Global NEST Journal, Vol 12, No 1, pp 27-39, 2010

OPTIMIZING THE PERFORMANCE OF KALMAN FILTER BASED STATISTICAL TIME-VARYING AIR QUALITY MODELS

K.I. HOI
K.V. YUEN
K.M. MOK*

Department of Civil and Environmental Engineering, University of Macau, Av. Padre Tomás Pereira Taipa, Macau, China

Received: 10/12/09
Accepted: 11/02/10
*to whom all correspondence should be addressed:
e-mail: KMMOK@umac.mo

ABSTRACT
In this study, the Bayesian approach is proposed to estimate the noise variances of Kalman filter based statistical models for predicting the daily averaged PM_{10} concentrations of a typical coastal city, Macau, with Latitude 22°10'N and Longitude 113°34'E. By using the measurements in 2001 and 2002, the Bayesian approach is capable to estimate the most probable values of the noise variances in the Kalman filter based prediction models. It turns out that the estimated process noise variance of the time-varying autoregressive model with exogenous inputs, TVAREX, is significantly (~76%) less than that of the time-varying autoregressive model of order 1, TVAR(1), since the TVAREX model incorporates important mechanisms which govern the daily averaged PM_{10} concentrations in Macau. By further using data between 2003 and 2005, the choice of the noise variances is shown to affect the model performance, measured by the root-mean-squared error, of the TVAR(*p*) model and the TVAREX model. In addition, the optimal estimates of noise variances obtained by Bayesian approach for both models are located in the region where the model performance is insensitive to the choice of noise variances. Furthermore, the Bayesian approach will be demonstrated to provide more reasonable estimates of noise variances compared to the noise variances found by simply minimizing the root-mean-squared prediction error of the model. By comparing the optimized TVAREX model and the TVAR(*p*) models in predicting the daily averaged PM_{10} concentrations between 2003 and 2005, it is found that the TVAREX model outperforms the TVAR(*p*) models in terms of the general performance and the episode capturing capability.

KEYWORDS: Bayesian inference, Kalman filter, Macau, PM_{10}, Time-varying models.

1. INTRODUCTION

Kalman filter (Kalman and Bucy, 1961) is a popular tool in the discipline of environmental science for predicting and filtering random signals. In addition, it can be used to update the uncertain parameters of environmental models once new measurements are obtained. Examples of application include air quality prediction (Hoi *et al.*, 2008; Hoi *et al.*, 2009), analysis of trace gas concentrations in ice core air bubbles (Trudinger *et al.*, 2002), estimation of nitrate reductase enzyme parameters in activated sludge (Hamilton *et al.*, 2008), parameter estimation for leaky aquifers (Yeh and Huang, 2005) and sea level forecasting (Choi *et al.*, 2002), etc. However, Kalman filter is usually applied with some prior assumption on the variances of the process noise and the measurement noise, which are difficult to be obtained in practice. Nevertheless, the choice of these values was shown to affect the performance of Kalman filter (Yuen *et al.*, 2007; Trudinger *et al.*, 2008). The objective of the present study is to apply the Bayesian approach to estimate the noise variances in order to optimize the performance of the predictive system. Bayesian approach is a probabilistic approach which allows one to obtain the most probable estimates of the unknown parameters of a system and

to quantify the associated uncertainties based on given data (Beck and Katafygiotis, 1998; Yuen 2010).

In order to demonstrate the proposed method, the Kalman filter will be implemented on two types of time-varying statistical air quality models. The first type is the time-varying autoregressive model, which is abbreviated as the TVAR model. The model performs the one-step-ahead prediction of the pollutant concentration by a linear combination of the concentrations of the previous steps. The unknown coefficients in this model are time-varying and to be traced by Kalman filter. The second statistical model type is called the time-varying autoregressive model with exogenous input, which is abbreviated as the TVAREX model. The TVAREX model is a combination of the time-varying autoregressive model and the exogenous inputs that depend on some selected meteorological properties on the day of prediction. The Bayesian approach will then be applied to both time-varying models to estimate the noise variances and a case study was provided in this paper as a reference. In the following section, the formulation of the Kalman filter based time-varying models is briefly described.

2. KALMAN FILTER BASED AIR QUALITY PREDICTION MODELS

2.1. TVAR(*p*) model

In this section, the Kalman filter is formulated for the time-varying autoregressive model of order *p*, which is abbreviated as the TVAR(*p*) model:

$$x_k = \phi_{1,k-1}x_{k-1} + \cdots + \phi_{p,k-1}x_{k-p} + f_{k-1} \tag{1}$$

where x_k denotes the daily averaged air pollutant concentration of the k^{th} day. The input f represents the unmodeled dynamics, and it is modeled as a Gaussian independent and identically distributed (i.i.d.) process with zero mean and variance σ_f^2. It represents the neglected factors that influence the pollutant concentration. In addition, it is assumed that the measurement of the pollutant concentration, denoted as z_k, is contaminated during the measurement process. The relationship between z_k and x_k is given by $z_k=x_k+n_k$, where n is the measurement noise and it is also modeled as Gaussian i.i.d. with zero mean and variance σ_n^2. Furthermore, the stochastic process f and n are assumed to be independent. The TVAR(*p*) model simply implies that the pollutant concentration of a day is a weighted sum of its concentrations of the previous p days with the weights being the time-varying coefficients ϕ_i. These unknown coefficients evolve according to the equation $\phi_{i,k} = \phi_{i,k-1} + w_{i,k-1}$, where $w_{i,k-1}$ denotes the variation made to the coefficient $\phi_{i,,k-1}$ at the $(k\text{-}1)^{\text{th}}$ time step. The stochastic process **w** is modeled as Gaussian i.i.d. with zero mean and covariance matrix $diag\left(\sigma_{w1}^2, \sigma_{w2}^2, \ldots, \sigma_{wp}^2\right)$. Now we define the state vector which contains the pollutant concentrations of different days and the unknown coefficients to be estimated:

$$\mathbf{Y}_k = \left[x_k, \ldots, x_{k-p+1}, \phi_{1,k}, \ldots, \phi_{p,k}\right]^T \tag{2}$$

Then, the measurement z_k can be expressed in terms of $\mathbf{Y}_k$ in the form $z_k = \mathbf{C}\mathbf{Y}_k + n_k$, where **C** is a row vector given by **C** = [1, $\mathbf{0}_{1\times(2p\text{-}1)}$]. Also, a process noise vector which contains the process noises f_k and $w_{i,k}$, i=1,…, p is defined as follows:

$$\mathbf{F}_k = \left[f_k, w_{1,k}, \ldots, w_{p,k}\right]^T \tag{3}$$

It is readily followed that $\mathbf{F}_k$ has zero mean and covariance matrix $\mathbf{Q} = diag\left(\sigma_f^2, \sigma_{w1}^2, \sigma_{w2}^2, \ldots, \sigma_{wp}^2\right)$. Then, the TVAR(*p*) model is linearized locally to a first order TVAR vector model:

$$\mathbf{Y}_k = \hat{\mathbf{A}}_{k-1|k-1}\mathbf{Y}_{k-1} + \mathbf{B}\mathbf{F}_{k-1} + \hat{\mathbf{G}}_{k-1|k-1} \tag{4}$$

