Using Hierarchical Bayes to Combine Spatial Air Quality Monitoring and Modeling Data

David WENDT, Steven BORTNICK, David PAUL, and Michele MORARA

Numerous scientific studies have shown associations between air pollutants and negative human health effects, as well as detrimental ecological effects. However, understanding the air pollution problem is difficult due to the interplay of diverse emission sources and complex atmospheric processes. As a result, the U.S. Environmental Protection Agency (USEPA) is interested in developing statistical models to estimate spatial gradients of air pollutants. Among other applications, such models can be used to define the spatial extent of episodes of unhealthy air quality or to illuminate relationships between different air pollutants. In this article, we present a hierarchical Bayesian approach to statistical spatial modeling. One key advantage to this approach is that it provides a means for combining two primary sources of ambient air quality data: air monitoring measurements and dispersion model predictions. By combining the monitor and model data, our approach takes advantage of the strengths of each. This results in a better understanding of the true spatial process than that provided by either information source on its own. We illustrate the approach using custom C++ code and ambient PM$_{2.5}$ concentration data for a two-week period in January 2000, collected over a large portion of the Eastern United States.

Key Words: Air pollution dispersion model; Air quality; Ambient air monitoring; Combining data; Hierarchical Bayes; Particulate matter; PM$_{2.5}$; Spatial model.

1. INTRODUCTION

Multiple scientific studies have shown associations between air pollutants, in particular airborne particles (i.e., particulate matter), and a range of acute and chronic health effects. Even at low levels, associations have been observed between air pollutants, including particulate matter, ozone, and chemical compounds designated as air toxics, and human health endpoints such as mortality, hospitalization for respiratory and heart disease, aggravation of asthma, incidence and duration of respiratory symptoms, lung function, and restricted activity (Wilson and Spengler, 1996). Sensitive groups that appear to be at greater risk to such effects include the elderly, individuals with cardiopulmonary disease such as asthma, and children (USEPA, 1999). In addition to health problems, particulate matter is the major cause of reduced visibility in many parts of the United States, and also causes soiling and damage to materials such as monuments and statues (USEPA, 2000).

1.1 COMPLEXITY OF THE AIR QUALITY PROBLEM

Air pollution, in general, has been defined as an atmospheric condition in which substances present at concentrations higher than their normal ambient (clean atmospheric) levels produce significant effects on humans, animals, vegetation, or materials (Seinfeld, 1986; Arya, 1999). Atmospheric processes such as transport, diffusion, chemical transformations, and
removal processes complicate understanding the extent of the air pollution problem. The problem is further complicated by a geographically diverse array of emission sources that produce air pollutants, ranging from urban and industrial sources such as power generation, industrial facilities, and motor vehicles; to agricultural and other rural sources such as plowing, pesticide application, and harvesting; to natural sources such as volcanic eruptions, forest fires, and sea spray (Arya, 1999).

Particulate matter air pollution, specifically that involving those particles that are less than 2.5 micrometers in diameter (PM$_{2.5}$), is a particularly complex problem to understand. PM$_{2.5}$, also referred to as fine particles, encompasses a mixture of dust, dirt, soot, smoke, and liquid droplets (USEPA, 2000). It originates from many different anthropogenic and natural sources. For example, fine particles result from direct sources such as fuel combustion from motor vehicles, construction sites, and residential fireplaces and wood stoves. PM$_{2.5}$ may also be formed in the atmosphere from the chemical transformation of gaseous emissions such as sulfur dioxide (SO$_2$), oxides of nitrogen (NO$_x$), and volatile organic compounds (VOCs). For example, PM$_{2.5}$ may form indirectly from industrial processes when gases from burning fuels react with sunlight and water vapor. In addition, PM$_{2.5}$ exhibits varying chemical and physical compositions depending on geographic location, time of year, and meteorological conditions (Main and Roberts, 2001).

### 1.2 Air Quality Spatial Modeling Needs

Given the potential for detrimental health and environmental outcomes due to PM$_{2.5}$ and other air pollutants, and the complexity of the atmospheric processes impacting the environmental behavior of such pollutants, the USEPA is interested in developing models to estimate spatial gradients of air pollutants. An important component of such models is their ability to reasonably estimate the uncertainty associated with the resulting spatial predictions. As such, the USEPA is considering various statistical approaches to the problem including, but not limited to, univariate kriging; models that account for covariates such as meteorology, topology, or other air pollutants; and hierarchical Bayesian models.

In general, statistical spatial models can be useful for illuminating relationships between different air pollutants, quantifying linkages between air pollutants and meteorology, validating air pollution dispersion models, defining the spatial extent of episodes of dangerous air quality, and evaluating the effectiveness of air monitoring networks. For example, such models can demonstrate the impact on spatial prediction uncertainty due to moving, adding, or removing air-monitoring stations. Statistical spatial models may also contribute to providing a better understanding of local, regional, national, and global sources of air pollution, and to demonstrating the trade-offs of alternative emission source control strategies.

