SPATIAL TEMPORAL MODELLING OF PARTICULATE MATTER FOR HEALTH EFFECTS STUDIES

N. A. S. Hamm

Faculty of Geo-Information Science and Earth Observation (ITC),
University of Twente, Enschede, The Netherlands - n.hamm@utwente.nl

Commission VIII, WG VIII/2

KEY WORDS: dynamic model, Gaussian process, air quality, particulate mattter, health, low-cost sensors

ABSTRACT:

Epidemiological studies of the health effects of air pollution require estimation of individual exposure. It is not possible to obtain measurements at all relevant locations so it is necessary to predict at these space-time locations, either on the basis of dispersion from emission sources or by interpolating observations. This study used data obtained from a low-cost sensor network of 32 air quality monitoring stations in the Dutch city of Eindhoven, which make up the ILM (innovative air (quality) measurement system). These stations currently provide PM10 and PM2.5 (particulate matter less than 10 and 2.5 m in diameter), aggregated to hourly means. The data provide an unprecedented level of spatial and temporal detail for a city of this size. Despite these benefits the time series of measurements is characterized by missing values and noisy values. In this paper a space-time analysis is presented that is based on a dynamic model for the temporal component and a Gaussian process geostatistical for the spatial component. Spatial-temporal variability was dominated by the temporal component, although the spatial variability was also substantial. The model delivered accurate predictions for both isolated missing values and 24-hour periods of missing values (RMSE = 1.4 μg m^{-3} and 1.8 μg m^{-3} respectively). Outliers could be detected by comparison to the 95% prediction interval. The model shows promise for predicting missing values, outlier detection and for mapping to support health impact studies.

1. INTRODUCTION

Epidemiological studies of the health effects of air pollution require estimation of individual exposure. Such studies typically aim to identify the outdoor air pollution concentration at subjects residence, school or place of work. This is then used to quantify exposure. Identifying the air pollution concentration at relevant locations, for given points or periods in time, is challenging because it is not possible to obtain measurements at all relevant locations. It is therefore necessary to predict at these space-time locations, either on the basis of dispersion from emission sources or by interpolating observations. This study uses data obtained from a low-cost sensor network of 32 air quality monitoring stations in the Dutch city of Eindhoven, which make up the ILM (innovatief luchtmeetsysteem/innovative air (quality) measurement system) (Hamm et al., 2016). These stations currently provide PM10 and PM2.5 (particulate matter less than 10 and 2.5 m in diameter), aggregated to hourly means. The data provide an unprecedented level of spatial and temporal detail for a city of this size.

Analysis of space-time data has received a lot of attention in recent years. The ILM data can be considered discrete in time (they refer to specific one-hour periods) but continuous in space. The approach taken in this paper follows Gelfand et al. (2005) and Finley et al. (2012) and treats the data as arising from time series of spatial processes, where a dynamic model describes the temporal evolution of the data. This is described further in Section 3.

Various data quality challenges arise when analysing these data. First, for each sensor there may be missing observations ranging from isolated values to a series of several weeks (e.g., if an instrument is removed for maintenance). Second, the data are typically noisy by comparison to conventional observations and this may lead to unreliable observations. It is necessary to identify these unreliable observation and remove them or correct them. Third, it is necessary to identify the requirements for the epidemiological study in terms of the spatial and temporal resolution and the precision of the interpolated values. This study addressed the first two problems.