The matrix $\hat{\mathbf{A}}_{k\text{-}1|k\text{-}1}$ denotes the filtering estimator of the matrix **A** at the $(k\text{-}1)^{\text{th}}$ time step, conditional on the measurements $z_1,z_2,\ldots,z_{k\text{-}1}$. The same notation is applied to other estimators conditional on the given measurements. The matrices $\hat{\mathbf{A}}_{k\text{-}1|k\text{-}1}$ and **B** are given by:

$$\hat{\mathbf{A}}_{k-1|k-1}=\begin{bmatrix}\hat{\phi}_{1,k-1|k-1} & \cdots & \hat{\phi}_{p,k-1|k-1} & \hat{x}_{k-1|k-1} & \cdots & \hat{x}_{k-p|k-1}\\ & \mathbf{I}_{(p-1)\times(p-1)} & & & \mathbf{0}_{(p-1)\times(p+1)} & \\ & \mathbf{0}_{p\times p} & & & \mathbf{I}_{p\times p} & \end{bmatrix} \tag{5}$$

$$\mathbf{B}=\begin{bmatrix}1 & \mathbf{0}_{1\times p}\\ \mathbf{0}_{(p-1)\times(p+1)} & \\ \mathbf{0}_{p\times 1} & \mathbf{I}_{p\times p}\end{bmatrix} \tag{6}$$

where $\hat{\phi}_{i,k-1|k-1}$ and $\hat{x}_{k-1|k-1}$ denote the filtering estimator of the i th AR coefficient and the pollutant concentration at the (k-1)th time step, conditional on the measurements $z_1,z_2,\ldots,z_{k-1}$. It is noted that **B** is constant for any time step. The vector $\hat{\mathbf{G}}_{k-1|k-1}$ is used to compensate the linearization error:

$$\hat{\mathbf{G}}_{k-1|k-1}=\begin{bmatrix}-\hat{\phi}_{1,k-1|k-1}\hat{x}_{k-1|k-1}-\cdots-\hat{\phi}_{p,k-1|k-1}\hat{x}_{k-p|k-1}\\ \mathbf{0}_{(2p-1)\times 1}\end{bmatrix} \tag{7}$$

With this TVAR vector model, one can perform the predicting and filtering steps of the pollutant concentrations by the Kalman filter. The essential steps of the Kalman filter are to predict and filter the measured concentration alternatingly. When the measured concentrations up to the (k-1)th day are available, the predicting procedure is applied to give the one-step-ahead prediction of the pollutant concentration. The predicting state vector on the kth day can be estimated from the filtered state vector on the (k-1)th day:

$$\hat{\mathbf{Y}}_{k|k-1}=\begin{bmatrix}\hat{\phi}_{1,k-1|k-1}\hat{x}_{k-1|k-1}+\cdots+\hat{\phi}_{p,k-1|k-1}\hat{x}_{k-p|k-1}\\ \hat{x}_{k-1|k-1}\\ \vdots\\ \hat{x}_{k-p+1|k-1}\\ \hat{\phi}_{1,k-1|k-1}\\ \vdots\\ \hat{\phi}_{p,k-1|k-1}\end{bmatrix} \tag{8}$$

In addition, the uncertainty of the air quality prediction and the estimated model parameters can be quantified by the covariance matrix:

$$\mathbf{P}_{k|k-1}=\hat{\mathbf{A}}_{k-1|k-1}\mathbf{P}_{k-1|k-1}\hat{\mathbf{A}}^{T}_{k-1|k-1}+\mathbf{BQB}^{T} \tag{9}$$

When the measurement on the kth day is available, the pollutant concentration and the model parameters are updated as follows:

$$\hat{\mathbf{Y}}_{k|k}=\mathbf{P}_{k|k}\left(\mathbf{P}^{-1}_{k|k-1}\hat{\mathbf{Y}}_{k|k-1}+\sigma_n^{-2}\mathbf{C}^{T}z_k\right) \tag{10}$$

The uncertainty of the state estimation is represented by its covariance matrix:

$$\mathbf{P}_{k|k}=\mathbf{P}_{k|k-1}-\frac{1}{\sigma_n^2+\sigma_{x,k}^2}\mathbf{v}_k\mathbf{v}_k^{T} \tag{11}$$

where $\mathbf{v}_k$ and $\sigma^2_{x,k}$ are the first column vector and the (1,1) component of the one-step-ahead prediction covariance matrix $\mathbf{P}_{k|k-1}$, respectively.

2.2. TVAREX model

The TVAREX model is a time-varying autoregressive model with exogenous inputs:

$$x_k=\left[\phi_{1,k-1}x_{k-1}+\phi_{2,k-1}x'_{k-1}+\phi_{3,k-1}\exp\left(-\alpha u_k-\beta|\theta_k|\right)\right]\exp\left(-\phi_{4,k-1}r_k\right)+f_k \tag{12}$$

It is a statistical model particularly formed according to the nature of a typical coastal city, Macau. However, it is believed that this model is applicable to other coastal cities with slight modification of the model inputs as they are also influenced by similar physical mechanisms. In this model the symbols x_{k-1} and x'_{k-1} denote the daily averaged pollutant concentration of yesterday and the hourly averaged pollutant concentration before midnight, respectively. It is used to reflect the initial condition of the pollutant concentration on the next day. In addition, the symbols u_k and $|\theta_k|$ denote the magnitude and the absolute angle of the resultant wind velocity vector. The resultant wind velocity vector is obtained by the sum of the hourly wind velocity vectors on the day of prediction. The magnitude u_k is associated with the dispersion condition of the k^{th} day. The absolute angle of the resultant wind velocity vector $|\theta_k|$ represents the dominant wind direction on the day of prediction. For example, the 0° refers to the Geographic True North and an absolute angle of 30° denotes the angle of +/- 30° from it. Therefore, the entire range of the wind direction is [0°,180°]. The absolute resultant angle indicates the type of the replenishing air masses being transported to the modeled area. In this study the values of α and β are fixed and those values are specified through the optimization procedure. Assuming fixed values of α and β ensures that there is unique most plausible value for each time-varying coefficient on a given day. In this study, the values of α and β are taken to be 2928.8 hr km^{-1} and 297 deg^{-1}, respectively. The exponential term containing the daily rainfall index r_k, which is defined as the product of the daily rainfall amount and the duration of rainfall on the k^{th} day, is used as a discounting factor on the pollutant concentration for a rainy day. Finally, the term f represents the modeling error and it is modeled as Gaussian i.i.d. with zero mean and variance σ_f^2. By applying similar procedures shown in the section 2.1, the Kalman filter can be implemented on the TVAREX model.