To illustrate the utility of statistical spatial models in terms of USEPA’s needs, consider a simple illustration. Figure 1 presents a scenario of ambient air monitoring in two hypothetical U.S. counties. County B contains an ambient air pollution monitoring station while county A does not. Suppose further that the monitor in county B yields air quality measurements that would be classified in non-attainment (i.e., exceeding a designated health limit) with respect to the USEPA’s National Ambient Air Quality Standards (NAAQS) (40 CFR Part 50, 1997). Then under the USEPA’s current guidance for determining whether air quality meets the NAAQS, the
entirety of county B would be deemed in non-attainment (need reference here). Without any monitoring, county A would be assumed to have good or unknown air quality. Furthermore, given the monitor’s location and the shape of the two hypothetical counties, the non-attainment result could be considered more representative of air quality in county A than county B.

In summary, the USEPA’s current approach to determining ambient air quality can result in a lack of information for certain geographic areas (i.e., areas without monitoring stations). Furthermore, the information provided by an individual monitoring station is typically assumed to apply, without variation, to the area surrounding that station. Often, the extent of such an extrapolation is arbitrarily bound by political borders such as counties. Air pollution does not obey such boundaries. Appropriate statistical spatial models can fill in spatial gaps of air quality information and can provide a better representation of true spatial gradients. These models can also offer estimates of uncertainty associated with the modeled spatial field.

1.3 AIR QUALITY DATA

There are two fundamental approaches to generating data on ambient air quality. One approach is to deploy air-monitoring stations to directly measure concentrations of air pollutants. The main advantage of this approach in the case of many air pollutants is the relative accuracy of the measurements. The main disadvantage is the high cost associated with monitoring. It is particularly expensive to simultaneously monitor in a large number of different locations, hence, the data generated by most ambient air monitoring programs tends to be spatially sparse. Given the population exposure objectives of many monitoring programs such as those managed by the USEPA’s Office of Air Quality Planning & Standards, monitoring data also tend to be clustered in and around areas of high population density such as metropolitan statistical areas (MSAs). Although such monitoring networks provide data that are appropriate for their intended use, the data generally are not ideal for statistical spatial models.

The other approach to generating data on ambient air quality is the use of air pollution dispersion models. Such models tend to be deterministic, not statistical, in nature; although, many make use of probabilistic concepts, for example, Gaussian plume dispersion models (40 CFR Part 51, Appendix W, 1996). Dispersion models combine inventories of emissions data
with meteorological data and rely on the fundamental differential equations of the conservation of energy, mass, and momentum, which govern atmospheric motions and dynamics as well as pollutant transport and diffusion processes (Arya, 1999). The main advantage of these models is that they can generate spatially (and temporally) dense air pollutant concentration data. The main disadvantage is their potential inaccuracy due to emissions data uncertainties such as emissions calculation, reporting, completeness, and general data quality (Mangino, 1997).

1.4 A HIERARCHICAL BAYESIAN APPROACH TO COMBINING AIR QUALITY DATA

This article presents a hierarchical Bayesian approach to statistical spatial modeling. One key advantage to this approach is that it provides a means for combining the two primary sources of ambient air quality data in a defensible, probabilistic manner. The basic idea is that both sources of information, monitoring data and model predictions, represent imperfect (to varying degrees) attempts to measure some true unobserved spatial process. With this conceptual framework in mind, we develop a statistical hierarchical model that (1) accounts for the measurement uncertainties associated with each information source, (2) describes the unobserved true spatial process, including both large-scale spatial trend and small-scale variation, and (3) includes a prior distribution for a given model’s associated parameters.

Among other uses, a hierarchical Bayesian model can output a posterior predictive distribution for the true spatial process conditional on the combined sources of available air quality information. By combining the monitoring data and the model predictions, the model can take advantage of the strengths of each. Namely, the spatially dense set of model predictions are used to capture the spatial variation of the true unobserved process and the monitoring data are used, in a sense, to update the model predictions toward a more accurate depiction of the process. Combining the data sources in this manner can provide a synergistic effect, resulting in a better understanding of the true spatial process than that provided by either information source on its own.

While the air quality data and specific statistical modeling example we present in this article are unique, in general, our approach is not. An application of Bayesian melding, which suggested a similar spatial model to that presented here, combined air monitoring data and dispersion model predictions of ambient sulfur dioxide ($SO_2$) concentrations (Fuentes and Raftery, unpublished manuscript). This particular application demonstrated the utility of the approach for dispersion model evaluation, and focused on ways to address change of support issues (e.g., combining point-in-space monitoring data with spatially-averaged model predictions). In another application to combine multiple data sources, a non-Bayesian errors-in-variables approach was applied to several data sets of multi-variate, multi-media environmental measurements (Strauss, Carroll, Bortnick, Menkedick, and Schultz 2001). The authors’ approach in this case was similar to that presented here in that it made use of the concept of a true unobserved process measured with error. Finally, in an application of space-time modeling, a hierarchical Bayesian approach was used to combine wind field data from satellites and weather stations to model tropical ocean surface winds (Wikle, Milliff, Nychka, and Berliner 2001).

