Topics in Bayesian Spatiotemporal Prediction of Environmental Exposure

by

Philip A. White

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Alan E. Gelfand, Supervisor

_____
Fan Li

_____
Colin Rundel

_____
Benjamin Goldstein

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2019

Abstract

Topics in Bayesian Spatiotemporal Prediction of Environmental Exposure

by

Philip A. White

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Alan E. Gelfand, Supervisor

_____
Fan Li

_____
Colin Rundel

_____
Benjamin Goldstein

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2019

Abstract

We address predictive modeling for spatial and spatiotemporal modeling in a variety of settings. First, we discuss spatial and spatiotemporal data and corresponding model types used in later chapters. Specifically, we discuss Markov random fields, Gaussian processes, and Bayesian inference. Then, we outline the dissertation.

In Chapter 2, we consider the setting where areal unit data are only partially observed. First, we consider setting where a portion of the areal units have been observed, and we seek prediction of the remainder. Second, we leverage these ideas for model comparison where we fit models of interest to a portion of the data and hold out the rest for model comparison.

In Chapters 3 and 4, we consider pollution data from Mexico City in 2017. In Chapter 3 we forecast pollution emergencies. Mexico City defines pollution emergencies using thresholds that rely on regional maxima for ozone and for particulate matter with diameter less than 10 micrometers (PM_{10}). To predict local pollution emergencies and to assess compliance to Mexican ambient air quality standards, we analyze hourly ozone and PM_{10} measurements from 24 stations across Mexico City from 2017 using a bivariate spatiotemporal model. With this model, we predict future pollutant levels using current weather conditions and recent pollutant concentrations. Employing hourly pollutant projections, we predict regional maxima needed to estimate the probability of future pollution emergencies. We discuss how predicted compliance to legislated pollution limits varies across regions within

Mexico City in 2017.

In Chapter 4, we propose a continuous spatiotemporal model for Mexico City ozone levels that accounts for distinct daily seasonality, as well as variation across the city and over the peak ozone season (April and May) of 2017. To account for these patterns, we use covariance models over space, circles, and time. We review relevant existing covariance models and develop new classes of nonseparable covariance models appropriate for seasonal data collected at many locations. We compare the predictive performance of a variety of models that utilize various nonseparable covariance functions. We use the best model to predict hourly ozone levels at unmonitored locations in April and May to infer compliance with Mexican air quality standards and to estimate respiratory health risk associated with ozone exposure.

Acknowledgements

I want to thank Alan Gelfand, my advisor, for his work in mentoring and working with me. With his help, I feel I've developed greatly as a statistician. In addition, I want to thank my committee, Fan Li, Colin Rundel, and Ben Goldstein, for their work, time, and feedback to help improve my dissertation. I also want to thank others who I have worked with while at Duke who have helped me to develop as a statistician, including Emilio Porcu, Shane Reese, William Christensen, Candace Barrett, and Shannon Tass.

I would like to thank my friends and family who have given me so much support during this time. I want to thank my parents for all they have done for me. Most of all, I want to thank my wife for her support and sacrifices made for me during our time in Durham. I'm also grateful for joy that my boys, Simon, Asher, and Ezra brought to me during this time.

Contents

List of Tables

List of Figures

1

Introduction

1.1 Spatial and Spatiotemporal Data

Spatial data can generally be categorized into three types or groups: areal unit or block-level data, point-referenced (also called geocoded or geostatistical) data, and point pattern data. Each data type is treated with different types of models, and analyses of these different types often have different goals. In this dissertation, we only consider point-referenced and areal unit data. Throughout, I let $\mathcal{D} \subset \mathbb{R}^2$ denote the spatial domain.

When modeling spatiotemporal data, time must be treated as either discrete or continuous. Data in discrete-time are commonly modeled using autoregressive, moving average, or dynamic linear models. However, in many settings continuous time models are necessary or more natural. Differences in modeling discrete and continuous time are similar to the differences in modeling areal unit data and point-referenced data. In this dissertation, we consider both discrete and continuous time models for spatiotemporal data.

1.1.1 Areal Unit Data and Markov Random Fields

Areal unit or block-level data are common in many applications, including disease mapping, small area estimation, and image analysis. For these data, we denote observations as $Y(B_i)$ as some total or average over the block, where $B_i \subset \mathcal{D}$. For simplicity, we sometimes denote this as Y_i. Most often, areal unit data are modeled using Markov random fields (MRF) with the goal of spatial smoothing. However, when areal unit data arise from the average or aggregation of a continuous spatial process, then Gaussian process (GP) models may be appropriate. When process models are used, block-to-point or block-to-block predictions are readily available using multivariate normal/kriging theory (see Banerjee et al., 2014). Connections between these two modeling approaches are discussed in Lindgren et al. (2011). Further discussion on GPs is deferred to Section 1.1.2.

Markov random fields define a joint distribution of random variable through conditional distribution $p(V_i|V_j, j \neq i)$, where we let $\mathbf{V} = \{V_1, ..., V_n\}$ denote random variables defined by an MRF. The joint distribution for $\mathbf{V}$ is obtained by Brook's lemma (Brook, 1964). Any MRF has a corresponding Gibbs distribution, and the reverse is true as well (Hammersley and Clifford, 1971; Geman and Geman, 1984). Interactions or *potentials* $\phi^{(k)}$ between random variables are specified on *cliques* (groups) of size k. Specifically, the joint distribution,

$$p(V_1, ..., V_n) \propto \exp\left\{\gamma \sum_k \sum_{\boldsymbol{\alpha} \in \mathcal{M}_k} \phi^{(k)}(V_{\alpha_1}, ..., V_{\alpha_k})\right\}. \tag{1.1}$$

For cliques of size $k = 1$, there are no interactions (an observation only interacts with itself), and random variables are independent. For $k = 2$, there are two-way interactions, and the most common potential is the squared pair-wise difference $(V_i - V_j)^2$. Higher order cliques are rare and yield very complicated distributions. We use MRFs for spatial random effects as a part of a hierarchical model. Thus, we

use models similar to

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + V_i + \epsilon_i, \tag{1.2}$$

where $\mathbf{x}_i^T$ are covariates with corresponding coefficients $\boldsymbol{\beta}$, V_i are specified with by an MRF, and ϵ_i is noise or error.

Dating to Besag (1974), the conditionally autoregressive model (CAR) model is a common choice for hierarchical spatial models. Here, we introduce the Gaussian case that I use in this dissertation, where we assume that

$$p(V_i | V_j, j \neq i) = \mathcal{N}\left(\frac{1}{w_{i+}} \sum_j w_{ij} V_j, \frac{\tau^2}{w_{i+}}\right), \tag{1.3}$$

where w_{ij} represents weights or adjacencies, $w_{i+} = \sum_j w_{ij}$, and τ^2 is a scale parameter. Here, I assume symmetric weights $w_{ij} = w_{ji}$; however, this is not required. Weights w_{ij} are chosen depending on the problem. For example, weights (i) may simply be binary (1 if blocks are neighbors and 0 otherwise), (ii) could incorporate distance in some way, or (iii) could be specified to approximate partial differential equations. The joint distribution for the CAR model can be written two ways:

$$p(V_1, ...V_n) \propto \exp\left[-\frac{1}{2\tau^2} \mathbf{V}^T \mathbf{Q} \mathbf{V}\right] \tag{1.4}$$

$$\propto \exp\left[-\frac{1}{2\tau^2} \sum_{i \neq j} w_{ij} (V_i - V_j)^2\right], \tag{1.5}$$

where $\mathbf{Q} = D_w - W$, $(D_w)_{ii} = w_{i+}$, and the i^{th} row and j^{th} column of W is w_{ij}. The matrix W is often called the weights or adjacency matrix. Because $\mathbf{Q1} = \mathbf{0}$, meaning that this distribution is improper. When used hierarchically as a prior distribution for spatial random effects, the posterior distribution of $\mathbf{V}$ is proper.

Although it is not necessary to use a proper prior distribution for hierarchical spatial random effects, many proper CAR models are proposed in the literature.

The simplest modification is the ρ-CAR model (see, e.g., Banerjee et al., 2014), where the joint distribution is

$$p(V_1, ...V_n) \propto \exp\left[-\frac{1}{2\tau^2}\mathbf{V}^T(D_w - \rho W)\mathbf{V}\right], \tag{1.6}$$

where $\rho \in (0, 1)$. This ensure that the prior distribution is proper; however, the prior can no longer be interpreted as a smoother. Two other approaches involve adding diagonal matrices to $\mathbf{Q}$. The Besag-York-Mollié (BYM) model parameterizes the model as

$$p(V_1, ...V_n) \propto \exp\left[-\frac{1}{2\tau^2}\mathbf{V}^T\left((1-\phi)\mathbf{I} + \phi\mathbf{Q}^-\right)^{-1}\mathbf{V}\right], \tag{1.7}$$

where $\mathbf{Q}^-$ is a generalized inverse of $\mathbf{Q}$ and $\phi \in [0, 1]$ (Besag et al., 1991). The BYM model is equivalent to having a linear combination of spatial random effects and independent random effects. A similar model, posed by Leroux et al. (2000), is

$$p(V_1, ...V_n) \propto \exp\left[-\frac{1}{2\tau^2}\mathbf{V}^T\left((1-\phi)\mathbf{I} + \phi\mathbf{Q}\right)\mathbf{V}\right], \tag{1.8}$$

where again $\phi \in [0, 1]$.

1.1.2 Point-referenced data and Gaussian Processes

Point-referenced data exist in continuous space, and, in our case, we consider locations $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2$. More generally, point-referenced data can lie in $\mathbb{R}^d$ or on spheres $\mathbb{S}^d := \{\mathbf{x} \in \mathbb{R}^{d+1} : \|\mathbf{x}\| = r\}$. In this setting, observations $Y(\mathbf{s})$ are indexed by their location. We envision hierarchical models for point-referenced data that take the form

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{1.9}$$

where $\mathbf{x}(\mathbf{s})^T$ are covariates with coefficients $\boldsymbol{\beta}$, $w(\mathbf{s})$ are spatial random effects, and $\epsilon(\mathbf{s})$ is noise.

The key to modeling point-referenced data is specifying the spatial random effects $w(\mathbf{s})$. We choose $w(\mathbf{s})$ to be Gaussian processes (GPs). A Gaussian process is a stochastic process where any finite subset of random variables from the process are jointly Gaussian. The GP is fully specified by its mean and covariance function. For hierarchical modeling, it is common to assume that the mean function is zero for all $\mathbf{s}$. Thus, the central component of the point-referenced model is the covariance function. Intuitively, the covariance captures spatial similarity in the data. If covariance is only a function of the separation vector between observations $\mathbf{h} = \mathbf{s} - \mathbf{s}'$, then the covariance is called *stationary*. A covariance function is *isotropic* if it is only a function of the distance between observations. In the problems we address here, we assume that covariance functions are isotropic.

Likelihood computations for Gaussian process models require inverting an $n \times n$ matrix, making Bayesian Gaussian process models intractable with even moderate amounts of data (for example, $n > 10000$). Many have addressed this computational bottleneck using either low-rank or sparse matrix methods (see Heaton et al., 2018, for a review and comparison of these methods). Low-rank methods project the original process onto representative points or knots (see, e.g., Higdon, 2002; Banerjee et al., 2008; Cressie and Johannesson, 2008; Stein, 2008); however, these approaches often perform poorly for prediction as they often over-smooth (see Stein, 2014).

Alternatively, sparse methods either induce zeros in the covariance matrix using compactly supported covariance functions (see, e.g., Furrer et al., 2006; Kaufman et al., 2008; Bevilacqua et al., 2016) or in the precision matrix by assuming conditional independence (Vecchia, 1988; Stein et al., 2004). We ultimately favor conditional independence approaches because predictive performance is generally better (Heaton et al., 2018), and the class of valid covariance models is more expansive. Sparse precision methods date to Vecchia (1988) and were furthered by Stein et al. (2004) and Bevilacqua et al. (2012) to approximate the likelihood of the full GP model

using conditional likelihoods. Gramacy and Apley (2015) and Datta et al. (2016a) extend this work to process modeling. The nearest-neighbor Gaussian process is itself a Gaussian process (Datta et al., 2016a) and has good predictive performance relative to other fast GP methods (See Heaton et al., 2018).

The nearest-neighbor Gaussian process (NNGP) is derived from a *parent* Gaussian process and assumes that observations are conditionally independent given a subset of the other observations as a conditioning or neighborhood set (Datta et al., 2016a). Equivalently, the NNGP can be viewed as a directed acyclic graph (DAG) that induces sparsity in the precision matrix of the parent process by assuming conditional independence given neighborhood sets. To specify the NNGP model, one must select a covariance model (as we have discussed), a reference set $\mathcal{R}$, and conditioning or neighborhood sets $N(\mathbf{s}, t)$ for each observation. In large data settings, we assume that random effects follow an NNGP.

1.1.3 Spatiotemporal Data

In Chapter 2, we work only in the spatial domain; however, Chapters 3 and 4 deal with spatiotemporal data. In particular, we consider modeling in (1) discrete space and discrete time and (2) continuous space and continuous time.

In the first setting (discrete space and discrete time), we consider combinations of auto-regression time-series models with areal unit data models. These models can be formulated using traditional time series methods (Shumway and Stoffer, 2017) or dynamic linear models (West and Harrison, 1997) with time-varying spatial components (Banerjee et al., 2014). In this setting, we also address multivariate data using coregionalization of a CAR model (see Banerjee et al., 2014, for an overview of coregionalization).

In the second setting (continuous space and continuous time), spatiotemporal random effects are specified by a Gaussian process or nearest-neighbor Gaussian process

with a spatiotemporal covariance function. While the same ideas of stationarity and isotropy apply, *nonseparability* is an important concept for covariance functions over more than one domain (e.g. space and time). Separable covariance functions assume that the space-time covariance function can be written as the product of a spatial and a temporal covariance function. Using a separable covariance function implies that there are no interactions between spatial and temporal differences, an assumption that is likely not true for the application discussed. Thus, nonseparable covariance functions are an important consideration.

1.2 Bayesian Inference

For a Bayesian model, the goal of modeling is to update our prior belief about a process using data, which we denote as $\mathbf{Y}$ to align with our previous discussion. Firstly, we select an appropriate model or likelihood $p(\mathbf{Y}|\theta)$ for the data with parameters θ. Then, we select a prior distribution $\pi(\cdot)$ that captures our beliefs before observing data. When a process is well understood or studies, an informative prior distribution may be warranted and can be very useful. In other cases, it is common to use a weakly informative prior that gives very little preference to any region of the parameter space *a priori*. Weaker still, some attempt to eliminate the role of the prior completely by using a Jeffrey's prior or a reference prior. For more discussion on prior selection see Berger (2013); Gelman et al. (2014). The prior distribution is updated by the likelihood using Bayes rule, giving the posterior distribution

$$\pi(\theta|\mathbf{Y}) = \frac{p(\mathbf{Y}|\theta)\pi(\theta)}{\int_\theta p(\mathbf{Y}|\theta)d\pi(\theta)}. \tag{1.10}$$

In most cases, $\int_\theta p(\mathbf{Y}|\theta)d\pi(\theta)$ is intractable, so Markov chain Monte Carlo methods are used to sample from the posterior distribution. Other methods to approximate $p(\mathbf{Y}|\theta)\pi(\theta)$ to give a tractable integral and thus tractable inference include variational

Bayes or integrated nested Laplace approximations (see, e.g., Beal, 2003; Rue et al., 2009). For more details, see Gelman et al. (2014).

The focus here is on prediction using a Bayesian model. Thus, we rely on sampling from the posterior predictive distribution,

$$p\left(\mathbf{Y}_{new}|\mathbf{Y}\right) = \int p\left(\mathbf{Y}_{new}|\theta\right) p\left(\theta|\mathbf{Y}\right) \mathrm{d}\theta. \tag{1.11}$$

As with the posterior distribution, the posterior predictive distribution is rarely analytically available. In this study, we rely on composition sampling to obtain draws from the posterior predictive distribution (1.11).

1.3 Outline of Dissertation and Contributions

In this dissertation, I present three topics in spatial and spatiotemporal modeling with applications in environmental health. Each provides a contribution to the statistics literature motivated by the application presented. In Chapter 2, we consider the case of areal unit spatial data when only a subset of the data is observed and provide examples in variety of examples, including one in disease mapping. In the Chapters 3 and 4, we present two analyses of pollution monitoring data from Mexico City.

In Chapter 2, we consider the situation of areal unit data that are only partially observed, and we seek to infer about the unobserved units. Here, there are two primary contributions. First, we discuss modeling approaches for prediction for partially observed areal unit data. Our second contribution leverages the ideas of predicting unobserved areal units for model comparison. In some cases, minimizing an out-of-sample predictive criterion may be desired, but customarily modeling areal unit data comes with the goal of spatial smoothing (Banerjee et al., 2014), employing a complete dataset over all areal units. In such cases, missingness is not a concern. Moreover, under fitting to the full data, with no hold out data for validation, it will

be impossible to outperform independent local (unit-level) random effects if model performance is assessed by comparison of predicted with observed. For this reason, it is difficult to assess model performance when the primary modeling goal is smoothing (see Stern and Cressie, 1999, for early thoughts in this regard). Since visual assessment is qualitative, how one can quantify one smoothing relative to another? In this chapter, we address these problems with applications in image analysis, disease mapping, and a more challenging problem of assessing the performance of semiconductor chips. This work is published in *Spatial Statistics* (White et al., 2017) and is a collaborative work with Alan E. Gelfand and Theresa Utlaut.

In Chapters 3 and 4, we consider pollution data from Mexico City in 2017. In Chapter 3, we model coarse particulate matter and ozone levels jointly to predict pollution emergencies, as defined by the Mexico City's Atmospheric Environmental Contingency Program (Administración Pública de la Ciudad de México, 2016), and compliance to Mexican ambient air quality standards (Diario Oficial de la Federación, 2014a,b). Pollution emergencies are defined in terms of five regional maxima of ozone and coarse particulate matter levels that are derived from pollutant levels at 24 monitoring stations. Because these regional maxima are come from relatively few stations, we argue against using extreme value models for the maxima. Instead, we use a multivariate discrete-time, discrete-space model to forecast pollutant levels at all 24 sites. These pollution forecasts are then used to project emergency status and to predict compliance to nationally legislated standards.

The contribution here is to understand and predict how often Mexico City was at risk of a pollution emergency in terms of (1) the Atmospheric Environmental Contingency Program in Mexico City and (2) current Mexican ambient air quality standards. For both, we assess how the risk of dangerous pollution varies over city regions and over time. As discussed, emergency phases depend entirely upon pollutant maxima within each region. Furthermore, environmental alerts are often

summarized over coarser temporal scales, like days, rather than the measurement level (hours) or the three hours of evaluation (10 AM, 3 PM, and 8 PM). So, daily emergency phases depend on pollutant maxima over hours of evaluation and stations within each region. For Mexican ambient air quality standards, we can do inference at each of three natural spatial scales: station-level, region-level, or city-level. Again, we may be interested in exceedances occurring on a daily scale rather than hourly. Therefore, we again need maxima over time (and potentially space depending on the spatial scale selected). Here, predictions from our model serve two practical purposes: First, our predictions allow us to carry out probabilistic inference about pollution emergency states or national compliance issues. Second, if implemented in practice, our model could warn of potential pollution emergencies or compliance problems, allowing regional and city-wide adjustments and responses to be made earlier. This work was a collaboration with Alan E. Gelfand, Eliane Rodrigues, and Guadalupe Tzintzun, and a similar pre-print of this work is available at White et al. (2018).

In Chapter 4, we consider ground-level ozone levels during Mexico City's peak ozone season (April and May). This data set consists of hourly ozone levels at 24 stations; thus, we observe more than 35,000 ozone levels. In this chapter, we focus on assessing respiratory health risks attributable to ozone and compliance to nationally legislated ozone standards at locations and times where ozone levels were not monitored. Because spatiotemporal prediction is our primary focus, I work in continuous space-time. These data here exhibit strong daily seasonality, linear drift, and spatial patterns which we address using covariance functions on space $\times$ circles (time-of-day) $\times$ time. For computational tractability, we use a Vecchia approximation (Vecchia, 1988), specifically a nearest-neighbor Gaussian process (Datta et al., 2016a,b).

We provide the following three contributions: First, we account for daily seasonality through covariance modeling rather than through terms in the mean function

of the model and introduce new classes of appropriate covariance functions for our approach. Second, we discuss appropriate Vecchia approximations for covariance functions with seasonality within the nearest-neighbor Gaussian process framework. Our modeling approach allows scalable model fitting, prediction, and inference for the Mexico City ozone data. Third, we use this model to assess respiratory risks and compliance with Mexican ambient air quality standards at locations where ozone is not monitored, a contribution that is more scientific than statistical. This work was a collaboration with Emilio Porcu and has been accepted to *Environmetrics* (White and Porcu, 2019).

Lastly, we discuss and summarize the contributions in this dissertation in Section 5.

2

Prediction and Model Comparison for Areal Unit Data

2.1 Introduction

We consider the situation of areal unit data, so-called discrete multivariate spatial data, where we have areal units over a region which are only partially observed, and we seek to infer about the unobserved units. Applications we envision include disease mapping, small area estimation, image analysis, and our motivating, more challenging application, performance of semiconductor chips. Customarily, with areal unit data, the objective is spatial smoothing (Banerjee et al., 2014), employing a complete dataset over the units; missingness is not a concern. With a goal of smoothing it is difficult to assess model performance (see Stern and Cressie, 1999, for early thoughts in this regard). Since visual assessment is qualitative, we might ask how one can quantify one smoothing relative to another? Moreover, under fitting to the full data, with no hold out data for validation, if model performance is assessed by comparison of predicted with observed, it will be impossible to outperform independent local (unit-level) estimation. Smoothing does not seek to minimize a goodness

of fit criterion.

Here, we are interested in either of the following two scenarios. The first supposes that a substantial portion of our data is missing. For example, in the case of semiconductor chip data, we have the following setting. We have a *run* consisting of *lots* which are portions of a silicon ingot to which impurities are added in order to affect electrical properties. Each lot is sliced into thin *wafers*, and each wafer is partitioned into 195 areal units called *dies*, illustrated in Figure 2.1. After production, a die is tested with respect to meeting measures of performance, e.g., speed, reliability, stress, power usage, in order to determine whether it is acceptable for use as a semiconductor chip. It is infeasible to test all of the dies in all of the wafers within a lot. In practice, performance is measured typically for only a subset of the dies but predictive inference is sought regarding performance for all of the dies on all of the wafers. In our examples, 20% of the dies are observed; however, the sampling rate can vary significantly with the application. See Figure 2.1 for a diagram of the wafer and a set of measurement locations. This setting requires a challenging "nested" spatial model with predictive inference for the unmeasured (missing) dies.

The second scenario focuses on model comparison for areal unit data. For instance, in the disease mapping context, we observe counts of disease cases across areal units (see, e.g., Clayton and Kaldor, 1987; Mollié et al., 1996; Green and Richardson, 2002; Lawson, 2013). Typically, spatial random effects are introduced using Markov random field (MRF) models in the form of a conditionally autoregressive (CAR) specification (see below) in the log mean for the counts. Model comparison would seek to compare the various CAR models that have been proposed in the literature (e.g. Besag, 1974; Besag et al., 1991; Leroux et al., 2000; Dean et al., 2001). As above, these models provide smoothing, here, of relative risks. With models fitted to the full set of counts, how can we decide which smoothing of the relative risks is preferred? For example, minimizing a predictive mean square error criterion will

not be appropriate since we are not trying to *fit* the observed counts. Instead, using metrics such as predictive mean square error or rank probability scores (Gneiting and Raftery, 2007) with hold out data provides a potentially useful quantitative assessment. These metrics can be interpreted as out-of-sample *measures of smoothness*; their use in this context does not seem to be suggested in the literature. Illustratively, we might fit a given model to a portion of the units, selected at random, and predict for the remainder. In fact, we might do this several times to *average* over the randomness in the selection of fitting and validation units.

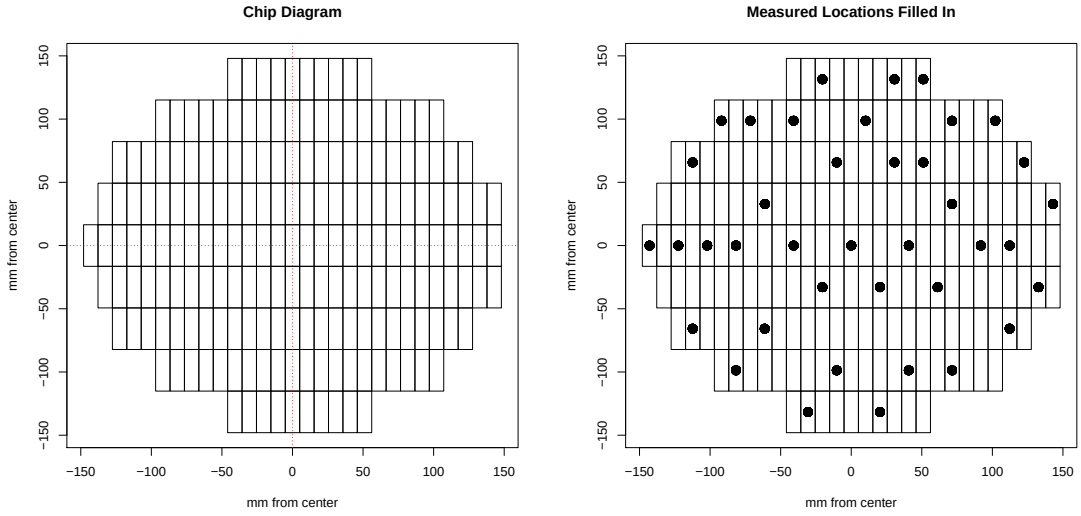


FIGURE 2.1: (Left) Diagram of dies on a silicon wafer, (Right) Locations of measured dies

Modeling for areal unit data arises according to the nature of the data. For example, with count data for the units, as in disease mapping, we think only in terms of measurements at areal scales. There exists a conceptual count for any subregion/areal unit of the study area, but we do not imagine a count at a point, i.e., there is no point-referenced surface of counts. Similarly, we might collect proportions as the data for the units. That is, for each unit, we can imagine a proportion of

a particular racial or ethnic group or a proportion of individuals older than 65. However, in either case, there is no proportion at a point. Such settings result in finite dimensional model specifications with MRFs providing the customary modeling (e.g. Rue and Held, 2005; Banerjee et al., 2014).

Alternatively, we can imagine areal unit measurements arising as averages of a surface over a region. That is, we now envision an entire surface, an uncountable number of random variables, over a region of interest. Such a surface is customarily viewed as a realization of a stochastic process, typically a Gaussian process (Cressie, 2015; Banerjee et al., 2014). The resulting areal unit observations then arise as *block* averages of the surface over the areal units of the study region (Banerjee et al., 2014). Familiar examples include averaged temperatures or averaged environmental exposures for grid cells. For the semiconductor chip data, with certain performance measurements, it would be appropriate to conceptualize a measurement surface over the wafer and what is observed for a die is an average value of the surface for that die.

We consider both cases here. In either one, we assume (i) the observations are available only at areal scales, (ii) that we have observations for only a subset of the areal units in the region of interest, and (iii) interest is in prediction for the remaining areal units whether they be missing or held out of the fitting. Again, this finite dimensional prediction setting does not seem to have received attention in the literature. Moreover, *kriging* with areal units in the case of a Markov random field (MRF) violates the assumptions associated with the MRF specification (see, e.g., Banerjee et al., 2014, Section 4.3). The joint distribution arising through multiplication of the conditional distribution for the unit for which kriging is sought given the observed units by the joint CAR distribution for the observed units is not the CAR model that would arise if the new unit and the observed units were jointly modeled as a CAR. The case of an underlying spatial surface leads to familiar block kriging; in

fact, here it is block-to-block kriging rather than point-to-block kriging (see Banerjee et al., 2014, in this regard). With application to the nested semiconductor chip scenario, we add substantial complexity to the modeling in order to obtain the required kriging.

A further challenge arises when we have multivariate measurements associated with each areal unit. In the disease mapping setting we can imagine counts of different types of cancers for each observed unit. In the environmental exposure context, we can imagine an average $PM_{2.5}$ level, an average ozone level, and an average carbon monoxide level over a grid cell. In the semiconductor chip setting, we may have several test measurements for each observed die and the measurements may be of different data types. Modeling for the count data case introduces multivariate CAR models (Carlin et al., 2003; Gelfand et al., 2003). Modeling for the exposure data employs (perhaps after transformation) multivariate Gaussian processes.

With lattice data and, more generally, irregularly shaped areal unit data, neighbor based modeling is usually employed. Such models provide the joint distribution for the areal units through neighbors, specifying a conditional distribution such that the expected value of the spatial variable at a given unit is an average of its neighboring units (using a suitable definition of neighbors). Such specifications are MRFs and date at least to Besag (1974). They can model continuous variables leading to so-called CAR models as well as discrete variables leading to, e.g., autologistic or Potts models (Besag, 1974; Hogmander and Møller, 1995; Hoeting et al., 2000). They provide joint distributions through local specifications. In fact, this is essentially the definition of a Markov random field, i.e., a local specification for each unit through conditional distributions such that, altogether, they produce a unique joint distribution. In practice, these distributions need not be proper. In fact, there is controversy over whether or not they should be improper (Banerjee et al., 2014). When improper they can not be used as models for data but can be moved to the

second stage of a hierarchical model and used as a prior for spatial random effects. In particular, with disease mapping data, the spatial structure is typically brought in through random effects at the second stage in the form of an improper CAR prior. By contrast, the foregoing autologistic and Potts models, having finite support for each areal unit, are always proper.

To accommodate the case of areal observations arising as averages over spatial surfaces, we employ block averages. Let $Z(\mathbf{s})$ denote the realization of the spatial process at location $\mathbf{s}$, for $\mathbf{s}$ in a region of interest D, e.g., a geographic region or a wafer. A realization of the process is in fact a random surface over D. For block $B \subset D$, the block average is defined as $Z(B) = \frac{1}{|B|}\int_B Z(\mathbf{s})d\mathbf{s}$ where $|B|$ denotes the area of B. The integration is not the usual integration of a function. Rather, it is an average of an uncountable number of random variables from a realization of a stochastic process, hence a random or stochastic integral. Such integrals are theoretically demanding (Kuo, 2006) but, practically, can be modeled. With a Gaussian process their distribution theory is complete; means, variances, and covariances for say $Z(B_1)$ and $Z(B_2)$ can be written down explicitly and multivariate normals emerge. However, they can only be handled approximately with regard to simulation and model fitting (see, e.g., Banerjee et al., 2014).

Simultaneous prediction over many areal units raises the issue of simultaneous inference. In particular, for the semiconductor chip setting with 20% sampling and say 5 lots, each lot with say 7 wafers and $0.8 \times 195 = 156$ unobserved dies per wafer we need to predict for 5460 dies, evidently a multiplicity problem[1]. Accordingly, we could utilize interval length corrections (e.g., Bonferroni or Tukey) or methods to bound false discovery rates to account for this. In practice, an expected error rate is selected and adopted. We do not address this issue further here.

[1] The National Institute of Standards and Technology (NIST) suggests seven wafers per lot, but notes that each lot could contain up to 25 wafers (Sematech, 2006).

We employ several examples here, both real and simulated, to illustrate the utility of the methods discussed for prediction and model validation. First, we simulate an image and impose varying levels of damage or missingness. Then, we reconstruct that image using a CAR model. This simulation serves as a proof of concept for the predictive CAR model. Then, we take a well-studied dataset of lung cancer deaths for all 88 Ohio counties from 1976-1996 (abstracted from a Centers for Disease Control database). With many randomly selected hold-out data sets of 22 counties, we investigate the predictive performance of various CAR models to compare smoothing. Lastly, we consider two semiconductor chip examples: one with a binary outcome at each die, the other with a continuous response surface that is observed at die-level. Because real data are proprietary, these data are simulated; however, the simulated data are representative of actual semiconductor chip data. The binary semiconductor chip outcomes are modeled using nested CAR specifications, while the continuous responses are modeled using nested Gaussian processes (Kaufman and Sain, 2010), and we investigate model comparison for both examples.

The format of the paper is as follows. We begin by discussing areal unit data in the context of MRFs with associated models and examples in Section 2.2. In Section 2.3, we present two MRF modeling examples, one with simulated data and one with real data, to illustrate the utility of the predictive CAR modeling strategy. Section 2.4 presents a semiconductor chip example where the outcome is binary, gives the predictive MRF models utilized to analyze this data, and discusses associated results. In Section 2.5, we offer a Gaussian process model for semiconductor chip data where the areal unit data arises from a continuous process. Lastly, in Section 2.6, we summarize our contributions and discuss future research problems they motivate.

2.2 The Markov random field setting

Markov random field models specify the joint distribution of a set of say n spatial random variables, $\mathbf{V} = (V_1, V_2, ..., V_n)$ through the set of conditional distributions denoted by $[V_i|\{V_j, j \sim i\}]$ where $j \sim i$ indicates that unit j is a neighbor of unit i with some definition of neighbors. When the conditional distributions are normal, this conditional mean is a weighted average of the V_j and these models are referred to as CAR models. In fact, a weight or proximity matrix, W is supplied such that entry w_{ij} provides the weight associated with V_j in the mean for V_i (evidently $w_{ii} = 0$). Typically, the weights are (i) normalized to sum to 1, (ii) positive but, for example, a random walk in two dimensions produces both positive and negative weights (Lindgren and Rue, 2008), and (iii) fixed but cases where they are random have been proposed (White and Ghosh, 2009; Ma et al., 2010; Berrocal et al., 2012; Lee and Mitchell, 2013). When, for each i, the w_{ij} sum to 1, the joint distribution is improper (Banerjee et al., 2014). More precisely, if we let D_W be diagonal with $D_{W\,ii} = w_{i+}$, where $w_{i+} = \sum_i w_{ij}$, then, the precision matrix of $\mathbf{V}$ is $Q = D_W - W$. This is improper since $\mathbf{1}Q = 0$. On irregular lattices, the weights can be modified so that $W_{ij} = 1/d_{ij}$, where d_{ij} is some distance metric separating unit i and j (Rue and Held, 2005). Neighbors can be first-order, i.e., sharing a border, or higher-order. Improper CAR models are often called intrinsic autoregressive (IAR) models and are smoothers, as the mean at any location is a weighted average of its neighbors. An important higher-order CAR model is defined by the two-dimensional second-order random walk (2D-RW2), which is used in MRF approximations to thin-plate splines and Gaussian processes (Rue and Held, 2005; Lindgren and Rue, 2008; Banerjee et al., 2014).

To remedy the impropriety, it has been proposed to scale the weights by ρ, an autoregression parameter in $(0, 1)$ (ρ-CAR model). This provides a proper CAR

distribution but leads to the expected value of V_i being less than the average of its neighbors which contradicts the local smoothing notion for the model. Other remedies have been proposed. Besag et al. (1991) (BYM) added a pure error term, ϵ_i to V_i, i.e., a vector $\boldsymbol{\epsilon}$, to the CAR model so that $(\mathbf{V}+\boldsymbol{\epsilon})|\mathbf{V}$ is a proper distribution. A different version, following from Leroux et al. (2000) and MacNab and Dean (2000), specifies $\Sigma_{\mathbf{V}} = \tau^2((1-\psi)I + \psi(D_W - W))^{-1}, \psi \in [0,1)$. Here, as $\psi \to 1$, we tend to the usual intrinsic CAR, $\psi = 0$ provides an independence model, and $\psi < 1$ provides diagonal dominance, hence a nonsingular matrix. We denote this model as CAR(τ^2, ψ) The parameter ψ enriches the CAR model but, as with the ρ-CAR models, $E(V_i)$ is less than the average of its neighbors.

Suppose we start with the model

$$Z_i = \mathbf{X}_i^T\boldsymbol{\beta} + V_i + \epsilon_i, \tag{2.1}$$

where V_i comes from one of the improper CAR (IAR) models described above, and the ϵ_i are pure Gaussian error (white noise) with variance σ^2. This model is valid for data Z_i since we have a conditionally independent first stage given $\{V_i\}$ and $\boldsymbol{\beta}$. So, we imagine the improper CAR prior at the second stage. This is a model for continuous real-valued observations (i.e. $Z_i \in \mathbb{R}$). Similarly, we can consider the case where the $\mathbf{Z}_i$ are multivariate, i.e., $\mathbf{Z}_i$ is a $q \times 1$, $\mathbf{X}_i$ is a $p \times q$ matrix, and the elements of $\mathbf{Z}_i$ are dependent. Then, we utilize a multivariate CAR (MCAR) (Carlin et al., 2003; Gelfand et al., 2003), specifying $\mathbf{Z}_i = \mathbf{X}_i^T\boldsymbol{\beta} + \boldsymbol{\phi}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\beta}$ is $p \times 1$, $\boldsymbol{\phi}_i$ and $\boldsymbol{\epsilon}_i$ are $q \times 1$, and $\boldsymbol{\phi}|\boldsymbol{\Lambda} \sim \text{MCAR}(0, \boldsymbol{\Lambda})$. This sort of model might be appropriate for data arising at grid cell level for multiple pollutants.

A model for binary spatial data, for example, a binary or two-color image or map, can be induced from (2.1). With a binary Y_i at each areal unit, we model Y_i through a probit specification ($P(Y_i = 1|\boldsymbol{\beta}, V_i) = \Phi(X_i^T\boldsymbol{\beta} + V_i)$). This model moves all spatial

modeling to the second level. Equivalently, we can introduce auxiliary variables

$$Y_i^* \sim N(X_i^T \boldsymbol{\beta} + V_i, 1), \tag{2.2}$$

where $Y_i = 1\,(0)$ if $Y_i^* \geq 0\,(< 0)$, and the Y_i are conditionally independent. This formulation is amenable to the Bayesian probit model offered by Albert and Chib (1993).

For the disease mapping setting, we assume the counts $Y_i \sim \text{Poisson}(\lambda_i)$ and are conditionally independent given λ_i. Here, $\lambda_i = E_i r_i$, where E_i is the expected count in unit i and r_i is the standardized rate for unit i. Then, we set

$$\log r_i = \mathbf{X}_i^T \boldsymbol{\beta} + V_i, \tag{2.3}$$

where, again, V_i is a CAR model. Again, we have moved the spatial dependence to the mean (on the link-transformed scale) as a hierarchical model.

Similar modeling has been proposed for other first stage areal unit data models with introduction of a second stage (hierarchical) CAR. One example is an ordered categorical first stage where we introduce latent cut points to define the ordinal categories (Albert and Chib, 1993). A second illustration considers extreme value data, for example, maximum annual temperature or other environmental exposure extremes, with a first stage general extreme value distribution (Sang and Gelfand, 2009). A third example could apply to census units providing rates or proportions. Such modeling could also be used for prediction in small area estimation settings where spatial smoothing is desired (Ghosh and Rao, 1994). We can also imagine higher-dimensional versions (e.g., neuroimaging), where we use voxels as the areal units but now require a CAR over 3-dimensional space. Some modeling details for this context can be found in, for example, Penny et al. (2005); Bowman et al. (2008); Derado et al. (2013). For any of these settings, prediction for missing areal units could become the objective apart from a strategy for model comparison.

The prediction that we propose for the CAR setting is done hierarchically with spatial dependence at the second stage and conditional independence at the first stage. This allows specifying areal unit dependence at the second stage over all units while introducing into the first stage only the units actually observed. More precisely, suppose the data vector $\mathbf{Y}$ is split into $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{miss}$. Then, within a Bayesian framework, the model fitting is

$$\Pi[Y_{obs,i}|\boldsymbol{\beta}, \boldsymbol{\theta}, V_{obs,i}][\mathbf{V}_{obs}, \mathbf{V}_{miss}|\tau^2, \psi][\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2, \psi]. \tag{2.4}$$

Here, $\boldsymbol{\theta}$ denotes any first stage model parameters and $[\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2, \psi]$ denotes the prior on all of the model parameters. This yields the posterior, $[\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{V}_{obs}, \mathbf{V}_{miss}, \tau^2, \psi|\mathbf{Y}_{obs}]$. Draws from this posterior distribution enable draws from the posterior predictive distribution for any $Y_i \in \mathbf{Y}_{miss}$ using composition sampling (Gelman et al., 2014). Therefore, the predictive distribution can be compared with the held out Y_i value using the criteria presented in Section 2.2.2 below.

An elementary question is whether, this posterior distribution is proper under an improper CAR prior where the data is only a partially observed set of the Z_i's (or the Y_i's). The answer is that, if all of the areal units are connected (no islands), we need only one observation to provide a proper posterior. Perhaps this result is not surprising since one Z_i is enough to "center" the V_i's. In any event, the result is demonstrated for the case of normal data below.

Proof. Under the normal formulation, where Y_i are conditionally normal and independent given ϕ_i, the posterior distribution $[\boldsymbol{\beta}, \tau^2, \sigma^2, \boldsymbol{\phi}_{\text{miss}}, \boldsymbol{\phi}_{\text{obs}}|\mathbf{Y}_{\text{obs}}]$ is proper when the observation vector $\mathbf{Y}$ is not fully observed. For simplicity, assume that $\mathbf{X} = \mathbf{0}$ and that τ^2 and σ^2 are constant. Assume that there are no missing values at islands – locations with no neighbors.

