Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computational Statistics and Data Analysis 57 (2013) 589–599

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Modeling respiratory illnesses with change point: A lesson from the SARS epidemic in Hong Kong

Heung Wong [a,*], Quanxi Shao [b], Wai-cheung Ip [a]

[a] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

[b] CSIRO Mathematics Informatics and Statistics, Private Bag No. 5, Wembley, WA 6913, Australia

ARTICLE INFO

Article history:
Received 12 January 2012
Received in revised form 24 May 2012
Accepted 31 July 2012
Available online 3 August 2012

Keywords:
Air pollution
Change point problem
Multiple index model
Non-parametric regression
Public health
Respiratory illness
Severe acute respiratory syndrome (SARS)

ABSTRACT

It is generally agreed that respiratory disease is closely related to ambient air quality and weather conditions. Besides, hygiene related factors such as the public health measures by the government and possible personal awareness in the community can also affect the spread of infectious respiratory diseases. However, there is no quantitative support for this conclusion, because of lack of quality data. The severe acute respiratory syndrome (or SARS) outbreak in 2003 triggered strict public health measures and personal awareness in the prevention of infectious respiratory diseases, providing us an opportunity to quantify the impact of hygiene related factors in the spread of the disease. In this paper, we model the number of the respiratory illnesses by a semiparametric model which models the environmental and weather impacts using a multiple index model and the impact of other public health measures and possible personal awareness using a growth curve with jump. Using data from Hong Kong, we found that public health measures contributed to about 39% of reduction in the number of respiratory illnesses during the SARS period. However, the impact of hygienically related factors eventually fades as time passes. The results provide indirect quantitative support to the usefulness of governmental campaigns to arouse the awareness of the public in staying away from transmission of respiratory diseases during the full outbreak of the disease. The results also show the fast fading of alertness of Hong Kong people towards the epidemic. Furthermore, our model also offers a way to model the impacts of environmental factors on respiratory diseases, when the data contains the effect of human intervention, by introducing the change point and growth curve to remove such an effect.

1. Introduction

It is widely agreed that the exposure to ambient air pollution may cause serious respiratory illnesses (Katsouyanni et al., 1996; Navrud, 2001; Aunan and Pan, 2004; Neuberger et al., 2004; Scoggins et al., 2004; Jalaludin et al., 2004; Maynard, 2004; Lee et al., 2009) and that weather conditions may also contribute to the seriousness (Smoyer et al., 2000). Several statistical models have been used to assess the impacts of air quality and weather factors on respiratory illnesses, including simple statistical correlation analysis, regression method, meta-analysis, spatial analysis and nonparametric statistical models such as the generalized additive model (GAM). However, quantifying the effects of various pollutants as well as weather conditions is always a difficult task due to the high unknown nonlinearities on the impact of these environmental and weather factors on the onset of the respiratory illnesses and possible interactions amongst these factors. Nonparametric

* Corresponding author.
E-mail addresses: mathwong@polyu.edu.hk (H. Wong), quanxi.shao@csiro.au (Q. Shao).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.07.029

regression is a powerful tool to model unknown functional relationships. With the consideration of a possible incubative period and/or cumulative effects, we need to consider the lags of potential factors that result in a large number of covariates in the regression. In order to avoid the so-called 'curse of dimensions' referring to the inaccuracy when the multivariate function is estimated nonparametrically, some statistical tools have been developed so as to let the model determine the best combinations in the sense of their contribution to the regression. Representative models include a varying/functional coefficient model (Chen and Tsay, 1993; Hastie and Tibshirani, 1993), regression graphics (Cook, 1998), sliced inverse regression (Li, 1991), single index model (Härdle and Stoker, 1989) and multiple index model (Xia et al., 2002). In this paper, we will adopt the multiple index model for our data, following Xia et al. (2002) and Shao et al. (2010).

Further complexity in modeling the number of respiratory illnesses is made by the effect due to hygienic related factors such as the public health measures by government and possible personal awareness in the community because they can also affect the spread of infectious respiratory diseases. Unfortunately, the effect by hygiene related factors has not been considered in literature, partly because of the lack of quality information which in turn makes it difficult to quantitatively measure and support the effect of hygiene related factors on respiratory illnesses.

From the data collected in Hong Kong, we observed an unusual change in the number of respiratory illnesses in early 2003, while all other air quality and weather data were relatively stable. The change coincides with the outbreak of severe acute respiratory syndrome (or SARS). The background and severity of this disease is briefed below. SARS is a respiratory disease in humans and is caused by the SARS-associated coronavirus (SARS–CoV). It was first reported in Asia in February 2003. Over the following few months, the disease spread to more than two dozen countries in North America, South America, Europe and Asia before the SARS global outbreak of 2003 was contained. According to the World Health Organization's (WHO) concluding report on 21 April 2004 (see http://www.who.int/csr/sars/country/table2004_04_21/en/index.html), this major outbreak consisted of 8096 known cases with 774 deaths (a mortality rate of 9.6%). Hong Kong was the most seriously affected area in terms of population size (6.8 million) and land area (1100 km^2). There were 1755 cases in total (1 case per 4000 population with 977 females and 778 males) with 299 deaths (a mortality rate of 17%). The source of infection was traced back to an infected Guangzhou professor who arrived in Hong Kong in late February 2003. During his stay in a hotel, the infection was spread to other guests and visitors, subsequently triggering off a chain of outbreaks in Hong Kong, Singapore, Canada and Vietnam. A detailed history and public health experiences can be found in Hong Kong's SARS Expert Committee Report (2003) and Cheng (2003). The official announcement of the SARS epidemic was on 8 March 2003. The last case was reported on 2 June 2003. After an observation period of 20 days, Hong Kong was removed from the WHO's list of areas with recent local transmission of SARS on 23 June 2003. The WHO's research concluded that SARS was spread by close person-to-person contact. SARS–CoV is thought to be transmitted most readily by respiratory droplets of infectious persons when droplets from the cough or sneeze of the infectious persons are propelled a short distance (generally up to 3 ft) through the air and deposited on the mucous membranes of the mouth, nose, or eyes of persons who are nearby. The virus also can be transmitted when a person touches a surface or object contaminated with infectious droplets and then touches his/her mouth, nose, or eye(s). In addition, it is possible that the SARS–CoV might spread more broadly through the air (airborne spread) or by other ways that are not known.

