MODERN SPACE/TIME GEOSTATISTICAL APPROACHES TO MAPPING POINT SOURCES OF POLLUTION AND INFECTIOUS DISEASE

William Benjamin Allshouse III

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Environmental Sciences and Engineering in the Gillings School of Global Public Health.

Chapel Hill
2014

Approved by:
Marc Serre
William Vizuete
Joachim Pleil
Bill Miller
Richard Smith
ABSTRACT

William Benjamin Allshouse III: Modern Space/Time Geostatistical Approaches to Mapping Point Sources of Pollution and Infectious Disease
(Under the direction of Marc Serre)

Point sources are defined in the context of environmental science as single identifiable locations that emit pollution into the environment. From a modeling perspective, they are “ground zero” - where a contaminant originates or is at a distance of 0 from the source - and they are desirable from the standpoint of mitigation strategies because the pollutant being produced must only be brought under control at the location of release. A set of studies were conducted to investigate whether modern geostatistical techniques, such as Bayesian Maximum Entropy which has the ability to incorporate non-Gaussian space/time data, can improve the estimates of a variable produced by point sources over that of the traditional kriging method. In Study 1, a single point source at the most famous Ground Zero is the focus in order to estimate atmospheric polycyclic aromatic hydrocarbons (PAHs) produced during the collapse and cleanup of the World Trade Center in New York City by modeling the mass fraction of PAH contained in PM$_{2.5}$. This PAH to PM$_{2.5}$ model is then applied to existing PM$_{2.5}$ monitors in the area to expand the number of estimated PAH measurements. Study 2 examines a situation where many point sources across space contribute to pollution in modeling hydrogen sulfide (H$_2$S) concentrations produced by industrial hog operations in an eastern North Carolina county with a high density of these facilities. Passive samplers that recorded H$_2$S were used to create a land use regression model to estimate individual source contribution to the community and then to produce geostatistical estimates after applying a non-Gaussian measurement error model to the data.
Finally, Study 3 researches an unorthodox “point source” by attempting to identify core areas – locations with elevated rates that perpetuate an infection – of syphilis and gonorrhea in North Carolina. To locate these “sources,” which have the ability to move in space and time, a Bayesian-derived non-Gaussian model for the error is used to improve estimation of incidence rates on a fine space/time scale. These estimates are then utilized in an outbreak detection algorithm so that a source of infection can be controlled before it increases and/or moves.
# TABLE OF CONTENTS

LIST OF TABLES.........................................................................................................................ix

LIST OF FIGURES.............................................................................................................................x

LIST OF ABBREVIATIONS.................................................................................................................xi

INTRODUCTION.................................................................................................................................1

CHAPTER 1: A SINGLE SOURCE – UTILIZING PM$_{2.5}$ AS SECONDARY INFORMATION FOR MAPPING ATMOSPHERIC POLYCYCLIC AROMATIC HYDROCARBONS FOLLOWING THE COLLAPSE OF THE WORLD TRADE CENTER......................................................3

  Overview...........................................................................................................................................3

  1.1 Introduction..................................................................................................................................4

  1.2 Methods......................................................................................................................................8

    1.2.1 Sampling..................................................................................................................................8

    1.2.2 Mass Fraction Model for WTC PAHs....................................................................................9

    1.2.3 Space/Time Mean Trend and Covariance..........................................................................10

    1.2.4 Concentration Mapping and Exposure Estimates............................................................12

    1.2.5 Validation............................................................................................................................13

  1.3 Results.......................................................................................................................................14

    1.3.1 Exploratory Analysis..............................................................................................................14

    1.3.2 Mass Fraction Model............................................................................................................15

    1.3.3 Space/Time Mean Trend and Covariance..........................................................................17
CHAPTER 3: SOURCES THAT MOVE IN SPACE AND TIME – THE BAYESIAN UNIFORM MODEL EXTENSION OF BME (BUMBME) FOR DELINEATING CORE AREAS OF SYPHILIS AND GONORRHEA………………..62

Overview………………………………………………………………………………………………………62
3.1 Introduction…………………………………………………………………………………………..63
3.2 Methods………………………………………………………………………………………………66
   3.2.1 Datasets…………………………………………………………………………………………66
   3.2.2 Latent Rate and Asymptotic Validation……………………………………………………..67
   3.2.3 Simple Kriging, Poisson Kriging, UMBME………………………………………………..68
   3.2.4 BUMBME………………………………………………………………………………………70
   3.2.5 Space/Time Offset and Covariance…………………………………………………………72
3.3 Results……………………………………………………………………………………………………73
   3.3.1 Summary……………………………………………………………………………………….73
   3.3.2 Model for BUMBME Prior Distribution……………………………………………………73
   3.3.3 Space/Time Estimation…………………………………………………………………………75
   3.3.4 Asymptotic Validation…………………………………………………………………………76
   3.3.5 Comparison of Methods in Outbreak Locations………………………………………..78
3.4 Discussion……………………………………………………………………………………………...80
References………………………………………………………………………………………………………87

CHAPTER 4: SOURCES THAT MOVE IN SPACE AND TIME – USING BUMBME POSTERIOR ESTIMATES TO DETECT OUTBREAKS OF SYPHILIS AND GONORRHEA ON A FINE SPACE/TIME SCALE………………..89

Overview………………………………………………………………………………………………………89
4.1 Introduction…………………………………………………………………………………………….90
4.2 Methods……………………………………………………………………………………………………93
LIST OF TABLES

Table 1.1 – Toxicity equivalency factors for particle-bound PAHs……………………………………..5
Table 1.2 – Covariance model parameters for each PAH.................................................................18
Table 2.1 – H₂S sampling locations by type..................................................................................36
Table 2.2 – Statistically significant variables in H₂S LUR model.................................................45
# LIST OF FIGURES

Figure 1.1 – Locations of PAH samplers near Ground Zero.................................................................9

Figure 1.2 – Spatial variability of log PAH, log PM$_{2.5}$, and log MF.........................................................16

Figure 1.3 – Covariance model for benzo(a)pyrene..............................................................................19

Figure 1.4 – Spatial and temporal validation of BME vs. kriging.............................................................20

Figure 1.5 – Time-integrated concentration and population burden of benzo(a)pyrene 150 days after 9/11.................................................................................................................................23

Figure 2.1 – Raw distance to source relationship for H$_2$S around hog CAFOs..........................................................44

Figure 2.2 – H$_2$S LUR model restricted to “edge” field...........................................................................46

Figure 2.3 – Measurement error model for H$_2$S.....................................................................................47

Figure 2.4 – Covariance model for H$_2$S – LUR offset............................................................................48

Figure 2.5 – Maps of H$_2$S concentration using different methodologies..................................................50

Figure 2.6 – Comparison of mean squared error for methods estimating H$_2$S..................................................51

Figure 3.1 – Syphilis and gonorrhea long-term vs. short term rate models..............................................74

Figure 3.2 – Syphilis and gonorrhea incidence rate in North Carolina for January 1, 2009 – July 1, 2009 using BUMBME..............................................................76

Figure 3.3 – MSE reduction relative to simple kriging for syphilis and gonorrhea using asymptotic validation.................................................................77

Figure 3.4 – Observed and estimated syphilis and gonorrhea over time..................................................79

Figure 4.1 – Syphilis and gonorrhea outbreak maps..............................................................................100

Figure 4.2 – Syphilis and gonorrhea outbreak detection ROC analysis.................................................101
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIRS</td>
<td>Aerometric Information Retrieval System</td>
</tr>
<tr>
<td>BME</td>
<td>Bayesian Maximum Entropy</td>
</tr>
<tr>
<td>BUMBME</td>
<td>Bayesian Uniform Model extension of BME</td>
</tr>
<tr>
<td>CAFO</td>
<td>concentrated animal feeding operation</td>
</tr>
<tr>
<td>CBG</td>
<td>census block group</td>
</tr>
<tr>
<td>H$_2$S</td>
<td>hydrogen sulfide</td>
</tr>
<tr>
<td>MEM</td>
<td>measurement error model</td>
</tr>
<tr>
<td>MSE</td>
<td>mean squared error</td>
</tr>
<tr>
<td>PAH</td>
<td>polycyclic aromatic hydrocarbon</td>
</tr>
<tr>
<td>PDF</td>
<td>probability density function</td>
</tr>
<tr>
<td>PM</td>
<td>particulate matter</td>
</tr>
<tr>
<td>ROC</td>
<td>receiver operating characteristic</td>
</tr>
<tr>
<td>STI</td>
<td>sexually transmitted infection</td>
</tr>
<tr>
<td>STRF</td>
<td>space/time random field</td>
</tr>
<tr>
<td>UMBME</td>
<td>Uniform Model extension of BME</td>
</tr>
<tr>
<td>WTC</td>
<td>World Trade Center</td>
</tr>
</tbody>
</table>
INTRODUCTION

Point sources are defined in the context of environmental science as single identifiable locations that emit pollution into the environment. From a modeling perspective, they are “ground zero” - where a contaminant originates or is at a distance of 0 from the source. The prototypical point source of pollution is a power plant that sends particulate into the air through a smokestack. The concentration of particulate is highest near the source and it decreases with distance as it disperses into the surrounding environment. Point sources are desirable from the standpoint of mitigation strategies because the pollutant being produced must only be brought under control at the location of release.