Proof: We first partition the proximity matrix $W = \begin{pmatrix} w_{oo} & w_{om} \\ w_{mo} & w_{mm} \end{pmatrix}$ and $D_w =$

$\begin{pmatrix} D_o & 0 \\ 0 & D_m \end{pmatrix}$. Note

$$
\begin{aligned}
[\boldsymbol{\phi}|\mathbf{Y}_{\text{obs}}] &\propto [\mathbf{Y}_{\text{obs}}|\boldsymbol{\phi}][\boldsymbol{\phi}] \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}_{\text{obs}}-\boldsymbol{\phi}_{\text{obs}})^T(\mathbf{Y}_{\text{obs}}-\boldsymbol{\phi}_{\text{obs}})\right\}\times \\
&\exp\left\{-\frac{1}{2\tau^2}\begin{pmatrix}\boldsymbol{\phi}_{\text{obs}}\\ \boldsymbol{\phi}_{\text{miss}}\end{pmatrix}^T(D_w-W)\begin{pmatrix}\boldsymbol{\phi}_{\text{obs}}\\ \boldsymbol{\phi}_{\text{miss}}\end{pmatrix}\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\phi}_{\text{obs}}^T\boldsymbol{\phi}_{\text{obs}}-2\boldsymbol{\phi}_{\text{obs}}^T\mathbf{Y}_{\text{obs}})\right\}\times \\
&\exp\left\{-\frac{1}{2\tau^2}\left(\boldsymbol{\phi}_{\text{obs}}^TD_o\boldsymbol{\phi}_{\text{obs}}+\boldsymbol{\phi}_{\text{miss}}^TD_m\boldsymbol{\phi}_{\text{miss}}+\begin{pmatrix}\boldsymbol{\phi}_{\text{obs}}\\ \boldsymbol{\phi}_{\text{miss}}\end{pmatrix}^TW\begin{pmatrix}\boldsymbol{\phi}_{\text{obs}}\\ \boldsymbol{\phi}_{\text{miss}}\end{pmatrix}\right)\right\} \\
&\propto \exp\Bigg\{-\frac{1}{2\tau^2}\Bigg(\begin{pmatrix}\boldsymbol{\phi}_{\text{obs}}\\ \boldsymbol{\phi}_{\text{miss}}\end{pmatrix}^T\left[\begin{pmatrix}D_o+\frac{\tau^2}{\sigma^2}I_{n_o} & 0\\ 0 & D_m\end{pmatrix}-W\right]\begin{pmatrix}\boldsymbol{\phi}_{\text{obs}}\\ \boldsymbol{\phi}_{\text{miss}}\end{pmatrix}- \\
&2\begin{pmatrix}\boldsymbol{\phi}_{\text{obs}}\\ \boldsymbol{\phi}_{\text{miss}}\end{pmatrix}^T\begin{pmatrix}\frac{\tau^2}{\sigma^2}\mathbf{Y}_{\text{obs}}\\ 0\end{pmatrix}\Bigg)\Bigg\} \\
&\propto \exp\left\{-\frac{1}{2\tau^2}\left[\boldsymbol{\phi}^T(D_w^*-W)\boldsymbol{\phi}-2\boldsymbol{\phi}^T\begin{pmatrix}\frac{\tau^2}{\sigma^2}\mathbf{Y}_{\text{obs}}\\ 0\end{pmatrix}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau^2}\left(\boldsymbol{\phi}-\begin{pmatrix}\frac{\tau^2}{\sigma^2}\mathbf{Y}_{\text{obs}}\\ 0\end{pmatrix}\right)^T(D_w^*-W)\left(\boldsymbol{\phi}-\begin{pmatrix}\frac{\tau^2}{\sigma^2}\mathbf{Y}_{\text{obs}}\\ 0\end{pmatrix}\right)\right\},
\end{aligned}
$$

which is normal if D_w^*-W is positive definite, which it is. Let ϕ_{ij} be the entries of D_w^*-W. Note $|\phi_{ii}|\geq\sum_{j\neq i}|\phi_{ij}|$ for all i. Specifically,

- For all i such that Y_i is observed $|\phi_{ii}|-\sum_{j\neq i}|\phi_{ij}|=\tau^2/\sigma^2$.
- For all i such that Y_i is not observed $|\phi_{ii}|-\sum_{j\neq i}|\phi_{ij}|=0$

Thus, if we observe at least one data, D_w^*-W is an irreducible, diagonal dominant matrix. Every irreducible, diagonal dominant matrix is positive definite (Feingold et al., 1962). Thus, the posterior is proper. □

Again, while the improper CAR prior distribution has no center (i.e. one could any constant to any Y_i without affecting the joint distribution), a single observation

is sufficient to center the CAR posterior, i.e., to ensure that the posterior distribution is proper.

2.2.1 Priors, fitting, prediction

Model fitting with the full data under a specification as in (2.4) requires priors on $\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2, \psi$. Usual choices for means and variances are vague normals and inverse gammas, respectively. Model fitting for the various complete data cases is well discussed in the literature and available through now well established software such as WinBUGS (Lunn et al., 2000), JAGS (Plummer et al., 2003), or STAN (Carpenter et al., 2016). The model fitting with missing Y's proceeds similarly. The only difference is that we only include the $Y_{obs,i}$ in the likelihood. For the normal model, a Gibbs sampler is readily available, and full conditional distributions are given in the supplementary material. When the likelihood is non-Gaussian, a Gibbs sampler is not generally available. For binary or categorical variables, the Bayesian probit model (Albert and Chib (1993)) allows us to use the full conditional distributions for the normal model. For multivariate data, model choices for $\boldsymbol{\Lambda}$ may be coregionalization (Gelfand et al., 2004) or simple Wishart priors. Given a Normal likelihood assumption and Wishart prior distributions for $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Lambda}^{-1}$, the full conditional distributions are given in the supplementary material.

Prediction proceeds through composition sampling (Gelman et al. (2014)),

$$[\mathbf{Y}_{miss}|\mathbf{Y}_{obs}] = \int [\mathbf{Y}_{miss}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{V}_{miss}][\mathbf{V}_{miss}|\mathbf{V}_{obs}, \tau^2, \psi][\mathbf{V}_{obs}, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2, \psi|\mathbf{Y}_{obs}]. \quad (2.5)$$

So, posterior draws of $(\mathbf{V}_{obs}, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2, \psi)$ enable a posterior draw for $\mathbf{V}_{miss}$ which, in turn, enables a posterior draw for $\mathbf{Y}_{miss}$.

2.2.2 Model comparison criteria

We propose comparing different CAR models using cross-validation. As above, we partition the data vector $\mathbf{Y}$ into $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{miss}$. We then fit the model and use draws from the posterior distribution to sample from the posterior predictive distribution for all $Y_i \in \mathbf{Y}_{miss}$. Therefore, the predictive distribution $[Y_i|\mathbf{Y}_{obs}]$ can be compared with the held out $Y_i = y_i \in \mathbf{Y}_{miss}$ value.

For continuous variables (e.g., Normal, t, or Gamma random variables), we propose the following criteria: $100 \times \alpha$ % predictive interval coverage, predictive mean square (PMSE) or absolute error (PMAE) ($(E(Y_i|\mathbf{Y}_{obs})-y_i)^2$ or $|E(Y_i|\mathbf{Y}_{obs})-y_i|$), and the continuous rank probability score (CRPS) (Gneiting and Raftery, 2007), where, with F_i denoting the predictive distribution for Y_i, $\text{CRPS}(F_i, y_i) = \int_{-\infty}^{\infty}(F_i(x)-\mathbf{1}(x \geq y_i))^2 dx = \mathbf{E}|Y_i - y_i| - \frac{1}{2}\mathbf{E}|Y_i - Y_i'|$. The last expression gives expectations that are immediately amenable to Monte Carlo integration using the posterior predictive samples of $Y_i|\mathbf{Y}_{obs}$. These criteria are summed over the hold out i's.

The criteria for count variables are similar: $100 \times \alpha\%$ predictive interval coverage, predictive mean square or absolute error, scaled MSE $\left[\frac{(y_i-E(Y_i|\mathbf{Y}_{obs}))^2}{E(Y_i|\mathbf{Y}_{obs})}\right]$, scaled MAE $\left[\frac{|y_i-E(Y_i|\mathbf{Y}_{obs})|}{E(Y_i|\mathbf{Y}_{obs})}\right]$, and rank probability score (RPS) (defined similarly to CRPS) (Gneiting and Raftery, 2007). Again, these are summed over the hold out i's. For binary or categorical variables, let FN denote the number of false negatives, TN the true negatives, FP the false positives, and TP the true positives. Then, the criteria we use for model comparison for binary random variables are predictive accuracy ($\frac{TP+TN}{TP+FP+TN+FN}$), sensitivity ($\frac{TP}{TP+FN}$), specificity ($\frac{TN}{TN+FP}$), and the Brier Score (BS $= \frac{1}{n}\sum_{i=1}^{n}(y_i - E(Y_i|Y_{\text{obs}}))^2$) (Brier, 1950; Gneiting and Raftery, 2007).

In summary, predictive interval coverage is useful with regard to assessing model adequacy. The other criteria are suitable for model comparison. However, we remind

the reader that here, with the modeling objective being smoothing, minimization need not provide model preference. As we noted in Section 2.1, comparison across models can be more naturally interpreted as comparison of the extent of smoothing. For example, higher MSE would suggest more smoothing. They can supplement visualization through maps in assessing model performance.

2.3 Examples

2.3.1 Image Reconstruction

As a simple first example, we simulate a single $64 \times 64 = 4096$ point image with a univariate continuous response z on a unit square using

$$z(x, y) = 20 + 4\sin(20x - .02) + 4\sin(10y) + 3\sin(30x + .2)\sin(15y) + \epsilon(x, y) \quad (2.6)$$

as the data generating mechanism, where $\epsilon(x, y) \overset{iid}{\sim} N(0, 1)$. However, we assume the model mechanism to be unknown. The realized image is plotted in Figure 2.2.

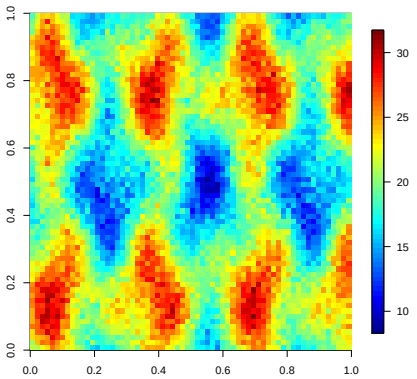


FIGURE 2.2: Simulated image from generative model.

Specifically, the model considered is:

$$Y_i = \beta_0 + \mathbf{V}_i + \epsilon_i, \quad (2.7)$$

where $\mathbf{V}_i$ is the improper CAR(τ^2) model with weights selected according to the 2D-RW2 (Rue and Held, 2005), and the ϵ_i are i.i.d. $N(0, \sigma^2)$. For this model, we take $\beta_0 \sim N(m_\beta, V_\beta)$, $\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma)$, and $\tau^2 \sim \text{IG}(a_\tau, b_\tau)$ where $a_\sigma = a_\tau = b_\sigma = b_\tau = 1$, $m_\beta = \mathbf{0}$, and $V_\beta = 10^{12}\mathbf{I}$. We give the reconstructed images and residual images for the 2D-RW2 CAR model in Figure 2.3 when 5%, 20%, and 50% of the image is observed. Note that the images reconstructed by the 2D-RW2 CAR model are very similar to the simulated image (See Figures 2.2 and 2.3), demonstrating that the CAR model can be used effectively for prediction. When the the level of missingness is higher, the reconstructed image is smoother relative to the images with less missingness. Additionally, note that the residual terms diminish as the levels of missingness decrease, as expected (See Figure 2.4).

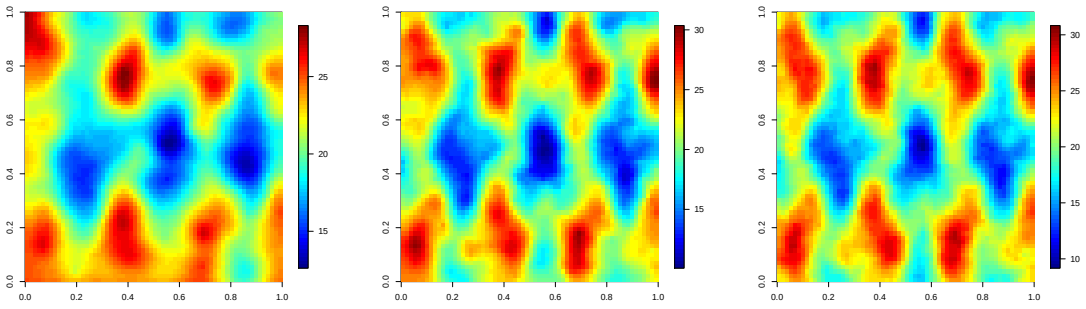


FIGURE 2.3: Mean reconstructed images under 2D-RW2 CAR model with (Left) 5% of data observed (Center) 20% of data observed (Right) 50% of data observed .

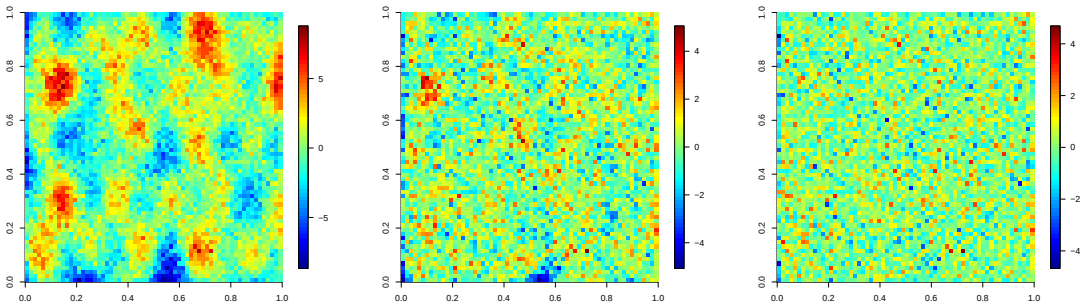


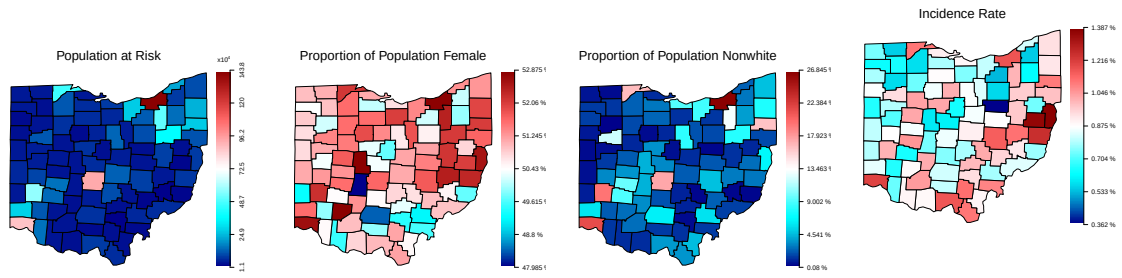
FIGURE 2.4: Residual images under 2D-RW2 CAR model with (Left) 5% of data observed (Center) 20% of data observed (Right) 50% of data observed .

2.3.2 Disease Mapping

For this example, we utilize Ohio county lung cancer deaths from 1976-1996 as an illustrative example. For each county, we have the number of lung cancer deaths, population, proportion of population that is non-white, and the proportion of population that is female for every year. In this analysis, aggregated lung cancer deaths from 1976-1996 for each county are the outcomes we model. The time-averaged population characteristics (population at risk) n_i, proportion of population that is non-white $p_{nw,i}$, and the proportion of population that is female $p_{f,i}$) are used as covariates. Population characteristics and lung cancer incidence rates are plotted in Figure 2.5.

The primary goal of the modeling in this example is to smooth the disease rate map in Figure 2.5d. We compare competing models for this example by holding out 22 randomly selected counties (25% of the data), fitting each model to the remaining data, and predicting the outcomes. Then, because we are smoothing the rates, we carry out model comparison on the incidence rates. Specifically, we compute PMSE and CRPS on these rates for the hold-out counties. Both of these quantities can be interpreted as measures of smoothing (i.e., the higher the PMSE and CRPS, the

more smoothing). We repeat this process 1000 times, each iteration having a unique random holdout set and associated measures of smoothness (PMSE and CRPS). Then, we take averages of these statistics in order to average over the randomness in the subset selection.



(a) Number at risk (b) % female (c) % nonwhite (d) Incidence rate

FIGURE 2.5: Summaries of Ohio population data.

All models have a similar structure:

$$y_i \sim \text{Pois}(n_i p_i) \tag{2.8}$$
$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{logit}(p_{nw,i}) + \beta_2 \text{logit}(p_{f,i}) + V_i.$$

The specification for the CAR random effect V_i differs with each model. Specifically, we consider the Besag (1974) CAR, a second-order CAR, the proper ρ-CAR model, and the Leroux et al. (2000) CAR(τ^2, ψ) model. For all models, we take

$$\boldsymbol{\beta} \sim N(m_\beta, V_\beta)$$
$$\tau^2 \sim \text{IG}(a_\tau, b_\tau), \tag{2.9}$$

where $a_\tau = b_\tau = 1$, $m_\beta = \mathbf{0}$, and $V_\beta = 10^{12}\mathbf{I}$. For the ρ-CAR model, we take $\rho \sim \text{Uniform}(0, 1)$; however, the posterior for this parameter was concentrated near 1 in all cases, as has been noted in the literature (see, e.g., Banerjee et al., 2014, Section 4.3). For the Leroux et al. (2000) CAR model, we take $\psi \sim \text{Uniform}(0, 1)$.

Using these prior distributions, we fit the specified models. Model comparison criteria results are given in Table 2.1.

Table 2.1: Model comparison statistics for the 1996 ohio lung cancer data. The Rel-PMSE and Rel-CRPS give the model smoothing relative to the nonspatial model.

	Non-spatial	Improper Besag (1974)	Proper ρ	Second-order	Leroux et al. (2000)
PMSE	3.27e-06	8.19e-05	3.75e-05	2.23e-05	2.78e-06
CRPS	1.37e-03	1.25e-03	1.23e-03	1.10e-03	9.57e-04
Rel-PMSE	1.00	25.04	11.47	6.82	0.85
Rel-CRPS	1.00	0.91	0.90	0.80	0.70

Table 2.1 shows that the Leroux et al. (2000) CAR model, which is proper, actually has lower PMSE than the non-spatial model, suggesting that it gives very little smoothing, which is reflected in the values of ψ observed (the 95% credible interval for ψ is $(0.0002, 0.0391)$). Every CAR model has lower (better) CRPS than the nonspatial model because they account for the overdispersion in the data relative to the Poisson model. The proper ρ-CAR model behaves more like the improper CAR models because the parameter ρ tends to 1 (i.e. it approaches the improper CAR model). Interestingly, the second-order CAR models has much lower MSE (less smoothing) than the Besag (1974) model but achieves very similar CRPS. This means that these models tend to predict better than the Besag (1974) CAR model, but the overall predictive performance (given by CRPS) of these models is comparable to the Besag (1974) model.

2.4 The Semiconductor chip setting

Here, we turn to the motivating example of semiconductor chip data requiring a nested spatial model. Because of the small physical size of semiconductor chips at die-level, exact and consistent control over the fabrication process is increasingly challenging (Mittal, 2016). As a result, the properties of the resulting materials can differ substantially. Electrical examinations (e.g., capacitance, voltage, resistance, and leakage) are carried out on each semiconducting device before and after slicing to

determine if they function properly (Diebold, 2001). A unit may be called a failure due to a single (univariate) measurement or some composite score derived from a multivariate measurement (this process is called product binning). The proportion of devices deemed successes is called the yield, but information about production yields is proprietary.

Again recall the structure. We have a run consisting of five lots, where each lot has many wafers and each wafer is partitioned into rectangular dies. Let i index lots, j index wafers, and k index dies on wafers. In the dataset, we observe $i = 1, 2, ..., 5$, $j = 1, 2, ..., 7$, and $k' = 1, 2, ..., 39$ from $k = 1, 2, ..., 195$. That is, at 20% sampling, we observe $5 \times 7 \times 39 = 1365$ dies and seek to predict to $5 \times 7 \times 156 = 5460$ dies. The same die locations are tested on every wafer. For this reason, we simulate, model, and analyze data according to this framework. In particular, here we simulate dies that pass or fail according to a probit model, where 50% of dies fail, on average. The full generative hierarchical model specification is given in A.1.

The model resembles an ANOVA structure. In particular, suppose Y_{ijk} denotes a binary response at die k within wafer j within lot i. We model Y_{ijk} through a probit specification, i.e., $P(Y_{ijk} = 1|\boldsymbol{\beta}, U_k, V_{ik}, W_{ijk}) = \Phi(\mathbf{X}_k^T\boldsymbol{\beta} + U_k + V_{ik} + W_{ijk})$. We imagine a quadratic trend surface, as in

$$\mathbf{x}_k^T = \left(1, x_{1k}, x_{1k}^2, x_{2k}, x_{2k}^2, x_{1k}x_{2k}, \sqrt{x_{1k}^2 + x_{2k}^2}\right), \tag{2.10}$$

common for all wafers, so that, at die k, the model specifies $\mu_k = \mathbf{x}_k^T\boldsymbol{\beta}$ where the entries in $\mathbf{X}_k$ arise at the centroid of die k, and $\boldsymbol{\beta}$ are the associated coefficients. U_k is a global standard CAR model over the dies with *variance* τ_U^2, intended to capture dependence across all wafers within all lots. The V_{ik} are lot-level standard (first-order) CARs over dies. These CARs are i.i.d. across lots with common *variance* τ_V^2. The intent is to introduce dependence between dies within a lot. The W_{ijk} are wafer within plot standard CARs over dies. Here, the W_{ijk} are independent across lots

(i) but are dependent between wafers (j) within a lot; the rationale is that wafers within a core are sliced consecutively and so are expected to be dependent. Below, we propose an *equicorrelated* version of dependence, i.e., within a lot the wafers are dependent but *exchangeable*; the rationale is that we do not know the order of cutting of the wafers. We consider comparison of four models here: (i) a trend surface only model, $\Phi(\mathbf{X}_k^T\boldsymbol{\beta})$, (ii) a trend plus global CAR model, $\Phi(\mathbf{X}_k^T\boldsymbol{\beta}+U_k)$, (iii) a trend plus global and lot CARs model, $\Phi(\mathbf{x}_k^T\boldsymbol{\beta}+U_k+V_{ik})$, and (iv) a trend plus global, lot, and wafer CARs model, as above, $\Phi(\mathbf{x}_k^T\boldsymbol{\beta}+U_k+V_{ik}+W_{ijk})$. This comparison allows us to examine whether CAR models improve prediction and how this changes with nesting structure. If this nesting is effective, the CAR models should outperform the trend-only model and the CAR models with additional nesting should outperform simpler models.

2.4.1 Priors, model fitting, prediction

Considering the full model, the improper prior distributions for the global and lot-level CAR effects are

$$[U|\tau_U^2] \propto \exp\left\{-\frac{1}{2\tau_U^2}U^TQU\right\} \tag{2.11}$$

$$[V_i|\tau_V^2] \propto \exp\left\{\frac{1}{2\tau_V^2}V_i^TQV_i\right\}, \tag{2.12}$$

where Q is, again, the precision matrix defined by the CAR structure. For the wafer-specific CAR effects, we introduce between wafer correlation within each lot via an equicorrelated structure. Let W_i be a concatenation of W_{ij} where the same dies from all intra-lot wafers are stacked together (i.e. $W_i = (W_{i11}, W_{i21}, ..., W_{i(J-1)K}, W_{iJK})^T$), then the prior CAR form is

$$[W_i|\tau_W^2] \propto \exp\left\{-\frac{1}{2\tau_W^2}W_i^T(Q\otimes T^{-1}(\delta))W_i\right\}, \tag{2.13}$$

where $T = (1-\delta)\mathbf{I}_{J\times J} + \delta \mathbf{1}_J \mathbf{1}_J^T$. Thus, each W_i is an improper multivariate CAR prior distribution, which we denote as $MCAR(0, \tau_W^2, T)$, obvious modification of the previous MCAR notation. The parameter δ controls the degree to which dies in neighboring positions on different intra-lot wafers affect any given die (i.e., the level of information sharing between intra-lot wafers). Its effect is seen through the full conditional distributions in the supplementary material. Note that T^{-1} exists if $\delta \in (-\frac{1}{p-1}, 1)$, and, since we assume $\delta \in [0,1]$, this is not an issue. Furthermore, because of the form of T, the inverse has a closed form

$$T^{-1} = \frac{1}{1-\delta}\left(\mathbf{I} - \frac{\delta}{((J-1)\delta+1)}\mathbf{1}\mathbf{1}^T\right). \tag{2.14}$$

Thus, the elements of T^{-1} denoted T_{ij}^{-1} are

$$T_{ij}^{-1} = \begin{cases} \frac{1+(J-2)\delta}{(1-\delta)((J-1)\delta+1)} & \text{if } i = j \\ \frac{-\delta}{(1-\delta)((J-1)\delta+1)} & \text{if } i \neq j \end{cases}. \tag{2.15}$$

Again, the intent of this specification is to capture dependence between wafers within a lot through an exchangeable specification. Formally, the model becomes:

$$\Pi_i \Pi_j \Pi_k [Y_{ijk} | \boldsymbol{\beta}, U_k, V_{ik}, W_{ijk}, \sigma^2][\mathbf{U}|\tau_U^2]\Pi_i[\mathbf{V}_i|\tau_V^2]\Pi_i[\{\mathbf{W}_{ij}\}|\tau_W^2, \delta]. \tag{2.16}$$

For this model, we take

$$\begin{aligned} \tau_U^2 &\sim IG(a_{\tau_U}, b_{\tau_U}) & \\ \tau_V^2 &\sim IG(a_{\tau_V}, b_{\tau_V}) & \delta &\sim \text{Unif}(0,1) \\ \tau_W^2 &\sim IG(a_{\tau_W}, b_{\tau_W}) & \boldsymbol{\beta} &\sim N(m_\beta, V_\beta), \end{aligned} \tag{2.17}$$

where $a_{\tau_U} = a_{\tau_V} = a_{\tau_W} = b_{\tau_U} = b_{\tau_V} = b_{\tau_W} = 1$, $m_\beta = \mathbf{0}$, and $V_\beta = 10^{12}\mathbf{I}$. Each of the posed models is fitted in the latent variable Bayesian probit model setting, proposed by Albert and Chib (1993). The full conditional distributions for fitting this

model are minor modifications to those given in the supplementary material. That is, latent Gaussian variables Z_{ijk} with variance 1 are introduced and are sampled from truncated normal distributions according to the associated Y_{ijk}. Again, prediction follows using composition sampling.

2.4.2 Results

The model comparison criteria for the four models discussed in Section 2.4 are given in Table 2.2. Specifically, we include overall prediction accuracy, sensitivity, specificity, and the Brier score. Unlike Section 2.3.2, we are not interested in measuring smoothing; instead the goal of the CAR modeling is to improve prediction. Note that, relative to the trend-only model, every CAR model has higher accuracy, specificity, and sensitivity and lower Brier scores. Similarly, we see higher accuracy, specificity, and sensitivity and lower Brier scores as we increase the number of levels in the nested CAR model. For this example, the nested CARs structure improves model performance relative to simpler CAR models and the trend-only model. We note that with a 50% failure rate under the simulation and only binary response to inform about the latent Gaussian specifications, we can not expect better performance than the table reveals.

Table 2.2: Model comparison criteria for nested CARs model for binary semiconductor chip data

	Accuracy	Sensitivity	Specificity	Brier Score
Trend Model	0.5468	0.5363	0.5568	0.2969
Trend+Global CAR	0.5527	0.5423	0.5628	0.2983
Trend+Global+Lot CAR	0.5876	0.5779	0.5971	0.2758
Trend+Global+Lot+Wafer CAR	0.6013	0.5919	0.6105	0.2722

2.5 The Gaussian processes setting

For the Gaussian process setting, we view the die-level observations as block averages arising from a response surface under a Gaussian process over locations on a

wafer. We find ourselves in a version of the functional ANOVA setting presented by Kaufman and Sain (2010). Again, we let i index lots, j index wafers, k index dies on wafers. Again, we observe 20% of the available dies $i = 1, 2, ..., 5$, $j = 1, 2, ..., 7$, and $k' = 1, 2, ..., 39$ from $k = 1, 2, ..., 195$. The generative model for this simulated data and the appearance of the data are given in A.2.

So, again we observe 1365 dies and seek to predict to 5460 dies. Here, the simulated data arises from a continuous surface (e.g., thickness, conductivity, voltage, frequency) but is assumed to be "measured" at die level. We take the wafer thickness to be the surface (in units of μm). That is, we imagine there is a $Y_{ij}(\mathbf{s})$ at every $\mathbf{s}$ on a wafer, and we think of die observations $Y_{ijk} \equiv Y_{ij}(B_k) = \int_{B_k} Y_{ij}(\mathbf{s})/|A|$ where $|A|$ is the common area for all of the dies B_k. In particular, we imagine a smooth mean surface that is observed with measurement error. Within die and across die variation is expected (Mittal, 2016).

We work with approximation to the block average, $\tilde{Y}_{ij}(B_k) = \frac{1}{m}\sum_{\ell=1}^{m} Y(\mathbf{s}_{k,\ell})$. Here, the die are $1mm \times 3.5mm$ rectangles and we choose $m = 5$ points to include the centroid and four others in fixed locations within the rectangle using a geometric lattice design (see Figure 2.6).

Turning to the modeling, we present a point-referenced specification with ANOVA structure which parallels the structure in the CAR version of the previous section. The intent is to capture a manufacturing process that produces spatially similar wafers within a lot and also spatially similar wafers across lots. This suggests a global spatial process with lot-to-lot variation and, within a lot, wafer-to-wafer variation. Again, we adopt a quadratic global trend surface in the x and y coordinates at location $\mathbf{s}$ in the global specification. A simple trend surface is not intended to well-explain wafer surfaces. Rather, inclusion may make it easier for the GP to make local adjustment than with a constant mean; more complicated trend surfaces could

be explored but the goal is not to build a complex parametric mean model. We write $\mu(\mathbf{s};\boldsymbol{\beta})$ to denote this trend surface and write the global surface as $\mu(\mathbf{s};\boldsymbol{\beta}) + U(\mathbf{s})$ where $U(\mathbf{s})$ is a mean 0 GP with covariance function $\sigma_U^2\rho(\mathbf{s}-\mathbf{s}';\phi_U)$. We employ an exponential covariance function. $U(\mathbf{s})$ is intended make all of the $Y_{ij}(\mathbf{s})$ dependent.

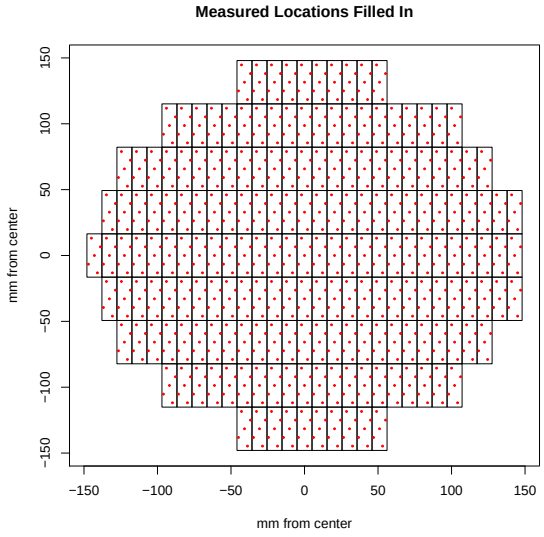


FIGURE 2.6: Lattice used on every wafer to approximate the continuous Gaussian process

Next, the lot level random effects are $V_i(\mathbf{s})$, each an independent mean 0 GP with covariance function $\sigma_V^2\rho(\mathbf{s}-\mathbf{s}';\phi_V)$. Then, we add the wafer random effects, $W_{ij}(\mathbf{s})$, These are assumed independent across lots i but dependent between wafers j within a lot, again because wafers are sliced consecutively from intra-lot core. Marginally, they are mean 0 GP's with covariance function $\sigma_W^2\rho(\mathbf{s}-\mathbf{s}';\phi_W)$. The dependence idea argues that the adjustment surface for wafer j and the adjustment surface for wafer j' within lot i might be dependent, similar to the model presented in Section 2.4. In fact, we will assume the $W_{ij}(\mathbf{s})$ are exchangeable (see below). Specifically, if we concatenate $\{W_{ij}\}$, so $W_i = (W_{i1}^T, ..., W_{iJ}^T)^T$, then $W_i \sim N(0, T \otimes \sigma_W^2 H(\phi_W))$, where T is defined as it is in Section 2.4.1.

So, altogether, we write our model as

$$Y_{ij}(\mathbf{s}) = \mu(\mathbf{s};\boldsymbol{\beta}) + U(\mathbf{s}) + V_i(\mathbf{s}) + W_{ij}(\mathbf{s}) + \epsilon_{ij}(\mathbf{s}), \tag{2.18}$$

where $\epsilon(\mathbf{s}) \overset{iid}{\sim} N(0,\tau^2)$ captures the measurement error. We can explicitly calculate the dependence structure:

$$\begin{aligned}\mathtt{cov}(Y_{ij}(\mathbf{s}), Y_{i'j'}(\mathbf{s}')) = \sigma_U^2\rho(\mathbf{s}-\mathbf{s}';\phi_U) + \sigma_V^2\rho(\mathbf{s}-\mathbf{s}';\phi_V)1(i=i') \\ + \sigma_W^2\rho(\mathbf{s}-\mathbf{s}';\phi_W)1(i=i')[1(j=j') + \delta 1(j\neq j')].\end{aligned} \tag{2.19}$$

Here, $0 < \delta < 1$ introduces the common dependence between wafers within lots. So, altogether, we have specified a 5×7 dimensional, stationary GP with complex dependence structure.

Finally, we can write out the full model. We assume the set of sampled dies is the same for all i, j and let $\mathbf{Y}_{ij}$ be the 39×1 vector of observations on wafer j in lot i. Because $m = 5$, $\mathbf{U}$ is 195×1 vector of $U(\mathbf{s})$ associated with the sampled dies. Similarly, let $\mathbf{V}_i$ be the 195×1 vector of $V(\mathbf{s})$ associated with the sampled dies and let $\mathbf{W}_{ij}$ be the 195×1 vector of $W(\mathbf{s})$ associated with the sampled dies. Furthermore, let $\sigma_U^2 H(\phi_U)$ be the 195×195 covariance matrix associated with $\mathbf{U}$, let $\sigma_V^2 H(\phi_V)$ be the 195×195 covariance matrix associated with the $\mathbf{V}_i$, and $\sigma_W^2 H(\phi_W)$ be the 195×195 covariance matrix associated with the $\mathbf{W}_{ij}$. Also, let $\boldsymbol{\mu}(\boldsymbol{\beta})$ be the 39×1 vector with entries being the trend surface at each of the sampled locations. For convenience, we choose to use the locations of the centroids instead of block averaging of the trend surface. We introduce an averaging matrix $\mathbf{A}$ that is 39×195, where

$$A_{kl} = \begin{cases} 1/m \text{ if } m \times (k-1) + 1 \leq l \leq m \times k \\ 0 \text{ otherwise.} \end{cases} \tag{2.20}$$

The $\mathbf{A}$ matrix averages the appropriate $m = 5$ GP components for the k-th block (die). Then, we have the following explicit model for the Y_{ijk} expressed through $\mathbf{Y}_{ij}$:

$$\mathbf{Y}_{ij} = \boldsymbol{\mu}(\boldsymbol{\beta}) + \mathbf{A}(\mathbf{U} + \mathbf{V}_i + \mathbf{W}_{ij}) + \boldsymbol{\epsilon}_{ij}. \tag{2.21}$$

As we did with the nested CAR example, we consider comparison of four models here: (i) a trend surface only model, $\mathbf{Y}_{ij} = \boldsymbol{\mu}_{ij}(\boldsymbol{\beta}) + \boldsymbol{\epsilon}_{ij}$, (ii) trend plus global GP, $\mathbf{Y}_{ij} = \boldsymbol{\mu}_{ij}(\boldsymbol{\beta}) + \mathbf{A}\mathbf{U} + \boldsymbol{\epsilon}_{ij}$, (iii) trend plus global and lot GPs, $\mathbf{Y}_{ij} = \boldsymbol{\mu}_{ij}(\boldsymbol{\beta}) + \mathbf{A}(\mathbf{U} + \mathbf{V}_i) + \boldsymbol{\epsilon}_{ij}$, and (iv) full nested GP, as above, $\mathbf{Y}_{ij} = \boldsymbol{\mu}_{ij}(\boldsymbol{\beta}) + \mathbf{A}(\mathbf{U} + \mathbf{V}_i + \mathbf{W}_{ij}) + \boldsymbol{\epsilon}_{ij}$. As in Section 2.4, this comparison will allow us to examine whether GP models improve prediction, which we expect, and how the nested GP structure improves prediction.

2.5.1 Priors, model fitting, prediction

Using the model offered in Section 2.5, we adopt the following prior distributions:

$$\begin{aligned} \sigma_U^2 &\sim IG(a_{\sigma_U}, b_{\sigma_U}) & \tau^2 &\sim IG(a_\tau, b_\tau) \\ \sigma_V^2 &\sim IG(a_{\sigma_V}, b_{\sigma_V}) & \delta &\sim \text{Unif}(0,1) \\ \sigma_W^2 &\sim IG(a_{\sigma_W}, b_{\sigma_W}) & \boldsymbol{\beta} &\sim N(m_\beta, V_\beta), \end{aligned} \tag{2.22}$$

where $a_{\sigma_U} = b_{\sigma_U} = a_{\sigma_V} = b_{\sigma_V} = a_{\sigma_W} = b_{\sigma_W} = a_\tau = b_\tau) = 1$, $m_\beta = \mathbf{0}$, and $V_\beta = 10^{12}\mathbf{I}$. For computational convenience and for identifiability (Zhang, 2004), we fix all $\phi = 10/\max(d)$, where $\max(d)$ is the maximum distance between two dies on a wafer (so, the range is $0.3\max(d)$). We studied sensitivity by fitting the model with $\phi \in \{1/\max(d), 3/\max(d), 20/\max(d)\}$ with indistinguishable results from $\phi = 10/\max(d)$. If treated as unknown, common prior choices for ϕ include uniform, gamma, and discrete distributions. All parameters besides decay parameters, ϕ's, and the equicorrelation parameter δ can be updated using a Gibbs sampler. The full conditional distributions are given in the supplementary material.

Turning to prediction, we need the distribution $[Y_{ij}(B_0)|\mathbf{Y}]$, where $\mathbf{Y}$ represents all observed dies. With composition sampling, we need $[Y_{ij}(B_0)|\mathbf{Y}, \texttt{parameters}]$. This is a normal distribution arising from the joint distribution and only needs $\mathbf{C}_U(\mathbf{s}_0)$, the vector of covariances of $U(\mathbf{s}_0)$ with $\mathbf{U}$, similarly, $\mathbf{C}_V(\mathbf{s}_0)$ and $\mathbf{C}_W(\mathbf{s}_0)$ as well as $\mu(\mathbf{s}_0; \boldsymbol{\beta})$.

2.5.2 Results

The results for the four models proposed in Section 2.5 are given in Table 2.3. The model comparison criteria used are 90% prediction interval coverage, PMSE, PMAE, and CRPS. Additionally, the 90% prediction interval width is included to show the effect of the sequential nesting on prediction certainty. As expected, all GP models outperform the trend-only model; however, there is a significant improvement when we include lot-specific GPs. When wafer level GPs are included, we see some predictive improvement, but small compared with the introduction of the lot-specific GPs. Altogether, in this example, the nested GP models predict increasingly better as more nested layers are added.

Table 2.3: Model comparison criteria for nested GPs model for binary semiconductor chip data

	90% Coverage	Interval Width	PMSE	PMAE	CRPS
Trend-only Model	0.8980	71.3361	494.4950	17.4637	12.4070
Trend+Global GP Model	0.8982	67.1237	432.9598	16.1502	11.5059
Trend+Global+Lot GP Model	0.9284	53.6796	250.4771	11.2362	8.3814
Full Nested GP Model	0.9317	53.5612	231.2235	10.5191	7.9982

2.6 Summary and future work

We have proposed areal unit prediction for two primary purposes: (i) prediction in the presence of partially sampled units and (ii) prediction for comparing smoothing under competing areal unit models. For the first purpose, we consider both MRF and GP models to explain the missingness. Through examples we have demonstrated improved predictive performance relative to a nonspatial model. For the second purpose, we have demonstrated predictive smoothing comparison using suitable predictive criteria. The initial motivation for this work arose from a challenge in assessing semiconductor chip performance under varying manufacturing schemes. We illustrate this scenario with a complex nested spatial model using a binary and

a continuous measure of performance. Extensions of this work could explore more complicated specifications of MRFs, including higher dimensional examples (e.g., voxels in neuroimaging). Also of interest are space-time problems where both purposes above need to be investigated dynamically. Additionally, significant work could be done in multivariate areal unit prediction, particularly where the responses are dependent but of different data types, e.g., categorical and continuous.

3

Pollution State Modeling for Mexico City

3.1 Introduction

Long-term exposure to air pollution is strongly linked with respiratory and cardiovascular disease and leads to increased mortality as well as hospital admissions (see, e.g., Brunekreef and Holgate, 2002). Particulate matter (PM) is defined to be solid particles and liquid droplets in the air. PM comes from direct emissions (primary particles) and chemical reactions between other pollutants (secondary particles). Particulate matter, and in particular PM with diameter less than 10 μm (PM_{10}), is known to increase human mortality and morbidity (see, e.g., Brunekreef and Holgate, 2002; Pope III and Dockery, 2006; Loomis et al., 2013; Hoek et al., 2013). Because PM generally has a short lifetime, urban and other high emission areas generally have higher concentrations of PM_{10} than rural areas (see Clements et al., 2012, as an example).

Unlike PM, ground-level ozone (O_3) is not emitted directly but is instead formed by chemical reactions between of nitrogen oxides and volatile organic compounds, a reaction that requires heat and sunshine (see, e.g., Sillman, 1999). Ozone is often

as high in rural areas as it is in urban areas (see, e.g., Angle and Sandhu, 1989; Sillman, 1999; Dueñas et al., 2004). Ozone is linked to a variety of negative health outcomes, including short-term respiratory events, long-term respiratory disease, increased mortality, and low birth weight (see, e.g., Lippmann, 1989; Salam et al., 2005; Bell et al., 2006; Weschler, 2006). Because of these adverse outcomes, regulatory agencies institute policies to monitor and limit pollution levels, especially PM_{10} and ozone. Urban areas are often monitored more closely to protect larger populations due to higher pollution levels found in urban environments (see, e.g., Heal and Hammonds, 2014). The detrimental health effects of air pollution in the Mexico City metropolitan area are well-studied (see Mage et al., 1996; Romieu et al., 1996; Hernández-Garduño et al., 1997; Loomis et al., 1999; Bravo-Alvarez and Torres-Jardón, 2002; Barraza-Villarreal et al., 2008; Riojas-Rodríguez et al., 2014). Thus, Mexican authorities have implemented a variety of regulations to control pollution levels in Mexico, and specifically in Mexico City.

In spite of several polices implemented by environmental authorities in Mexico and Mexico City over the past 30 years, the city and its metropolitan area still suffer with high levels of pollution (see, e.g., Bravo-Alvarez and Torres-Jardón, 2002; Zavala et al., 2009; Rodríguez et al., 2016; Davis, 2017; Instituto Nacional de Ecología y Cambio Climático (INECC), 2017; Gouveia et al., 2018). Some of the most recent measures implemented are new thresholds limiting ozone and PM_{10} concentrations nation-wide which decreased allowable pollution levels relative to previous thresholds (Diario Oficial de la Federación, 2014a,b). Thresholds are updated every five years based on current research on the effect of pollutants on human health. In these new standards, the ozone thresholds were reduced to 95 parts per billion (ppb) or, equivalently, 0.095 parts per million (ppm) for *hourly* ozone and 70 ppb for *eight-hour average* ozone (Diario Oficial de la Federación, 2014b). Additionally, the allowable 24-hour average PM_{10} concentration threshold was lowered to 75 micrograms per

cubic meter ($\mu g/m^3$) (Diario Oficial de la Federación, 2014a). These thresholds are not used to reduce pollution levels but are instead established to ensure human health protection and to evaluate air quality.

By comparison, the United States Environmental Protection Agency (EPA) limits 24-hour average PM_{10} concentration to not exceed 150 $\mu g/m^3$ and 8-hour average ozone concentration to not exceed 70 ppb. (101st United States Congress, 1990). The European Union restricts 24-hour average PM_{10} concentration to not exceed 50 $\mu g/m^3$ and 8-hour average ozone concentration to not exceed 120 ppb (European Environment Agency, 2016). Thus, Mexican ambient air quality standards (which we denote MAAQS) are progressive when compared to American and European standards. Mexico City's pollution emergencies, however, are not related to the Mexican national standards and instead use more permissive thresholds.

