Stat 525 Project: A Bayesian Spatial-Temporal Model for Atmospheric $PM_{2.5}$ and Mortality

Kenny Flagg

December 9, 2016

1 Introduction

Air pollution is rampant in urban regions around the world, with many adverse effects on human and environmental health. One component of air pollution is particulate matter, or solid particles suspended in the air. Particles small enough to be inhaled (diameter less than 10 μm, known as PM_{10}) can cause respiratory irritation and disease; fine particulate (diameter less than 2.5 μm, $PM_{2.5}$) can be respired into the blood stream and lead to serious conditions including asthma, bronchitis, pneumonia, lung cancer, and toxic effects of the constituent chemicals in the particulate (Charlesworth, De Miguel, and Ordóñez 2011). Setting aside possible health consequences, $PM_{2.5}$ itself has been the subject of much investigation because its exact chemical characteristics vary across space and time depending on the chemicals in the local environment, and the natural and human mechanisms that disperse particles into the air (Özkaynak et al. 2013). Therefore, individuals exposed to the same particulate concentration at different times or places may be have been exposed to different chemicals, so measurements made at air quality monitoring stations may be poor proxies for individual exposure.

Nonetheless, the association between air quality and human health at the level of a general population is an important concern for policymakers, so many authors have attempted to connect atmospheric $PM_{2.5}$ concentration with mortality among large, heterogeneous populations via sophisticated statistical models. Borja-Aburto et al. (1998) used an overdispersed Poisson regression to model daily morality counts in Mexico City as a function of total $PM_{2.5}$ and other pollutant concentrations. Dominici et al. (2002) used hierarchical Bayesian model to study the effects of $PM_{2.5}$ and PM_{10} concentration on daily mortality counts by city and region for 88 cities and seven regions in the United States. Krall et al. (2013) used a similar hierarchical model for daily mortality counts in 72 U.S. communities in six regions, but used separate concentration measurements of seven $PM_{2.5}$ constituents as predictors.

Choi, Fuentes, and Reich (2009) intended to improve upon models like those mentioned above by incorporating spatial and temporal smoothing. Their goal was to estimate the association between daily mortality counts and total atmospheric $PM_{2.5}$ concentration on several timescales across the entire state of North Carolina, and in particular to test the "harvesting hypothesis" that short-term increases in exposure have a larger effect on frail people than healthy people. They included demographic and weather variables, arguing that smoothing was necessary because the weather and particulate data were collected at different sites, and that modeling spatial and temporal trends should account for variation in $PM_{2.5}$ composition. They use their model to estimate relative risk of mortality at the county level. This paper serves to explain their model and discuss its strengths and weaknesses in the context of other epidemiological approaches.

2 The Spatial-Temporal Model

The response variable is the daily count of natural and cardiovascular mortalities for each combination of demographic variables, in each county in North Carolina, on each day of the year 2001. The explanatory variable of interest is the $PM_{2.5}$ concentration, measured in $\mu g/m^3$ every three days at 41 sites around North Carolina. Age, sex, race/ethnicity, ozone concentration, daily minimum temperature, daily maximum temperature, dew point, wind speed, and atmospheric pressure are identified as possible confounders and controlled for in the model. The authors are not clear how they evaluated the harvesting hypotheses, but from the variables described, it seems most likely that they did so by looking for an interaction between age and $PM_{2.5}$. Age was used as a categorical variable, with categories for children 14 years and under, adults 15 to 64, and seniors 65 and older.

The model is comprised of two stages (Figure 1). The first stage predicts the daily mean $PM_{2.5}$ concentration for each county. The second stage models daily mortality count with mean $PM_{2.5}$ as a predictor.

