Handbook of Environmental and Ecological Statistics

Alan Gelfand, Montse Fuentes, Jennifer A. Hoeting, Richard L. Smith

Statistical methods for exposure assessment

Publication details


Montse Fuentes, Brian J Reich, Yen-Ning Huang

Published online on: 15 Sep 2017

How to cite: Montse Fuentes, Brian J Reich, Yen-Ning Huang. 15 Sep 2017, Statistical methods for exposure assessment from: Handbook of Environmental and Ecological Statistics CRC Press

Accessed on: 07 Dec 2023


PLEASE SCROLL DOWN FOR DOCUMENT
19

Statistical methods for exposure assessment

Montse Fuentes
Virginia Commonwealth University

Brian J Reich
North Carolina State University, Raleigh, North Carolina

Yen-Ning Huang
Indiana University, Bloomington, Indiana

CONTENTS
19.1 Defining Exposure .............................................................. 447
19.2 Spatiotemporal Mapping of Monitoring Data ................................ 449
  19.2.1 K-Nearest Neighbor (KNN) interpolation ............................ 449
  19.2.2 Inverse Distance Weighting (IDW) ................................. 450
  19.2.3 Kriging ................................................................. 450
  19.2.4 Bayesian Interpolation ................................................ 452
  19.2.5 Comparison of methods .............................................. 453
  19.2.6 Case study: PM$_{2.5}$ data in the Eastern US .................... 453
19.3 Spatiotemporal extensions .................................................... 455
19.4 Data Fusion ........................................................................ 458
  19.4.1 Calibration of Computer Models ...................................... 458
  19.4.2 Spatial Downscaler ..................................................... 459
  19.4.3 Spatial Bayesian Melding ............................................... 459
  19.4.4 Case Study: Statistical downscale of PM$_{2.5}$ .................... 460
  Bibliography .................................................................... 461

19.1 Defining Exposure

Assessing the exposure of an individual or population is a challenging statistical task [21]. Borrowing from the concept of the genome in genetics, recently the concept of an exposome [13] has emerged to help explain the complexities of exposure assessment and guide research in this area. According to the Centers for Disease Control and Prevention [1], “The exposome can be defined as the measure of all the exposures of an individual in a lifetime and how those exposures relate to health.” A statistical analysis of exposure must consider complexities such as simultaneous exposure from multiple contaminants, spatiotemporal aggregation, and multiple sources of uncertainty. A data collection plan and subsequent statistical analysis relies on clearly defining the exposure of interest, and the precise definition of exposure and exposure metrics is crucial for interpreting the results and limitations of a study.

The receptor-based definition of exposure begins by defining the concentration of the contaminant at location $s$ and time $t$, $C(s, t)$. The exposure of an individual between times
$t_0$ and $t_1$ is then defined as the cumulative amount of the contaminant that reaches the individual,

$$\int_{t_0}^{t_1} C[u(t), t] dt,$$

(19.1)

where $u(t)$ is the spatial location of the individual at time $t$. Taking this concept a step further, the dose is the amount of the contaminant that is actually ingested by the individual. Depending on the nature of the contaminant and individual, these three summaries can be quite different, and thus the concentration-response, exposure-response, and dose-response curves used to quantify the health effect of the contaminant can have substantially different meanings.

For example, in Section 19.2 we consider the example of exposure to fine particulate matter (PM$_{2.5}$). In this example, we use PM$_{2.5}$ concentrations measured at several locations in the US to estimate the concentration surface for all $s$. We note however that these estimated concentrations may only be crude estimates of exposure and dose. Even two individuals living in the same residence may have different exposures depending on their commuting pattern and amount of time spent outdoors. For a thorough discussion of individual exposure see Chapter XI. Similarly, two individuals with identical exposures may have different doses depending in their inhalation rate or metabolism.

As indicated by (19.1), the level of aggregation plays a key role in exposure assessment. In some cases, concentration data are only available in aggregates, and in other settings aggregation is performed to reduce uncertainty in the concentration estimates or to better represent exposure. The most common form of aggregation is temporal, which can be used to distinguish between acute exposure over a narrow time window, usually close to the health event of interest, and cumulative exposure over a longer period. For example, we may compute the average concentration over the previous few days to measure acute exposure to explain asthma events, or the average concentration over the first trimester to measure a mother's cumulative exposure during pregnancy. Spatial aggregation has similar benefits. A common motivation for spatial aggregation is to match the scale of a health endpoint defined at a regional level. For example, the county-average concentration may be deemed the most relevant measure of exposure to analyze the number of events in the county. Even in cases where the objective is to measure exposure for an individual, if the exact location of the individual over time (i.e., $u(t)$ in (19.1)) is unknown, a spatial average may be a reasonable proxy for exposure.

