Bayesian Spatio-Temporal Modelling to Deliver More Accurate and Instantaneous Air Pollution Forecasts

Sujit K Sahu

Abstract Air pollution is known to have a significant health impact particularly on people suffering from asthma and other forms of respiratory diseases. In the US ozone pollution is a huge concern during summer months because strong sunlight and hot weather result in harmful ozone concentrations in the atmosphere. Many urban and suburban areas have high levels of ozone concentrations, but many rural areas also have high ozone levels as winds carry emissions hundreds of miles from their sources. With air quality changing day to day, and even hour to hour, the challenge is to devise a model that could provide more accurate forecasts in real time. A Bayesian hierarchical space-time model is proposed and is validated to be the most accurate one that reduces forecasting error up to a third. The method combines observational air monitoring data with a forecast numerical model output to create a statistical model that could be used to provide very accurate forecast maps for the current eight-hour average and the next day maximum eight-hour average ozone concentration levels. The method is fully Bayesian and is able to instantly update the 8-hour map at the current hour (upon receiving monitor data for the current hour) and forecast the map for several hours ahead. Consequently, children and vulnerable people suffering from respiratory illnesses could gain potential health benefits by limiting their exposure to potentially harmful air pollution by reducing their outdoor activity when levels are high.

1 Introduction

Air quality changes very fast in space and time as airborne particles and harmful gases are transported by the prevailing weather conditions and human activity, such as motoring, in the immediate neighbourhood and beyond. For example, dust particles originating from the Sahara desert have been known to pollute the air in the UK and Europe in 2014 and 2015. Thus episodes in air pollution can occur in a study region for activities and natural phenomena taking place in areas even 1000s of miles apart. How then can air pollution levels be forecast accurately so that at risk

Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, UK. e-mail: S.K.Sahu@soton.ac.uk

people, i.e. children and those suffering from respiratory illnesses can be alerted to exposure risk?

Air quality models have been developed based on chemical transport models and those for atmospheric air dispersion systems. In the United State of America (USA), national air quality forecasts and near real-time predictive spatial maps are provided to the general public through the EPA-AIRNow web site: http://airnow.gov/. Current and next day particulate matter and ozone (O_3) air quality forecasts for over 200 U.S. cities are now provided on a daily basis. These forecast maps, however, are based primarily on the output of a computer simulation model known as the Eta-CMAQ model, see e.g. http://www.epa.gov/asmdnerl/CMAQ/. These models use emission inventories, meteorological information, and land use to estimate average pollution levels for gridded cells (12 km^2) over successive time periods. However, it is well known that these computer models may produce biased output and, as a result, this may lead to inaccurate pollution forecasts [3].

Monitoring data, on the other hand, provide much better air quality information since those are based on actual measurements and thus are free from biases in the computer model output. However, the monitoring sites are often sparsely located and irregularly spaced over large areas such as the Eastern US which is the study region of interest in this chapter. The sparsity limits accurate air quality information for areas away from the monitoring sites. Surely, from an individual's view point the most relevant air quality information must be the one where he/she lives or works and not at or near the monitoring sites. The problem of finding accurate air quality information in space and time still remains even after obtaining data from a monitoring network. This problem is further exacerbated by the need to forecast air quality so that preventive steps can be taken to limit exposure.

The need for prediction of air quality in both space and time naturally leads to the consideration of statistical modelling as candidate solutions. The main contribution behind the current impact case study is the development of a statistical spatio-temporal model that combines information from both the numerical model (Eta-CMAQ) and real time data from the monitoring sites. The model, implemented in a Bayesian inference framework, is computationally efficient and produces instantaneous forecast maps of hourly ozone concentration levels. The space-time model lends itself to closed form analytic Bayesian posterior predictive distributions for spatial interpolation of ozone concentration level for the past hours, current hour and forecast for future hours. These predictive distributions provide instantaneous spatial interpolation maps which could be used in a real-time environment such as the U.S. EPA AIRNow system. The predictive distributions are used to obtain the eight-hour average map which is the average of the past four hours, current hour and three hours ahead. The forecasts are evaluated by using the model fitted to a two week test data set.