3. BAYESIAN INFERENCE FOR THE NOISE PARAMETERS

As mentioned above, Kalman filter is usually applied with some prior assumption on the variances of the process noise σ_f^2 and the measurement noise σ_n^2, which are difficult to be obtained in practice. The choice of these values was shown to affect the performance of Kalman filter (Yuen *et al.*, 2007; Trudinger *et al.*, 2008). In this study, the Bayesian approach is proposed here to identify these uncertain parameters. Bayesian approach is a probabilistic approach which allows one to obtain the most probable estimates of the unknown parameters of a model and to quantify the associated uncertainties based on given data (Beck and Katafygiotis, 1998; Yan *et al.*, 2009; Yuen and Mu, 2009; Yuen, 2010). First, we define the uncertain parameter vector which contains the process noise and the measurement noise variances:

$$\boldsymbol{\theta} = \left[\sigma_f^2, \sigma_n^2\right]^T \tag{13}$$

By using the Bayes' theorem and given the measurement of the pollutant concentrations $\mathbf{D}=\{z_1,...,z_N\}$, the posterior probability density function of the uncertain parameters is given by:

$$p(\boldsymbol{\theta} \mid \mathbf{D}) = c_0\, p(\mathbf{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{14}$$

where c_0 is a normalizing constant such that the volume of the right side over the parametric space Θ is unity; and $p(\mathbf{D} \mid \boldsymbol{\theta})$ is the likelihood function that represents the level of data fitting. $p(\boldsymbol{\theta})$ is the prior distribution of the uncertain parameters and it reflects the prior knowledge of the user on the uncertain parameters. As a result, the optimal estimate of the parameter vector $\hat{\boldsymbol{\theta}}$ is obtained by maximizing the posterior probability density function $p(\boldsymbol{\theta}|\mathbf{D})$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} p(\boldsymbol{\theta} \mid \mathbf{D}) = \arg\max_{\boldsymbol{\theta}\in\Theta} p(\mathbf{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{15}$$

In this study, a noninformative prior is chosen so it is equivalent to use the maximum likelihood criterion for choosing the optimal parameter vector. The likelihood function can be expressed as the product of the PDFs of the measured concentration z_k conditional on the

parameter vector $\boldsymbol{\theta}$ and all previous measurements $z_1,\ldots,z_{k-1}$, (Yuen and Katafygiotis, 2001; Yuen *et al.,* 2002):

$$p(\mathbf{D} \mid \boldsymbol{\theta}) = \prod_{k=2}^{N} p(z_k \mid z_1,\ldots,z_{k-1},\boldsymbol{\theta}) \tag{16}$$

The measured concentration z_k conditional on all previous measurements and the parameter vector is normally distributed

$$p(z_k \mid z_1,\ldots,z_{k-1},\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)(\sigma_{x,k}^2+\sigma_n^2)}} \exp\left(-\frac{(z_k-\hat{x}_{k|k-1})^2}{2(\sigma_{x,k}^2+\sigma_n^2)}\right) \tag{17}$$

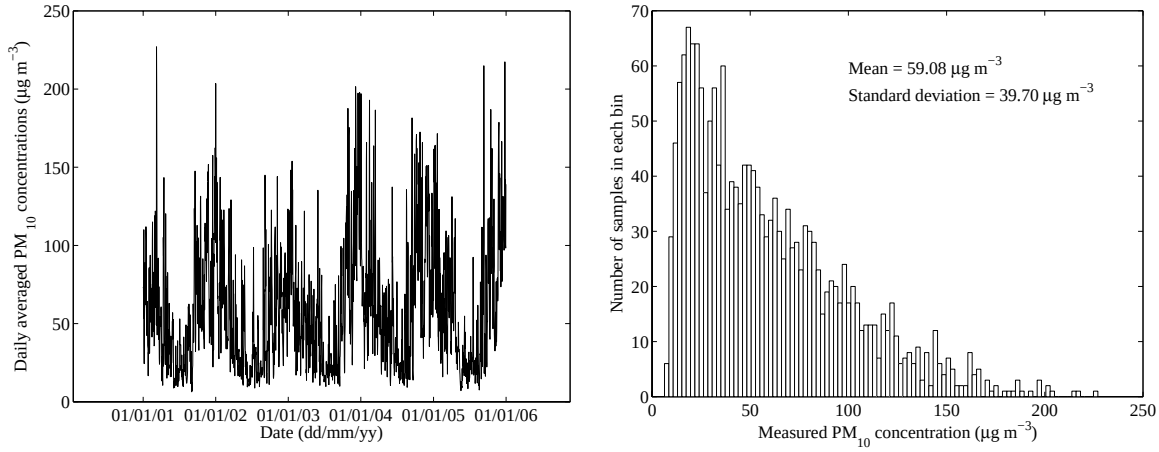
Although the process noise variance does not explicitly appear in the expression, its influence on the likelihood function can be reflected through the variance $\sigma_{x,k}^2$. For a noninformative prior distribution, the optimal parameter vector is found by minimizing the objective function $J(\boldsymbol{\theta}) = -\ln p(\mathbf{D} \mid \boldsymbol{\theta})$ over the search space:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} J(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}\in\Theta}\left\{\frac{N-1}{2}\ln 2\pi + \frac{1}{2}\sum_{k=2}^{N}\left[\ln(\sigma_{x,k}^2+\sigma_n^2) + \frac{(z_k-\hat{x}_{k|k-1})^2}{\sigma_{x,k}^2+\sigma_n^2}\right]\right\} \tag{18}$$

Furthermore, the uncertainty of the estimation can be quantified by using the posterior PDF, e.g., to calculate the standard deviation or the contours with equal probability density.