In addition to spatiotemporal aggregation, other forms of aggregation can be used depending on the nature of the contaminant. For example, PM$_{2.5}$ is defined as the total concentration of particles with diameter below 2.5 micrometers, and is thus an aggregate of a complex mixture of contaminants. The total PM$_{2.5}$ concentration can be decomposed in several interesting and useful ways. Recent results [12] suggest that ultrafine particles with diameter less than 100 nanometers may in fact have stronger health effect than larger particles. From a policy perspective, it may be more relevant to decompose PM$_{2.5}$ by source rather than size, e.g., concentration of particles attributed to motor vehicle traffic versus biomass burning [19]. Finally, PM$_{2.5}$ is a complex mixture of different types of particles, such as nitrate, sulfate, and organic carbon. There is no reason to assume that all types of particles have the same health effects, and so recent work has begun to study the effects of individual particle types [11].

The type of exposure that can be assessed is determined by the nature of the data collected. Concentration is arguably the easiest quantity to estimate. In response to the U.S. Clean Air Act, the U.S. EPA has established an extensive network to monitor the ambient concentration of several air pollutants. These networks provide hourly or daily measurements at fixed spatial locations, and combined with spatiotemporal interpolation methods
give estimates of the concentration surface $C(s, t)$ or aggregated summaries. However, with these indirect measurements it is difficult to assess the exposure of an individual. Direct measurements such as wearable devices are less abundant and more expensive, but provide valuable information about individual exposure, and the relationship between concentration and exposure. Measurements of the dose ingested by the individual typically require even more invasive sampling, such as hair or blood samples, but provide the most reliable data to establish dose-response curves.

### 19.2 Spatiotemporal Mapping of Monitoring Data

A key task in an exposure assessment is to estimate the concentration function $C(s, t)$ (denoted in this section as $Y(s, t)$ as is common in the spatiotemporal literature) for relevant locations $s$ and times $t$. One approach is to develop a method to estimate the concentration for all $s$ and $t$. The estimated concentration function can then be used to monitor changes over time, identify spatial hotspots, extract concentrations for specific locations, and compute aggregate measures such as (19.1). Collecting data for all spatiotemporal locations is often infeasible (with some notable exceptions such as remote sensing, satellite, or computer simulation data; see Section 19.4), and we focus on the case where data are collected at a finite number of locations and statistical algorithms are used to interpolate the measurements to create a concentration function applicable for all $s$ and $t$. In this section we discuss three popular spatial interpolation methods: $K$-nearest neighbors, inverse distance weighting, and Kriging. Spatial methods for a single time point are discussed in general, with emphasis on quantifying uncertainty, and then demonstrated with an application to PM$_{2.5}$ data. Spatiotemporal extensions are also provided.

#### 19.2.1 $K$-Nearest Neighbor (KNN) interpolation

The simplest algorithm to estimate concentration at an unmonitored locations is to use the measured value for the nearest monitoring location. The estimated nearest-neighbor concentration surface is then piecewise constant over Voronoi polygons [14] around the monitors. This is a special case of $K$-nearest neighbor (KNN) interpolation (with $K = 1$). Let $Y(s)$ be the observed data at location $s \subset \mathbb{R}^2$ and $s_1, \ldots, s_N$ be the $N$ monitor locations. The KNN estimate at an unobserved location $s_0$ is the average of $K$ nearest observations. If $d_i(s_0) = ||s_0 - s_i||$ denotes the distance between $s_0$ and $s_i$, then the KNN estimator can be written

$$\hat{Y}(s_0) = \sum_{i=1}^{N} w_i Y(s_i)$$

where $w_i = \frac{1}{K}$ if $s_i$ belongs to the nearest $K$ neighbors of $s_0$ and 0 otherwise. The variance of $\hat{Y}(s_0)$ is

$$\text{Var}[\hat{Y}(s_0)] = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \text{Cov}(s_i, s_j)$$

where $\text{Cov}(s_i, s_j)$ is the spatial covariance function. Applying this variance requires an estimate of the spatial covariance function. As described further in Section 19.2.3, if the covariance function is known then Kriging is the optimal method for spatial prediction. Therefore, in KNN is typically conducted without assuming that the covariance function is known, making uncertainty difficult to quantify. In addition, it has been shown that using the nearest monitoring site to estimate air pollutant exposures performs poorly in complex urban environments [10].
19.2.2 Inverse Distance Weighting (IDW)