2. STUDY AREA AND DATA

The study site is Eindhoven, a city in the south of the Netherlands (municipal population 220,000, municipal area 90 km^2). Eindhoven is home to the AiREAS initative, which is a cooperative venture that unites industry, local government, universities, small business and civic organizations towards the goal of a healthy and sustainable city (Close, 2016). As part of this initiative an innovative outdoor low-cost was installed in 2012, comprised of 32 Airboxes (Figure 1). Each Airbox contains a control and telecommunication unit as well as low-cost sensors that measure the concentration of different air pollutants. In this paper, the focus is on particulate matter (PM), particularly PM10 (PM $<$ 10μm in diameter), which is measured at every Airbox using an optical system. The locations of the sensors were chosen to reflect (i) sources of pollution (e.g., roads, junctions) (ii) places where people live and spend their time (iii) background locations. All sensors were calibrated against a beta attenuation monitoring (BAM) instrument prior to installation in autumn/winter 2013 (Hamm et al., 2016). The BAM instrument is part of the official Dutch air quality monitoring network, maintained by RIVM (National Institute for Public Health and the Environment). Background information to the ILM is given by Hamm et al. (2016). The data used for this study were hourly data for a 2-week period from 1-14 October 2014 where a complete set of hourly data for PM1, PM2.5 and PM10 were available. This led to 336 observations (24 hours × 14 days) at 32 locations.

In addition to the air quality data, hourly meteorological data

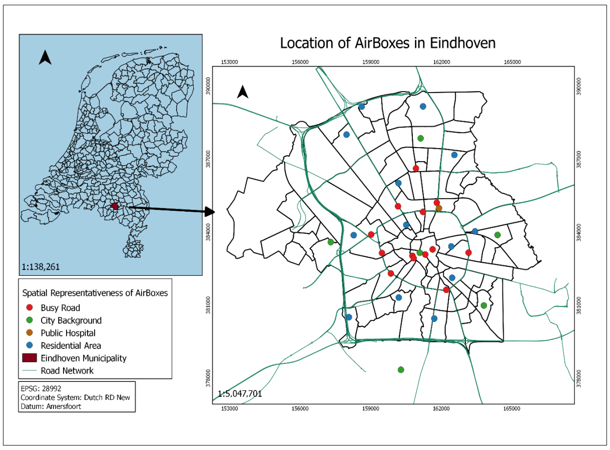

Figure 1. Location of airboxes in Eindhoven

were available for a single station in Eindhoven, maintained by the Royal Netherlands Meteorological Institute (KNMI).

3. METHODS

Consider that $y_t(\mathbf{s})$ is the observation (e.g., PM10) at location, $\mathbf{s}$, and time, t. The model is built up firstly through a *measurement equation*:

$$y_t(\mathbf{s}) = \mathbf{x}_t(\mathbf{s})'\boldsymbol{\beta}_t + u_t(\mathbf{s}) + \varepsilon_t \tag{1}$$

where $\mathbf{x}_t(\mathbf{s})$ is a vector of p covariates, which may vary in space, and $\boldsymbol{\beta}_t$ is a $p \times 1$ vector of regression coefficients which are constant in space for a given time. The term $u_t(\mathbf{s})$ is a space-time varying intercept whilst $\varepsilon_t \overset{\text{ind}}{\sim} N(0, \tau_t^2)$ is the spatially and temporally uncorrelated error. Next *transition equations* model the time varying regression coefficients, β_t:

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t,\ \boldsymbol{\eta}_t \overset{\text{iid}}{\sim} N(0, \Sigma_\eta) \tag{2}$$

and the space-time varying intercept, $u_t(\mathbf{s})$:

$$u_t(\mathbf{s}) = u_{t-1}(\mathbf{s}) + w_t(\mathbf{s}) \tag{3}$$

where $w_t(\mathbf{s}) \sim GP(0, C_t(\cdot, \boldsymbol{\theta}_t))$ for $t = 1, \ldots, n_t$ and GP refers to a spatial *Gaussian Process* where $C_t(\cdot, \boldsymbol{\theta}_t))$ is the spatial covariance function. For a covariance function with a single decay parameter, $\boldsymbol{\theta}_t = (\sigma_t^2, \phi_t)$, where ϕ is the spatial decay parameter, often referred to as the range. Considering the commonly used exponential correlation function, $C_t(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}_t)) = \sigma_t^2 \rho(\mathbf{s}_1, \mathbf{s}_2; \phi_t) = \sigma_t^2 \exp(-\phi_t ||\mathbf{s}_1 - \mathbf{s}_2||)$, where $\mathbf{h} = ||\mathbf{s}_1 - \mathbf{s}_2||$ is the Euclidean lag distance between $\mathbf{s}_1$ and $\mathbf{s}_2$.

