Environ Ecol Stat
DOI 10.1007/s10651-012-0192-5

# Prediction of non-methane hydrocarbons in Kuwait using regression and Bayesian kriged Kalman model

**Fahimah A. Al-Awadhi · Ali Alhajraf**

Received: 31 January 2011 / Revised: 23 February 2012

**Abstract** This article describes the hierarchical Bayesian approach for predicting average hourly concentrations of ambient non-methane hydrocarbons (NMHC) in Kuwait where records of six monitor stations located in different sites are observed at successive time points. Our objective is to predict the concentration level of NMHC in unmonitored areas. Here an attempt is made for the prediction of unmeasured concentration of NMHC at two additional locations in Kuwait. We will implement a kriged Kalman filter (KKF) hierarchical Bayesian approach assuming a Gaussian random field, a technique that allows the pooling of data from different sites in order to predict the exposure of the NMHC in different regions of Kuwait. In order to increase the accuracy of the KKF we will use other statistical models such as imputation, regression, principal components, and time series analysis in our approach. We considered four different types of imputation techniques to address the missing data. At the primary level, the logarithmic field is modeled as a trend plus Gaussian stochastic residual model. The trend model depends on hourly meteorological predictors which are common to all sites. The residuals are then modeled using KKF, and the prediction equation is derived conditioned on adjoining hours. On this basis we developed a spatial predictive distribution for these residuals at unmonitored sites. By transforming the predicted residuals back to the original data scales, we can impute Kuwait's hourly non-methane hydrocarbons field.

F. A. Al-Awadhi (✉)
Department of Statistics and Operations Research, Kuwait University, P.O. Box 5969,
Khaldiya 13060, Kuwait
e-mail: falawadi@kuc01.kuniv.edu.kw

A. Alhajraf
Department of Biomedical Sciences, College of Nursing, The Public Authority for Applied Education
and Training, P.O. Box 23167, Khaldiya 13092, Kuwait
e-mail: af.alhajraf@paaet.edu.kw

Published online: 03 May 2012

**Keywords** Imputation · Space–time models · Spatial interpolation · Kriging · Kalman filter · Hierarchical model · Spatial prediction · Non-methane hydrocarbons · Air pollution

## 1 Introduction

The need to develop adequate spatial environmental models to address environmental issues has grown rapidly in recent years around the world. Statistical modeling and prediction of one or more pollutants generated at each of a regular series of timepoints over non-overlapping regions of the same geographical domain have been a concern of statisticians. For example, Carroll et al. (1997) studied ozone exposure in Harris county, Texas, Tonellato (2001) analyzed the concentration of carbon monoxide $CO_2$ in the city of Venice, Shaddick and Wakefield (2002) considered sulphur dioxide $SO_2$, nitrogen oxide NO, carbon monoxide CO and $PM_{10}$, for London over a 4-year time period, Zidek et al. (2002) dealt with particulate matter $PM_{10}$ for the city of Vancouver; Kibria et al. (2002) predicted $PM_{2.5}$ exposure for Philadelphia, and Huerta et al. (2004) considered hourly readings of concentration of ozone $O_3$ over Mexico City. Al-Awadhi and Al-Awadhi (2006) dealt with daily readings of non-methane hydrocarbons for the state of Kuwait along with nitrogen oxide, carbon monoxide and sulfur dioxide.

Temporal and spatial interpolation were first approached using the method of Kriging, for example Rouhani and Mayers (1990), Mardia and Goodall (1993), Cressie (1993) and Mardia et al. (1998). Bayesian methodology for both temporal and spatial interpolation has been developed in the past decade as an alternative to space–time Kriging, for example, Le and Zidek (1992) and Brown et al. (1994). Both approaches assume an underlying spatial process with responses that are functions of their locations.

This paper is motivated by an analysis of air quality data for the state of Kuwait to predict the concentration level of non-methane hydrocarbon compounds (NMHC) in unmonitored areas. It applies the Bayesian kriged Kalman filter (BKKF) method to the problem of modeling fields of NMHC. Non-methane hydrocarbon compounds are organic molecules present in the atmosphere produced by the processes of extracting and refining oil in Kuwait, including combustion and volatilization. Internal combustion engines with unburned fuel emissions, solvents, particularly paints, liquefied petroleum gas or natural gas leakage are the main sources of NMHC in addition to industrial and domestic sources, such as decoration, chemical factories and power plants. All the above mentioned activities have effects on the environment and on public health, as confirmed by ecologists in Kuwait (Alsayed 2008, 2009). Owing to their high reactivity and flux, increasing non-methane hydrocarbons lead to an increased production of ozone. Kuwait Environment Public Authority (EPA) has specified the concentration of non-methane hydrocarbons for morning hours shall not exceed 0.24 ppm. AbdulWahab and Bouhamra (2004), Al-Awadhi et al. (2005) have shown that the levels of NMHC exceed the proposed ambient air quality standard for residential areas in Kuwait. Al-Awadhi (2011) used a multivariate Bayesian approach for predicting hourly exposure of NMHC assuming that the response will follow a multivariate Gaussian distribution with an isotropic covariance matrix. The lifetimes of the non-methane hydrocarbons vary from a few minutes to several months for some of the lighter alkane. Therefore, a temporal effect exists between the hourly records up to certain limits. To reduce the computation time, she used a multivariate approach of the spatial process to predict hourly non-methane levels for the six stations at a given time using the $p$ proceeding hours instead of using all available data. Comparing different $p$-preceding hours to predict the next level led to a conclusion that $p = 4$ consecutive hours will capture enough information to predict the level of concentration of NMHC for Kuwait with a satisfactory precision level.

In this paper, we consider a hierarchical Bayesian kriged-Kalman filtering (BKKF) model introduced by Sahu and Mardia (2005). Spatial attribute estimation and the associated accuracy depend on the available sampling design and statistical inference modeling. Both the theoretical analysis and the empirical study show that the mean Kriging technique outperforms other commonly-used techniques. We use deterministic and stochastic approaches in our model. We build a model for NMHC based on linear regression to predict the non-methane hydrocarbons for unmeasured sites in Kuwait. In the regression model, we accommodate different meteorological effects. For the resulting stochastic residual, we follow the BKKF approach that uses the information gathered from monitor sites to predict unmeasured levels at unmonitored sites. We will compare our result with the stochastic approach where we ignore modeling the physical and chemical processes governing the spatial and temporal evolutions of pollutant concentrations represented by the meteorology variables.

This paper is organized as follows: the general description of the data and some explanatory analysis is discussed in Sect. 2. In Sect. 3, the general trend model is given and explained. Detailed KKF model is described in Sect. 4 and the posterior and the conditional distributions are derived. Section 5 illustrates the application and the results of our proposed method. Spatial prediction for unmonitored sites is shown in Sect. 6 with a discussion in Sect. 7.

## 2 Exploratory analysis of Kuwait pollution data

The data were collected from six fixed air monitoring stations operated for several years by EPA and distributed throughout the state of Kuwait. The levels of concentration of NMHC along with different atmospheric variables are considered from 1 July 2004 to 5 September 2004. The data are recorded every 5 min continuously, 24 h a day. The 5-min values are likely subject to extreme measurement error, hence we will consider aggregating the data to an hourly record to obtain a 24-h reading per day for each station.

Kuwait is a small flat desert country influenced significantly by high speed wind; thus the atmospheric concentration levels of its areas are interrelated. Therefore, it is important to pool the information gathered from different monitored sites to model the concentration of NMHC.