Inverse distance weighting (also known as kernel smoothing) is an extension to KNN that weights locations according to their distance from the prediction location. The IDW estimate of $Y(s_0)$ has the form of (19.2) but with a different choice of weights. To encode spatial structure, the weights decrease with the distance from $s_0$. The weights are often written as a function of symmetric kernel function $k(s_i, s_0) = k(s_0, s_i) \geq 0$ and forced to sum to one, i.e., $\sum_{i=1}^{N} w_i(s_0) = 1$, using the transformation

$$w_i(s_0) = \frac{k(s_i, s_0)}{\sum_{j=1}^{N} k(s_j, s_0)}. \quad (19.4)$$

There are many possible kernel functions. Two popular options are the Gaussian (squared exponential) kernel $k(s_i, s_0) = \exp\{-[d_i(s_0)/\rho]^2\}$ and the power kernel $k(s_i, s_0) = d_i(s_0)^{-1/\nu}$, where $\rho > 0$ is the kernel bandwidth and controls the degree of spatial smoothness. For large datasets, these kernels can be slow because they are a global average over all $N$ observations. When $N$ is large, computation can be improved by using a compact kernel that is zero for large $d_i(s)$ so that the IDW estimate becomes a local average. As with the choice of covariance function in Kriging (Section 19.2.3), it is important to carefully choose the parametric form of the kernel function $k$ and bandwidth parameter $\rho$, and test for sensitivity to these specifications. Also, variance of the IDW estimate is given by (19.3), and is thus difficult to compute without estimates of the spatial covariance function.

19.2.3 Kriging

Kriging is a powerful interpolation method described in detail in Chapter X, and briefly here. Kriging can be viewed as the predictions that arise from modeling the data as a realization of a Gaussian process defined by its mean function $E[Y(s)] = \mu(s)$ and covariance function $\text{Cov}[Y(s), Y(s')] = \rho(s, s')$; in this section we assume that the covariance function is stationary and isotropic, i.e., $\rho(s, s') = \rho(h)$ where $h$ is the distance between $s$ and $s'$. This assumption implies that the covariance is the same for all pairs of points separated by the same distance, and is invariant to translations and rotations. A Kriging analysis proceeds by estimating the mean and covariance function in the model-building step, and with estimates in hand the Kriging equations are applied for prediction at new locations.

We assume parametric forms for both the mean and covariance functions. Let the mean function be $\mu(s) = X(s)^T \beta$, where $X(s)$ is a $p$-vector of spatial covariates. For example, in Chapter X2 the spatial land-use regression models include local covariates such as distance to a roadway or a point source. Depending on assumptions about the mean function, Kriging can be classified as simple, ordinary, or universal Kriging. In simple kriging we assume the mean function $\mu(s)$ is known at all locations and so Kriging prediction can be applied on the mean-zero residuals. In ordinary Kriging the mean is an unknown constant $\mu(s) = \beta_0$ that needs to be estimated from the data, and universal Kriging includes covariates $X(s)$ and requires estimating the coefficient vector $\beta$.

Many options are available, but we use the Matern spatial covariance function [18]

$$\rho(h) = \begin{cases} \tau^2 + \sigma^2 & \text{for } h = 0 \\ \sigma^2 \frac{\Gamma(\nu-\frac{1}{2})}{\Gamma(\nu)} \left( \frac{2 \sqrt{h}}{\rho} \right) K_{\nu} \left( \sqrt{2 \frac{h}{\rho}} \right) & \text{for } h > 0 \end{cases} \quad (19.5)$$

where $K$ is the modified Bessel function of the second kind. The Matern has four parameters $\theta = (\tau^2, \sigma^2, \rho, \nu)$: the nugget $\tau^2 > 0$ effects only variance at distance zero, e.g., measurement error; the partial sill $\sigma^2 > 0$ dictates the overall spatial variance; $\rho > 0$ is the spatial range.
and determines the strength of long-range spatial correlation; and $\nu > 0$ measures the local spatial correlation and smoothness.

Closely related to the covariance function is the variogram, defined as $E\{[Y(s) - Y(s')]^2\} = \gamma(s, s')$. For the stationary and isotropic model the variogram is a function of only the distance between locations, i.e., $\gamma(s, s') = \gamma(h)$. When the covariance exists, it is related to the variogram as $\gamma(h) = \rho(0) - \rho(h)$. For example, the Matern variogram is $\gamma(0) = 0$ and

$$
\gamma(h) = \tau^2 + \sigma^2 \left[ 1 - \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \sqrt{\frac{2\nu}{\rho}} h \right) K_\nu \left( \sqrt{\frac{2\nu}{\rho}} h \right) \right] \quad (19.6)
$$

for $h > 0$.