Estimating the distribution of a given variable across space is often done with traditional geostatistics, such as the simple kriging method, which estimates the variable at unmeasured locations by weighting measurements within a neighborhood of that location according to a covariance model. This method can easily be extended to space/time so that estimates can be produced as the variable changes over time. With the development of modern geostatistical methodologies, such as Bayesian Maximum Entropy, Gaussian and non-Gaussian distributions of error can be incorporated into spatiotemporal models, producing better estimates.

This thesis hypothesizes that using modern geostatistical techniques can improve the estimates of a variable produced by point sources over that of the traditional kriging method. It starts with a case involving a single point source and the most famous Ground Zero – estimating atmospheric polycyclic aromatic hydrocarbons (PAHs) produced by the collapse and cleanup of the World Trade Center in New York City – by using the fraction of PAH to PM$_{2.5}$ to use PM$_{2.5}$
monitors to expand the number of estimated PAH measurements. The second part of the thesis examines a situation where many point sources across space contribute to pollution. The pork industry changed dramatically in the 1980s and 1990s, replacing traditional hog farms with industrial operations which house hundreds or thousands of swine, store the waste in pits, and then spray it onto land to drain the pits. In order to estimate the hydrogen sulfide concentration in a community affected by a high density of these operations, passive samplers were placed to record two week average measurements. These values were used to create a land use regression model and then to produce geostatistical estimates that include a non-Gaussian measurement error model. The final two chapters investigate an unorthodox “point source” by attempting to identify core areas – locations with elevated rates that perpetuate an infection – of syphilis and gonorrhea in North Carolina. To locate these “sources,” which have the ability to move in space and time, a Bayesian-derived non-Gaussian model for the error is used to improve estimation of incidence rates on a fine space/time scale. These estimates are then incorporated into an outbreak detection algorithm so that a source of infection can be controlled or eliminated before it increases and/or moves.
Overview

On September 11, 2001, terrorists crashed airplanes into the two main World Trade Center towers, located in the lower Manhattan section of New York City, causing the buildings to collapse. The fires that ensued in the following days and the large amount of diesel equipment needed to clean-up the area created a point source of polycyclic aromatic hydrocarbons (PAHs), pollutants which have been classified as carcinogenic and teratogenic, in the middle of a major population center. These pollutants, which are a component of PM$_{2.5}$, were measured in PM$_{2.5}$ samples in the area immediately around where the towers collapsed. The mass fraction of PAH to PM$_{2.5}$ was used to estimate PAH across a larger spatial domain and this method was compared to the traditional simple kriging method that only used actual PAH measurements. From a public health perspective, it is important to characterize the exposure to this pollutant since it was in the middle of a major metropolitan area and affected a large number of people, since the consequences of this pollution source might not be seen for some time.
1.1 Introduction

Extensive research has been conducted on effects resulting from exposure to ambient particulate matter. Particulate matter has been linked to cardiovascular diseases, respiratory problems, and reproductive effects. A large body of work on particulate matter focuses on atmospheric particles less than 10 microns in size (PM$_{10}$); more recently, research has been extended to investigation of fine particulate matter (particles less than 2.5 microns in aerodynamic diameter, PM$_{2.5}$), which travel deeper into the lungs and increase the risks of health effects. The overwhelming evidence that high concentrations of atmospheric particulate matter (PM) are associated with adverse health effects led the United States Environmental Protection Agency (EPA) to create the Aerometric Information Retrieval System (AIRS) in order to document ambient PM levels for purposes of data storage, retrieval, and interpretation. This is a nationwide system of stations that typically monitor daily concentrations of PM. Since the effects of exposure to this criteria pollutant are well established, research is starting to focus on what compounds in the PM drive the associations.

One class that could be contributing to adverse health outcomes is polycyclic aromatic hydrocarbons (PAHs). PAHs are produced by incomplete combustion during the process of burning fossil fuels (Caricchia et al. 1999). Many of these compounds have been classified as carcinogenic, mutagenic, and teratogenic by U.S. EPA (http://cfpub.epa.gov/ncea/iris/index.cfm) and IARC (http://monographs.iarc.fr/). Sixteen are identified as representative of this class by the EPA. Nine of these 16 are typically particle-bound compounds and were the focus of this study: benz(a)anthracene, chrysene, benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, indeno(1,2,3-c,d)pyrene, dibenzo(a,h)anthracene, benzo(g,h,i)perylene, and benzo(e)pyrene. Since PAHs are attached to
particles, they make up a fraction of the PM collected by filters from the EPA AIRS monitors (Guo et al. 2003; Vardar and Noll 2003). The toxicity of these compounds relative to benzo(a)pyrene can be found in Table 1.1.

**Toxicity equivalency factors for particle-bound PAHs**

<table>
<thead>
<tr>
<th>PAH</th>
<th>Toxicity Equivalency Factor (TEF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>benz(a)anthracene</td>
<td>0.1</td>
</tr>
<tr>
<td>chrysene</td>
<td>0.001</td>
</tr>
<tr>
<td>benzo(b)fluoranthene</td>
<td>0.1</td>
</tr>
<tr>
<td>benzo(k)fluoranthene</td>
<td>0.01</td>
</tr>
<tr>
<td>benzo(a)pyrene</td>
<td>1</td>
</tr>
<tr>
<td>indeno(1,2,3-c,d)pyrene</td>
<td>0.1</td>
</tr>
<tr>
<td>dibenzo(a,h)anthracene</td>
<td>1</td>
</tr>
<tr>
<td>benzo(g,h,i)perylene</td>
<td>0.01</td>
</tr>
<tr>
<td>benzo(e)pyrene</td>
<td>0</td>
</tr>
</tbody>
</table>

*Table 1.1* Toxicity equivalency factors for particle-bound PAHs relative to benzo(a)pyrene.

Although PAHs are identified as carcinogens, non-cancerous endpoints also show associations with these compounds. These include reproductive and developmental effects due to fetal PAH exposure (Perera et al. 2002; Berkowitz et al. 2003; Landrigan et al. 2004; Lederman et al. 2004; Miller et al. 2004; Tonne et al. 2004; Bocskay et al. 2005; Wolff et al. 2005).

One study approximated that only 10% of the mother’s PAH exposure reached the fetus, even though the fetus had a similar percentage of detectable PAH-DNA adducts. This evidence suggests that the fetus could be more susceptible to PAHs compared to the mother by 10-fold (Perera et al. 2005). In another study, an African-American cohort of women delivered babies with significantly lower birth weight and smaller head circumference when exposed to higher levels of PAHs. Both of these outcomes increase risk for further developmental complications, and therefore can be considered adverse (Perera et al. 2003).
The reproductive outcome of intrauterine growth restriction (IUGR), which has been linked to developmental problems, has also shown associations with exposure to PAHs (Berkowitz et al. 2003). Previous studies suggested that there might be a link between IUGR and PM. One study, however, compared areas with similar PAH concentrations and different levels of particles, finding the rate of IUGR was comparable. This suggests that PAHs, not PM, could be the pollutant of concern (Dejmek et al. 2000).

The collapse of the World Trade Center (WTC) and other buildings in its immediate area on September 11, 2001 produced a large plume of dust and gases that was visible for miles around over a period of several days. The dust and debris of crushed building materials from the collapse of the towers blanketed the lower Manhattan area. Although large particles were in the majority, the smaller particles that have the ability to travel deep into the lungs are of greatest concern (Fagan et al. 2002). It has been estimated that 11,000 tons of PM$_{2.5}$ were emitted into the air around the area. This included high concentrations of particle-bound PAHs (Lioy et al. 2002; Christen 2003; Edelman et al. 2003; McGee et al. 2003; Swartz et al. 2003; Wolff et al. 2005).

The immediate source of PAHs at the WTC site was burning jet fuel from the two planes that crashed into the towers. Subsequently, the primary source shifted to the fires that continued burning up to 100 days after the terrorist attacks. The fuel that produced these fires was estimated to be 490,000 liters of transformer oil, 380,000 liters of heating and diesel oil, 100,000 tons of organic debris, 91,000 liters of jet fuel, and gasoline from the large number of cars that were parked in lots underneath the WTC (Dalton 2003; McCallister 2002; Nordgren et al. 2002). Once these fires were extinguished, construction vehicles and other diesel equipment used for the massive clean-up effort continued to be a source of PAHs in lower Manhattan until
approximately 150 days after the towers’ collapse. After this time, PAHs in the area began to return to background concentrations (Pleil et al. 2004a).

Given that PAHs are possibly associated with several adverse health outcomes and the WTC collapse was a large source of these compounds, it is important to characterize the population’s exposure to PAHs following 9/11. Since exposure changes based on the space/time point where an individual is located, it is necessary to produce space/time maps for fully describing concentrations. This is typically conducted by measuring the pollutant at certain space/time points and using this data to interpolate values at unmeasured locations. Specific measurements for PAHs are not performed routinely as they are not part of the National Ambient Air Quality Standards (NAAQS) monitoring requirements, but instead are performed only episodically at few locations for specific assessment projects. As such, developing the link between PAHs and the more available PM$_{2.5}$ data is an important exposure and risk assessment tool.