Each of the six monitoring stations, Mansouriya (Site 1), Rabiya (Site 2), Riqqa (Site 3), Um-Alhayman (Site 4), Jahra (Site 5) and Um-Alayesh (Site 6), is located in areas that differ significantly in their surroundings, geographical characteristics and the sources of air pollution. Figure 1 shows the locations of these monitoring stations along with their surroundings.

Mansouriya is a typical urban residential and commercial area that is impacted by heavy traffic of the capital city. It gives a clear picture of pollutants from automobiles and local sources. The source of pollutants for the second monitoring station, Rabiya, are Kuwait airport, a sewage plant and Al-Ray industrial area. Riqqa is a semi-coastal urban area surrounded by petroleum refineries and industrial plants. The pollutants in Riqqa are influenced by the sea breeze. Um-Alhayman is a coastal urban region situated in the southern part of Kuwait. It is surrounded by oil wells, an oil port, petrochemical and oil industry plants, as well as other plants. Jahra is an urban residential and commercial area situated in the north of Kuwait surrounded by a power plant, oil fields, an industrial area and a sewage plant. Um-Alayesh is a deserted rural area in the northern part of Kuwait.

At most stations, some data are missing. The percentage of missing data at different stations is given for Site 1 (14.96 %), Site 2 (10.49 %), Site 3 (7.76 %), Site 4 (11.77 %), Site 5 (18.35 %) and Site 6 (24.76 %). Missing data is an ubiquitous problem in evaluating long term experimental measurements. Different techniques can be used to impute the missing values to keep the full sample size (see for example West and Harrison 1997; Rubin and Little 2002; Zhu et al. 2003; Fitzmaurice et al. 2009; Pollice and Lasinio 2009). Some imputation methods are simple and some are sophisticated. So far, no method has been commonly accepted and the imputation methods used are largely dependent on the researchers' choice. An important criteria to consider is

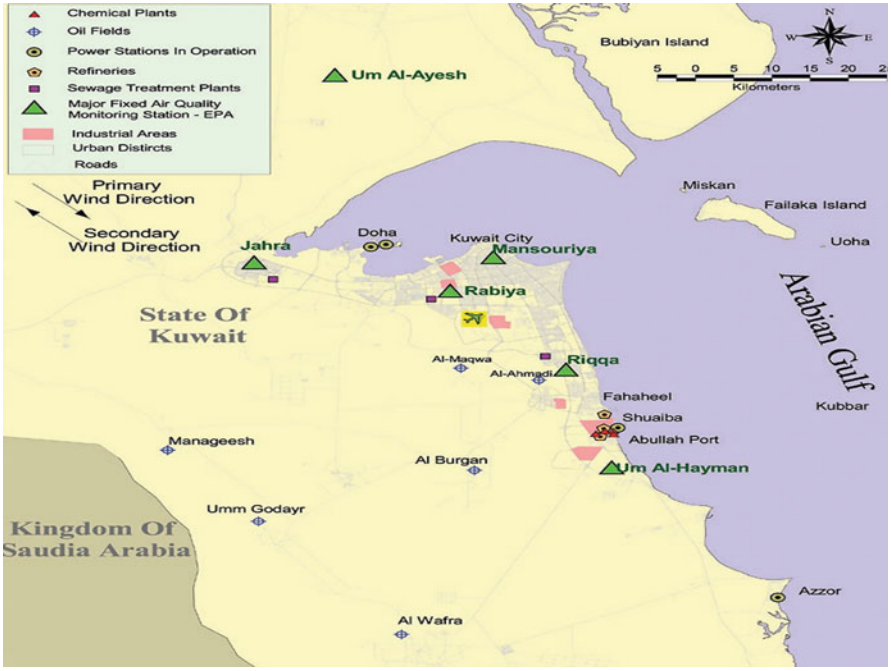

**Fig. 1** Kuwait map verifying location of the six monitoring stations along with their surroundings

the standard deviation or the correlations between the observations. We do not want to underestimate or overestimate; that is, we want to keep the nature of the records unaffected as much as we can.

We used four different approaches for imputation; imputation 1: simple random imputation, imputation 2: regression prediction using the meteorological regressors such as wind speed, relative humidity, temperature and week days. Imputation 3: linear spatial regression which predicts the missing data in a specific station using the NMHC data of other stations. This method is a modification of using the mean of the relative measures of the different sites to replace the missing of each site and it overcomes the under-estimation of the standard deviation of the data. Imputation 4 applies MCMC method assuming the data from multivariate normal distribution. Data augmentation is applied to Bayesian inference with missing data to converge to a distribution $P(Y_{mis}|Y_{obs})$ (Dempster et al. 1977) using the expectation-maximization (EM) to provide a good starting point.

The four different imputation methods produced different results. We display a sample of time series plots of the resulted NMHC data in Figs. 2 and 3 for station 1 and station 6 using the four imputation methods as well as the original data set. Our criteria for selecting the optimal method is to keep least changes in the variance-covariance matrix of the data. We used log-likelihood ratio statistic for testing the equity of the variance-covariance matrices of the original data set and the resulting data after the imputation. All had significant results, the least chi-quare testing value occurred when applying imputation 3, hence linear spatial regression imputation is considered.

The data is transformed by taking the natural logarithm to be a normally distributed random response, after which some statistical techniques such as trend surface plot,

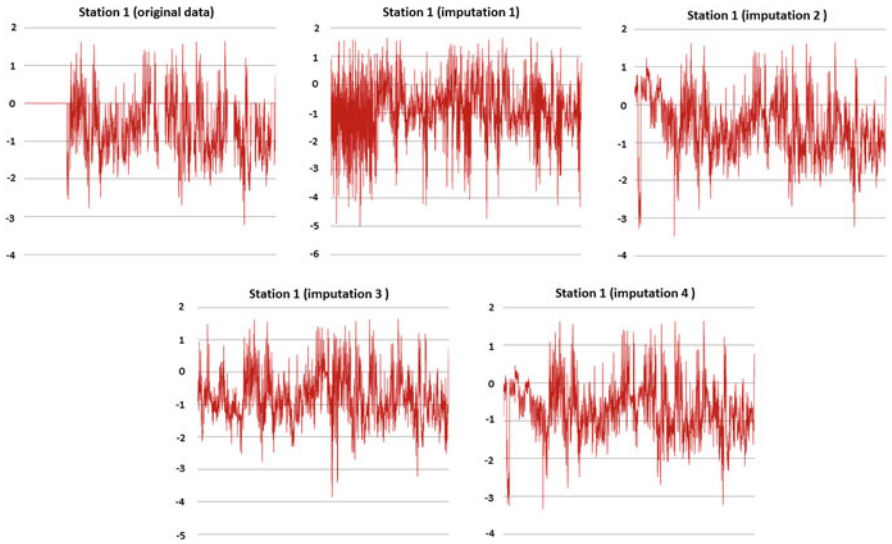

**Fig. 2** NMHC data for station 1 using different imputation methods (imputation 1: simple random imputation, imputation 2: regression prediction using the meteorological regressors, imputation 3: linear spatial regression, imputation 4: MCMC imputation)