A useful graphical tool to determine whether a particular spatial covariance function such as the Matern fits the data well is the empirical variogram. For example, under the assumption that the process is intrinsically stationary with isotropic variogram $\gamma(h)$, the non-parametric empirical variogram is [5]:

$$
\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} [Y(s_i) - Y(s_j)]^2 \quad (19.7)
$$

where $N(h)$ denotes the set of pairs $(i, j)$ such that $||s_i - s_j|| = (h - \epsilon, h + \epsilon)$, $\epsilon > 0$ is the bin width, and $|N(h)|$ is the number of pairs in $N(h)$. The empirical variogram is computed and plotted for a grid of distances $h \in \{h_1, h_2, \ldots\}$ and compared to the model-based variogram $\gamma(h)$ in (19.6) to determine if the parametric model is appropriate.

With the assumptions of normality, stationarity, and parametric mean and covariance functions, the likelihood of the observations at the $N$ sample locations $Y = [Y(s_1), \ldots, Y(s_N)]^T$ is simply the multivariate normal with mean vector $X\beta$ and covariance matrix $\Sigma(\theta)$, where $X = [X(s_1)^T, \ldots, X(s_N)^T]^T$ is the $N \times p$ covariate matrix and $\Sigma(\theta)$ is the $N \times N$ matrix with $(i, j)$ element equal to $\rho(||s_i - s_j||)$ evaluated at $\theta$. The likelihood function is then

$$
l(\beta, \theta) = -\frac{1}{2} \log(|\Sigma(\theta)|) - \frac{1}{2} (Y - X\beta)^T \Sigma(\theta)^{-1} (Y - X\beta).
$$

The parameters $\beta$ and $\theta$ can therefore be estimated by maximum likelihood analysis, although this can be slow for large $N$ because it requires many operations on the $N \times N$ covariance matrix $\Sigma(\theta)$.

Given parameter estimates $\hat{\beta}$ and $\hat{\theta}$, the Kriging prediction for unmonitored location $s_0$ is the conditional mean of $Y(s_0)$ given the monitor data $Y$,

$$
\hat{Y}(s_0) = X(s_0)\hat{\beta} + \Sigma_0(\hat{\theta})\Sigma(\hat{\theta})^{-1} \left( Y - X\hat{\beta} \right) \quad (19.8)
$$

where $\Sigma_0$ is the $N$ vector of covariances between $Y(s_0)$ and $Y$. Similarly, the prediction variance is simply the variance of the condition distribution of $Y(s)$ given $Y$,

$$
\text{Var}[\hat{Y}(s_0)] = \rho(0) - \Sigma_0(\hat{\theta})\Sigma(\hat{\theta})^{-1}\Sigma_0(\hat{\theta}). \quad (19.9)
$$

Alternatively, the Kriging predictor can be written

$$
\hat{Y}(s_0) = X(s_0)^T\hat{\beta} + \sum_{i=1}^{N} w_i(s_0) \left[ Y(s_i) - X(s_i)^T\hat{\beta} \right] \quad (19.10)
$$

where the weights $w_i(s_0)$ are the $N$ elements of $\Sigma_0(\hat{\theta})\Sigma(\hat{\theta})^{-1}$, which can be determined using
the variogram function $\gamma$ evaluated at $\hat{\theta}$. Therefore, as in (19.2) the Kriging prediction (19.8) is a linear combination of the observations. Rather than specifying the weights directly as in KNN or IDW, the Kriging weights are determined indirectly via estimates of the mean and covariance functions. It can be shown that under the assumed mean and covariance functions, these weights are optimal in the sense that they give the best linear unbiased predictions (BLUP).

Kriging naturally leads to spatial-average concentration estimates and uncertainty quantification for these aggregate estimates. Let $Y(B) = \frac{1}{|B|} \int_B Y(s) ds$ be the spatial-average concentration over region $B$ (with area $|B|$). The integral can be approximated arbitrarily well using a finite average

$$Y(B) \approx \frac{1}{J} \sum_{j=1}^{J} Y(s_{0j})$$

where the $J$ grid points $s_{0j}$ cover $B$ and $Y_0 = [Y(s_{01}),...,Y(s_{0J})]^T$ are the concentrations at the $J$ grid points. The Kriging estimate for $Y_0$ is

$$Y_0 = X_0 \hat{\beta} + \Sigma_0(\hat{\theta})\Sigma(\hat{\theta})^{-1}(Y - X\hat{\beta})$$  \hspace{1cm} (19.11)

where $X_0$ is the $J \times p$ covariate matrix for the grid points and $\Sigma_0(\hat{\theta})$ is the $J \times N$ matrix defining the covariance between $Y_0$ and $Y$ with $(j,i)$ element defined by $\rho$ with parameters set to $\hat{\theta}$. The predicted value is then