Implementation followed the approach set out in Finley et al. (2012). The model parameters were estimated in a Bayesian framework using Markov Chain Monte Carlo (MCMC) simulation. Non-informative priors were used for all paramters: Normal distributions for the β's, uniform distributes for the ϕ's, inverse Gamma distributions for the τ^2's and σ^2's and inverse Wishart for Σ_η. The MCMC was implemented using `spBayes` (Finley et al., 2007) in the R software (R Core Team, 2016) for a chain of length 20,000 with the first 15,000 being discarded as burn-in. Chain trace plots were examined for convergence and inference was conducted using the remaining 5000 samples.

Three experiments were conducted.

1. 500 measurements were removed at random. This experiment recreates the sitution of isolated missing observations ($\sim$ 5% of all observations) that need to be filled-in in order to have a complete time series.
2. Three complete days of measurements were removed (day 8 for sensors 9, 12 and 18). This experiment recreates the situation where sensors are removed for extended periods.
3. The results were queried for outlying observations, which were then cleaned from the dataset.

4. RESULTS

For this research, space varying covariates were not available. However, exploratory analysis of the temporal variability in PM10 against the meteorological variables did not reveal any clear correlation. There was also no clear association between the type of monitoring station (e.g., background, busy street, residential street) and PM10 concentration. Subsequent analysis proceeded without covariates in Equation 1, hence $p = 1$ and $x_t = 1$.

4.1 Experiment 1

Figure 2 shows the time-series of the temporally varying mean, β_t. This can be interpreted as the dynamic time signal in the data. The time specific estimates of the variance parameters (τ_t^2, σ_t^2 and ϕ) are shown in Figure 3. There was clear evidence of spatial structure at most points in time, with a typical median value of $\sigma_t^2/(\tau_t^2 + \sigma_t^2) = 0.6$. Large values of τ_t^2 or low values of ϕ were typically associated with outliers (see Section 4.3). The temporal component was larger than the spatial component with $\Sigma_\eta = 6.13(5.24, 7.19)$ (values in parentheses give the 95% credible interval) and σ_t^2 as illustrated in Figure 3.

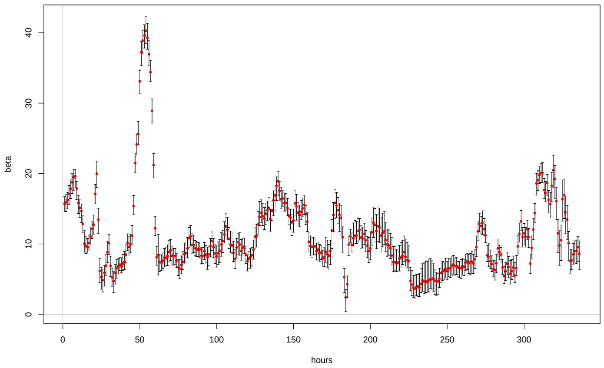

Figure 2. Variation in β_t over time (median (red dot) and 95% credible interval).

An example time-series plot of the predicted values is shown in Figure 4 for Airbox 16. Note that β_t is the same at all locations (see Equation 2, whereas $u_t(\mathbf{s})$ differs between locations (see Equation 3) and the predicted value is the sum of these two components. The isolated missing values were not observable on the prediction plots; however, the measured values were consistent with the 95% credible interval and the RMSE was 1.4 μg m^{-3} across the 500 missing values.