2 Models

Modelling development is for observed data from the, n say, monitoring sites denoted by $\mathbf{s}_1, \ldots, \mathbf{s}_n$ where each $\mathbf{s}_i$ is described either by a latitude and longitude pair or equivalently a northing and easting pair. Observed data often have high variability which causes problems in prediction (e.g. a negative value) using Gaussian error distribution. To address that, we model data on the square-root scale but report all predictions at the original scale for ease of interpretation. Let $Z(\mathbf{s},t)$ denote the observed square-root ozone concentration, in parts per billion (ppb) units at location $\mathbf{s}$ and at hour t for $t = 1,\ldots,T$ where we take $T = 168$ corresponding to a seven day modelling period that captures a full weekly cycle.

The Eta-CMAQ forecasts are proposed to be used as a regressor in the model so that we can use the best estimates so far to train the model. These forecasts fill in the gaps in space where monitoring data are not available and the regression method improves the accuracy by using these in conjunction with the ground truth revealed by the observations.

There is, however, a potential problem in using the Eta-CMAQ forecasts since those correspond to an average value on a 12-kilometre square grid-cell while the monitoring data are observed at a point level, $\mathbf{s}$, described by a latitude-longitude pair. This general problem is the 'change of support problem' and the method used to solve the problem is known as 'downscaling', see e.g. [1] and [2]. We follow [5] and use $x(\mathbf{s},t)$ (in ppb units) to denote the square-root of the Eta-CMAQ ozone forecast value at the unique grid cell covering the site $\mathbf{s}$ and at time t.

Ozone concentration data often shows strong diurnal patterns and we model using a different hourly intercept for each of the 24 hours in a day. Let $\xi(t) = \beta_j$ denote the hourly intercept, where the hour $t(=1,\ldots,T)$ corresponds to the jth hour of the day, $j = 1,\ldots,24$. In addition, a weekday/weekend indicator, $q(t)$ taking value 1 if the hour t is within a weekday and 0 otherwise is also used as a regressor.

The full model is written as observed data as the total of a mean and a random error and is given by:

$$Z(\mathbf{s}_i,t) = \beta_0 x(\mathbf{s}_i,t) + \xi(t) + \beta_q q(t) + w(\mathbf{s}_i,t), \quad i = 1,\ldots,n, \quad t = 1,\ldots,T, \tag{1}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_{24}, \beta_q)'$ contains $p = 26$ unknown regression parameters and $w(\mathbf{s}_i,t)$ is a space-time error term.

The error term $w(\mathbf{s}_i,t)$ is assumed to be a zero-mean spatio-temporal process with a covariance structure, given by:

$$\mathrm{Cov}\big\{w(\mathbf{s}_i,t_k),\ w(\mathbf{s}_j,t_l)\big\} = \sigma_w^2\ \rho_s(|\mathbf{s}_i - \mathbf{s}_j|;\phi_s)\ \rho_t(|t_k - t_l|;\phi_t). \tag{2}$$

We write $\mathbf{w}$ to denote the vector of all the nT $w(s_i,t)$'s. Let $H(\phi) = \Sigma_s \otimes \Sigma_t$ where the $n \times n$ spatial correlation matrix Σ_s has elements $\rho_s(|\mathbf{s}_i - \mathbf{s}_j|;\phi_s)$, for $i, j = 1,\ldots,n$ and $T \times T$ temporal correlation matrix Σ_t has elements $\rho_t(|t_k - t_l|;\phi_t)$, for $k, l = 1,\ldots,T$. Here $A \otimes B$ denotes the Kronecker product of the two matrices A and B. This model reduces to the usual regression model with independent errors when we take $H(\phi) =$

I, the identity matrix. This can be achieved by choosing $\rho_s(d;\phi_s) = \rho_t(d,\phi_t) = 1$ if $d = 0$ and 0 otherwise. This independent error regression model is compared with the spatio-temporal model in Section 3.

