the minimal number of parameters. The same principle applies to the AIC value.

3.3 Model validation

Since each set of input sequence gives a particular best model, the overall best model is obtained by checking how well the particular best models fit the other set of data, i.e. the model are cross validated with the different validation data. This cross validation is quantified using the Fit measure, which is the normalized root mean square error fitness value, defined as

$$Fit = 100 \times \left(1 - \frac{\|y - \hat{y}\|}{\|y - \bar{y}\|}\right), \tag{6}$$

where y is the output of the validation data, $\hat{y}$ is the output given by the estimated model, and $\bar{y}$ is the mean value of the validation data.

The full system identification and model validation process is shown in Figure 3. See that once one simulation is completed, the Input/Output data obtained is used to estimate an OE model in a loop where the orders of the OE model are varied up to a predefined maximum model order M. The best OE model is determined by the $AICc$ value. This process is repeated for a new simulation. Finally, the best model from each simulation is cross-validated against the input and output data from each of the other simulations to determine which OE-model that has the best average Fit value against every simulation made.

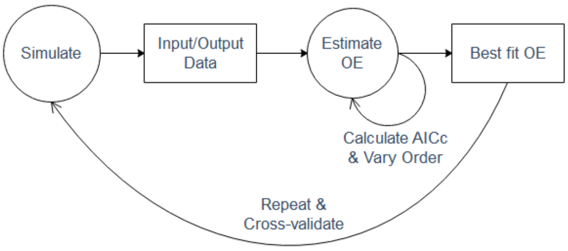

Figure 3. System identification process. A simulation is done and the data from this is collected and used to estimate a model. During model estimation different model orders are tried and each model is evaluated based on the AICc value. Multiple similar simulations are then done and the best model from each simulation is validated against input and output data from other simulations.

The BSM1 was simulated using the Matlab/Simulink®platform, version R2017a. Matlab was also used to estimate the coefficients of the OE-model running the `oe` command with default parameter settings.

4 Results

The best SISO case was estimated by running $i = 100$ cross-realizations of input-output data with model order $M = 10$. Figure 4(a) shows how good the input-output data fit for each of the different cross-fit realizations applied to the SISO case. In the boxplots, the median value is represented by a red line, the edges of the box are the 25th and 75th percentile, the end of the lines are the extreme points that the algorithm consider not to be outliers and outliers are plotted as a red + sign. The best model order obtained was an OE(6,10,1) from trial 93, which gave an average fit of 71.4% with the following coefficients:

$$\begin{cases} B(q^{-1}) & = -0.133q^{-1} + 0.429q^{-2} - 0.545q^{-3} \\ & +0.371q^{-4} - 0.156q^{-5} + 0.034q^{-6}, \\ F(q^{-1}) & = 1 - 3.556q^{-1} + 5.154q^{-2} - 4.172q^{-3} \\ & +2.216q^{-4} - 0.686q^{-5} - 0.17q^{-6} \\ & +0.601q^{-7} - 0.768q^{-8} + 0.523q^{-9} \\ & -0.142q^{-10}. \end{cases} \tag{7}$$