4.2 Experiment 2

Figure 5 shows prediction at Airboxes 9. For Airbox 9 the measurements were removed on Day 8 for the modelling stage. The median prediction closely matched the observed values (RMSE=1.8 μg m^{-3}) although the credible interval was clearly wider than at

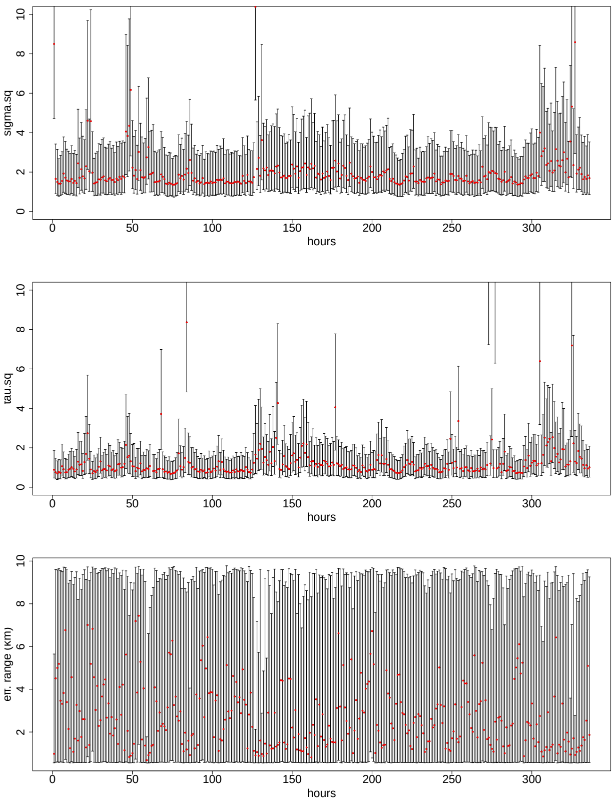

Figure 3. Variation in σ_t^2 (top), τ_t^2 (middle), and $3/\phi$ (effective range) (bottom) over time (median and 95% credible interval).

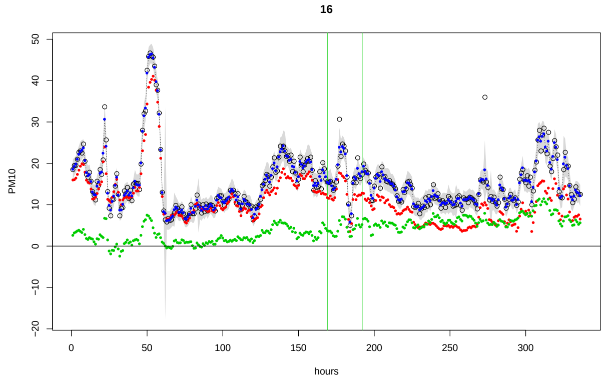

Figure 4. Predictions of PM10 at Airbox 16. Included are the 95% prediction interval (grey region), median β_t (red dots), median $u_t(\mathbf{s}_9)$ (green dots), median prediction (i.e., $\beta_t + u_t(\mathbf{s}_9)$) (blue dots), observed value (open black circles).

other points in time or for Airbox 16 (Figure 4), where observations were not removed. This is as expected, since no measurements were available to support prediction.

4.3 Experiment 3

As well as prediciting missing values, this study also yielded results that are useful for identifying outliers. Figure 4 shows that there are some isolated observations that lie outside the 95% credible interval. Following the approach of Zhang et al. (2012) these may be considered outliers. Further, the very wide credible interval at hour 62 (Figure5) arose because a single extreme value at Airbox 6 (PM10 = 83 μg m^{-3}) leading to a large estimated value for τ_{62}^2. Removing these values and re-doing the modelling removes these outliers and very large value of τ_{62}^2.

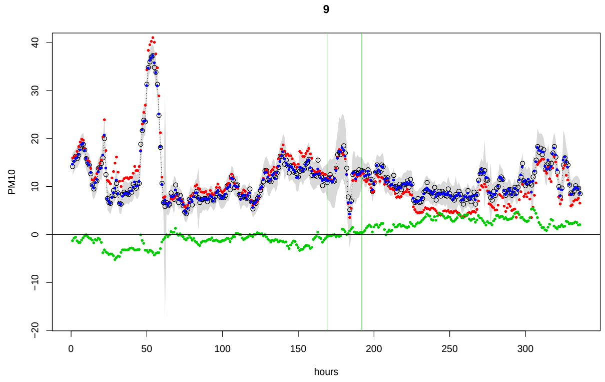

Figure 5. Predictions of PM10 at Airbox 9. Vertical lines indicate Day 8 where the observations were predicted but not included in the modelling stage. Other details are as in Figure 4.