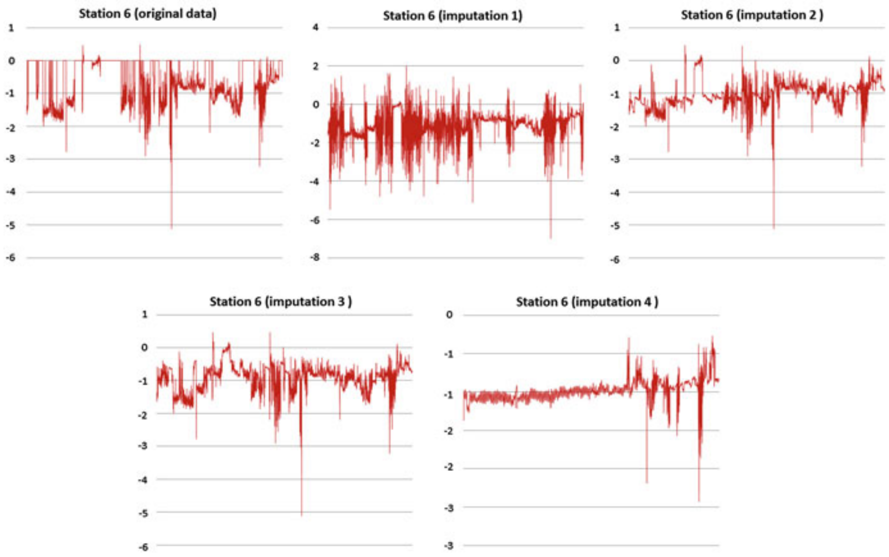

**Fig. 3** NMHC data for station 6 using different imputation methods (as above if Fig. 2)

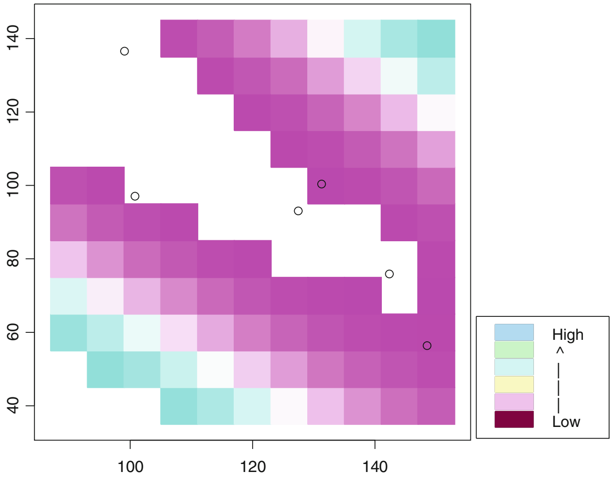

**Fig. 4** Trend surfaces for the NMHC in Kuwait by *gray scale plot*

correlation and time series plot are applied to identify the features of the data. Trend surfaces describe the behavior of the data with respect to space. For our data, it is obvious by observing Figs. 4 and 5 that the pollutant levels increase in the south-west near the refineries and oil wells. Moreover the increase is in the direction of prevailing wind directions north-west and south-east.

Table 1 displays the variances of the logged NMHC for the six stations on the diagonal and the values of the correlation coefficient below the diagonal. The symbol (*) on the values of the correlation coefficients in the table indicates the significance

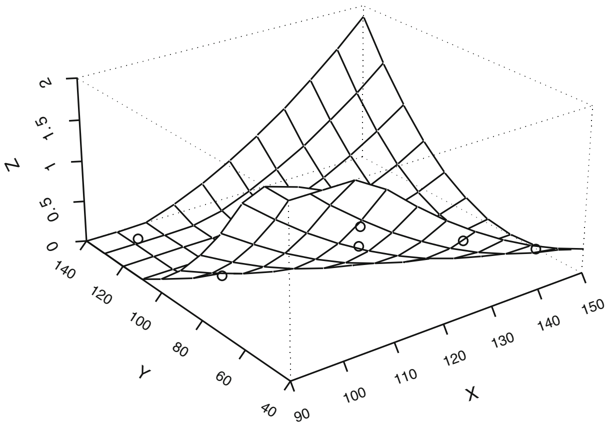

**Fig. 5** Trend surfaces for the NMHC in Kuwait by perspective plot

**Table 1** The variance of the logged NMHC on the diagonal for the six stations and the values of the correlation between them below the diagonal

| | Site 1 | Site 2 | Site 3 | Site 4 | Site 5 | Site 6 |
|---|---|---|---|---|---|---|
| Site 1 | 0.420 | | | | | |
| Site 2 | 0.815* | 0.459 | | | | |
| Site 3 | 0.293* | 0.288* | 0.860 | | | |
| Site 4 | 0.084* | 0.167* | 0.059* | 0.709 | | |
| Site 5 | 0.533* | 0.609* | 0.395* | 0.317* | 0.400 | |
| Site 6 | 0.214* | 0.157* | 0.380* | −0.053* | 0.348* | 0.229 |

The symbol (*) indicates the significance of the correlation values at 0.05 level

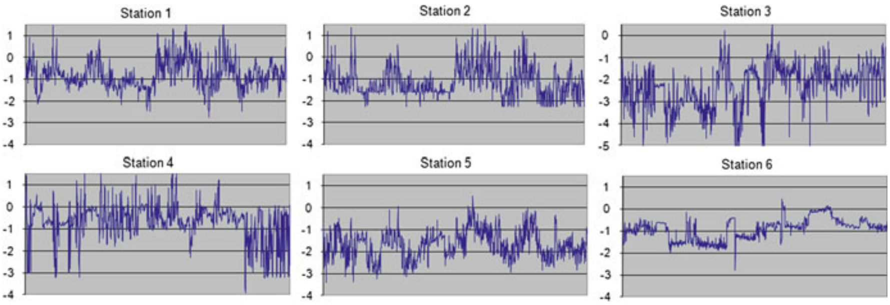

**Fig. 6** Time series plots for the logged NMHC for the six stations (*top*: stations 1–3, *bottom*: stations 4–6)

at 0.05 level. Site 3 has the largest variation, which was expected as it is surrounded by many factories and is affected by the sea breeze and pollution from the other sites. The lowest variation is recorded for Site 6 a deserted rural area. The high values of the correlation coefficients between Site 1, Site 2 and Site 5 show the influence of factors like wind on the concentration of the pollutants in nearby places.

The analysis of time series plots illustrating the hourly variations in the average concentration of NMHC for the six sites (Fig. 6) demonstrates that it follows different patterns for various stations. The atmospheric concentration of NMHC at each station is affected by many meteorological and geographical factors including climatology, traffic, height of surrounding buildings, power plants, oil refineries, and emissions from factories. These factors play a crucial role in determining the level of concentration of NMHC. The trends of NMHC for Sites 1 and 2 are alike due to the similarities of the two areas. For Site 6 the trend is found to be nearly constant.

In the next section, we describe the underlying space–time model for the NMHC in Kuwait, and in particular we find a trend model that incorporates not only structural time trends but meteorology as well.

## 3 The trend model

For a data set with an apparent trend, both ordinary kriging and universal kriging may considered (Journel and Rossi 1989). Their estimators appear as a sum of a trend and residual components. The trend component in the ordinary kriging approach is expressed as a constant term while in the universal kriging it is split into a number of pre-specified subcomponents. Let $Y^*(s, t)$ be the logarithm of the NMHC level at hour $t = 1, \ldots, T$, and space $s = 1, \ldots, S$. Our spatial-temporal model for hourly ambient NMHC has the simple form

$$Y^*(s, t) = \mathcal{R}(s, t) + Y(s, t) \tag{3.1}$$

Here $\mathcal{R}(s, t)$ represents the site-consistent deterministic component that incorporates hourly and daily effects, $Y(s, t)$, the error term which represents the spatial-temporal correlation structure due to other sources of variation or factors.

Temperature (temp), relative humidity (rh) and wind speed (ws) recorded at Kuwait Airport are considered for the study. All the variables were standardized to reduce the collinearity among the variables included in the models. These selected meteorological components ($M$) were then incorporated into the full trend model. Time is another variable compounded in this model; therefore hours ($O$), day ($D$) and the linear trend ($L$) is added. Thus the model is given as

$$\mathcal{R}(t, s) = \mu + O(s, t) + D(s, t) + L(s, t) + M(s, t), \tag{3.2}$$

where $\mu$ is the overall effect.