For infectious diseases, efficient prevention before major outbreak and treatment during outbreak are primary ways of disease control. However, community- or population-based health protection should be another important way to stop the spread of such infectious diseases but has not been addressed seriously before. During the SARS epidemic, Hong Kong's Department of Health introduced a large-scale health awareness program to publicize the importance of personal and environmental hygiene. The actions included wearing face masks in public by everyone, washing hands regularly (especially before touching mouth and eyes and having food). Official actions included cleaning surfaces in public areas such as handles and buttons with disinfectants regularly, and temperature checks at airports and other immigration ports. An educational campaign for personal and environmental hygiene was widely promoted by TV advertisements and posters in public areas (Leisure Cultural Services Department, 2003; Housing Authority, 2003; Leung et al., 2003). Such public health measures showed great success in controlling the disease. Some measures in public areas have been kept in force until now.

Therefore, the data provide us an opportunity to model the effect of strict hygiene related factors on respiratory illnesses. For the purpose of public health, it is very important and interesting to see how effective these public health measures are. By developing a method to handle data with sudden change(s) in modeling, we can also assess the change of the number of illness before and after the SARS epidemic. The results can be used as an indirect support of the effectiveness of such efforts by governmental and community campaigns.

The paper is organized as follows. Section 2 is a brief description of the data set. Model and estimation procedures are given in Section 3, followed by a statistical test to quantify the efficacy of community-based hygiene campaign. The results are given in Section 4, and Section 5 concludes the paper.

2. Data set

Hong Kong became a Special Administrative Region of the People's Republic of China on July 1, 1997, after a century and a half of British administration. It is located in the south of Mainland China with a population of 6.8 million and an area of 1103 km^2 covering Hong Kong Island and the more rural New Territories. Its climate is sub-tropical with temperatures between 10 °C in winter and 33 °C in summer.

Fig. 1. Map of Hong Kong.

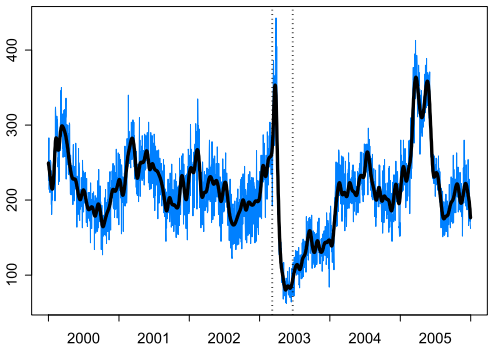

Fig. 2. Time series plot of the number of respiratory illnesses for Hong Kong. The trend lines (thick line) are provided by a smoothing spline. The two dotted vertical lines indicate the start and end of the SARS period.

Air pollution data were obtained from the Environmental Protection Department of HKSAR (www.epd.gov.hk). These include the daily average level of sulfur dioxide (SO_2, μgm^{-3}), respirable suspended particulates (RSP μgm^{-3}), nitrogen oxides (NO_x, μgm^{-3}), nitrogen dioxide (NO_2, μgm^{-3}), ozone (OZ, μgm^{-3}). Weather data were obtained from Hong Kong Observatory, including temperature (Temp, degree Celsius) and relative humidity (Humi). The same variables were used by Xia et al. (2002) and Shao et al. (2010). Daily admission figures to public hospitals in Hong Kong due to respiratory illnesses were obtained from the Hospital Authority of HKSAR. The cases included all respiratory illnesses. These figures can be regarded as the daily number of illnesses due to respiratory diseases. Although the Air Pollution Index is usually reported to the public, we do not use it in this study as it is secondary data derived from the highest indices of several key pollutants.

Daily data over six years were collected for this study, from 01/01/2000 to 31/12/2005. The numbers of respiratory illnesses are plotted in Figs. 1 and 2, from which a sudden drop in the number of respiratory illnesses can be seen in early 2003 during the outbreak of SARS. The data was divided into two parts: Pre-SARS period refers to the data before 8 March 2003 and post-SARS after 23 June 2003 (inclusive). We do not use the data during the SARS epidemic as it is a transition period for the change. The summary statistics for individual variables are given in Table 1 and box plots are given in Fig. 3. We can see that the temperature and humidity are very similar during the pre- and post-SARS period. However, SO_2, RSP and Ozone increase for all regions while NOs and NO_2 are relatively stable. In contrast, the number of illnesses decreases.

Both Xia et al. (2002) and Shao et al. (2010) considered the weekly pattern of the respiratory illnesses. It is possible that the weekly pattern in the number of the respiratory illnesses confounds with similar patterns in air pollution. To check if it is the case, the box-plots of all the variables for each day of the week is given in Fig. 4, from which it can be seen that only the number of admitted patients shows some weekly pattern, while all the air pollutants/weather variables show no weekly pattern (in fact, they remain quite constant over the day of the week). Therefore, the weekly pattern in the number of the respiratory illnesses could be modeled together with the pollutants/weather variables.

Table 1
Summary statistics for individual variables.

	Cases	SO_2	NO_x	NO_2	RSP	Ozone	Temp	Humi
Before SARS								
Minimum	122	2.84	23.04	15.65	13.98	3.10	7.25	41.25
1st quartile	193	9.76	78.47	44.79	31.48	19.19	19.12	73.38
Median	218	14.29	101.40	55.23	45.26	28.25	23.82	79.25
Mean	220	16.95	115.80	57.63	51.04	31.80	22.96	78.01
3rd quartile	246	20.93	139.20	67.71	65.69	41.94	27.32	84.50
Maximum	350	106.10	441.90	162.10	172.60	99.63	30.48	97.25
STD	38.91	11.14	53.49	19.32	24.89	15.70	5.03	9.72
After SARS								
Minimum	2	2.88	26.95	17.66	13.42	5.56	8.60	19.00
1st quartile	166	12.68	74.30	42.01	31.32	18.69	19.46	71.88
Median	206	18.35	98.70	55.63	52.86	31.56	25.08	78.50
Mean	207	22.60	107.80	58.23	58.07	35.45	23.54	76.92
3rd quartile	238	26.22	127.00	71.80	78.71	48.37	27.94	84.00
Maximum	413	140.50	419.00	160.90	207.30	123.50	31.90	100.00
STD	64.11	16.43	47.63	22.11	31.32	19.89	5.23	11.21

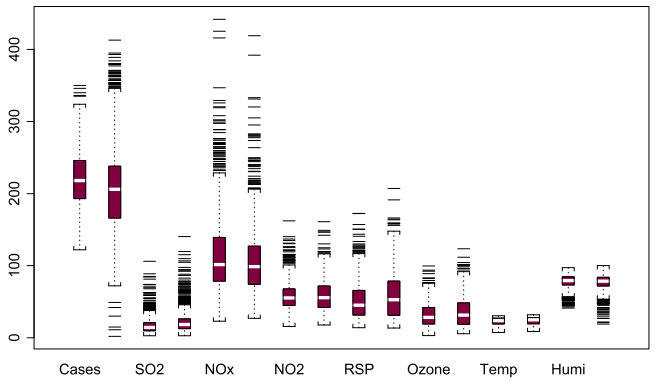

Fig. 3. Box-plot to compare daily values of variables before and after the SARS epidemic. Note: For each item on the horizontal axis, the two box-plots refer to values before and after SARS.