4. CASE STUDY

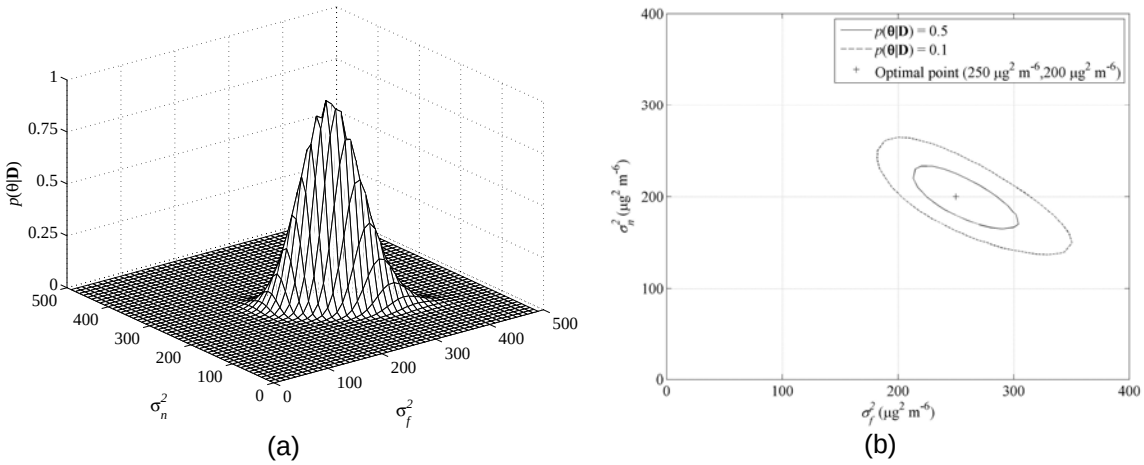
The Kalman filter based air quality prediction models are tested in a typical coastal city, Macau, with Latitude 22°10'N and Longitude 113°34'E. The data consists of daily averaged concentrations of the adverse air pollutant, PM_{10} (Konovalov *et al.*, 2009; Politis *et al.*, 2008; Shan *et al.*, 2009) and the meteorological conditions including the wind speed, the wind direction, and the amount of precipitation recorded at an ambient air quality monitoring station between 2001 and 2005. The station has an altitude of 158.2 m, so its air quality and meteorological measurements are considered to be representative of the general background conditions for the whole city. Figure 1a shows the time series of the measured daily averaged PM_{10} concentrations. It is noted that the time series has a distinct seasonal pattern which is related to the seasonal variation of wind conditions in Macau (Mok and Hoi, 2005). Figure 1b shows the histogram of the measured PM_{10} concentrations. It is noted that the histogram is unimodal and positively skewed with the mean concentration of 59.08 μg m^{-3} and the standard deviation of 39.70 μg m^{-3}.



(a) (b)

Figure 1. (a) Time-series and (b) histogram of measured daily averaged PM_{10} concentrations

Figure 2a shows the posterior PDF of the noise variances conditional on the data between 2001 and 2002 using the TVAR(1) model. It is noted that the posterior PDF is a unimodal distribution. Figure 2b shows the associated contour plot of the posterior PDF. The optimal estimate of the parameter vector for the TVAR(1) model is (250 μg^2 m^{-6},200 μg^2 m^{-6}). The uncertainty of the optimal estimate is represented by the contours with equal probability density. Figure 3 shows the optimal estimates of the parameter vector for different model order of the TVAR(*p*) models from 1 to 10 and the best-fit straight line across the optimal estimates. It is noted that the estimated process noise variance is generally increased, while the estimated measurement noise variance is decreased when *p* is increased from 1 to 3.



(a) (b)

Figure 2. (a) Posterior PDF and (b) contour plot of noise variances for TVAR(1) model

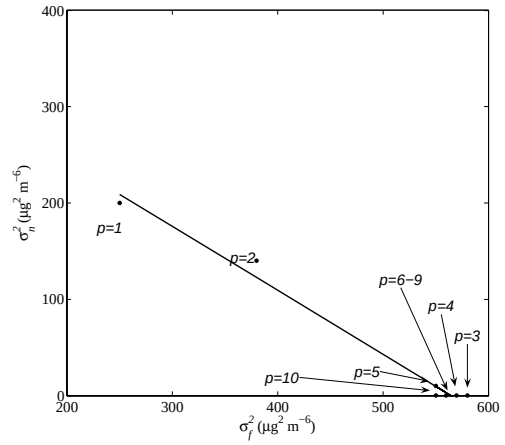


Figure 3. Optimal estimates of the parameter vector for the TVAR(*p*) model

Only small variation in the optimal estimates of the parameter vector is observed when *p* is larger than 3. A higher-order TVAR model tends to have smaller optimal estimate of measurement noise variance since it has more adjustable parameters and hence higher capability to fit the data. Therefore, the fitting error is reduced and smaller optimal estimate of measurement noise variance is obtained. However, the higher-order models may have unnecessarily too many uncertain parameters and overfit the data. Therefore, the process noise of the model is correspondingly increased. Judging from the estimated process noise variance, the TVAR(*p*) model with lower model order is comparatively more representative of the air quality system. To further reinforce the observed trend, the Akaike information criterion (AIC) (Akaike, 1976; Abdel-Aziz and Frey, 2003; Monson, 2009; Peng *et al.*, 2006) was adopted to perform model class selection of the TVAR(*p*) models. The AIC derived for the time-varying models is shown below:

$$AIC = N\ln(2\pi) + \sum_{k=1}^{N}\left[\ln\left(\varsigma_{x,k}^2 + \sigma_n^2\right) + \frac{\left(z_k - \hat{x}_{k|k}\right)^2}{\varsigma_{x,k}^2 + \sigma_n^2}\right] + 2N_{\varphi} \quad (18)$$

where the symbols $\hat{x}_{k|k}$ and $\varsigma_{x,k}^2$ represent the filtered daily averaged PM_{10} concentration and its variance on the k^{th} day, and N_{ϕ} denotes the number of uncertain parameters in the model. The first two terms in the expression represent the goodness of fit, while the last term penalizes the model with more uncertain parameters. In general, a more efficient and robust model tends to have lower value of AIC. Figure 4 shows the natural logarithm of AIC for the TVAR(*p*) models with different model orders from 1 to 10. It is noted that a general increasing trend of ln(AIC) is observed when *p* is increased from 1 to 10. According to the Akaike information criterion, the most parsimonious model TVAR(1) should be chosen among those 10 possible candidates and the increasing trend further supports the aforementioned judgement. Figures 5a and 5b shows the posterior PDF and the contour plot for the noise variances of the TVAREX model. The optimal estimate of the parameter vector for the TVAREX model is (60 $\mu g^2\ m^{-6}$,220 $\mu g^2\ m^{-6}$). The process noise variance of the TVAREX model is significantly (~76%) less than that of the TVAR(1) model since the exogenous inputs of the TVAREX model reflect the influencing physics controlling the variation of PM_{10} concentration in Macau. Therefore, the model class becomes more representative and this causes the process noise to be reduced.