In general, the potential usefulness of hierarchical Bayesian modeling for the environmental sciences is well documented (Berliner 2000). More specifically, it has been demonstrated that there are numerous benefits to using a hierarchical approach, Bayesian or
otherwise, for spatial modeling (Royle and Berliner 1999). Also, in an overview of statistical advances in the environmental sciences, hierarchical Bayesian methods are presented as a useful means to combine multiple data sources (Piegorsch, Smith, Edwards, and Smith 1998). When data sources include deterministic models, Bayesian melding has been proposed to account for both the inputs and outputs of such models (Poole and Raftery 2000). Using the term Bayesian calibration, Kennedy and O’Hagan (2001) have also addressed using observed data to account for uncertainty and discrepancies in model predictions. For additional examples of hierarchical Bayesian models applied in an environmental context, refer to (Berliner, Wikle, and Cressie 2000; Wikle, Berliner, and Cressie 1998; Royle, Berliner, Wikle, and Milliff 1998).

2. BASIC MODELING THEORY

Most spatial statistical models are based on the theory of stochastic processes, just as most time series models are rooted in stochastic process theory. The one exception is a linear or non-linear regression function containing as many predictors as possible. In this case, no effort is usually made to explicitly model the error process. This works fine if the researcher is merely interested in predicting values within the domain of the data. On the other hand, forecasting either outside of the spatial domain or at a different spatial scale requires that the stochastic dependence of spatially “close” data points be taken into account. The same is true if parameter interpretation is important. Consequently, it is typically very important to account for spatial dependencies.

There are basically four broad types of spatial data:

1. Point Patterns/Lattice Data
2. Sampled Data Based on Quadrants
3. Continuous Spatial Processes
4. Sampled Data based on Rectangular Grids

At least one unique type of spatial analysis is available for each, and usually there are several between which the researcher may choose.

The purpose of this section is to provide the reader with brief introductions to several types of spatial models. There are many types of spatial models and the ones discussed here are not intended to be exhaustive. Specifically, we will consider both point patterns and continuous spatial processes, along with several types of models appropriate for each. A discussion of the Bayesian perspective is provided in the next subsection.

2.1 BAYESIAN PHILOSOPHY AND ADVANTAGES

Bayesian methods of statistical inference have multiple strengths. The label ‘Bayesian’ is derived from the fact that Bayes’ theorem provides the mechanism by which statistical inference is performed. In this context, Bayes’ theorem takes the form

\[
g(\theta | x) = \frac{f(x | \theta) g(\theta)}{h(x)}
\]  

(2.1)

whose elements have the following interpretations:
The likelihood and marginal distribution of the data are generally well understood. The elements $g(\theta)$ and $g(\theta | x)$ deserve more explanation, as many readers will be unfamiliar with their interpretation. Examining the likelihood, $f(x | \theta)$, reveals that $\theta$ refers to the parameters of the model. For example, if the data from an experiment is normally distributed, then $\theta$ refers to the mean $\mu$ and the variance $\sigma^2$ of the data. Since $\theta$ is a vector of likelihood parameters, $g(\theta)$ can only refer to the distribution of these parameters. It is at this point that Bayesian methods differ the most from frequentist methods. Frequentist inference assumes that the parameters in the likelihood are fixed, and that it is the job of the data to provide estimates of these fixed parameters. Under these assumptions, the more data one has available, the narrower the confidence intervals about these parameters will be, so that an infinite amount of actual data will yield exact estimates of these parameters.

Bayesian inference treats parameters the same way that it treats the data – parameters are themselves assumed to have a distribution, so that they are no longer fixed. Various techniques have been proposed about how to choose the distribution of these parameters. However, even if a researcher is unwilling to make a strong assumption about the location and shape of a parameter distribution, there is no impediment to analysis. He may simply choose a very flat, spread out distribution for $g(\theta)$.

The function $g(\theta | x)$ has an elegant interpretation – it represents the distribution of the parameters of the likelihood after the data $x$ has been taken into account. In a very real sense, $g(\theta | x)$ captures the information about $\theta$ that is contained in the data $x$. Also, by choosing a flat, spread out $g(\theta)$, the data $x$ are allowed to say as much as possible about $\theta$ through $g(\theta | x)$, since $g(\theta)$ does not interfere in this process.

In addition to offering elegant interpretations, Bayesian methods of inference never need to rely on asymptotic arguments. It is often the case with frequentist methods of inference that no suitable distribution for the realized portion of the sample space can be found. In these instances, frequentist inference relies heavily on determining what would happen if an infinite number of observations from the observed process happened to be available. Bayesian methods, on the other hand, can always find the distribution $g(\theta | x)$ and thereby avoid the need to appeal to theoretical, infinite sets of data – Bayesian inference makes use of only the data that is actually available. Note that Bayesian inference is not dependent on the type of analyses to be performed. Bayesian techniques are equally valid with simple univariate parameter estimation and with complex spatial modeling applications. The only difference between such scenarios appears in the level of computational complexity required to obtain an estimate.
2.2 Specific Spatial Models

Cressie (1993) and others have dedicated a great deal of effort investigating and detailing spatial models. We recognize that we cannot begin to summarize this broad literature characterizing spatial behavior in this document, thus we focus on two specific classes of methods for describing spatial relationships in this section. The first class of methods, autoregressive models, relates spatial values to their neighbors, rather than as part of the whole space. The second class of methods, trend surfaces, attempts to describe the entire surface with a set of coherent equations. Each set of methods has different strengths and weaknesses and we address the technical details of each briefly here. For each class of methods, two examples are given.