PM$_{2.5}$ samplers from the AIRS network were in the New York area before 9/11, and the EPA installed four additional samplers for measuring particulate matter specifically around Ground Zero. A new method had previously been developed at EPA for assessing particle bound PAHs from PM$_{2.5}$ filters (Pleil et al. 2004b) and was applied to the archived filters from the four additional stations, producing a small amount of particle-bound PAH data.

These co-located measurements of PAHs and PM$_{2.5}$ from the monitoring stations around Ground Zero allowed a mass fraction approach for creating soft (probabilistic) PAH data to be implemented and tested. An estimated mean and variance of each PAH to PM$_{2.5}$ at the co-located space/time locations were used to estimate PAH values at EPA AIRS PM$_{2.5}$ monitoring station locations and produce more complete space/time maps for exposure assessment purposes.
1.2 Methods

1.2.1 Sampling

The EPA placed air samplers for a variety of species in lower Manhattan after the collapse of the WTC. Filters from four of the ambient PM$_{2.5}$ samplers were archived and later tested for particle bound PAHs. The nine representative PAHs were benz(a)anthracene, chrysene, benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, indeno(1,2,3-c,d)pyrene, dibenzo(a,h)anthracene, benzo(g,h,i)perylene, and benzo(e)pyrene. Three of the samplers (A,C,K) were located at ground level immediately around Ground Zero (as close as the fence would allow), and the fourth sampler (B) was placed on the 16$^{th}$ floor of a building that was approximately 1 km from the site of the towers (Figure 1.1). This method produced daily averages for each of the nine PAHs from September 22, 2001 until March 27, 2002. The space/time locations of these samplers are referred to as $p_{\text{hard}}$, because particle-bound PAHs at these locations are measured without noticeable errors.

Additionally, the EPA AIRS monitors in the New York/New Jersey area collected hourly concentrations of PM$_{2.5}$. These hourly measurements were converted to daily average values of ambient PM$_{2.5}$. These space/time points are locations where daily PAH data is derived from the value of daily PM$_{2.5}$. Since the PAH data at these locations is not directly measured and includes measurement error in the form of a distribution, it is referred to as $p_{\text{soft}}$. 
Locations of PAH samplers near Ground Zero

![Map showing locations of PAH samplers near Ground Zero](image)

**Figure 1.1** Samplers A, C, and K were placed around the fence line of Ground Zero. Sampler B was located about 1km away on the 16th floor of the EPA building.

### 1.2.2 Mass Fraction Model for WTC PAHs

*The mathematical formulation of the mass fraction spatiotemporal geostatistics model for particle-bound compounds can be found in Appendix A.*

The log-mass fraction (log-MF) $w_{\text{hard},i}$ at each space/time location $p_i \in p_{\text{hard}}$ was calculated using the following formulae:

$$w_{\text{hard},i} = \ln(\text{PAH}_{\text{hard},i} \ (\text{ng/m}^3) / \text{PM}_{2.5\text{hard},i} \ (\mu\text{g/m}^3)) \quad (1.1)$$

where $\text{PAH}_{\text{hard},i}$ and $\text{PM}_{2.5\text{hard},i}$ are daily average concentrations. An estimator for the mean and variance of the log-MF at the soft data points $p_j \in p_{\text{soft}}$ were obtained non-parametrically (Eqs.
A6 and A7), using a space/time distance \( d(p_i,p_j) \leq D \) that corresponds to a 60-day moving window with a spatial radius encompassing 10 km around each point \( p_j \) (i.e. \( d(p_i,p_j) = ||s_i-s_j|| \) if \( |t_i-t_j| \leq 30 \) days, \( d(p_i,p_j) = \infty \) otherwise, and \( D=10 \) km).

The estimated mean and variance of the log-MF were used to derive a soft log-PAH datum from each measured log-PM\(_{2.5}\) value (Eqs. A8 and A9). This produced a Gaussian PDF (Eq. A10) describing the uncertainty in the soft data for log-PAHs at the soft data points \( p_j \in p_{\text{soft}} \).

The use of the AIRS PM\(_{2.5}\) data allows us to extend the geographic area over which PAHs can be estimated. In the WTC application of this model, the PAH data is only available near the WTC site, while the PM\(_{2.5}\) data is available over a much wider area. In order to address this data limitation we limited the AIRS PM\(_{2.5}\) stations used to be within \( D=10 \) km of the WTC site. The 10km radius contains a high density of tall buildings, similar to that around the WTC site. We believe that a larger radius would include areas where winds are not blocked by buildings and that PAH and PM\(_{2.5}\) would behave differently than around the WTC. This constraint offers a mechanism to ensure that the PM\(_{2.5}\) stations used were within the same overall air shed as the WTC site where fires and other typical urban sources are present. However, this constraint can be relaxed for other applications where PAH data is not limited spatially as long as there are sufficient hard data to establish the ratios.

### 1.2.3 Space/Time Mean Trend and Covariance

The directly-measured log-PAH values at the WTC and the log-PAH values derived from PM\(_{2.5}\) at EPA AIRS locations were used to obtain the deterministic global log-PAH mean trend
function for each of the nine compounds. This global mean trend function was obtained using an empirical procedure that provides the benefit of using information provided by the data, and leads to a model with realistic physical characteristics of the plume. We used an additive separable space/time function for this global mean trend.

The spatial component was calculated empirically by averaging the log-PAH values at each monitoring station over the study period. These averages were smoothed using an exponential filter with a moving window radius that increases with the distance from the WTC site. This method served two purposes. It put more weight on the directly measured log-PAH values located at the WTC, and it created a physically realistic plume-like spatial component centered at the WTC site that is known to have existed from aerial photography evidence after the towers collapsed. The temporal component was calculated using a previous model developed for PAHs at the WTC site. One should refer to the article by Pleil et al. (2004a) for an in-depth description of this temporal component.

The effect of the additive space/time mean trend function in log space is a PAH plume centered at the WTC that flattens over time, slowly returning to its pre 9/11 background concentration. This space/time mean trend was removed from the hard and soft log-PAH data, with geostatistical estimation procedures (Christakos 1990, 2000) conducted on the residuals.

The general knowledge about the log-PAH residual field includes its covariance function. This function is estimated using classical geostatistical estimators based on the hard data and the expected value of the soft data. Since the global mean trend captures the spatial and temporal non-stationarity, the residual field is spatially and temporally stationary. In other words, the covariance of the residual field between points $p$ and $p'$ is only a function of the spatial distance and time difference between these points.
1.2.4 Concentration Mapping and Exposure Estimates

The Bayesian Maximum Entropy (BME) framework (Christakos 1990, 2000) was used to process the general knowledge (ie. mean and covariance), hard data (i.e. directly measured) and soft data (i.e. estimated from mass fraction) for the residual log-PAH field, and to obtain a posterior PDF characterizing the residual log-PAH concentration at estimation points distributed across space and time in the New York/New Jersey area. The median PAH concentration and the lower and upper bound of the 68% confidence interval are obtained by adding back the global mean trend and log-back transforming the corresponding statistical estimator for the residual log-PAH.

The numerical implementation of this work was done using the BMElib package (Serre and Christakos 1999; Christakos et al. 2002) written in MATLAB™. The library makes it easy to integrate the general and site specific knowledge bases developed in this work, and to obtain the BME median estimate of PAH concentrations. Readers are referred to Appendix A for a thorough description of the BME framework.

Two types of maps were then constructed to display the PAH concentrations following the WTC disaster. The first type of map is an aggregate estimate of the concentration that one would be exposed to at a given point for a given residence time. The second type of map calculates a potential population burden that accounts for population density.

Specifically, the time-integrated PAH concentration can be calculated as:

\[
Time \text{ Integrated } PAH \text{ Concentration} = (\Sigma PAH) \times (DAI) \times (TEF) \tag{1.2}
\]

where the time integration term \(\Sigma PAH \ ((\text{ng}\cdot\text{day})/\text{m}^3)\) is the sum of the BME median estimates of daily concentrations of a given PAH up to the day of interest at a given geographical location,
DAI is the daily air intake (assumed to be 11 m$^3$/day for a person at rest), and TEF (unitless) is the toxicity equivalency factor relative to benzo(a)pyrene. These maps show the geographic areas that had the highest PAH concentrations following 9/11 after normalizing to the toxicity of benzo(a)pyrene.

The maps for potential population burden are created by multiplying the time integrated PAH concentration by the population density. They are calculated using equation (1.3), where PD (persons/mi$^2$) is the population density:

$$\text{Time Integrated PAH Population Burden} = (\sum \text{PAH})*(\text{DAI})*(\text{TEF})*(\text{PD})$$  \hspace{1cm} (1.3)

These maps show the population impact of PAH concentrations. Areas with higher population density show a higher population burden compared to a similar time integrated PAH concentration, but lower population density.

1.2.5 Validation

Validation was performed to check the performance of the mass fraction model. This was done using two methods: a spatial validation and a temporal validation. To validate the mass fraction model in space, the hard data from one of the Ground Zero monitoring stations measuring co-located PAHs and PM$_{2.5}$ were removed and re-estimated by the mass fraction model. Then PAHs derived from PM$_{2.5}$ were compared to the space/time kriging model that only used the directly measured PAH information. Reduction of the mean squared error (MSE) relative to the space/time kriging model was used as the measure of success. A temporal validation was also conducted to compare models. This method involved removing all PAH information for a given time window and re-estimating. The mass fraction model used directly-
measured PAH data from days not removed and PAH data derived from PM$_{2.5}$. This was compared to space/time kriging, which generated estimates based on PAH values from days not removed. The reduction in the relative MSE was again the measure of success.