3. Methods

Let Y be our response variable, which is the number of daily admission to a regional hospital due to respiratory illnesses, and $\mathbf{X} = (X_1, \ldots, X_p)$ be the row-vector of covariates (with possible lags) affecting the response variable. The potential covariates in our study include (SO_2, NO_x, RSP, OZ, Temp, Humi). Let t_{start} (=8 March 2003) denote the full outbreak of the SARS epidemic and t_{end} (23 June 2003) the end of the SARS epidemic.

In order to evaluate the effectiveness of public health measures and air quality on the reduction of respiratory illnesses, with consideration of weekly effect, we consider the following candidate models:

$$Y_t = \lambda_t + \sum_{k=1}^{6} \alpha_k D_{k,t} + g(\mathbf{X}) + \varepsilon_t \tag{1}$$

with

$$\lambda_t = \begin{cases} 0, & \text{if } t \le t_{start}, \\ \xi_1 + \dfrac{\xi_3}{1 + \exp\{-\xi_2(t - \xi_4)\}} & \text{if } t \ge t_{end}. \end{cases} \tag{2}$$

The individual terms in the model are considered as follows. (i) ε_t is a zero-mean, constant variance error term. (ii) λ_t measures the impact of public health measures with ξ_1 and ξ_3 being real values representing the lower and upper asymptotes, respectively, ξ_2 being a positive value representing the growth rate and ξ_4 being a real value representing the time of maximum growth. That is, the effect of public health measures as measured by the shift in the total number of illnesses. By noting the magnitude of impact decreases eventually over time due to many factors including behavioral and memory dynamics, the growth curves are widely used in psychological research to model the changes over time (Mcardle and Nesselroade, 2002). The logistic growth curve is a special case of the generalized Richard growth curve (Richard, 1959). A negative value of ξ_1 indicates a reduction of the number of illnesses, meaning an effective impact of public health measures.

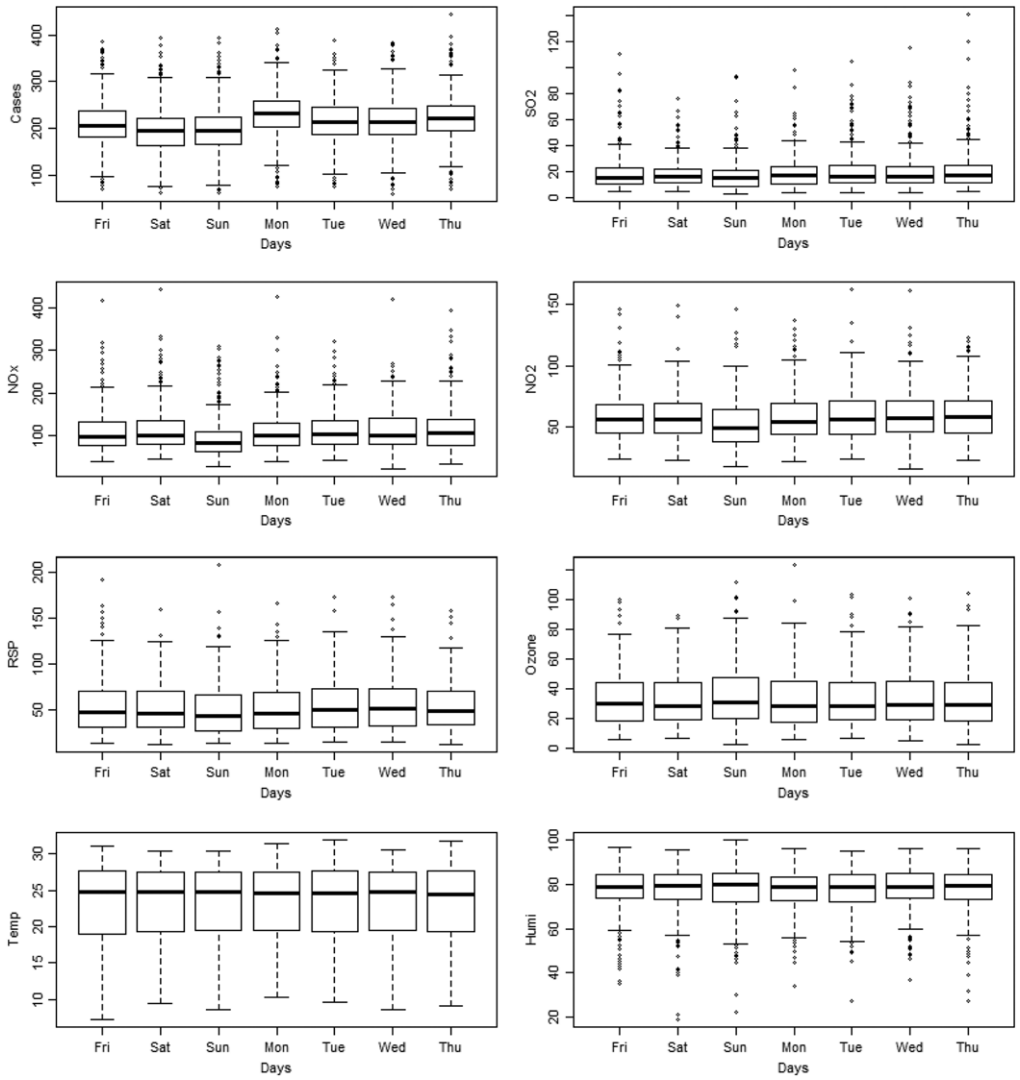

Fig. 4. Box-plot to compare daily values of variables for each day of the week.