Table 2 gives the regression statistics for the six sites. Wind speed (ws), time ($t$) and temperature (temp) have significant influence on the records. The hour of the day and the day of the week may reflect the influence of traffic. It is noted from the regression analysis (not shown here) that comparatively higher values are shown during late night hours, as the NMHC concentration rises from 5 p.m. till early morning and then declines during the day time, signifying a negative relation with temperature. Moreover, in these hours the humidity traps the pollutants from dispersing. The least concentration occurs during 1–3 p.m. For the weekly cycle, the highest levels occur

**Table 2** Regression statistics for logged NMHC at Sites 1–6

| Source | DF | $MS_1$ | $MS_2$ | $MS_3$ | $MS_4$ | $MS_5$ | $MS_6$ |
|---|---|---|---|---|---|---|---|
| Constant | 1 | 49.58 | 203.19 | 1, 261.48 | 200.76 | 466.40 | 2395.27 |
| Week | 6 | 10.55 | 22.05 | 2.564 | 7.91 | 1.031 | 1.657 |
| Hour | 23 | 13.51 | 7.84 | 17.83 | 19.24 | 1.80 | 1.05* |
| $t$ | 1 | 172.69 | 16.30 | 66.35 | 1, 529.38 | 179.82 | 10.75 |
| Temp | 1 | 44.97 | 3.46 | 8.50 | 152.78 | 0.53* | 2.00 |
| rh | 1 | 24.58 | 0.35* | 146.07 | 248.92 | 2.24 | 4.36 |
| ws | 1 | 873.60 | 375.21 | 808.35 | 266.73 | 33.62 | 26.68 |
| Temp × rh | 1 | 21.24 | 9.07 | 1.05* | 34.94 | 3.03 | 69.18 |
| Temp × ws | 1 | 181.68 | 1.22 | 10.03 | 16.41 | 19.63 | 1.01* |
| rh × ws | 1 | 173.06 | 1.60 | 4.19 | 33.41 | 2.13 | 9.16 |

The symbol (*) indicates the non significance at 0.05

**Table 3** The variance of $Y$ on the diagonal and the correlation between the residuals of the six sites below the diagonal

| | Site 1 | Site 2 | Site 3 | Site 4 | Site 5 | Site 6 |
|---|---|---|---|---|---|---|
| Site 1 | 0.240 | | | | | |
| Site 2 | 0.229* | 0.255 | | | | |
| Site 3 | 0.161* | 0.170* | 0.493 | | | |
| Site 4 | 0.071* | 0.052* | 0.002 | 0.253 | | |
| Site 5 | 0.136* | 0.294* | 0.064* | 0.105* | 0.239 | |
| Site 6 | 0.162* | 0.110* | 0.282* | −0.020 | 0.118* | 0.141 |

The symbol (*) indicates the significance of the correlation values at 0.05 level

during the weekend (Friday and Saturday), and the lowest toward the end of the week. Site 6, being a rural area, shows a uniform level of NMHC, confirming that emissions play an important role in building up NMHC levels.

The correlation matrix for the residuals, $Y(t, s)$, resulting from the regression analysis is given in Table 3. The highest variation is for Site 4 followed by Site 3. A reduction is noticed in the values of the correlation without changing their order. The spatial-temporal effect exists and is dealt with using the kriged Kalman filter approach.

## 4 The kriged Kalman filter approach (KKF)

The kriged Kalman filter approach (Mardia et al. 1998) combines the kriging method, named after the mining engineer D. G. Krige in the 1950s and which deals with the spatial effects, with the Kalman filtering method, which concerns the temporal effects. Kriging was discussed by many authors such as Ripley (1988), Cressie (1993), and Chilès and Delfiner (1999). It is used to make an inference about an unobserved value of a random process $Y(\cdot) = (Y(s_1), \dots, Y(s_N))$ based on an observed data at $N$

known spatial locations $\{s_1, \dots, s_N\}$. It assumes that the data $Y$ is composed of two main components: a mean term and a zero mean stochastic error,

$$Y(s) = \mu(s) + \varepsilon(s), s \in S. \tag{4.1}$$

Assuming the process $Y(\cdot)$ is Gaussian and that the conditional expectation of $Y(\cdot)$ given the data $Y$; $E(Y(s)|Y) = \mu(s)$ is linear in $Y$ and depends only on $\mu(s) = E(Y(s))$, and the covariance function of the process is of the second order stationary and isotropic as well,

$$cov(Y(s), Y(s')) = \sigma(s, s') = \sigma(s' - s).$$

The general form of the kriging assumes that the mean of $Y$ is an unknown linear combination of $q$ functions $\left\{f_1(s), \dots, f_q(s)\right\}$, that is $Y(s) = f(s)^{\mathsf{T}}\beta + \epsilon(s)$, where $\beta = (\beta_1, \dots, \beta q)^{\mathsf{T}}$ is a vector of coefficients of length $q$ which represents the number of spatial components. Hence, under the assumption of Gaussian random field

$$E(Y(s)|Y) = f(s)^{\mathsf{T}}\beta + \sigma(s)^{\mathsf{T}}\Sigma^{-1}(y - F\beta), \tag{4.2}$$

where $(\Sigma)_{ij} = \sigma(s_i, s_j), \sigma(s)^{\mathsf{T}} = (\sigma(s, s_1), \dots, \sigma(s, s_N))$ and the $i$th row of $F$ is given by $(f(s_i))^{\mathsf{T}}, i = 1, \dots, q$, the optimal predictor for the unobserved value $Y(s)$ is the conditional expectation of $Y(s)$ given $Y = y$,

$$\hat{Y}(s) = E(Y(s)|Y = y) = f(s)^{\mathsf{T}}Ay + \sigma(s)^{\mathsf{T}}By, \tag{4.3}$$

where $B = \Sigma^{-1} - \Sigma^{-1}F(F^{\mathsf{T}}\Sigma^{-1}F)^{-1}F^{\mathsf{T}}\Sigma^{-1}$ and $A = (F^{\mathsf{T}}\ \Sigma^{-1}F)^{-1}F^{\mathsf{T}}\Sigma^{-1}$. The Kalman filter originally developed by Kalman (1960) and Kalman and Bucy (1961) is also practiced by many statisticians (see Huang and Cressie 1996; Meinhold and Singpurwalla 1983).

Let $Y_t$, $t = 1, \dots, T$ be the observed values of $N$-variables of interest at time $t$. The observation equation which specifies an assumed linear relationship between $Y_t$ and an unobserved quantities $\alpha_t$ of dimension $p$ is given by

$$Y_t = H\alpha_t + \varepsilon_t, \tag{4.4}$$

where $H$ is a known $N \times p$ matrix. The error term is assumed to have a multivariate normal distribution (MVN) with zero mean and covariance $\Sigma_\epsilon$. The state equation $\alpha_t$ changes over time,

$$\alpha_t = P\alpha_{t-1} + K\eta_t, \tag{4.5}$$

where $P$ is the transition matrix of dimension $p \times p$, $K$ represents the innovation parameter matrix of dimension $p \times p$ and the disturbance term $\eta_t$ is assumed to have a MVN distribution with zero mean and covariance $\Sigma_\eta$ of dimension $p \times p$.