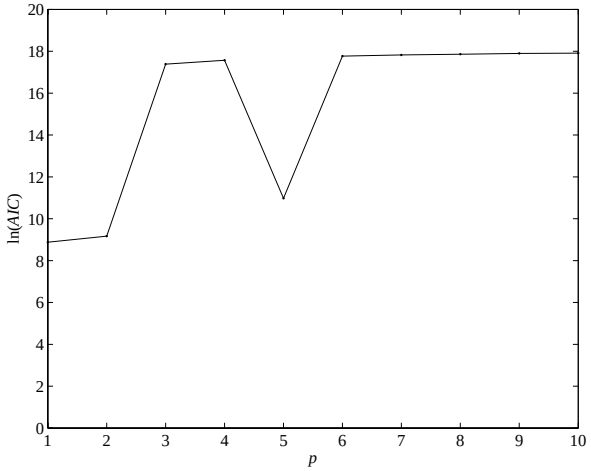
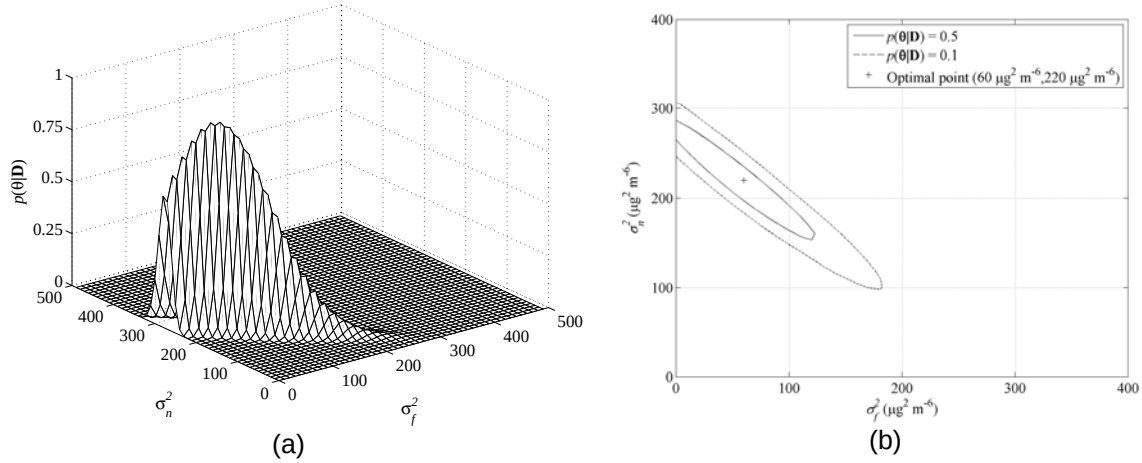


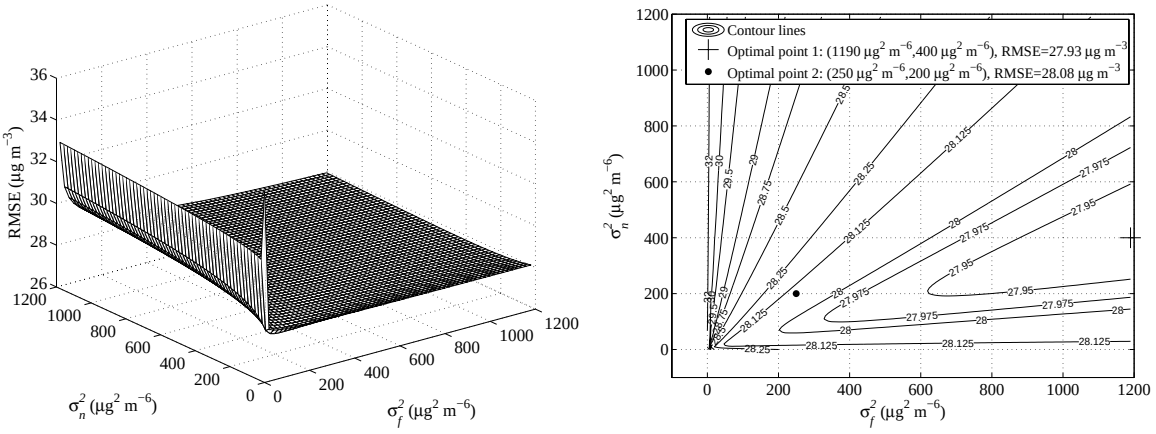
Figure 4. Natural logarithm of AIC for the TVAR(*p*) models



(a) (b)

Figure 5. (a) Posterior PDF and (b) contour plot of noise variances for TVAREX model

The TVAR(1) model and the TVAREX model are evaluated by using the data between 2003 and 2005. Figure 6a shows the root-mean-square error (RMSE) of the TVAR(1) model associated with different combinations of the process noise and the measurement noise variance. It is noted that the performance of the TVAR(1) model is more sensitive to the choice of the process noise variance than the choice of the measurement noise variance as the RMSE of the model increases rapidly for small values of process noise variance. This illustrates that the selection of noise parameters for Kalman filter is important. Figure 6b shows the corresponding contour plot of the RMSE. The optimal point 1 (+) represents the optimal estimates of the noise variances corresponding to the minimum RMSE, whereas the optimal point 2 (.) represents the optimal estimates of the noise variances by the Bayesian approach. It is noted that both points are located at the region where the model performance is insensitive to the choice of noise variances. The RMSE of the TVAR(1) model evaluated at the optimal noise variances by Bayesian approach is approximately 0.54% higher than the minimum RMSE. Although the RMSE evaluated at the optimal noise variances by Bayesian approach is slightly higher than the minimum RMSE, the standard deviation of the process noise corresponding to the minimum RMSE is about 86.90% of the root mean square of the daily averaged PM_{10} concentrations. This value is unreasonably high since it implies that the fluctuation of the daily averaged PM_{10} concentrations is mostly contributed by the process noise.

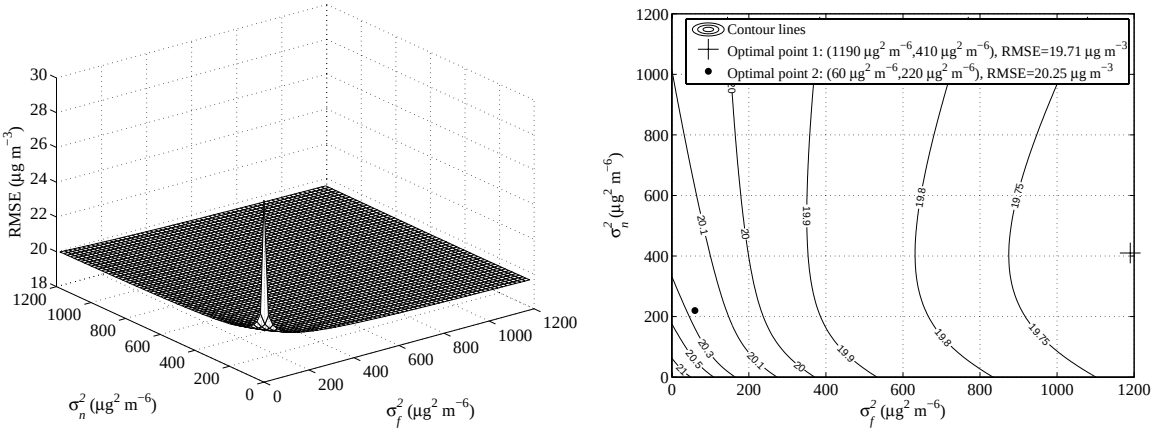


(a) (b)