A positive value of ξ_3 means that the impact of public health measures will fade and vanish over time. We do not use the full Richard curve because it involves complicated computing due to a limit or embedded cases (Shao, 2002) and the special case here is good enough to fit the data. Bear in mind that we do not use the data during the SARS epidemic as it is highly unstable. (iii) Due to the hospital system in Hong Kong where only emergency departments are in normal service during weekends, the effect of the day-of-the-week is indicated in previous studies (Xia et al., 2002; Xia and Tong, 2006). Here the dummy variables are defined as

$$D_{k,t} = \begin{cases} 1, & t \text{ is the } k\text{th day in the week,} \\ 0 & \text{Otherwise,} \end{cases}$$

with day 1 to 6 referring to Saturday to Thursday respectively and Friday as the reference day. The second term measures the effect of the day-of-the-week which is also important (Dominici et al., 2000; Forster De and Solomon, 2003; Bruckmann and Wichmann-Fiebig, 1997; Xia and Tong, 2006). (iv) The effect of environmental and weather factors on the illnesses is characterized by a Multiple Index Model defined as

$$g(\mathbf{X}) = g(\beta_{11}X_1 + \cdots + \beta_{p1}X_p, \ldots, \beta_{1M}X_1 + \cdots + \beta_{pM}X_p) = g(\mathbf{XB}) \tag{3}$$

with $M < p$ and $E(\varepsilon|X) = 0$ almost surely. Here $g(\ )$ is an unknown smooth function and $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)$ is the coefficient matrix with $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots \beta_{pj})^T$ being the jth direction or components, $()^T$denotes the transpose of a vector or matrix. Without loss of generality, the standardized version of $\mathbf{X} = (X_1, \ldots, X_p)$ is used in the model, i.e. $E(X_j) = 0$ and $Var(X_j) = 1$ $(j = 1, \ldots, p)$. To ensure the uniqueness of the coefficient matrix, one should impose some identification condition on $\mathbf{B}$. Usually it is assumed that $\mathbf{B}^T\mathbf{B} = I_{M\times M}$, that is, $\mathbf{B}$ is an orthogonal matrix.

Model estimation

Xia et al. (2002) used the minimum average variance estimation technique (MAVE) to estimate $\mathbf{B}$ in the multiple index model and then cross validation (CV) to determine D. A brief outline of the method is given below.

For a fixed D, the minimum average variance estimation is to find $\mathbf{B}$ by

$$\min_{\mathbf{B}} \left[E\left\{Y - E(Y|\mathbf{XB})\right\}^2\right] \tag{4}$$

with the constraint that $\mathbf{B}^T\mathbf{B} = I_{M\times M}$. The name MAVE is given in light of the following equation

$$E\left\{Y - E(Y|\mathbf{XB})\right\}^2 = E\left[E\left\{Y - E(Y|\mathbf{XB})\right\}^2 |\mathbf{XB}\right] = \sigma^2_{\mathbf{XB}}, \tag{5}$$

where the last term is the conditional variance with a given $\mathbf{XB}$. There are two conditional expectations we need to deal with. Assume that there are n observations Y_i with corresponding covariates $\mathbf{X}_i = (X_{1,i}, \ldots, X_{p,i})$ $(i = 1,\ 2, \ldots, n)$. For any given $\boldsymbol{X}_0 = (X_{1,0}, \ldots, X_{p,0})$, the inner expectation at $\boldsymbol{X}_0$ can be approximated by a local linear expansion of $E(Y_i|\mathbf{X}_i\mathbf{B})$. That is

$$E(Y_i|\mathbf{X}_i\mathbf{B}) \approx \lambda_{t_i} + \sum_{k=1}^{6} \alpha_k D_{k,t_i} + a + (\mathbf{X}_i - \mathbf{X}_0)\mathbf{B}b^T, \tag{6}$$

where $a = g(\mathbf{X}_0\mathbf{B})$ and $b = \partial g(\mathbf{V})/\partial\mathbf{V}$ evaluated at $\mathbf{V} = \mathbf{X}_0\mathbf{B}$. Note that the outer conditional expectation is in fact the conditional variance with a given $\mathbf{XB}$. Following the idea of local linear smooth estimation, we can estimate $\sigma^2_{\mathbf{XB}}$ by exploiting the approximation

$$\sum_{i=1}^{n} w_{i0}\left\{Y_i - E(Y_i|\mathbf{X}_i\mathbf{B})\right\}^2 \approx \sum_{i=1}^{n} w_{i0}\left[Y_i - \left\{\lambda_{t_i} + \sum_{k=1}^{6}\alpha_k D_{k,t_i} + +a + (\mathbf{X}_i - \mathbf{X}_0)\mathbf{B}b^T\right\}\right]^2, \tag{7}$$

where $w_{i0} \geq 0$ are some weights with $\sum_{i=1}^{n} w_{i0} = 1$. Usually kernel smoothing is employed by defining the weights as

$$w_{i0} = K_h\left\{(\mathbf{X}_i - \mathbf{X}_0)\mathbf{B}\right\} \Big/ \sum_{l=1}^{n} K_h\left\{(\mathbf{X}_l - \mathbf{X}_0)\mathbf{B}\right\}, \tag{8}$$

where $K_h(\bullet) = h^D K_h(\bullet/h)$ with D being the dimension of kernel function $K(\bullet)$ and h the bandwidth. On the basis of the above approximation of conditional expectations, the MAVE estimate of the coefficient matrix $\mathbf{B}$ is obtained by solving the minimization

$$\min_{\mathbf{B}:\mathbf{B}^T\mathbf{B}=I}\left\{\sum_{j=1}^{n}\hat{\sigma}^2_{\mathbf{X}_j\mathbf{B}}\right\} = \min_{\substack{\mathbf{B}:\mathbf{B}^T\mathbf{B}=I\\ a_j, b_j, j=1,\ldots,n}}\left(\sum_{j=1}^{n}\sum_{i=1}^{n} w_{ij}\left[Y_i - \left\{\lambda_i + \sum_{k=1}^{6}\alpha_k D_{k,t_i} + a_j + (\mathbf{X}_i - \mathbf{X}_j)\mathbf{B}b_j^T\right\}\right]^2\right), \tag{9}$$

where $b_j = (b_{j1}, \ldots, b_{jD})$ is a row vector. Note that the constraint $\mathbf{B}^T\mathbf{B} = I$ makes the above minimization more difficult. Furthermore, the weights depend on the choice of coefficient matrix $\mathbf{B}$. We can use an iteration procedure to find the solution. A suggested stop rule for $\mathbf{B}$ is $m(\mathbf{B}^{(t-1)}, \mathbf{B}^{(t)}) = \left\|\left\{I - \mathbf{B}^{(t-1)}(\mathbf{B}^{(t-1)})^T\right\}\mathbf{B}^{(t)}\right\| < \delta$ and $\left|\lambda_i^{(t)} - \lambda_i^{(t-1)}\right| < \delta$, where δ is a predetermined tolerance value.