The kriged and Kalman filter model (KKF) which combines both the kriging and Kalman filter approaches is expressed as

$$Y(s,t) = h(s)^{\mathsf{T}}\alpha_t + \varepsilon_t, \tag{4.6}$$

where $\alpha_t$ as given in the state equation (4.5), $Y(s,t)$ denotes the observation at time $t$ and site $s$, $h(s) = (h_s, \ldots, h_{sp})^{\mathsf{T}}$, $\alpha_t = (\alpha_{t1}, \ldots, \alpha_{tp})^{\mathsf{T}}$ and $\varepsilon_t \backsim N(0, \sigma_\varepsilon^2)$. It is assumed that $\varepsilon_t$ and $\eta_t$ are independent for all $t$ and that the initial information about $\alpha_0$ can be quantified as MVN distribution with zero mean and covariance matrix $\Sigma_0$.

The transition matrix $P$ can be obtained using the dynamic linear model (Mardia et al. 1998) or by converting an ARMA model to state space model (Aoki 1990). The matrix $H = h(s)^{\mathsf{T}}$ is of size $N \times p$ and represents the spatial component in the model. The first $q$ columns in $H$ are the coefficients of the trend field and the remaining $p - q$ columns are the principle fields and represent the spatial directions relative to the covariance function

$$\Sigma_\gamma = \sigma_\gamma(s_i, s_j) = \sigma_\gamma^2 e^{-\lambda d_{ij}}, \quad \lambda > 0, \tag{4.7}$$

where $d_{ij}$ is the Euclidean distance between $s_i$ and $s_j$,$\lambda$,and $\sigma_\gamma^2$ and unknown parameters.

If $F$ is the matrix of the first $q$ columns of $H$ and $\Sigma_\gamma$ is the covariance matrix calculated using the above covariance function then, the bending energy matrix $B$ can be determined by $B = \Sigma_\gamma^{-1} - \Sigma_\gamma^{-1} F(F^{\mathsf{T}}\Sigma_\gamma^{-1}F)^{-1}F^{\mathsf{T}}\Sigma_\gamma^{-1}$. The $p-q$ columns of $H$ are corresponding to the smallest non-zero eigen values of $B$, which means that the principal fields with high degree of variability are chosen.

## 4.1 The Bayesian KKF model

The formulation of KKF model in a Bayesian framework is mentioned here by giving the prior specifications of the unobserved parameters in the model, determining the likelihood function, performing the posterior analysis and deriving the full conditional distributions of the parameters.

### *4.1.1 Prior specification*

It is assumed that $\sigma_\epsilon^2$, $\Sigma_0$ and $\Sigma_\eta$ are independent. As our likelihood is a member of an exponential family, it is common to use a conjugate prior which is from the same family; hence, we describe our prior knowledge about $\tau_\epsilon = \frac{1}{\sigma_\epsilon^2}$ by a noninformative gamma distribution, and set its parameters $a$ and $b$ to 0.001. Let $\Sigma_0 = C_0 I$ and set $C_0$ to be 100. Moreover, $Q_\eta = \Sigma_\eta^{-1}$ is assumed to have Wishart prior distribution with parameters $a_\eta$ and $\beta_\eta$, where $a_\eta$ is the prior degrees of freedom which is chosen to be 2 and $\beta_\eta$ is a positive definite matrix and it is chosen to be $0.01I$ where $I$ denotes a

$p \times p$ identity matrix. Consequently, the joint prior distribution for $\alpha_0$, $\tau_\epsilon$ and $Q_\eta$ is given by

$$\begin{aligned} \pi\left(\alpha_0, \tau_\epsilon, Q_\eta\right) &= \pi\left(\alpha_0\right)\pi(\tau_\epsilon|a, b)\pi(Q_\eta|a_\eta, \beta_\eta\right) \\ &= C_0^{-\frac{p}{2}} e^{-\frac{1}{2}\alpha_0^T (C_0 I)^{-1}\alpha_0} \frac{b^a}{\Gamma(a)} \tau_\epsilon^{a-1} e^{-b\tau_e} |\beta_\eta|^{\frac{a\eta}{2}} |Q_\eta|^{\frac{a\eta-p-1}{2}} e^{-\frac{1}{2}tr(\beta_\eta Q_\eta)}. \\ &= \frac{b^a}{\Gamma(a)} C_0^{-\frac{p}{2}} |\beta_\eta|^{\frac{a\eta}{2}} |Q_\eta|^{\frac{a\eta-p-1}{2}} \tau_\epsilon^{a-1} e^{-\frac{1}{2}\{\alpha_0^T (C_0 I)^{-1}\alpha_0 + tr(\beta_\eta Q_\eta)\} - b\tau_\epsilon}. \end{aligned} \tag{4.8}$$

### *4.1.2 Likelihood function*

We assume that $\epsilon_t \sim MVN(0, \tau_\epsilon^{-1} I)$ and $K\eta_t \sim MVN(0, KQ_\eta^{-1}K^\mathsf{T})$. The likelihood function for the KKF model given the observations, $L(\alpha_0, \alpha_1, \ldots, \alpha_t, \ldots, \alpha_T, \tau_\epsilon, Q_\eta; Y_1 \ldots, Y_T)$, is proportional to

$$\begin{aligned} &\prod_{t=1}^{T} \frac{1}{|\tau_\epsilon^{-1} I|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\, \epsilon_t^T \tau_\epsilon I\, \epsilon_t\right\} \frac{1}{|KQ_\eta^{-1}K^\mathsf{T}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\, (K\eta_t)^T (K^\mathsf{T})^{-1} Q_\eta K^{-1}\, (K\eta_t)\right\} \\ &= \prod_{t=1}^{T} |K|^{-\frac{1}{2}} |Q_\eta|^{\frac{1}{2}}\, |K^\mathsf{T}|^{-\frac{1}{2}}\; \tau_\epsilon^{\frac{n}{2}} \exp\left\{-\frac{1}{2} \epsilon_t^T \tau_\epsilon I\, \epsilon_t + (K\eta_t)^T (K^\mathsf{T})^{-1} Q_\eta K^{-1}\, (K\eta_t)\right\}. \end{aligned}$$

Substitute $\epsilon_t = Y_t - H\alpha_t$ and $K\eta_t = \alpha_t - P\alpha_{t-1}$, to get

$$\begin{aligned} &\prod_{t=1}^{T} |K|^{-\frac{1}{2}} |Q_\eta|^{\frac{1}{2}}\, |K^\mathsf{T}|^{-\frac{1}{2}}\, \tau_\epsilon^{\frac{n}{2}} \\ &\quad \exp\left\{-\frac{1}{2}[(y_t - H\alpha_t)^\mathsf{T} \tau_\epsilon I (y_t - H\alpha_t) + (\alpha_t - P\alpha_{t-1})^\mathsf{T} (K^\mathsf{T})^{-1} Q_\eta K^{-1} (\alpha_t - P\alpha_{t-1})].\right\} \end{aligned}$$

After simple calculations, we get

$$\begin{aligned} &|K|^{-\frac{T}{2}} |Q_\eta|^{\frac{T}{2}}\, |K^\mathsf{T}|^{-\frac{T}{2}}\; \tau_\epsilon^{\frac{Tn}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{T}[(y_t - H\alpha_t)^\mathsf{T}\right. \\ &\quad \left. \tau_\epsilon I (y_t - H\alpha_t) + (\alpha_t - P\alpha_{t-1})^\mathsf{T} (K^\mathsf{T})^{-1} Q_\eta K^{-1} (\alpha_t - P\alpha_{t-1})]\right\}. \end{aligned} \tag{4.9}$$