$$\hat{Y}(B) = \frac{1}{J} \sum_{j=1}^{J} \hat{Y}(s_{0j})$$

where $\hat{Y}(s_{0j})$ is the Kriging prediction in (19.8). The variance of this estimate is

$$\text{Var}[\hat{Y}(B)] = \frac{1}{J^2} \sum_{j=1}^{J} \sum_{k=1}^{J} \text{Cov}[\hat{Y}(s_{0j}),\hat{Y}(s_{0k})],$$ \hspace{1cm} (19.12)

where $\text{Cov}[\hat{Y}(s_{0j}),\hat{Y}(s_{0k})]$ is the $(j,k)$ element of $\text{Cov}(\hat{Y}_0) = \Sigma_0(\hat{\theta}) - \Sigma_0(\hat{\theta})\Sigma(\hat{\theta})^{-1}\Sigma_0(\hat{\theta})^T$ for $J \times J$ covariance matrix $\Sigma_{00}(\hat{\theta})$.

### 19.2.4 Bayesian Interpolation

The covariance matrix, $\Sigma$, in the kriging method is treated as known. The estimated covariance, $\Sigma$, replaces $\Sigma$ in the kriging formula resulting in the classical kriging interpolator. In this case the uncertainty in the covariance structure is ignored which leads to overconfidence in the interpolated values. In Bayesian kriging, uncertainty about $\Sigma$ (as well as all other parameters in the model) is quantified. For instance, [22] apply universal Bayesian kriging to estimate the residential exposure to air pollutants in a high-risk area providing an assessment of model uncertainty.

In classical kriging, $\theta$ are estimated using likelihood approaches or empirical methods and then treated as known for interpolation. In Bayesian kriging, we introduce a prior distribution on the parameters and update the distribution using the data. The solution to Bayesian kriging is the posterior predictive distribution of $Y(s_0,t_0)$ given the observations $Y$:

$$\pi(Y(s_0,t_0)|Y) \propto \int \pi(Y(s_0,t_0)|Y,\theta)\pi(\theta|Y)d\theta$$ \hspace{1cm} (19.13)
Statistical methods for exposure assessment

where \( \pi(\theta | Y) \) is the posterior of the model parameters \( \theta \). As a helpful reference, [2] provide a tutorial for Bayesian kriging using WinBUGS.

Spatial models often assume the outcomes follow normal distributions. However, the Gaussian assumption in the standard kriging approaches may be overly-restrictive for air pollution data, which often display erratic behavior, such as rapid changes in time or space. Bayesian nonparametric methods avoid dependence on distributional parametric assumptions by working with probability models on function spaces. In particular, [16] introduce a nonparametric model extending the stick-breaking prior of [17], which is frequently used in Bayesian modelling to capture uncertainty in the parametric form of an outcome. The stick-breaking prior is extended to the spatial setting by assigning each location a different, unknown distribution, and smoothing the distributions in space with a series of kernel functions. This results in a flexible spatial model, as different kernel functions lead to different relationships between the distributions at nearby locations. This model is also computationally convenient because it avoids inverting large matrices, which often hinders analysis of large spatial data sets.

19.2.5 Comparison of methods

We have described three procedures for spatial prediction: \( K \)-nearest neighbors (KNN), inverse distance weighting (IDW), and Kriging. All three methods use a linear combination of the observations as prediction, all three have tuning parameters (number of neighbors, choice of kernel/covariance, etc) that need to be selected carefully. KNN and IDW have the clear advantage of simplicity, and can be computed efficiently for large datasets because (assuming compact kernel for KNN) they use only local measurements for prediction. Limitations of these methods are that it is difficult to incorporate spatial covariates and challenging to quantify uncertainty without declaring a parametric covariance function. If a parametric covariance is assumed, then Kriging is the natural choice because it results in the optimal spatial predictions under the presumed model. Kriging also has the advantage of being able to include spatial covariates and, especially using the Bayesian framework, full characterization of uncertainty. Therefore, despite the mathematical and computational burden, Bayesian Kriging is our preferred method for spatial interpolation.

19.2.6 Case study: PM\(_{2.5}\) data in the Eastern US

We use PM\(_{2.5}\) data from August 21, 2006 to illustrate the spatial interpolation methods. Figure 19.1 (top) plots the data at the 229 monitoring location in the Eastern US. The data are pooled over the Interagency Monitoring of PROtected Visual Environment (IMPROVE) network and the Speciation Trends Network (STN).