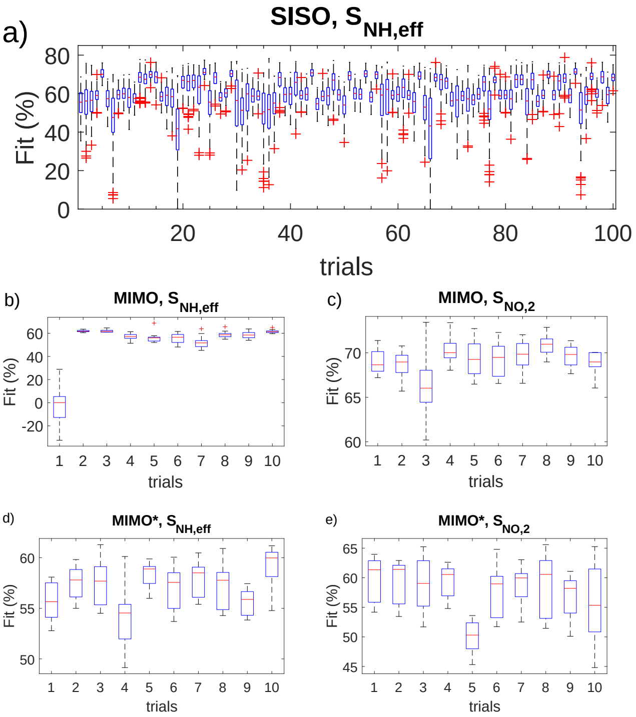

Figure 4. Boxplot of cross-validation for several trials in the SISO, MIMO and MIMO* cases. (a) SISO case, (b)-(c) Output $S_{NH,eff}$ and $S_{NO,2}$ of MIMO case, (d)-(e) Output $S_{NH,eff}$ and $S_{NO,2}$ of MIMO* case. Red line is the median value, the edges of the box are the 25th and 75th percentile, the end of the lines are the extreme points that the algorithm consider not to be outliers and outliers are plotted as a red '+' sign.

The same procedure was applied for the MIMO case. In this case, $i = 10$ and $M = 3$ which is lower than the SISO case due to the increasing processing time. Figure 4(b)-(c) show how good the input-output data fit for each of the different cross-fit realizations applied to the MIMO case. The best model order obtained was from trial 2, which gave a sum average fit of 65.2% and has the following form:

$$\mathrm{OE}\left(\begin{bmatrix} 3 & 3 \\ 2 & 3 \end{bmatrix}, \begin{bmatrix} 3 & 3 \\ 1 & 3 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right),$$

https://doi.org/10.3384/ecp18153243

with the coefficients for $\hat{y}_1(t)$ ($S_{NH,eff}$):

$$\begin{cases} B_1(q^{-1}) & = -0.1197q^{-1} - 0.04353q^{-2} + 0.1633q^{-3}, \\ B_2(q^{-1}) & = 2.767 \times 10^{-5}q^{-1} - 5.52 \times 10^{-5}q^{-2} \\ & +2.754 \times 10^{-5}q^{-3}, \\ F_1(q^{-1}) & = 1 - 0.9046q^{-1} - 0.2211q^{-2} + 0.1258q^{-3}, \\ F_2(q^{-1}) & = 1 - 2.482q^{-1} + 2.021q^{-2} - 0.5382q^{-3}. \end{cases} \tag{8}$$

and for $\hat{y}_2(t)$ ($S_{NO,2}$):

$$\begin{cases} B_1(q^{-1}) & = -0.1747q^{-1} + 0.1817q^{-2} \\ B_2(q^{-1}) & = 4.156 \times 10^{-5}q^{-1} - 8.21 \times 10^{-5}q^{-2} \\ & +4.054 \times 10^{-5}q^{-3}, \\ F_1(q^{-1}) & = 1 - 0.9173q^{-1}, \\ F_2(q^{-1}) & = 1 - 2.601q^{-1} + 2.213q^{-2} - 0.6124q^{-3}. \end{cases} \tag{9}$$

The same procedure was applied for the MIMO* case with influent flow rate. In this case, $i = 10$ and $M = 2$ which is even lower than the MIMO case due to the further increase of processing time. Figure 4(d)-(e) show how good the input-output data fit for each of the different cross-fit realizations applied to the MIMO* case. The best model order obtained was from trial 2, which gave a sum average fit of 58.5% and has the following form:

$$\mathrm{OE}\left(\begin{bmatrix} 2 & 2 & 2 \\ 1 & 2 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\right),$$

with the coefficients for $\hat{y}_1(t)$ ($S_{NH,eff}$):

$$\begin{cases} B_1(q^{-1}) & = -0.5266q^{-1} + 0.4534q^{-2}, \\ B_2(q^{-1}) & = 1.513 \times 10^{-5}q^{-1} - 1.54 \times 10^{-5}q^{-2}, \\ B_3(q^{-1}) & = -3.791 \times 10^{-5}q^{-1} + 5.487 \times 10^{-5}q^{-2}, \\ F_1(q^{-1}) & = 1 - 0.9664q^{-1}, \\ F_2(q^{-1}) & = 1 - 1.614q^{-1} + 0.6388q^{-2}, \\ F_3(q^{-1}) & = 1 - 1.768q^{-1} + 0.8218q^{-2}, \end{cases} \tag{10}$$