1.3 Results

1.3.1 Exploratory Analysis

An exploratory analysis of the data was conducted to get a sense of the behavior throughout space and time. The statistical distributions for each of the nine PAHs of interest were highly skewed from normality, as expected for this type of environmental data, and a log transformation was therefore implemented.

The concentrations of all pollutants should have generally decreased as a function of space/time distance with respect to the WTC catastrophe. Factors such as fires, wind, and other pollution sources were expected to cause deviations in this trend. Based on available data and visual evidence, a large plume of contaminants (including PAHs and PM$_{2.5}$) existed at the space/time point immediately following the towers’ collapse at Ground Zero. This plume decreased as the space/time distance increased. PAH concentrations in the 10 days immediately following the towers’ collapse are not available. During that time, the community, city, State, and Federal focus was on rescue and recovery; samplers were not deployed until Sept. 21. We assume that in those few days that the PAH and PM$_{2.5}$ levels would have been the highest and most variable, but there is no objective evidence available. Log-PAH concentrations were highly variable over the entire study period (until 200 days after 9/11). It is possible that a decrease in the variability of log-concentration would have been evident if PAH data from day 0-10 were
available. All PAHs show higher values at monitoring stations A, C, and K (those closest to Ground Zero) and lowest values at station B (290 Broadway, approximately 1km away).

Because PM\(_{2.5}\) was the basis for the creation of soft PAH information at unmeasured space/time locations, it is important to characterize its behavior near the WTC site as well. Predictably, the temporal trend is very similar to that of PAHs. PM\(_{2.5}\) displays high variability over the study period with a decrease in concentration over time. The values reflect higher concentrations compared to the time series of PAHs since the PAHs make up a fraction of the PM\(_{2.5}\) mass.

1.3.2 Mass Fraction Model

The co-located PAH and PM\(_{2.5}\) data were used to model the log-MF of PAH to PM\(_{2.5}\). The estimated distribution of the log-MF for each PAH compound was derived from the 60-day moving window approach described previously. A 60-day moving window was used to capture a more realistic estimate of the behavior at unmeasured locations due to the high temporal variability. In addition to the fact that both PAHs and PM\(_{2.5}\) decreased over time after 9/11, the ratio of each PAH to PM\(_{2.5}\) decreased as well. The one exception to this downward trend was benzo(g,h,i)perylene. This anomaly in the data was interpreted previously to demonstrate the changing pattern of PAHs as a reflection of changing of dominant combustion source producing the various compounds (Pleil et al. 2006).
Spatial variability of log PAH, log PM$_{2.5}$, and log MF

Figure 1.2 (a) The change in mean by distance (using the WTC sites as a baseline) for log PM, log PAH and log MF for all 9 PAH compounds. (b) The vertical bars at distance 0km (WTC sites) and 0.9km (Station B) represent the mean +/- one standard deviation for the log-transformed benzo(a)pyrene data and its corresponding mass fraction collected at those respective monitoring locations. The horizontal dotted line depicts +/- one standard deviation of the mass fraction mean based on the pooled data. (c) Similarly, this is shown for the compound indeno(1,2,3-c,d)pyrene.

In addition to temporal changes, we also investigated the spatial gradient of log-PAH, log-PM$_{2.5}$, and the log-MF by analyzing the differences in the data at Station B compared to the WTC monitoring sites (A, C, K). As expected, the average of PM$_{2.5}$ and all nine PAHs declined from the WTC sites to Station B (Figure 1.2a). Monitoring sites at greater distances should show a further decline in these pollutants. However, a consistent spatial trend was not evident for the log-MF. For some compounds, there was an increase in the ratio of PAH to PM$_{2.5}$. Figures 1.2b and 1.2c help to illustrate this point as there is a slight decrease in the log-MF for benzo(a)pyrene while there is an increase of similar magnitude of the log-MF for indeno(1,2,3-c,d)pyrene. Both
are easily contained within one standard deviation of their respective mass fraction mean based on the pooled data. If we can assume that the enrichment of PAHs in PM$_{2.5}$ is due to the same source at the WTC, then there is no evidence to reject that the slope of the spatial gradient with respect to the mass fraction is different from 0 within a reasonable distance of the source. Therefore, our mass fraction model that is used to derive soft PAH data from surrounding PM$_{2.5}$ stations is based on the assumption that the MF is homogenous across space, while decreasing as the number of days from 9/11 increases.

The log-MF data points tend to be within one standard deviation of the estimated mean for the model, showing that the 60-day moving window is a reasonable method to use. One notable exception is the compound dibenzo(a,h)anthracene, which has a set of values that consistently lie well below the expected distribution. Therefore, soft data created for this PAH tends to be less reliable than others. We observed that in general, this PAH is present at lower absolute value than the other target compounds and is therefore more difficult to measure.

These log-MF distributions were used to derive the soft data for log-PAH concentrations from PM$_{2.5}$ measurements at EPA AIRS monitoring stations from the New York metropolitan area. Using the soft data from these additional stations allowed for an increased spatial domain that covered lower Manhattan and surrounding areas.

1.3.3 Space/Time Mean Trend and Covariance

The approach described in section 1.2.3 was used to obtain the space/time global mean trend, describing the consistent spatial and temporal trends in PAHs produced by the WTC catastrophe. This global mean trend was removed from the log-transformed data to create the
data available for the homogenous/stationary residual field. The shape of this global mean trend is a plume with its peak at Ground Zero on September 11, 2001. As time increases from day 0 (9/11), the plume decreases while spreading over space.

The data of the homogenous/stationary residual field were then used to obtain estimates of covariance for various spatial and temporal lags. The data used in this estimation included the hard values from the WTC as well as the mean of the log-PAH distribution of residuals derived from PM$_{2.5}$. Use of the soft data was necessary to have a covariance that described the dependence between two points at distances greater than the small spatial extent covered by the WTC data. Nested exponential/exponential space/time separable covariance models produced the best fit to the data for all nine PAHs (Eq. 1.4).

\[
c_X(r,\tau) = c_1 \exp\left(\frac{-3r}{a_{r1}}\right) \exp\left(\frac{-3r}{a_{r2}}\right) + c_2 \exp\left(\frac{-3r}{a_{r1}}\right) \exp\left(\frac{-3r}{a_{r2}}\right) + c_3 \exp\left(\frac{-3r}{a_{r1}}\right) \exp\left(\frac{-3r}{a_{r2}}\right)
\]

where \( r \) and \( \tau \) are the spatial and temporal lags, respectively, between space/time points, the \( a_r \) are spatial ranges, and the \( a_t \) are temporal ranges. The parameters fit to this model for each of the nine PAHs are shown in Table 1.2, and the typical shape of the spatial and temporal components are illustrated by the covariance model for benzo(a)pyrene (Figure 1.3).

**Covariance model parameters for each PAH**

<table>
<thead>
<tr>
<th>PAH</th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>a$_{r1}$</th>
<th>a$_{r2}$</th>
<th>a$_t$</th>
<th>a$_{t2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>benz(a)anthracene</td>
<td>0.08</td>
<td>0.20</td>
<td>0.04</td>
<td>0.03</td>
<td>10</td>
<td>100</td>
<td>3</td>
</tr>
<tr>
<td>chrysene</td>
<td>0.11</td>
<td>0.23</td>
<td>0.02</td>
<td>0.03</td>
<td>10</td>
<td>100</td>
<td>3</td>
</tr>
<tr>
<td>benzo(b)fluoranthene</td>
<td>0.32</td>
<td>0.18</td>
<td>0.08</td>
<td>0.03</td>
<td>100</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>benzo(k)fluoranthene</td>
<td>0.26</td>
<td>0.12</td>
<td>0.04</td>
<td>0.03</td>
<td>3</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>benzo(a)pyrene</td>
<td>0.09</td>
<td>0.25</td>
<td>0.01</td>
<td>0.03</td>
<td>10</td>
<td>100</td>
<td>3</td>
</tr>
<tr>
<td>indeno(1,2,3-c,d)pyrene</td>
<td>0.28</td>
<td>0.10</td>
<td>0.12</td>
<td>10</td>
<td>0.02</td>
<td>3</td>
<td>200</td>
</tr>
<tr>
<td>dibenzo(a,h)anthracene</td>
<td>0.18</td>
<td>0.30</td>
<td>0.15</td>
<td>0.05</td>
<td>10</td>
<td>75</td>
<td>2</td>
</tr>
<tr>
<td>benzo(g,h,i)perylene</td>
<td>0.25</td>
<td>0.10</td>
<td>0.05</td>
<td>0.04</td>
<td>3</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>benzo(e)pyrene</td>
<td>0.24</td>
<td>0.10</td>
<td>0.11</td>
<td>0.03</td>
<td>2</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.2 Parameters for the covariance functions for each PAH**
Covariance model for benzo(a)pyrene

![Covariance of benzo(a)pyrene](image)

**Figure 1.3** Nested exponential/exponential space/time separable covariance model for benzo(a)pyrene

### 1.3.4 Comparison of Methods

The estimation of PAHs using BME with soft data was compared to the alternative of space/time kriging using only hard PAH data. Space/time kriging was considered the best possible method available that does not use the additional information from PM$_{2.5}$. Spatial and temporal validations were conducted in order to compare the two methods. Both comparisons showed a reduction in MSE when the soft PAH data was included in the estimation.