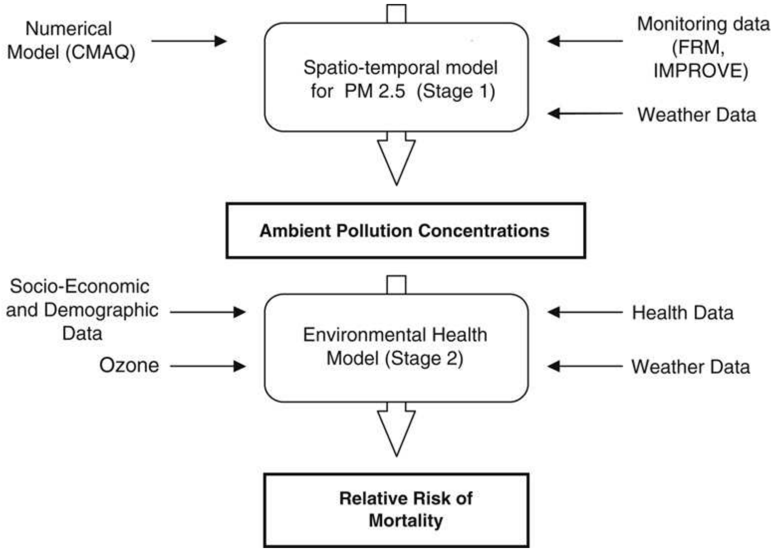

Numerical Model (CMAQ)
Spatio-temporal model for PM 2.5 (Stage 1)
Monitoring data (FRM, IMPROVE)
Weather Data
Ambient Pollution Concentrations
Socio-Economic and Demographic Data
Ozone
Environmental Health Model (Stage 2)
Health Data
Weather Data
Relative Risk of Mortality

Figure 1: Diagram of the two-stage model, from Choi et al. Multiple sources of $PM_{2.5}$ concentration measurements are combined with weather data to predict fine particulate concentration across North Carolina. The predictions are then combined with the weather data and demographic information to estimate relative risk of mortality.

Stage 1: Fine Particulate Matter

The first stage assumes that $\text{PM}_{2.5}$ is measured with error, so the observed value measured at location $\mathbf{s}$ and day t is

$$\widehat{Z}(\mathbf{s},t) = Z(\mathbf{s},t) + e(\mathbf{s},t); \quad e(\mathbf{s},t) \sim \text{N}(0,\sigma^2)$$

where $Z(\mathbf{s},t)$ is the latent true $\text{PM}_{2.5}$ concentration and $e(\mathbf{s},t)$ is measurement error. They actually had three sources of $\text{PM}_{2.5}$ data using different collection methods, so σ^2 is allowed to vary by data source.

The latent value is treated in a linear regression

$$Z(\mathbf{s},t) = \mathbf{M}^T(\mathbf{s},t)\boldsymbol{\zeta} + e_Z(\mathbf{s},t)$$

with $\mathbf{M}(\mathbf{s},t)$ being a vector of the weather covariates, and $\boldsymbol{\zeta}$ is the vector of model coefficients. The random variation term $e_Z(\mathbf{s},t)$ follows a normal distribution, with a covariance structure such that values at the same location on different days have an AR(1) structure with lag 1 correlation ψ_Z, and observations at different locations on the same day have an exponential covariance function with sill σ_Z^2, range ϕ_Z, and nugget 0. The authors selected this covariance structure after some exploratory analysis.

The prior distribution for σ is Unif$(0,5)$, based on the documentation about the measurement equipment. The priors $\sigma_Z \sim \text{Unif}(0,100)$, $\psi_Z \sim \text{N}(0,10)$, and $\phi_Z \sim \text{Unif}(0,500)$ are meant to be uninformative. The authors do not mention the prior for $\boldsymbol{\zeta}$.

This latent variable construction allows the $\text{PM}_{2.5}$ concentrations to be smoothed across space and time. The stage 1 model is used to predict $Z(\mathbf{s},t)$ across the state, and then the posterior predictions are averaged by county to produce a time series $Z_j(t)$ of daily $\text{PM}_{2.5}$ concentrations for each county j.

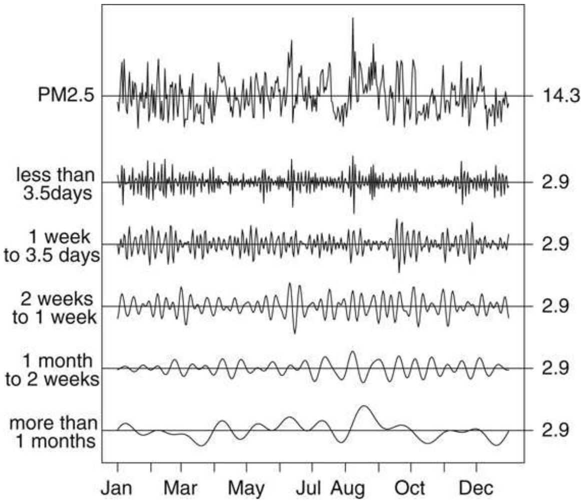


Figure 2: An example predicted $\text{PM}_{2.5}$ time series for Wake County. The complete time series (top) is decomposed into five components with variation on different timescales. The five components are used as predictors in the second stage.

The authors briefly mention (without providing details) that the weather data and $\text{PM}_{2.5}$ data were measured at different locations and on different days, so they used a similar "stage 0" model to predict $\mathbf{M}(\mathbf{s}, t)$ at the locations and times of the particulate measurements.