2.2.1 Trend Surfaces

A polynomial trend surface is the simplest of all spatial models to fit to data. It is an attempt to separate the large-scale trend from the error process in such a way that the resulting error process is random throughout the spatial domain. In general, a polynomial trend surface model can be written as

\[ f(\mathbf{s}) = f(x, y) = \sum_{m+n \leq p} c_{mn}x^m y^n + \varepsilon(x, y) \]  

for a two-dimensional spatial statistical model of degree \( p \). It is traditional in the frequentist literature to estimate these models using least-squares, but the Bayesian framework requires a likelihood approach. Bayesian polynomial trend surface models typically assume a normal likelihood (so that the errors are normally distributed) but this is not a requirement.

There are three important aspects to polynomial trend models. First, they are highly influenced by clusters of points in space. If points are “clumped” then a polynomial trend surface model will tend to overfit those regions of space exhibiting the “clumping.” Second, the edges of a polynomial trend model tend to fluctuate wildly as terms are added or taken away from the model. This implies that inference in the interior of the spatial domain is much safer than inference along the edges. Third, it is important when using polynomial models of degree \( p \) to keep all terms of degree \( p \) or less. Failure to do so can lead to a model that will change if the orientation of the axes is shifted.

A much more sophisticated model is the kriging model originally proposed by Matheron in 1963. The basic kriging model can be written as

\[ f(\mathbf{s}) = \mu(\mathbf{s}) + \varepsilon(\mathbf{s}) \]

\[ \text{Cov}(\mathbf{s}_j, \mathbf{s}_p) = E\left[\{f(\mathbf{s}_j) - \mu(\mathbf{s}_j)\}\{f(\mathbf{s}_p) - \mu(\mathbf{s}_p)\}\right] \]  

so that the modeler is explicitly taking into account the covariation of data values located “close” to each other in space.

The reader will note that the term \( \mu(\mathbf{s}) \) represents the large-scale spatial variation in the
data and is usually modeled by a trend surface of some type. This implies that the major
difference between polynomial trend surfaces and kriging is that kriging models try to account
for the spatial dependencies in the data through some sort of variogram or covariance structure.

The reader will also note that both of the trend surface models discussed here can be
highly useful for describing large scale behavior in spatial regions. However, as noted, they may
be most useful at small degree, where over-fitting is unlikely. It is also worth noting that each of
the models presented here can be extended naturally to include a third-spatial dimension, time, or
other covariates when the data indicates such extensions would be appropriate. In the case of
polynomial trend surfaces, the appropriate variables are included are regression terms. In the case
of kriging, techniques such as co-kriging or kriging with external drift may need to be
incorporated.

2.2.2 Auto Regressive Models

Point patterns that may be considered to derive from either a regular or irregular lattice are often
modeled using Simultaneous Auto Regressive models (SAR) or Conditional Auto Regressive
models (CAR). (Note that Auto Regressive models are closely related to ARMA or ARIMA
models in time series.) In general, CAR models are usually given as

\[
E(Z_i \mid Z_j, j \neq i) = \mu_i + \sum_j C_{ij}(Z_j - \mu_j)
\]

\[
Var(Z_i \mid Z_j, j \neq i) = \sigma^2
\]

\[
\eta_i = Z_i - \mu_i - C(Z - \mu)
\]

\[
C = [C_{ij}]
\]

where the matrix \( C = [C_{ij}] \) is defined similarly to \( S = [S_{ij}] \).

By way of contrast, SAR models can be written in the form

\[
Z(s_i) = \mu(s_i) + \sum_j S_{ij} \{ Z(s_j) - \mu(s_j) \} + \varepsilon(s_i)
\]

\[
\varepsilon(s_i) \sim \text{indep. } N(0, \sigma^2)
\]

in which the matrix \( S = [S_{ij}] \) specifies how the discrete regions in the spatial domain are related
to each other.

The reader should note the primary difference between SAR and CAR models: SAR
models focus on \( P(Z_i) \) while CAR models focus on \( P(Z_i \mid \{Z_j\}) \). It is a fact that all SAR
processes may be written as CAR processes. This follows because one may always write
\( C = S + S^T - S^T S \). The reverse is often true, but not necessarily, so that CAR processes are more
general than SAR processes.