Figure 6. (a) RMSE and (b) contour plot for the TVAR(1) model with different assumptions of process noise and measurement noise variances

Figure 7a shows the root-mean-square error (RMSE) of the TVAREX model for different process noise and measurement noise variances. It is noted that the performance of the TVAREX model is not sensitive to the choices of noise variances except at the region with small process noise and measurement noise variance. Figure 7b shows the corresponding contour plot. The optimal point 1 (+) represents the optimal estimates of the noise variances corresponding to the minimum RMSE, whereas the optimal point 2 (.) represents the optimal estimates of the noise variances by the Bayesian approach. Both are located at the region where the model performance is not sensitive to the choice of noise variances. It is surprising that the optimal process noise variance which associates with the minimum RMSE of the TVAREX model is the same as that of the TVAR(1) model. As mentioned earlier, the TVAREX model is more representative than the TVAR(*p*) model since the TVAREX model includes the meteorological conditions which reflect the influencing physics controlling the variation of PM_{10} concentrations in Macau. Therefore, the process noise of the TVAREX model is expected to be smaller than that of the TVAR(*p*) model. However, the process noise variance corresponding to the minimum RMSE is unreasonable and the optimal estimates of the noise variances by the Bayesian approach is more reliable even though its RMSE is slightly (~2.74%) higher than the minimum RMSE. The evidence illustrates the reliability of the Bayesian approach. Previously, the TVAR(1) model and the TVAREX model were compared from the perspectives of the noise variances. Now, both models are further evaluated by

comparing the general performance and their abilities to capture the pollution episodes. The optimal estimates of the noise variances were assumed for both models.



(a) (b)

Figure 7. (a) RMSE and (b) contour plot for the TVAREX model with different assumptions of process noise and measurement noise variances

Figure 8 shows the time series of measured daily averaged PM_{10} concentrations and the predictions by the TVAR(1) model between 2003 and 2005, respectively. The solid line represents the measurements whereas the dashed line represents the predictions. It is noted

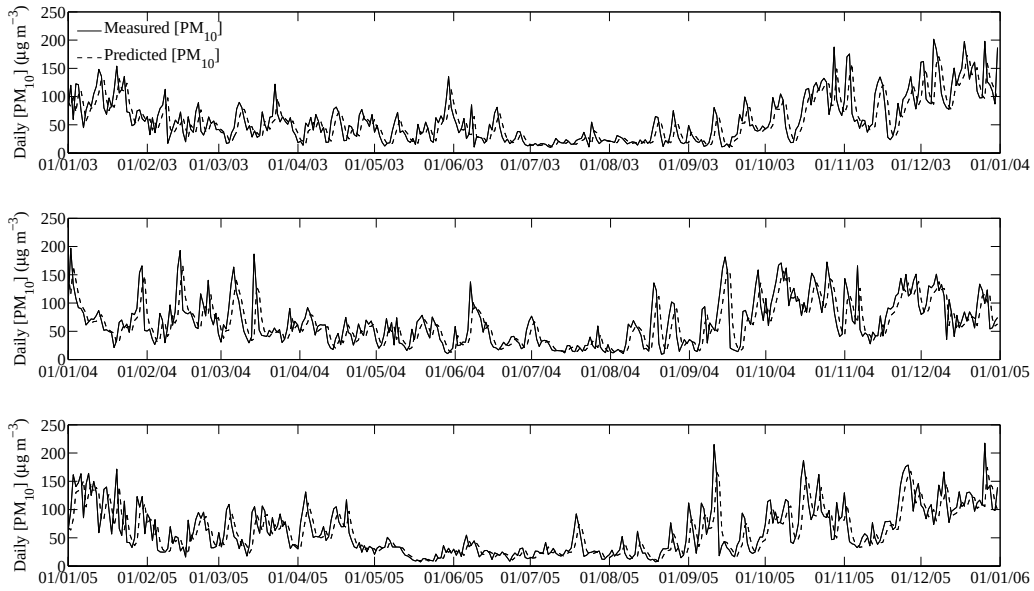


Figure 8. Measured daily averaged PM_{10} concentrations and predictions by TVAR(1) model between 2003 and 2005

that there is a time-delay problem associated with the TVAR(1) model, i.e., the trend of the predictions generally lags behind the trend of the measurements. The problem appears since the prediction is based solely on its own past history. Those influencing factors such as the dispersion condition on the day of prediction and the nature of replenishing air masses which can be continental or oceanic are treated as the unknown inputs. Therefore, the process noise is large with respect to the RMS of the signal and this causes the predicted signal to be delayed. Figure 9 shows the scatterplot of the TVAR(1) predicted daily averaged PM_{10} concentration against its measurement. A 45° straight line is also drawn on the figure for reference. A point falling on the 45° line implies a perfect match between the measured and predicted PM_{10} concentration. It is noted that a large portion of the points are lying close to

the 45° line, indicating that the model results are acceptable. However, some of the points are located far away from the line. As mentioned above, the predictions generally lag behind the measurements. Large prediction error is expected when there is large variation in the daily averaged PM_{10} concentrations. Therefore, those points lying far above or below the 45° line are associated with the onset and retreat of the PM_{10} episode. Figure 10 shows the time series of the measured daily averaged PM_{10} concentrations and the predictions by the TVAREX model. It is found that the time delay problem is generally resolved. Figure 11 shows the scatterplot of the predicted daily averaged PM_{10} concentration by the TVAREX model

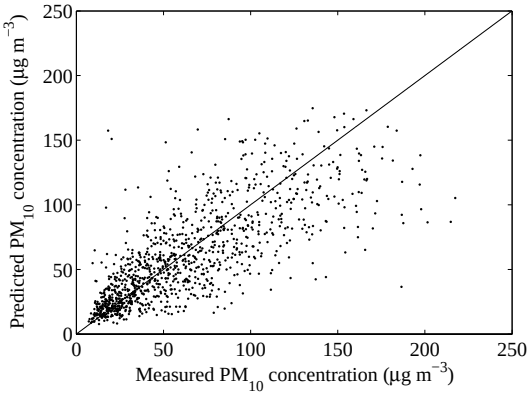


Figure 9. Predicted PM_{10} concentrations by TVAR(1) model against its measurements