Mexico City's thresholds, established by the Atmospheric Environmental Contingency Program in Mexico City, are used to indicate times when pollutant concentrations are high enough to cause significant damage to human health (Administración Pública de la Ciudad de México, 2016). Thus, the goals of Mexico City's Atmospheric Environmental Contingency Program differ from those specified for Mexico's ambient air quality standards. When emergency phases (or events) are activated, the aim is to control emission levels to decrease air pollution and its harmful effects to the population. It is worth mentioning that thresholds have decreased significantly. For instance, the thresholds for declaring the equivalent emergencies 1995-2000 were 1.5-2 times the current limits, depending on the type of emergency declared (Departamento del Distrito Federal et al., 1996).

Mexico City and its metropolitan area are split into five regions: northeast (NE), northwest (NW), central (CE), southeast (SE), and southwest (SW). Within these five regions, there are 24 monitoring stations that record both hourly ozone and PM_{10} levels during the year 2017. To control the health risks associated with high

ozone and PM_{10}, environmental alerts are declared if either hourly ozone or 24-hour average PM_{10} levels exceed certain pollutant-specific thresholds which rely on regulatory suggestions that differ from those presented in Diario Oficial de la Federación (2014a,b). Depending on the levels of the pollutant, either a phase I or a phase II alert is declared. Phase I is declared when hourly ozone exceeds L_1^O =0.154 ppm (154 ppb) or 24-hour average PM_{10} exceeds $L_1^{PM} = 214\ \mu g/m^3$. During a phase I emergency, people are encouraged to limit outdoor time, exercise, smoking, and consumption of gas. Additionally, several transportation protocols are instituted to reduce vehicular emissions. Similarly, phase II is declared when hourly ozone exceeds $L_2^O = 0.204$ ppm (204 ppb) or 24-hour average PM_{10} exceeds $L_2^{PM} = 354\ \mu g/m^3$. Phase II institutes stricter protocols than phase I, including restricting circulation of official vehicles and strictly limiting civilian and commercial emissions. See Administración Pública de la Ciudad de México (2016) for details regarding Mexico City's pollution emergency phases.

Compared to MAAQS, Mexico City's Atmospheric Environmental Contingency Program thresholds are more tolerant of high pollution levels. The phase I thresholds for ozone are 1.6 times the Mexican legal limits, while the phase I thresholds for PM_{10} are almost three times MAAQS. The phase II thresholds are roughly two and five times MAAQS for ozone and PM_{10}, respectively. The protocols for phase I or phase II are the same regardless of the pollutant that triggered the alert. If ozone thresholds are exceeded in any region, i.e., the maximum over any station within the region, then emergency phases are declared city-wide (i.e., in all regions), where the phase is determined by which threshold (L_1^O or L_2^O) was exceeded.

On the other hand, PM_{10} exceedances could trigger regional or city-wide phase alerts, depending on which stations exceed the allowable limits. More explicitly, if the maximum 24-hour average over stations within the same region exceeds a PM_{10} threshold, then the environmental alert is declared only in that region. However,

if the maxima for two or more regions exceed a given PM_{10} threshold, then the environmental alert is declared over the entire metropolitan area (i.e. in all five regions). This description is summarized in Table 3.1.

Table 3.1: Description of Mexico City emergency phase alerts. Note that ozone thresholds are for hourly ozone, while PM_{10} limits are for 24-hour running average PM_{10}.

Phase	Region-wide Alert	City-wide Alert
None	• $PM_{10} < L_1^{PM}$ and $O_3 < L_1^O$ for all regions • No higher-order alerts supersede	• $PM_{10} < L_1^{PM}$ and $O_3 < L_1^O$ for all regions
I	• $PM_{10} \geq L_1^{PM}$ within the region • And no higher-order alerts supersede	• $O_3 \geq L_1^O$ for any region • Or $PM_{10} \geq L_1^{PM}$ for two or more regions • And no higher-order alerts supersede
II	• $PM_{10} \geq L_2^{PM}$ within the region	• $O_3 \geq L_2^O$ for any region • Or $PM_{20} \geq L_2^{PM}$ for two or more regions

Pollution emergency phases are only suspended when pollution levels for every station drop below phase I thresholds (i.e. the conditions for no phase alerts are met). For practical reasons, evaluation of the emergency phases is carried out three times daily at 10 AM, 3 PM, and 8 PM (Administración Pública de la Ciudad de México, 2016). Ultimately, however, phase activation and suspension are dependent on meteorological forecasts in addition to observed pollution levels. Because the additional meteorological criteria are not explicitly outlined, we do not attempt to predict actual phase occurrence but instead quantify the *risk* of a phase occurrence.

The contribution here is to understand and predict how often Mexico City was at risk of a pollution emergency in terms of (1) the Atmospheric Environmental Contingency Program in Mexico City and (2) current Mexican ambient air quality standards. For both, we assess how the risk of dangerous pollution varies over city regions and over time. As described above, for Mexico City's Atmospheric Environmental Contingency Program, alerts are triggered when one or more stations exceeds thresholds. Thus, emergency phases depend entirely upon pollutant maxima within each region. Furthermore, environmental alerts are often summarized over coarser temporal scales, like days, rather than the measurement level (hours) or the three

hours of evaluation (10 AM, 3 PM, and 8 PM). So, daily emergency phases depend on pollutant maxima over hours of evaluation and stations within each region. Regarding MAAQS, we can do inference at each of three natural spatial scales: station-level, region-level, or city-level. Again, we may be interested in exceedances occurring on a daily scale rather than hourly. Therefore, we again need maxima over time (and potentially space depending on the spatial scale selected).

In summary, the foregoing tasks addressed here, the analyses of emergency contingency plan and Mexican ambient air quality standards, rely on the same pollution level data. Therefore, we develop, using model choice over a selection of models, a single hierarchical bivariate spatiotemporal model for hourly ozone and PM_{10} levels. Predictions from our model serve two practical purposes: First, our predictions allow us to carry out probabilistic inference about pollution emergency states or national compliance issues. Second, if implemented in practice, our model could warn of potential pollution emergencies or compliance problems, allowing regional and city-wide adjustments and responses to be made earlier. From this model, all prediction and inference regarding emergency phases and legislation-based exceedances becomes a post-model fitting exercise, as we demonstrate.

By now there is a rich literature on modeling both O_3 and PM at both coarse ($10\mu m$) and fine ($2.5\mu m$) scale. Here, we highlight some examples relevant to our analysis. Sahu et al. (2007) use a hierarchical space-time model to model square-root ozone with the goal of assessing long term trends in ozone in Ohio. Cocchi et al. (2007) adopt a hierarchical model for log-PM_{10} concentrations to characterize the effect of meteorological conditions on the PM_{10} process and to estimate PM_{10} at unmonitored locations. Berrocal et al. (2010) model square-root ozone using data of two types, output from numerical models and data collected from monitoring networks, that are misaligned on spatial scales using spatially-varying regression coefficients (Gelfand et al., 2003). Huang et al. (2018) model log-PM_{10} and log-

nitrogen dioxide (NO_2) jointly and assessed the effect of these pollutants on health outcomes in Scotland. Similarly, we adopt a hierarchical space-time model with site-specific regression and auto-regressive coefficients for square-root ozone and log-PM_{10} concentrations in Mexico City, Mexico.

In this paper, we start by presenting and discussing the Mexico City pollution dataset in Section 3.2, highlighting data characteristics that inform modeling decisions. In Section 3.3, we discuss modeling decisions, model fitting, inference, and selection. We present and discuss results in the context of both Mexico City's Atmospheric Environmental Contingency Program and MAAQS in Section 3.4. In this section, we first present a comprehensive analysis of Mexico City's phase alert system, predicting pollution levels and associated phase levels at 10 AM, 3 PM, and 8 PM to mirror the actual phase activation and suspension procedure. Then, we carry out inference on MAAQS exceedances and compare these results to emergency phase predictions to demonstrate differences between these standards. We provide concluding discussion regarding our results and statistical modeling in Section 3.5.

3.2 Mexico City Pollution Dataset

In this dataset, we have hourly ozone and PM_{10} measurements at $N_s = 24$ stations across Mexico City, Mexico for the duration of 2017. Ozone and PM_{10} measurements are obtained minute by minute at each station, and the hourly measurement reported is an average of the 60 minute-by-minute measurements. Let Y_{it}^O, Y_{it}^{PM} denote ozone and PM_{10} levels, respectively, at station i and time t with units of hours. Consequently, we observe measurements over $N_t = 8760$ times at each station, giving $N = 210240$ pairs of ozone and PM_{10} concentration across the 24 stations, across the entire year.[1] Relative humidity (RH) and temperature (TMP) are measured over

[1] Missing hourly measurements were imputed using the corresponding measurements at the nearest station within the same region. If no stations in that region recorded a measurement at that time,

the same space-time grid as ozone and PM_{10} and are used as explanatory variables for both ozone and PM_{10}.

As mentioned above, Mexico City is partitioned into five regions which are employed for defining environmental alert phases (See Section 3.1). Abbreviated station names, corresponding regions, and annual summaries of pollution levels are given in Table 3.2. Station locations are plotted in Figure 3.1 using the R package GGMAP (Kahle and Wickham, 2013). Besides being on the main wind path (from NE to SW) and therefore receiving many ozone precursors from the NE region, the CE region is heavily-trafficked by automobile. The SW region, located and the end of the NE-SW wind corridor, receives ozone produced along this wind path, and this ozone stays trapped in the SW region due to mountains on its southwest boundary.

Note that the number of stations in each region differs. Moreover, pollution levels appear to vary over the regions. The northeast region has the highest average PM_{10}, while the southeast and southwest regions have the highest average ozone. Hence, the regional maxima, used for phase alerts, are expected to have very different hourly distributions. Note that, with the maxima being taken over a small number of stations in each region, there is no reason to attempt to employ extreme value theory here. We model at the station-level rather than at the regional level so that the regional maximum distributions are *induced* by the station-level modeling. In this regard, because ozone and PM_{10} are strictly non-negative, we consider modeling the station data using either transformations to $\mathbb{R}$ or using strictly positive data models.

then the nearest station in a different region provided the missing value. This was done prior to our receiving the data for analysis.

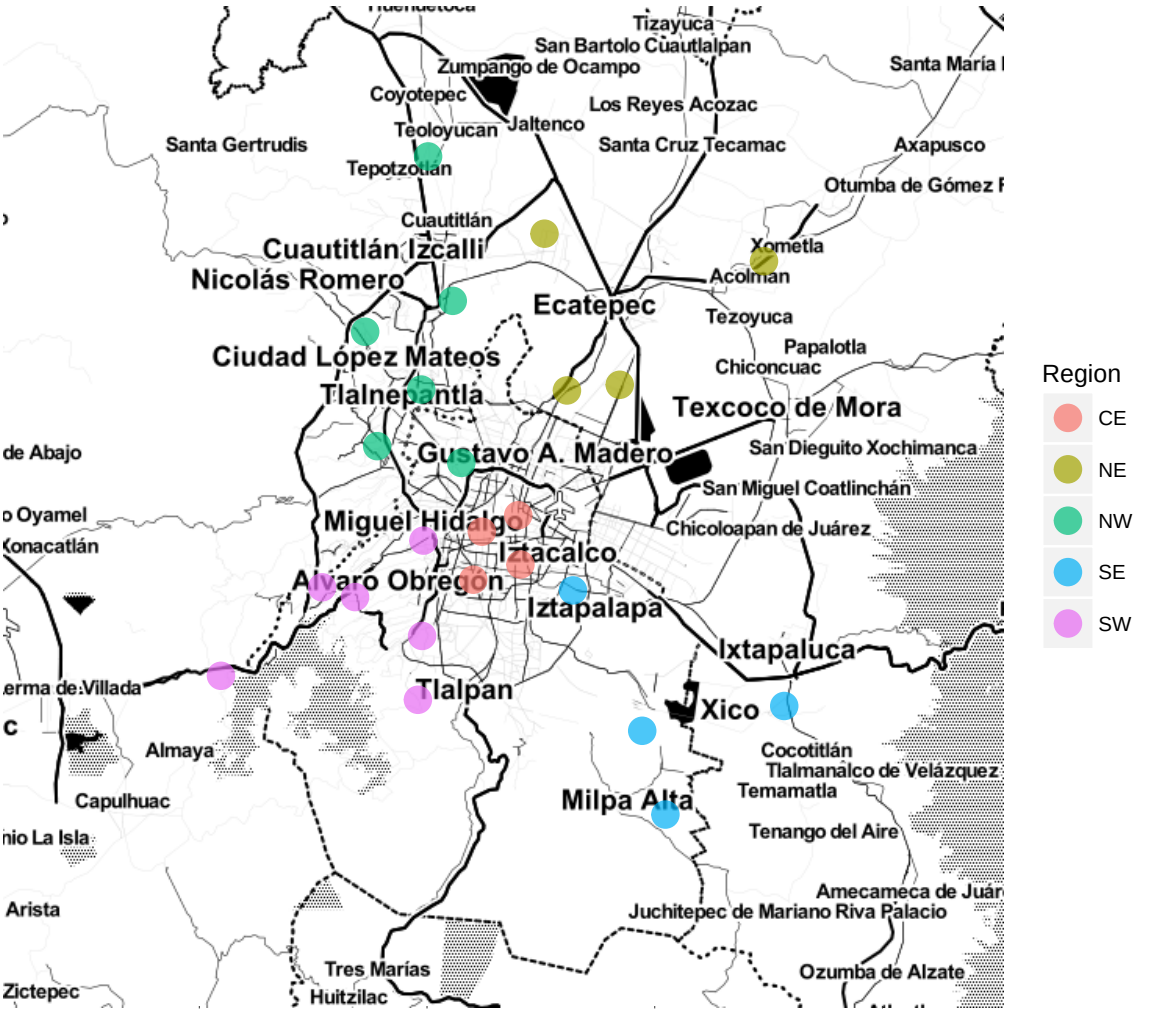


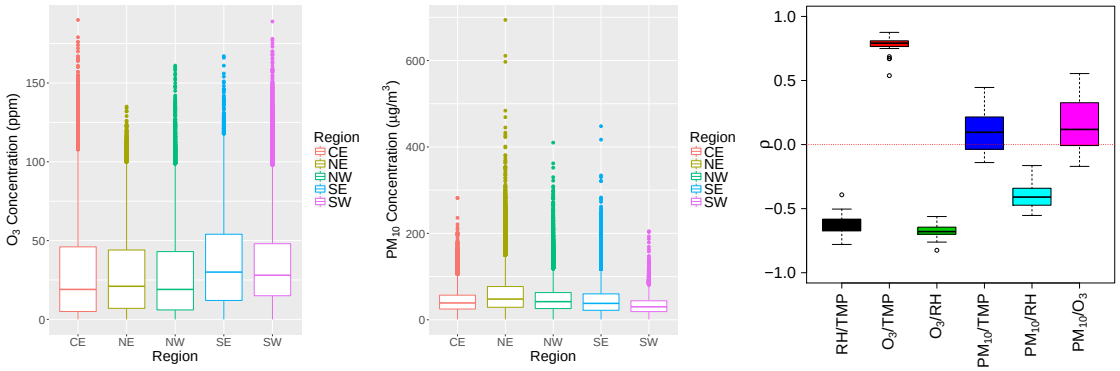
FIGURE 3.1: Station locations with regional labels

Region-specific box-plots for ozone and PM_{10} are plotted in Figures 3.2a and 3.2b. For ozone, the SE region has the highest mean, but the CE and SW regions have the most extreme values. Note that the NE region has the highest average PM_{10}, as well as the most extreme values. This is because the NE region houses a large industrial section that generates many direct pollutants, including particulate matter. To explore the relationships between covariates (RH and TMP), outcomes (ozone and

PM_{10}), and covariates and outcomes, we compute the station-specific Spearman's ρ for all covariate-covariate, covariate-outcome, and outcome-outcome relationships. As a rank correlation, Spearman's ρ avoids concern regarding transformations and outlying values. We plot these site-specific correlation coefficients in Figure 3.2c.

Table 3.2: Station names and regions. Average ozone and PM_{10} across regions are given.

Region	Stations	Station Names	Annual Average Ozone	Annual Average PM_{10}
Northeast	4	ACO, SAG, VIF, XAL	28 ppb	59 $\mu g/m^3$
Northwest	6	ATI, CAM, CUT, FAC, TLA, TLI	27 ppb	49 $\mu g/m^3$
Central	4	BJU, HGM, IZT, MER	29 ppb	44 $\mu g/m^3$
Southeast	4	CHO, MPA,TAH, UIZ	36 ppb	45 $\mu g/m^3$
Southwest	6	AJM, CUA, INN MGH, PED, SFE	34 ppb	33 $\mu g/m^3$



(a) Region O_3 boxplot (b) Regional PM_{10} boxplot (c) Correlation boxplots

FIGURE 3.2: Region-specific boxplots for (Left) ozone and (Center) PM_{10}. (Right) Site-specific Spearman's ρ for (from left to right) relative humidity and temperature, ozone and temperature, ozone and relative humidity, PM_{10} and temperature, PM_{10} and relative humidity, and PM_{10} and ozone.

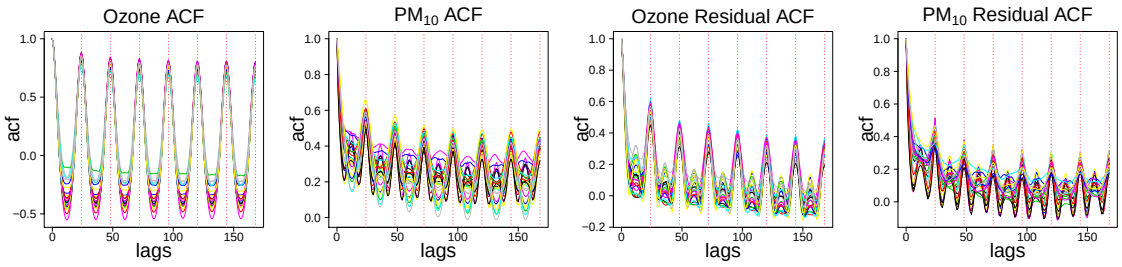
The relationships between ozone and covariates (RH and TMP) are strong for all sites, while the relationships between PM_{10} and covariates (RH and TMP) vary much more across sites. However, there appears to be a strong negative correlation

between RH and PM_{10} for all locations. On the other hand, Spearman's ρ varies greatly across sites for TMP and PM_{10}. Similarly, there appear to be important relationships between ozone and PM_{10} depending on the site, motivating the use of a joint model for ozone and PM_{10}. The variability of the outcome, covariate, and outcome-covariate relationships across stations motivates a hierarchical model for covariate effects.

We find strong daily and weekly patterns for both pollutants. Using residual analysis for a model with site-specific effects for meteorological covariates, we still observe strong seasonal patterns for both day and week. Additionally, our preliminary analyses reveal strong correlation between the variance of residuals and the mean for both pollutants. This correlation could be addressed through modeling in a variety of ways. First, and most simply, one could use a variance stabilizing transformation (VST) to address the correlation between the mean and variance (e.g. log, square-root, Box-Cox) as was done by, for example, Sahu et al. (2007); Cocchi et al. (2007); Berrocal et al. (2010); Huang et al. (2018). Alternatively, we could use heteroscedastic models that specify variance directly as a function of hour or month. We consider both modeling approaches (transformations and heteroscedasticity) in Section 3.3.3. This exploratory analysis fleshed out below.

For preliminary examination of the temporal pattern of model residuals, we fit a model with site-specific regression coefficients effects for relative humidity and temperature on hourly ozone and PM_{10} concentrations. Although phase alerts depend on averaged PM_{10} levels and Mexican ambient air quality standards depend on averaged O_3 and PM_{10} levels, we carry out all modeling and exploratory analyses on the hourly pollutant levels. In Figure 3.3, we supply the empirical autocorrelation function (ACF) for both pollutants and their model residuals for each site. It is evident that daily seasonality is very strong for both pollutants. However, the ACF has somewhat similar behavior across sites but also varies significantly across sites,

suggesting the need for a hierarchical time-series specification. It should be noted that ACF plots serve as exploratory tools. However, they are not as useful for selecting specific lags (e.g. seasonality). To capture seasonality in the data, we consider models that use autoregressive (AR) terms of one day (24 hours) and one week (168 hours). These seasonal AR terms, in conjunction with AR terms of lower order, account for overall changes over the year in Figures 3.4b and 3.5b.



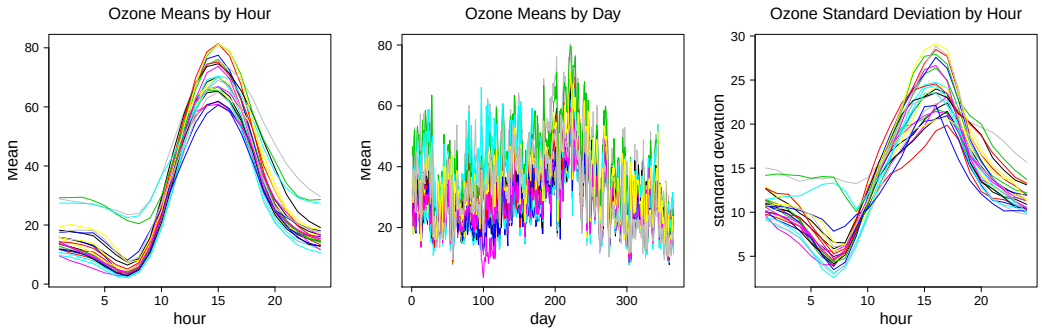
(a) O_3 ACF (b) PM_{10} ACF (c) O_3 Res-ACF (d) PM_{10} Res-ACF

FIGURE 3.3: Site-specific autocorrelation functions for ozone and PM_{10} for one week of lags. Here, we use Res-ACF to denote the autocorrelation function of the residuals.

We also examined partial ACF plots to gain insight into which autoregressive lags may be necessary in the model; however, we found them to be uninformative given the strong seasonality of the data. Ultimately, we use out-of-sample predictive performance to select the autoregressive structure for ozone and PM_{10}.

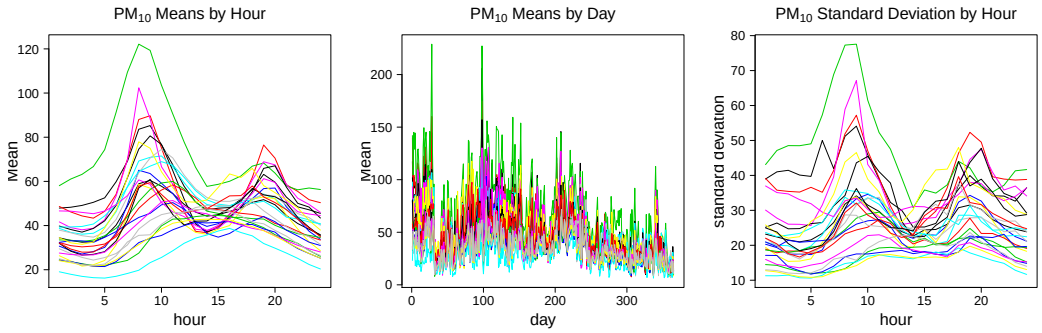
We plot site-specific means for both pollutants in two ways. We consider hourly means (Figures 3.4a and 3.5a), aggregated over the year, and daily means over the course of the year (Figures 3.4b and 3.5b). Figure 3.4a show that ozone concentrations generally peak around 4 pm (the warmest time of the day). In general, the highest ozone levels in Mexico City occur during spring months and June. In 2017, May, June, and July have the highest average ozone levels (see Figure 3.4b). Figure 3.5a shows two daily peaks in PM_{10} concentrations corresponding to commuting hours; however, annual trends for PM_{10} are less clear in Figure 3.5b. We also plot

standard deviation of ozone and PM_{10} concentrations as a function of hour of the day for all stations in Figures 3.4c and 3.5c. Note that there is strong correlation between the standard deviation and the mean for both pollutants (compare Figures 3.4a and 3.4c and compare Figures 3.5a and 3.5c). This correlation could be addressed through modeling in a variety of ways. First, and most simply, one could use a variance stabilizing transformation (VST) to address the correlation between the mean and variance (e.g. log, square-root, Box-Cox). Alternatively, we could use heteroscedastic models that specify variance directly as a function of hour or month.



(a) Hourly ozone means (b) Daily ozone means (c) Hourly ozone std. dev.

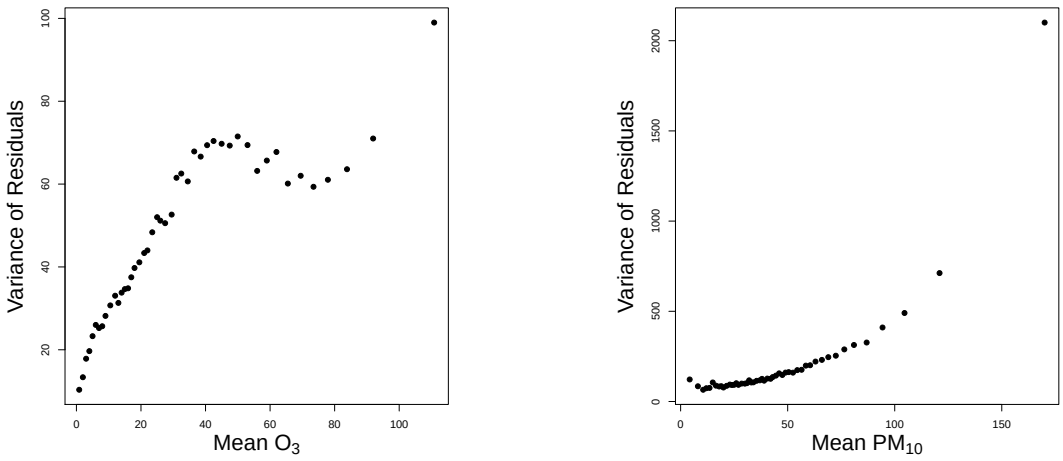
FIGURE 3.4: Site-specific means by hour averaged over the year



(a) Hourly PM_{10} means (b) Daily PM_{10} means (c) Hourly PM_{10} std. dev.

FIGURE 3.5: Site-specific hourly means, means averaged over every day, and standard deviations over hour of the day

To further investigate the relationship between the mean and variance, we fit a simple AR model with site-specific regression and AR coefficients and explore the residuals as a function of the mean. We include lags 1, 2, 24, and 168. After fitting the model, we bin observations according to their mean and calculate the variance of the associated residuals within that bin. The results for ozone and PM_{10} concentration are in Figure 3.6. These plots suggest that using VST's may effectively address the correlation between the mean and variance of model residuals. Specifically, the mean-variance relationship for ozone is strong (and approximately linear for values less than 50 ppb), and the mean-variance relationship for PM_{10} appears to be approximately quadratic. Linear mean-variance relationships are stabilized by a square-root transformation, and quadratic mean-variance associations are roughly removed using log transformations (see, for example, Section 3.3. in Hocking, 2013).



(a) O_3 mean-variance (b) PM_{10} mean-variance

FIGURE 3.6: Binned residual variance for ozone and PM_{10} plotted against mean.

3.3 Methods and Models

Given the exploratory analysis, a time-series analysis is certainly warranted. Because the data are collected hourly, because the exposure standards are at the scale of hours (or functions of hours), and because we can identify useful discrete lags which are difficult to capture with covariance specifications, we elect to work with discrete time rather than continuous time. Additionally, our exploratory analysis suggests that a model using either a VST or time-varying variance may describe the data more accurately than models using non-transformed data or homoscedastic models. As a result, we envision the model for these data to be

$$Y_{it}^{O} = \mathbf{x}_{i(t-1)}^{T}\boldsymbol{\beta}_{1i} + \mathbf{L}_{it}^{O^T}\boldsymbol{\gamma}_{1i} + \psi_{1i} + \epsilon_{1it} \tag{3.1}$$

$$Y_{it}^{PM} = \mathbf{x}_{i(t-1)}^{T}\boldsymbol{\beta}_{2i} + \mathbf{L}_{it}^{PM^T}\boldsymbol{\gamma}_{2i} + \psi_{2i} + \epsilon_{2it},$$

where Y_{it}^{O} is ozone concentration (or square-root ozone) and Y_{it}^{PM} is PM_{10} concentration (or log-PM_{10}) at site i and hour t. Here, $\mathbf{x}_{i(t-1)}$ includes an intercept, temperature, and relative humidity at site i and time $t-1$. We use covariates

from the previous hour because one of the primary purposes of this model is one-hour-ahead predictions for pollutants and corresponding phase alerts and exceedance probabilities and $\mathbf{x}_{it}$ will not be available for such prediction.

The parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_{11}, ..., \boldsymbol{\beta}_{1N_s}, \boldsymbol{\beta}_{21}, ..., \boldsymbol{\beta}_{2N_s})$ are station-specific regression coefficients for both PM_{10} and ozone. Because we imagine that the effect of humidity on pollutant concentrations is similar from region to region, we model regression coefficients exchangeably and hierarchically, centering effects on respective common means (see Gelman et al., 2014, for introductory thoughts on such hierarchical modeling). We define $\mathbf{L}_{it}^{O}$ and $\mathbf{L}_{it}^{PM}$ to be generic vectors of the lagged observations for ozone and PM_{10}, respectively with $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{11}, ..., \boldsymbol{\gamma}_{1N_s}, \boldsymbol{\gamma}_{21}, ..., \boldsymbol{\gamma}_{2N_s})$ as corresponding site-specific autoregressive coefficients. Lags for observations in early January 2017 ($\mathbf{L}_{it}^{O}$ and $\mathbf{L}_{it}^{PM}$) may depend upon observations from December 2016. The choice of components of $\mathbf{L}_{it}^{O}$ and $\mathbf{L}_{it}^{PM}$ becomes the model choice issue which we take up in Section 3.3. As with $\boldsymbol{\beta}$, we model $\boldsymbol{\gamma}$ hierarchically. Then, we have pure error terms, $\epsilon_{1it} \overset{iid}{\sim} N(0, \sigma_1^2)$ and $\epsilon_{2it} \overset{iid}{\sim} N(0, \sigma_2^2)$, or $\epsilon_{1it} \overset{ind}{\sim} N(0, \sigma_{1t}^2)$ and $\epsilon_{2it} \overset{ind}{\sim} N(0, \sigma_{2t}^2)$ for the heteroscedastic formulation.

Finally, to bring in spatial structure across the sites, jointly, ψ_{1i}, ψ_{2i} follow a bivariate conditionally autoregressive (CAR) model using coregionalization of two independent CAR models V_{1i} and V_{2i} (see Rue and Held, 2005; Banerjee et al., 2014). Coregionalization allows flexible, multivariate modeling (see, e.g., Matheron, 1982; Grzebyk and Wackernagel, 1994; Wackernagel, 1994; Banerjee et al., 2014). Explicitly,

$$\begin{pmatrix} \psi_{1i} \\ \psi_{2i} \end{pmatrix} = \mathbf{A}_{\psi}(V_{1i}, V_{2i})^T$$

$$\mathbf{A}_{\psi} = \begin{pmatrix} a_{11}^{(\psi)} & 0 \\ a_{12}^{(\psi)} & a_{22}^{(\psi)}, \end{pmatrix}$$

where $\mathbf{V}_1 = (V_{11}, V_{12}, ..., V_{1N_s})^T$ and $\mathbf{V}_2 = (V_{21}, V_{22}, ..., V_{2N_s})^T$. Equivalently, we

can view $a_{11}^{(\psi)}$ and $a_{22}^{(\psi)}$ as scale parameters for the $\mathbf{V}_1$ and $\mathbf{V}_2$, a fact we use in model fitting (See Appendix B.1). Because there are not natural borders or edges shared between stations, it is natural that $\mathbf{V}_1$ and $\mathbf{V}_2$ would use an inverse distance-dependent proximity matrix which we denote $\mathbf{W}$. We assume the same distant-dependent CAR structure for both $\mathbf{V}_1$ and $\mathbf{V}_2$, where weights are proportional to $\exp(-ad)$ with d denoting the distance between locations and a being the inverse of the maximum distance between stations. For common proximity matrix $\mathbf{W}$, if we let D_W be diagonal with $(D_W)_{ii} = w_{i+}$, where $w_{i+} = \sum_i \mathbf{W}_{ij}$, then, the (unscaled) precision matrix of $\mathbf{V}_1$ and $\mathbf{V}_2$ is $Q = D_W - W$.

3.3.1 Priors, Model Fitting, and Prediction

We model regression and autoregressive coefficients $\boldsymbol{\beta}_{1i}$, $\boldsymbol{\beta}_{2i}$, $\boldsymbol{\gamma}_{1i}$, and $\boldsymbol{\gamma}_{2i}$ hierarchically,

$$\begin{array}{lll}
\boldsymbol{\beta}_{1i} \sim N(\boldsymbol{\beta}_{01}, \Sigma_{\beta_1}), & \boldsymbol{\gamma}_{1i} \sim N(\boldsymbol{\gamma}_{01}, \Sigma_{\gamma_1}), & \Sigma_{\beta_1} \sim IW(10^3\,\mathbf{I}, p+1), \\
\boldsymbol{\beta}_{2i} \sim N(\boldsymbol{\beta}_{02}, \Sigma_{\beta_2}), & \boldsymbol{\gamma}_{2i} \sim N(\boldsymbol{\gamma}_{02}, \Sigma_{\gamma_2}), & \Sigma_{\beta_2} \sim IW(10^3\,\mathbf{I}, p+1), \\
\boldsymbol{\beta}_{01} \sim N(\mathbf{0}, 10^3\,\mathbf{I}), & \boldsymbol{\gamma}_{01} \sim N(\mathbf{0}, 10^3\,\mathbf{I}), & \Sigma_{\gamma_1} \sim IW(10^3\,\mathbf{I}, n_{1l}+1), \\
\boldsymbol{\beta}_{02} \sim N(\mathbf{0}, 10^3\,\mathbf{I}), & \boldsymbol{\gamma}_{02} \sim N(\mathbf{0}, 10^3\,\mathbf{I}), & \Sigma_{\gamma_2} \sim IW(10^3\,\mathbf{I}, n_{2l}+1),
\end{array} \tag{3.2}$$

where $p = 3$ is the number of regressors including the intercept, n_{1l} is the number of lags for ozone, and n_{2l} is the number of lags for PM_{10}. By this, we assume that station-specific regression and autoregression coefficients are exchangeable. For the variance terms in the likelihood and the CAR prior, we assume that

$$\begin{array}{ll}
\sigma_1^2 \sim IG(1,1), & {a_{11}^{(\psi)}}^2 \sim IG(1,1), \\
\sigma_2^2 \sim IG(1,1), & {a_{22}^{(\psi)}}^2 \sim IG(1,1).
\end{array} \tag{3.3}$$

Lastly, we assume ${a_{12}^{(\psi)}}^2 \sim N(0, 10^3)$. Model fitting details via a Gibbs sampler are given in Appendices B.1 and B.2. This model could be fit sequentially, but this

would require us to update model parameters 8760 times for each step an MCMC sampler, once for each hour in 2017.

Prediction can be done in two ways, each with a different purpose: one predicts at unobserved values within the time range of the data (missing data) or one can predict future observations (forecasting). The former is viewed as *retrospective* prediction, filling in missing data over sites and times. The latter is viewed as prospective, predicting the next hour given the data up to the current hour. We are interested in MAAQS exceedances for specific months on an hourly scale. In this case, we only train model parameters on data observed prior to our predictions. Instead of a fully sequential model fitting where the model is updated hourly, we fit the model up to the last hour of the previous month. This model is then used to predict for pollutant levels for the upcoming month, making all prediction in this setting prospective. When carrying out inference for all days simultaneously, we fit the model to all the data once. For pollution and phase predictions, we limit our prediction to 10 AM, 3 PM, and 8 PM, each day to match the times of phase activation and suspension. Even though we make one-hour-ahead predictions, our predictions depend on model parameters that are trained using all the data; thus, our phase analysis is, in a sense, retrospective even though predictions are prospective. This model allows us to make probabilistic inference about reaching the conditions for environmental phase alerts in Mexico City and about Mexico City's compliance with MAAQS.

In the missing data context, we suppose that arbitrary Y_{it}^{O} or Y_{it}^{PM} is unobserved. This could be due to limited sampling or for model validation on a holdout dataset, but we only take this predictive approach when comparing models in Section 3.3.3. Each held-out observation is updated or imputed as a part of model fitting using a Gibbs sampler (See Appendix B.3 for details).

To predict future pollution measurements using our model, Equation 3.1, we use

the following formula:

$$Y^O_{i(t+1)} = \mathbf{x}^T_{it}\boldsymbol{\beta}_{1i} + {\mathbf{L}^O_{i(t+1)}}^T\boldsymbol{\gamma}_{1i} + \psi_{1i} + \epsilon_{1i(t+1)} \tag{3.4}$$

$$Y^{PM}_{i(t+1)} = \mathbf{x}^T_{it}\boldsymbol{\beta}_{2i} + {\mathbf{L}^{PM}_{i(t+1)}}^T\boldsymbol{\gamma}_{2i} + \psi_{2i} + \epsilon_{2i(t+1)}.$$

Note that one-step-ahead predictions do not rely on future covariates. We use this type of prediction for both inferential tasks (See Sections 3.4.1 and 3.4.2).

In our setting, predicted phase alerts come from the one-hour-ahead ozone predictions ($\hat{Y}^O_{it}$) and the predicted 24-hour average PM_{10} concentration ($\widehat{\overline{Y}^{PM}_{it}}$), where $\widehat{\overline{Y}^{PM}_{it}}$ is the average of the 23 most recent observed PM_{10} concentrations ($Y^{PM}_{i(t-1)}$, ..., $Y^{PM}_{i(t-23)}$) and the forecasted PM_{10} level ($\hat{Y}^{PM}_{it}$). Nationally legislated thresholds depend on 24-hour average PM_{10} and on 8-hour average O_3. Similar to $\widehat{\overline{Y}^{PM}_{it}}$, predicted 8-hour average ozone concentration ($\widehat{\overline{Y}^{O}_{it}}$) is an average of a one-hour-ahead prediction and the previous seven ozone measurements ($Y^O_{i(t-1)}$, ..., $Y^O_{i(t-7)}$). For predictions of both $\widehat{\overline{Y}^{PM}_{it}}$ and $\widehat{\overline{Y}^{O}_{it}}$ on January 1, 2017, we rely on hourly observations from December 31, 2016.

3.3.2 Posterior Inference

The primary inferential goal for this dataset is to assess how often the Mexico City metropolitan area (1) is at risk for declaring phase I or II emergencies and (2) exceeds MAAQS. For each task, we take different modeling approaches, as discussed in Section 3.3.1. To analyze the risk of phase I and II emergencies, we fit the model to all the data. In contrast, when examining pollution level exceedances, we fit the model sequentially. For both tasks, we use one-step-ahead predictions for ozone and PM_{10} concentrations. These posterior predictions allow us to carry out probabilistic inference on emergency phases and MAAQS exceedances to assess how often the Mexico City metropolitan area was at risk of a pollution emergency and how often

pollution levels were unsafe according to Mexican federal guidelines (Diario Oficial de la Federación, 2014a,b).

To define useful quantities, let j index region and d index day, such that each station $i \in j$ and each hour $t \in d$. Additionally, we define $\overline{Y_{it}}^{PM}$ to be the 24-hour running average of PM_{10} at time t and station i. We define the following maxima:

$$\begin{aligned} Z_{jt}^{O} &= \max_{i \in j} Y_{it}^{O} & W_{jd}^{O} &= \max_{t \in d} \max_{i \in j} Y_{it}^{O} \\ Z_{jt}^{PM} &= \max_{i \in j} \overline{Y}_{it}^{PM} & W_{jd}^{PM} &= \max_{t \in d} \max_{i \in j} \overline{Y}_{it}^{PM}, \end{aligned} \tag{3.5}$$

where Z's are regional maxima for any hour t and W's are daily regional maxima. It is important to clarify that although these maxima often rely on data observed prior to time t or day d, these quantities are used to define exceedances at time t or day d. There is limited literature about the distributions and properties of maxima for correlated random variables (see Gupta et al., 1985; Ho and Hsing, 1996). However, these examples are too constrained for our application. In the spatial literature, modeling extreme values, and sometimes maxima, is well-studied (see, e.g., Sang and Gelfand, 2009, 2010; Davison et al., 2012); however, these approaches generally invoke generalized extreme value (GEV) distribution models. As noted in Section 2, our inference depends on relatively few maxima over few sites or hours, so GEV theory is not applicable. In fact, using the definitions in (3.5), we do not model the maxima directly. Instead, we obtain the derived posterior predictive distribution for Z_{jt}^{O}, Z_{jt}^{PM}, W_{jd}^{O}, and W_{jd}^{PM} from posterior predictive samples of Y_{it}^{O} and Y_{it}^{PM}.

The states of Mexico City's phase alert system $S_{jt} \in \{0, 1, 2\}$ are completely determined by Z_{jt}^{O} and Z_{jt}^{PM} (see Section 3.1 and Table 3.1). To obtain the maximum phase alert for a day d in some region j ($\max_{t \in d} S_{jt}$), we use W_{jd}^{O} and W_{jd}^{PM}. One may also wish to infer the distribution of the highest phase alert in any region on day d ($\max_{j} \max_{t \in d} S_{jt}$). All these derived posterior quantities can be obtained after model

fitting. Using derived posterior predictive distributions for various maxima, as well as associated phase states and threshold exceedances, we can compute hourly and daily probabilities of (possible) phase alerts and pollution exceedances regionally and city-wide. The utility of these probabilities is the insight they can provide regarding how often the Mexico City metropolitan area is at risk of a pollution emergency, even if phase alerts were not enacted due to meteorological forecasts.

Inference for the first task, analysis of the phase emergencies, requires analysis of regional pollution levels at 10 AM, 3 PM, and 8 PM. Specifically, phase states depend on maxima of stations over regions. If we carry out inference on a daily scale, double maxima are needed, maxima over hours and stations within regions. Because phase alerts are based upon one-hour ozone measurements and 24-hour average PM_{10} (Administración Pública de la Ciudad de México, 2016), we predict these averages as described in Section 3.3.1. These predictions allow us to compute predictions for derived quantities $\widehat{Z_{jt}^{O}}$, $\widehat{Z_{jt}^{PM}}$, $\widehat{W_{jd}^{O}}$, and $\widehat{W_{jd}^{PM}}$, which in turn define phase state predictions $\widehat{S_{jt}}$. Again, inference on phase predictions is our primary goal.

We also carry out similar inference on MAAQS exceedance for ozone and PM_{10}. Again, we are interested in regional (i.e. stations within a specified region) and city-level (i.e. at any station in the city) exceedance, hourly and daily. Similar, but not identical to phase alerts, inference for pollution exceedances relies upon maxima of one-hour ozone, eight-hour average ozone $\overline{Y_{it}}^{O}$, and 24-hour average PM_{10}. For eight-hour average ozone, we define $Z_{jt}^{\overline{O}}$ to be the regional maxima at time t and $W_{jd}^{\overline{O}}$ to be the daily maxima for region j. In this case, thresholds are much lower than the thresholds Atmospheric Environmental Contingency Program in Mexico City (Diario Oficial de la Federación, 2014a,b; Administración Pública de la Ciudad de México, 2016). Because nationally legislated ozone and PM_{10} thresholds were specified to avoid reaching unsafe pollution levels, comparisons to MAAQS indicate how often

Mexico City reaches unsafe pollution levels without triggering any city protocols. Such comparisons highlight important differences in how pollution emergencies are defined in Mexico City relative to nationally legislated levels.