To determine the number M of efficient dimensions, cross validation is used by Xia et al. (2002) as

$$CV(m) = n^{-1}\sum_{j=1}^{n}\left(y_j - \hat{a}_{0d,j}\right)^2, \tag{10}$$

where

$$\hat{a}_{0d,j} = \hat{\lambda}_{t_j} + \sum_{k=1}^{6}\alpha_k D_{k,t_j} + \sum_{i=1,i\neq j}^{n} K_{h_d}^{(i,j)}(y_i - \hat{\lambda}_i) \Big/ \sum_{i=1,i\neq j}^{n} K_{h_d}^{(i,j)}, \tag{11}$$

with $K_{h_d}^{(i,j)} = K_{h_d}\left\{(\mathbf{X}_i - \mathbf{X}_j)\mathbf{B}\right\}$. The d with the smallest $CV(m)$ value is the estimated dimension.

Computing issues

Note that the conditional expectation in (7) and (8) does not involve the time t of observation. We can apply the algorithm developed by Xia et al. (2002) for the multiple index model but with more steps for the estimation of other parameters. The algorithm has the following steps.

Step 1. Initialize the estimate of the coefficient $\mathbf{B}$ in the multiple Index part $g(\mathbf{XB})$ as $\mathbf{B} = \mathbf{I}$.

Step 2. Initialize the parameter estimates of $\xi = (\xi_1, \xi_2, \xi_3, \xi_4)$ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_6)$ as $\xi^{[0]}$ and $\alpha^{[0]}$. Good estimates can be the least squares estimates of the model $Y = \lambda_t + \sum_{k=1}^{6}\alpha_k D_{k,t}$.

Step 3. In the mth iteration ($m = 1, 2, \ldots$), compute

$$Y^* = Y - \lambda_t(\xi^{[m-1]}) - \sum_{k=1}^{6} \alpha_k^{[m-1]} D_{k,t},$$

and update the coefficient $\mathbf{B}$ of the multiple index part $g(\mathbf{XB})$ by MAVE with model $Y^* = g(\mathbf{XB}^{[m]})$. The updated estimate is denoted as $\mathbf{B}^{[m]}$.

Step 4. Compute

$$Y^{**} = Y - g(\mathbf{XB}^{[m]}),$$

and update the parameter estimates of ξ and α by the least squares method. The updated estimate is denoted as $\xi^{[m]}$ and $\alpha^{[m]}$.

Step 5. Check the differences between the current estimates and previous estimates. If the maximum difference is less than a predetermined tolerance value (which is 0.001 in this study), then stop the iteration. Otherwise, return to Step 3.

Based on the above five steps, estimation for the parameters and the unknown function in model (1) (denoted as $\hat{\xi}$, $\hat{\alpha}$ and $\hat{g}(\cdot)$ respectively) can be obtained. To construct the confidence bounds of each parameter in the growth curve and weekly pattern, we employ the bootstrap method (see, for example Efron and Tibshirani, 1993) as below.

Step I. Randomly resample the model residuals $\{\hat{\varepsilon}_t\}$ $t = 1, \ldots, T$, where

$$\hat{\varepsilon}_t = Y_t - \lambda_t(\hat{\xi}) - \sum_{k=1}^{6} \hat{\alpha}_k D_{kt} - \hat{g}(X\hat{B})$$

with replacement to form a new residual series ε_t^*. Note that the superscript '*' denotes something calculated from bootstrap resampling.

Step II. Add ε_t^* to the fitted model output $\hat{Y}_t$ to form the new values $Y_t^* = \hat{Y}_t + \varepsilon_t^*$. Because the new residual series is generated with replacement, any model residual $\hat{\varepsilon}_t$ can be sampled more than one time or not sampled at all.

Step III. Calibrate the bootstrap sample $\{\mathbf{X}_t, Y_t^*\}$ $(t = 1, \ldots, T)$ to obtain a bootstrap estimator $(\hat{\xi}^*, \hat{\alpha}^*)$ of parameter vector (ξ, α).

Step IV. Repeat the bootstrap sampling for R times. In this study the resampling is repeated 500 times.

Since the bootstrapping produces a large number of estimates for each parameter (say ξ_1 for example), the bootstrap confidence interval (or percentile interval) for ξ_1 can be estimated by simply taking quantiles from empirical marginal distribution $\hat{F}_{\xi_1}$. To do this, we first obtain the ordered bootstrap estimates $\{\xi_{11}^*, \ldots, \xi_{1R}^*\}$ derived from the bootstrap resampling method. The two-sided confidence interval at level α is then given by $[\xi_{1R(\alpha/2)}^*, \xi_{1R(1-\alpha/2)}^*]$. Note that unlike standard asymptotic confidence intervals, percentile intervals obtained from bootstrapping will not generally be symmetric around the parameter estimate.

4. Results

The cumulative effects of pollutants and weather factors on the respiratory illnesses have been recognized by many researchers (Schwartz, 2000; Zeger et al., 1999; Dominici et al., 2003; Xia et al., 2002; Xia and Tong, 2006; Shao et al., 2010). Relevant investigations on the situation were conducted by (Xia et al., 2002; Xia and Tong, 2006; Shao et al., 2010). In our data analysis, the 'weekend-effect' and the change of hygiene habits, together with the environmental effect, are modeled by (3). Log-transformation is applied to the number of respiratory cases before model fitting.

The cumulative effects are incorporated into the model by using lagged variables. For ease of implementation and interpretation, we use the same lags for all variables in our model (Xia et al., 2002; Shao et al., 2010). That is, for each time point t, $\mathbf{X} = (X_{1t}, \ldots, X_{p,t})$ is the collection of all the environmental and weather factors with lag l, where $X_{i,t} = (x_{i,t}, \ldots, x_{i,t-l})$ with $x_{i,t}$ being observed value of the ith variable at time t and l is an integer to be estimated. The long-term effect of changing hygiene habits is measured by the expected amount of change (jump) on the number of admissions after SARS epidemic. For a given lag l, the efficient dimension M and the optimal bandwidth h are determined by the cross validation criterion given in Eq. (10). We did all computations in Matlab (see http://www.mathworks.com/products/matlab/).