Spatial interpolation based on the nearest neighbor and inverse distance weighting (with Gaussian kernel and bandwidth chosen based on 5-fold cross validation) are plotted in Figure 19.2. The nearest neighbor interpolation is piecewise constant, whereas the IDW interpolation is smooth. Bayesian Kriging estimates are plotted in Figure 19.3 (top row). In this analysis the mean function at location \( s = (s_1, s_2) \) is taken to be \( \mu(s) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 \) and the spatial covariance is Matern with nugget \( \sigma^2 \), partial sill \( \tau^2 \), range \( \rho \) and smoothness \( \nu \). We select flat prior for the mean coefficients, and \( \sigma^2, \tau^2, \rho \sim \text{InvGamma}(0.01,0.01) \) and \( \nu \sim \text{Unif}(0.5,10) \). The predictions are fairly similar to the IDW predictions. The prediction standard deviation is the largest in Texas, Mississippi, and North Dakota where the monitors are the most sparse.

The interpolated surfaces in Figures 19.2 and 19.3 can be aggregated to estimate the state-average concentrations, \( \frac{1}{|B|} \int_B C(s) \, ds \), where \( B \) is the region that defines the state,
FIGURE 19.1
Fine particulate matter concentrations from the STN and IMPROVE monitoring networks (top) and CMAQ model output (bottom) on August 21, 2006.
FIGURE 19.2
Fine particulate matter concentrations from the STN and IMPROVE monitoring networks on August, 21 2006 interpolated using the nearest monitoring station and smoothed using inverse distance weighting.

\[ |B| \] is the state’s area, and \( C(s) \) is the concentration at location \( s \). The integrals are approximated in Figure 19.4 by computing estimates of \( C(s) \) on a 200 × 200 rectangular grid of points covering the spatial domain and averaging the predictions within each state. The IDW and Kriging estimates both show that the largest state-averages are in the southeast, with the largest average in Alabama. The largest uncertainty is in Texas and Mississippi where the data are sparse. However, standard deviation of the state averages are considerably smaller than the standard deviations at individual locations (Figure 19.3), illustrating a benefit of spatial aggregation. The top left panel of Figure 19.4 shows the simple average of all monitors in the state. The state with the highest simple average is Mississippi, which has only a single monitor (Figure 19.1). The contrast between the simple average and the Kriging estimates illustrates the advantage of spatial smoothing to borrow strength across nearby observations and appropriately weight the monitors.

19.3 Spatiotemporal extensions

Next, we present two possible approaches, suggested by [9], to include temporal interactions in the kriging model, either by considering time as the third dimension, or by fitting a covariance function with both spatial and temporal components.

It is straightforward to generalize the idea of purely spatial kriging to spatiotemporal kriging with time as the third dimension. In the case of purely spatial process, \( h \) in (19.7) can be calculated using Euclidean distances. If we add time as the third dimension, we can adapt the calculation of \( h \) to accommodate the additional dimension as long as we take into account the different scales between the temporal dimension and the spatial dimension. [9] give more details on methods to optimize the scaling.

The second approach for including temporal interactions is to introduce a spatiotemporal covariance function. The empirical spatiotemporal semivariogram can be extended from (19.7), as follows,

\[
\hat{\gamma}(h,u) = \frac{1}{2|N(h,u)|} \sum_{N(h,u)} (Y(s_i, t_k) - Y(s_j, t_l))^2
\]

(19.14)
FIGURE 19.3
Estimated (and prediction standard deviation) particulate matter concentration via Bayesian Kriging (“KR”; top) and Bayesian downscaling (“DS”; middle) for August, 21 2006. The bottom row plots the difference in predicted values and ratios of prediction standard deviations.
FIGURE 19.4
Estimated state-average fine particulate matter concentration via sample means of the monitor values (top left), and means of interpolated values using inverse distance weighting, Bayesian Kriging and Bayesian downscaling for August, 21 2006. The final row maps standard deviation for the Bayesian methods.
where in this case \( N(h) \) denotes the set of pairs of observations separated by distance \( h \) and time \( u \). Model fitting of the theoretical parametric spatiotemporal variogram can be performed under different assumptions of the spatiotemporal covariance function. The solution for simple spatiotemporal kriging is:

\[
\hat{Y}(s_0, t_0) = \mu(s_0, t_0) + k_0' \Sigma^{-1}(Y - \mu) \tag{19.15}
\]

where \( Y = (Y(s_1), \ldots, Y(s_n))' \) represent all the spatiotemporal observations, \( \mu = E(Y) \), \( \mu(s_0, t_0) = E[Y(s_0, t_0)] \), \( \Sigma = \text{Cov}(Y) \) and \( k_0' = \text{Cov}[Y(s_0, t_0), Y] \). The solutions for spatiotemporal ordinary kriging and universal kriging can be derived similarly as in purely spatial case. When the observations are measured with error, we can add \( \tau^2 I(h = 0, u = 0) \) to the spatiotemporal covariance function to account for the nugget effect.