### *4.1.3 Posterior analysis*

Given the likelihood function (4.9) and prior specifications (4.8), the posterior distribution $\pi(\alpha_0, \alpha_1, \ldots, \alpha_t, \ldots, \alpha_T, \tau_\epsilon, Q_\eta | Y_1 \ldots, Y_T)$ is proportional to

$$\frac{b^a}{\Gamma(a)} C_0^{-\frac{p}{2}} |K|^{-\frac{T}{2}} |Q_\eta|^{\frac{T+a_\eta-p-1}{2}} |K^\top|^{-\frac{T}{2}} |\beta_\eta|^{\frac{a_\eta}{2}} \tau_\epsilon^{\frac{Tn}{2}+a-1}$$
$$\exp\left\{-\frac{1}{2}\left\{\sum_{t=1}^{T}[(y_t - H\alpha_t)^\top \tau_\epsilon I (y_t - H\alpha_t) + (\alpha_t - P\alpha_{t-1})^\top \left(K^\top\right)^{-1}\right.\right.$$
$$\left.\left. Q_\eta K^{-1}(\alpha_t - P\alpha_{t-1})] + \alpha_0^\top (C_0 I)^{-1}\alpha_0 + tr(\beta_\eta Q_\eta)\right\} - b\tau_\epsilon\right\} \tag{4.10}$$

### *4.1.4 Conditional distributions*

The full conditional distributions are needed for the Gibbs sampler algorithm. The conditional distribution of $\alpha_0$ is

$$\pi(\alpha_0|\alpha_1, \ldots, \alpha_T, \tau_\epsilon, Q_\eta) = \pi(\alpha_0|\alpha_1) \sim MVN\ (\mu_0, \Sigma_0) \tag{4.11}$$

where

$$\mu_0 = (P^\top(K^\top)^{-1}Q_\eta K^{-1}P + (C_0 I)^{-1})^{-1}(P^\top(K^\top)^{-1}Q_\eta K^{-1}\alpha_1),$$

and

$$\Sigma_0^{-1} = P^\top(K^\top)^{-1}Q_\eta K^{-1}P + (C_0 I)^{-1}.$$

The full condition distribution of $\alpha_t$ given $\alpha_0, \ldots, \alpha_{t\ -1}, \alpha_{t\ +1}, \ldots, \alpha_T, \tau_\epsilon, Q_\eta$, for $0 < t < T$, is

$$\pi(\alpha_t|\alpha_0, \ldots, \alpha_{t-1}, \alpha_{t+1}, \ldots, \alpha_T, \tau_\epsilon, Q_\eta) \sim MVN\ (\mu_t, \Sigma_t), \tag{4.12}$$

where

$$\Sigma_t^{-1} = H^\top \tau_\epsilon I H + (K^\top)^{-1}Q_\eta K^{-1} + P^\top(K^\top)^{-1}Q_\eta K^{-1}P,$$

and

$$\mu_t = (H^\top \tau_\epsilon I H + (K^\top)^{-1}Q_\eta K^{-1} + P^\top(K^\top)^{-1}Q_\eta K^{-1}P)^{-1}$$
$$(H^\top \tau_\epsilon I y_t + (K^\top)^{-1}Q_\eta K^{-1}P\alpha_{t-1} + P^\top(K^\top)^{-1}Q_\eta K^{-1}\alpha_{t+1}).$$

The full conditional distribution of $\alpha_T$ is

$$\pi(\alpha_T|\alpha_0, \ldots, \alpha_{T-1}, \tau_\epsilon, Q_\eta) \sim MVN\,(\mu_T, \Sigma_T), \tag{4.13}$$

where

$$\Sigma_T^{-1} = H^\top \tau_\epsilon I H + (K^\top)^{-1} Q_\eta K^{-1},$$

and

$$\mu_T = (H^\top \tau_\epsilon I H + (K^\top)^{-1} Q_\eta K^{-1})^{-1} (H^\top \tau_\epsilon I y_T + (K^\top)^{-1} Q_\eta K^{-1} P \alpha_{T-1}).$$

The full conditional distribution of $\tau_\epsilon$ given all other parameters is Gamma distribution,

$$\pi(\tau_\epsilon | \alpha_0, \ldots, \alpha_T, Q_\eta) \sim \Gamma(\alpha, \theta) \tag{4.14}$$

where $\alpha = a + \frac{T_n}{2}$ and $\theta = \frac{1}{2} \sum_{t=1}^{T} (y_t - H\alpha_t)^\top (y_t - H\alpha_t) + b$.

The full conditional distribution of $Q_\eta$ is Wishart,

$$\pi(Q_\eta | \alpha_0, \ldots, \alpha_T, \tau_\epsilon) \sim W_p(m, R), \tag{4.15}$$

where $m = a_\eta + T$ and $R = \beta_\eta + \sum_{t=1}^{T} (K^\top)^{-1} (\alpha_t - P\alpha_{t-1})(\alpha_t - P\alpha_{t-1})^\top K^{-1}$.

## 5 Application

We consider the error component resulting from the regression equation (3.1). The number of time points, $T$, is 1,609 and the number of sites is equal to 6. A comparison of Table 3 with Table 1 reveals that some spatial correlation has been removed but some significant correlation still exists. The next step is to constructing the KKF model for the error component.

### 5.1 Implementation of Bayesian KKF model

We apply principal component analysis for the possibility of reducing the dimensionality of the data and to look for relationships that were not previously suspected.

A spectral decomposition of the covariance matrix yields six principal components with eigenvalues 0.198, 0.149, 0.107, 0.043, 0.024 and 0.017 respectively. The first principal component explains 36.7 % of the total variation, the first two components 64.4 %, the first three components 84.4 %, the first four components 92.4 % and the first five components 96.9 %. A scree plot in Fig. 7 shows the importance of the four principal components.

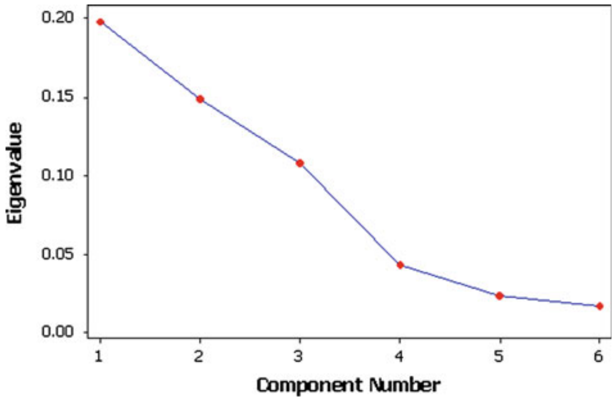

**Fig. 7** A screen plot shows the importance of the principal components

For the matrix $H$ we choose $q = 1$ because a constant field reflects overall levels in the series, and $p - q = 4$ which are the principal fields. This matrix is given by

$$H = \begin{bmatrix} 1 & -0.567 & 0.275 & 0.354 & -0.049 \\ 1 & 0.638 & -0.165 & -0.233 & -0.135 \\ 1 & -0.330 & -0.587 & -0.128 & -0.421 \\ 1 & 0.252 & 0.654 & 0.108 & -0.208 \\ 1 & -0.243 & 0.132 & -0.664 & 0.617 \\ 1 & 0.199 & -0.327 & 0.593 & 0.615 \end{bmatrix}.$$

The matrix of the distance between sites $d$ is given

$$d = \begin{bmatrix} 0 & & & & & \\ 30.66 & 0 & & & & \\ 26.99 & 8.29 & 0 & & & \\ 46.64 & 26.88 & 22.68 & 0 & & \\ 62.71 & 47.30 & 42.27 & 20.50 & 0 & \\ 39.56 & 48.44 & 52.00 & 74.54 & 94.23 & 0 \end{bmatrix}.$$