3.3.3 Model Selection

In this subsection, we describe the model selection process leading to the model under which we carry out all the inference described in the previous subsections. Our model selection decision centers around answering how many and which lagged terms should be used in our spatiotemporal model. In our exploratory analyses, we argued that the variance of ozone and PM_{10} vary with the time of day and time of year. We indicated that this could be remedied in one of two ways: (1) using variance-stabilizing transformations to address correlation between mean and variance in the data or (2) modeling the heteroscedasticity directly. For transformation approaches, we consider modeling the data on different scales (truncated, log, and square root) to stabilize the mean-variance correlation. To answer these modeling questions, we hold out 10% of both pollutants and treat these as missing data. Specifically, the locations and times of the hold-out data are selected at random, and both ozone and PM_{10} are held-out at these location-time pairs so that model comparison can be made using joint predictions. We make predictions at these held-out observations and compare competing models based on several criteria: predictive mean squared error $(E(Y_i|\mathbf{Y}_{obs}) - y_i)^2$ (PMSE) or mean absolute error $|E(Y_i|\mathbf{Y}_{obs}) - y_i|$ (PMAE), $100 \times \alpha$ % prediction interval coverage, and continuous rank probability scores (CRPS) (Gneiting and Raftery, 2007), where

$$\text{CRPS}(F_i, y_i) = \int_{-\infty}^{\infty} (F_i(x) - \mathbf{1}(x \geq y_i))^2 dx = \mathbf{E}|Y_i - y_i| - \frac{1}{2}\mathbf{E}|Y_i - Y_{i'}|. \qquad (3.6)$$

Because we are utilizing MCMC to fit our model, we use posterior predictive samples for a Monte Carlo approximation of CRPS using an empirical CDF approximation

(see, e.g., Krüger et al., 2016),

$$\text{CRPS}(\hat{F}_i^{\text{ECDF}}, y_i) = \frac{1}{M}\sum_{j=1}^{M}|Y_j - y_i| - \frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}|Y_j - Y_k|, \tag{3.7}$$

where M is the number of MCMC samples used, Y_i are predictions, and y_i are observed values. We then average $\text{CRPS}(\hat{F}_i^{\text{ECDF}}, y_i)$ over all held-out data. In addition to being a proper scoring rule (Gneiting and Raftery, 2007), because CRPS considers how well the entire predictive distribution matches the observed data rather than only the predictive mean (MAE and MSE) or quantiles (prediction interval coverage), we prefer it as selection criterion. For multivariate predictions, as we have in this analysis, we consider the energy score (ES), which is a multivariate generalization of CRPS. For a set of multivariate predictions $\mathbf{Y}$, ES is defined as

$$\text{ES}(P, \mathbf{y}) = \frac{1}{2}E_P\left\|\mathbf{Y} - \mathbf{Y}'\right\|^{\beta} - E_P\left\|\mathbf{Y} - \mathbf{y}\right\|^{\beta}, \tag{3.8}$$

where $\mathbf{y}$ is an observation, $\beta \in (0, 2)$, and P is a probability measure (Gneiting and Raftery, 2007). It is common to fix $\beta = 1$ (see, e.g., Gneiting et al., 2008; Jordan et al., 2017). For a set of M MCMC predictions $\mathbf{Y} = \mathbf{Y}_1, ..., \mathbf{Y}_M$ for a held-out observation $\mathbf{y}$, the empirical ES reduces to

$$\text{ES}(\mathbf{Y}, \mathbf{y}) = \frac{1}{M}\sum_{j=1}^{M}\left\|\mathbf{Y}_j - \mathbf{y}\right\| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left\|\mathbf{Y}_i - \mathbf{Y}_j\right\|, \tag{3.9}$$

as was discussed in Gneiting et al. (2008). Energy scores are scale-sensitive, meaning that if one of the variables has a much larger scale than other of the variables, it dominates the norms in Equation 3.9. In our data, PM_{10} concentration in $\mu g/m^3$ takes values larger than ozone in ppb. To assure that predictions for each pollutant are similarly weighted, we standardize the predictions and hold-out values for each pollutant (i.e. subtract the sample mean and divide by the sample standard deviation). Like CRPS, we average ES over all held-out data.

Interestingly, the heteroscedastic models with variance that varies over hour of the day and month of the year performed uniformly worse than homoscedastic counterparts that used VST's to stabilize the mean-variance correlation. Testing various combinations of square-root transformations, log transformations, and truncated distributions, we found that models using the square-root transformation for ozone and the log transformation for PM_{10} gave the best predictive performance. So, for model selection, we only give the results for six models which use the square-root transformation for ozone and the log-transformation for PM_{10} but differ in terms of which lags are included in the model. The results of this comparison are given in Table 3.3.

Table 3.3: Predictive model comparison. The "Lags" label indicates which lags are used for both outcomes. "ES," "CRPS," "MSE," "MAE," and "Cov" head columns giving ES, CRPS, MSE, MAE, and 90% prediction interval coverage. Best performances are indicated with bold text.

Lags	ES	O_3 CRPS	O_3 RMSE	O_3 MAE	O_3 Cov	PM_{10} CRPS	PM_{10} RMSE	PM_{10} MAE	PM_{10} Cov
(1,2)	0.2552	2.5392	5.0448	3.3409	0.8867	6.7263	14.5427	8.7725	0.9228
(1,2,24)	0.2513	2.5158	4.9709	3.3035	0.8925	6.6176	14.1469	8.6189	**0.9217**
(1,2,24,168)	**0.2505**	**2.5140**	**4.9614**	**3.2982**	**0.8941**	**6.5947**	**14.0922**	**8.6285**	0.9229
(1,2,12)	0.2540	2.5298	5.0274	3.3244	0.8887	6.6959	14.4314	8.7864	0.9230
(1,2,12,24)	0.2509	2.5168	4.9726	3.3115	0.8917	6.6035	14.0981	8.6296	0.9220
(1,2,12,24,168)	0.2507	2.5154	4.9651	3.2993	**0.8941**	6.5976	14.0972	8.6378	0.9225

We further note that in preliminary modeling, we found that models which included a lag-3 and other higher order lags or that excluded lag-2 saw no improvement in terms of prediction; thus, we arrived at the models included in Table 3.3. Given these results, we argue that the best model for ozone and PM_{10} uses lags 1, 2, 24, and 168. So, the ensuing results are presented for this model.

3.4 Results and Discussion

We present our inference based on a joint model for ozone and PM_{10} with four lags (1, 2, 24, and 168). We use a Gibbs sampler to obtain 100,000 posterior samples after a burn-in of 10,000 iterations. Posterior parameter inference validates many of the

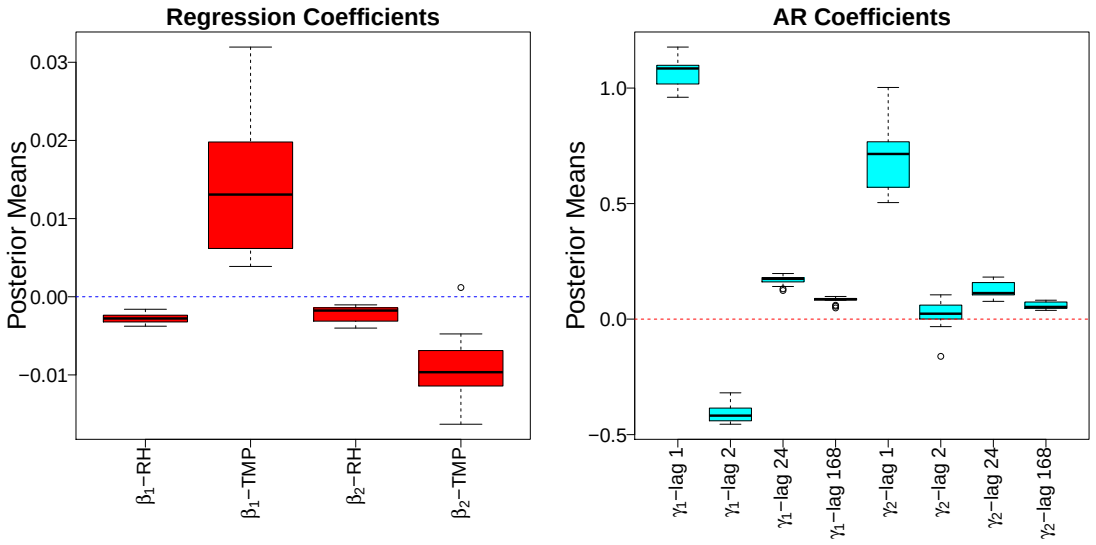
modeling decisions suggested by our exploratory analysis in Section 3.2 and discussed in Section 3.3. Posterior summaries for covariance parameters (σ_1^2, σ_2^2, $a_{11}^{(\psi)}$, $a_{12}^{(\psi)}$, and $a_{22}^{(\psi)}$) and the overall means for hierarchical coefficient parameters ($\boldsymbol{\beta}_{01}$,$\boldsymbol{\beta}_{02}$,$\boldsymbol{\gamma}_{01}$,$\boldsymbol{\gamma}_{02}$) are given in Table 3.4.

Because $a_{12}^{(\psi)}$ is significantly positive, this indicates that regions with higher spatial random effects for ozone generally have higher random effects for PM_{10}, as suggested in our data exploration. The inference given by $\boldsymbol{\beta}_{01}$ confirms that ozone is negatively related to RH and positively related to TMP, again as we noted in our exploratory analyses. For PM_{10}, we see a negative relationship with RH and TMP via $\boldsymbol{\beta}_{02}$, while the only relationship that was evident in our exploration was the negative relationship with RH. The autoregressive terms for PM_{10} are positive, and the lag-1 and lag-24 terms are largest. For ozone, the autoregressive terms for the lags 1, 24, and 168 are positive, but the lag-2 coefficient is negative which tempers the effect of the lag-1 coefficient. While $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\beta}_{02}$ represent the average relationships between covariates and ozone and PM_{10}, each site has unique covariate effects. We provide box plots for the posterior means of site-specific regression and AR coefficients in Figure 3.7 (each box displays the 24 site-specific posterior means for that coefficient). In general, the site-specific coefficients are in the same direction as the overall effect, as we would expect. Interestingly, the effect of temperature varies significantly between locations. The site-specific AR coefficients generally are tightly clustered except the lag 1 coefficient for PM_{10}.

Table 3.4: Posterior summaries for covariance parameters and overall or common means for the hierarchical regression and autoregression coefficients. $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\gamma}_{01}$ are interpreted with respect to the square-root ozone scale. $\boldsymbol{\beta}_{02}$ and $\boldsymbol{\gamma}_{02}$ are interpreted as effects on PM_{10} on the log-scale.

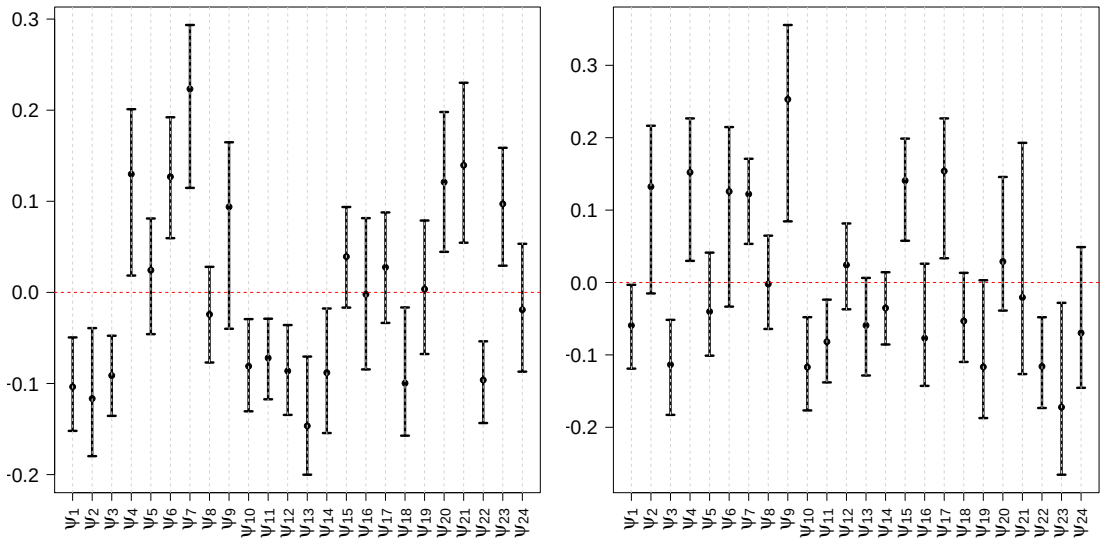
	Mean	Std Dev	2.5%	97.5%
σ_1^2	0.5180	0.0016	0.5148	0.5211
σ_2^2	0.1490	0.0005	0.1481	0.1499
$a_{11}^{(\psi)}$	0.5102	0.0857	0.3633	0.6988
$a_{22}^{(\psi)}$	0.5203	0.1093	0.3176	0.7438
$a_{12}^{(\psi)}$	0.4723	0.2622	0.0224	1.0090
$\boldsymbol{\beta}_{01}$ (Intercept)	0.4503	0.0087	0.4330	0.4668
$\boldsymbol{\beta}_{01}$ (RH)	-0.0027	0.0002	-0.0031	-0.0024
$\boldsymbol{\beta}_{01}$ (TMP)	0.0138	0.0019	0.0100	0.0176
$\boldsymbol{\beta}_{02}$ (Intercept)	0.3573	0.0124	0.3300	0.3836
$\boldsymbol{\beta}_{02}$ (RH)	-0.0022	0.0002	-0.0026	-0.0018
$\boldsymbol{\beta}_{02}$ (TMP)	-0.0093	0.0009	-0.0110	-0.0076
$\boldsymbol{\gamma}_{01}$ (lag 1)	1.0649	0.0113	1.0432	1.0878
$\boldsymbol{\gamma}_{01}$ (lag 2)	-0.4079	0.0080	-0.4238	-0.3921
$\boldsymbol{\gamma}_{01}$ (lag 24)	0.1683	0.0041	0.1603	0.1763
$\boldsymbol{\gamma}_{01}$ (lag 168)	0.0835	0.0028	0.0780	0.0889
$\boldsymbol{\gamma}_{02}$ (lag 1)	0.6952	0.0250	0.6446	0.7442
$\boldsymbol{\gamma}_{02}$ (lag 2)	0.0227	0.0112	0.0011	0.0448
$\boldsymbol{\gamma}_{02}$ (lag 24)	0.1261	0.0065	0.1130	0.1391
$\boldsymbol{\gamma}_{02}$ (lag 168)	0.0572	0.0033	0.0507	0.0638

We plot the posterior means and credible intervals for ψ_{1i} and ψ_{2i} in Figure 3.8. For most sites, the 95% credible intervals for ozone's spatial random effects exclude 0. By contrast, the credible intervals for PM_{10}'s spatial random effects include 0 for 10 of the 24 stations.



(a) Regression coefficients (b) AR coefficients

FIGURE 3.7: Posterior means for site-specific regression and AR coefficients for ozone and PM_{10}. Each box displays the 24 site-specific posterior means for that coefficient.



(a) CAR Effects for Ozone (b) CAR Effects for PM_{10}

FIGURE 3.8: Posterior means and credible intervals for site-specific CAR random effects for ozone and PM_{10}.

Because we have $N = 210240$ observations, the predictive space is large ($2 \times N \approx$

4×10^5). Thus, we thin the posterior predictive samples using every 10^{th} sample. By thinning, we make 10,000 roughly independent predictions. These predictions are used to carry out analyses in Sections 3.4.1 and 3.4.2.

3.4.1 Analysis of the Phase Alert System

In this section, we analyze Mexico City's phase alert system to identify when the Mexico City metropolitan area was predicted to be at risk for pollution emergencies. For this, we use one-hour-ahead predictions for pollution levels each day at the three decision times reference in Section 3.1: 10 AM, 3 PM, and 8 PM. Thus, our analysis predicts at three hours per day, altogether 1095 hours in 2017. This allows us to assess probabilities of the risk of phase alerts given the most recent weather conditions and pollution levels. Again, we note that the risk of a phase alert is not the same as a phase alert. As discussed above, we use parameter values trained on the entire dataset which enables effective prediction in early months. Because the pollutant thresholds for triggering phase alerts are very high, most of the year has very low probabilities for phase activation. In May of 2017, however, Mexico City was featured prominently in the news for having dangerously high ozone levels which led to an activation of a phase I pollution emergency. Phase probabilities aggregated over regions ($P(\max_j S_{jd} = k)$ for state k) are displayed in Figure 3.9. Regional phase I probabilities for each day ($P(S_{jd} = k)$ for phase k) are given in Figure 3.10. In Figure 3.10, we do not show phase II probabilities because they are so low. Additionally, we only display region NE compared to other regions because all other regions overlap (See Figure 3.10). Both plots (Figures 3.9 and 3.10) show high probabilities ($> 1/2$) of phase I activation from May 16th to May 25, coinciding with the time of the actually declared phase I emergency. Because this phase I alert was triggered by ozone levels, the emergency was declared city-wide, as indicated by the agreement of regional curves in Figure 3.10.

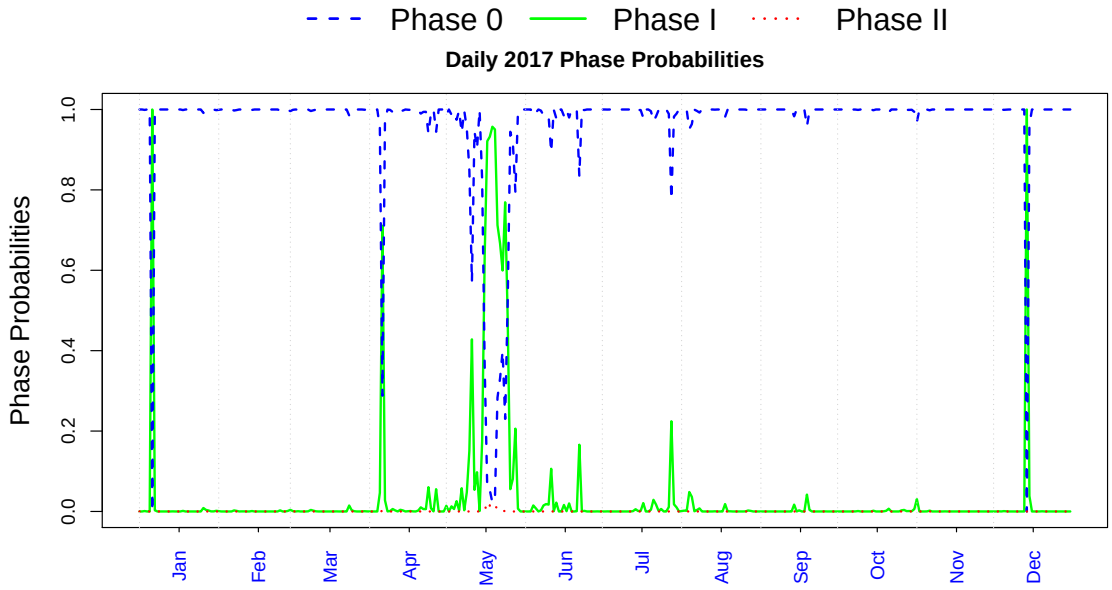


FIGURE 3.9: Phase probabilities in Mexico City, aggregated over all regions

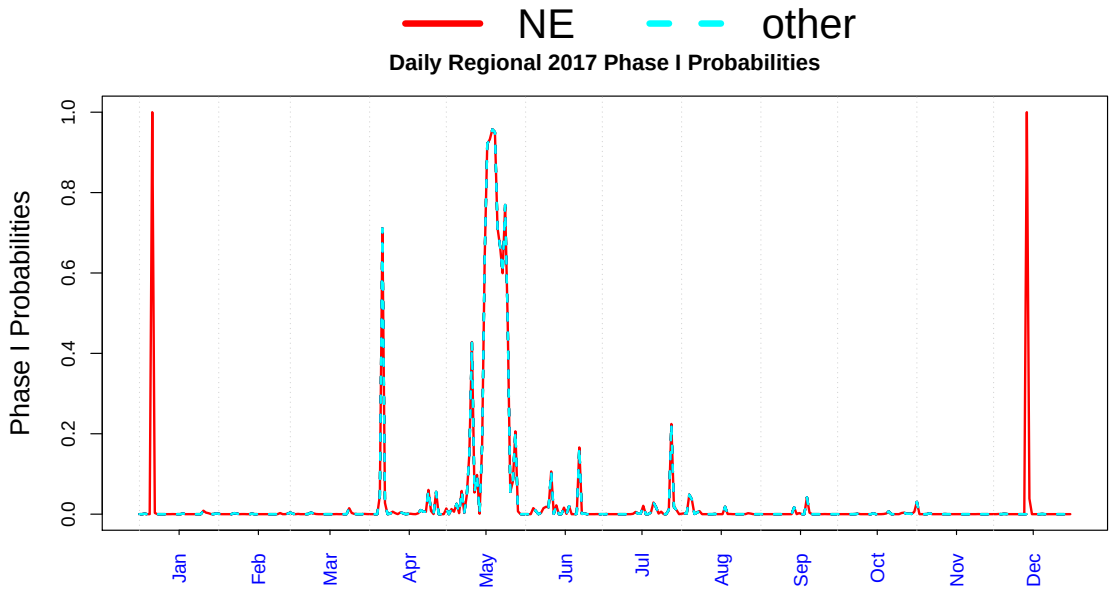


FIGURE 3.10: Daily phase I probabilities for Mexico City over the year by region. Phase II probabilities are not included because they are uniformly low.

On April 6th, predicted ozone levels were sufficient to trigger a phase I emergency

city-wide, although a phase alert was not declared. On only two occasions, one in January and one in December, was any region at risk of activating the emergency contingency plan due to PM_{10} levels. These high phase I probabilities were limited to the northeast region (See the red peaks in Figure 3.10). Again, it is worth noting that a phase I alert triggered by PM_{10} corresponds to PM_{10} levels that are nearly three times the levels specified as safe by Mexican legislation (Diario Oficial de la Federación, 2014b).

In Table 3.5, we provide posterior means and 95% credible intervals for the number of hours and days for which the Mexico City metropolitan area is at risk for pollution emergencies ($\sum_d \mathbf{1}(S_{jd} = k)$ for phase k). The first thing to notice is that there are very few hours and days when the metropolitan area or its sub-regions are at risk of pollution emergencies. Note that the posterior predictive mean for risk of a pollution emergency is 11 days and 11 hours for the central, northwest, southeast, and southwest regions. These counts are not necessarily reflective of conditions in central and northwest regions. Instead, these counts are indicative of predicted city-wide phase I alerts due to predicted ozone exceedances in the southeast and southwest regions, one in April and 10 in May (see Figures 3.9 and 3.10). The northeast region is the only region that had more average predicted hours and days of pollution emergency than other regions. We predict six hours of risk for phase I emergencies in the northeast region due to PM_{10} levels over two non-consecutive days, one day in January and one in December. Because the northeast region was the only region where a predicted phase alert was triggered by PM_{10}, the predicted risk of a phase alert was limited to the northeast region. No phase alert was declared even though predicted phase probabilities were equal to one. Thus, the reason for not declaring an emergency must be attributed to meteorological conditions. While we do know the exact rationale for not declaring a phase emergency, we speculate that the emergency was not declared because these predicted phase risks were transient,

lasting only one day each.

Table 3.5: One-hour-ahead posterior predictive estimates for the (Top) Number of hours in each phase state for each region (Bottom) Number of days for which that phase state was attained (the maxima attained each day). Posterior means and 95% credible intervals are given for each region, using ± or parentheses. The "Any" label indicates that this is the maximum across regions.

	Hours (total of 1095 —— 3 hours / day)					
	CE	NE	NW	SE	SW	Any
No Phase	1084 ± 4	1078 ± 4	1084 ± 4	1084 ± 4	1084 ± 4	1078 ± 4
Phase I	11 ± 4	17 ± 4	11 ± 4	11 ± 4	11 ± 4	17 ± 4
Phase II	0.09 (0,1)	0.09 (0,1)	0.09 (0,1)	0.09 (0,1)	0.09 (0,1)	0.09 (0,1)
	Days (total of 365)					
No Phase	354 ± 4	352 ± 4	354 ± 4	354 ± 4	354 ± 4	352 ± 4
Phase I	11 ± 4	13 ± 4	11 ± 4	11 ± 4	11 ± 4	13 ± 4
Phase II	0.09 (0,1)	0.09 (0,1)	0.09 (0,1)	0.09 (0,1)	0.09 (0,1)	0.09 (0,1)

3.4.2 Comparison of Mexico City to Mexican Legislated Thresholds

In this section, we examine the probability that maxima within regions exceed MAAQS on a given day (W_{jd}^{O} and W_{jd}^{PM} from Section 3.3.2). In contrast to phase alert probabilities, which are generally very low, exceedance probabilities are often high through much of the year. Because MAAQS are more reflective of healthy levels of ozone and PM_{10}, comparison between the exceedance probabilities and emergency phase probabilities highlights how often Mexico City has harmful pollution levels without triggering phase alerts. Additionally, this analysis gives insight into the probability of triggering phase alerts in Mexico City if MAAQS were adopted for Mexico City's Atmospheric Environmental Contingency Program. For our purposes, we group either type of ozone exceedance, one or eight-hour, together. We focus on three months, April, August, and December, to illustrate how exceedance probabilities change over the course of the year. April and August are warm months, and ozone creation needs heat. August is the wettest month of the year in Mexico City, on average. Rainfall tends to clear out PM, so PM_{10} levels are expected to be low in August. April, on the other hand, precedes the rainy season and is normally dry.

December is cold and dry. These months naturally contrast each other by illustrating how yearly climate affects pollution levels and the probability of exceeding Mexican ambient air quality standards. We plot regional daily exceedance probabilities for ozone ($P(W_{jd}^{O} > 95 \text{ ppb} \cup W_{jd}^{\overline{O}} > 70 \text{ ppb})$) and PM_{10} ($P(W_{jd}^{PM} > 75\ \mu g/m^3)$) over April (Figure 3.11), August (Figure 3.12), and December (Figure 3.13).

Recall that April had low phase probabilities except for April 6th when phase I probabilities spiked in all regions due to high ozone levels. Unsurprisingly, the daily exceedance probabilities for ozone are high for all regions in April. The northern regions (NE and NW) have the lowest ozone exceedance probabilities but are still above 1/2 most of the month. Daily PM_{10} exceedance probabilities vary over the month, with probabilities near one before the 13th, zero from the 13th of April to April 19th, and again high toward the end of the month.

In our phase analysis, we showed that August had uniformly low probabilities of predicted pollution emergencies. Because August is in the rainy season, we expected that it would have low PM_{10}. This is confirmed by our analysis with August having low daily PM_{10} exceedance probabilities, with the exception of a three days (8/11, 8/15, 8/16). Daily ozone exceedance, on the other hand, is high over most of the month for three regions (CE, SE, SW). Like in April, the northern regions have lower probability of ozone exceedance.

In December, we showed that phase probabilities were predicted to be low with the exception of a single peak in phase I probabilities in the northeast region due to high PM_{10} concentrations. Because the phase I PM_{10} threshold is nearly three times Mexican ambient air quality standards, it is unsurprising to observe high exceedance probabilities for PM_{10} in the northeast region. It is, however, interesting that four of the five regions have predicted daily exceedance probabilities of one (or very close to one) for 20 or more days. Ozone has many periods of low exceedance probabilities but does exhibit high daily exceedance probabilities overall.

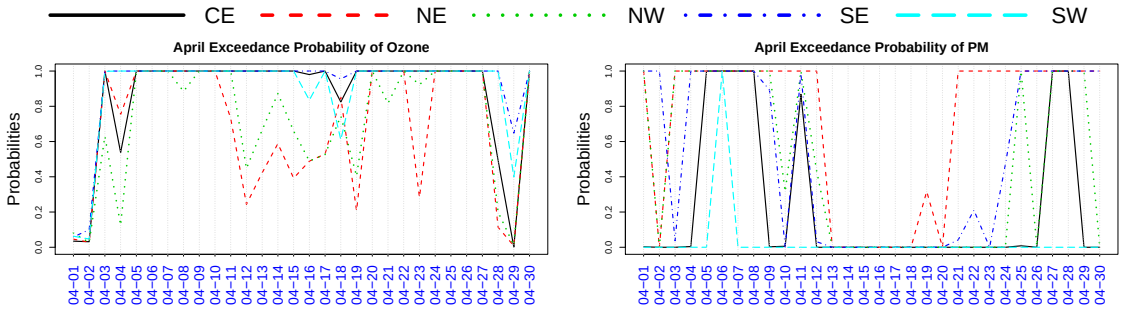


FIGURE 3.11: Exceedance probabilities of (Left) ozone and (Right) PM_{10} for April for each region on a daily level. The colors indicate regions: black represents CE, NE is red, NW is green, SE in blue, and SW is cyan.

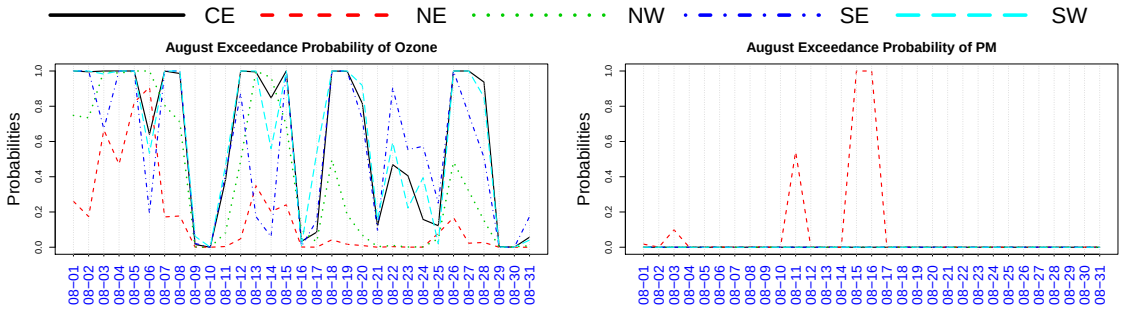


FIGURE 3.12: Exceedance probabilities of (Left) ozone and (Right) PM_{10} for August for each region on a daily level. The colors indicate regions: black represents CE, NE is red, NW is green, SE in blue, and SW is cyan.

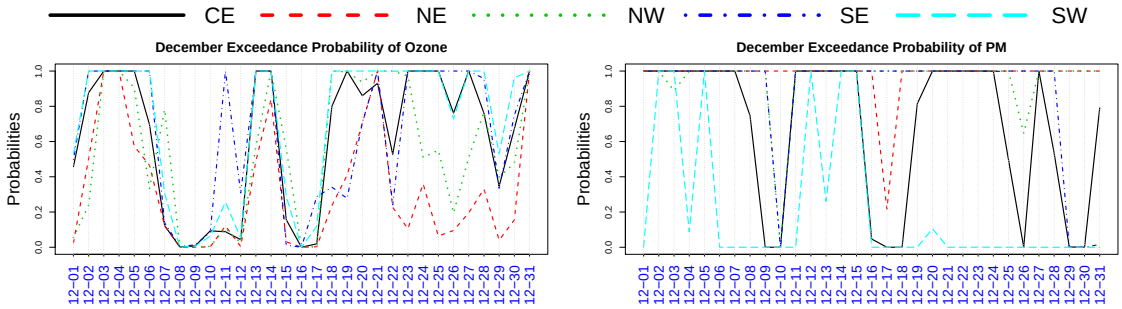


FIGURE 3.13: Exceedance probabilities of (Left) ozone and (Right) PM_{10} for December for each region on a daily level. The colors indicate regions: black represents CE, NE is red, NW is green, SE in blue, and SW is cyan.

We continue the prospective analysis for all months except January, fitting the model up until the last hour of the previous month to predict pollution exceedance for the month of interest. Because we fit the model sequentially, prospective predictions for January are poor because the model has not been trained on data for these times. Using these predictions, we give posterior means and 95% credible intervals for the one-hour-ahead predicted proportion of hours and days of exceedance for each region (i.e. $P(Z_{jt}^{O} > 95 \text{ ppb} \cup Z_{jt}^{\overline{O}} > 70 \text{ ppb})$, $P(Z_{jt}^{PM} > 75\ \mu g/m^3)$, $P(W_{jd}^{O} > 95 \text{ ppb} \cup W_{jd}^{\overline{O}} > 70 \text{ ppb})$, and $P(W_{jd}^{PM} > 75\ \mu g/m^3)$, as defined in Section 3.3.2). The results for ozone are given in Table 3.6, and the estimates for PM_{10} are presented in Table 3.7. For ozone, the proportion of exceedances in both hours and days decreases as latitude increases, with northern regions showing nearly half as many exceedances as the southern regions, on average. The trend for PM_{10} is less clear, although there is significant variability across regions. The northeast region has many more predicted PM_{10} exceedances than any other region. This is due to the large industrial economy located within this region. By contrast, the southwest region has, comparatively, very few PM_{10} exceedances.

Table 3.6: One-hour-ahead posterior predictive estimates for the (Left) Proportion of hours where either of the Mexican legislated ozone limits (one-hour or eight-hour) were exceeded (Right) Proportion of days where either of the Mexican legislated ozone limits were exceeded. Posterior means and 95% credible intervals are given for each region. The "Any" label indicates that at least one region has an exceedance for one or more location for the time level (hour or day).

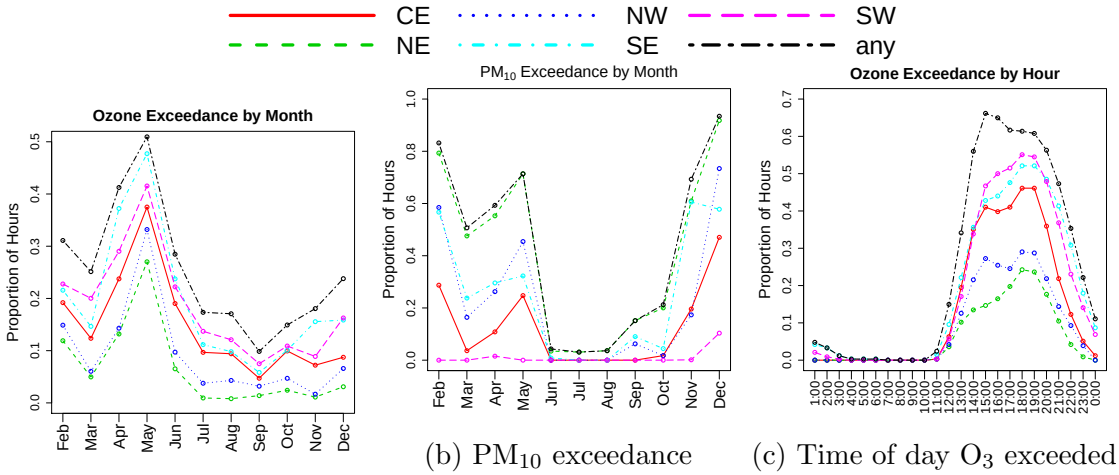
Ozone	Hours (total of 8016)						Days (total of 334)					
	CE	NE	NW	SE	SW	Any	CE	NE	NW	SE	SW	Any
Mean	0.147	0.066	0.093	0.194	0.186	0.252	0.651	0.367	0.499	0.671	0.696	0.794
2.5%	0.143	0.064	0.090	0.190	0.183	0.249	0.626	0.338	0.467	0.647	0.674	0.773
97.5%	0.150	0.069	0.096	0.197	0.189	0.256	0.677	0.395	0.530	0.698	0.719	0.814

Table 3.7: One-hour-ahead posterior predictive estimates for the (Left) Proportion of hours where the Mexican legislated 24-hour PM_{10} limits were exceeded (Right) Proportion of days where either of the Mexican legislated PM_{10} limits were exceeded. The "Any" label indicates that at least one region has an exceedance for one or more location for the time level (hour or day).

PM_{10}	Hours (total of 8016)						Days (total of 334)					
	CE	NE	NW	SE	SW	Any	CE	NE	NW	SE	SW	Any
Mean	0.123	0.408	0.221	0.247	0.0112	0.429	0.218	0.5223	0.333	0.370	0.026	0.535
2.5%	0.122	0.406	0.219	0.245	0.0106	0.427	0.210	0.512	0.323	0.362	0.021	0.524
97.5%	0.125	0.410	0.223	0.249	0.0119	0.430	0.228	0.533	0.341	0.377	0.030	0.545

Lastly, we discuss the proportion of predicted hourly exceedances as a function of the month of the year and of the hour of the day. The summaries for ozone by month and hour-of-day are plotted in Figure 3.14, while we display PM_{10} exceedances only by month (Figure 3.14b). We do not plot PM_{10} exceedances as function of the hour of the day because PM_{10} exceedances depend on 24-hour averages; thus, trends over time-of-day are not meaningful. As a function of month, the patterns of ozone and PM_{10} exceedances are clear. For ozone, the proportion of exceedances reaches a peak in May and is high in March, April, and June. We attribute these high ozone levels to warm times of the year that are dry compared to the rainy season (June-August). PM_{10} exceedance appears to co-vary strongly with the rainy season as well, which is captured by relative humidity in our model. In particular, June, July, August, and September have almost no exceedances for PM_{10}. The coldest months (December and February) have higher probabilities of PM_{10} exceedance than warmer months that are similarly dry. Mexico City's pollution output is higher during winter festivities like

Our Lady of Guadalupe, Christmas, and New Year due to fireworks and increased motor traffic. In conjunction with increased pollution output, pollution exceedances in cold months are also due to thermal inversion that traps pollution in the Valley of Mexico where Mexico City lies. Ozone exceedances also tend to peak in the afternoon to evening. Because ozone levels can exceed thresholds for either one-hour or eight-hour average ozone, we expect two peaks in ozone exceedance as a function of hour. The one-hour peak occurs around 4 PM (16:00) when the temperature is highest. The peak of eight-hour average ozone peaks around 7 or 8 PM (19:00 or 20:00), after eight hours of relatively high ozone levels. These peaks can be seen in Figure 3.14c.



(a) O_3 exceedance (b) PM_{10} exceedance (c) Time of day O_3 exceeded

FIGURE 3.14: Posterior predictive means for the proportion of hours of exceedance as function of month (Left and Center) and hour of the day (Right).

3.5 Conclusions and Future Work

We have discussed the monitoring network for ozone and PM_{10} within Mexico City and proposed a joint spatiotemporal model for ozone and PM_{10} concentrations. This model was used to predict future pollutant concentrations. Our predictions were then used to obtain derived distributions for regional maxima of ozone and PM_{10} which

are needed to determine Mexico City's pollution emergency phases. Additionally, our predictions are used to assess compliance with MAAQS. We find that predicted risk of pollution emergency is rare and are predicted for only a few periods of 2017. By contrast, we demonstrate that predicted exceedance of Mexico's ambient air quality standards is common.

In future work, we will attempt to operationalize our model so that it can be used in practice. This would require real time (hourly) fitting of the model as new measurements are available. Our modeling is amenable to sequential updating as well as possible parallelization though considerable optimization remains before this could be implemented in practice. Once implemented, our model could warn of potential pollution emergencies or compliance issues, allowing regional and city-wide adjustments, warnings, responses, and decision-making to be made earlier. Our model could incorporate weather forecasts to perhaps more accurately forecast pollutant levels farther ahead than our one-hour-ahead predictions.

4

Nonseparable Covariance Models on Circles Cross Time: A Study of Mexico City Ozone

4.1 Introduction

Ground-level ozone is linked with short- and long-term health risks in general (see, e.g., Bell et al., 2006) and in Mexico City specifically (see Riojas-Rodríguez et al., 2014). In statistics, many have studied daily ozone levels (e.g., Sahu et al., 2007; Berrocal et al., 2010; Huang et al., 2018). Sahu et al. (2007) utilize a dynamic spatiotemporal model for square-root ozone concentrations in Ohio to assess pollution trends over time. Berrocal et al. (2010) propose a bivariate spatiotemporal model for square-root ozone levels and the log concentrations of fine particulate matter that combines monitoring data with output from numerical models. Huang et al. (2018) use a multivariate space-time model for daily pollutant levels to estimate health risks and associated uncertainty in pollution exposure. Like Chiogna and Pauli (2011) and Arisido (2016), we argue for using finer than daily time scales and treating time as continuous to quantify short-term health risks because short-term spikes in ozone levels increase respiratory health risks.

The Mexico City ozone monitoring data presented here consist of hourly measurements from 24 stations in April and May of 2017, Mexico City's peak ozone season (SEDEMA, 2017). The data that support the findings of this study are available under Datos/Horarios/Contaminante at `http://www.aire.cdmx.gob.mx/default.php`. In total, this dataset has of 35136 ozone measurements. Ozone levels vary greatly over this time period and across Mexico City. Strong daily seasonality or periodicity of ozone levels is perhaps the most prominent feature of these data (see Section 4.2 for more discussion). Ozone levels tend to peak in the afternoon, fall at night, and reach a minimum in the morning. We develop methods to address the following characteristics of the ozone data: spatial patterns over Mexico City, temporal variability over April and May (long-term temporal trends), and daily seasonality (refer to Figures 4.2 and 4.3). Of primary interest in our analysis is predicting ozone levels at unmonitored locations to estimate compliance with national ambient air quality standards and respiratory health risk at those unmonitored locations. Thus, our model must permit predictions at any location within the spatial boundaries of the monitoring network at any time in April or May.

Here, we provide three primary contributions to account for the attributes of the data and to meet the goals of our analysis. First, we account for daily seasonality directly through the covariance model, along with space and time, to model spatiotemporal patterns like those in these data (see Figures 4.3). Accounting for seasonality using autoregressive or dynamic terms in the mean function instead of the covariance function is common (e.g., see West and Harrison, 1997; Prado and West, 2010; Shumway and Stoffer, 2017), and models like these have even been applied to Mexico City ozone levels (Huerta et al., 2004; White et al., 2018). Our approach, however, allows richer relationships between linear time lags and daily seasonal patterns than discrete time-series models and are more natural for continuously-varying spatiotemporal data (see Stein, 2005).

Specifically, we capture daily seasonal patterns in ozone in Mexico City by modeling time as a quantity that lies on a circle (i.e. a 24-hour clock) and refer to this as *circular time*. We define *circular time lag* to be the minimum angle between two points on a 24-hour clock. Circular time lags are used to account for daily seasonality in our model and are used in conjunction with linear time lags that lie on the real line and spatial differences that lie in the spatial domain (Mexico City) for covariance modeling. To clarify this approach, we plot and compare circular time lags and linear time lags in Figure 4.1. We use both circular and linear time lags to define the covariance between ozone observations to account for autocovariance patterns like those we observe in these data (see Figure 4.3c), showing both periodicity and decay. Shirota et al. (2017) use circular time for spatiotemporal point process modeling but do not use linear time lags in conjunction with circular time lags, something that we argue is essential for modeling these data.