There exists a rich literature on the study of spatiotemporal covariance functions, \([6]\) provide examples of various spatiotemporal covariance functions with different properties.

### 19.4 Data Fusion

The use of monitoring data facilitates the estimation of trends of pollution levels across space and time. However, the air quality physical models can be a valuable and powerful tool to addressing gaps if not enough monitoring data are available. The air quality models, based on the dynamics and mechanics of atmospheric processes, typically provide information at higher temporal and spatial resolution than data from observational networks. Though, errors and biases in these deterministic models are inevitable due to simplified or neglected physical processes or mathematical approximations used in the physical parameterization. Therefore the introduction of statistical models to combine different sources of information, physical models and observations, known as data fusion, is a powerful tool to obtain improved maps of air pollution and reduce the prediction error of the air pollution predictive surfaces.

#### 19.4.1 Calibration of Computer Models

To compensate for the limitations of the field data and understand the behavior of complex physical processes, deterministic model outputs are increasingly used in environmental research. As opposed to statistical models, deterministic models are simulations based on differential equations which attempt to represent the underlying physical processes. Using a large number of grid cells, they generate averaged concentrations with high spatial and temporal resolution and without missing values. Ideally, such model outputs would help fill the space-time gaps in observational data. However, the model output are simulations, and the uncertainty about them should be characterized. The various sources of uncertainty are classified as low quality of emissions data, model inadequacy and residual variability. As result, to use model output, particularly to provide assistance in control strategy selections, it may be necessary first to calibrate the model.

Air quality models, such as EPA’s Community Multi-scale Air Quality (CMAQ) model \([4]\), provide one of the most reliable tools for air quality assessment. CMAQ simulates ambient air concentrations of numerous pollutants on a myriad of spatial and temporal scales. These model simulations are designed to support both regulatory assessments by EPA program as well as scientific studies. One particularly important use of numerical models is forecasting, for which model outputs are used as predictor values in statistical models when
monitoring measurements are not available in the future. The evaluation of physically based computer models, in particular in the context of air quality applications is crucial to assist in control strategy selection. A rigorous statistical assessment of model performance is needed, since selecting the wrong control strategy has costly economic and social consequences. The objective comparison of mean and variances of modeled air pollution concentrations with the ones obtained from observed field data is the common approach for assessment of model performance.

However, outputs from numerical models of geophysical phenomena are different from observations for a number of reasons, including problems with model inputs, imperfect and/or inadequate descriptions of processes in the model and errors from numerical simulation methods [20]. It is also worth to note that the measurements are made at specific locations, whereas CMAQ concentrations represent a volume-average value (corresponding to the volume of the grid cell). This discrepancy in spatial representativeness is a fundamental source of uncertainty when calibrating models [7]. Thus, quantifying the discrepancies and subsequently adjusting outputs in order to achieve compatibility with the observations are critical to improve and utilize numerical models.

A standard approach to calibration is to simply shift and scale the numerical model output by matching the sample mean and standard deviation. One drawback of this strategy is that it fails to calibrate properly the tails of the modeled air pollution distribution, and improving the ability of these numerical models to characterize high pollution events is of critical interest for air quality management. Therefore, due to the differences in the tail of the distributions and the desire to accurately model extreme events, calibrating sample quantile levels instead of raw data has been applied in the literature [23].

19.4.2 Spatial Downscaler

A spatial downscaler is a statistical framework to combine disparate spatial data from observations and physical models in order to improve the spatial prediction of a process of interest. [3] develop a fully model-based strategy within a Bayesian framework to downscale air quality numerical model output to a point level. In their static spatial model, the observations are regressed on the numerical model output using spatially-varying coefficients that are specified through a correlated spatial Gaussian process.

Let $\tilde{Y}(B, t)$ be the CMAQ values over grid cell $B$ at time $t$, all the points $s$ within the same grid cell are assigned the same CMAQ value. The observed data $Y(s, t)$ and the CMAQ output are related as follows:

$$Y(s, t) = \beta_0(s, t) + \beta_1(s, t)\tilde{Y}(B, t) + \epsilon(s, t)$$ (19.16)

where the bias terms $\beta_0(s, t)$ and $\beta_1(s, t)$ are functions of space and time. In order to interpolate to a new location of interest $s_0$ at time $t_0$, we obtain the posterior predictive distribution of $Y(s_0, t_0)$ given the data $Y(s_i, t_j)$ and $Y(B_k, t_j)$.