Auto-regressive models define the values at specific points of spatial surfaces in terms of the observations nearest to the location of interest. This makes auto-regressive models particularly useful for describing small-scale variation in surfaces. In addition, this definition in terms of neighbors means that auto-regressive models are more resistant to the weaknesses of polynomial trend surfaces. They are more robust to clumping and data and less inclined to suffer unstable edge effects. Both SAR and CAR models have natural extensions to a third spatial dimension. However, only the SAR extends easily to incorporate temporal information. This apparent disparity is a direct result of the uni-directional nature of time. Namely, while it is reasonable to speak of a location being affected by any other location near it in any direction, it is less reasonable to speak of measurements taken at the current time as being impacted by the, as of yet unobserved, future.

We shall use a SAR model to represent the small-scale variation in the example presented below. Specifically, we shall let \( \mu() = 0 \), choose the spatial relationship between locations to be the same in all directions and independent of location, and use the SAR structure to model the variation term associated with the spatial variability. Specifically, we write the model used in this paper as follows:

\[
\varepsilon(s_i) \sim N(\text{mean} = \Sigma \eta(i) \rho(s_i) / N_i, \text{precision} = \tau_{AR}), \tag{2.7}
\]

where \( \eta(i) \) is the set of nearest neighbors to \( s_i \), \( \rho \) is a correlation terms describing the degree of association between neighbors, and \( N_i \) is the number of neighbors in that set. Note that we incorporate the sum into the mean of the random term in this representation of the SAR structure. This follows directly from the properties of the normal probability distribution.

### 3. PREDICTIVE MODELS AND MONITOR OBSERVATIONS

In this section, we describe the basic principles of Bayesian hierarchical modeling and how those principles might be applied to the combination of information from predictive models and monitor observations. Consider a set of observations from a predictive model, \( z_{PM}(s_i) \) for \( i = 1, \ldots, m \), and a set of observations from a monitoring network, \( z_{MO}(s_i) \) for \( i = 1, \ldots, n \). Each of these sets of observations represents observations of the true state process \( z(s) \), \( s \) in some region \( D \) of interest.

The Bayesian statistical approach offers a coherent mechanism for using scientific reasoning, learning from data, and prediction, all in a framework that manages the various sources of uncertainty. The modifier “hierarchical” refers to the development of complex stochastic models, built conditionally at various levels. The following summary of hierarchical Bayesian modeling is useful in describing this framework. The joint statistical distribution for all unknowns is developed as the product of three components, namely,

Assume we are modeling a two-dimensional process. Write \( s = (s_1, s_2) \) and define a polynomial of order \( k \) by \( s^{[k]} \theta_k \), where \( s^{[k]} = (s_1^k, s_1^{k-1}s_2, s_1^{k-2}s_2^2, \ldots, s_1s_2^{k-1}, s_2^k) \) and \( \theta_k = (\theta_{k,0}, \theta_{k,1}, \ldots, \theta_{k,k}) \). Let \([i, j]\) identify a particular grid cell and \( \mu[i, j] \) be the value of the underlying spatial process averaged across that grid cell.
3.1 MEASUREMENT UNCERTAINTY

It this example, there are two sources of information: observations from the CMAQ predictive model and observations from the FRM monitoring stations. We will examine how the observations are related to the underlying spatial process separately.

Define the observations from the predictive model, conditional on the true surface as follows:

\[ z_{PM}[i, j] = a + \mu[i, j] + \epsilon_{PM}[i, j], \]

where \( \epsilon_{PM}[i, j] \) are independent realizations of a zero-mean Gaussian random variate with precision \( \tau_{PM} \). We may later relax the assumption of independence on \( \epsilon_{PM}[i, j] \) by allowing a spatial covariance structure. We assume that \( z_{PM}[i, j] \) is defined on the same regular, partitioning lattice as \( \mu[i, j] \). More generically, \( z_{PM}[i, j] \) can be defined on tiles. We will explore this model in more detail at another time.

Note that predictive models represent good science. Our assumption of bias upon these modeling observations is meant in no way to imply otherwise. Indeed, predictive models have many strengths, probably identifying the shape of the surface and the behavior near point sources more accurately that monitoring observations. However, they are not required to be unbiased, and are likely to be less precise than the monitoring observations due to data limitations, such as missing source information.

Now consider monitoring observations, defining them as follows:

\[ z_{MO}(s) = \mu[i, j] + \epsilon_{MO}(s), \]

where \( \epsilon_{MO}(s) \) are independent realizations of a zero-mean Gaussian random variate with precision \( \tau_{MO} \). We assume the \( z_{MO}(s) \) is defined on set of points contained with in the region of interest. The reader should be aware that each \( s \) is, by definition, within a grid cell. Thus in each case, we let \( [i, j] \) be the grid cell that contains \( s \). In this formulation, the error term accounts for observation error, as well as, error associated with change of scale. (Recall that \( \mu[i, j] \) is defined for an areal average, while \( z_{MO}(s) \) refers to a point observation.)