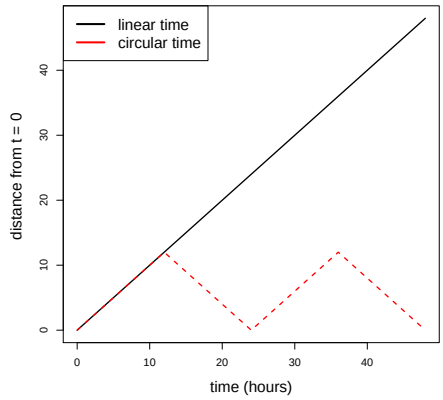


FIGURE 4.1: This represents our two ideas of temporal differences: linear time lags (solid and black) and circular time lags (red and dashed).

A Gaussian process with a covariance model over space, linear time, and circular time is a natural model for our goals (see, e.g., Banerjee et al., 2014). However, a fully Bayesian Gaussian process model with 35136 ozone measurements is not

computationally feasible. For this reason, we use the nearest-neighbor Gaussian process (NNGP) (Datta et al., 2016a) as a scalable alternative to the full Gaussian process that preserves predictive performance (Heaton et al., 2018).

Our second contribution focuses on adapting nearest-neighbor Gaussian processes (Datta et al., 2016a) for daily seasonality. This model assumes that spatial random effects are conditionally independent given a small subset of the other random effects (called neighbors), giving computational tractability. We discuss how to select neighbors when covariance does not monotonically decay with increasing linear time lags (see Jones and Zhang, 1997; Stein et al., 2004; Gramacy and Apley, 2015; Datta et al., 2016b, for discussion about neighbor selection for monotone decay). A particularly relevant example is Stein et al. (2004). In this paper, they work with spatial data that are collected in a "sawtooth" pattern that is conceptually similar to periodic temporal patterns in our data. In this setting, they argue for including neighbors in periodic spatial increments, as well as some distant neighbors. Similarly, we argue for including neighbors that correspond to periodic peaks in the covariance. Our approach yields computational benefits that make space-time NNGP modeling more scalable.

Our third and last contribution is scientific and less statistical. Our statistical analysis provides a greater understanding of the possible health risks associated with ozone levels at locations where ozone is not monitored. Posterior predictions from our Bayesian spatiotemporal model allow us to estimate the respiratory health risk (Chiogna and Pauli, 2011) and non-compliance with Mexico's ambient air quality standards (Diario Oficial de la Federación, 2014b) in areas of Mexico City where ozone is not monitored. To be clear, using posterior predictions in this way is not statistically novel; however, this framework allows us to propagate uncertainty from our ozone model into our inference on respiratory risk and non-compliance with federal air quality standards.

In summary, we provide the following contributions: First, we account for daily seasonality through covariance modeling rather than through terms in the mean function of the model and introduce new classes of appropriate covariance functions for our approach. Second, we adapt the nearest-neighbor Gaussian process for covariance functions with seasonality, allowing for scalable model fitting, prediction, and inference for the Mexico City ozone data. Third, we use this model to assess respiratory risks and compliance with Mexican ambient air quality standards at locations where ozone is not monitored.

We continue by discussing the Mexico City pollution monitoring data in Section 4.2. Then, we discuss relevant covariance classes for these data and present new nonseparable covariance classes for circles and linear time in Section 4.3. Using these covariance classes, we discuss modeling details (Section 4.4.1), neighbor selection (Section 4.4.2), model fitting (Section 4.4.3), and prediction and inference (Section 4.4.4). We address compliance with Mexican air quality standards and respiratory health risks associated with ground-level ozone in Section 4.5. Lastly, we give concluding remarks, comment on our approach, and discuss future extensions in Section 4.6.

4.2 Mexico City Ozone Monitoring Data

In this dataset, we have hourly ozone measurements for April and May of 2017 at $n_s = 24$ monitoring stations across Mexico City, Mexico. At each station, we have measurements at $n_t = 1464$ hours, giving $n = 35136$ observations in total[1]. Similar to Sahu et al. (2007) and Berrocal et al. (2010), we found that using square-root ozone reduced correlation between the variance of model residuals and the mean and led to

[1] Although ozone levels are given hourly, these hourly quantities are derived is an average of the 60 minute-by-minute measurements. Missing measurements were imputed prior to receiving the data using the measurements at the nearest station within the same region. If no simultaneous measurements in the same region were available, then the missing measurement was filled in using the nearest station in a different region.

better predictive performance than modeling on the original scale. On the square-root scale, the sample mean and variance are 6.07 and 6.41, respectively. Because ozone formation requires heat and sunlight (Sillman, 1999), we use temperature as an explanatory variable for ozone. In addition, we use relative humidity in our model because rain clears out ozone and high humidity is associated with decreased sunlight. Both relative humidity and temperature are measured hourly at the same 24 stations as ozone.

We first examine the variability in ozone levels across the city by plotting station locations with their mean ozone levels over April and May in Figure 4.2. In general, ozone levels decrease as we move northward but not uniformly as central Mexico City has the lowest ozone values. Peak ozone levels in the south are largely explained by a wind corridor that flows northeast to southwest, moving ozone and ozone precursors produced along this wind path to southern parts of Mexico City. This ozone is then trapped in the south by mountains on Mexico City's southwest boundary.

To illustrate how ozone levels vary over April and May, we plot daily means and maxima in Figure 4.3a. Over April and May, the highest and lowest daily maxima differ by a factor of three, while the highest and lowest daily means differ by more than a factor of two. Peak levels of ozone occur in May, but there are also significant peaks in April. Figure 4.3b plots ozone averages as a function of hour of the day and demonstrates a clear peak in ozone levels around 2:00 or 3:00 p.m.

To examine whether the temporal patterns in ozone can be adequately explained by only using temperature and humidity, we fit a linear model to ozone using relative humidity and temperature as covariates. Using this model, we examine the autocorrelation of the model residuals for each of the 24 locations (see Figure 4.3c). For each site, the temporal autocorrelation pattern peaks every 24 hours but decays overall. Thus, a purely seasonal or purely decaying covariance model would be insufficient for these data. We demonstrate this point empirically in Section 4.5. Together, the

plots in Figure 4.3 motivate our covariance discussion in Section 4.3.

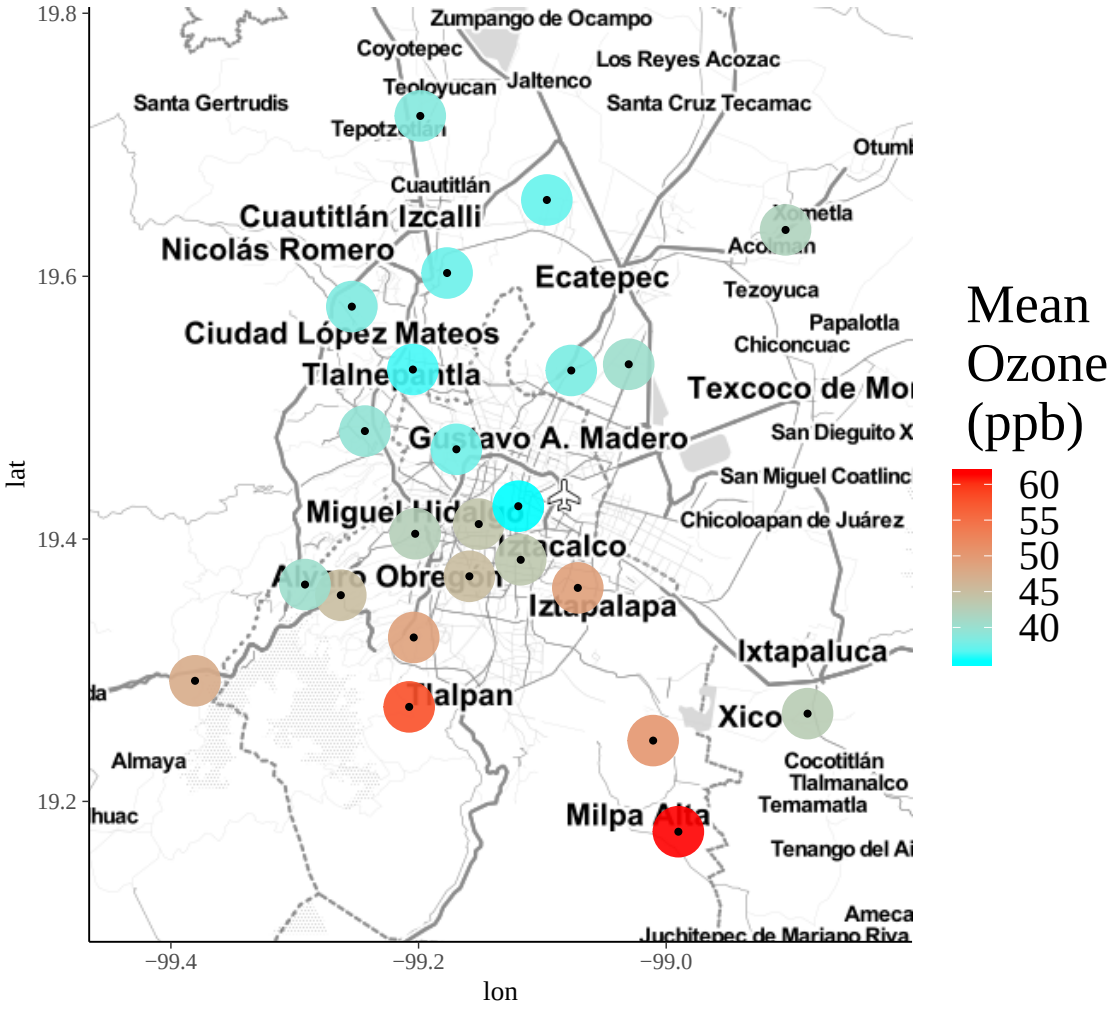
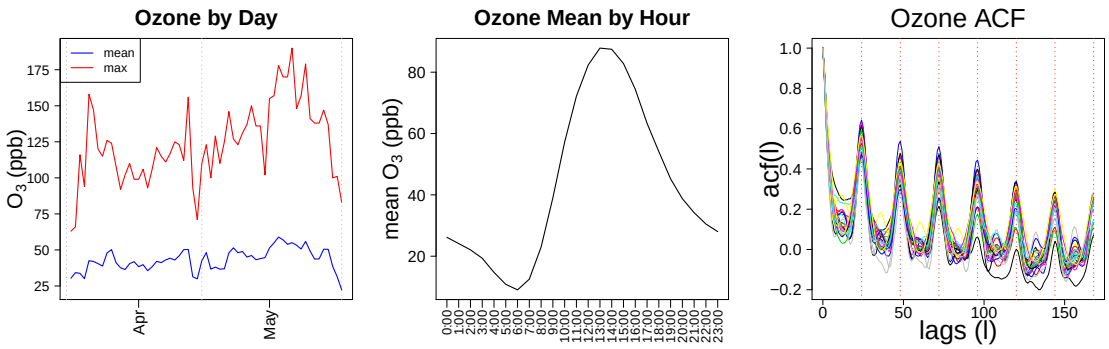


FIGURE 4.2: Station locations with mean ozone coded by the color of the point with low to high ozone indicated by cyan to red.



(a) Mean and max over days (b) Mean ozone over hours (c) ACF for ozone

FIGURE 4.3: (Left) Mean and maximum ozone levels for all stations, each day (Center) Mean ozone over the hour of the day, averaging over days in April and May and stations (Right) Autocorrelation function (ACF) of the model residuals for each station using relative humidity and temperature as explanatory variables. Each curve represents the ACF of each station.

4.3 Covariance Models

We present our covariance modeling approach prior to discussing other components of our model because the covariance model is key to all other model components. Given our goal of predicting ozone levels at unmonitored locations and motivated by the autocovariance patterns displayed and discussed in Section 4.2, the overall goals of our covariance model are to (i) explain similarities in ozone levels due to circular time (daily seasonality), (ii) account for long-term temporal changes in ozone levels, and (iii) capture spatial patterns. To meet these goals, we consider covariance models on space, circular time, and linear time. In Section 4.3.1, we outline the types of covariance models that we consider. In Section 4.3.2, we propose two classes that allow nonseparable relationships between linear and circular time. Then, we provide specific examples in Section 4.3.3 that are compared in Section 4.5.

4.3.1 Covariance Modeling Approach

We start with some notation and background. Consider two space-time pairs: $(\mathbf{s}, t)$, $(\mathbf{s}', t') \in \mathcal{D} \times \mathbb{R}$, where $\mathcal{D} \subset \mathbb{R}^2$ is our spatial domain (Mexico City) indexed by eastings and northings (in units of km) and $t \in \mathbb{R}$ is the time of the observation (in hours from the start of April). Additionally, let $\mathbf{h} = \mathbf{s} - \mathbf{s}'$ and $u = t - t'$ be the spatial and temporal lags, respectively. As discussed, we capture daily seasonality using circular time (the unit circle is used without loss of generality). For this, let $\mathbb{S}^1 \subset \mathbb{R}^2$ be the unit circle and $\theta : \mathbb{S}^1 \times \mathbb{S}^1 \to [0, \pi]$ be the minimum angle between any two points on a circle. The angle θ is defined as $\theta(u) = \min\left(\frac{\pi(|u| \mod 24)}{12}, 2\pi - \frac{\pi(|u| \mod 24)}{12}\right)$, where we use $\mod$ to indicate modular division. A one-hour difference in circular time corresponds to an angular difference of $\frac{\pi}{12}$. Together, $(\|\mathbf{h}\|, \theta, |u|) \in [0, \infty) \times [0, \pi] \times [0, \infty)$ are the inputs of the covariance models that we consider, where $\|\cdot\|$ denotes Euclidean distance.

There are direct approaches for periodic covariance functions that use sinusoidal functions of linear time lags. For fixed period P, MacKay (1998) uses mapping, $u \mapsto \left(\sin\left(\frac{P}{2\pi}u\right), \cos\left(\frac{P}{2\pi}u\right)\right)$, in the squared exponential correlation function to obtain a periodic correlation model, $C(u) = \exp\left(-\frac{2\sin^2\left(\frac{\pi u}{P}\right)}{\alpha^2}\right)$, where α acts as a range parameter. Rasmussen and Williams (2006) propose a covariance function that decays away from periodicity by taking the product of periodic and decaying squared-exponential covariance functions, $C(u) = \exp\left(-\frac{\sin^2\left(\frac{\pi u}{P}\right)}{\alpha_1^2} - \frac{u^2}{\alpha_2^2}\right)$, where α_1 and α_2 are range parameters for the periodic and decaying components of the model. Solin and Särkkä (2014) show an explicit link between some Gaussian process specifications with separable periodic and decaying covariance functions and state-space models. Covariance models using sinusoidal functions of linear temporal lags are,

however, limited to separable models that ignore potentially important interactions between linear time lags and circular time lags. Because our goal is to capture more general nonseparable relationships between linear and circular time lags, we turn to covariance functions on circular time and linear time.

We consider both separable and nonseparable covariance models on $\mathcal{D} \times \mathbb{S}^1 \times \mathbb{R}$ for Mexico City ozone. Separable models assume that there are no interactions between the subspaces of our domain (i.e. there are no space/linear time, space/circular time, or linear/circular time interactions). In the separable case, the covariance function on $\mathcal{D} \times \mathbb{S}^1 \times \mathbb{R}$ is the product of three covariance functions, one on each space. We expect that a separability assumption is unrealistic and hypothesize that covariance models incorporating interactions between sub-domains may enhance our predictive ability (see, e.g., Cressie and Huang, 1999; Gneiting, 2002; Kolovos et al., 2004; Gneiting et al., 2006, for similar arguments). Nonseparable models have generally been used to capture space/linear-time interactions (see, e.g., Gneiting, 2002; Porcu et al., 2016; Shirota et al., 2017; White and Porcu, 2018); however, we also explore nonseparable relationships between linear time and circular time. To address the potential nonseparable relationships present in these data (i.e. space and linear time, space and circular time, and linear-circular time interactions), we explore a variety of models that allow more realistic covariance relationships than separable covariance functions.

Working on $\mathcal{D} \times \mathbb{S}^1 \times \mathbb{R}$, we use possible covariance models have the form:

$$C(\|\mathbf{h}\|, \theta, |u|), \qquad \mathbf{h} \in \mathbb{R}^2, \theta \in [0, \pi], u \in \mathbb{R}, \tag{4.1}$$

where $C : [0, \infty) \times [0, \pi] \times [0, \infty) \to \mathbb{R}$ ensures that $C(\|\mathbf{h}\|, \theta, |u|)$ is positive-definite. To construct covariance functions of type (4.1), we propose covariance functions that are *partially* nonseparable. As an abuse of notation, we write h for $\|\mathbf{h}\|$ and u for $|u|$ and consider the following constructions for nonseparability:

(**A**) $(\mathcal{D}, \mathbb{R})$ - space and linear time nonseparability - $C(h, \theta, u) = C_1(h, u)C_2(\theta)$

(**B**) $(\mathcal{D}, \mathbb{S}^1)$ - space and circular time nonseparability - $C(h, \theta, u) = C_3(h, \theta)C_4(u)$

(**C**) $(\mathbb{S}^1, \mathbb{R})$ - circular and linear time nonseparability - $C(h, \theta, u) = C_5(\theta, u)C_6(h)$

We also consider complete separability, $C(h, \theta, u) = C_6(h)C_4(u)C_2(\theta)$, to motivate using nonseparable models for these data. All constructions here account for daily seasonality, long-term changes over April and May, and spatial autocorrelation. The partial nonseparability of each construction provides flexibility in different ways. To create valid covariance models of these types, one must select C_i, $i = 1, \ldots, 6$, to be valid covariance functions on their respective spaces. We focus on possible covariance selections for constructions (**A**), (**B**), and (**C**) for the remainder of this section and in Section 4.3.3.

We first define the Gneiting class of functions (Gneiting, 2002), $\mathcal{G}_{dim} : [0, \infty)^2 \to \mathbb{R}_+$, as

$$\mathcal{G}_{dim}(x_1, x_2) := \frac{1}{\psi(x_2)^{dim/2}} \varphi\left(\frac{x_1}{\psi(x_2)}\right), \qquad x_1, x_2 \geq 0, \tag{4.2}$$

where dim is the dimension of the space over which x_1 is computed and x_1, x_2 are placeholders for h^2, u^2, or θ. A similar class is proposed by Porcu et al. (2016), which we call the modified Gneiting class, $\mathcal{P} : [0, \infty) \times [0, \pi] \to \mathbb{R}$, and is given by

$$\mathcal{P}(x_1, x_2) := \frac{1}{\psi_{[0,\pi]}(x_2)^{1/2}} \varphi\left(\frac{x_1}{\psi_{[0,\pi]}(x_2)}\right), \qquad x_1 \geq 0, x_2 \in [0, \pi]. \tag{4.3}$$

For both classes, the function $\varphi : [0, \infty) \to \mathbb{R}_+$ is completely monotonic; that is, φ is infinitely differentiable on $(0, \infty)$, satisfying $(-1)^n \varphi^{(n)}(t) \geq 0$, $n \in \mathbb{N}$. The function ψ is strictly positive and has a completely monotonic derivative. Here, $\psi_{[0,\pi]}$ denotes the restriction of ψ to the interval $[0, \pi]$. We provide selections for $\varphi(\cdot)$ and $\psi(\cdot)$ in Tables 4.1 and 4.2.

Table 4.1: Examples of functions that are strictly positive and have a completely monotonic derivative

ψ	Expression	Parameters
Dagum	$\psi(t) = 1 + \left(\frac{t^\beta}{1+t^\beta}\right)^\tau$	$\beta, \tau \in (0, 1]$
Gen. Cauchy	$\psi(t) = (1 + t^\alpha)^{\beta/\alpha}$	$\alpha \in (0, 1]$, $\beta \leq \alpha$
Power	$\psi(t) = c + t^\alpha$	$\alpha \in (0, 1]$, $c > 0$

Table 4.2: Examples of completely monotonic functions

φ	Expression	Parameters
Dagum	$\varphi(t) = 1 - \left(\frac{t^\beta}{1+t^\beta}\right)^\tau$	$\beta, \tau \in (0, 1]$
Matérn	Equation (4) - Manuscript	$0 < \nu \leq 1/2$
Gen. Cauchy	$\varphi(t) = (1 + t^\alpha)^{-\beta/\alpha}$	$\alpha \in (0, 1]$, $\beta > 0$
Pow. Expon,	$\varphi(t) = \exp(-t^\alpha)$	$\alpha \in (0, 1]$

The Gneiting class $\mathcal{G}_{dim}$ in (4.2) provides a general class of nonseparable space-linear time covariance functions. Thus, it can be used for construction (**A**), i.e. $C_1(h, u) = \mathcal{G}_2(h^2, u^2)$. For construction (**A**), we limit our selections to $\mathcal{G}_{dim}$, although other examples exist in the literature. For construction (**B**), space-circular time nonseparability, Shirota et al. (2017) propose using $C_3(h, \theta) = \mathcal{G}_2(h^2, \theta)$ to account for the interaction between circular time lags and spatial differences. Other appropriate models for $C_3(h, \theta)$ are given by Theorem 2 of White and Porcu (2018), which shows $C_3(h, \theta) = \mathcal{G}_2(\theta, h^2)$ is positive-definite, and Theorem 1 of Porcu et al. (2016), which proves that $C_3(h, \theta) = \mathcal{P}(h^2, \theta)$ is a valid covariance function. The classes proposed in Porcu et al. (2016) and White and Porcu (2018) can also be used for circular-linear time nonseparability (i.e. for construction (**C**), replace h with u to obtain $C_5(\theta, u)$). Additional choices for $C_5(\theta, u)$ of construction (**C**) are the core of our theoretical contributions in Section 4.3.2.

For marginal covariance functions needed to complete covariance constructions

(**A**), (**B**), and (**C**), we turn to the Matérn class $\mathcal{M}_{\alpha,\nu} : [0,\infty) \to \mathbb{R}_+$, defined as

$$\mathcal{M}_{\alpha,\nu}(x) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{x}{\alpha}\right)^{\nu} \mathcal{K}_{\nu}\left(\frac{x}{\alpha}\right), \qquad \alpha, \nu > 0, \tag{4.4}$$

where $\mathcal{K}_\nu$ is the MacDonald function (Gradshteyn and Ryzhik, 2007). For linear time, $C_4(u) = \mathcal{M}_{\alpha,\nu}(u)$ is valid for any $\alpha, \nu > 0$ (Stein, 1999). To use the Matérn covariance function on a circle (i.e. $C_2(\theta) = \mathcal{M}_{\alpha,\nu}(\theta)$), the parameter ν must be in $(0, 1/2]$ (see Gneiting, 2013). Lastly, the Matérn covariance function $C_6(h) = \mathcal{M}_{\alpha,\nu}(h)$ is valid on $\mathbb{R}^2$ when $\alpha, \nu > 0$ (Stein, 1999).

4.3.2 Covariance Functions for Circular and Linear Time

In addition to the current literature on spheres cross time, which we reviewed in Section 4.3.1, we propose two classes of nonseparable covariance functions that are unique to $\mathbb{S}^1 \times \mathbb{R}$ and are motivated by the autocorrelation patterns in Figure 4.3c in Section 4.2, exhibiting both seasonality and decay. We propose two classes of nonseparable covariance functions on $\mathbb{S}^1 \times \mathbb{R}$ to allow nonseparable relationships between linear and circular time lags. For this, we define the variogram of an intrinsically stationary process, $\gamma : \mathbb{R} \to [0,\infty)$, as $\gamma(u) = \frac{1}{2}\text{var}\left(Z(t+u) - Z(t)\right)$, $t, u \in \mathbb{R}$. Recall that, for any covariance function, the ratio $\rho(\theta, u) = C(\theta, u)/C(0,0)$ is a correlation function. Lastly, we define sinh as the hyperbolic sine function.

Theorem 1. *Let $\gamma : \mathbb{R} \to \mathbb{R}_+$ be a variogram and $\rho : \mathbb{R} \to [-1, 1]$ be a correlation function.*

(I) *Let C be defined as*

$$C(\theta, u) = \frac{1}{\gamma(u)} + \frac{\pi}{2} \frac{\sinh\left[\sqrt{\gamma(u)}(\pi - \theta)\right]}{\sinh\left[\sqrt{\gamma(u)}\pi\right]} \qquad \theta \in [0, \pi], u \in \mathbb{R}. \tag{4.5}$$

Then, $C(\theta, u)/C(0,0)$ is a correlation function.

(II) *Let C be defined as*

$$C(\theta, u) = \exp\{\rho(u)\cos(\theta) - 1\}\cos\{\rho(u)\sin\theta\} \qquad \theta \in [0, \pi], u \in \mathbb{R}.. \quad (4.6)$$

Then, $C(\theta, u)$ is a correlation function.

Proof. We start by noting that arguments in Berg and Porcu (2017) show that the functions C defined at points **(I)** and **(II)** are positive definite if and only if they can both be written as

$$C(\theta, u) = \sum_{k=0}^{\infty} b_k(u)\cos(k\theta), \qquad (\theta, u) \in [0, \pi] \times \mathbb{R}, \quad (4.7)$$

where uniquely determined sequence of positive definite functions $\{b_k(\cdot)\}_{k=0}^{\infty}$ has $\sum_k b_k(0) < \infty$. To show **(I)**, we resort to [1.445.2] in Gradshteyn and Ryzhik (2007):

$$\sum_{k=1}^{\infty} \frac{\cos(kx)}{(k^2 + a^2)} = \frac{\pi}{2}\frac{\sinh(a(\pi - x))}{\sinh(\pi a)}, \qquad a > 0, x \in \mathbb{R}.$$

The function in Equation (4.5) admits an expansion of the type (4.7) with coefficients $b_k(u) = (k^2 + \gamma(u))^{-1}$ for $k = 0, 1, \ldots$. It can be namely verified that $\sum_k b_k(0)$ is finite. Thus, the proof is completed. As for Assertion **(II)**, Equation (4.6) comes straight by considering the expansion in [1.449.2] in Gradshteyn and Ryzhik (2007):

$$\sum_{k=0}^{\infty} \frac{p^k \cos(kx)}{k!} = \mathrm{e}^{p\cos(x)}\cos(p\sin x),$$

which is absolutely convergent provided $p^2 \leq 1$. We now replace p with the correlation function $\rho(\cdot)$ to obtain (4.6). The proof is completed. □

Examples of variograms $\gamma(\cdot)$ are given in Tables 4.1 and 4.3. The classes presented in Theorem 1 do not decay to zero as t gets large. If eventual decay to zero is preferred, then the model $C(\theta, u) = C_5(\theta, u)C_2(u)$ could be used, where $C_2(u)$ decays

to zero as a function of time. We found that this construction improves predictive performance for these data.

Table 4.3: Examples of variograms $\gamma(u) = c_0 + c_1 f(u/c_t)$, where c_0, c_1 and c_t are strictly positive parameters

Model	Function $f(u)$	Parameters
Power	$\left(\frac{u}{c_t}\right)^{\alpha}$	$\alpha \in (0,1)$
Cauchy	$\left\{1 - [1 + (u)^{\alpha}]^{-\lambda}\right\}$	$\lambda > 0$, $\alpha \in (0,1]$
Exponential	$(1 - e^{-u})$	
Gaussian	$\left(1 - e^{-(u)^2}\right)$	
Matérn	$\left[1 - \frac{1}{2^{\nu-1}\Gamma(\nu)} u^{\nu} K_{\nu}(u)\right]$	$\nu > 0$

4.3.3 Covariance Examples

In this section, we provide four specific covariance functions that we compare in Section 4.5. Example 1 in Table 4.4 comes from Gneiting (2002) and is used for construction (**A**) (space and linear time nonseparability). For space and circular time nonseparability (**B**), we give a Example 2 in 4.4 from Shirota et al. (2017). Lastly, we provide two covariance functions for circular and linear time nonseparability (construction (**C**)) in Table 4.4, Examples 3 and 4. Example 3 is adapted from White et al. (2018), while Example 4 comes from Theorem 1. Other models of these types were considered; however, these represent the models with the best predictive performance. As discussed, we complete constructions (**A**), (**B**), and (**C**) using the Matérn covariance function (4.4).

Table 4.4: Covariance examples (Ex.) presented for model selection. Results for this comparison are presented in Section 4.5.

Ex.	Covariance Function	Parameters
1	$C(h,u) = \frac{\sigma^2}{\left(1+\left(\frac{u}{c_t}\right)^\alpha\right)^{\delta+\beta d/2}}\left(1+\frac{h^{2\gamma}}{c_s^{2\gamma}\left(1+\left(\frac{u}{c_t}\right)^\alpha\right)^{\beta\gamma}}\right)^{-\lambda}$	$\delta, \lambda > 0$, $\beta, \gamma \in (0,1]$, $\alpha \in (0,2]$
2	$C(h,\theta) = \frac{\sigma^2}{\left(1+\left(\frac{\theta}{c_t}\right)^\alpha\right)^{\delta+\beta d/2}}\left(1+\frac{h^{2\gamma}}{c_s^{2\gamma}\left(1+\left(\frac{\theta}{c_t}\right)^\alpha\right)^{\beta\gamma}}\right)^{-\lambda}$	$\delta, \lambda > 0$, $\beta, \gamma \in (0,1]$, $\alpha \in (0,2]$
3	$C(\theta,u) = \frac{\sigma^2}{\left(1+\left(\frac{\lvert u\rvert}{c_t}\right)^\alpha\right)^{\delta+\beta d/2}}\left(1+\frac{\theta^{\gamma}}{c_s^{\gamma}\left(1+\left(\frac{\lvert u\rvert}{c_t}\right)^\alpha\right)^{\beta\gamma}}\right)^{-\lambda}$	$\delta > 0$, $\beta, \gamma \in (0,1]$, $\alpha \in (0,2]$, $\lambda > 0$
4	$C(\theta,u) = \exp\left\{\exp\left[-\left(\frac{u}{c_{t_1}}\right)^\alpha\right]\cos(\theta) - \frac{u}{c_{t_2}} - 1\right\} \times$ $\cos\left\{\exp\left[-\left(\frac{u}{c_{t_1}}\right)^\alpha\right]\sin(\theta)\right\}$	$\alpha \in (0,2]$, $c_{t_1}, c_{t_2}, \lambda > 0$

4.4 Methods and Models

In this section, we detail our modeling approach. First, let $\mathbf{x}(\mathbf{s},t)$ be relative humidity and temperature and $\boldsymbol{\beta}$ be corresponding regression coefficients. Using covariance models discussed in Section 4.3, we specify $w(\mathbf{s},t)$ as a mean-zero NNGP and $\epsilon(\mathbf{s},t)$ as Gaussian error with variance τ^2. We also define $\sigma^2 = C(0,0,0)$. Combined, we envision a hierarchical spatiotemporal model for hourly square-root ozone $\sqrt{Y(\mathbf{s},t)}$ measured at the location-time pair $(\mathbf{s},t)$ as

$$\begin{aligned}\sqrt{Y(\mathbf{s},t)} &= \mathbf{x}(\mathbf{s},t)^\top\boldsymbol{\beta} + w(\mathbf{s},t) + \epsilon(\mathbf{s},t), \qquad (4.8)\\ w(\mathbf{s},t) &\sim \text{NNGP}\left[0, C\left(h,\theta,u\right)\right],\\ \epsilon(\mathbf{s},t) &\overset{iid}{\sim} \mathcal{N}(0,\tau^2).\end{aligned}$$

In Section 4.4.1, we introduce the nearest-neighbor Gaussian process that is used for spatial random effects. We discuss how neighbors are selected in Section 4.4.2 for model fitting and prediction. Then, we present prior distributions and model fitting in Section 4.4.3. Lastly, we discuss prediction and inference in Section 4.4.4,

including model selection (Section 4.4.4), compliance to ozone standards (Section 4.4.4), and assessing respiratory risk (Section 4.4.4).

4.4.1 Nearest-Neighbor Gaussian Process Model

Because we are modeling the data in continuous space-time, it is natural to specify random effects through a functional prior, most commonly a Gaussian process (GP) (see, e.g., Rasmussen and Williams, 2006; Banerjee et al., 2014). As discussed, the primary modeling decision for GP models is the covariance function. Likelihood computations for GP models require inverting an $n \times n$ matrix, making hierarchical Bayesian GP models intractable with even moderate amounts of data (e.g., $n > 10000$). Since our dataset has $n = 35136$ observations, the full GP is not feasible.

Many have addressed this computational bottleneck using either low-rank or sparse matrix methods (see Heaton et al., 2018, for a review and comparison of these methods). Low-rank methods project the original process onto representative points or knots (see, e.g., Higdon, 2002; Banerjee et al., 2008; Cressie and Johannesson, 2008; Stein, 2008); however, these approaches often perform poorly for prediction as they often over-smooth (see Stein, 2014). Alternatively, sparse methods either induce zeros in the covariance matrix using compactly supported covariance functions (see, e.g., Furrer et al., 2006; Kaufman et al., 2008; Bevilacqua et al., 2016) or in the precision matrix by assuming conditional independence (Vecchia, 1988; Stein et al., 2004). We ultimately favor approaches that assume conditional independence because predictive performance is generally better (Heaton et al., 2018) and the class of valid covariance models is more expansive.

Sparse precision methods date to Vecchia (1988) and are used to approximate the likelihood of the full GP model using conditional probability. See Stein et al. (2004) and Bevilacqua et al. (2012) for additional discussion on Vecchia approximations. Gramacy and Apley (2015) and Datta et al. (2016a) extend Vecchia approximations

to process modeling. The nearest-neighbor Gaussian process is itself a Gaussian process (Datta et al., 2016a) and has good predictive performance relative to other fast GP methods (See Heaton et al., 2018).

The nearest-neighbor Gaussian process (NNGP) is derived from a *parent* Gaussian process and assumes that random variables are conditionally independent given a *conditioning or neighborhood set*: a subset of the other random variables (Datta et al., 2016a). Equivalently, the NNGP can be viewed as a directed acyclic graph (DAG) that induces sparsity in the precision matrix of the parent process by assuming conditional independence given neighborhood sets. To specify the NNGP model, one must select a covariance model, a reference set $\mathcal{R}$, and conditioning or neighborhood sets $N(\mathbf{s}, t)$ for each observation in the reference set $\mathcal{R}$.

To define the NNGP, let the reference set $\mathcal{R} = \mathcal{S} \times \mathcal{T}$, where $\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_{n_s}\}$ are the locations of the n_s ozone monitoring sites and $\mathcal{T} = \{t_1, ..., t_{n_t}\}$ denote the n_t recorded time points at each spatial location. For the Mexico City application, we select the reference set to be the location-time pairs of $n_t = 1464$ hourly measurements at the $n_s = 24$ monitoring sites. The reference set $\mathcal{R}$ must be ordered to properly specify the model. Time provides natural ordering, and we impose spatial ordering using latitude, south to north.

To complete the specification of the NNGP, we define neighborhood or conditioning sets $N(\mathbf{s}_i, t_j)$ for each location-time pair in the reference set. This set $N(\mathbf{s}_i, t_j)$ consists of up to m neighbors of $(\mathbf{s}_i, t_j)$ chosen from observations ordered before $(\mathbf{s}_i, t_j)$, and we denote the collection of all neighborhood sets as $N_{\mathcal{R}} = \{N(\mathbf{s}_i, t_j) : i = 1, ..., n_s, j = 1, ..., n_t\}$. We discuss in detail how these conditioning sets are chosen in Section 4.4.2. Together, $\mathcal{R}$ and $N_{\mathcal{R}}$ define a Gaussian directed acyclic graph

(DAG) prior distribution for the spatial random effects $\mathbf{w}_{\mathcal{R}}$,

$$p(\mathbf{w}_{\mathcal{R}}) = \prod_{i=1}^{n_s}\prod_{j=1}^{n_t} p(\mathbf{w}(\mathbf{s}_i,t_j)|\mathbf{w}_{N(\mathbf{s}_i,t_j)}) = \prod_{i=1}^{n_s}\prod_{j=1}^{n_t} \mathcal{N}(\mathbf{w}(\mathbf{s}_i,t_j)|\mathbf{B}_{(\mathbf{s}_i,t_j)}\mathbf{w}_{N(\mathbf{s}_i,t_i)}, \mathbf{F}_{(\mathbf{s}_i,t_j)}), \tag{4.9}$$

where $\mathbf{B}_{\mathbf{s}_i,t_j} = C_{(\mathbf{s}_i,t_j),N(\mathbf{s}_i,t_j)}C^{-1}_{N(\mathbf{s}_i,t_j)}$, $\mathbf{F}_{(\mathbf{s}_i,t_j)} = \sigma^2 - \mathbf{B}_{(\mathbf{s}_i,t_j)}C_{N(\mathbf{s}_i,t_j),(\mathbf{s}_i,t_j)}$, $\mathbf{w}_{N(\mathbf{s}_i,t_j)}$ is the subset of $\mathbf{w}_{\mathcal{R}}$ corresponding to neighbors $N((\mathbf{s}_i,t_j))$, $C_{(\mathbf{s}_i,t_j),N(\mathbf{s}_i,t_j)}$ is an $m-$vector with covariances between $(\mathbf{s}_i,t_j)$ and its neighbors, and $C_{N(\mathbf{s}_i,t_j)}$ is an $m \times m$ matrix with the covariances between the neighbors of $(\mathbf{s}_i,t_j)$. The elements of these covariance matrices are calculated using covariance functions discussed in Section 4.3.3. Although the random effects of the NNGP have a conditional mean of $\mathbf{B}_{(\mathbf{s}_i,t_j)}\mathbf{w}_{N(\mathbf{s}_i,t_i)}$, each random effect has a mean of zero marginally (see Datta et al., 2016a).

4.4.2 Neighbor Selection

Neighbor selection is challenging in part because "best" neighborhood sets vary depending on the covariance function and associated parameters (see, e.g., Vecchia, 1988; Datta et al., 2016b). In our case, this is particularly challenging because we use covariance functions that account for daily seasonality (see Section 4.3). Datta et al. (2016a) use nearest neighbors in the conditioning set, while Jones and Zhang (1997) and Datta et al. (2016b) argue for including the most correlated neighbors in the conditioning set. Stein et al. (2004), however, show that both nearest-neighbor and most-correlated conditioning set selections can lead to models that are suboptimal. Accordingly, they include some distant neighbors in the conditioning set. There are, however, computational benefits to selecting nearest- or most-correlated neighbors (see Stein, 2005; Datta et al., 2016b). Due to the seasonality in ozone levels, we take a similar approach to Stein et al. (2004) for selecting neighbors for the nearest-neighbor Gaussian process. Our argument is simple: conditioning sets that include observations at and near periodic peaks have better predictive performance.

We construct neighborhood sets for each observation in $\mathcal{R}$ using simultaneous and past observations. For simultaneous neighbors, we select the six nearest available spatial neighbors from locations ordered previous to that observation in the reference set (we select fewer neighbors if fewer than six are available). We also include six neighbors (the five nearest-neighbors and the same location) for lags 1, 2, 23, 24, 25, and 168 hours back in the conditioning set, when they are available (e.g. at $t = 167$ hours, we do not have lags from 168 hours back). These lags and the number of spatial neighbors were chosen based on a sensitivity analysis presented in the Appendix C.1, where out-of-sample predictive performance is compared for various conditioning sets. For these data, we find that selecting six spatial neighbors and using lags of 1, 2, 23, 24, 25, and 168 hours each accounted for an improvement of about 3.5% in predictive performance, compared to using only two spatial neighbors or only most-recent neighbors (7% accounting for both lag and spatial neighbor selection). This highlights how important carefully selecting conditioning sets is in this setting.

To make our neighbor selection procedure concrete, we diagram our approach using time and a one-dimensional space. Because we impose ordering among spatial locations (south to north) to specify the NNGP, ozone monitoring sites can be viewed similarly to the one-dimensional space diagrammed here. For an observation at $(\mathbf{s}, t)$, we show its conditioning set for a nearest-neighbor approach and our approach in Figure 4.4. The spatial neighbors at the same time as the observation are not identical to the spatial neighbors at past times because our spatial ordering prohibits selecting simultaneous neighbors north of the observation.

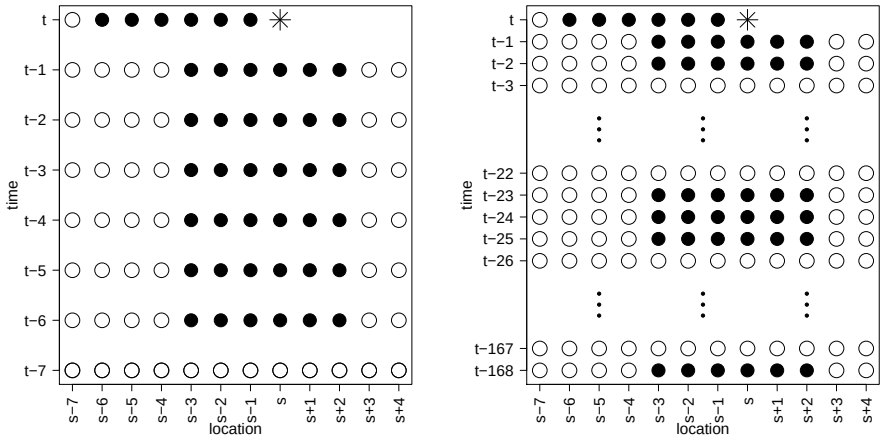


FIGURE 4.4: Here filled circles represent observations in the conditioning set (i.e. selected neighbors). Empty circles represent observations not included in the conditioning set. (Left) Nearest-neighbor conditioning set for an observations at $(\mathbf{s}, t)$. (Right) Our conditioning set for location $(\mathbf{s}, t)$ that uses neighbors at and near periodic covariance peaks and a distant neighbor (one-week back).

Because we use the same spatiotemporal structure in our conditioning sets for all observations at the same location (i.e. the same temporal lag spacing with the same spatial neighbors), given that all the lags are available, we obtain many of the computational benefits suggested by Stein (2005). In particular, $\mathbf{B}_{(\mathbf{s}_i,t_i)} = \mathbf{B}_{(\mathbf{s}_i,t_j)}$ and $\mathbf{F}_{(\mathbf{s}_i,t_i)} = \mathbf{F}_{(\mathbf{s}_i,t_j)}$ when t_i and t_j are greater than the maximum lag (168 hours back in our case). In fact, this equality holds so long as there is not an additional lag added to the model between t_i and t_j. For example, $\mathbf{F}_{(\mathbf{s}_i,167)} = \mathbf{F}_{(\mathbf{s}_i,168)}$, but $\mathbf{F}_{(\mathbf{s}_i,168)} \neq \mathbf{F}_{(\mathbf{s}_i,169)}$ because there is an additional lag available at time $t = 169$. Using these types of conditioning sets, with the same spatial neighbors at n_{lags} temporal lags, one needs to store and carry out operations on $n_s \times (n_{\text{lags}} + 1)$ covariance matrices of size $m \times m$ or smaller. In comparison, the standard NNGP requires operations on n matrices of size $m \times m$. Matrix operations for these repeated matrices are unnecessary, and the computational gains are significant.