For the spatial validation, Station B (the one furthest from the WTC) was removed and re-estimated using both estimation methods. BME with the mass fraction approach reduced the MSE between 7\% - 22\% for the nine PAHs of interest (Figure 1.4a). When the other stations were removed (A, C, and K), the results were similar, but smaller reductions in the mean squared errors were observed. The reduction in MSE was greater in Station B because it was further
away from the WTC, thereby providing a greater opportunity for contrast between the two methods. Although this still only represented a small area for showing the estimation improvement when using soft data, we presume that as distance from the WTC increases, one could expect that the contrast between methods would further increase, leading to a larger reduction in MSE.

**Spatial and temporal validation of BME vs. kriging**

![Spatial Validation](image1.png) ![Temporal Validation](image2.png)

**Figure 1.4** (a) The spatial validation showing reduction in the MSE using BME compared to space/time kriging. The number for PAH refers to: 1) benz(a)anthracene, 2) chrysene, 3) benzo(b)fluoranthene, 4) benzo(k)fluoranthene, 5) benzo(a)pyrene, 6) indeno(1,2,3-c,d)pyrene, 7) dibenzo(a,h)anthracene, 8) benzo(g,h,i)perylene, and 9) benzo(e)pyrene. (b) The temporal validation showing reduction in the MSE for benzo(a)pyrene.

A similar comparison between methods was conducted temporally by removing data over different periods of time and re-estimating. There was an exponential improvement in the MSE positively associated with the number of days of PAH data removed, increasing to a 50% reduction with 30 days removed. This trend is likely a result of the large variance component contributed by time. Using soft data greatly reduces this temporal variability, giving a more
dramatic effect than the spatial validation. All nine PAHs showed similar trends to that of benzo(a)pyrene (Figure 1.4b). This illustrates the importance of the mass fraction model for obtaining information about PAHs from available PM$_{2.5}$ data when only sparse PAH data is available.

1.3.5 Exposure Mapping Results

The availability of PM$_{2.5}$ data across the New York metropolitan area allowed concentration maps to be created that incorporate the soft PAH information. Maps were constructed for each day in the study period and made into a movie. However, the daily PAH maps are highly variable and not as informative as maps aggregated over time. Hence we map the time-integrated PAH concentration (Eq. 1.2) to characterize the areas of New York that were most affected by the higher levels of PAHs in the atmosphere. The time-integrated PAH concentration for benzo(a)pyrene 150 days after 9/11 is shown in Figures 1.5a and 1.5b.

The areas of highest time-aggregated PAH concentrations are those around Ground Zero and the areas to the south and west. These maps show that a hypothetical person at rest with average lung capacity standing at Ground Zero would have inhaled over 1000 $ng$ of benzo(a)pyrene during the 150 days after 9/11. This is not meant to be an extensive exposure assessment, but is simply a tool to help illustrate the potential impact of PAH pollution from the WTC. This does not account for human activity patterns which could lower the estimate (real people did not stand still at ground zero for 150 days), nor does it account for the initial 10 days when we assume higher levels could raise the estimate.
As mentioned earlier, there are no reliable hard data for the first 10 days after 9/11 because the main focus was on human rescue efforts and sampling sites were not yet established. We expect that workers in the area likely experienced higher inhalation exposures than the general public as they were typically doing difficult tasks requiring higher pulmonary ventilation rates. This issue is beyond the scope of this work, but has been addressed in other publications contrasting the public with firefighters, rescue workers, cleanup crews, construction and sanitation workers (Dahlgren et al. 2007; Herbert et al. 2006; Fireman et al. 2004; Lorber et al. 2007; Landrigan et al. 2004).

Including residential population density shows a different picture for the potential effect of PAHs. These maps can help estimate the population/economic burden resulting from elevated PAH levels. Multiplying the residential population density by the time integrated concentration (Eq. 1.3) produces the maps of residential population burden shown in Figures 1.5c and 1.5d. Since Ground Zero and areas west are typically non-residential areas, these places do not show high levels of residential population burden. The highly populated residential area of eastern lower Manhattan had the greatest population based exposure based on these calculations. Similarly, highly populated areas east and west of Central Park although with lower concentrations, show a large population impact due to the high population density of these areas.
Time-integrated concentration and population burden of benzo(a)pyrene 150 days after 9/11

Figure 1.5 (a,b) Time-integrated benzo(a)pyrene concentration (ng inhaled) during the 150 days after 9/11 as calculated by Eq 1.2; (c,d) the time-integrated benzo(a)pyrene residential population burden (\((ng \times persons)/mi^2\)) for the 150 days after 9/11 as calculated by Eq 1.3.
The residential population density maps represent where the population resides during non-working hours. If daytime population densities were available, a much different map would emerge. The financial district (around the WTC site), for example, could display the highest population burden during the daytime hours due to its high density of office buildings but this would require additional information regarding infiltration and ventilation rates.

1.4 Discussion

This research produces the first maps to date displaying estimates of the space/time concentrations and variability of increased PAH levels after the collapse of the WTC. Using the limited amount of PAH data available, a mass fraction approach to producing soft PAH data made it possible to extend the geographical area covered by the PAH maps.

The maps of PAHs in lower Manhattan from this study provide advanced atmospheric concentration estimates for these pollutants not available in previous studies concerning health outcomes due to the environmental air pollution from the WTC. Other measures for WTC PAH contributions include the visible plume of debris and the number of PAH-DNA adducts. However, neither of these methods captures the comprehensive nature of PAH exposure for individuals in the New York area following 9/11. Our space/time maps allow for the integration of daily PAH and PM$_{2.5}$ data to estimate time-integrated PAH concentrations that provide a measure for ecologic exposure assignments.

There are obvious deficiencies with only using these maps for an individual’s PAH exposure. This assessment does not capture the indoor PAH concentrations that a person might encounter. An individual’s habits (such as cooking) and how tightly their home is sealed will
help determine their indoor PAH exposure. Our maps created for outdoor exposure cannot account for these factors and could underestimate or overestimate based on how an individual spends their time.

As mentioned earlier, directly measured PAH data are not available for the 10 days immediately following the towers’ collapse due to the logistics of the disaster response and rescue. Using a regression model developed previously, an estimate of these compounds was included for these days as well (Pleil et al. 2004a). However, we cannot validate our model during this period without hard PAH data. This likely causes under-prediction in our PAH exposure maps as the mass fractions of PAHs were likely higher during these initial days than the value estimated starting on day 10.

Finally, the maps regarding the population impact due to PAHs refer to population estimates where people live, not where they work. Although detailed information regarding where people spend their workday were not available to us, many New Yorkers typically spend the majority of their day within a relatively small distance from their home. Therefore, the maps of population impact can generally show where the most people would likely encounter PAHs. If detailed information about how one spends their day were available, these maps could be refined to characterize their exposure.

Research studies have confirmed that exposures to fine particulate matter as a whole are related to a variety of adverse health outcomes without specific regard to their precise composition. This work provides an additional tool for deducing PM$_{2.5}$ effects by considering the enrichment of the particles with PAHs which, as a group, demonstrate carcinogenic, mutagenic, and reproductive effect associations. The use of ambient air pollution monitors could
play an integral part of this research in order to conduct large studies where the use of personal samplers is impractical.

The mass fraction approach for estimating PAH concentrations from archived PM$_{2.5}$ samples has several benefits over the traditional geostatistical method of space/time kriging using only directly-measured PAH data. Our method allows for greater spatial and temporal coverage in an exposure assessment for PAHs. The additional cost is only a function of the number of PM$_{2.5}$ samples analyzed and not the need for more, expensive monitors. By creating a parametric model based on this method, it may become possible to predict PAH concentrations using weather and readings of other atmospheric pollutants. This would reduce additional costs to virtually zero. In addition to the approach developed here for a massive, but transient event, future work should also focus on long-term ambient exposures as produced by numerous smaller point sources (factories, incinerators, refineries, forest fires, agricultural burning), long range transport of PM$_{2.5}$, and distributed or line sources (vehicle traffic in congested urban areas, busy highways, ship channel traffic, airports). There are several reasons why the mass fraction method could perform even better under these circumstances. In contrast to the high variability of WTC aftermath where fires continuously flared up and diminished, PAH and PM$_{2.5}$ concentrations should be more stable in other settings, and decrease the variance of the mass fraction, leading to better estimates of PAHs. Also, New York City is unique compared to other areas due to its high density of tall buildings. These structures block and channel winds, affecting the transport of PAHs. This has the effect of decreasing the reliability of interpolation because places close in space might not be highly correlated due to the structures between them. For these reasons, estimation of PAHs using the mass fraction method should improve in other situations.
However, it is difficult to predict how our approach will perform elsewhere. While a large amount of these pollutants were produced in the disaster, this dataset had a very limited spatial domain which included only four monitoring stations around the immediate WTC area. This left a small spatial area in which to perform the validation procedure. While it is possible this inflated the performance of our method, the station used for validation recorded PAH values that were consistently an order of magnitude lower than those around the WTC fence. Therefore, it is reasonable to assume that the method would improve if PAH data outside lower Manhattan were available for validation.