and for $\hat{y}_2(t)$ ($S_{NO,2}$):

$$\begin{cases} B_1(q^{-1}) & = 0.0091q^{-1}, \\ B_2(q^{-1}) & = 2.735 \times 10^{-5}q^{-1} + 2.898 \times 10^{-6}q^{-2}, \\ B_3(q^{-1}) & = -4.192 \times 10^{-5}q^{-1}, \\ F_1(q^{-1}) & = 1 - 0.9977q^{-1}, \\ F_2(q^{-1}) & = 1 - 0.7495q^{-1}, \\ F_3(q^{-1}) & = 1 - 0.889q^{-1}. \end{cases} \tag{11}$$

In figure 5(a)-(e) each of the best models is validated against a new simulation. For SISO and MIMO validation, new random sequences are generated for the constant influent scenario, whereas for MIMO* validation new random sequences are generated for the dry influent scenario.

5 Discussions

When more input and output variables are added to the model the complexity increase and the fit goes down. But since more data is considered and being taken into account in these more complex models, they should also be able to handle variations and changes of the data better. The more complex models could also be further improved by increasing the order of the models if higher processing power or more time was used for the system identification process. Another possibility would be to find correlations between the input variables to the OE model estimator namely nb and nf which could eliminate the need to try different orders of some variables and thus the time required to estimate the model. It would also be possible to have variating order of the different inputs variables since some of the output coefficients are very low while others are higher.

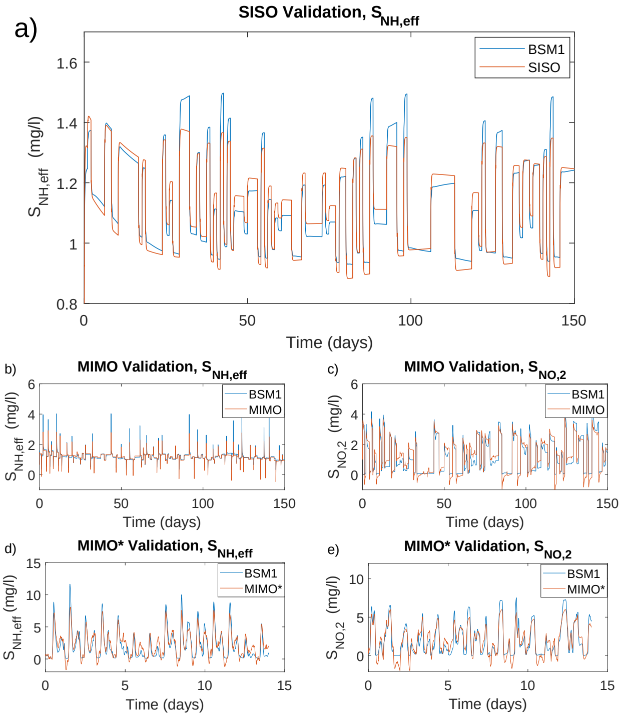

Figure 5. Validation of how well the output data from each of the winning models fit against the output of a new simulation of the same type as the ones used for system identification. Meaning that the input data from this new simulation is used as the input to the model and the output from the simulation is compared with the output of the simulation. (a) SISO case, (b)-(c) Output $S_{NH,eff}$ and $S_{NO,2}$ of MIMO case, (d)-(e) Output $S_{NH,eff}$ and $S_{NO,2}$ of MIMO* case.

A potential use of system identification is from Model Predictive Control (MPC). Usually, the controllers installed in an ASP are based on a proportional-integral (PI) controllers which regulates the air flow rate in the aeration tanks using the feedback from the effluent ammonia concentration. MPC could enhance the response of a given process since it deals with multivariate constrained control problems in an optimal way. MPC has already been tested in ASP models with good results, see for example Foscoliano et al. (2016); Mulas et al. (2015).