5. DISCUSSION AND CONCLUSIONS

The lack of correlation with meteorological variables should not be taken as a general result. It may simply be that there was no correlation within this time period. This should be re-evaluated when implementing the model for other time periods. Future efforts should consider space-varying covariates as well as time-varying covariates that do not have a spatial component (e.g., via Equation 2. Space-time varying covariates have recently become available (Dash, 2016) and will be evaluated in future implementations. Dash (2016) also used a dispersion model output as a covariate, an approach which has been successful at coarser resolutions (Akita et al., 2014; Hamm et al., 2015).

The proposed model yielded accurate predictions of isolated missing values (Experiment 1) as well single days of missing values (Experiment 3). Although further evaluation is required, this approach shows promise as a method that could be used to fill in missing values and provide a complete time series to users. A further step is to make predictions at unsampled locations, leading to the production of space-time maps, as proposed by (Gelfand et al., 2005) and (Finley et al., 2012). When such maps are to be used in the context of environmental epidemiological studies, further investigation would be required to identify the required space-time resolution, as well as the required accuracy.

Finally, outliers were identified in an interactive fashion by comparing the observed values to the predictions and the associated 95% credible interval. Isolated outliers were identified and removed from the dataset. These could then be replaced by the predicted values. Extreme outliers could influence inference (e.g., hour 62 at Airbox 6), so it was necessary to re-run the analysis after removing these outliers. This approach presented in this paper requires further evaluation but shows promise as method for interactive removal of outliers. A future step would be to use it for automated outlier detection.

This paper has addressed the initial stages of space-time modelling of particulate matter for a novel low-cost sensor network that delivers observations at a fine spatial and temporal resolution. Future work needs to identify the spatial and temporal resolution that is achievable for predictive mapping. This is important because it will influence the health questions that can be addressed.

References

Akita, Y., Baldasano, J. M., Beelen, R., Cirach, M., de Hoogh, K., Hoek, G., Nieuwenhuijsen, M., Serre, M. L. and de Nazelle, A., 2014. Large scale air pollution estimation method combining land use regression and chemical transport modeling in a geostatistical framework. *Environmental Science & Technology* 48(8), pp. 4452–4459.

Close, J.-P. (ed.), 2016. *AiREAS: Sustainocracy for a Healthy City*. Springer. DOI: 10.1007/978-3-319-26940-5, Dordrecht.

Dash, I., 2016. Space-time observations for city level air quality modelling and mapping. Master's thesis, University of Twente, The Netherlands.

Finley, A., Banerjee, S. and Carlin, B., 2007. spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software* 19(1), pp. doi: 10.18637/jss.v019.i04.

Finley, A. O., Banerjee, S. and Gelfand, A. E., 2012. Bayesian dynamic modeling for large space-time datasets using Gaussian predictive processes. *Journal of Geographical Systems* 14(1), pp. 29–47.

Gelfand, A. E., Banerjee, S. and Gamerman, D., 2005. Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics* 16(5), pp. 465–479.

Hamm, N. A. S., Finley, A. O., Schaap, M. and Stein, A., 2015. A spatially varying coefficient model for mapping air quality at the European scale. *Atmospheric Environment* 102, pp. 393–405.

Hamm, N. A. S., van Lochem, M., Hoek, G., Otjes, R., van der Sterren, S. and Verhoeven, H., 2016. The invisible made visible: Science and technology. In: J.-P. Close (ed.), *AiREAS: Sustainocracy for a Healthy City: The Invisible made Visible Phase 1*, Springer, Dordrecht, pp. 51–77.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Zhang, Y., Hamm, N. A. S., Meratnia, N., Stein, A., van de Voort, M. and Havinga, P. J. M., 2012. Statistics-based outlier detection for wireless sensor networks. *International Journal of Geographical Information Science* 26(8), pp. 1373–1392.