A limitation of our approach in the WTC situation is that the mass fraction model was created only using data from lower Manhattan in order to estimate concentrations outside the model’s spatial domain. While it is true that prediction reliability from PM$_{2.5}$ will decrease outside of the WTC, a counterargument is that some information about PAHs from PM$_{2.5}$ is much better than no information at all in these areas. The exponential shape of the temporal validation supports this hypothesis.

This application used a normal distribution to characterize the soft data at points outside Ground Zero. It would have also been possible to use the actual statistical distributions of the log-MF. This would be beneficial for those compounds whose log-MF showed an obvious departure from normality. Dibenzo(a,h)anthracene, which produced a bimodal distribution, would have been the only compound in this study that probably would have benefited from this type of analysis. Even with the challenges posed by this dataset and study area, the mass fraction method for creating additional soft PAH information appears to make considerable improvements over the alternative method of space/time kriging. The validation should show even greater reductions in MSE using more spatially diverse data in another area.
The concept of using the mass fraction to predict PAHs from readily available PM$_{2.5}$ data could become a useful regulatory or epidemiological assessment tool because there are obvious cost savings by not having to deploy special samplers and make separate PAH measurements at all space/time monitoring locations. Future data sets that include more spatially diverse PAH data would be useful in developing a better model. Applying this model to a rare disaster with many contaminants such as the WTC is a worst-case scenario for evaluation. However, despite the complexity of the dataset, the mass fraction appears to be very useful compared to traditional kriging methods. We expect that the accuracy and value of state-wide assessments, in the absence of extraordinary events, would benefit even more from these methods.

In conclusion, the mass fraction approach for creating soft data from PM$_{2.5}$ appears to be a promising method for estimating concentrations of PAHs in future studies. Since the technology exists for measuring PAHs in archived PM$_{2.5}$ filters, this is a logical approach to extract information from a dense network of available PM$_{2.5}$ monitors. In addition, this method could be applied to other particle bound pollutants such as polychlorinated biphenyls (PCBs) and dioxins that constitute a fraction of atmospheric particulate matter.
REFERENCES


Pleil JD, Vette AF, Johnson BA, Rappaport SM (2004a) Air levels of carcinogenic polycyclic aromatic hydrocarbons after the World Trade Center disaster. PNAS 101(32):11685-11688

Pleil JD, Vette AF, Rappaport SM (2004b) Assaying particle-bound polycyclic aromatic hydrocarbons from archived PM$_{2.5}$ filters. Journal of Chromatography A 1033(1):9-17


CHAPTER 2

MULTIPLE SOURCES IN SPACE – USING LAND USE REGRESSION TO ESTIMATE ATMOSPHERIC HYDROGEN SULFIDE IN AN AREA WITH A HIGH DENSITY OF INDUSTRIAL HOG OPERATIONS

Overview

Pork production in the United States has moved to an industrial model where a large number of swine are confined in a building. The waste from the hogs in these buildings is captured and sent to a holding pit that must be pumped out periodically. In North Carolina, the waste is often sprayed on adjacent fields. While this method of waste treatment should in and of itself cause environmental concerns, the problem is multiplied by the fact that hog operations are concentrated in rural areas of eastern North Carolina such that some counties have hundreds of these point sources of pollution. Passive samplers were placed in one such area to measure two week average concentrations of hydrogen sulfide (H$_2$S). A land use regression model was developed to estimate the H$_2$S concentration contributed by so many point sources within the area, and the geostatistical estimates produced were updated with a non-Gaussian measurement error model. This again was compared to the traditional method of simple kriging to quantify the improvement in estimation from using a modern geostatistical approach. Since residents surrounding these operations have continually voiced concerns about smells and adverse health effects, we use H$_2$S as a marker for the atmospheric mixture of pollutants produced by these sources to gauge their influence across space and over time.
2.1 Introduction

During the 1980s and 1990s, hog production in North Carolina changed from a traditional model to an industrial one where swine were kept in concentrated animal feeding operations (CAFOs). A typical CAFO has buildings that can house hundreds or thousands of animals. The waste in these buildings is collected and stored in open air pits and then to maintain freeboard required by permits, most operations spray the waste on adjacent fields.

The CAFO open waste pit model has created numerous environmental concerns from contaminants in the air, water, and soil and health effects such as additional stress, irritated eyes, nose, and throat, increases in asthma, and acute blood pressure elevation (Wing et al. 2013; Mirabelli et al. 2006; Merchant et al. 2005; Bullers 2005; Hodne 2001; Wing and Wolf 2000; Cole et al. 2000; Thu et al. 1997; Schiffman et al. 1995). In addition to adverse health outcomes, CAFOs can contribute to decreased quality of life, lower property values and suppressed economic development (Thu 2002; Palmquist et al. 1997; Thu 1996; Durrenberger and Thu 1996).

In North Carolina, there are now a large number of hog CAFOs, particularly in the eastern part of the state (NCDENR 2007). This area is typically rural, with a lower socioeconomic status and higher Black and Hispanic populations compared to the rest of the state. The highest density of these operations is in the south central part of the North Carolina coastal plain region, especially in the area of Sampson and Duplin counties. Locations in this area can have hundreds of CAFOs within a radius of just 30 km. Thus regardless of wind direction, residences are usually downwind from CAFOs at any given time.

There are a number of pollutants produced by hog CAFOs (Donham et al. 2006; NAS 2002; ISG 2002). These include hydrogen sulfide (H₂S), which is a colorless gas characterized
by an odor similar to that of a rotten egg. \( \text{H}_2\text{S} \) is produced by the anaerobic breakdown of organic matter which occurs in the pit containing hog waste as well as being emitted when that waste is dispersed. The odor threshold of \( \text{H}_2\text{S} \) has been reported from 0.5-300 ppb. Exposure to \( \text{H}_2\text{S} \) at moderate to high concentrations can cause eye and throat irritation, cough, and nausea, and acute exposure to very high concentrations of \( \text{H}_2\text{S} \) can result in death (ATSDR 2006). Occupational exposure limits are not to exceed 20 ppm, however little is known regarding the effects of chronic low level exposure to the compound, especially in the presence of other air pollutants (OSHA 2014; ATSDR 2006).

This study was part of a larger project in which researchers partnered with a local community organization in eastern North Carolina. Pollutant data, biological samples, and residents’ own accounts of their health over time were collected to investigate the impacts of industrial hog operations on a community living with a high density of CAFOs. \( \text{H}_2\text{S} \) was one exposure of interest in this study because accurate estimates of its concentration in space and time could be used to correlate with health effects reported by residents. While three single point monitors (SPMs) were available to record air pollutant measurements in real-time, we complemented these devices with Radiello \( \text{H}_2\text{S} \) passive samplers to provide greater spatial coverage across the study area in a cost-effective manner (Pavilonis et al. 2013; Radiello 2007). In this study, we use the average \( \text{H}_2\text{S} \) concentrations collected by the passive samplers to develop a land use regression (LUR) model to indicate whether CAFOs are the source of \( \text{H}_2\text{S} \) in the study area, and we then conduct a comparison of methods to find which one produces the most accurate estimates of \( \text{H}_2\text{S} \) across space and time.
2.2 Materials and Methods

2.2.1 Sampling

We deployed 67 Radiello passive samplers designed to collect H$_2$S at 53 unique space/time locations from January to June 2007. The leadership of our partner community group provided expertise regarding sampling locations and connected us with individuals who were willing to have passive samplers placed on their property. Each passive device was placed approximately 1.5-2.0 meters above the ground (breathing height) during one of six 2-3 week sampling phases. The coordinates of each passive sampler’s spatial location were recorded with a Global Positioning System (GPS). Placement of passive samplers was designed to estimate several types of variability including device measurement error and the changes in H$_2$S concentration over space and time.

**H$_2$S sampling locations by type**

<table>
<thead>
<tr>
<th>Sampling Phase Number</th>
<th>Number of Unique S/T Measurements</th>
<th>Triplicate*</th>
<th>Single</th>
<th>Edge Field</th>
<th>Blanks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>0</td>
<td>7</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>S/T Locations</td>
<td>53</td>
<td>7</td>
<td>18</td>
<td>28</td>
<td>--</td>
</tr>
<tr>
<td>Samplers</td>
<td>72</td>
<td>21</td>
<td>18</td>
<td>28</td>
<td>5</td>
</tr>
</tbody>
</table>

*Each triplicate measure represents 3 samplers in one spatial location

**Table 2.1** The number of passive samplers used during each of 6 sampling phases. A total of 72 samplers measured at 53 space/time locations with triplicate and single samplers measuring at spatial locations during one time period and edge field measurements occurring at spatial locations over multiple time periods.
Each sampler used in the field can be categorized as a triplicate, single, edge field, or field blank (Table 2.1). Seven of the space/time sites had samplers placed in triplicate in order to quantify the sampling error. These triplicate sites, as well as the single sampler locations, were only monitored during one sampling phase and placed in reasonable proximity (within a few km) to one another for a given sampling period where residents would allow us access to their property. There was one area where we were able to gain access to land adjacent to a CAFO waste pit. In this “edge” field, we placed devices during each of the 6 sampling periods in order to analyze the distance to source relationship as well as the changes in H$_2$S from one period to the next. Finally, a field blank was used during the collection period of each phase to quantify any contamination during transport (except for Phase 5 when the field blank was damaged).