The optimal results have lag $l = 7$ and $M = 3$, which are the same as those given before (Xia et al., 2002; Shao et al., 2010). The parameter estimates for the SARS effect (modeled by a growth curve) and the weekly effect are given in Table 2. The optimal bandwidth is $h = 0.339$. It can be seen from Table 2 that all the parameters in the growth curve part are statistically significant at the level of 0.05. The expected difference of the number of respiratory cases (after log-transformation) due to the change of public health measures is estimated as $\hat{\xi}_1 = -0.4977$ (the second term in the growth curve is very small in the beginning of the public health measures), which means that the number of total respiratory cases after the SARS epidemic reduces 39.21% in comparison with the expected number of respiratory cases (original data) before the SARS epidemic. The shape and 95% confidence limits of the growth curve which model the effect of SARS over time is given in Fig. 5 which

Table 2
Parameter estimates and 95% confidence bounds.

Parameter	ξ_1	ξ_2	ξ_3	ξ_4	α_1	α_2	α_3	α_4	α_5	α_6
Estimate	−0.498	0.027	0.565	252.25	−0.109	−0.121	0.074	0.018	0.007	0.046
2.5% bound	−0.524	0.021	0.524	203.25	−0.159	−0.167	0.023	−0.031	−0.046	0.002
97.5% bound	−0.460	0.034	0.593	327.08	−0.055	−0.061	0.132	0.075	0.063	0.106

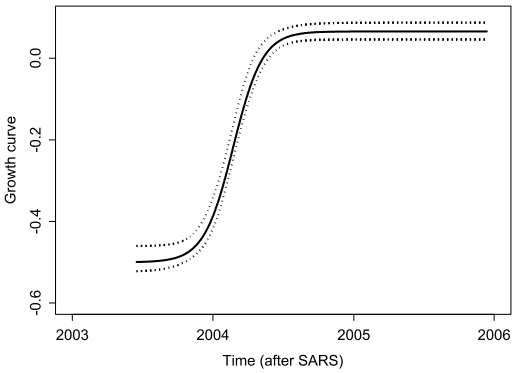

Fig. 5. The shape (solid line) and 95% confidence limits (dotted lines) of the growth curve which model the effect of SARS over time.

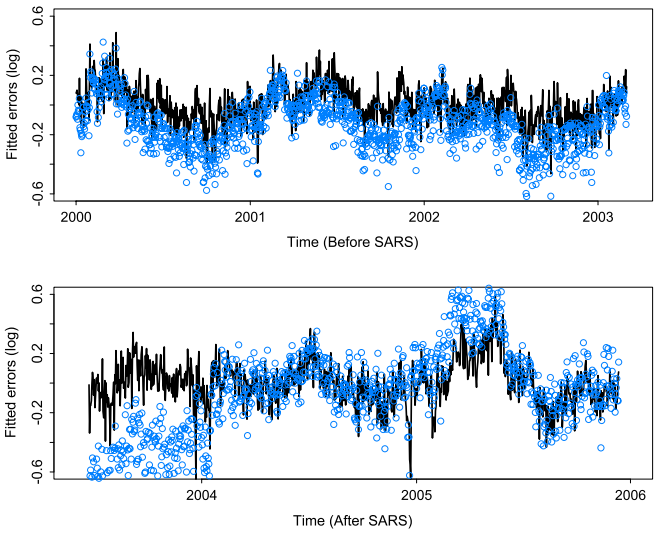

Fig. 6. Time series plots of the fitted errors (difference between observed and fitted number of cases) for models with (line) and without change (dots). The results are separated by before (top panel) and after (bottom panel) SARS.

clearly shows the significance of the public health measure due to the SARS outbreak on the number of administration. However, the impact eventually fades over time as modeled by the logistic growth curve. The impact is obvious as nearly all the parameter estimates are significantly different from zero at 95% of the confidence level (Table 2).

It can also be seen from Table 2 that the weekly effect is statistically significant, with an obvious lower number of administration during Saturday and Sunday but higher on Monday.

The cross-validation $CV = 0.0502$. The root mean squared error between observed and fitted values is $RMSE = 0.1414$, which translates to the root mean squared error $RMSE = 31.25$ for the original number of the cases. Therefore, the model works reasonably well to capture the change of the number of hospital administrations. The residuals are plotted in Fig. 6. This reduction in the number of respiratory diseases is not an accidental event. It has come from a combination of better public awareness and the efforts of the HKSAR government.

A by-product of our model is the assessment of the weekend-effect and the impact of different factors on the number of admitted patients. The change of ratios of the total respiratory cases for each day of the week to that for Friday is given in Table 2. We can see that the number of reported cases on Saturday is less than Friday and decreases a little further on Sunday, before a big increase on Monday. Tuesday and Wednesday are only slightly more than Friday before more cases on

Table 3
Coefficients of the first efficient dimension reduction direction with Lag $l = 7$, $M = 3$ and $h = 0.332$. The coefficients of greater than 0.1 are highlighted by **bold** font.

Lag	0	1	2	3	4	5	6	7
Dimension one								
SO_2	**0.18198**	**0.13521**	0.07044	0.08680	**0.10318**	**0.14647**	0.08144	**0.21273**
RSP	−0.12792	−0.04405	−0.07415	**−0.16826**	0.05507	−0.01655	0.06469	−0.02638
NO_x	−0.02260	−0.07716	0.00683	0.00786	−0.07636	**−0.13606**	−0.03518	−0.03128
NO_2	−0.09867	−0.03139	−0.07528	0.02517	**−0.10034**	−0.08479	**−0.15833**	**−0.24851**
Ozone	**0.16884**	**0.10748**	0.06374	**0.15246**	−0.02362	−0.05829	0.06656	−0.05566
Temp	**0.43964**	−0.01396	**−0.10179**	**0.14753**	**−0.25698**	0.03630	−0.08213	**−0.32950**
Humi	**0.21626**	0.07775	**0.13371**	0.05545	**0.10276**	**0.12552**	**0.17745**	**0.13737**
Dimension two								
SO_2	0.03920	−0.01234	**−0.14083**	**−0.16270**	**−0.14157**	**−0.10755**	**−0.13330**	−0.09973
RSP	0.01641	−0.00204	0.05754	0.04197	**0.12598**	0.07726	0.06560	0.02284
NO_x	**0.19583**	**0.13806**	**0.27809**	**0.26501**	**0.19721**	**0.16392**	**0.21650**	**0.14070**
NO_2	**−0.22760**	−0.08043	**−0.22063**	−0.06629	**−0.24412**	**−0.14617**	**−0.16537**	−0.06376
Ozone	0.04157	−0.02062	−0.01193	0.02108	**−0.16967**	**−0.12276**	0.00378	−0.01231
Temp	**0.26097**	−0.01078	0.07124	**0.26321**	0.05637	−0.05566	0.05563	0.04100
Humi	−0.08768	**−0.13989**	**0.13102**	**0.11494**	**0.12031**	**−0.10616**	−0.04013	**−0.11437**
Dimension three								
SO_2	0.01869	0.09162	**0.14565**	**0.14309**	**0.12004**	0.07965	−0.01265	**−0.12548**
RSP	0.09027	**0.15887**	0.06233	**0.22867**	0.09941	0.08186	0.01825	−0.00017
CO	−0.01685	0.06540	−0.08793	−0.01226	**0.11379**	**0.18284**	**0.10150**	**0.13091**
NO_2	**0.20283**	**0.25812**	−0.06601	**−0.33088**	**−0.17358**	**−0.21485**	**0.10228**	**0.14089**
Ozone	−0.01332	0.05816	0.02803	−0.01551	**0.14001**	**0.15330**	0.03493	**0.19790**
Temp	0.06353	**−0.12861**	0.04030	**−0.20063**	−0.03444	**−0.35266**	**−0.17107**	**−0.27707**
Humi	0.01586	0.04363	0.01929	0.03201	0.04432	0.02327	−0.03352	0.05263