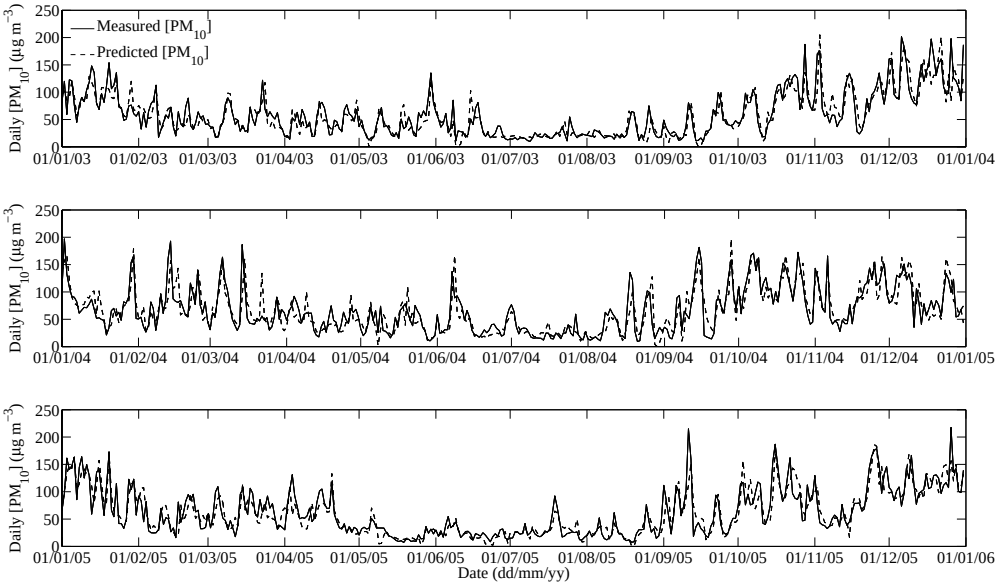


Figure 10. Measured daily averaged PM_{10} concentrations and predictions by TVAREX model between 2003 and 2005

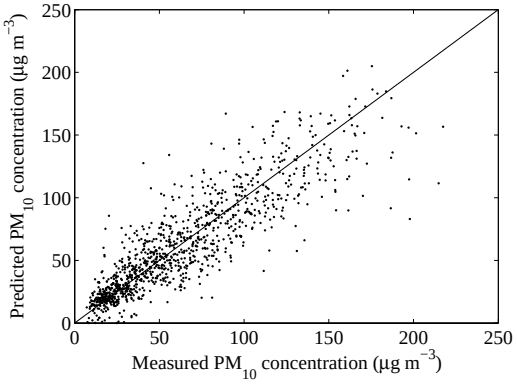


Figure 11. Predicted PM_{10} concentrations by TVAREX model against its measurements

against its measurements. In comparison to the scatterplot generated from results of the TVAR(1) model, it is found that the points become more concentrated around the 45° line. It echoes the observation of the improvement in the time-delay problem as shown in Figure 10 as well as the reduction of the process noise variance commented before.

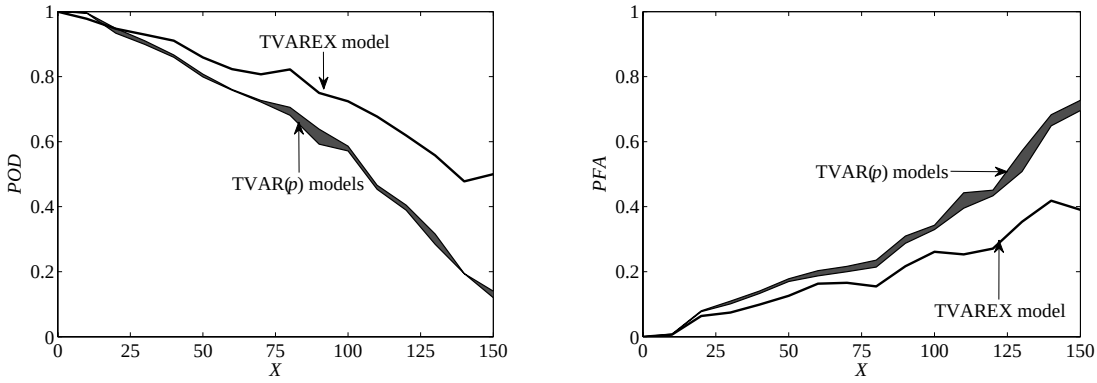
To further compare their performances of capturing the pollution episodes, two performance indicators, namely the probability of detection (*POD*) and the probability of false alarm (*PFA*) are calculated for reference. The probability of detection is defined as follows:

$$POD = P\left(\hat{x}_{k|k-1} \geq X \mid z_k \geq X\right) \tag{19}$$

It is equal to the probability that the model can produce a prediction which is greater than or equal to the threshold *X* when it is given that the daily averaged [PM_{10}] is greater than or equal to the threshold. The probability of false alarm (*PFA*) is defined as follows:

$$PFA = P\left(z_k < X \mid \hat{x}_{k|k-1} \geq X\right) \tag{20}$$

It is equal to the probability that the daily averaged [PM_{10}] is below the threshold *X* when it is given that the model produces a prediction which is greater than or equal to the threshold. Figures 12(a) and 12(b) show the *POD* and *PFA* when the threshold X takes on different values between 0 and 150. It is noted that the *POD* values of the TVAREX model are mostly higher than the *POD* envelope of the TVAR(*p*) models, while the *PFA* values of the TVAREX



(a) (b)

Figure 12. (a) *POD*, (b) *PFA* for TVAR(*p*) models and TVAREX model

model are far below the *PFA* envelope. Therefore, the TVAREX model is more efficient than the TVAR(*p*) models in capturing the episode conditions with lower frequency of false alarm. In fact, the TVAREX model is more efficient in capturing the episodic conditions since it was shown in the previous literatures (Chang *et al.*, 2007; Feng *et al.*, 2007; Lee and Hills, 2003; Lee and Savtchenko, 2006) that the high PM_{10} concentrations occurring in Macau and nearby cities of the Pearl River Delta were due to fine weather (no rainfall), poor dispersion condition, and the northerly air masses, and those factors were taken into consideration by the exogenous inputs, thus reducing the process noise of the model.