19.4.3 Spatial Bayesian Melding

[8] develop a Bayesian approach to address the air quality numerical models evaluation problem, and show how it can be used to remove the bias in the numerical model output. This Bayesian framework provides a natural approach to compare data from very different sources taking into account different uncertainties, and it also provides posterior distributions of quantities of interest that can be used for scientific inference. The monitoring data is not treated as the “ground truth”. Instead, it is assumed that there is some smooth underlying (but unobserved) field that measures the “true” concentration/flux of the pollutant at each location. Monitoring data are these “true values” plus some measurement
error. The CMAQ output can also be written in terms of this true underlying (unobservable) process, with some parameters that explain the bias and microscale noise in CMAQ. The truth is assumed to be a smooth underlying spatial process with some parameters that explain the large scale and short scale dependency structure of the air pollutants. CMAQ is evaluated by comparing the distribution of the monitoring data at a given location, to the predictive posterior distribution of CMAQ at that given point in space and time. The bias in CMAQ is quantified by obtaining the posterior distribution of the bias parameters given the monitoring data and CMAQ output.

More specifically, we model the true underlying air pollution of interest using both observed data and numerical model output. We do not consider ground measurements to be the “true” values because they are measured with error. Thus, we denote the observed pollution values at location $s \in D_1$ on day $t \in D_2$ from the monitoring stations by $Y(s, t)$, where $D_1 = \{s : s_1, \ldots, s_N\} \subset \mathbb{R}^2$ and $D_2 = \{t : 1, \ldots, T\} \subset \mathbb{R}$, and it is modeled as

$$Y(s, t) = \beta_0 + \beta_1 \tilde{Y}(B) + \epsilon(s),$$

where $Z(s, t)$ is the unobserved “true” underlying spatial-temporal process at location $s$ and at time $t$. The measurement error $e_F(s, t) \sim N(0, \sigma_F^2)$ is assumed to be independent of the true underlying process.

Since the CMAQ values are averages over grid squares, not point measurements, we model the CMAQ values, $\tilde{Y}(B, t)$, where subregions $B_1, \ldots, B_b$ cover the spatial domain $B$, as follows:

$$\tilde{Y}(B_i, t) = a(B_i) + \frac{1}{|B_i|} \int_{B_i} Z(s, t) ds + e_N(B_i, t),$$

where $a(B_i)$ is the additive bias of the CMAQ output in subregion $B_i$ and is assumed to be a polynomial function of the centroid of the subregion, $s_i$, with a vector of coefficients, $a_0$.

The true underlying process $Z$ is modeled as a function of the weather covariates:

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

where $M(s, t)$ is a vector of meteorological variables.

In order to predict $Z(s_0, t_0)$, the true pollution values at space $s_0$ and time $t_0$, given the data, we obtain the posterior predictive distribution of $Z(s_0, t_0)$ given the data.

The hierarchical structure of the Bayesian melding approach, allows a characterization of the different types of error and biases in the sources of information, data and physical models, with regard to an underlying true process. Those different type of error structures and biases due in part to the change of support problem get combined and become unidentifiable in the downscaler approach.

### 19.4.4 Case Study: Statistical downscale of PM$_{2.5}$

Figure 19.1 (bottom) plots the 12km × 12km CMAQ model output for August 21, 2006. Both the CMAQ model output and monitoring data have the same broad-scale trends with the largest concentrations in Kentucky and the southeast. There is a strong linear relationship between a monitor’s observation and the CMAQ output from the grid cell that contains the monitor (Figure 19.5, left). However, there are clear differences between these two data products. The CMAQ output displays several local features not seen in the monitor data, such as local hotspots in Montana and Wyoming. Also, the CMAQ output underestimates the extremely large air pollution values (Figure 19.5, right).

To combine the two data sources, we fit the model in (19.16) but with spatially-constant intercept and slope, i.e.,

$$Y(s) = \beta_0 + \beta_1 \tilde{Y}(B) + \epsilon(s),$$
where $\epsilon$'s Matern covariance function and all priors are the same as the Bayesian Kriging model described in Section 19.2.6. The differences between Bayesian Kriging and Bayesian downscaling methods are shown in Figure 19.3 (bottom row). The advantage of incorporating CMAQ is evident in the 15\% reduction in prediction standard deviation in Texas where monitoring data is sparse.

The local CMAQ hotspots in Montana and Wyoming are also present in the downscaling predictions. [15] explore the value of incorporating local CMAQ output by comparing the agreement between monitoring data and model output for different spatial resolutions. They find that CMAQ output and monitor data have similar large-scale spatial patterns, but disparate fine-scale structure, and propose a multi-resolution downscaler to avoid using CMAQ inappropriately at fine scales.

**Bibliography**