The prior specification for the unknown parameters $\sigma_{\epsilon}^2$, $\Sigma_0$and $\Sigma_{\eta}$, $\lambda = 0.05$ and $\sigma_{\gamma}^2 = 0.5$. The $P$ and $K$ matrices which are used in the state equation are obtained by identifying an ARMA model from the averages of the coefficients of ARMA models of all sites. The AIC criterion selects an order three auto-regressive model $AR(3)$ with no moving average components hence, $K$ is a zero matrix. This model is converted to a state-space model as mentioned in Aoki (1990) and it is found that

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0.675 & 0.014 & 0.058 & 0 & 0 \end{bmatrix}.$$

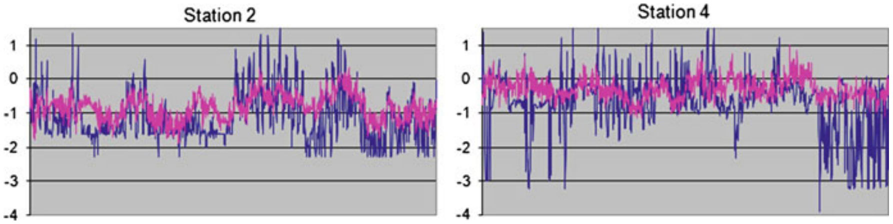

**Fig. 8** Cross validation test for Site 2 and Site 4 without using the meteorological variables

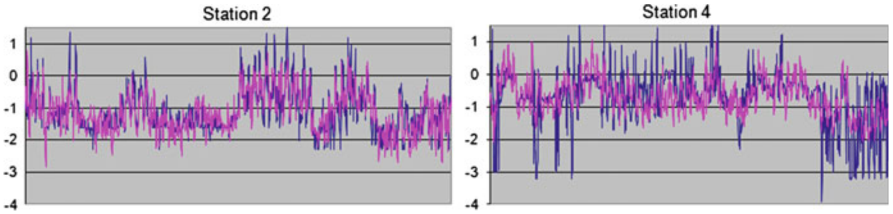

**Fig. 9** Cross validation for Site 2 and Site 4 using the meteorological variables

### 5.2 Cross validation

We apply cross validation to test the accuracy and the efficiency of our approach. The variation of the data differs between the stations. Records of Site 4 have high variance while the readings of Site 2 have low variance, therefore we use these two sites for cross validation. The predicted values for the two sites using the other stations, is obtained and compared with the actual values to observe the efficiency of our method with and without adding meteorological variables. The actual and the predicted values for Site 2 and Site 4 are plotted without using the meteorological variables in Fig. 8 and with using meteorological variables in Fig. 9. Using the meteorological variables, the resulted sum square error for Site 2 and Site 4 are 272.03 and 806.64 respectively. Without the meteorological variables, the resulted sum square error for Site 2 and Site 4 are 651.11 and 1,475.36 respectively. The significant decrease in sum square error implies good evidence of the improvement of the accuracy of the predicted data and supports the efficiency of our modified extended model where we added meteorological variables.

The error components $Y(s, t), t = 1, \ldots, 1,609; s = 1, \ldots, 6$ and the predicted values $\hat{Y}(s, t)$ using Bayesian KKF model and incorporating meteorological variables are plotted in Fig. 10 and the NMHC data $Y^*(s, t)$ and the predicted values $\hat{Y}^*(s, t)$ are plotted in Fig. 11.

Table 4 displays the sum of squares for the predicted terms of different stations for the stochastic terms $Y$ and for NMHC data $Y^*$ at the six different sites. Site 6 has the smallest sum of squares, as its data is approximately uniformly distributed followed by Site 1, 2 and 5 respectively. Site 4 has the largest sum of squares, reflecting the large variation within its data.

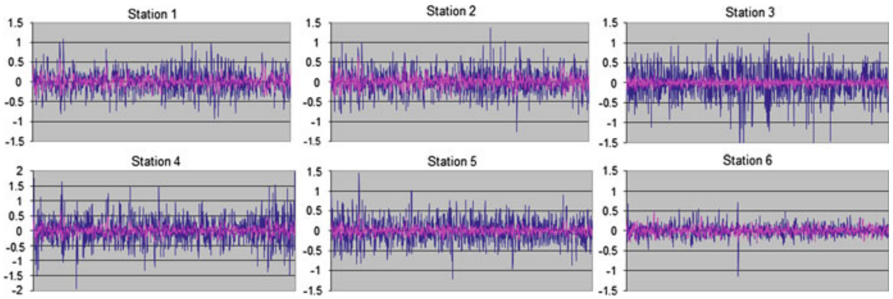

**Fig. 10** Times series plots for the stochastic residual $Y(s, t)$ and $\hat{Y}(s, t)s = 1, \ldots, 6, t = 1, \ldots, 1, 609$ for the six stations

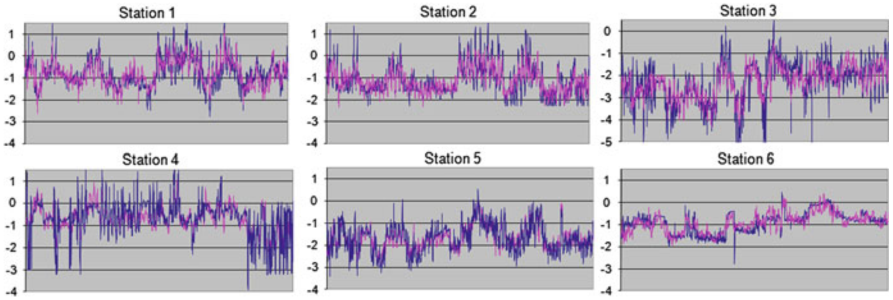

**Fig. 11** Times series plots for the NMHC data $Y^*(s, t)$ and $\hat{Y}^*(s, t)$, $s = 1, \ldots, 6, t = 1, \ldots, 1, 609$ for the six stations

**Table 4** Sum of squares for the predicted terms of different stations for the stochastic terms $Y$ and for NMHC data $Y^*$ at different stations

| | Site 1 | Site 2 | Site 3 | Site 4 | Site 5 | Site 6 |
|---|---|---|---|---|---|---|
| $Y$ | 62.42 | 74.49 | 152.54 | 254.99 | 77.51 | 33.05 |
| $Y^*$ | 245.99 | 288.27 | 535.45 | 632.34 | 224.05 | 145.97 |

## 6 Spatial prediction

Our aim is to predict the NMHC values for the unmonitored sites at any given time point. The covariates for these sites are same as before, the model is flexible to accommodate different covariate values for each station. The algorithm extends to add two additional locations plus the six existing monitoring sites in Kuwait. Figure 12 shows the proposed locations named Position 7 and Position 8. We first obtain the spatial covariance matrix $\Sigma_\gamma$ in (4.7) using the Euclidean distances between the two new locations and the other six locations through their latitude and longitude.

The forecast of $Y_t^*$ at the proposed sites is shown in Fig. 13. One can observe more fluctuation in Position 8 than in Position 7 because of the position of each site. Position 8 is located close to stations 1, 2 and 3 and is affected by these high variance stations. On the other hand, Position 7 is located between site 5 and site 6, which are more homogenous stations and hence the variation is not high.