Note that much effort has gone, and continues to go, into accurate monitoring. As a result of intensive quality control and quality assurance efforts, it is reasonable to assume that monitoring observations are unbiased and that they are measured with high precision. On the other hand, monitoring measurements may not match the support of interest. For instance, they are point measurements as opposed to grid cell averages. In addition, monitors tend to be located away from point sources and rural areas and can thus not give much information about these sorts of regions.

Notice that in the above formulations we are not saying that the \( z \)'s are marginally independent. Indeed, these data are likely to be highly dependent. Instead, we assume they are conditionally independent, given the true values of the underlying process. This conditional-independence assumption is appropriate if we can argue that the measurement errors associated with the two measurement processes are independent. This seems a reasonable assumption under
3.2 Spatial Variability

Spatial variability can be imagined as divided into two components. One component is regional scale variability. This is large-scale trend that can be described with a simple polynomial or other basic structure. The second component is neighborhood scale variability. This is a smaller scale trend with general continuity, but much more “texture” than implied by the regional trend alone. We shall use a SAR structure as described earlier to model neighborhood scale variability. Note that there are a number of ways to model spatial variability, the formulation below merely represents the model we chose for this example.

The key in this step is to build a stochastic process for the true state of the property of interest. Our basic assumption is that the property has a large-scale trend that is well approximated by a low-degree polynomial and that the small-scale trend is well described by a SAR structure. That is,

$$\mu_{i, j} = S'\Theta + \epsilon_{i, j},$$

where $S = (s_0^0, s_1^1, \ldots, s^{p_p})'$, $\Theta = (\theta_0, \theta_1, \ldots, \theta_p)$; $s$ is the centroid of $[i, j]$; and $\epsilon_{i, j}$ has a SAR structure as described earlier (2.7), with parameters $\rho$ and $\tau_{AR}$.

3.3 Parameters and Prior Knowledge

We now define the parameters that have been left undefined thus far. As we do so, we make a number of assumptions. First, we assume that the precision of the predictive model is an order of magnitude smaller than that of the monitoring data. This assumption is based expert knowledge provided by air quality scientists familiar with the practical properties of both predictive models and air quality monitors. Second, we assume bias in the predictive model, relative to the unbiased monitoring data, again based on expert knowledge discussed earlier. Third, we assume that locations in space exhibit more similar pollutant levels (Tobler’s 1st Law of Geography). This assumption leads to a restriction in the SAR model that the correlation be non-negative. Other parameter assumptions are as follows:

- $\Theta \sim \text{Gaussian}(0, \Sigma_\Theta)$. Early simulation experiments indicated that the $\theta$’s are highly correlated. Thus, we allow $\Sigma_\Theta$ to have a non-diagonal covariance structure and assume a Wishart prior.
- $\tau_{PM}$, $\tau_{MO}$, and $\tau_{AR}$ are diffusely distributed Gamma random variates, with restrictions as noted above.
- $a$ is a diffuse zero-mean Normal random variate.
- $\rho$ is distributed as a Uniform random variate, constrained to the range $(0, 1)$.

Note that wherever information about a parameter is unavailable, or limited, we use a diffuse structure to model the parameter. By diffuse, in this context, we refer to a probability distribution that spreads probability as evenly as possible across as wide as possible a selection of events.

The essence of this hierarchical approach is that we separate the issue of modeling the
data, and the errors implicit in them, from the modeling of the spatial process, the underlying state of interest. Bayes’ Theorem provides a mechanism for developing the posterior distribution of the unknowns, conditional on the observations. Specifically, our main interest in this paper is on the true state, given the data, and to a lesser extent, on the nature of the bias on the predictive model, given everything else.

The selection, or modeling, of prior distributions depends upon expert knowledge of the parameters, statistical convenience, or modeling intuition. It is often desirable to choose a prior distribution that is as “flat” as possible (the extreme case being a non-informative prior), especially in situations where little is known about the parameters of interest.

4. EXAMPLE

For the purpose of these experiments, we have chosen to focus on the average distribution small particulate matter over a period of time in a section of the southeastern United States. Specifically, we focus on January 13 through January 25 during the 2000 calendar year and on a spatial region bounded by latitudes 28.55 and 42.29 and longitudes -97.71 and -77.72. For the purpose of this paper, we assume that this region is approximately rectangular. Given the relatively small scale of the region, this is a reasonable assumption.

The modeling observational data was drawn from the set of PM$_{2.5}$ monitoring sites within the region defined above. Each of the monitoring locations reported data for a different subset of days during the targeted time period. Inspection of the frequency of the reporting indicated a three-day cycle. That is, every third day, a major of monitors appeared to provide PM$_{2.5}$ measurements. Therefore, we chose to focus on only those monitors that reported observations on all of the following five dates in January 2000: 13, 16, 19, 22, and 25. There were 256 monitors in the region of interest that reported data on all five of the days above. In the end, the data used for analysis were the average of the daily observations made January 13, January 16, January 19, January 22, and January 25, 2000, at the 256 monitoring sites chosen from the region of interest, and their associated latitudes and longitudes.