The lab analysis instructions provided by Radiello were followed in order to obtain the mass of sulfide ions (ug) in each sample (see Appendix B) (Radiello 2007). The mass of sulfide ions from each passive device’s corresponding field and lab blanks, which characterized any contamination during travel and analysis, respectively, were removed from this value to obtain the corrected mass of sulfide ions in the exposed sample, which we denote as $M_{ij}$ for the $i$-th sampler collected during the $j$-th sampling period. The average H$_2$S concentration (ppb) for each sample ($Z_{ij}$) was then calculated using the average temperature (in degrees Kelvin) during the sampling time $T_{ij}$ (in minutes) using Eq. 2.1.

$$Z_{ij} = \frac{1000 \times M_{ij}}{0.096^{\frac{T_{ij}}{298}}} T_{ij}$$ (2.1)

### 2.2.2 Public Data Sources

We downloaded the spatial database of CAFOs which is maintained by the North Carolina Department of Environment and Natural Resources (NCDENR) and used ArcGIS 9.2 to
select the facilities from this database listed as swine CAFOs. In addition to the latitude and longitude coordinates of these operations, information regarding the approximate number of animals and the number of lagoons was also available (NCDENR 2007).

Hourly weather data was obtained for the weather station closest to the study area that reports to the National Oceanic and Atmospheric Administration’s (NOAA) National Climatic Data Center. Available weather variables from this database include temperature, dew point, relative humidity, wind speed, wind direction, pressure, and precipitation (NOAA 2011). Since there was only one station that was reasonably close to our passive sampler locations, hourly temperature, dew point, relative humidity, pressure, and precipitation were averaged over the collection period for each respective sampling device.

2.2.3 Land Use Regression Model

A Land Use Regression (LUR) model was developed to test the hypothesis that CAFOs are a source of atmospheric H₂S in the study area, to estimate the distance (or range) over which a single CAFO contributes to the H₂S concentration, and to serve as a space/time offset to remove from the observed data for the purpose of geostatistical estimation (Messier et al. 2012). We denote the space/time random field (STRF) of the instantaneous H₂S concentration as \( Y \) and the time-averaged concentration of H₂S as \( Z \). The variable \( Z \) is measured for passive sampler \( ij \) at spatial coordinate \( s_i = (\text{longitude}_i, \text{latitude}_i) \) and time \( t_j +/− T_{ij} / 2 \), where \( T_{ij} \) is the length of time (in minutes) that sampler was exposed in the field. Thus, with \( u \in (t_j − \frac{T_{ij}}{2}, t_j + \frac{T_{ij}}{2}) \) we obtain
The concentration of time-averaged H$_2$S for passive sampler $ij$ can then be expressed as dependent on a set of explanatory variables $X^{(m)}_{ij}$ variables so that:

$$Z_{ij} = \beta_0 + \beta_1 X_{ij}^{(1)} + \beta_2 X_{ij}^{(2)} + \cdots + \beta_m X_{ij}^{(m)} + \epsilon_{ij} \quad (2.2)$$

, where $\beta_1, \ldots, \beta_m$ are the linear regression coefficients for explanatory variables $X_{ij}^{(1)}, \ldots, X_{ij}^{(m)}$, respectively, and $\epsilon_{ij}$ is the error term associated with the respective sample. We expect the primary explanatory variable of interest to be the sum of wind-weighted exponential decay functions (Eq. 2.3), where each individual exponential decay function represents a contribution from a surrounding hog CAFO. For passive sampler $ij$, the contribution of CAFO $k$ at hour $l$ is given a weight $w_{ijkl}$ defined in Eq. 2.4 such that the sum of the exponentially decaying contributions from CAFOs is equal to:

$$X_{ij}^{(1)} = \sum_{l=1}^{h_{ij}} \sum_{k=1}^{n} w_{ijkl} \exp\left(-3 \frac{D_{ijk}}{a_r} \right) \quad (2.3)$$

$$w_{ijkl} = \frac{1 + \cos(\alpha_{ijkl})}{2} \quad (2.4)$$

, where $D_{ijk}$ is the distance (in km) between sampler $ij$ and CAFO $k$, $a_r$ is the decay range of the exponential function (in km, where H$_2$S declines to 5% of the concentration produced at the CAFO), and $\alpha_{ijkl}$ is the difference between the wind direction (in degrees at time $l$) and the direction from CAFO $k$ to sampler $ij$. For example, if a location is directly downwind from a CAFO, it is given a weight of 1 whereas a location directly upwind from that CAFO is given a weight of 0. Additionally, $n$ is the number of CAFOs in North Carolina and $h_{ij}$ is the number of hours that sampler $ij$ was exposed. The rest of the explanatory variables were chosen through stepwise regression as those weather variables that were statistically significant ($\alpha = 0.05$). The
spatial range, \( a_r \), was chosen as the one which optimized the coefficient of determination \((R^2)\) of the full model. Since only one NOAA weather station was reasonably close to the study area, the non-wind weather variables do not change over space and their average over time \( T_{ij} \) was used. The set of estimated regression coefficients \((\hat{\beta}_0, \ldots, \hat{\beta}_m)\) can be used to construct the LUR model for H\(_2\)S, which for any spatial location \( s \) and time \( t \) can be expressed as:

\[
L_z(s, t) = \hat{\beta}_0 + \hat{\beta}_1 X^{(1)}(s, t) + \hat{\beta}_2 X^{(2)}(s, t) + \cdots + \hat{\beta}_m X^{(m)}(s, t) \quad (2.5)
\]

### 2.2.4 Space/Time Estimation of Hydrogen Sulfide

We utilized six methods for estimating the average concentration of atmospheric H\(_2\)S in the study area. The first method, which we refer to as constant mean (CM), assumes that the H\(_2\)S concentration at any space/time location can be reasonably approximated by the average concentration of all passive samplers. We utilize this method as the null hypothesis that hog CAFOs do not contribute to atmospheric H\(_2\)S and there is simply a background concentration present throughout the study area. Our second method (LUR) uses the space/time LUR model (Eq. 2.5) to estimate atmospheric H\(_2\)S. The model takes into account the contribution of the sources as well as the space/time varying wind direction, and the time-varying weather variables. While it uses the space/time data for the purpose of developing the global space/time LUR model, this method does not allow for geostatistical estimation where the data also serves to influence the estimation of H\(_2\)S in the local space/time neighborhood of their location.

Our third and fourth methods allow for local influence from each space/time data point by using ordinary space/time kriging. Kriging is a widely-used geostatistical method that produces the best linear unbiased estimator (BLUE) at a given unmeasured location by weighting
the data according to a space/time covariance function. We use a global space/time offset 
\( o_{ij}(s,t) \) for each sampler value to create the \( H_2S \) offset data \( (X_{ij}) \) (Eq. 2.6).

\[
X_{ij} = Z_{ij} - o(s_i, t_j) \quad (2.6)
\]

We then model the variability and uncertainty associated with \( X(s, t) \) using a
homogeneous/stationary STRF for which the set of observed values \( X_{ij} \) represents one realization.

For the third method, which we refer to as kriging-constant mean offset (K-CM), we offset each
space/time data point by the average measured value of \( H_2S \) as in the CM method. In the fourth
method which we refer to as kriging-LUR offset (K-LUR), we of
set each data point by the
space/time LUR model in Eq. 2.5. A space/time separable exponential covariance model was fit
to the offset data (Eq. 2.7), which captures the space/time variability of the STRF.

\[
c_X(r, \tau) = c_1 \exp \left( -\frac{3r}{d_{r1}} \right) \exp \left( -\frac{3\tau}{d_{\tau1}} \right) \quad (2.7)
\]

A least squares approach was used to fit a model for each offset method (CM and LUR). The
kriging methods allow local influence by the measured data and assume that each measured data
point is the true value of \( H_2S \) and that it is measured without error (hard data).

The fifth and sixth methods attempt to account for errors in the measurements given by
the passive samplers. We account for two types of errors: (1) measurements that fall below the
detection limit of the sampling device and (2) the measurement error of readings that are above
the detection limit. The detection limit is defined as the average blank concentration plus 3 times
the blank standard deviation (IUPAC 1997). For measurements that fall below the detection
limit, we replace them by a truncated lognormal distribution where the non-truncated lognormal
distribution is obtained as the best fit to measured values \( Z \). This distribution is truncated at the
detection limit and the portion of the probability density function (pdf) below the detection limit is re-normalized (Messier et al. 2012).

For the measurements above the detection limit, we created a measurement error model (MEM) that uses the field triplicates and lab and field blanks to predict the standard deviation (Wilson and Serre 2007).

\[ \sigma_{Z_{ij}} = \sigma_0 + kZ_{ij} \quad (2.8) \]

The MEM (Eq. 2.8) shows that we expect the standard deviation of the time-averaged H\textsubscript{2}S measurement from passive sampler \( ij \) \( (\sigma_{Z_{ij}}) \) to be equal to the error associated with very small values \( (\sigma_0) \) plus a coefficient \( k \) that increases the measurement error as the measured value \( Z_{ij} \) increases. Since concentration cannot be negative, a Gaussian distribution with mean equal to the measured value \( Z_{ij} \) and variance given by our MEM was truncated below 0 and re-normalized. A space/time separable exponential/exponential covariance model was again fit, this time to the hardened value from our distributional (soft) data.