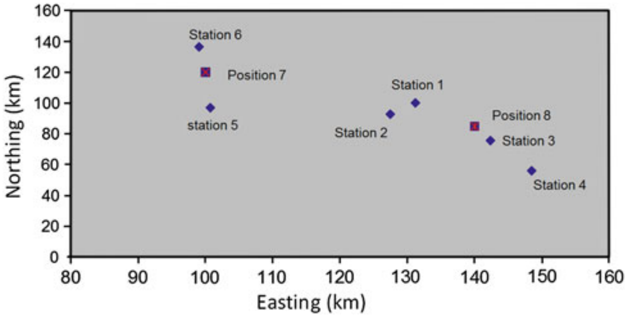

**Fig. 12** Position of the six monitor sites and the two additional sites

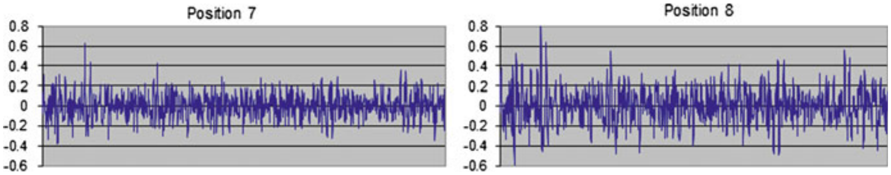

**Fig. 13** Time series for the predicted values of NMHC for the two new sites

## 7 Discussion

In this study, we have extended the Bayesian KKF model to incorporate a regression technique and accommodate meteorological variables. The spatio-temporal model is used to predict the NMHC for the unmonitored sites using the data obtained from regions of Kuwait. The accuracy of the predicted values is better when adding these covariates. The proposed model is efficient and easily accounts for the behavior of the space–time process and allows inclusion of other additional covariates. It is flexible even when applying different values for these variables for different sites. We used the model to predict the values for two unmonitored sites, but it can be extended to more sites at any given time point. Spatial-temporal data usually suffers from missing values and has high variation. To overcome the missing value problem, we applied and compared four different imputation methods and found the optimal one for this study. For the high variation problem, one can use $t$ student distribution as an alternative to the normal likelihood distribution; however, the complexity of the resulting model should be tested against the efficiency of the model.

**Acknowledgments** The authors appreciate the valuable comments of the referees and the associate editor, which greatly improved the paper.

## References

AbdulWahab SA, Bouhamra WS (2004) Diurnal variations of air pollution from motor vehicles in residential area. Int J Environ Stud 61(1):73–98

Al-Awadhi FA (2011) A multivariate prediction of spatial process with non-stationary covariance for Kuwait non-methane hydrocarbons levels. Environ Ecol Stat 18:57–77

Al-Awadhi F, Al-Awadhi S (2006) Spatial-temporal model for ambient air pollutants in the State of Kuwait. Environmetrics 17:1–14

Al-Awadhi S, Al-Awadhi F, Al-Jarrallah R (2005) Advanced statistical techniques applied for the air pollution data of Kuwait. Technical report. Kuwait University, Safat

Alsayed H (2008) Pollution level at Um-Alhayman is hazardeous for the 40,000 citizen. Alwatan newspaper, 5 Nov 2008. 16, Kuwait

Alsayed H (2009) Assessing the enviromental situation in Um-Alhayman. Alwatan newspaper, 28 Feb 2009. 4, Kuwait (2009)

Aoki M (1990) State space modeling of time series. Springer, New York

Brown PJ, Le ND, Zidek JV (1994) Multivariate spatial interpolation and exposure to air pollutants. Can J Stat 22:489–509

Carroll RJ, Chen R, George EI, Li TH, Newton HJ, Schimiediche H, Wang N (1997) Ozone exposure and population density in Harris county Texas. J Am Stat Assoc 92:392–413

Chilès JP, Delfiner P (1999) Geostatistics modeling spatial uncertainty. Wiley, New York

Cressie NAC (1993) Statistics for spatial data. Wiley, New York

Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Ser B (Methodological) 39(1):1–38

Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G (2009) Longitudinal data analysis. Part V: incomplete data. Chapman and Hall/CRC, London

Huang HC, Cressie N (1996) Spatio-temporal prediction of snow water equivalent using the kalman filter. Comput Stat Data Anal 22:159–175

Huerta G, Sanso B, Stroud JR (2004) A spatiotemporal model for Mexico City ozone levels. J Appl Stat 53(2):231–248

Journel AG, Rossi ME (1989) When do we need a trend model in kriging? Math Geol 21(7):715–739

Kalman R (1960) A new approach to linear filtering and prediction problems. Trans ASME J Basic Eng 82(7):35–45

Kalman R, Bucy R (1961) New results in filtering and prediction theory. Trans ASME J Basic Eng 83:95–108

Kibria G, Sun L, Zidek JV, Le ND (2002) Bayesian spatial prediction of random space time field with application to mapping PM2.5 exposure. J Am Stat Assoc 457:101–112

Le ND, Zidek JV (1992) Interpolation with uncertain spatial covariances: a Bayesian alternative to Kriging. J Multivar Anal 43:351–374

Mardia KV, Goodall CR (1993) Spatial-temporal analysis of multivariate enviromental monitoringmodel data. In: Multivariate enviromental statistics. Elsevier, Amsterdam, pp 347–386

Mardia KV, Goodall CR, Redfern EJ, Alonso FJ (1998) The kriged Kalman filter with discussion. Multivar Environ Stat Test 7:217–252

Meinhold RJ, Singpurwalla ND (1983) Understanding the kalman filter. Am Stat 37(2):123–127

Pollice A, Lasinio GJ (2009) Two approaches to imputation and adjustment of air quality data from a composite monitoring network. J Data Sci 7:43–59

Ripley BD (1988) Statistical inference for spatial processes. Cambridge University Press, Cambridge

Rouhani S, Mayers DE (1990) Problems in space–time kriging of geohydrological data. Math Geol 22:611–623

Rubin DB, Little RJA (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York

Sahu SK, Mardia KV (2005) Bayesian kriged Kalman model for short-term forecasting of air pollution levels. J R Stat Soc Ser C (Applied Statistics) 54(1):223–244

Shaddick G, Wakefield J (2002) Modelling multiple pollutants and multiple sites. Appl Stat 51:351–372

Tonellato SF (2001) A multivariate time series model for the analysis and prediction of carbon monoxide atmospheric concentrations. Appl Stat 50(2):187–200

West M, Harrison PJ (1997) Bayesian forecasting and dynamic models, 2nd edn. Springer, New York

Zhu L, Carlin BP, Gelfand AE (2003) Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. Environmetrics 14:537–557

Zidek J, Sun L, Le N, Ozkaynak H (2002) Contending with space–time interaction in the spatial prediction of pollution: Vancouver's hourly ambient $\mathrm{PM}_{10}$ field. Envirometrics 13:595–613

## Author Biographies

**Fahimah A. Al-Awadhi** graduated in Statistics and Computer Science from the Kuwait University, Kuwait. In 1998, she joined the Department of Mathematical Sciences at the University of Bath in Bath, UK, as a

Ph.D. candidate. She received the Ph.D. degree in Statistics from University of Bath in 2001, her dissertation titled "Statistical Image Analysis and Confocal Microscopy". From 2002 to 2006, she was an Assistant Professor in the Department of Statistics and Operations Research at Kuwait University, Kuwait. Since 2007, she has been an Associate Professor in the same University. Her research activities are focused on Bayesian Statistics, spatial temporal modeling and Stochastic Processes.

**Ali Alhajraf** graduated in Computer Science and Statistics from the University of Kuwait, Kuwait, in 1989. In 1996, he got the master degree in Statistics and Operations Research from the University of Kuwait, Kuwait. He received the Ph.D. degree in Statistics from University of Leeds, UK, in 2005. From 2005 to 2010, he was a Teacher Assistant in the Department of Statistics at the University of Kuwait. Since 2010, he has been an Assistant Professor in the College of Nursing at the Public Authority For Applied Education and Training, Kuwait. His research activities are focused on spatial temporal modeling and Monte Carlo methods in statistics inference.