Stage 2: Mortality

The second stage relates the mortality counts to the $\text{PM}_{2.5}$ concentrations and possible confounders. By way of a Fourier transform, the $Z_j(t)$ time series are decomposed into 5 different timescales, to capture variation with periods less than 3.5 days, between 3.5 days and one week, one week to two weeks, two weeks to one month, and longer than one month (Figure 2).

The mortality count $Y_j(t)$ in county j on day t is assumed to follow the generalized Poisson distribution of Famoye (1993), with mean $\mu_j(t)$ and variance $\mu_j(t)[1 + \alpha\mu_j(t)]^2$, where α is a dispersion parameter. A generalized additive mixed model is used, with

$$\log(\mu_j(t)) = \gamma_j + \mathbf{x}_j^T(t)\boldsymbol{\beta} + S(\mathbf{M}_j(t)) \tag{1}$$

where γ_j is a normally distributed random intercept for county j with mean μ_{γ_j}, and $S(\mathbf{M}_j(t))$ is a smooth spline function of the meteorological variables. To account for dependence of adjacent counties, the γ_j have a conditional autoregressive structure (CAR) with scale σ_γ and reaction parameter ρ (Banerjee, Carlin, and Gelfand 2004).

The final piece of this model is a linear fixed-effect component, where $\mathbf{x}_j^T(t)$ is a row of the design matrix that includes the $\text{PM}_{2.5}$ time series components, demographic variables, and interactions. Although the notation does not make this clear, the coefficient vector $\boldsymbol{\beta}$ is allowed to vary by season (winter, spring, fall, summer) and county.

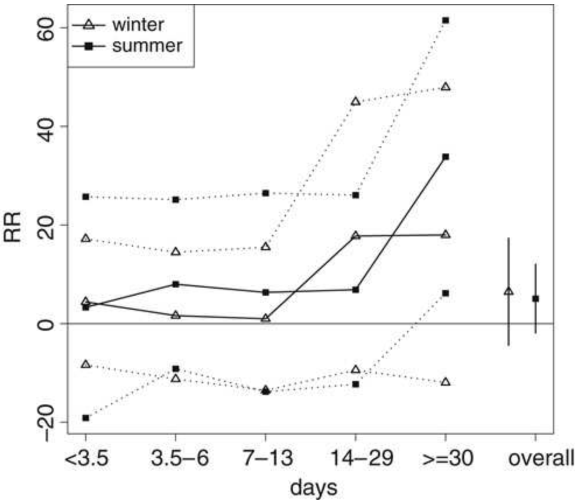
Choi et al. again chose uninformative priors, with $\mu_{\gamma_j} \sim \text{N}(0, 100)$ and $1/\sigma_\gamma^2 \sim \text{Gamma}(0.5, 0.0005)$. The prior distribution of ρ is uniform with bounds that "guarantee that the variance matrix [is] positive definite" following the suggestion of Banerjee et al. For $\boldsymbol{\beta}$, they use a special case of multivariate CAR, called the multivariate intrinsic autoregressive (MIAR) prior (Gelfand and Vounatsou 2003). The MIAR prior assumes some spatial smoothness among coefficients for adjacent counties. They do not mention the prior distribution of α.

3 Results from Choi et al.

Choi et al. fit their model using Markov chain Monte Carlo implemented in WinBUGS and R. They ran two chains for 2,000 iterations after a burn-in period of 3,000 iterations. This seems insufficient to adequately characterize the joint posterior distribution of such a complicated model (recall that this posterior includes daily $\text{PM}_{2.5}$ concentrations for every county in North Carolina), but they comment that the fitting took "a couple days" and their results section appears abbreviated, suggesting their main focus was on explaining and demonstrating the model. The authors say they assessed convergence using the Gelman-Rubin $\widehat{R}$, autocorrelation functions, and traceplots, but did not present any of these.

Results were summarized by the posterior mean and SD of log relative risk by season and timescale. This was calculated as 1,000 times the appropriate regression coefficient, and interpreted as the

Figure 3: A plot of posterior log relative risk of mortality for Wake County, showing the mean and 95% prediction bounds. Posterior log relative risk is higher for the longest timescale that for the other timescales, but only one interval excludes zero. The plots for other counties show similar results.



percent increase in mortality rate per 10 μg/m^3 increase in. PM$_{2.5}$ concentration. Posterior log relative risk was generally larger in winter and summer than in spring and fall, and larger for the longer timescales (Figure 3). The authors stated that the potential confounders did not have large interactions with fine particulate concentration on any timescale, nor did they have strong main effects. They found no evidence of harvesting.