For prediction at unmonitored locations and times, any subset of the reference

set $\mathcal{R}$ (all space-time pairs in the data) can be used in the conditioning set. Because our goal is understanding past risks associated with high ozone levels (i.e. we are not interested in forecasting risk and compliance here), we use past and future lags from $\mathcal{R}$ to specify the neighborhood or conditioning set $N(\mathbf{s}, t)$. Specifically, we use the six nearest spatial neighbors at lags 1, 2, 23, 24, 25, and 168 hours back and forward, as well as the six nearest spatial neighbors at the same time as the prediction. Because we condition on both past and future observations, these predictions conceptually resemble the full conditional distributions used for held-out data.

4.4.3 Prior Distributions and Model Fitting

We begin here by discussing the prior distributions for our model parameters. We use inverse-gamma prior distributions for τ^2 and σ^2 with 2.1 and 10 for shape and rate parameters. The prior mean and variance for τ^2 and σ^2 are near 9 and over 800, respectively. Since the mean and variance of square-root ozone levels in April and May are both near six, these selections are not overly informative. For bounded covariance parameters, we use uniform prior distributions over the support of the parameter. For parameters that are strictly positive, we use gamma prior distributions with shape and rate of 0.01. See Table 4.4 to see the domains of the parameters for each covariance function that we consider. We assume that regression coefficients *a priori* follow independent normal distributions with mean zero and variance 10^3.

For this model formulation, a Gibbs sampler is readily available for regression coefficients $\boldsymbol{\beta}$, variance parameters τ^2 and σ^2, and spatial random effects $w(\mathbf{s}_i, t_j)$. For this model, we provide full conditional distributions for the Gibbs sampler in Appendix C.2. With the exception of σ^2 and τ^2, we update covariance function parameters using the Metropolis-Hastings algorithm with a multivariate normal proposal distribution with covariance estimated throughout the burn-in of the sampler. Letting $\boldsymbol{\eta}$ denote all model parameters, our Markov chain Monte Carlo (both Gibbs

and Metropolis-Hastings updates together) model fitting yields samples from the joint posterior distribution $\boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, ..., \boldsymbol{\eta}^{(K)} \sim \pi(\boldsymbol{\eta}|\mathbf{Y}_{obs})$, where K is the number of posterior samples obtained and $\mathbf{Y}_{obs}$ are the all observed ozone levels. Computational improvements could be made by extending the collapsed sampler in Finley et al. (2018) to the space-time setting.

4.4.4 Prediction and Inference

For this problem, we use prediction in two ways: for model selection (predictive comparisons at held-out locations and times in $\mathcal{R}$) and to predict ozone at unmonitored locations and times. Predictive performance at held-out locations and times is used for model selection. Using this model, we predict ozone at unmonitored locations and times to assess respiratory health risks and compliance to legislated ozone thresholds. For either predictive goal (model selection using hold-out data or prediction at unmonitored locations and/or times), consider the location-time pair $(\mathbf{s}, t)$. At this location-time pair, we predict from the posterior predictive distribution,

$$p\left(Y(\mathbf{s},t)|\mathbf{Y}_{obs}\right) = \int p\left(Y(\mathbf{s},t)|\boldsymbol{\eta}\right) p\left(\boldsymbol{\eta}|\mathbf{Y}_{obs}\right) \mathrm{d}\boldsymbol{\eta}, \tag{4.10}$$

using composition sampling (see, e.g., Gelman et al., 2014). Composition sampling provides a Monte Carlo approximation of (4.10), where posterior samples are used to simulate $\sqrt{Y(\mathbf{s},t)^{(k)}}$ from the model $\mathcal{N}\left(\mathbf{x}(\mathbf{s},t)^T\boldsymbol{\beta}^{(k)} + w(\mathbf{s},t)^{(k)}, \tau^{2(k)}\right)$. To obtain ozone predictions $Y(\mathbf{s},t)^{(k)}$, we square posterior predictive samples of square-root ozone. In this regard, predictions for hold-out data and unmonitored sites are obtained similarly; however, they differ in how the conditioning or neighborhood set $N(\mathbf{s},t)$ is specified and how random effects are predicted.

For prediction at an unmonitored location-time pair $(\mathbf{s}, t)$, predictions are made after model fitting on a grid of 201 locations over Mexico City, which we denote $\mathcal{S}_g$. This grid is uniformly distributed over the convex hull of the monitoring network. At

each grid location $\mathbf{s} \in \mathcal{S}_g$, we predict ozone for each hour in April and May. Relative humidity and temperature are not measured over our prediction grid. Prediction or interpolation of $\mathbf{x}(\mathbf{s}, t)$ could be done by utilizing a multivariate NNGP model for ozone, temperature, and relative humidity; however, this would be computationally expensive. For simplicity, we estimate of $\mathbf{x}(\mathbf{s}, t)$ at unmonitored locations using a weighted average of all simultaneously observed covariates with weights proportional to inverse squared distance. Random effects at unmonitored location-time pairs are sampled from a conditional normal distribution,

$$w(\mathbf{s}, t)^{(k)} | \mathbf{w}^{(k)}_{N(\mathbf{s},t)} \sim \mathcal{N}\left(C_{(\mathbf{s},t),N(\mathbf{s},t)} C^{-1}_{N(\mathbf{s},t)} \mathbf{w}^{(k)}_{N(\mathbf{s},t)}, \sigma^2 - C_{(\mathbf{s},t),N(\mathbf{s},t)} C^{-1}_{N(\mathbf{s},t)} C^{\top}_{(\mathbf{s},t;),N(\mathbf{s},t)}\right), \tag{4.11}$$

where $\mathbf{w}^{(k)}_{N(\mathbf{s},t)}$ is a subset of the k^{th} posterior sample of $\mathbf{w}_{\mathcal{R}}$ and the covariance matrices are computed using the k^{th} posterior sample.

Model Selection

Although we fit models on the square-root scale, we compare predictive performance on the original scale. To compare models, we randomly select 20% of the hours in April and May, hold out all observations (i.e. all monitoring stations) at those times, and predict ozone levels at these held-out locations and times (see Section 4.4.3 for details). We include held-out location-time pairs in the reference set. Therefore, conditioning sets are specified as they would be for data in Section 4.4.2, and prediction becomes part of model fitting. In particular, random effects $w(\mathbf{s}, t)$ are sampled within the Gibbs sampler described in Appendix C.2.

This hold-out approach allows us to use both univariate and multivariate criteria. Letting y_i be a held-out observation and n_h be the number of held out data, we

compare models using root mean squared prediction error (RMSPE)

$$\sqrt{\frac{1}{n_h}\sum_i (y_i - \mathbf{E}(Y_i|\mathbf{Y}_{obs}))^2},$$

mean absolute prediction error (MAPE)

$$\frac{1}{n_h}\sum_i |y_i - \mathbf{E}(Y_i|\mathbf{Y}_{obs})|,$$

and $100 \times (1-\alpha)\%$ prediction interval coverage (CVG)

$$\frac{1}{n_h}\sum_i \mathbf{1}\left(y_i > Y_{i,\alpha/2} \;\&\; y_i < Y_{i,1-\alpha/2}\right),$$

where $Y_{i,\alpha/2}$ and $Y_{i,1-\alpha/2}$ represent the end points of a $100 \times (1-\alpha)\%$ prediction interval for the held-out observation y_i and $\mathbf{1}(\cdot)$ is the indicator function. We also use continuous ranked probability scores (CRPS) (Gneiting and Raftery, 2007),

$$\text{CRPS}(F_i, y_i) = \int_{-\infty}^{\infty} (F_i(x) - \mathbf{1}(x \geq y_i))^2 \mathrm{d}x = \mathbf{E}|Y_i - y_i| - \frac{1}{2}\mathbf{E}|Y_i - Y_{i'}|, \quad (4.12)$$

for model comparison using a Monte Carlo approximation of CRPS (see, e.g., Krüger et al., 2016), $\text{CRPS}(\hat{F}_i, y_i) = \frac{1}{K}\sum_{j=1}^{K}|Y_j - y_i| - \frac{1}{2K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}|Y_j - Y_l|$, using K posterior predictions. We average $\text{CRPS}(\hat{F}_i, y_i)$ over held-out data to obtain a single value. Continuous ranked probability scores compare the entire predictive distribution to held-out a value, rather than only the predicted mean (MAPE and RMSPE) or quantiles (CVG). Furthermore, CRPS is a proper scoring rule (Gneiting and Raftery, 2007); thus, we prefer it as a model selection criterion.

To assess predictions for all observations at a held-out hour jointly, we use the energy score (ES), a multivariate generalization of CRPS. For a set of multivariate predictions $\mathbf{Y}$, ES is

$$\text{ES}(\Omega, \mathbf{y}) = \frac{1}{2}E_\Omega \left\|\mathbf{Y} - \mathbf{Y}'\right\|^q - E_\Omega \left\|\mathbf{Y} - \mathbf{y}\right\|^q, \quad (4.13)$$

where $\mathbf{y}$ is an observation, $q \in (0, 2)$, and Ω is the probability measure of the predictive distribution (Gneiting and Raftery, 2007). As is common, we fix $q = 1$ (see Gneiting et al., 2008; Jordan et al., 2017). For K predictions $\{\mathbf{Y}_1, ..., \mathbf{Y}_K\}$ for a held-out observation $\mathbf{y}$, the empirical ES reduces to $\text{ES}(\mathbf{Y}, \mathbf{y}) = \frac{1}{K}\sum_{j=1}^{K} \|\mathbf{Y}_j - \mathbf{y}\| - \frac{1}{2K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} \|\mathbf{Y}_i - \mathbf{Y}_j\|$, as was done in Gneiting et al. (2008). Like CRPS, ES is a proper scoring rule (Gneiting and Raftery, 2007) and is averaged over held-out data.

Inference on Compliance to Air Quality Standards

For compliance, we discuss the probability of exceeding nationally legislated limits, 95 parts per billion (ppb) for *hourly* ozone and 70 ppb for *eight-hour average* ozone (Diario Oficial de la Federación, 2014b). To assess compliance to both legislated thresholds for ozone, we consider one-hour and eight-hour average ozone exceedances together and formally define the unknown exceedance status for an unmonitored location $\mathbf{s} \in \mathcal{S}_g$ and hour t as $E(\mathbf{s}, t) = \mathbf{1}\left(Y(\mathbf{s}, t) > 95 \text{ or } \frac{1}{8}\sum_{i=0}^{7} Y(\mathbf{s}, t-i) > 70\right)$. Estimates of exceedance can be used to assess the probability of exceedance for each grid location on an hourly or daily level. To present estimated exceedance rates in Section 4.5, we summarize exceedances as (i) a function of day, averaging over our prediction grid $\mathcal{S}_g$, and (ii) as a function of space, averaging over the hours in April and May.

When considering exceedances over days, we estimate the proportion of locations in $\mathcal{S}_g$ that exceed Mexican ambient ozone standards at least once on a given day d. For the k^{th} posterior sample, this is computed as

$$\frac{1}{|\mathcal{S}_g|}\sum_{\mathbf{s}\in\mathcal{S}_g} \mathbf{1}\left(\sum_{t\in d} E(\mathbf{s}, t)^{(k)} > 0\right), \tag{4.14}$$

where $|\mathcal{S}_g| = 201$ locations. Using all samples of (4.14), we obtain posterior mean

and credible intervals for the proportion of the city exceeding ozone thresholds at least once that day. To summarize exceedance probabilities spatially, we average over hourly exceedances to estimate the proportion of hours that a given location is non-compliant to Mexican ambient ozone standards,

$$\frac{1}{n_t}\sum_{t=1}^{n}\mathbf{1}\left(E(\mathbf{s},t)^{(k)}>0\right), \tag{4.15}$$

where, as before, the K estimates of (4.15) provide a predictive distribution for the estimated proportion of hours not in compliance for $\mathbf{s}\in\mathcal{S}_g$.

Respiratory Risk

For health outcomes, we use methods in Chiogna and Pauli (2011) that give a daily risk score using hourly ozone levels, thus accounting for health risks due to short-term peaks in ozone due to daily seasonality. In their paper, Chiogna and Pauli (2011) model the short-term effects of summer ozone levels on respiratory hospital admissions, comparing 115 models using cross-validation. Although the risk model of Chiogna and Pauli (2011) does not include space, we apply their approach to each prediction location $\mathbf{s}\in\mathcal{S}_g$ and each day d. Similar to compliance, we assess respiratory health risks over space and time in Section 4.5. Using this approach, we compare respiratory health risks during the peak ozone season to the average risk over the entirety of 2017.

Chiogna and Pauli (2011) found that the best respiratory risk model included three measures associated with the ozone threshold $T = 60$ ppb: number of the hours exceeding T on day d (we call this $H(\mathbf{s},d)$), the difference between the maximum daily ozone level and T (max - threshold) $D(\mathbf{s},d)$, and the lagged average nighttime ozone from the last three days $O_n(\mathbf{s},d)$ (nighttime is defined to be 9 PM and 8 AM). We used fewer days to calculate $O_n(\mathbf{s},d)$ if data from the last three nights were not available (e.g. April 2 only has one night to estimate $O_n(\mathbf{s},d)$). Using these

quantities, the relative risk of respiratory hospital admissions $r(\mathbf{s}, d)$ on day d at location $\mathbf{s}$ is

$$r(\mathbf{s}, d) = 0.864 \exp\left(5.020 \times 10^{-4} H(\mathbf{s}, d) D(\mathbf{s}, d) + 5.714 \times 10^{-3} O_n(\mathbf{s}, d)\right), \quad (4.16)$$

where 5.020×10^{-4} and 5.714×10^{-3} are regression coefficients for $H(\mathbf{s}, d)D(\mathbf{s}, d)$ and $O_n(\mathbf{s}, d)$, respectively. The scaling factor 0.864 makes the average risk at the 24 ozone monitoring stations over 2017 equal to one, preserving the interpretation of relative risk compared to average risk over the year. The scale of the coefficients differs from those presented in Chiogna and Pauli (2011) because we use ozone levels in ppb instead of $\mu g/m^3$. Importantly, we do not observe $H(\mathbf{s}, d)$, $D(\mathbf{s}, d)$, and $O_n(\mathbf{s}, d)$ at unmonitored locations; thus, these must be estimated using posterior predictions over the space-time grid $\mathcal{S}_g \times \{1, .., n_t\}$. For a single set of predictions, we clarify how the quantities $H(\mathbf{s}, d)$, $D(\mathbf{s}, d)$, and $O_n(\mathbf{s}, d)$ are estimated. For the k^{th} set of ozone predictions, we estimate $H(\mathbf{s}, d)^{(k)} = \sum_{t \in d} \mathbf{1}\left(Y(\mathbf{s}, t)^{(k)} > T\right)$, $D(\mathbf{s}, d)^{(k)} = \max_{t \in d}\left(Y(\mathbf{s}, t)^{(k)}\right) - T$, and $O_n(\mathbf{s}, d)^{(k)} = \frac{1}{3}\sum_{i=0}^{2} \frac{1}{11} \sum_{t \in \text{NGT}_{d-i}} Y(\mathbf{s}, t)^{(k)}$, where NGT_d contains the nighttime hours spanning days d and $d-1$ (9 PM on day $d-1$ to 8 AM on day d). These quantities are computed using each of the K sets of posterior predictions to estimate risk $r(\mathbf{s}, d)^{(k)}$ at each location $\mathbf{s} \in \mathcal{S}_g$ and day d in April and May and are used to summarize respiratory health risks over space and time.

Similar to compliance, we assess respiratory health risks over space and time. To summarize spatial patterns, we estimate the mean (4.17) and maximum (4.18) risk,

$$\frac{1}{61} \sum_{d}^{61} r(\mathbf{s}, d)^{(k)} \qquad \text{and} \quad (4.17)$$

$$\max_{d} \left(r(\mathbf{s}, d)^{(k)}\right), \quad (4.18)$$

for each $\mathbf{s} \in \mathcal{S}_g$. We also compute the mean (4.19) and maximum risk (4.20) for each

day, given by

$$\frac{1}{|\mathcal{S}_g|}\sum_{\mathbf{s}\in\mathcal{S}_g} r(\mathbf{s},d)^{(k)} \qquad \text{and} \tag{4.19}$$

$$\max_{\mathbf{s}\in\mathcal{S}_g}\left(r(\mathbf{s},d)^{(k)}\right), \tag{4.20}$$

respectively. Equations (4.17)-(4.20) are computed for each posterior predictive sample and can be used to estimate risks over Mexico City in April and May with uncertainty.

4.5 Results and Discussions

As discussed, we select a covariance model by holding out all observations at 20% of the hours in April and May. This hold-out approach allows us to compare models using the energy score, as well as univariate criteria like CRPS, RMSPE, MAPE, and 90% CVG. While RMSPE and prediction interval coverage are the most common criteria, we rely most upon CRPS and ES because they are proper scoring rules (Gneiting and Raftery, 2007) and compare the whole predictive distribution to held-out values. Predictions are made using 25,000 posterior predictions after a burn-in of 5,000 iterations. Model fitting is done on a single core of an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz server using Rcpp (Eddelbuettel et al., 2011) to integrate C++ functionality into R. Computational improvements are possible through parallelization for linear algebra operations.

We present results for a variety of covariance models on $\mathcal{D}\times\mathbb{S}^1\times\mathbb{R}$ or a subset of this space, given in Table 4.5. We first consider a completely separable covariance model on $\mathcal{D}\times\mathbb{S}^1\times\mathbb{R}$, Model 1, which is used to motivate the need for nonseparable models. We then consider two (space-linear time) nonseparable examples from Gneiting (2002). The first, given as Model 2, ignores seasonality. We also consider this same covariance model multiplied by an exponential covariance function that

takes θ as an argument (Model 3), an example of construction (**A**) in Section 4.3.1. Then, we consider two examples based on the class of space-circular time nonseparable covariance models presented by Shirota et al. (2017) (Models 4 and 5), where Model 4 ignores temporal decay and Model 5 is an example of construction (**B**). Then, we give two examples of nonseparable circular cross linear time covariance models, one from White and Porcu (2018) and the other from Theorem 1, where both models are multiplied by an exponential spatial covariance function (Models 6 and 7). Both Models 6 and 7 are examples of construction (**C**). Other examples of constructions (**A**), (**B**), and (**C**) are considered; however, those we present represent the models with the best predictive performance.

Table 4.5: Model selection results for the Mexico City ozone data. Parentheses in the "domain" column indicate that those quantities are nonseparable in the covariance model. The lowest ES and CRPS are bolded.

Model	Equation	Domain	ES	CRPS	MAPE	RMSPE	90% CVG
1	$\left(\mathcal{M}_{\alpha_k,1/2}\right)^3$	$\mathcal{D}\times\mathbb{S}^1\times\mathbb{R}$	21.122	3.466	4.640	6.463	0.920
2	(Example 1)	$(\mathcal{D}\times\mathbb{R})$	24.101	4.008	5.376	7.222	0.934
3	(Example 1) $\times\ \mathcal{M}_{\alpha,1/2}$	$(\mathcal{D}\times\mathbb{R})\times\mathbb{S}^1$	22.044	3.638	4.890	6.738	0.924
4	(Example 2)	$(\mathcal{D}\times\mathbb{S}^1)$	25.673	4.288	5.820	7.846	0.922
5	(Example 2) $\times\ \mathcal{M}_{\alpha,1/2}$	$(\mathcal{D}\times\mathbb{S}^1)\times\mathbb{R}$	19.829	3.235	4.316	6.073	0.909
6	$\mathcal{M}_{\alpha,1/2}\times$ (Example 3)	$\mathcal{D}\times(\mathbb{S}^1\times\mathbb{R})$	21.049	3.461	4.631	6.436	0.926
7	$\mathcal{M}_{\alpha_1,1/2}\times$ (Example 4)	$\mathcal{D}\times(\mathbb{S}^1\times\mathbb{R})$	**19.334**	**3.159**	4.183	5.881	0.922

All models except Model 4 use the neighbor selection described in Section 4.4.2, conditioning on at most 42 neighbors. For Model 4, we exclude lags 24 and 168 at the location of the observation to give positive-definite covariance matrices and instead use lags 3, 167, and 169 to compensate. The predictive performance of each model are given in Table 4.5.

The results highlight the importance of modeling the entire space ($\mathcal{D}\times\mathbb{S}^1\times\mathbb{R}$) and utilizing nonseparable models. Models that exclude one of these subspaces (Models 2 and 4) have the worst predictive performance. The next worse model (Model 3) used construction (**A**), nonseparability in space and linear time. The completely separable model (Model 1) was slightly better in prediction than Model 3 but worse than models

with nonseparability on circular time and another domain (constructions (**B**) and (**C**)). Our variation of Shirota et al. (2017) with space-circular time nonseparability (Model 5 using construction (**B**)) was second best in prediction. Ultimately, we use Model 7, derived from Theorem 1 and using construction (**C**), for our analysis because it performed best in terms of ES, CRPS, MAPE, and RMSPE on our hold-out data.

Explicitly, the final covariance model is

$$C(h,\theta,u) = \exp\left\{\exp\left[-\left(\frac{u}{c_{t_1}}\right)^{\alpha}\right]\cos(\theta) - \frac{u}{c_{t_2}} - \frac{h}{c_s} - 1\right\} \times \cos\left\{\exp\left[-\left(\frac{u}{c_{t_1}}\right)^{\alpha}\right]\sin(\theta)\right\}, \tag{4.21}$$

where $(h,\theta,|u|) \in [0,\infty) \times [0,\pi] \times [0,\infty)$, $c_{t_1}, c_{t_2}, c_s > 0$ and $\alpha \in (0,2]$. The parameters c_{t_1} and c_{t_2} govern temporal decay of the covariance function. While neither c_{t_1} nor c_{t_2} are uniquely interpretable, c_{t_2} assures that the covariance decays to 0 as u gets large. c_s is the spatial range parameter, and α is a smoothness parameter. The parameter c_s is most accurately understood as the spatial range for simultaneous observations. One benefit of this model is that it has few parameters compared with some of its competitors.

Posterior summaries for regression coefficients and covariance parameters are given in Table 4.6. The regression parameters agree with our hypotheses that relative humidity is negatively related to ozone levels and the temperature is positively associated with ozone. The mean of the spatial range parameter c_s is 22.27 km. The posterior means of the temporal range parameters are both around 200 hours, meaning that the correlation for this model persists over many days.

Table 4.6: Posterior summaries for model parameters

	Mean	Standard Deviation	2.5%	97.5%
β_0	6.1643	0.2837	5.7134	6.6673
β_{RH}	-0.0226	0.0012	-0.0244	-0.0200
β_{TMP}	0.1025	0.0066	0.0890	0.1139
c_{t_1}	182.2633	6.6341	169.8709	196.6384
c_{t_2}	214.1469	12.1319	191.2321	239.3512
c_s	22.2692	0.5536	21.2732	23.4618
α	0.6084	0.0075	0.5949	0.6235
τ^2	0.1024	0.0019	0.0985	0.1059
σ^2	1.9895	0.0417	1.9093	2.0705
$\sigma^2/(\sigma^2+\tau^2)$	0.9510	0.0014	0.9484	0.9537

Using posterior predictions for April and May 2017, we assess probabilities of exceeding Mexican ambient air quality standards and examine respiratory risk rates during Mexico City's peak ground-level ozone season. To summarize exceedance results temporally, we plot posterior means and 95% credible intervals for the estimated proportion of the city that exceeds national ozone standards at least once on that day in Figure 4.5. These quantities are computed using (4.14). In April, there are several peaks in exceedance probability; however, from May 6 to May 28, the proportion of the city with at least one daily ozone exceedance is near one.

For each predictive location, we plot the estimated proportion of hours when Mexican ground-level ozone standards (either one-hour or eight-hour ozone) are exceeded, using (4.15), in Figure 4.6a. We estimate that locations in southern Mexico City exceed national ozone regulations nearly three times as often as north, west, and central Mexico City.

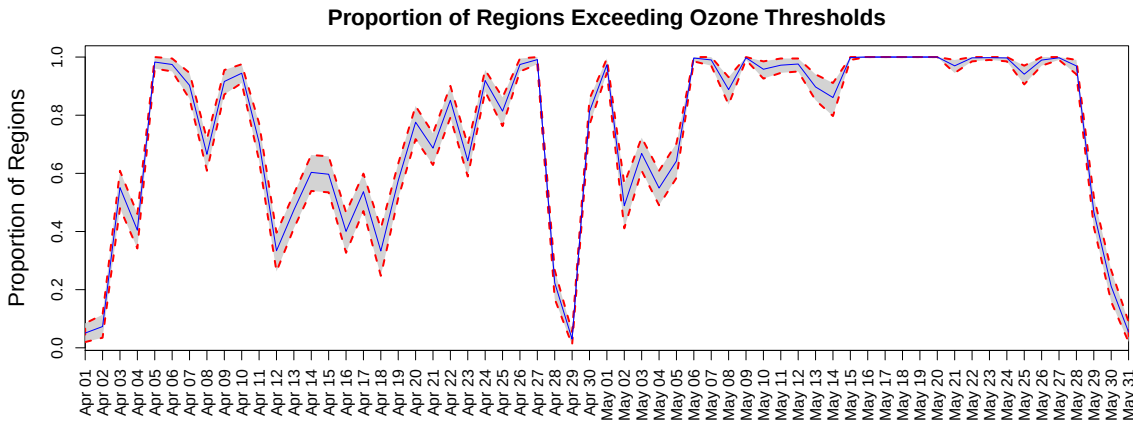
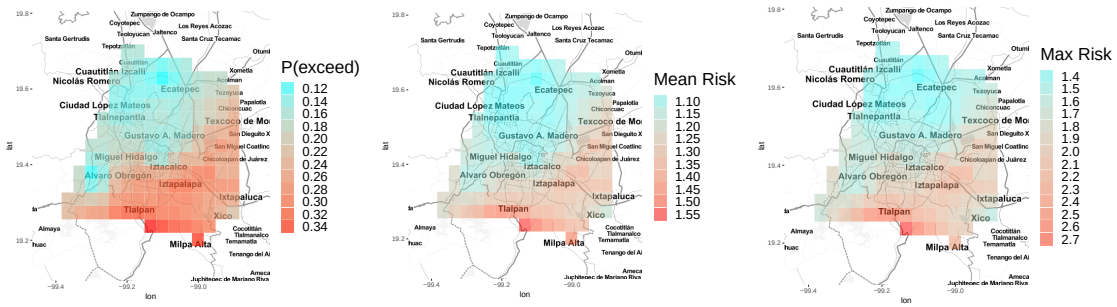


FIGURE 4.5: Estimated proportion of the city that exceeds either one-hour and eight-hour average ozone limits at least once that day. These are computed using (4.14).

In Figures 4.6b, we give the mean relative risk averaged over all days in April and May using (4.17), and we plot the estimated maximum relative risk over the same time period, calculated using (4.18), in Figure 4.6c. Although the spatial patterns of the means and maxima are similar, the scale of these plots differs significantly. In terms of mean risk, the most extreme regions have respiratory risks about 50% higher than the least extreme regions. In contrast, regions of the highest maximum risk in the peak season are nearly twice that of regions with the lowest maximum risk. These plots demonstrate the degree of increased respiratory risk determined solely by where one lives or works.

The estimated mean and maximum respiratory health risk over Mexico City (with 95% credible intervals) are computed using (4.19) and (4.20). These are plotted as a function of the day in Figure 4.7. Ozone risk peaks May 17 to May 23 in terms of both the mean and maximum, but the changes in the maximum risk are much more drastic than those in mean risk. These maxima correspond to extreme ozone levels in south Mexico City where the estimated risk is 2.7 times the annual average and

nearly two times the mean risk on the same day.



(a) Exceedance probability (b) Mean risk (c) Maximum risk

FIGURE 4.6: Spatial summaries of (Left) exceedance probability computed using (4.14), and (Center and Right) respiratory risk for ozone, using (4.17) and (4.18). The exceedance probability considers both one-hour and eight-hour average ozone limits.

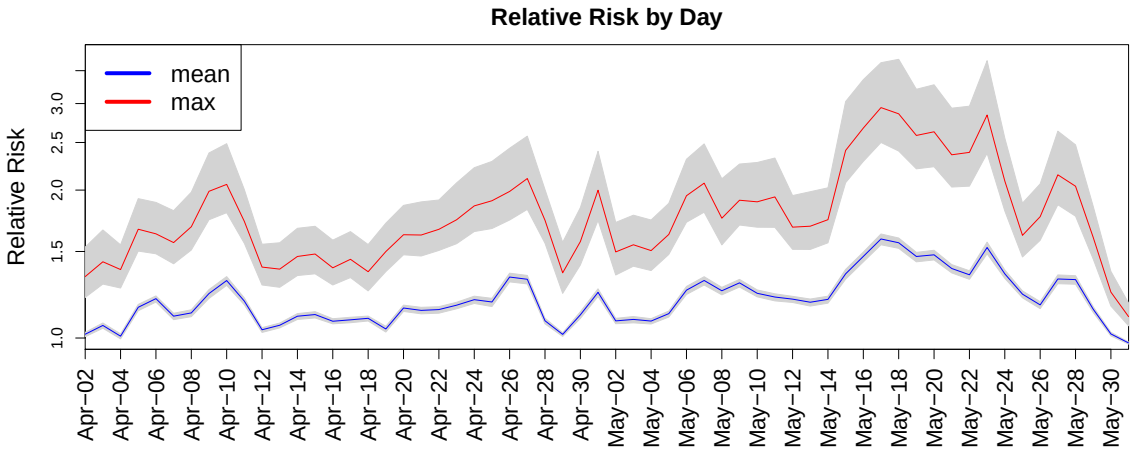


FIGURE 4.7: Estimated mean and maximum respiratory risk, computed using (4.19) and (4.20).

These results highlight how extreme the space-time variability in both exceedance probability and respiratory health risk is within Mexico City during its peak ozone

season. According to the risk model from Chiogna and Pauli (2011), those living in regions with extreme ozone levels are about 50% more likely to be admitted to the hospital due to ozone exposure compared to those in less polluted areas. Although this is not a component of the risk model, increased respiratory risk impacts at-risk populations (e.g. the elderly) at a much higher rate than healthier sub-populations (see, e.g., Bell et al., 2004).

4.6 Discussion and Conclusions

We have discussed Mexico City's ozone monitoring network and analyzed ozone levels from April and May of 2017. For these data, we discuss existing classes of covariance functions and develop new classes of covariance models for circular time, linear time, and spatial patterns in the data. We apply covariance models to the Mexico City ozone data using nearest-neighbor Gaussian processes and discuss how neighbors are selected when seasonality is present. We select a model using predictive criteria on a hold-out dataset and use this model for prediction to assess compliance with Mexican ambient air quality standards and respiratory risk due to ozone exposure. In this analysis, we identify regions and times where exceedance probabilities and respiratory risk are particularly high.

In this paper, we do not forecast pollution levels and instead use both past and future lags for prediction over Mexico City. In many settings, however, forecasting respiratory risk and compliance at unmonitored sites may be an important goal. To forecast ozone levels, conditioning sets for prediction at arbitrary location-time pairs must be limited to past observations in the reference set. Our model fitting is currently too slow to be used to forecast in real time; however, future work could adapt computational improvements for spatial modeling suggested by Finley et al. (2018), Guinness (2018), and Katzfuss and Guinness (2017) to space-time modeling.

We also note that by adopting the risk model of Chiogna and Pauli (2011) that

(i) we assume that the effects of ozone in Milan, Italy are like those in Mexico City, Mexico and (ii) we adopt their choice of a threshold where ozone begins to be harmful, while it should be noted that several question using thresholds like this altogether (See, e.g., Kim et al., 2004). Noting these limitations, this model enables us to estimate respiratory risk on a daily level using predicted hourly ozone levels. Another potential weakness of our model is that inference and predictions in this manuscript are limited to 2017. Accordingly, an interesting follow-up study could assess how exceedance probabilities and risk assessments vary year-to-year.

Here, we have considered models that are partially nonseparable over $\mathcal{D} \times \mathbb{S}^1 \times \mathbb{R}$ but not over all three spaces, which could motivate future work. Additional theoretical work could address how the constraints on the functions φ and ψ in Shirota et al. (2017) can be relaxed while preserving positive-definiteness.

5

Conclusions

In this dissertation, I focused on predictive spatial and spatiotemporal modeling in three settings with applications in environmental exposure. For each topic, we modeled in a Bayesian framework because it provides a natural framework for prediction that allows rigorous uncertainty quantification. This is particularly advantageous when inferring the status of functions of predicted values. In these situations, asymptotic results (e.g. the delta method) are unlikely to provide adequate approximations. Therefore, Bayesian modeling is preferred as it allows principled and accurate predictive inference.

In Chapter 2, we considered the goal of analyzing areal unit data that are partially observed. First, we discussed modeling approaches for prediction for partially observed areal unit data. To illustrate this goal, we provided examples in image reconstruction and testing semiconducting chips. Our second contribution used the ideas of predicting unobserved areal units for model comparison. In some cases, minimizing an out-of-sample predictive criterion may be desired, but customarily modeling areal unit data comes with the goal of spatial smoothing, employing a complete dataset over all areal units. In these settings, minimizing prediction error

is rarely the goal of interest. To illustrate this, we provided an example of selecting a conditional autoregressive model for lung cancer rates in Ohio. Discussion on model comparison for smoothing is limited, and this discussion contributed to this gap in the literature. In addition, we applied our model comparison discussion to our analyses of semiconducting chips. In this case, the goal was minimizing predictive error, differing from our discussion on selecting a smoother.

In Chapter 3, we jointly modeled coarse particulate matter and ozone levels jointly to predict pollution emergencies, as defined by the Mexico City's Atmospheric Environmental Contingency Program, and compliance to Mexican ambient air quality standards. Because pollution emergencies are defined in terms of five regional maxima of ozone and coarse particulate matter levels, we must forecast these maxima of correlated random variables. Because these regional maxima come from relatively few stations, we used a multivariate discrete-time, discrete-space model to forecast pollutant levels at all 24 sites rather than extreme value models for the maxima. These pollution forecasts are then used to project emergency status and to predict compliance to nationally legislated standards.

In Chapter 4, we considered ground-level ozone levels during Mexico City's peak ozone season (April and May). Here, we focused on assessing respiratory health risks attributable to ozone and compliance to nationally legislated ozone standards at locations and times where ozone levels were not monitored. Because spatiotemporal prediction was the primary goal, we worked in continuous space-time. To model these data, we provided three primary contributions: First, we accounted for daily seasonality through covariance modeling rather than through terms in the mean function of the model. For this approach, we introduced new classes of appropriate covariance functions. Second, we discussed appropriate Vecchia approximations for covariance functions with seasonality within the nearest-neighbor Gaussian process framework. Our modeling approach allowed scalable model fitting, prediction, and

inference for the Mexico City ozone data. Third, we used this model to assess respiratory risks and compliance with Mexican ambient air quality standards at locations where ozone is not monitored, a contribution that is more scientific than statistical.

Appendix A

Appendix for Prediction and Model Comparison for Areal Unit Data

A.1 Generative Model for Binary Semiconductor Chip Data

Here, we present the simulation method for the binary variable used in Section 2.2. As in A.2, we draw a uniform random variable for each site $p_{ijk} \sim \text{Uniform}(0,1)$ that is used to generate the response. Let $r_k = \sqrt{x_{1k}^2 + x_{2k}^2}$ be the distance from the center of the semiconductor chip (distances are in units of 100 μm). Using μ_{ijk} from A.2, we generate a latent random variable Z_{ijk}^*

$$Z_{ijk}^* = \mu_{ijk} + U_k + V_{ik} + W_{ijk} + \epsilon_{ijk}, \tag{A.1}$$

where

$$\begin{aligned}
\mu_{ijk} &= \beta_0 + \alpha_i + \gamma_j + D_{ijk} r_k + (1 - D_{ijk}) r_k/2 \qquad \text{(A.2)} \\
\alpha_i &\sim N(0, 400) \\
\gamma_j &\sim N(0, 100) \\
D_{ijk} &= \mathbf{1}(x < 0, y > 0, p_{ijk} < 0.8, r_k \in (50, 130)) \\
U &\sim CAR(\tau_U^2)
\end{aligned}$$

$$V_i \overset{iid}{\sim} CAR(\tau_V^2)$$

$$W_{ij} \overset{iid}{\sim} CAR(\tau_W^2)$$

$$\epsilon_{ijk} \overset{iid}{\sim} N(0,1),$$

where the CAR models are Besag (1974) CAR models with a sum-to-zero constraint to center the distribution. Then, we set $\tau_U^2 = 100$, $\tau_V^2 = 64$, and $\tau_W^2 = 1$. We then center and scale Z_{ijk}^*, and define this to be Z_{ijk}. Then, we draw a random binary variable Y_{ijk} using a probit specification (i.e. $P(Y_{ijk} = 1|Z_{ijk}) = \Phi(Z_{ijk})$). Again, this generates wafers that generally have the highest probability of failure in the upper left quadrant at distances between 5 and 13 mm from the center of the wafer. Then, we added lot and wafer specific spatial noise and pure error to the latent surface.

A.2 Generative Model for Continuous Semiconductor Chip Data

Here, we present the simulation method for the continuous variable used in Section 2.5. We first present the GP simulation because the CAR simulation uses components from this simulation. We draw a uniform random variable for each site $p_{ijk} \sim$ Uniform$(0,1)$ that is used to generate the response. Let $r_k = \sqrt{x_{1k}^2 + x_{2k}^2}$ be the distance from the center of the semiconductor chip (distances are in units of 100 μm). We simulate a normal random variable Y_{ijk} of the following form:

$$Y_{ijk} = \mu_{ijk} + U_k + V_{ik} + W_{ijk} + \epsilon_{ijk}, \tag{A.3}$$

where

$$\mu_{ijk} = \beta_0 + \alpha_i + \gamma_j + D_{ijk} r_k + (1 - D_{ijk}) r_k/2 \tag{A.4}$$

$$\beta_0 = 100$$

$$\alpha_i \sim N(0, 400)$$

$$
\begin{aligned}
\gamma_j &\sim N(0, 100) \\
D_{ijk} &= \mathbf{1}(x < 0, y > 0, p_{ijk} < 0.8, r_k \in (50, 130)) \\
U &\sim N(0, \sigma_U^2 H_U(\phi_U)) \\
V_i &\overset{iid}{\sim} N(0, \sigma_V^2 H_V(\phi_V)) \\
W_{ij} &\overset{iid}{\sim} N(0, \sigma_W^2 H_W(\phi_W)) \\
\epsilon_{ijk} &\overset{iid}{\sim} N(0, 1)
\end{aligned}
$$

and all correlation matrices H are defined are generated using exponential covariance functions with $\phi_U = \phi_V = \phi_W = 3/\max(d)$, $\sigma_U^2 = 100$, $\sigma_V^2 = 25$, and $\sigma_W^2 = 9$. Effectively, this generates wafers that that peak in the upper left quadrant at distances between 5 and 13 mm from the center of the wafer. Then, we added lot and wafer specific spatial noise and pure error to the generated surface.

A.3 Full Conditional Distributions

A.3.1 Normal CAR model

The model that we consider in this case is

$$
\begin{aligned}
Y_i|\boldsymbol{\beta}, \mathbf{V}, \sigma^2, \tau^2 &\overset{ind}{\sim} N(\mathbf{x}_i^T\boldsymbol{\beta} + V_i, \sigma^2) \\
\mathbf{V}|\tau^2 &\sim \text{CAR}(0, \tau^2)
\end{aligned}
\qquad
\begin{aligned}
\boldsymbol{\beta} &\sim N(m_\beta, s_\beta^2) \\
\tau^2 &\sim IG(a_\tau, b_\tau) \\
\sigma^2 &\sim IG(a_\sigma, b_\sigma).
\end{aligned}
\tag{A.5}
$$

For the full conditional distributions under this model, we define a binary variable $o_i \in \{0, 1\}$, where $o_i = 1$ if the i^{th} die is observed. The full distributions are

$$
\begin{aligned}
\boldsymbol{\beta}|\mathbf{V}, \sigma^2, \tau^2, \mathbf{Y}_{\text{obs}} &\sim N(V_\beta^* m_\beta^*, V_\beta^*) \\
V_i|\boldsymbol{\beta}, \sigma^2, \tau^2, \mathbf{Y}_{\text{obs}}, V_{j\neq i} &\sim N(V_{V_i}^* m_{V_i}^*, V_{V_i}^*) \\
\sigma^2|\mathbf{V}, \boldsymbol{\beta}, \tau^2, \mathbf{Y}_{\text{obs}} &\sim IG(a_\sigma^*, b_\sigma^*) \\
\tau^2|\mathbf{V}, \boldsymbol{\beta}, \sigma^2, \mathbf{Y}_{\text{obs}} &\sim IG(a_\tau^*, b_\tau^*)
\end{aligned}
\tag{A.6}
$$

where

$$
\begin{aligned}
V_\beta^* &= (\mathbf{X}_{\text{obs}}^T \mathbf{X}_{\text{obs}}/\sigma^2 + V_\beta^{-1})^{-1} & a_\tau^* &= a_\tau + n/2 \\
m_\beta^* &= \mathbf{X}_{\text{obs}}^T (\mathbf{y}_{\text{obs}} - \mathbf{V}_{\text{obs}})/\sigma^2 + V_\beta^{-1} m_\beta & b_\tau^* &= b_\tau + \frac{1}{2}\mathbf{V}^T (D_w - W)\mathbf{V} \\
V_{V_i}^* &= \left(o_i \frac{1}{\sigma^2} + \frac{w_{i+}}{\tau^2} \right)^{-1} & a_\sigma^* &= a_\sigma + \frac{1}{2}\sum_i o_i \\
m_{V_i}^* &= o_i \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma^2} \right) + \frac{1}{\tau^2} \sum_{j \sim i} w_{ij} V_j & b_\sigma^* &= b_\sigma + \frac{1}{2}\sum_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta} - V_i)^2 o_i
\end{aligned}
\tag{A.7}
$$

In the case that sampler mixes to slowly, the updates for $\mathbf{V}$ can be made jointly.

A.3.2 Multivariate CAR model

In this case, we take Y_i is multivariate normal and conditionaly independent given V_i, and V_i takes the form of a multivariate CAR model. The model is defined as

$$
\begin{aligned}
Y_i | \boldsymbol{\beta}, \boldsymbol{\phi}, \Sigma, \tau^2 &\sim N(\mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{V}_i, \Sigma) \\
\mathbf{V} | \boldsymbol{\Lambda} &\sim \text{MCAR}(0, \boldsymbol{\Lambda})
\end{aligned}
\qquad
\begin{aligned}
\boldsymbol{\beta} &\sim N(m_\beta, V_\beta) \\
\Sigma^{-1} &\sim \text{Wishart}(\mathbf{V}_\Sigma, \nu_\Sigma) \\
\Lambda^{-1} &\sim \text{Wishart}(\mathbf{V}_\Lambda, \nu_\Lambda),
\end{aligned}
\tag{A.8}
$$

where MCAR is the multivariate CAR prior. Like the simple Gaussian case, this model has Gibbs updates for model fitting.