![Figure 2. Map of air quality monitor locations and grid cell centers in the spatial region of interest.](image-url)
The predictive modeling data is drawn from the results of the Multi-Stage Community Air Quality Modeling System (CMAQ). This system produces three-dimensional results, but we shall focus only on the surface data for this application. Results from the REMSAD modeling are also available, but we shall focus on CMAQ data for initial investigations. For our purposes, we average across the data generated for the dates January 13, 2000 through January 25, 2000, inclusive. Data is not available for January 23, 2000.

CMAQ is an air quality modeling system, designed to be the core of the Models-3 framework. It is a multi-pollutant, multi-scale air quality modeling system capable of simulating the transport and chemistry of ozone, particulate matter, and acidic and nutrient pollutant species across large geographic regions. The Atmospheric Modeling Division of the EPA National Exposure Research Laboratory in Research Triangle Park, North Carolina used three-dimensional Eulerian modeling simulations to develop CMAQ. It is used by regulatory modelers and model developers, researchers, state and government agencies, and academia. CMAQ has a flexible modeling structure that is capable of modeling complex atmospheric processes affecting transformation, transport, and deposition of air pollutants using a powerful user-friendly graphical interface. It accepts as input meteorological data, emissions data, photolysis rates, land-use category information, and initial and boundary conditions. CMAQ produces as the following as simulated output: air quality conditions, process rate contributing to species concentrations, process and reaction rates files, wet and dry deposition, visibility, and emissions rates.

The CMAQ model produces a matrix of results located in terms of the distance in kilometers from the grid projection center point at -90 degrees longitude and 40 degrees latitude. In this scale, the x-values range from –918 km to 1782 km and the y-values from -1242 km to 1062 km. The matrix represents an approximate uniform lattice with approximately 36 km spacing. Converting the boundaries of the monitoring observations to this ‘distance’ scale, we see that the x-values for the monitoring observations range (approximately) from -658 km to 1046 km and the y-values range (approximately) from -1274 km to 255 km. As a result, we consider only a subset of the available CMAQ data that approximates these boundaries.
4.1 Preliminary Data Investigation

In order to determine a strictly spatial empirical prior on the Process Equation, we consider the temporal average of the monitoring observations. The following table compares the rank, p, of S, the polynomial in s used to describe z(\(s\)), the number of parameters needed to specify that polynomial, the resulting Sum of Squared Errors, SSE, resulting from such a fit, and the adjusted R\(^2\) value, describing the proportion of variance explained by the model, after adjusting for the number of terms in the model. The adjusted R\(^2\) is based on \(n = 256\) data points and a total sum of squares (n) of 50,319.68.

<table>
<thead>
<tr>
<th>degree (p)</th>
<th>rank of (\Theta) (k)</th>
<th>SSE</th>
<th>adjusted R(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1889.473</td>
<td>0.9623028</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>1732.492</td>
<td>0.9650216</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>1335.479</td>
<td>0.9726005</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>1123.166</td>
<td>0.9764803</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>1036.466</td>
<td>0.9777441</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td>899.6735</td>
<td>0.9800909</td>
</tr>
<tr>
<td>7</td>
<td>35</td>
<td>793.7644</td>
<td>0.9817987</td>
</tr>
<tr>
<td>8</td>
<td>44</td>
<td>737.7217</td>
<td>0.9823657</td>
</tr>
<tr>
<td>9</td>
<td>54</td>
<td>710.6579</td>
<td>0.9821716</td>
</tr>
</tbody>
</table>

Table 1: Proposed polynomial fits to the monitoring observations and associated measures of fit.

This table suggests that even a simple linear (in \(s\)) model explains most of the variation in the process. It also suggests that a polynomial of greater than degree eight may over fit the model. In our initial explorations, we choose a linear model because, while the adjusted R\(^2\) is at a maximized at degree eight, the effective improvement is very small between degree one and degree eight. Thus, we choose, for our initial modeling, \(p = 1\).

4.2 Specific Methodology for Combining Information

Referring back to the discussion of the Bayesian paradigm earlier, observe that the paradigm has many similarities to the scientific method. The scientific method is an iterative process, wherein one starts by formulating a hypothesis that is rooted in current knowledge. Under Bayesian methods, the prior distributions represent the current knowledge about a problem and are often based on literature searches, expert opinion, and the like. In the scientific method, one collects data against which the hypothesis may be tested and then updates one’s knowledge about the truth in light of that data. This is analogous, under the Bayesian paradigm, to updating the priors using the likelihood, which is based on the observed data, to form posterior
distributions. Finally, one can repeat the scientific process. Likewise, updated posteriors, under Bayes, can serve as prior distributions for future studies.

These similarities, combined with the discussion about Bayesian methods earlier in this paper, lead us to use Bayesian techniques to combine the two sources of information presented in this example. Note however, the direct integration of the denominator of Bayes’ theorem is often difficult when a model has a large number of parameters. (Though we have not referred to them as such, this is the case in this problem, as each grid cell value is a parameter from the point of view of Bayesian estimation.) Fortunately, various methods have been developed to numerically estimate the posterior distribution. Among these methods are Monte Carlo integration (which is similar to weighted bootstrapping) and Markov-Chain Monte Carlo (MCMC) methods like Gibbs sampling and Metropolis-Hastings algorithms. Each of these methods has its own strengths and weakness and both tend to be computationally intensive and to require expert programming skills.