The design of MPC involves a system identification of

https://doi.org/10.3384/ecp18153243

the process, where the aim is to achieve a good model of the process in order to get a good control design (Foscoliano et al., 2016). Empirical models have been used for performing the system identification of ASPs, where a pre-defined model order is assumed, see some examples in Vrečko et al. (2004). System identification of simplified ASPs has been carried out by Chistiakova et al. (2017), dealing with linear and non-linear models.

Future studies will be to analyze how well the selected models work as models for MPC and how the fit of the models affects the MPC performance.

Another aspect to consider is different MIMO structures. For example, considering the effluent nitrate concentration. This might require a non-linear system identification of the process.

6 Conclusions

An OE model was used as a black-box model that would describe the input to output relationship of an ASP. A model with a good fit against several scenarios could be obtained by using the rather simple approach of calculating an OE model from input-output data and then cross-validate this model against several similar trials. A model for a SISO scenario could be calculated without much effort to a high model order, whereas the computation time using this method increases fast as the number of input and output variables increase. To reduce the computation time it was required to reduce the cross-validation trials and/or reduce the maximum order of the model which has a negative impact of the fit between the obtained model and the data. Some other black-box models would be investigated, however they might increase the number of parameters to identify.

7 Acknowledgments

This work has received founding from EU Horizon 2020, program SPIRE-2 (Sustainable Process Industry Research), under the project FUDIPO (Future Directions of Production Planning and Optimized Energy - and Process Industries).

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi:10.1109/tac.1974.1100705.

J. Alex, L. Benedetti, J. Copp, K.V. Gernaey, U. Jeppsson, I. Nopens, M.-N. Pons, L. Rieger, C. Rosen, J.P. Steyer, P. Vanrolleghem, and S. Winkler. Benchmark simulation model no. 1 (BSM1). Technical Report (TEIE-7229)/1-62/(2008), Dept. of Industrial Electrical Engineering and Automation. Lund University, 2008.

T. Chistiakova, B. Carlsson, and T. Wigren. Nonlinear modelling of the dissolved oxygen to ammonium dynamics in a nitrifying activated sludge process. In *Instrumentation, Control and Automation - ICA2017*, pages 85–93, Quebec, Canada, 2017.

M. Ekman. Bilinear black-box identification and MPC of the activated sludge process. *Journal of Process Control*, 18(7-8):643–653, 2008. doi:10.1016/j.jprocont.2007.12.006.

C. Foscoliano, S. Del Vigo, M. Mulas, and S. Tronci. Predictive control of an activated sludge process for long term operation. *Chemical Engineering Journal*, 304:1031–1044, 2016. doi:10.1016/j.cej.2016.07.018.

M. Henze, C. Grady, W. Gujer, G. Marais, and T. Matsuo. Activated sludge model no. 1 - scientific and technical report. Technical report, IAWQ, London, UK, 1987.

C. M. Hurvich and C-L Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989. doi:10.1093/biomet/76.2.297.

L. Ljung. *System Identification: Theory for the User (2nd Edition)*. Prentice Hall, 1999. ISBN 0136566952.

M. Mulas, S. Tronci, F. Corona, H. Haimi, P. Lindell, M. Heinonen, R. Vahala, and R. Baratti. Predictive control of an activated sludge process: An application to the Viikinmäki wastewater treatment plant. *Journal of Process Control*, 35:89–100, 2015. doi:10.1016/j.jprocont.2015.08.005.

I. Takács, G.G. Patry, and D. Nolasco. A dynamic model of the clarification-thickening process. *Water Research*, 25(10): 1263–1271, 1991. doi:10.1016/0043-1354(91)90066-y.

D. Vrečko, N. Hvala, and S. Gerlšič. Multivariable predictive control of an activated sludge process with nitrogen removal. *IFAC Proceedings Volumes*, 37(3):505–510, 2004. doi:10.1016/s1474-6670(17)32632-0.

T. Wigren. User choices and model validation in system identification using nonlinear wiener models. *IFAC Proceedings Volumes*, 36(16):837–842, 2003. doi:10.1016/s1474-6670(17)34864-4.