4 Discussion

The analysis by Choi, Fuentes, and Reich has some limitations specific to their model and implementation. More generally, their study illustrates a trade-off between making broad generalizations about a heterogeneous population versus learning detailed information about certain groups or individuals. However, it has a strength in that it provides a quick summary of the association between mortality and atmospheric fine particulate across a large region containing diverse individuals.

Some issues specific to Choi et al. include difficulty in reproducing their results without knowing all of the priors used. Also, as with many studies, endless comments could be made about the choice of potential confounders to control for (smoker/non-smoker status is a curious omission).

On a similar note, total PM$_{2.5}$ concentration and mortality counts are an simplification of a complex issue, and they should be examined jointly with other outcomes and air quality variables. For example, Vinikoor-Imler, Davis, and Luben (2011) looked at the association between lung cancer incidence and PM$_{2.5}$ in North Carolina. Choi, Fuentes, and Reich's approach certainly could be used to model disease incidence and to use measurements of multiple PM$_{2.5}$ constituents.

Studies that draw subjects from a broadly-defined population will typically average over lots of potentially useful information. As Özkaynak et al. (2013) discuss, air quality monitoring stations fail to capture the details of individuals' exposure, which depends on behavior and occupation, among many other things. This study cannot identify where or how people were exposed, what they were exposed to, or what health problems they had as a result.

On the other side of the generalization versus detail trade-off, Riediker et al. (2004) looked at a select population of highly exposed individuals. They studied young male North Carolina highway patrol officers with air quality measurement equipment placed in their patrol cars. The officers had blood samples drawn daily, and were constantly monitored by electrocardiograms. This study certainly had limitations. Only nine officers were observed for four days each, but detailed information was produced about exposure to numerous different substances and many health-related variables. Such information combined with rigorous physiological or chemical explanations of the mechanics of exposure could do better at identifying risk factors than large-scale studies, particularly if used in cohort studies with groups who differ in exposure level. However, this type of study lacks the ability to make immediate inference to the general population.

Despite the drawbacks, complex models like that of Choi et al. are very useful. Organizations concerned with public health need to know where possibly hazardous fine particulate is concentrated, and when and where people are dying and getting sick. This model provides exactly that information, and can be used to identify populations for more intensive future study.

References

Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand (2004). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall.

Borja-Aburto, Víctor H et al. (1998). "Mortality and ambient fine particles in southwest Mexico City, 1993-1995". In: *Environmental Health Perspectives* 106.12, p. 849.

Charlesworth, S, E De Miguel, and A Ordóñez (2011). "A review of the distribution of particulate trace elements in urban terrestrial environments and its application to considerations of risk". In: *Environmental Geochemistry and Health* 33.2, pp. 103–123.

Choi, Jungsoon, Montserrat Fuentes, and Brian J Reich (2009). "Spatial–temporal association between fine particulate matter and daily mortality". In: *Computational Statistics and Data Analysis* 53.8, pp. 2989–3000.

Dominici, Francesca et al. (2002). "Air pollution and mortality: estimating regional and national dose-response relationships". In: *Journal of the American Statistical Association* 97.457, pp. 100–111.

Famoye, Felix (1993). "Restricted generalized Poisson regression model". In: *Communications in Statistics-Theory and Methods* 22.5, pp. 1335–1354.

Gelfand, Alan E and Penelope Vounatsou (2003). "Proper multivariate conditional autoregressive models for spatial data analysis". In: *Biostatistics* 4.1, pp. 11–15.

Krall, Jenna R et al. (2013). "Short-term exposure to particulate matter constituents and mortality in a national study of US urban communities". In: *Environmental Health Perspectives* 121 (10), pp. 1148–1153.

Özkaynak, Halûk et al. (2013). "Air pollution exposure prediction approaches used in air pollution epidemiology studies". In: *Journal of Exposure Science and Environmental Epidemiology* 23.6, pp. 566–572.

Riediker, Michael et al. (2004). “Particulate matter exposure in cars is associated with cardiovascular effects in healthy young men”. In: *American Journal of Respiratory and Critical Care Medicine* 169.8, pp. 934–940.

Vinikoor-Imler, Lisa C, J Allen Davis, and Thomas J Luben (2011). “An ecologic analysis of county-level PM2.5 concentrations and lung cancer incidence and mortality”. In: *International Journal of Environmental Research and Public Health* 8.6, pp. 1865–1871.