After evaluating the strengths and weaknesses of the various techniques, we choose to work with MCMC algorithms. As a result, we developed custom code in C to implement both Gibbs sampling and Metropolis-Hasting steps to evaluate the hierarchical model presented in this paper. These algorithms and other recently developed efficiencies allowed us to take advantage of high speed processing. The results presented for this example were computed in less than 2 minutes. That is, we are able to compute approximately 2500 grid cell observations in a matter of minutes. Scaled upward, our simulations indicate that we could conduct similar analyses on 2 million observations in less than a day. Thus, perhaps unlike earlier Bayesian efforts, the bulk of the time and effort required to complete applications such as the one presented here is in the development of good models, not in the calculation of the parameter estimates.

4.3 Results

Several advantages are obtained by rigorously combining monitoring and modeling information. The revised estimate provides a ‘better’ picture of the surface in general and allows more accurate model validation. An estimate of uncertainty is provided everywhere. The methods presented here can are fairly generic to a wide variety of environmental modeling situations. (Figures 3 and 4 are screen shots of a custom movie application demonstrating the results of this analysis. This movie can be downloaded from http://somewhere.at.Battelle.) Compare Figures 3 and 4. Observe that the peaks have been, in general, brought down, while the overall land surface has been brought up in value. This suggests that the hierarchical Bayesian predictions have captured information about background particular matter concentrations that were not included in the emissions inventories used to create the CMAQ predictions. Also observe that there are fewer monitoring observations visible above the predictive surface in Figure 4. This suggests that the overall surface is unbiased with respect to the monitors. (Rotating the image shows that about the same number of monitoring observations are below the surface as above.)
Figure 4: Plot of the Bayesian predictions (surface) and the monitoring observations (spheres).

Figure 5: Various comparisons of Bayesian predictions to original data.
Observe in the above plots that our solution is consistent with both the monitoring data and the bias-adjusted model predictions. In each graph, the predictions that result from the Bayesian analysis are consistent with the original observations (bias-adjusted as necessary). The reader may observe that the lower left hand plot indicates some potential deviation between the CMAQ predictions and the posterior cell means for higher particulate matter concentrations. We conducted a number a preliminary investigations into this phenomenon and, in each case, discarded more complicated models that did not greatly improve this fit. Also consider the plot in the lower right-hand corner of Figure 5. This illustrates the posterior distribution of a single cell prediction. In addition to recognizing that the prediction is consistent with the original monitor observation, as well as, the bias-adjusted CMAQ model prediction, this plot depicts the uncertainty associated with the hierarchical Bayesian estimate. Such a posterior density is available for every grid cell in the region of interest. (Also note that the movie discussed earlier has an option for representing the uncertainty along with the posterior mean. In this case, color illustrates the level of uncertainty while the height illustrates the mean.)

5. DISCUSSION AND CONCLUSIONS

In summary, the hierarchical Bayesian approach facilitates combining different sources and types of information in a quantitative, rigorous, and defensible fashion. These methods can provide more accurate solutions and will provide better estimates of uncertainty in many cases. In this application, we demonstrated the greatest possible extent of model-to-monitor integration. Further, as part of the process, evaluation of the predictive model came for free.

We present a relatively simple spatial model for the purposes of example in this paper. More complex spatial relationships and hierarchical relationships could be added simply to address specific concerns of a given stakeholder. For instance, a more complex bias-estimate, perhaps dependent on location, land-use, or particulate matter concentrations, might be used to address the slight deviation seen at the high end of the posterior – CMAQ prediction plot in Figure 5. While we treated correlations as unidirectional in our final SAR model for this example, directional correlations, putting different strengths of relations on the north-south direction as opposed to the east-west direction is another natural extension that we investigated in our preliminary analyses.

There are several other natural extensions for the techniques presented in this paper. One extension is incorporating spatial and temporal predictive models with spatial and temporal monitoring observations to provide spatial temporal prediction of air pollutants and other spatial processes. Another extension is the combination of information from more than two inputs. For example, there are multiple predictive models for dioxin. Combining the spatial information of all these models with dioxin monitoring observations will incorporate the strengths of all the data sources. Further extensions and applications include multimedia modeling, source apportionment, estimation of fugitive emissions, and incorporation of satellite images.

ACKNOWLEDGEMENTS
We wish to thank Shelly Eberly and Bill Cox of the USEPA, Office of Air Quality Planning & Standards for providing a data set of PM2.5 air monitoring data, and for emphasizing spatial modeling as an important tool for supporting the USEPA’s management of the air pollution problem. We would also like to thank Kirk Baker and Mike Koerber for providing a data set of PM2.5 model predictions. This research was supported solely by Battelle as part of its internal research and development program.

REFERENCES