5. CONCLUSION

The Bayesian approach was proposed to find the optimal estimates of noise parameters for the Kalman filter based air quality prediction system. By optimizing the objective function with respect to the noise variances, the Bayesian methodology allows the most probable values of noise variances to be obtained and the associated uncertainties to be quantified. Throughout the case study, the Bayesian approach was demonstrated to be capable to estimate the most probable noise variances of the Kalman filter based TVAR(*p*) model and TVAREX model for the prediction of daily averaged PM_{10} concentrations in Macau between 2001 and 2002. It was found that the estimated process noise variance of the TVAREX model is less than that of the TVAR(1) model since the TVAREX model reflects more influencing physics which controls the variation of daily averaged PM_{10} concentrations in Macau. By further using data between 2003 and 2005, the choice of the noise variances was demonstrated to affect the

performance, which was indicated by the root-mean-squared error, of the TVAR(*p*) model and the TVAREX model. In addition, it was found that the optimal estimates of noise variances obtained by Bayesian approach for both models were located in the region where the model performance is not sensitive to the choice of noise variances. The Bayesian approach was demonstrated to provide more reasonable estimates of noise variances compared to the noise variances found by simply minimizing the root-mean-squared prediction error of the model. The evidences illustrated the reliability of the approach. By comparing the optimized TVAREX model and the TVAR(*p*) models in predicting the daily averaged PM_{10} concentrations between 2003 and 2005, it was found that the TVAREX model outperformed the TVAR(*p*) models in terms of the general performance and the episode capturing ability.

6. ACKNOWLEDGEMENTS

The financial support from the Fundo para o Desenvolvimento das Ciências e da Tecnologia (FDCT) under grant 052/2005/A and the Research Committee of University of Macau under grant RG-UL/07-08S/Y0/MKM/FST is gratefully acknowledged. The SMG of Macau is thanked for supplying the data.

REFERENCES

Akaike, H. (1976), A new look at the statistical identification model, *IEEE Transactions on Automatic Control*, **19**, 716-723.

Abdel-Aziz, A. and Frey, H.C. (2003) Development of hourly probabilistic utility NO_x emission inventories using time series techniques: Part I-univariate approach, *Atmospheric Environment*, **37**, 5379-5389.

Beck J.L. and Katafygiotis L.S. (1998), Updating models and their uncertainties. I: Bayesian statistical framework, *Journal of Engineering Mechanics*, **124**, 455-461.

Choi I.C., Mok K.M. and Tam S.C. (2002), Solving harmonic sea-level model with Kalman Filter: a Macau case study. *Carbonate Beaches 2000*, December 5-8 2000, Key Largo, Florida, USA, 38-52.

Chang S.W., Mok K.M. and Yuen K.V. (2007), Association of PM_{10} pollution episodes with the meteorological conditions in Macau. *Proceedings of the 10th International Conference on Environmental Science and Technology*, Kos Island, Greece, 5-7 September 2007, pp. 90-95.

Feng Y., Wang A., Wu D. and Xu X. (2007), The influence of tropical cyclone Melor on PM_{10} concentrations during an aerosol episode over the Pearl River Delta region of China: Numerical modeling versus observational analysis, *Atmospheric Environment*, **41**, 4349-4365.

Hamilton R., Braun B., Koopman B. and Svoronos S.A. (2008), Estimation of nitrate reductase enzyme parameters in activated sludge using an extended Kalman filter algorithm, *Water Research*, **42**, 1889-1896.

Hoi K.I., Yuen K.V. and Mok K.M. (2008), Kalman filter based prediction system for wintertime PM_{10} concentrations in Macau, *Global NEST Journal*, **10**, 140-150.

Hoi K.I., Yuen K.V. and Mok K.M. (2009), Prediction of daily averaged PM_{10} concentrations by statistical time-varying model, *Atmospheric Environment*, **43**, 2579-2581..

Kalman R.E. and Bucy R.S. (1961), New results in linear filtering and prediction theory. *Transactions of ASME, Journal of Basic Engineering*, **83**, 95-108.

Konovalov I.B., Beekmann M., Meleux F., Dutot A. and Foret G. (2009), Combining deterministic and statistical approaches for PM_{10} forecasting in Europe, *Atmospheric Environment*, **43**, 6425-6434.

Lee Y.C. and Hills P.R. (2003), Cool season pollution episodes in Hong Kong, 1996-2002, *Atmospheric Environment*, **37**, 2927-2939.

Lee Y.C. and Savtchenko A. (2006), Relationship between air pollution in Hong Kong and in the Pearl River Delta Region of South China in 2003 and 2004: An analysis, *Journal of Applied Meteorology and Climatology*, **45**, 269-282.

Monson, B.A. (2009) Trend reversal of mercury concentrations in Piscivorous fish from Minnesota Lakes: 1982-2006, *Environmental Science and Technology*, **43**, 1750-1755.

Mok K.M., Hoi K.I. (2005), Effect of meteorological conditions on PM_{10} concentrations: A study in Macau, *Environmental Monitoring and Assessment*, **102**, 201-223.

Peng R.D., Dominici, F. and Louis T.A. (2006), Model choice in time series studies of air pollution and mortality, *Journal of Royal Statistical Society: Series A (Statistics in Society)*, **169**, 179-203.

Politis M., Pilinis C. and Lekkas T.D. (2008). Ultrafine particles (UEP) and health effects. Dangerous. Like no other PM? Review and analysis, *Global NEST Journal*, **10**, 439-452.

Shan W., Lu H., Huo S., Huang Z. and You L. (2009), Analysis of a high PM_{10} episode observed at a coastal site nearby Shanghai, China, *Environmental Monitoring and Assessment*, doi: 10.1007/s10661-009-0838-4.

Trudinger C.M., Enting I.G. and Rayner P.J. (2002), Kalman filter analysis of ice core data 1. Method development and testing the statistics, *Journal of Geophysical Research - Atmospheres*, **107**, doi:10.1029/2001JD001111.

Trudinger C.M., Raupach M.R., Rayner P.J. and Enting I.G. (2008), Using the Kalman filter for parameter estimation in biogeochemical models, *Environmetrics*, **19**, 849-870.

Yuen K. V., Beck J.L. and Katafygiotis L.S. (2002), Probabilistic Approach for Modal Updating Using Nonstationary Noisy Response Measurements Only, *Earthquake Engineering and Structural Dynamics,* **31**, 1007-1023.

Yeh H.D. and Huang Y.C. (2005), Parameter estimation for leaky aquifers using the extended Kalman filter, and considering model and data measurement uncertainties, *Journal of Hydrology*, **302**, 28-45.

Yuen K.V., Hoi K.I. and Mok K.M. (2007), Selection of noise parameters for Kalman filter, *Journal of Earthquake Engineering and Engineering Vibration*, **6**, 49-56.

Yuen K.V. and Katafygiotis L.S. (2001), Bayesian time-domain approach for model updating using ambient data, *Probabilistic Engineering Mechanics*, **16**, 219-231.

Yuen K.V. and Mu H.Q. (2009), Seismic hazard analysis from a Bayesian perspective, *Journal of Disaster Advances*, **2**, 36-42.

Yan W.M., Yuen K.V. and Yoon G.L. (2009), Bayesian probabilistic approach for the correlations of compressibility index for marine clays, *Journal of Geotechnical and Geoenvironmental Engineering (ASCE)*, **135**, 1932-1940.

Yuen K.V. (2010), Bayesian methods for structural dynamics and civil engineering. John Wiley and Sons.