We take the two ρ's to be exponential covariance functions, i.e., $\rho_s(d;\phi_s) = \exp(-\phi_s|d|)$ and $\rho_t(d;\phi_t) = \exp(-\phi_t|d|)$. Estimation of the spatial decay parameters, ϕ_s and ϕ_t, is generally problematic since those are weakly identified by the model and in Section 3 we choose optimal values of ϕ using a validation mean square error criterion.

The Bayesian model is completed by specifying prior distributions for β and σ_w^2. For convenience, we work with the precision $\tau_w^2 = 1/\sigma_w^2$. The joint prior distribution of β, τ_w^2 is assumed to be:

$$\pi(\beta,\tau_w^2) = N\left(\beta_m, \frac{V}{\tau_w^2}\right) G(a_w, b_w),$$

where β_m, $p \times 1$, and V, $p \times p$, are suitable hyper-parameters and τ_w^2 follows the gamma distribution $G(a_w, b_w)$ with mean a_w/b_w. In our implementation we take $a_w = 2$ and $b_w = 1$ to have a proper prior specification. We take β_m to be the null vector and $V = 10^4 I$ to have a vague prior on the regression parameter β.

2.1 Posterior distributions

Model (1) can be written as

$$\mathbf{Z} \sim N\left(X\beta, \sigma_w^2 H(\phi)\right)$$

where $\mathbf{Z}$, $nT \times 1$, contains all the data and X is the associated $nT \times p$ design matrix. Any missing value in $\mathbf{Z}$ must be replaced by an appropriate average of the space-time observations. The joint posterior distribution of β and τ_w^2, $\pi\left(\beta, \tau_w^2|\mathbf{z}\right)$, is:

$$\propto \left(\tau_w^2\right)^{\frac{nT+p}{2}+a_w-1} \exp\left[-\frac{\tau_w^2}{2}\left\{(\mathbf{z}-X\beta)' H^{-1}(\phi)(\mathbf{z}-X\beta) + (\beta-\beta_m)'V^{-1}(\beta-\beta_m) + 2b_w\right\}\right].$$

By direct integration the marginal posterior distributions are obtained as follows:

$$\beta|\mathbf{z} \sim t\left(\beta^*, 2b_w^* \frac{V^*}{nT+2a_w}, nT+2a_w\right), \quad \tau_w^2|\mathbf{z} \sim G\left(nT/2 + a_w, b_w^*\right) \tag{3}$$

where

$$V^* = \left(V^{-1} + X'H^{-1}(\phi)X\right)^{-1}, \quad \beta^* = V^*\left(V^{-1}\beta_m + X'H^{-1}(\phi)\mathbf{z}\right)$$

and

$$b_w^* = b_w + \left\{\beta_m' V^{-1}\beta_m + \mathbf{z}'H^{-1}(\phi)\mathbf{z} - (\beta^*)'(V^*)^{-1}(\beta^*)\right\}/2.$$

Here $t(\mu, \Sigma, \nu)$ denotes the multivariate t distribution with ν degrees of freedom having location parameter μ and scale parameter Σ. We use the marginal posterior distributions (3) to make inference. Specifically, β^* provides the point estimates for the parameter β. We obtain a credible interval for the component, β_k, $k = 1, \ldots, p$ by using its marginal posterior distribution which is a t-distribution with $nT + 2a_w$ degrees of freedom having mean β_k^* and scale parameter λ_k^2 where $\lambda_k^2 = \frac{2b_w^*}{nT+2a_w} V_{kk}^*$ where V_{kk}^* is the kth diagonal entry of V^*. Similarly we estimate σ_w^2 by the posterior expectation $E(1/\tau_w^2|\mathbf{z}) = \frac{b_w^*}{nT/2+a_w-1}$ which follows from the properties of the Gamma distribution.