Table 4
Summary statistics of the fitted errors (difference between observed and fitted number of cases) for models with and without change. The model is fitted using log-transformed data.

	Overall		Before SARS		After SARS	
	With	Without	With	Without	With	Without
Minimum	−0.950	−1.034	−0.489	−0.447	−0.950	−1.034
1st quartile	−0.101	−0.098	−0.083	−0.054	−0.113	−0.164
Median	0.001	0.004	−0.003	0.024	0.001	−0.022
Mean	0.000	0.000	0.000	0.029	0.000	−0.036
3rd quartile	0.099	0.104	0.088	0.111	0.120	0.092
Maximum	0.626	0.573	0.430	0.552	0.626	0.573
STD	0.168	0.180	0.140	0.142	0.199	0.213

Thursday. Friday has the smallest reported cases in comparison with other weekdays, perhaps again due to the weekend effect. We do not tend to study the detailed holiday effect any further here.

The impact of different factors on the number of admitted patients can be assessed by the coefficients $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)$ of the index, which are given in Table 3. We pay special attention to the large coefficients associated with the factors. Following (Xia et al., 2002; Shao et al., 2010), we regard coefficients with absolute value greater than 0.1 as significant coefficients. With this rationale, a simple way to interpret the results is to look for coefficients larger than 0.1. Let us call these coefficients as large coefficients. Bear in mind that the first dimension accounts the most contribution to the model. We can see that SO_2, temperature and humidity are important factors for diseases, compared with NO_2 and OZ which have moderate contributions. The RSP and NO_x are less important to the respiratory diseases in Hong Kong. The second dimension is set to model the variance which is not included in the first dimension. We can see that for the second dimension, SO_2, NO_x and Humidity are the major factors and RSP still does not show its importance. For the third dimension, all the factors show similar importance except humidity which has no coefficient larger than 0.1.

As a comparison, we report here the best model for $\lambda = 0$ (i.e., the change of hygiene habits does not affect the respiratory cases), with $h = 0.2417$ and $CV = 0.05924$. The root mean squared error between observed and fitted values is $RMSE = 0.1802$, which translates to the root mean squared error $RMSE = 37.25$ for the original number of the cases. To see the difference between the models with and without change, we provide the fitted errors (differences between the observed and fitted numbers of cases) in time series plots in Fig. 6 and summary statistics in Tables 3 and 4. We can see visually from Fig. 6 that the model without change tends to under-fit the observed numbers for the data before SARS and over-fit the observed numbers for the data after SARS. This can be confirmed by the summary statistics in Table 3. The model with change has smaller residuals with smaller standard deviation for the overall data, the data before SARS and the data after SARS. For the model with change, the summary statistics do not show any trend when the data are splitinto two part

(before and after SARS). However, for the model without change, both the mean and median prediction errors are positive for the data before SARS but are negative for the data after SARS, reassuring the importance of incorporating change in the model in order to capture the change after the SARS outbreak.

5. Conclusion and discussion

In this paper, we develop a model to assess the effect of ambient air quality and weather conditions on respiratory diseases with consideration of the impact of the change of hygiene related factors by using the multiple index model with change, and model the change after the introduction of hygiene related measures (after the SARS outbreak in the case of Hong Kong data) by a growth curve. The model consists of three additive parts: the nonparametric multiple index model part takes account of the impact of pollutants and climatic factors, a growth curve after the SARS outbreak models the impact due to public health measures and a categorical variable models the weekly effect. The data from Hong Kong are used for this study as the 2003 SARS epidemic initiated a change of public health measures in the community due to government effort and public awareness. In addition, another possible driving force for maintaining individual hygiene habits during the epidemic was the power of the media. During the time of the SARS epidemic, all media (including newspapers, radio and TV) reported daily death cases due to SARS. This was probably a most effective way to alert people of the dreadful disease and the importance of having good hygienic habits. It was well observed that during the SARS period the people of Hong Kong followed the advice of the government on hygienic measures closely, such as wearing face masks in the public, washing hands frequently, avoiding going to the public in case of having a flu, etc. The effects of pollutants and climatic factors on respiratory diseases are complicated due to the latent effect, nonlinearity and interactions between factors. The multiple index model is used to model such complicated relationship due to it flexibility and ability of handling a large number of dependent variables. The results provide indirect quantitative support to the governmental awareness program during the outbreak of the epidemic, and also the effectiveness of the media in disease control. Another important observation, however, is that the growth curve starts to increase sharply near the end of 2003 (see Fig. 5), and has flattened before mid-2004. This means that the number of illnesses had reverted to pre-SARS levels in about 6 months' time. This in turn reflects the forgetfulness of Hong Kong people, or people in many large cities. Large city people are occupied with many attractions, when the immediate threat of an epidemic fades, their alertness will wane fast. The media's interest in the epidemic is also short-lived if it does not continue to cause casualties. This suggests strongly that public health organizations and officers have to frequently remind people to keep up with their hygienic habits, so as to minimize the damage of infectious diseases.