Let $\mathbf{V} = \begin{pmatrix} V_1^T \\ \vdots \\ V_n^T \end{pmatrix}$ be the $n \times q$ matrix of all spatial random effects. The full conditionals distributions under the MCAR model is

$$
\begin{aligned}
\boldsymbol{\beta} | \cdots &\sim N(V_\beta^* m_\beta^*, V_\beta^*) \\
\mathbf{V}_i | \cdots &\sim N(V_{V_i}^* m_{V_i}^*, V_{V_i}^*) \\
\Sigma^{-1} | \cdots &\sim \text{Wishart}(\mathbf{V}_\Sigma^*, \nu_\Sigma^*) \\
\Lambda^{-1} | \cdots &\sim \text{Wishart}(\mathbf{V}_\Lambda^*, \nu_\Lambda^*),
\end{aligned}
\tag{A.9}
$$

where

$$
\begin{aligned}
V_\beta^* &= \left(\sum_{i=1}^{n} \mathbf{X}_i \boldsymbol{\Sigma}^{-1} \mathbf{X}_i^T o_i + \mathbf{V}_\beta^{-1}\right)^{-1} & \mathbf{V}_\Sigma^* &= V_\Sigma^{-1} + \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \boldsymbol{\beta} - \mathbf{V}_i)(Y_i - \mathbf{X}_i^T \boldsymbol{\beta} - \mathbf{V}_i)^T \\
m_\beta^* &= \sum_{i=1}^{n} \mathbf{X}_i (Y_i - \mathbf{V}_i) o_i + \mathbf{V}_\beta^{-1} \mathbf{m}_\beta & \nu_\Sigma^* &= \nu_\Sigma + n_o \\
V_{V_i}^* &= \left(\boldsymbol{\Sigma}^{-1} o_i + \boldsymbol{\Lambda}^{-1} w_{i+}\right)^{-1} & \mathbf{V}_\Lambda^* &= V_\Lambda^{-1} + \mathbf{V}^T (D_W - W) \mathbf{V} \\
m_{V_i}^* &= \boldsymbol{\Sigma}^{-1} (Y_i - X_i^T \boldsymbol{\beta}) o_i + \Lambda^{-1} \sum_{j \sim i} w_{ij} \mathbf{V}_j & \nu_\Lambda^* &= n + \nu_\Lambda.
\end{aligned}
\tag{A.10}
$$

A.3.3 Nested CAR model

The model that we consider in this case is

$$
\begin{aligned}
Y_{ijk} | \boldsymbol{\beta}, \mathbf{U}, \mathbf{V}, \mathbf{W}, \sigma^2 &\overset{ind}{\sim} N(\mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{U}_k + \mathbf{V}_{ik} + \mathbf{W}_{ijk}, \sigma^2) \\
\mathbf{U} | \tau_U^2 &\sim \text{CAR}(0, \tau_U^2) \\
\mathbf{V}_i | \tau_U^2 &\overset{ind}{\sim} \text{CAR}(0, \tau_V^2) \\
\mathbf{W}_i | \tau_W^2 &\overset{ind}{\sim} \text{MCAR}(0, \tau_W^2, T) \\
\boldsymbol{\beta} &\sim N(m_\beta, V_\beta)
\end{aligned}
\qquad
\begin{aligned}
\tau_U^2 &\sim IG(a_{\tau_U}, b_{\tau_U}) \\
\tau_V^2 &\sim IG(a_{\tau_V}, b_{\tau_V}) \\
\tau_W^2 &\sim IG(a_{\tau_W}, b_{\tau_W}) \\
\sigma^2 &\sim IG(a_\sigma, b_\sigma).
\end{aligned}
\tag{A.11}
$$

We define a binary variable $o_{ijk} \in \{0, 1\}$, where $o_{ijk} = 1$ if the k^{th} die on the j^{th} wafer in the i^{th} core is observed. Additionally, let K indicate the number of dies on each wafers, J_i be the number of wafers in each lot, and I be the number of lots. For this problem, let $\phi_{ijk} = U_k + V_{ik} + W_{ijk}$ be the total spatial random effect. Again, let the subscript "obs" indicated observed quantities. Note that w_{kl} refers to elements of the proximity matrix W_N (maybe we rename this P to avoid confusion with W_{ijk}).

$$
\begin{aligned}
\boldsymbol{\beta} | \sigma^2, \tau^2, U, V, W, \mathbf{Y}_{\text{obs}} &\sim N(V_\beta^* m_\beta^*, V_\beta^*) \\
\sigma^2 | \boldsymbol{\beta}, \tau^2, U, V, W, \mathbf{Y}_{\text{obs}} &\sim IG(a_\sigma, b_\sigma) \\
\tau_U^2 | \boldsymbol{\beta}, \sigma^2, U, V, W, \mathbf{Y}_{\text{obs}} &\sim IG(a_{\tau_U}, b_{\tau_U}) \\
\tau_V^2 | \boldsymbol{\beta}, \sigma^2, U, V, W, \mathbf{Y}_{\text{obs}} &\sim IG(a_{\tau_V}, b_{\tau_V})
\end{aligned}
\tag{A.12}
$$

$$\tau_W^2|\boldsymbol{\beta},\sigma^2,U,V,W,\mathbf{Y}_{\text{obs}} \sim IG(a_{\tau_W},b_{\tau_W})$$

$$W_{ijk}|\boldsymbol{\beta},\sigma^2,\tau^2,U,V_i,W_{ijk'\neq k},\mathbf{Y}_{\text{obs}} \sim N(V^*_{W_{ijk}}m^*_{W_{ijk}},V^*_{W_{ijk}})$$

$$V_{ik}|\boldsymbol{\beta},\sigma^2,\tau^2,U,V_{ik'\neq k},W_i,\mathbf{Y}_{\text{obs}} \sim N(V^*_{V_{ik}}m^*_{V_{ik}},V^*_{V_{ik}})$$

$$U_k|\boldsymbol{\beta},\sigma^2,\tau^2,U_{k'\neq k},V,W,\mathbf{Y}_{\text{obs}} \sim N(V^*_{U_k}m^*_{U_k},V^*_{U_k})$$

where

$$\begin{aligned}
V^*_\beta &= (\sum_{ijk}\mathbf{x}_{ijk}\mathbf{x}_{ijk}^T o_{ijk}/\sigma^2 + V_\beta^{-1})^{-1} \\
m^*_\beta &= \sum_{i,j,k}\mathbf{x}_{ijk}(y_{ijk}-\phi_{ijk})o_{ijk}/\sigma^2 + V_\beta^{-1}m_\beta \\
V^*_{U_k} &= \left(\frac{\sum_{i,j}o_{ijk}}{\sigma^2}+\frac{w_{k+}}{\tau_U^2}\right)^{-1} \\
m^*_{U_k} &= \sum_{i,j}o_{ijk}\left(\frac{y_{ijk}-\mathbf{x}_{ijk}^T\boldsymbol{\beta}-V_{ik}-W_{ijk}}{\sigma^2}\right)+ \\
&\quad \frac{1}{\tau^2}\sum_{k'\sim k}w_{kk'}U_{k'} \\
V^*_{V_{ik}} &= \left(\frac{\sum_j o_{ijk}}{\sigma^2}+\frac{w_{k+}}{\tau_V^2}\right)^{-1} \\
m^*_{V_{ik}} &= \sum_j o_{ijk}\left(\frac{y_{ijk}-\mathbf{x}_{ijk}^T\boldsymbol{\beta}-U_k-W_{ijk}}{\sigma^2}\right)+ \\
&\quad \frac{1}{\tau^2}\sum_{k'\sim k}w_{kk'}V_{ik'} \\
V^*_{W_{ijk}} &= \left(\frac{o_{ijk}}{\sigma^2}+\frac{w_{k+}}{\tau_W^2}\right)^{-1},\ \text{if } \delta=0 \\
m^*_{W_{ijk}} &= o_{ijk}\left(\frac{y_{ijk}-\mathbf{x}_{ijk}^T\boldsymbol{\beta}-U_k-V_{ik}}{\sigma^2}\right)+ \\
&\quad \frac{1}{\tau^2}\sum_{k'\sim k}w_{kk'}W_{ijk'},\ \text{if } \delta=0 \\
a^*_{\tau_U} &= a_{\tau_U}+\frac{K}{2} \\
b^*_{\tau_U} &= b_{\tau_U}+\frac{1}{2}U^T(D_w-W_N)U \\
a^*_{\tau_V} &= a_{\tau_V}+\frac{K\times I}{2} \\
b^*_{\tau_V} &= b_{\tau_V}+\frac{1}{2}\sum_i V_i^T Q V_i \\
a^*_{\tau_W} &= a_{\tau_W}+\frac{K\times\sum_i J_i}{2} \\
b^*_{\tau_W} &= b_{\tau_W}+\frac{1}{2}\sum_i W_i^T(Q\otimes T^{-1}(\delta))W_i \\
b^*_{\tau_W} &= b_{\tau_W}+\frac{1}{2}\sum_{i,j} W_{i,j}^T Q W_i,\ \text{if } \delta=0 \\
a^*_\sigma &= a_\sigma+\frac{K\times\sum_i J_i}{2} \\
b^*_\sigma &= b_\sigma+\frac{1}{2}\sum_{i,j,k}(y_{ijk}-\mathbf{x}_{ijk}^T\boldsymbol{\beta}-\phi_{ijk})^2 o_{ijk}
\end{aligned} \tag{A.13}$$

If $\delta \neq 0$, then $m^*_{W_{ijk}}$ and $V^*_{W_{ijk}}$ take on a more complicated form:

$$V^*_{W_{ijk}} = \left(\frac{o_{ijk}}{\sigma^2}+\frac{T_{jj}^{-1}w_{k+}}{\tau_W^2}\right)^{-1} \tag{A.14}$$

$$m^*_{W_{ijk}} = o_{ijk}\left(\frac{y_{ijk}-\mathbf{x}_{ijk}^T\boldsymbol{\beta}-U_k-V_{ik}}{\sigma^2}\right)+$$

$$\frac{1}{\tau^2}\left[\sum_{j'=1}^{J} T_{j'j}^{-1} \sum_{k' \sim k} w_{k'k} W_{ij'k'} - w_{k+} \sum_{j' \neq j} T_{j'j}^{-1} W_{ij'k}\right].$$

A.3.4 Nested GP model

For this section, the full conditional distributions are for the nested Gaussian process model. The full conditional distributions are of the following form.

$$\begin{aligned}
\boldsymbol{\beta}|\cdots &\sim N(V_\beta^* m_\beta^*, V_\beta^*) \\
U|\cdots &\sim N(V_U^* m_U^*, V_U^*) \\
V_i|\cdots &\sim N(V_{V_i}^* m_{V_i^*}, V_{V_i}^*) \\
W_i|\cdots &\sim N(V_{W_i}^* m_W^*, V_{W_i}^*) \\
\sigma_U^2|\cdots &\sim \text{IG}(a_{\sigma_U}^*, b_{\sigma_U}^*) \\
\sigma_V^2|\cdots &\sim \text{IG}(a_{\sigma_V}^*, b_{\sigma_V}^*) \\
\sigma_W^2|\cdots &\sim \text{IG}(a_{\sigma_W}^*, b_{\sigma_W}^*) \\
\tau^2|\cdots &\sim \text{IG}(a_\tau^*, b_\tau^*),
\end{aligned} \tag{A.15}$$

where

$$\begin{aligned}
m_\beta^* &= \frac{1}{\tau^2}\sum_i \sum_j \mathbf{X}_{ij}^T(\mathbf{y}_{ij} - \mathbf{A}\boldsymbol{\phi}_{ij}) \\
V_\beta^* &= \left(\frac{1}{s_\beta^2} + \frac{I \times J(\mathbf{X}^T\mathbf{X})}{\tau^2}\right)^{-1} \\
m_U^* &= \frac{1}{\tau^2}\mathbf{A}^T \sum_i \sum_j (\mathbf{y}_{ij} - \mathbf{A}(V_i + W_{ij})) \\
V_U^* &= \left(H_U(\phi_U)^{-1}/\sigma_U^2 + \frac{I \times J(\mathbf{A}^T\mathbf{A})}{\tau^2}\right)^{-1} \\
m_{V_i^*} &= \frac{1}{\tau^2}\mathbf{A}^T \sum_j (\mathbf{y}_{ij} - \mathbf{A}(V_i + W_{ij})) \\
V_{V_i^*} &= \left(H_V(\phi_V)^{-1}/\sigma_V^2 + \frac{J(\mathbf{A}^T\mathbf{A})}{\tau^2}\right)^{-1}
\end{aligned} \tag{A.16}$$

$$m_{W_i^*} = \frac{1}{\tau^2}(1_J \otimes \mathbf{A})^T(\mathbf{y}_i - (1_J \otimes \mathbf{A})(1_J \otimes (V_i + U))$$

$$V_{W_i^*} = \left(\frac{T(\delta)^{-1} \otimes H_W(\phi_W)^{-1}}{\sigma_W^2} + \frac{(1_J \otimes \mathbf{A})^T(1_J \otimes \mathbf{A})}{\tau^2}\right)^{-1}$$

$$a^*_{\sigma_U} = a_{\sigma_U} + \frac{m \times n_{ij}}{2}$$

$$b^*_{\sigma_U} = b_{\sigma_U} + \frac{1}{2}U^T H_U^{-1}(\phi_U)U$$

$$a^*_{\sigma_V} = a_{\sigma_V} + \frac{m\sum_j n_{ij}}{2}$$

$$b^*_{\sigma_V} = b_{\sigma_V} + \frac{1}{2}\sum_i V_i^T H_V^{-1}(\phi_V)V_i$$

$$a^*_{\sigma_W} = a_{\sigma_W} + \frac{m\sum_i\sum_j n_{ij}}{2}$$

$$b^*_{\sigma^*_W} = b_{\sigma_W} + \frac{1}{2}\sum_i W_i^T(T^{-1}(\delta) \otimes H_W^{-1}(\phi_W))W_i$$

$$a^*_\tau = a_\tau + \frac{1}{2}\sum_i\sum_j n_{ij}$$

$$b^*_\tau = b_\tau + \frac{1}{2}\sum_i\sum_j(\mathbf{y}_{ij} - X_{ij}\boldsymbol{\beta} - \mathbf{A}\boldsymbol{\phi}_{ij})^T(\mathbf{y}_{ij} - X_{ij}\boldsymbol{\beta} - \mathbf{A}\boldsymbol{\phi}_{ij}),$$

where n_{ij} is the number of dies observed on the j^{th} wafer in the i^{th} lot.

A.4 Intel GP Simulation Data

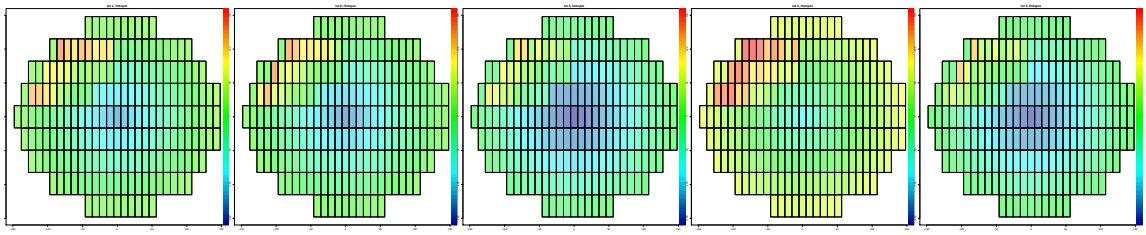


FIGURE A.1: For each lot, ordered 1 to 5, left to right, averages for each die across wafers in the lot. Note that the scale blue to red indicates small to larger means.

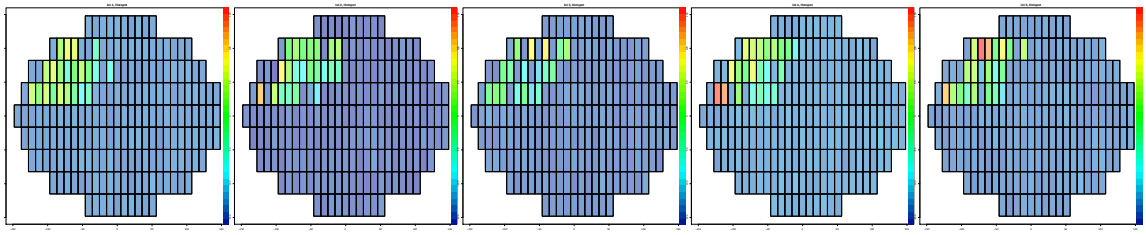


FIGURE A.2: For each lot, ordered 1 to 5, left to right, standard deviation for each die across wafers in the lot. Note that the scale blue to red indicates small to larger standard deviations.

A.5 Comparisons between INLA and MCMC Model Fitting

A.5.1 Image Reconstruction (Section 3.1)

We fit the model using INLA in addition to our previous MCMC model fitting. For reference, we provide the reconstructed images presented in the manuscript under the MCMC results section. Corresponding results found with INLA are given in the INLA results section.

MCMC results

Note that when the the level of missingness is higher, the reconstructed image is smoother relative to the images with less missingness. Additionally, note that the

residual terms diminish as the levels of missingness decrease, as expected.

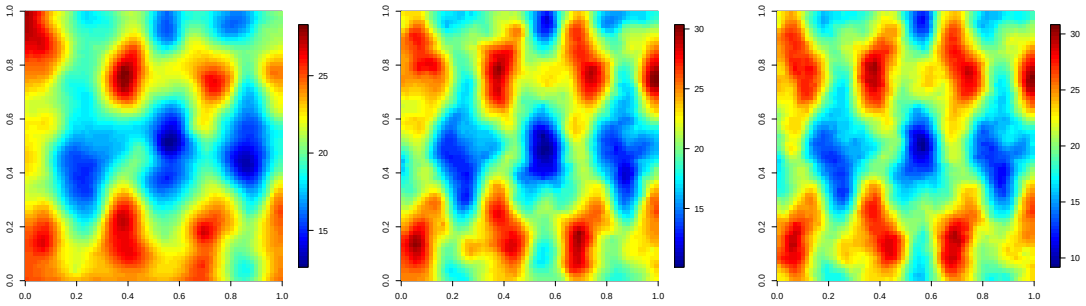


FIGURE A.3: Mean reconstructed images under 2D-RW2 CAR model fit using a Gibbs sampler with (Left) 5% of data observed (Center) 20% of data observed (Right) 50% of data observed .

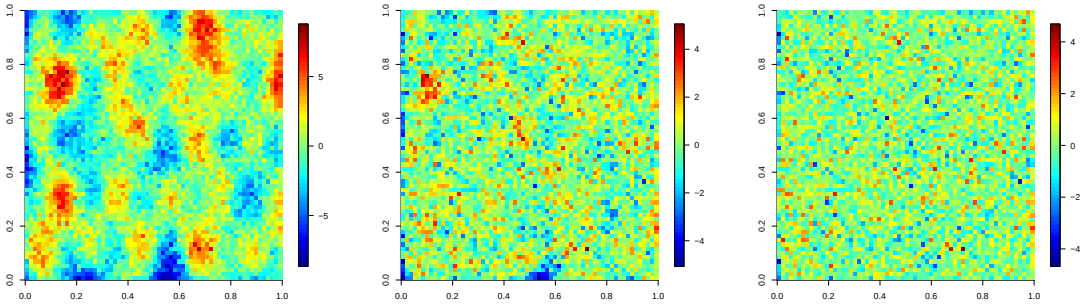


FIGURE A.4: Residual images under 2D-RW2 CAR model fit using a Gibbs sampler with (Left) 5% of data observed (Center) 20% of data observed (Right) 50% of data observed .

INLA results

For this example, the INLA model reconstructs images similarly to those given using MCMC model fitting. Additionally, the pattern in residual terms diminishes as the levels of missingness decrease, as expected. So, INLA and MCMC yield similar results in this application.

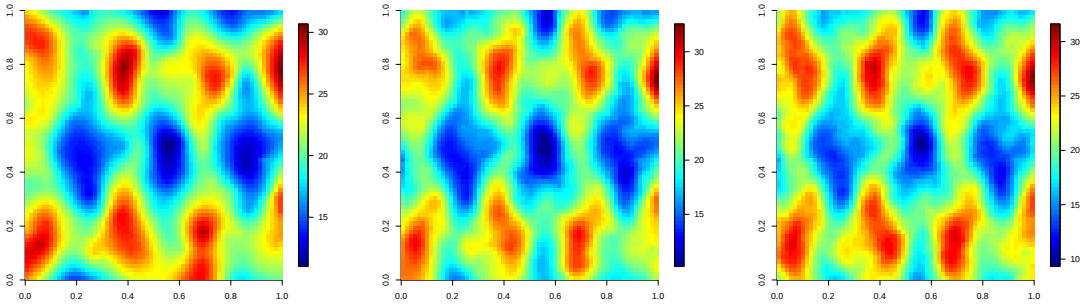


FIGURE A.5: Mean reconstructed images under 2D-RW2 CAR model using INLA with (Left) 5% of data observed (Center) 20% of data observed (Right) 50% of data observed .

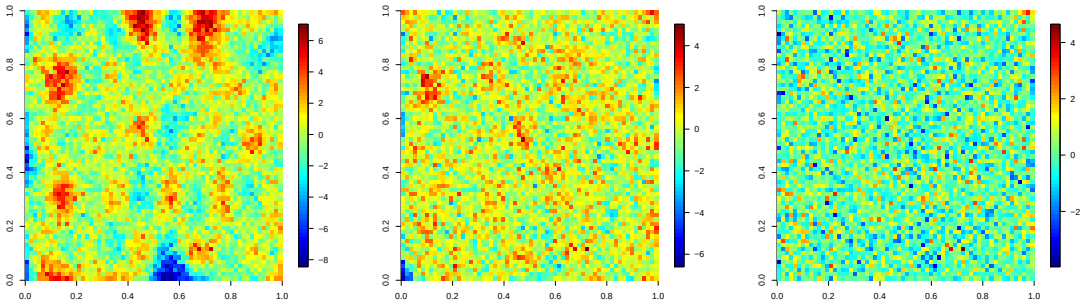


FIGURE A.6: Residual images under 2D-RW2 CAR model using INLA with (Left) 5% of data observed (Center) 20% of data observed (Right) 50% of data observed .

A.5.2 Ohio Example – Model Comparison (Section 3.2)

In the submitted manuscript, we corrected errors in Table 1. A corrected table is included under MCMC results for reference (as well as in the resubmitted manuscript). All other results have been checked.

MCMC results

Recall that the models fit have the following form All models have a similar structure:

$$y_i \sim \text{Pois}(n_i p_i)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{logit}(p_{nw,i}) + \beta_2 \text{logit}(p_{f,i}) + V_i.$$

Because the CAR model smooths on the probability of disease incidence, model comparison was done comparing p_i to the observed disease proportion. The result for the Ohio lung cancer data is given below. And

Table A.1: Model comparison statistics for the 1996 ohio lung cancer data using MCMC model fitting. The relative PMSE and CRPS give the model smoothing relative to the nonspatial model.

	Non-spatial	Improper Besag 1974	Proper ρ	Second-order	Leroux 2000
PMSE	3.20e-06	2.33e-06	2.49e-06	2.17e-06	2.24e-06
CRPS	1.35e-03	8.68e-04	9.00e-04	8.46e-04	8.90e-04
Relative PMSE	1.00	0.73	0.78	0.68	0.70
Relative CRPS	1.00	0.64	0.67	0.63	0.66

INLA results

Because the model formulation proposed in the manuscript was not available using R-INLA, we fit a model with a similar structure that used the supported log-link function:

$$y_i \sim \text{Pois}(E_i e^{x_i^T \boldsymbol{\beta} + V_i}),$$

where E_i is the expected number of counts for each county (i.e. $y_i n_i / \sum_i n_i$). We prefer the formulation given within the manuscript because it does not treat the outcome as an offset, using the outcome y_i as a part of the model for itself. Although these formulations are not identical, they are likely to yield similar results. Again, we do model comparison was the observed disease proportion to the estimated rate $E_i e^{x_i^T \boldsymbol{\beta} + V_i} / n_i$. Instead of drawing from the approximate posterior induced by INLA model fitting, we only use posterior means to compare PMSE. The results for the Ohio lung cancer data is given below.

Table A.2: Model comparison statistics for the 1996 ohio lung cancer data using INLA model fitting. The relative PMSE and CRPS give the model smoothing relative to the nonspatial model.

	Non-spatial	Improper Besag 1974	Proper ρ	Second-order	Leroux 2000
PMSE	3.19e-06	2.25e-06	2.22e-06	3.12e-06	2.24e-06
Relative PMSE	1.00	0.70	0.69	0.98	0.70

Again, note that the output is quite similar to that given by MCMC even though the models are not identical. The exception is the second-order CAR model. However, given that we are not fitting exactly the same model, some differences are not surprising.

A.5.3 Microchip Example with Binary Outcome

We compare model fitting for the microchip example with a binary outcome. We only report accuracy, sensitivity, and specificity for INLA, as we do not obtain samples from the approximated posterior. Recall that for this section, Y_{ijk} denotes a binary response at die k within wafer j within lot i. We model Y_{ijk} through a probit specification, i.e., $P(Y_{ijk} = 1|\boldsymbol{\beta}, U_k, V_{ik}, W_{ijk}) = \Phi(\mathbf{X}_k^T\boldsymbol{\beta} + U_k + V_{ik} + W_{ijk})$, where CAR terms ($U$, V, W) are introduced sequentially.

MCMC results

Using a Gibbs sampler with a probit specification and a data augmentation scheme, we present the following results given in the manuscript. For each iteration of the Gibbs sampler, we compute accuracy, sensitivity, and specificity (we add this clarification to the resubmitted manuscript). This process gives different results than we would obtain making a single prediction with a posterior mean or averaging many predictions to compute accuracy, sensitivity, and specificity.

Table A.3: Model comparison criteria for nested CARs model for binary semiconductor chip data using MCMC.

	Accuracy	Sensitivity	Specificity	Brier Score
Trend Model	0.5468	0.5363	0.5568	0.2969
Trend+Global CAR	0.5527	0.5423	0.5628	0.2983
Trend+Global+Lot CAR	0.5876	0.5779	0.5971	0.2758
Trend+Global+Lot+Wafer CAR	0.6013	0.5919	0.6105	0.2722

INLA results

Using INLA with a probit specification, we found the following results. We exclude the equicorrelated structure for simplicity. To match the procedure using for the MCMC results, we generate many sets of predictions using an approximate posterior distribution given by R-INLA (using the mean and standard deviation). Then, using each prediction set we compute accuracy, sensitivity, and specificity, and average over prediction sets. This calculation accounts for parameter uncertainty differently from using MCMC samples, but the process is similar. The results are given below. We do not include the Brier score, as we do not have posterior samples.

Table A.4: [

	Accuracy	Sensitivity	Specificity
Trend Model	0.5564	0.5507	0.5729
Trend+Global CAR	0.5556	0.5511	0.5726
Trend+Global+Lot CAR	0.5591	0.5553	0.5757
Trend+Global+Lot+Wafer CAR	0.5596	0.5555	0.5766

Model comparison for chip analysis for continous variable using Gib]Model comparison criteria for nested CARs model for binary semiconductor chip data using INLA.

Using INLA, the trend-only model has better prediction than seen using MCMC. However, the predictive gains using spatially structured models is almost completely lost using INLA.

A.5.4 Summary

It appears that INLA certainly gives computational advantages and gives similar predictive results in some circumstances. However, it performed significantly worse than MCMC for the microchip example, perhaps due to model complexity.

Appendix B

Appendix to Pollution State Modeling for Mexico City

B.1 Full Conditional Distributions for AR Model

We give the full conditional distributions for the model specified in Section 3.3. We give some additional details here to clarify model fitting. Because $\mathbf{V}_1$ and $\mathbf{V}_2$ are independent *a priori*, the joint prior distribution for $\mathbf{V}_1$ and $\mathbf{V}_2$ is

$$[\mathbf{V}_1, \mathbf{V}_2] \propto \exp\left(-\frac{1}{2}\mathbf{V}_1^T Q \mathbf{V}_1\right) \exp\left(-\frac{1}{2}\mathbf{V}_2^T Q \mathbf{V}_2\right). \tag{B.1}$$

The induced joint prior distribution of $\begin{pmatrix}\boldsymbol{\psi}_1 \\ \boldsymbol{\psi}_2\end{pmatrix}$, where $\boldsymbol{\psi}_1 = (\psi_{11}, \psi_{12}, ..., \psi_{1N_s})^T$ and $\boldsymbol{\psi}_2 = (\psi_{21}, \psi_{22}, ..., \psi_{2N_s})^T$, is used for model fitting and can be represented as

$$\begin{aligned}[\boldsymbol{\psi}_1, \boldsymbol{\psi}_2 | A_\psi] &= [\boldsymbol{\psi}_1 | A_\psi][\boldsymbol{\psi}_2 | \boldsymbol{\psi}_1, A_\psi] \\ &\propto \exp\left(-\frac{1}{2{a_{11}^{(\psi)}}^2}\boldsymbol{\psi}_1^T Q \boldsymbol{\psi}_1\right) \exp\left(-\frac{1}{2{a_{22}^{(\psi)}}^2}\left(\boldsymbol{\psi}_2 - \frac{a_{12}^{(\psi)}}{a_{11}^{(\psi)}}\boldsymbol{\psi}_1\right)^T Q \left(\boldsymbol{\psi}_2 - \frac{a_{12}^{(\psi)}}{a_{11}^{(\psi)}}\boldsymbol{\psi}_1\right)\right).\end{aligned} \tag{B.2}$$

For this section, let $\theta|\cdots$ indicate the full conditional distribution of θ, where

θ is an arbitrary parameter. For several quantities, we combine site-specific variables. For example, let $Y_t^O = (Y_{1t}^O, ..., Y_{N_s t}^O)^T$, $Y_t^{PM} = (Y_{1t}^{PM}, ..., Y_{N_s t}^{PM})^T$, $\mathbf{X}_t = \text{blockdiag}(\mathbf{x}_{it})$ and $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, ..., \boldsymbol{\beta}_{kN_s})^T$. In addition to previous terms, we also let $\mathbf{L}_t = \text{blockdiag}(\mathbf{L}_{it})$ and $\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}_{k1}, ..., \boldsymbol{\gamma}_{kN_s})^T$. The full conditional distributions for this model are provided below.

$$\beta_{1i}|\cdots \sim N(V^*_{\beta_{1i}} m^*_{\beta_{1i}}, V^*_{\beta_{1i}}) \qquad \gamma_{1i}|\cdots \sim N(V^*_{\gamma_{1i}} m^*_{\gamma_{1i}}, V^*_{\gamma_{1i}})$$

$$\beta_{2i}|\cdots \sim N(V^*_{\beta_{2i}} m^*_{\beta_{2i}}, V^*_{\beta_{2i}}) \qquad \gamma_{2i}|\cdots \sim N(V^*_{\gamma_{2i}} m^*_{\gamma_{2i}}, V^*_{\gamma_{2i}})$$

$$\beta_{01}|\cdots \sim N(V^*_{\beta_{01}} m^*_{\beta_{01}}, V^*_{\beta_{01}}) \qquad \gamma_{01}|\cdots \sim N(V^*_{\gamma_{01}} m^*_{\gamma_{01}}, V^*_{\gamma_{01}})$$

$$\beta_{02}|\cdots \sim N(V^*_{\beta_{02}} m^*_{\beta_{02}}, V^*_{\beta_{02}}) \qquad \gamma_{02}|\cdots \sim N(V^*_{\gamma_{02}} m^*_{\gamma_{02}}, V^*_{\gamma_{02}})$$

$$\Sigma_{\beta_1}|\cdots \sim IW(M^*_{\beta_1}, \nu^*_{\beta_1}) \qquad \Sigma_{\gamma_1}|\cdots \sim IW(M^*_{\gamma_1}, \nu^*_{\gamma_1})$$

$$\Sigma_{\beta_2}|\cdots \sim IW(M^*_{\beta_2}, \nu^*_{\beta_2}) \qquad \Sigma_{\gamma_2}|\cdots \sim IW(M^*_{\gamma_2}, \nu^*_{\gamma_2})$$

$$\sigma_1^2|\cdots \sim IG(a^*_{\sigma_1}, b^*_{\sigma_1})$$

$$\sigma_2^2|\cdots \sim IG(a^*_{\sigma_2}, b^*_{\sigma_2})$$

$$\mathbf{V}_1|\cdots \sim N(V^*_{V_1} m^*_{V_1}, V^*_{V_1})$$

$$\mathbf{V}_2|\cdots \sim N(V^*_{V_2} m^*_{V_2}, V^*_{V_2})$$

$${a_{11}^{(\psi)}}^2|\cdots \sim IG(a^*_{a_1}, b^*_{a_1}),$$

$${a_{22}^{(\psi)}}^2|\cdots \sim IG(a^*_{a_2}, b^*_{a_2}),$$

$$a_{12}^{(\psi)}|\cdots \sim N(V^*_{\psi} m^*_{\psi}, V^*_{\psi})$$

with

$$a^*_{\sigma_1} = 1 + \frac{N_s \times N_t}{2}$$

$$b^*_{\sigma_1} = 1 + \frac{1}{2}\sum_{t=1}^{N_t}\sum_{i=1}^{N_s}(Y_{it}^O - \mathbf{x}_{it}^T\boldsymbol{\beta}_{1i} - {\mathbf{L}_{it}^O}^T\boldsymbol{\gamma}_{1i} - a_{11}^{(\psi)}V_{1i})^2$$

$$a^*_{\sigma_2} = 1 + \frac{N_s \times N_t}{2}$$

$$b^*_{\sigma_2} = 1 + \frac{1}{2}\sum_{t=1}^{N_t}\sum_{i=1}^{N_s}(Y_{it}^{PM} - \mathbf{x}_{it}^T\boldsymbol{\beta}_{2i} - {\mathbf{L}_{it}^{PM}}^T\boldsymbol{\gamma}_{2i} - a_{12}^{(\psi)}V_{1i} - a_{22}^{(\psi)}V_{2i})^2$$

$$m^*_{\beta_{1i}} = \Sigma^{-1}_{\beta_1}\beta_{01} + \frac{1}{\sigma_1^2}\sum_{t=1}^{N_t}\mathbf{x}_{it}(Y_{it}^O - {\mathbf{L}_{it}^O}^T\boldsymbol{\gamma}_{1i} - a_{11}^{(\psi)}V_{1i})$$

$$V^*_{\beta_{1i}} = \left(\Sigma^{-1}_{\beta_1} + \frac{1}{\sigma_1^2}\sum_{t=1}^{N_t}\mathbf{x}_{it}\mathbf{x}_{it}^T\right)^{-1}$$

$$m^*_{\beta_{2i}} = \Sigma^{-1}_{\beta_2}\beta_{02} + \frac{1}{\sigma_2^2}\sum_{t=1}^{N_t}\mathbf{x}_{it}(Y_{it}^{PM} - {\mathbf{L}_{it}^{PM}}^T\boldsymbol{\gamma}_{2i} - a_{12}^{(\psi)}V_{1i} - a_{22}^{(\psi)}V_{2i})$$

$$V^*_{\beta_{2i}} = \left(\Sigma^{-1}_{\beta_2} + \frac{1}{\sigma^2_2} \sum_{t=1}^{N_t} \mathbf{x}_{it} \mathbf{x}^T_{it} \right)^{-1}$$

$$m^*_{\beta_{01}} = \sum_{i=1}^{N_s} \Sigma^{-1}_{\beta_1} \boldsymbol{\beta}_{1i}$$

$$V^*_{\beta_{01}} = \left(N_s \Sigma^{-1}_{\beta_1} + 10^{-3} \mathbf{I} \right)^{-1}$$

$$m^*_{\beta_{02}} = \sum_{i=1}^{N_s} \Sigma^{-1}_{\beta_2} \boldsymbol{\beta}_{2i}$$

$$V^*_{\beta_{02}} = \left(N_s \Sigma^{-1}_{\beta_2} + 10^{-3} \mathbf{I} \right)^{-1}$$

$$M^*_{\beta_1} = 10^3 \mathbf{I} + \sum_{i=1}^{N_s} (\boldsymbol{\beta}_{1i} - \boldsymbol{\beta}_{01})(\boldsymbol{\beta}_{1i} - \boldsymbol{\beta}_{01})^T$$

$$\nu^*_{\beta_1} = N_s + p + 1$$

$$M^*_{\beta_2} = 10^3 \mathbf{I} + \sum_{i=1}^{N_s} (\boldsymbol{\beta}_{2i} - \boldsymbol{\beta}_{02})(\boldsymbol{\beta}_{2i} - \boldsymbol{\beta}_{02})^T$$

$$\nu^*_{\beta_2} = N_s + p + 1$$

$$m^*_{\gamma_{1i}} = \Sigma^{-1}_{\gamma_1} \boldsymbol{\gamma}_{01} + \frac{1}{\sigma^2_1} \sum_{t=1}^{N_t} \mathbf{L}^O_{it} (Y^O_{it} - \mathbf{x}^T_{it} \boldsymbol{\beta}_{1i} - a^{(\psi)}_{11} V_{1i})$$

$$V^*_{\gamma_{1i}} = \left(\Sigma^{-1}_{\gamma_1} + \frac{1}{\sigma^2_1} \sum_{t=1}^{N_t} \mathbf{L}^O_{it} \mathbf{L}^{O^T}_{it} \right)^{-1}$$

$$m^*_{\gamma_{2i}} = \Sigma^{-1}_{\gamma_2} \boldsymbol{\gamma}_{02} + \frac{1}{\sigma^2_2} \sum_{t=1}^{N_t} \mathbf{L}^{PM}_{it} (Y^{PM}_{it} - \mathbf{x}^T_{it} \boldsymbol{\beta}_{2i} - a^{(\psi)}_{12} V_{1i} - a^{(\psi)}_{22} V_{2i})$$

$$\Sigma^*_{\gamma_{2i}} = \left(\Sigma^{-1}_{\gamma_2} + \frac{1}{\sigma^2_2} \sum_{t=1}^{N_t} \mathbf{L}^{PM}_{it} \mathbf{L}^{PM^T}_{it} \right)^{-1}$$

$$m^*_{\gamma_{01}} = \sum_{i=1}^{N_s} \Sigma^{-1}_{\gamma_1} \boldsymbol{\gamma}_{1i}$$

$$V^*_{\gamma_{01}} = \left(N_s \Sigma^{-1}_{\gamma_1} + 10^{-3} \mathbf{I} \right)^{-1}$$

$$m^*_{\gamma_{02}} = \sum_{i=1}^{N_s} \Sigma^{-1}_{\gamma_2} \boldsymbol{\gamma}_{2i}$$

$$V^*_{\gamma_{02}} = \left(N_s \Sigma^{-1}_{\gamma_2} + 10^{-3}\,\mathbf{I}\right)^{-1}$$

$$M^*_{\gamma_1} = 10^3\,\mathbf{I} + \sum_{i=1}^{N_s} (\boldsymbol{\gamma}_{1i} - \boldsymbol{\gamma}_{01})(\boldsymbol{\gamma}_{1i} - \boldsymbol{\gamma}_{01})^T$$

$$\nu^*_{\gamma_1} = N_s + n_{1l} + 1$$

$$M^*_{\gamma_2} = 10^3\,\mathbf{I} + \sum_{i=1}^{N_s} (\boldsymbol{\gamma}_{2i} - \boldsymbol{\gamma}_{02})(\boldsymbol{\gamma}_{2i} - \boldsymbol{\gamma}_{02})^T$$

$$\nu^*_{\gamma_2} = N_s + n_{2l} + 1$$

$$m^*_{V_1} = \frac{a^{(\psi)}_{11}}{\sigma_1^2} \sum_{i=1}^{N_t} (Y^O_t - \mathbf{X}_t \boldsymbol{\beta}_1 - \mathbf{L}^O_t \boldsymbol{\gamma}_1) +$$

$$\frac{a^{(\psi)}_{12}}{\sigma_2^2} \sum_{i=1}^{N_t} (Y^{PM}_t - \mathbf{X}_t \boldsymbol{\beta}_2 - \mathbf{L}^{PM}_t \boldsymbol{\gamma}_2 - a^{(\psi)}_{22} \mathbf{V}_2)$$

$$V^*_{V_1} = \left(Q_1 + \left[\frac{{a^{(\psi)}_{11}}^2 N_t}{\sigma_1^2} + \frac{{a^{(\psi)}_{12}}^2 N_t}{\sigma_2^2} \right] \mathbf{I} \right)^{-1}$$

$$m^*_{V_2} = \frac{a^{(\psi)}_{22}}{\sigma_2^2} \sum_{i=1}^{N_t} (Y^{PM}_t - \mathbf{X}_t \boldsymbol{\beta}_2 - \mathbf{L}^{PM}_t \boldsymbol{\gamma}_2 - a^{(\psi)}_{12} \mathbf{V}_1)$$

$$V^*_{V_2} = \left(Q_2 + \mathbf{I} \frac{{a^{(\psi)}_{22}}^2 N_t}{\sigma_2^2} \right)^{-1}$$

$$a^*_{a_1} = 1 + N_s/2$$

$$b^*_{a_1} = 1 + \frac{1}{2} \boldsymbol{\psi}_1^T Q_1 \boldsymbol{\psi}_1$$

$$a^*_{a_2} = 1 + N_s/2$$

$$b^*_{a_2} = 1 + \frac{1}{2} \left(\boldsymbol{\psi}_2 - \frac{a^{(\psi)}_{12}}{a^{(\psi)}_{11}} \boldsymbol{\psi}_1 \right)^T Q \left(\boldsymbol{\psi}_2 - \frac{a^{(\psi)}_{12}}{a^{(\psi)}_{11}} \boldsymbol{\psi}_1 \right)$$

$$m^*_{\psi} = \frac{1}{\sigma_2^2} \sum_{t=1}^{N_t} \sum_{i=1}^{N_s} V_{1i} (Y^{PM}_{it} - \mathbf{L}^{PM^T}_{it} \boldsymbol{\gamma}_{2i} - \mathbf{x}^T_{it} \beta_{2i} - a^{(\psi)}_{22} V_{2i})$$

$$V^*_{\psi} = \left(10^{-3} + \frac{1}{\sigma_2^2} \sum_{t=1}^{N_t} \sum_{i=1}^{N_s} V^2_{1i} \right)^{-1}$$

B.2 Full Conditional Distributions for Heteroscedastic AR Model

We rely on some of the details presented in Appendix B.1 for the CAR terms. We give the full conditional distributions for the heteroscedastic model specified in Section 3.3 where the variance is a function of the hour of the day $h(t)$. For this section, let $\theta|\cdots$ indicate the full conditional distribution of θ, where θ is an arbitrary parameter. We again combine several quantities for site-specific variables. Let $Y_t^O = (Y_{1t}^O, ..., Y_{N_s t}^O)^T$, $Y_t^{PM} = (Y_{1t}^{PM}, ..., Y_{N_s t}^{PM})^T$, $\mathbf{X}_t = \text{blockdiag}(\mathbf{x}_{it})$ and $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, ..., \boldsymbol{\beta}_{kN_s})^T$. In addition to previous terms, we also let $\mathbf{L}_t = \text{blockdiag}(\mathbf{L}_{it})$ and $\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}_{k1}, ..., \boldsymbol{\gamma}_{kN_s})^T$. The full conditional distributions for this model are provided below. If posterior parameters are not given below, then they are identical to those given for the homoscedastic model in Appendix B.1.