2.2 Predictive distribution for forecasting

Using the above models we interpolate the spatial surface at any time point t' in the future or in the past. Let the p-dimensional vector of values of the regression variables at this new location-time combination be given by $\mathbf{x}_0$. We first construct the joint distribution:

$$\begin{pmatrix} Z(s', t') \\ \mathbf{Z} \end{pmatrix} \sim N \left\{ \begin{pmatrix} \mathbf{x}_0' \beta \\ X\beta \end{pmatrix}, \, \sigma_w^2 \begin{pmatrix} 1 & \Sigma_{12} \\ \Sigma_{21} & H(\phi) \end{pmatrix} \right\},$$

where $\Sigma_{21} = \Sigma_{12}'$ and Σ_{12} is the nT dimensional vector with elements given by $\sigma_s(\mathbf{s}_i - \mathbf{s}')\sigma_t(t - t')$ where $\sigma_s(\mathbf{s}_i - \mathbf{s}') = \rho_s(|\mathbf{s}_i - \mathbf{s}'|; \phi_s)$ and $\sigma_t(t - t') = \rho_t(|t - t'|; \phi_t)$. Now we obtain the conditional distribution

$$Z(s', t') \,|\mathbf{z}, \beta, \sigma_w^2 \sim N\left\{ \mathbf{x}_0'\beta + \Sigma_{12} H^{-1}(\phi)(\mathbf{z} - X\beta), \sigma_w^2 \left(1 - \Sigma_{12} H^{-1}(\phi) \Sigma_{21}\right) \right\}.$$

By integrating out β and τ_w^2 from the above distribution we obtain the predictive distribution given by:

$$Z(\mathbf{s}', t')|\mathbf{z} \sim t \left(\mathbf{x}_0'\beta^* + \Sigma_{12} H^{-1}(\phi)(\mathbf{z} - X\beta^*), 2b_w^* \frac{C(\mathbf{s}', t') + \mathbf{g}' V^* \mathbf{g}}{nT + 2a_w}, nT + 2a_w \right) \tag{4}$$

where $\mathbf{g}' = \mathbf{x}_0' - \Sigma_{12} H^{-1}(\phi) X$. Observe that we model ozone on the square root scale. Hence the predictions using the posterior predictive distribution (4) will be on the square-root scale as well. We can predict on the original scale by evaluating:

$$\begin{aligned} E(Z^2(\mathbf{s}', t')|\mathbf{z}) &= \{E(Z(\mathbf{s}', t')|\mathbf{z})\}^2 + \mathrm{Var}\{Z(\mathbf{s}', t')|\mathbf{z}\} \\ &= \left\{ \mathbf{x}_0'\beta^* + \Sigma_{12} H^{-1}(\phi)(\mathbf{z} - X\beta^*) \right\}^2 + 2b_w^* \tfrac{C(\mathbf{s}', t') + \mathbf{g}' V^* \mathbf{g}}{nT + 2a_w - 2}. \end{aligned}$$

Further details of the predictive distributions and the computations are provided in [5].

3 Validation analysis

The model and the forecasts are validated using the root mean square error (RMSE) for the forecasts $\hat{Y}_j$ for the observed Y_j, on the original scale, for $j = 1, \ldots, m$ where m denotes the number of validation observations and j is the index that represent a unique space and time combination. The RMSE is given by:

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{j=1}^{m} \left(Y_j - \hat{Y}_j \right)^2}. \tag{5}$$

In our illustration, we use data from $n = 694$ sites in a study region in the eastern US. We use the RMSE criterion (5) to select the optimal values of the spatial and temporal decay parameters ϕ_s and ϕ_t. For selecting ϕ_s the candidate effective ranges ($\approx 3/\phi_s$) were taken as 3, 6, 30, 60 and 600 kilometres. For selecting the temporal decay parameter ϕ_t we searched corresponding to effective ranges of 3, 6, 9, 12 and 24 hours. The optimal selection of these two parameters the only tuning required in the whole procedure. The optimal values of these parameters must be found for each case of model based spatial interpolation and forecasting. However, the RMSE criterion cannot be calculated when it is necessary to forecast values in the future. In such cases, we recommend to use the optimal values of ϕ_s and ϕ_t for forecasting the most recent observed values by pretending those to be as yet unobserved.