In terms of modeling practice, it is noted that the growth curve function becomes positive after the summer 2004 due to a short upward trend in the first half of year 2005. Although it is possible to restrict the growth curve function to the maximum of zero by letting $\xi_3 \leq -\xi_1$, it is not a good exercise to implement such a restriction before clearer understanding of the short upward trend. Furthermore, our model setting has no problem to include more parametric terms such as a linear function of time for long-term trends.

Acknowledgments

The authors thank the Hospital Authority of Hong Kong for providing the hospital admission data. The research of Heung Wong and Wai-Cheung Ip was supported by a GRF grant of the Hong Kong Research Grant Council PolyU5029/11P. Thanks to Mr. Xingfa Zhang for his computing assistance and Mark Palmer for his valuable comments and suggestions during our internal review. Thanks also to two anonymous referees for their constructive comments.

References

Aunan, K., Pan, X.-C., 2004. Exposure-response functions for health effects of ambient air pollution applicable for China — a meta-analysis. Sci. Total Environ. 329, 3–16.

Bruckmann, P., Wichmann-Fiebig, W., 1997. The efficiency of short term actions to abate summer smog: results from field studies and model calculations. Eurotrac Newsletter 19, 2–9.

Chen, R., Tsay, R.S., 1993. Functional-coefficient autoregressive models. J. Amer. Statist. Assoc. 88 (421), 298–308.

Cheng, C., 2003. Report on the public responses to the SARS outbreak in Hong Kong. Available from http://www.ust.hk/src/Research_e.html.

Cook, D., 1998. Regression Graphics: Ideas for Studying Regressions Through Graphics. Wiley, New York.

Dominici, F., Mcdermott, A., Zeger, S.L., Samet, J.M., 2003. Airborne particulate matter and mortality: time-scale effects in four U.S. cities. Am. J. Epidemiol. 157, 1055–1065.

Dominici, F., Samet, J.M., Zeger, S.L., 2000. Combining evidence on air pollution and daily mortality from the 20 largest U.S. cities: a hierarchical modelling strategy. J. Roy. Statist. Soc. A 163, 263–302.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Forster De, F.P.M., Solomon, S., 2003. Observations of a 'weekend effect' in diurnal temperature range. Proc. Natl. Acad. Sci. 100, 11225–11230.

Härdle, W., Stoker, T.M., 1989. Investigating smooth multiple regression by method of average derivatives. J. Amer. Statist. Assoc. 84, 986–995.

Hastie, T., Tibshirani, R., 1993. Varying-coefficient models (with discussion). J. Roy. Statist. Soc. B 55 (4), 757–796.

Housing Authority, 2003. HA shopping centre activities postponed. Press release 28 March 2003; Available from http://www.info.gov.hk/gia/general/200303/28/0328183.htm.

Jalaludin, B.B., O'Toole, B.I., Leeder, S.R., 2004. Acute effects of urban ambient air pollution on respiratory symptoms, asthma medication use, and doctor visits for asthma in a cohort of Australian children. Environ. Res. 95, 32–42.

Katsouyanni, K., Schwartz, J., Spix, C., Touloumi, G., Zanobetti, A., Wojtyniak, B., Tobias, A., Pönkä, A., Medina, S., Bachrova, L., Anderson, H.R., 1996. Short term effect of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. J. Epidemiol. Comm. Health 50 (Suppl. 1), S12–S18.

Lee, D., Ferguson, C., Mitchell, R., 2009. Air pollution and health in Scotland: a multicity study. Biostatistics 10 (3), 409–423.
Leisure Cultural Services Department, 2003. Additional precautionary measures at LCSD facilities and functions. 2003; Press release 27 March. Available from http://www.info.gov.hk/gia/general/200303/27/0327250.htm.
Leung, G.M., Lam, T.H., Ho, L.M., Ho, S.Y., Chan, B.H., Wong, I.O., et al., 2003. The impact of community psychological responses on outbreak control for severe acute respiratory syndrome in Hong Kong. J. Epidemiol. Comm. Health 57, 857–863.
Li, K.C., 1991. Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc. 86, 316–342.
Maynard, R., 2004. Key airborne pollutants-the impact on health. Sci. Total Environ. 334–335, 9–13.
Mcardle, J.J., Nesselroade, J.R., 2002. Growth curve analysis in contemporary psychological research. In: Schinka, J.A., Velicer, W.E., Weiner, I.B. (Eds.), Handbook of Psychology Vol 2: Research Methods in Psychology. pp. 447–480.
Navrud, S., 2001. Valuing health impacts from air pollution in Europe: new empirical evidence on Morbidity. Environ. Resour. Econom. 20, 305–329.
Neuberger, M., Schimek, M.G., Horak Jr., F., Moshammer, H., Kundi, M., Frischer, T., Gomiscek, B., Puxbaum, H., Hauck, H., 2004. Acute effects of particulate matter on respiratory diseases, symptoms and functions: epidemiological results of the Austrian Project on Health Effects of Particulate Matter (AUPHEP). Atmos. Environ. 38, 3971–3981.
Richard, F.J., 1959. A flexible growth function for empirical use. J. Exp. Botany 10, 290–300.
Schwartz, J., 2000. Harvesting and long term exposure effects in the relation between air pollution and mortality. Am. J. Epidemiol. 151, 440–448.
Scoggins, A., Kjellstrom, T., Fisher, G., Connor, J., Gimson, N., 2004. Spatial analysis of annual air pollution exposure and mortality. Sci. Total Environ. 321, 71–85.
Shao, Q., 2002. A reparameterization method for embedded models. Comm. Statist. Theory Methods 31 (5), 683–697.
Shao, Q., Wong, H., Ip, W.C., Li, M., 2010. Effect of ambient air pollution on respiratory illness in Hong Kong: a regional study. Environmetrics 21, 173–188.
Smoyer, K.E., Kalkstein, L.S., Greene, J.S., Ye, H., 2000. The impact of weather and pollution on human mortality in Birmingham, Alabama and Philadelphia. In. J. Clim. 20, 881–897.
Xia, Y., Tong, H., 2006. Cumulative effect of air pollution on public health. Stat. Med. 25, 3548–3559.
Xia, Y., Tong, H., Li, W.K., Zhu, L., 2002. An adaptive estimation of dimension reduction space. J. Roy. Statist. Soc. B 64 (3), 363–410.
Zeger, S.L., Dominici, F., Samet, J., 1999. Harvesting-resistant estimates of pollution effects on mortality. Epidemiology 10, 171–175.