$$\beta_{1i}|\cdots \sim N(V^*_{\beta_{1i}} m^*_{\beta_{1i}}, V^*_{\beta_{1i}}) \qquad \gamma_{1i}|\cdots \sim N(V^*_{\gamma_{1i}} m^*_{\gamma_{1i}}, V^*_{\gamma_{1i}})$$

$$\beta_{2i}|\cdots \sim N(V^*_{\beta_{2i}} m^*_{\beta_{2i}}, V^*_{\beta_{2i}}) \qquad \gamma_{2i}|\cdots \sim N(V^*_{\gamma_{2i}} m^*_{\gamma_{2i}}, V^*_{\gamma_{2i}})$$

$$\beta_{01}|\cdots \sim N(V^*_{\beta_{01}} m^*_{\beta_{01}}, V^*_{\beta_{01}}) \qquad \gamma_{01}|\cdots \sim N(V^*_{\gamma_{01}} m^*_{\gamma_{01}}, V^*_{\gamma_{01}})$$

$$\beta_{02}|\cdots \sim N(V^*_{\beta_{02}} m^*_{\beta_{02}}, V^*_{\beta_{02}}) \qquad \gamma_{02}|\cdots \sim N(V^*_{\gamma_{02}} m^*_{\gamma_{02}}, V^*_{\gamma_{02}})$$

$$\Sigma_{\beta_1}|\cdots \sim IW(M^*_{\beta_1}, \nu^*_{\beta_1}) \qquad \Sigma_{\gamma_1}|\cdots \sim IW(M^*_{\gamma_1}, \nu^*_{\gamma_1})$$

$$\Sigma_{\beta_2}|\cdots \sim IW(M^*_{\beta_2}, \nu^*_{\beta_2}) \qquad \Sigma_{\gamma_2}|\cdots \sim IW(M^*_{\gamma_2}, \nu^*_{\gamma_2})$$

$$\sigma^2_{1q}|\cdots \sim IG(a^*_{\sigma_{1q}}, b^*_{\sigma_{1q}})$$

$$\sigma^2_{2q}|\cdots \sim IG(a^*_{\sigma_{2q}}, b^*_{\sigma_{2q}})$$

$$\mathbf{V}_1|\cdots \sim N(V^*_{V_1} m^*_{V_1}, V^*_{V_1})$$

$$\mathbf{V}_2|\cdots \sim N(V^*_{V_2} m^*_{V_2}, V^*_{V_2})$$

$$a_{11}^{(\psi)^2}|\cdots \sim IG(a^*_{a_1}, b^*_{a_1}),$$

$$a_{22}^{(\psi)^2}|\cdots \sim IG(a^*_{a_2}, b^*_{a_2}),$$

$$a_{12}^{(\psi)}|\cdots \sim N(V^*_{\psi} m^*_{\psi}, V^*_{\psi})$$

with

$$a^*_{\sigma_{1q}} = 1 + \frac{N_s \times N_t}{2N_h}$$

$$b^*_{\sigma_{1q}} = 1 + \frac{1}{2}\sum_{t=1}^{N_t}\sum_{i=1}^{N_s}(Y^O_{it} - \mathbf{x}^T_{it}\boldsymbol{\beta}_{1i} - \mathbf{L}^{O^T}_{it}\boldsymbol{\gamma}_{1i} - a_{11}^{(\psi)}V_{1i})^2\mathbf{1}(h(t) = q)$$

$$a^*_{\sigma_{2q}} = 1 + \frac{N_s \times N_t}{2N_h}$$

$$b^*_{\sigma_{2q}} = 1 + \frac{1}{2}\sum_{t=1}^{N_t}\sum_{i=1}^{N_s}(Y^{PM}_{it} - \mathbf{x}^T_{it}\boldsymbol{\beta}_{2i} - \mathbf{L}^{PM^T}_{it}\boldsymbol{\gamma}_{2i} - a^{(\psi)}_{12}V_{1i} - a^{(\psi)}_{22}V_{2i})^2\mathbf{1}(h(t) = q)$$

$$m^*_{\beta_{1i}} = \Sigma^{-1}_{\beta_1}\beta_{01} + \sum_{t=1}^{N_t}\frac{1}{\sigma^2_{1h(t)}}\mathbf{x}_{it}(Y^O_{it} - \mathbf{L}^{O^T}_{it}\boldsymbol{\gamma}_{1i} - a^{(\psi)}_{11}V_{1i})$$

$$V^*_{\beta_{1i}} = \left(\Sigma^{-1}_{\beta_1} + \sum_{t=1}^{N_t}\frac{1}{\sigma^2_{1h(t)}}\mathbf{x}_{it}\mathbf{x}^T_{it}\right)^{-1}$$

$$m^*_{\beta_{2i}} = \Sigma^{-1}_{\beta_2}\beta_{02} + \sum_{t=1}^{N_t}\frac{1}{\sigma^2_{2h(t)}}\mathbf{x}_{it}(Y^{PM}_{it} - \mathbf{L}^{PM^T}_{it}\boldsymbol{\gamma}_{2i} - a^{(\psi)}_{12}V_{1i} - a^{(\psi)}_{22}V_{2i})$$

$$V^*_{\beta_{2i}} = \left(\Sigma^{-1}_{\beta_2} + \sum_{t=1}^{N_t}\frac{1}{\sigma^2_{2h(t)}}\mathbf{x}_{it}\mathbf{x}^T_{it}\right)^{-1}$$

$$m^*_{\gamma_{1i}} = \Sigma^{-1}_{\gamma_1}\boldsymbol{\gamma}_{01} + \sum_{t=1}^{N_t}\frac{1}{\sigma^2_{1h(t)}}\mathbf{L}^O_{it}(Y^O_{it} - \mathbf{x}^T_{it}\boldsymbol{\beta}_{1i} - a^{(\psi)}_{11}V_{1i})$$

$$V^*_{\gamma_{1i}} = \left(\Sigma^{-1}_{\gamma_1} + \sum_{t=1}^{N_t}\frac{1}{\sigma^2_{1h(t)}}\mathbf{L}^O_{it}\mathbf{L}^{O^T}_{it}\right)^{-1}$$

$$m^*_{\gamma_{2i}} = \Sigma^{-1}_{\gamma_2}\boldsymbol{\gamma}_{02} + \sum_{t=1}^{N_t}\frac{1}{\sigma^2_{2h(t)}}\mathbf{L}^{PM}_{it}(Y^{PM}_{it} - \mathbf{x}^T_{it}\boldsymbol{\beta}_{2i} - a^{(\psi)}_{12}V_{1i} - a^{(\psi)}_{22}V_{2i})$$

$$V^*_{\gamma_{2i}} = \left(\Sigma^{-1}_{\gamma_2} + \sum_{t=1}^{N_t}\frac{1}{\sigma^2_{2h(t)}}\mathbf{L}^{PM}_{it}\mathbf{L}^{PM^T}_{it}\right)^{-1}$$

$$m^*_{V_1} = a^{(\psi)}_{11}\sum_{t=1}^{N_t}\frac{1}{\sigma^2_{1h(t)}}(Y^O_t - \mathbf{X}_t\boldsymbol{\beta}_1 - \mathbf{L}^O_t\boldsymbol{\gamma}_1) +$$

$$a^{(\psi)}_{12}\sum_{t=1}^{N_t}\frac{1}{\sigma^2_{2h(t)}}(Y^{PM}_t - \mathbf{X}_t\boldsymbol{\beta}_2 - \mathbf{L}^{PM}_t\boldsymbol{\gamma}_2 - a^{(\psi)}_{22}\mathbf{V}_2)$$

$$V^*_{V_1} = \left(Q_1 + \left[a^{(\psi)^2}_{11}\frac{N_t}{N_h}\sum_{q=1}^{24}\sigma^2_{1q} + a^{(\psi)^2}_{12}\frac{N_t}{N_h}\sum_{q=1}^{24}\sigma^2_{2q}\right]\mathbf{I}\right)^{-1}$$

$$m^*_{V_2} = \frac{a^{(\psi)}_{22}}{\sigma^2_2}\sum_{i=1}^{N_t}(Y^{PM}_t - \mathbf{X}_t\boldsymbol{\beta}_2 - \mathbf{L}^{PM}_t\boldsymbol{\gamma}_2 - a^{(\psi)}_{12}\mathbf{V}_1)$$

$$V_{V_2}^* = \left(Q_2 + \mathbf{I} a_{22}^{(\psi)^2} \frac{N_t}{N_h} \sum_{q=1}^{24} \sigma_{2q}^2 \right)^{-1}$$

$$m_\psi^* = \sum_{t=1}^{N_t} \frac{1}{\sigma_{2h(t)}^2} \sum_{i=1}^{N_s} V_{1i} (Y_{it}^{PM} - \mathbf{L}_{it}^{PM^T} \boldsymbol{\gamma}_{2i} - \mathbf{x}_{it}^T \beta_{2i} - a_{22}^{(\psi)} V_{2i})$$

$$V_\psi^* = \left(10^{-3} + \sum_{t=1}^{N_t} \frac{1}{\sigma_{2h(t)}^2} \sum_{i=1}^{N_s} V_{1i}^2 \right)^{-1}$$

B.3 Prediction of Held-out Data

For model validation, we hold out 10% of the data and impute or update these held-out values each step of the Gibbs sampler which is described below.

$$\mu_{1it} = \mathbf{x}_{i(t-1)}^T \boldsymbol{\beta}_{1i} + \mathbf{L}_{it}^{O^T} \boldsymbol{\gamma}_{1i} + \psi_{1i},$$

$$\mu_{2it} = \mathbf{x}_{i(t-1)}^T \boldsymbol{\beta}_{2i} + \mathbf{L}_{it}^{PM^T} \boldsymbol{\gamma}_{2i} + \psi_{2i},$$

and let $Y_{it}^O | \cdots$ and $Y_{it}^{PM} | \cdots$ denote the full conditional distributions of missing observations. For the heteroscedastic model, the full conditional distributions for the missing data are

$$Y_{it}^O | \cdots \sim N(\tau_{1it}^* \mu_{1it}^*, \tau_{1it}^*)$$

$$Y_{it}^{PM} | \cdots \sim N(\tau_{2it}^* \mu_{2it}^*, \tau_{2it}^*)$$

with

$$\tau_{1it}^* = \left(\frac{1}{\sigma_{1t}^2} + \sum_{j=1}^{n_{1l}} \frac{\gamma_{1j}^2}{\sigma_{1(t+l_{1j})}^2} \right)^{-1}$$

$$\mu_{1it}^* = \mu_{1it} + \sum_{j=1}^{n_{1l}} \frac{\gamma_{1ij} (Y_{i(t+l_{1j})}^O - m_{1i(t+l_{1j})} + \gamma_{1ij} Y_{it}^O)}{\sigma_{1(t+l_{1j})}^2}$$

$$\tau_{2it}^* = \left(\frac{1}{\sigma_{2t}^2} + \sum_{j=1}^{n_{2l}} \frac{\gamma_{2j}^2}{\sigma_{2(t+l_{2j})}^2} \right)^{-1}$$

$$\mu^*_{2it} = \mu_{2it} + \sum_{j=1}^{n_{2l}} \frac{\gamma_{2ij}(Y^{PM}_{i(t+l_{2j})} - m_{2i(t+l_{2j})} + \gamma_{2ij}Y^{PM}_{it})}{\sigma^2_{2(t+l_{2j})}},$$

where l_{1j} is the j^{th} lag for ozone with coefficient γ_{1ij} and l_{2j} is the j^{th} lag for PM_{10} with coefficient γ_{1ij}. The imputation method for the homoscedastic model is a special case of the heteroscedastic model.

Appendix C

Appendix for Nonseparable Covariance Models on Circles Cross Time: A Study of Mexico City Ozone

C.1 Sensitivity Analysis for Conditioning Sets

The number of spatial neighbors used and lags used were chosen using out-of-sample predictive criteria. In this section, we present the results of this selection process. This assessment also acts as a sensitivity analysis to examine how much predictive performance depends on the number of spatial neighbors and the temporal lags chosen. Here, we present only a subset of the results for brevity. While we carried out this analysis on a grid of the number of spatial neighbors, we present results at the cross-section of the best number of spatial neighbors and the best set of temporal lags. n addition, we only present the results for our selected "best" covariance model. The hold-out sample is the same as that given in the main document.

Recall that the "best" covariance model is

$$C(h,\theta,u) = \exp\left\{\exp\left[-\left(\frac{|u|}{c_t}\right)^{\alpha}\right]\cos(\theta) - \frac{h}{c_s} - 1\right\}\cos\left\{\exp\left[-\left(\frac{|u|}{c_t}\right)^{\alpha}\right]\sin(\theta)\right\},$$

where $(h,\theta,u) \in [0,\infty) \times [0,\pi] \times [0,\infty)$, $c_t, c_s > 0$ and $\alpha \in (0,2]$. The parameter c_t

governs the temporal range, c_s is the spatial range parameter, and α is a smoothness parameter.

We first assess how many spatial neighbors should be used in the conditioning set. We summarize the results using temporal lags or 1, 2, 23, 24, 25, and 168 hours back in our conditioning set (this represents the best performing set of temporal lag, as we discuss later). As described in the neighbor selection section in the manuscript, at lags greater than 0 (i.e. past observations), we condition on the $m-1$ nearest locations and the station where the observation what taken (past observations at the same station). For lag 0 or simultaneous observations, we can only condition on other stations that are ordered prior to the monitoring station in the reference set $\mathcal{S}$. Recall that we ordered stations by latitude from south to north. The results are given in Table C.1.

From these results, we found that using six spatial neighbor improved predictive performance relative to the same model using fewer spatial neighbors. One could similarly argue that five neighbors is sufficient; however, we choose six to be conservative. We also found that using more spatial neighbors than six either gave no benefit or marginally decreased predictive performance. The model with six neighbors per lag (the best model) was 3.6% better than the worst model (with two neighbors per lag) in terms of ES and 3.3% using CRPS. Ultimately, we conclude that using six neighbors is sufficient for this data.

Table C.1: Out-of-sample predictive performance as a function of the number of spatial neighbors used.

$m =$	ES	CRPS	MAE	RMSE	90% CVG
2	20.05	3.27	4.34	6.14	0.92
3	19.61	3.20	4.26	5.99	0.92
4	19.63	3.21	4.26	5.99	0.92
5	19.41	3.17	4.20	5.92	0.92
6	19.33	3.16	4.18	5.89	0.92
7	19.36	3.16	4.19	5.89	0.92
8	19.43	3.17	4.21	5.92	0.92
9	19.35	3.16	4.18	5.88	0.92

For all sets of lags, we fixed the number of lags used so that no model is given the intrinsic advantage of using more information. At each lag we use six spatial neighbors. In this analysis, we found that there were three lag selections that yielded significant gains in predictive performance relative to a nearest temporal neighbor strategy. First, including a lag for 24 hours (one day) back gave predictive improvements. However, including 48 and 72 hour lags (two and three day) appeared to give poorer predictive performance. Second, including neighbors near the 24 hour peak also improved predictive performance. Lastly, we found that some models that had the one week lag (168 hours) had better performance than similar models that used a different lag instead. These results are given in Table C.2. The best set of lags is 3.5% better than the worst model (using nearest temporal neighbors) in terms of ES and 3.4% better in terms of CRPS. Thus, we conclude that lag selection is important to model performance.

Table C.2: Out-of-sample predictive performance as a function of the lags used.

Lags	Description	ES	CRPS	MAE	RMSE	90% CVG
0-6	nearest-neighbor	20.02	3.27	4.31	6.09	0.92
0-4,24,48	nearest,local peaks	19.79	3.24	4.29	6.03	0.92
0-3,24,48,72	nearest,local peaks	19.91	3.26	4.31	6.04	0.92
0-3,24,48,168	nearest, local peaks, distant peak	19.95	3.27	4.32	6.05	0.92
0-2,12,24,48,168	nearest, local peaks, distant peak	19.55	3.20	4.21	5.94	0.93
0-2,23-25,168	nearest, near local peaks, distant peak	19.33	3.16	4.18	5.89	0.92
0-3,23,25,168	nearest, near local peaks, distant peak	19.38	3.16	4.21	5.91	0.91

Using these results, we ultimately argue for a model using six spatial neighbors

at lags 1, 2, 23, 24, 25, and 168 hours back.

C.2 Gibbs Sampler for Nearest-Neighbor Gaussian Process

For the model posed in Section 4.4.1, we provide full conditional distributions. We define o_i to be an indicator for whether a point in the reference set was observed, in our case whether it was held-out. The full conditional distributions for the Gibbs sampler, which we denote $\cdot \mid \cdots$, are

$$\boldsymbol{\beta} \mid \cdots \sim \mathcal{N}(V_\beta^* m_\beta^*, V_\beta^*) \qquad \sigma^2 \mid \cdots \sim IG(a_V^*, b_V^*)$$
$$\tau^2 \mid \cdots \sim IG(a_\tau^*, b_\tau^*) \qquad \mathbf{w}_i \mid \cdots \sim \mathcal{N}(V_{w_i}^* m_{w_i}^*, V_{w_i}^*)$$

$$V_\beta^* = \left(\mathbf{X}^\top \mathbf{X}/\tau^2 + V_\beta^{-1}\right)^{-1}$$

$$m_\beta^* = V_\beta^{-1} m_\beta + \sum_i \mathbf{x}_i (y_i - w_i)/\tau^2$$

$$V_{w_i}^* = \left(\mathbf{1}(o_i = 1)/\tau^2 + F_{s_i}^{-1} + \sum_{t:U(s_i)} B_{t,s_i}^\top F_t^{-1} B_{t,s_i} \right)^{-1}$$

$$m_{w_i}^* = \mathbf{1}(o_i = 1)(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})/\tau^2 + F_{s_i}^{-1} B_{s_i} \mathbf{w}_{N(s_i)} + \sum_{q:U(s_i,t_i)} B_{q,s_i}^\top F_q^{-1} \mathbf{a}_{q,s_i}$$

$$a_V^* = a_V + n/2$$

$$b_V^* = b_V + \sum_i (\mathbf{w}_i - B_{s_i} \mathbf{w}_{N(s_i)})^\top (\mathbf{w}_i - B_{s_i} \mathbf{w}_{N(s_i)})/R_{s_i}$$

$$a_\tau^* = a_\tau + \frac{1}{2} \sum_i \mathbf{1}(o_i = 1)$$

$$b_\tau^* = b_\tau + \frac{1}{2} \sum_i (y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - w_i)^2 \mathbf{1}(o_i = 1),$$

and $\mathbf{a}_{t,s_i}$ is as it is defined in Datta et al. (2016a).

C.3 Discussion for Exponential Covariance Functions

Suppose that we specify $w(t) \sim GP(0, C)$, where C is the exponential covariance function $\sigma^2 e^{-\phi|t-t'|}$. Let $\mathbf{T}$ be a sequence of points $t_1 < t_2 < \cdots < t_n$, where $t_i \in \mathbb{R}$, and denote $\mathbf{w} = (w(t_1), ..., w(t_n))^T$. Then, for $\Delta t_i = t_i - t_{i-1}$, the joint density of $\mathbf{w}$ is

$$\mathcal{N}\left[w(t_1); 0, \sigma^2\right] \prod_{i=2}^{n} \mathcal{N}\left[w(t_i); e^{-\phi \Delta t_i} w(t_{i-1}), \sigma^2\left(1 - e^{-2\phi \Delta t_i}\right)\right]. \tag{C.1}$$

This covariance function corresponds to the Ornstein-Uhlenbeck (OU) process. More generally, the Matérn covariance function with $\nu = p - 1/2$ corresponds to a continuous-time auto-regressive (CAR) model with lag p for $p \in \mathbb{N}$ (see, e.g., Brockwell, 2001; Rasmussen and Williams, 2006).

Bibliography

101st United States Congress, 1990. Clean Air Act Amendments of 1990. Public Law 101-549. 104 Stat. 2399.

Administración Pública de la Ciudad de México, 2016. Órgano de difusión del gobierno de la ciudad de méxico.

Albert, J. H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American statistical Association 88 (422), 669–679.

Angle, R., Sandhu, H., 1989. Urban and rural ozone concentrations in alberta, canada. Atmospheric Environment (1967) 23 (1), 215–221.

Arisido, M. W., 2016. Functional Measure of Ozone Exposure to Model Short-term Health Effects. Environmetrics 27 (5), 306–317.

Banerjee, S., Carlin, B. P., Gelfand, A. E., 2014. Hierarchical modeling and analysis for spatial data. Crc Press.

Banerjee, S., Gelfand, A. E., Finley, A. O., Sang, H., 2008. Gaussian Predictive Process Models for Large Spatial Data Sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (4), 825–848.

Barraza-Villarreal, A., Sunyer, J., Hernandez-Cadena, L., Escamilla-Nuñez, M. C., Sienra-Monge, J. J., Ramírez-Aguilar, M., Cortez-Lugo, M., Holguin, F., Diaz-Sánchez, D., Olin, A. C., 2008. Air Pollution, Airway Inflammation, and Lung Function in a Cohort Study of Mexico City Schoolchildren. Environmental Health Perspectives 116 (6), 832.

Beal, M. J., 2003. Variational algorithms for approximate Bayesian inference. University of London London.

Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., Dominici, F., 2004. Ozone and Short-term Mortality in 95 US Urban Communities, 1987-2000. Journal of the American Medical Association 292 (19), 2372–2378.

Bell, M. L., Peng, R. D., Dominici, F., 2006. The Exposure–Response Curve for Ozone and Risk of Mortality and the Adequacy of Current Ozone Regulations. Environmental Health Perspectives 114 (4), 532.

Berg, C., Porcu, E., 2017. From Schoenberg Coefficients to Schoenberg Functions. Constructive Approximation 45 (2), 217–241.

Berger, J. O., 2013. Statistical decision theory and Bayesian analysis. Springer Science & Business Media.

Berrocal, V. J., Gelfand, A. E., Holland, D. M., 2010. A Spatio-Temporal Downscaler for Output from Numerical Models. Journal of Agricultural, Biological, and Environmental Statistics 15 (2), 176–197.

Berrocal, V. J., Gelfand, A. E., Holland, D. M., 2012. Space-time data fusion under error in computer model output: An application to modeling air quality. Biometrics 68 (3), 837–848.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological), 192–236.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics (with discussion). Annals of the Institute of Statistical Mathematics, 1–59.

Bevilacqua, M., Faouzi, T., Furrer, R., Porcu, E., 2016. Estimation and prediction using generalized wendland covariance functions under fixed domain asymptotics. arXiv preprint arXiv:1607.06921.

Bevilacqua, M., Gaetan, C., Mateu, J., Porcu, E., 2012. Estimating Space and Space-Time Covariance Functions for Large Data Sets: A Weighted Composite Likelihood Approach. Journal of the American Statistical Association 107 (497), 268–280.

Bowman, F. D., Caffo, B., Bassett, S. S., Kilts, C., 2008. A bayesian hierarchical framework for spatial modeling of fmri data. NeuroImage 39 (1), 146–156.

Bravo-Alvarez, H., Torres-Jardón, R., 2002. Air Pollution Levels and Trends in the Mexico City Metropolitan Area. In: Urban Air Pollution and Forests. Springer, pp. 121–159.

Brier, G. W., 1950. Verification of forecasts expressed in terms of probability. Monthly weather review 78 (1), 1–3.

Brockwell, P., 2001. Continuous-Time ARMA Processes. Handbook of statistics 19, 249–276.

Brook, D., 1964. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. Biometrika 51 (3/4), 481–483.

Brunekreef, B., Holgate, S. T., 2002. Air pollution and health. The lancet 360 (9341), 1233–1242.

Carlin, B. P., Banerjee, S., et al., 2003. Hierarchical multivariate car models for spatio-temporally correlated survival data. Bayesian statistics 7, 45–63.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., Riddell, A., 2016. Stan: A probabilistic programming language. Journal of Statistical Software 20.

Chiogna, M., Pauli, F., 2011. Modelling Short-term Effects of Ozone on Morbidity: An Application to the Cty of Milano, Italy, 1995–2003. Environmental and Ecological Statistics 18 (1), 169–184.

Clayton, D., Kaldor, J., 1987. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics, 671–681.

Clements, N., Piedrahita, R., Ortega, J., Peel, J. L., Hannigan, M., Miller, S. L., Milford, J. B., 2012. Characterization and nonparametric regression of rural and urban coarse particulate matter mass concentrations in northeastern colorado. Aerosol Science and Technology 46 (1), 108–123.

Cocchi, D., Greco, F., Trivisano, C., 2007. Hierarchical space-time modelling of pm10 pollution. Atmospheric environment 41 (3), 532–542.

Cressie, N., 2015. Statistics for spatial data. John Wiley & Sons.

Cressie, N., Huang, H.-C., 1999. Classes of nonseparable, spatio-temporal stationary covariance functions. Journal of the American Statistical Association 94 (448), 1330–1339.

Cressie, N., Johannesson, G., 2008. Fixed Rank Kriging for Very Large Spatial Data Sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (1), 209–226.

Datta, A., Banerjee, S., Finley, A. O., Gelfand, A. E., 2016a. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. Journal of the American Statistical Association 111 (514), 800–812.

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A., Schaap, M., 2016b. Nonseparable dynamic nearest neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. The Annals of Applied Statistics 10 (3), 1286–1316.

Davis, L. W., 2017. Saturday driving restrictions fail to improve air quality in mexico city. Scientific Reports 7, 41652.

Davison, A. C., Padoan, S. A., Ribatet, M., 2012. Statistical modeling of spatial extremes. Statistical science 27 (2), 161–186.

Dean, C., Ugarte, M., Militino, A., 2001. Detecting interaction between random region and fixed age effects in disease mapping. Biometrics 57 (1), 197–202.

Departamento del Distrito Federal, Gobierno del Estado de México, Secretaría de Recursos Naturales y Pesca, Secretaría de Salud, 1996. Programa para mejorar la calidad del aire en el valle de mexico 1995-2000.

Derado, G., Bowman, F. D., Zhang, L., 2013. Predicting brain activity using a bayesian spatial model. Statistical methods in medical research 22 (4), 382–397.

Diario Oficial de la Federación, 2014a. Norma Oficial Mexicana NOM-020-SSA1-2014.

Diario Oficial de la Federación, 2014b. Norma Oficial Mexicana NOM-025-SSA1-2014.

Diebold, A. C., 2001. Handbook of silicon semiconductor metrology. CRC Press.

Direccion de Monitoreo Atmosferico. SEDEMA, Ciudad de México, 2017. Indice de la Calidad del Aire (horarios), http://www.aire.cdmx.gob.mx.
URL `http://www.aire.cdmx.gob.mx`

Dueñas, C., Fernández, M., Canete, S., Carretero, J., Liger, E., 2004. Analyses of ozone in urban and rural sites in málaga (spain). Chemosphere 56 (6), 631–639.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., Bates, D., 2011. Rcpp: Seamless r and c++ integration. Journal of Statistical Software 40 (8), 1–18.

European Environment Agency, 2016. Exceedance of air quality limit values in urban areas.

Feingold, D. G., Varga, R. S., et al., 1962. Block diagonally dominant matrices and generalizations of the gerschgorin circle theorem. Pacific J. Math 12 (4), 1241–1250.

Finley, A. O., Datta, A., Cook, B. C., Morton, D. C., Andersen, H. E., Banerjee, S., 2018. Efficient Algorithms for Bayesian Nearest Neighbor Gaussian Processes. arXiv preprint arXiv:1702.00434.

Furrer, R., Genton, M. G., Nychka, D., 2006. Covariance Tapering for Interpolation of Large Spatial Datasets. Journal of Computational and Graphical Statistics 15 (3), 502–523.

Gelfand, A. E., Kim, H.-J., Sirmans, C., Banerjee, S., 2003. Spatial modeling with spatially varying coefficient processes. Journal of the American Statistical Association 98 (462), 387–396.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., Sirmans, C., 2004. Nonstationary multivariate process modeling through spatially varying coregionalization. Test 13 (2), 263–312.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., 2014. Bayesian Data Analysis. Vol. 2. CRC press.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence (6), 721–741.

Ghosh, M., Rao, J., 1994. Small area estimation: an appraisal. Statistical science, 55–76.

Gneiting, T., 2002. Nonseparable, Stationary Covariance Functions for Space–Time Data. Journal of the American Statistical Association 97 (458), 590–600.

Gneiting, T., 2013. Strictly and Non-Strictly Positive Definite Functions on Spheres. Bernoulli 19 (4), 1327–1349.

Gneiting, T., Genton, M. G., Guttorp, P., 2006. Geostatistical space-time models, stationarity, separability, and full symmetry. Monographs On Statistics and Applied Probability 107, 151.

Gneiting, T., Raftery, A. E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102 (477), 359–378.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., Johnson, N. A., 2008. Assessing Probabilistic Forecasts of Multivariate Quantities, with an Application to Ensemble Predictions of Surface Winds. Test 17 (2), 211.

Gouveia, N., Junger, W. L., Romieu, I., Cifuentes, L. A., de Leon, A. P., Vera, J., Strappa, V., Hurtado-Díaz, M., Miranda-Soberanis, V., Rojas-Bracho, L., 2018. Effects of air pollution on infant and children respiratory mortality in four large latin-american cities. Environmental Pollution 232, 385–391.

Gradshteyn, I. S., Ryzhik, I. M., 2007. Tables of Integrals, Series, and Products, seventh Edition. Academic Press, Amsterdam.

Gramacy, R. B., Apley, D. W., 2015. Local Gaussian Process Approximation for Large Computer Experiments. Journal of Computational and Graphical Statistics 24 (2), 561–578.

Green, P. J., Richardson, S., 2002. Hidden markov models and disease mapping. Journal of the American statistical association 97 (460), 1055–1070.

Grzebyk, M., Wackernagel, H., 1994. Multivariate analysis and spatial/temporal scales: real and complex models. In: Proceedings of the XVIIth International Biometrics Conference. Vol. 1. Citeseer, pp. 19–33.

Guinness, J., 2018. Permutation and grouping methods for sharpening gaussian process approximations. Technometrics (just-accepted).

Gupta, S. S., Panchapakesan, S., Sohn, J. K., 1985. On the distribution of the studentized maximum of equally correlated normal random variables. Communications in Statistics-Simulation and Computation 14 (1), 103–135.

Hammersley, J. M., Clifford, P., 1971. Markov fields on finite graphs and lattices.

Heal, M. R., Hammonds, M. D., 2014. Insights into the composition and sources of rural, urban and roadside carbonaceous pm10. Environmental science & technology 48 (16), 8995–9003.

Heaton, M. J., Datta, A., Finley, A., Furrer, R., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., 2018. Methods for Analyzing Large Spatial Data: A Review and Comparison. arXiv preprint arXiv:1710.05013.

Hernández-Garduño, E., Pérez-Neria, J., Paccagnella, A. M., Piña-García, M. A., Munguía-Castro, M., Catalán-Vázquez, M., Rojas-Ramos, M., 1997. Air Pollution and Respiratory Health in Mexico City. Journal of Occupational and Environmental Medicine 39 (4), 299–307.

Higdon, D., 2002. Space and Space-Time Modeling Using Process Convolutions. In: Quantitative Methods for Current Environmental Issues. Springer, pp. 37–56.

Ho, H.-C., Hsing, T., 1996. On the asymptotic joint distribution of the sum and maximum of stationary normal random variables. Journal of applied probability 33 (1), 138–145.

Hocking, R. R., 2013. Methods and applications of linear models: regression and the analysis of variance. John Wiley & Sons.

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., Kaufman, J. D., 2013. Long-term air pollution exposure and cardio-respiratory mortality: a review. Environmental Health 12 (1), 43.

Hoeting, J. A., Leecaster, M., Bowden, D., 2000. An improved model for spatially correlated binary responses. Journal of agricultural, biological, and environmental statistics, 102–114.

Hogmander, H., Møller, J., 1995. Estimating distribution maps from atlas data using methods of statistical image analysis. Biometrics, 393–404.

Huang, G., Lee, D., Scott, E. M., 2018. Multivariate Space-Time Modelling of Multiple Air Pollutants and Their Health Effects Accounting for Exposure Uncertainty. Statistics in Medicine 37 (7), 1134–1148.

Huerta, G., Sansó, B., Stroud, J. R., 2004. A spatiotemporal model for mexico city ozone levels. Journal of the Royal Statistical Society: Series C (Applied Statistics) 53 (2), 231–248.

Instituto Nacional de Ecología y Cambio Climático (INECC), Diciembre 2017. Informe Nacional de Calidad del Aire 2016, México. Coordinación General de Contaminación y Salud Ambiental, Dirección de Investigación sobre la Calidad del Aire y los Contaminantes Climáticos. Ciudad de México.

Jones, R. H., Zhang, Y., 1997. Models for continuous stationary space-time processes. In: Modelling longitudinal and spatially correlated data. Springer, pp. 289–298.

Jordan, A., Krüger, F., Lerch, S., 2017. Evaluating Probabilistic Forecasts with the R Package scoringRules. arXiv preprint arXiv:1709.04743.

Kahle, D., Wickham, H., 2013. ggmap: Spatial Viualization with ggplot2. The R Journal 5 (1), 144–161.

Katzfuss, M., Guinness, J., 2017. A General Framework for Vecchia Approximations of Gaussian Processes. arXiv preprint arXiv:1708.06302.

Kaufman, C. G., Sain, S. R., 2010. Bayesian functional {ANOVA} modeling using gaussian process prior distributions. Bayesian Analysis 5 (1), 123–149.

Kaufman, C. G., Schervish, M. J., Nychka, D. W., 2008. Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets. Journal of the American Statistical Association 103 (484), 1545–1555.

Kim, S.-Y., Lee, J.-T., Hong, Y.-C., Ahn, K.-J., Kim, H., 2004. Determining the Threshold Effect of Ozone on Daily Mortality: An Analysis of Ozone and Mortality in Seoul, Korea, 1995–1999. Environmental Research 94 (2), 113–119.

Kolovos, A., Christakos, G., Hristopulos, D. T., Serre, M. L., 2004. Methods for generating non-separable spatiotemporal covariance models with potential environmental applications. Advances in Water Resources 27 (8), 815–830.

Krüger, F., Lerch, S., Thorarinsdottir, T. L., Gneiting, T., 2016. Probabilistic forecasting and comparative model assessment based on markov chain monte carlo output. arXiv preprint arXiv:1608.06802.

Kuo, H.-H., 2006. Introduction to stochastic integration. Springer Science & Business Media.

Lawson, A. B., 2013. Bayesian disease mapping: hierarchical modeling in spatial epidemiology. CRC press.

Lee, D., Mitchell, R., 2013. Locally adaptive spatial smoothing using conditional auto-regressive models. Journal of the Royal Statistical Society: Series C (Applied Statistics) 62 (4), 593–608.

Leroux, B. G., Lei, X., Breslow, N., 2000. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: Statistical models in epidemiology, the environment, and clinical trials. Springer, pp. 179–191.

Lindgren, F., Rue, H., 2008. On the second-order random walk model for irregular locations. Scandinavian journal of statistics 35 (4), 691–700.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (4), 423–498.

Lippmann, M., 1989. Health Effects of Ozone a Critical Review. Japca 39 (5), 672–695.

Loomis, D., Castillejos, M., Gold, D. R., McDonnell, W., Borja-Aburto, V. H., 1999. Air Pollution and Infant Mortality in Mexico City. Epidemiology, 118–123.

Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H., Straif, K., 2013. The carcinogenicity of outdoor air pollution. The lancet oncology 14 (13), 1262–1263.

Lunn, D. J., Thomas, A., Best, N., Spiegelhalter, D., 2000. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. Statistics and computing 10 (4), 325–337.

Ma, H., Carlin, B. P., Banerjee, S., 2010. Hierarchical and joint site-edge methods for medicare hospice service region boundary analysis. Biometrics 66 (2), 355–364.

MacKay, D. J., 1998. Introduction to Gaussian Processes. NATO ASI Series F Computer and Systems Sciences 168, 133–166.

MacNab, Y., Dean, C., 2000. Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models. Statistics in Medicine 19 (17-18), 2421–2435.

Mage, D., Ozolins, G., Peterson, P., Webster, A., Orthofer, R., Vandeweerd, V., Gwynne, M., 1996. Urban Air Pollution in Megacities of the World. Atmospheric Environment 30 (5), 681–686.

Matheron, G., 1982. Pour une analyse krigeante des données régionalisées. Centre de Géostatistique, Report N-732, Fontainebleau.

Mittal, S., 2016. A survey of architectural techniques for managing process variation. ACM Computing Surveys (CSUR) 48 (4), 54.

Mollié, A., Gilks, W., Richardson, S., Spiegelhalter, D., 1996. Bayesian mapping of disease. Markov chain Monte Carlo in practice 1, 359–379.

Penny, W. D., Trujillo-Barreto, N. J., Friston, K. J., 2005. Bayesian fmri time series analysis with spatial priors. NeuroImage 24 (2), 350–362.

Plummer, M., et al., 2003. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing. Vol. 124. Vienna, p. 125.

Pope III, C. A., Dockery, D. W., 2006. Health effects of fine particulate air pollution: lines that connect. Journal of the air & waste management association 56 (6), 709–742.

Porcu, E., Bevilacqua, M., Genton, M. G., 2016. Spatio-Temporal Covariance and Cross-Covariance Functions of the Great Circle Distance on a Sphere. Journal of the American Statistical Association 111 (514), 888–898.

Prado, R., West, M., 2010. Time Series: Modeling, Computation, and Inference. CRC Press.

Rasmussen, C. E., Williams, C. K. I., 2006. Gaussian Processes for Machine Learning. The MIT Press, Cambridge, MA.

Riojas-Rodríguez, H., Álamo-Hernández, U., Texcalac-Sangrador, J. L., Romieu, I., 2014. Health Impact Assessment of Decreases in PM10 and Ozone Concentrations in the Mexico City Metropolitan Area: A Basis for a New Air Quality Management Program. Salud Pública de México 56, 579–591.

Rodríguez, S., Huerta, G., Reyes, H., 2016. A study of trends for mexico city ozone extremes: 2001-2014. Atmósfera 29 (2), 107–120.

Romieu, I., Meneses, F., Ruiz, S., Sienra, J. J., Huerta, J., White, M. C., Etzel, R. A., 1996. Effects of Air Pollution on the Respiratory Health of Asthmatic Children lLving in Mexico City. American Journal of Respiratory and Critical Care Medicine 154 (2), 300–307.

Rue, H., Held, L., 2005. Gaussian Markov random fields: theory and applications. CRC press.

Rue, H., Martino, S., Chopin, N., 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71 (2), 319–392.

Sahu, S. K., Gelfand, A. E., Holland, D. M., 2007. High-Resolution Space–Time Ozone Modeling for Assessing Trends. Journal of the American Statistical Association 102 (480), 1221–1234.

Salam, M. T., Millstein, J., Li, Y.-F., Lurmann, F. W., Margolis, H. G., Gilliland, F. D., 2005. Birth Outcomes and Prenatal Exposure to Ozone, Carbon Monoxide, and Particulate Matter: Results from the Children's Health Study. Environmental Health Perspectives 113 (11), 1638.

Sang, H., Gelfand, A. E., 2009. Hierarchical modeling for extreme values observed over space and time. Environmental and ecological statistics 16 (3), 407–426.

Sang, H., Gelfand, A. E., 2010. Continuous spatial process models for spatial extreme values. Journal of agricultural, biological, and environmental statistics 15 (1), 49–65.

Sematech, N., 2006. Engineering statistics handbook. NIST SEMATECH.

Shirota, S., Gelfand, A. E., et al., 2017. Space and Circular Time Log Gaussian Cox Processes with Application to Crime Event Data. The Annals of Applied Statistics 11 (2), 481–503.

Shumway, R. H., Stoffer, D. S., 2017. Time Series Analysis and Its Applications. Springer.

Sillman, S., 1999. The Relation Between Ozone, NOx and Hydrocarbons in Urban and Polluted Rural Environments. Atmospheric Environment 33 (12), 1821–1845.

Solin, A., Särkkä, S., 2014. Explicit Link between Periodic Covariance Functions and State Space Models. In: Artificial Intelligence and Statistics. pp. 904–912.

Stein, M. L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer Science & Business Media.

Stein, M. L., 2005. Statistical Methods for Regular Monitoring Data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (5), 667–687.

Stein, M. L., 2008. A Modeling Approach for Large Spatial Datasets. Journal of the Korean Statistical Society 37 (1), 3–10.

Stein, M. L., 2014. Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data. Spatial Statistics 8, 1–19.

Stein, M. L., Chi, Z., Welty, L. J., 2004. Approximating Likelihoods for Large Spatial Data Sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66 (2), 275–296.

Stern, H., Cressie, N., 1999. Disease mapping and risk assessment for public health. Inference for Extremes in Disease Mapping, 61–82.

Vecchia, A. V., 1988. Estimation and Model Identification for Continuous Spatial Processes. Journal of the Royal Statistical Society. Series B (Methodological), 297–312.

Wackernagel, H., 1994. Cokriging versus kriging in regionalized multivariate data analysis. Geoderma 62 (1-3), 83–92.

Weschler, C. J., 2006. Ozone's Impact on Public Health: Contributions from Indoor Exposures to Ozone and Products of Ozone-Initiated Chemistry. Environmental Health Perspectives 114 (10), 1489.

West, M., Harrison, J., 1997. Bayesian Forecasting and Dynamic Models. Springer-Verlag New York, Inc., New York, NY, USA.

White, G., Ghosh, S. K., 2009. A stochastic neighborhood conditional autoregressive model for spatial data. Computational statistics & data analysis 53 (8), 3033–3046.

White, P., Gelfand, A., Utlaut, T., 2017. Prediction and model comparison for areal unit data. Spatial Statistics 22, 89–106.

White, P. A., Gelfand, A. E., Rodrigues, E. R., Tzintzun, G., 2018. Pollution State Modeling for Mexico City. arXiv preprint arXiv:1807.03935.

White, P. A., Porcu, E., 2018. Towards a Complete Picture of Covariance Functions on Spheres Cross Time. arXiv preprint arXiv:1807.04272.

White, P. A., Porcu, E., 2019. Nonseparable Covariance Models on Circles Cross Time: A Study of Mexico City Ozone. Environmetrics.

Zavala, á., Herndon, S., Wood, E., Onasch, T., Knighton, W., Marr, L. C., Kolb, C., Molina, L., 2009. Evaluation of mobile emissions contributions to mexico city's emissions inventory using on-road and cross-road emission measurements and ambient data. Atmospheric Chemistry and Physics 9 (17), 6305–6317.

Zhang, H., 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. Journal of the American Statistical Association 99 (465), 250–261.

Biography

Philip Andrew White completed a B.S. in Applied Physics in 2014 and an M.S. in Statistics from Brigham Young University in 2015. He plans to graduate with his Ph.D. in Statistical Science from Duke University in May 2019. Upon graduation, he will begin as an Assistant Professor in the Department of Statistics at Brigham Young University.