Figure 1 illustrates the RMSE of the forecasts for one hour ahead at the 694 fitting sites. Here one hour ahead forecasts are obtained for 11 hours from 6AM to 4PM for 7 days. At each forecasting occasion the data from previous seven days (i.e. 168 hours) have been used and the optimal values of the tuning parameters are found using method described above. On average, the RMSEs for the Bayesian model based forecasts are a third lower than the same for the Eta-CMAQ forecasts and are about half of the same for the forecasts based on simple linear regression method. [5] illustrate the accuracy of the forecasts in further detail. In conclusion, it is expected that forecasting using optimal Bayesian space-time model will have much better accuracy than other methods which do not explictly take space-time correlation into account.

4 Discussion

Millions of Americans with respiratory illnesses could gain potential health benefits from improved air pollution forecasting methods. The air quality forecasts developed are up to three times more accurate than previous forecasts (as illustrated here) and this means that people can limit their exposure to potentially harmful air pollution by reducing their outdoor activity when levels are high. The ability to limit exposure to high levels of air pollution can have a positive impact on long-term health and also has an economic impact as the need for medication, doctors and hospital admissions is reduced.

The accuracy of the forecasts can be increased by more complex modelling as has been claimed by [4]. However, such approaches require iterative model fitting methods, such as Markov Chain Monte Carlo (MCMC), since the posterior predictive distributions are not available in closed form unlike the case here. Being iterative, the MCMC methods require considerably more execution time to estimate model parameters and the methods also need convergence monitoring, thus eliminating their potential use in real-time forecasting environments. The proposed methods can be fully automated requiring no user input or intervention. The two weeks test data set is available from the author upon request.

Acknowledgements The author sincerely thanks Prof James Vickers for help in writing this article.

References

1. Banerjee, S., Carlin, B. P., Gelfand, A. E.: Hierarchical Modeling and Analysis for Spatial Data. CRC Press, Boca Raton, 2nd edition (2015).
2. Gelfand, A. E. and Sahu, S. K. Combining Monitoring Data and Computer model Output in Assessing Environmental Exposure. In: O'Hagan, A. and West, M. (eds) Handbook of Applied Bayesian Analysis, pp 482-510, Oxford University Press, Oxford (2009).
3. Kang, D., Mathur, R., Rao, S. T., Yu, S. Bias adjustment techniques for improving ozone air quality forecasts. Journal of Geophysical Research, 113, D23308, doi:10.1029/2008JD010151, (2008).
4. Paci, L, Gelfand, A. E., Holland, D. M. Spatio-temporal modeling for real-time ozone forecasting. Spat. Stat. **1**:79-93, (2013).
5. Sahu, S. K., Yip, S. and Holland, D. M. (2011) A fast Bayesian method for updating and forecasting hourly ozone levels. Env. Ecol. Stat. **18**, 185-207, (2011).

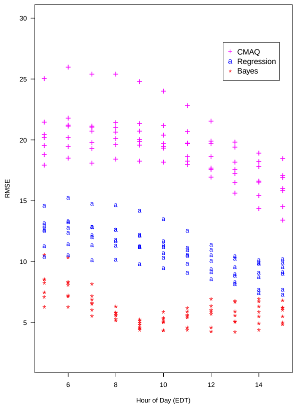

Fig. 1 The RMSE's of the forecasts for the 8-hour averages at the current hour at each hour from 6AM to 4PM for three different forecasting methods.