The Bayesian Maximum Entropy (BME) framework and its mathematical implementation were used to process the general knowledge (ie. mean and covariance), and non-Gaussian soft data for the STRF of \( X \) and to obtain a posterior PDF characterizing \( X \) at estimation points distributed across space and time in the study area (Christakos et al. 2002; Christakos 2000; Serre and Christakos 1999; Christakos 1990). The median H\textsubscript{2}S concentration and the lower and upper bound of the 68\% confidence interval are obtained by adding back the offset for the corresponding methods used for the respective BME-CM and BME-LUR methods.

The numerical implementation of this work was done using the BMElib package written in MATLAB\textsuperscript{TM} (Christakos et al. 2002; Serre and Christakos 1999). The library makes it easy to
integrate the general and site specific knowledge bases developed in this work, and to obtain the BME median estimate of H$_2$S concentrations.

2.2.5 Comparison of Methods

A validation was performed to compare the estimation performance among the six methods. For the non-geostatistical methods (CM and LUR), the mean squared error (MSE, defined in Eq. 2.9) for method $q$ is calculated using the difference between the observed value ($Z_{ij}$) and the model estimate ($Z_{ij}^{*(q)}$) at all space/time locations where data was measured ($m$) by a passive sampler.

$$MSE^{(q)} = \frac{1}{m} \sum_{i=1}^{m} (Z_{ij}^{*(q)} - Z_{ij})^2 \quad (2.9)$$

For the four geostatistical methods, each point $Z_{ij}$ was removed from the dataset and that space/time location was re-estimated using a given method $q$, yielding the estimate $Z_{ij}^{*(q)}$.

2.3 Results

2.3.1 Summary

The blank-corrected measured passive sampler data (N=67) ranged from 0.17 – 3.24 ppb for average H$_2$S concentration and the spatial distance to nearest CAFO as given by the NCDENR ranged from 0.10 – 1.83 km. The average value recorded by our blank samples was 0.07 ppb. A distance to source relationship is evident for the un-modeled H$_2$S data, however there were distinct differences from one sampling period to the next, indicating that there are time-dependent factors which are important to explaining H$_2$S (Figure 2.1).
The raw distance to closest source relationship (not accounting for wind or other factors) shows that H₂S concentration and distance to closest CAFO are inversely related. The strength of this relationship changes due to time-varying components and contribution from other CAFOs located further out.

2.3.2 Land Use Regression Model

Explanatory variables and their first-order interactions were included in the LUR model if they were statistically significant ($\alpha=0.05$) by the stepwise inclusion/exclusion method using PROC REG in SAS Version 9.2. The resulting components of the global LUR model are given in Table 2.2. We evaluated the full LUR model for different distances of the CAFO decay range parameter $a_r$ (Eq. 2.3) ranging from 0-20 km at intervals of 0.1 km. The maximum $R^2$ value of 0.62 occurred at a spatial range of 1.0 km. The CAFO sources (sum of wind-weighted exponential decay functions) were deemed to be the most significant variable in the model. All five of the NOAA weather variables were also statistically significant. The CAFOs, temperature,
relative humidity, pressure, and precipitation were positively associated with H$_2$S whereas dew point was negatively associated with H$_2$S.

**Statistically significant variables in H$_2$S LUR model**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>F-Statistic</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-268.04 (ppb)</td>
<td>19.71</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Unit-less Sum of Wind-Weighted Exponential Decay Functions with a$_{r}$ = 1.0 km</td>
<td>2.64 (ppb)</td>
<td>40.56</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Temperature (F)</td>
<td>0.99 (ppb)</td>
<td>10.53</td>
<td>0.0022</td>
</tr>
<tr>
<td>Dew Point (F)</td>
<td>-1.06 (ppb)</td>
<td>9.49</td>
<td>0.0035</td>
</tr>
<tr>
<td>Relative Humidity (%)</td>
<td>0.58 (ppb)</td>
<td>6.70</td>
<td>0.0129</td>
</tr>
<tr>
<td>Pressure (Pa)</td>
<td>7.32 (ppb)</td>
<td>12.39</td>
<td>0.0010</td>
</tr>
<tr>
<td>Precipitation (in)</td>
<td>197.10 (ppb)</td>
<td>8.39</td>
<td>0.0058</td>
</tr>
</tbody>
</table>

*Table 2.2* The statistically significant components in the LUR model for predicting time-averaged H$_2$S achieved an $R^2$=0.62.

We also limited our LUR model to H$_2$S data from the “edge” field where we were able to place 5 samplers close to a CAFO waste pit and at increasing distances in one direction so that the second closest operation was always at least 0.5 km further from the sampler than the target operation. We created an “edge” model consistent with the full global LUR model, but limited to data collected in the edge field, and accounting only for the nearest CAFO (i.e. $n=1$ in Eq. 2.3) since the edge field is mainly influenced by a single operation. This helped to ascertain whether the exponential decay function is a good model. The edge model had an $R^2$ value of 0.80 using a spatial range of 0.7 km and clearly demonstrates the off-site migration of H$_2$S (Figure 2.2).
H₂S LUR model restricted to “edge” field

Figure 2.2 The LUR model for the “edge” field which should have minimal influence from other CAFOs achieved a maximum $R^2$ of 0.80 for a spatial range of 0.7 km. The LUR model is valid for community level exposures greater than 100m from a CAFO.

2.3.3 Distributional (Soft) Data

Distributional (soft) data were created to implement our BME geostatistical estimation methods. Without respect to distance from CAFO, the observed $Z_{ij}$ values were lognormally distributed. A lognormal distribution with a mean from the associated normal distribution of -0.60 ppb and variance of 0.42 ppb was found to be the best fit pdf for all $Z_{ij}$ measured values. A version of this distribution that was truncated above the detection limit and re-normalized was used to replace observations that fell below the detection limit. We estimated a detection limit of 0.22 ppb, which was obtained as the mean of the blanks plus 3 times the standard deviation of the blanks. For observations above the detection limit, we used the MEM described in Eq. 2.8 to
estimate the standard deviation of the distributional (soft) data. The lab and field blanks were used to obtain the value $\sigma_0$ which represents the standard deviation at very low measured values of H$_2$S and our locations where samplers were placed in triplicate were used to estimate the coefficient $k$ which represents the increase in standard deviation as the measured value $Z_{ij}$ increases. Given that $\sigma_0$ was found to be 0.05 ppb, the least squares estimate for $k$ was 0.05, so that we obtain the standard deviation for measurement $Z_{ij}$ as $\sigma_{Z_{ij}} = 0.05 + 0.05 \cdot Z_{ij}$ (Figure 2.3). Using the MEM, we obtain a Gaussian distribution with a mean of $Z_{ij}$ and variance of $\sigma^2_{Z_{ij}}$ as the distributional (soft) data used to model the uncertainty associated with each measured value above the detection limit. Each of these distributions was truncated below 0, and re-normalized, making them non-Gaussian.

**Measurement Error Model for H$_2$S**

*Figure 2.3* A measurement error model (MEM) was used to obtain the standard deviation of the soft data that were above the detection limit of 0.22 ppb. The standard deviation of the blank values was used to estimate $\sigma_0$ (intercept) and the mean and standard deviations of locations with triplicates were used to estimate $k$ (slope).
2.3.4 Space/Time Covariance

The offset-removed data $X_{ij}$, where we use the subscript $cm$ when referring to the constant mean offset data and $lur$ for the land use regression offset data, had an experimental covariance that was exponential for each offset method in both space and time. The modeled covariance parameters for the constant mean offset had a sill ($c_{1,cm}$) of 0.44 ppb, a spatial range ($a_{r1,cm}$) of 0.95 km, and a temporal range ($a_{t1,cm}$) of 110 days. The parameters for the model fit to the land use regression data had a sill ($c_{1,lur}$) of 0.17 ppb, a spatial range ($a_{r1,lur}$) of 0.27 km, and a temporal range ($a_{t1,lur}$) of 73 days (Figure 2.4).

Covariance model for H$_2$S – LUR offset

![Covariance model for H$_2$S – LUR offset](image)

Figure 2.4 The experimental and modeled space/time separable covariance function for the H$_2$S LUR offset data.
2.3.5 Exposure Mapping Results

As expected there were distinct differences in the visual representations of the maps produced by the different offsets. As depicted in Figure 2.5, which shows H$_2$S estimates around observed values (sampling locations) for phase 4 where all sampling locations were in reasonably close proximity, the CM methods (left side of the figure) estimate H$_2$S concentrations that are not obviously dependent on the locations of CAFOs. The geostatistical CM methods do show weak correlation with the CAFO locations based on the influence of higher measured values closer to CAFOs than those which are further away.

The LUR methods clearly show the contribution of H$_2$S attributed to CAFOs, with the highest concentrations corresponding to where CAFOs are clustered together. While there are differences among the estimates produced by the three LUR methods, it is clear that the good fit of the LUR model make these variations difficult to distinguish. Among these methods, the K-LUR and BME-LUR geostatistical approaches do a better job of matching the estimate to the measured values due to the neighborhood influence of each passive sampler as evidenced by the results of the validation.
Maps of H$_2$S concentration using different methodologies

### Constant Mean Methods

- **H$_2$S Constant Mean: Phase 4**

### Land Use Regression Methods

- **H$_2$S Land Use Regression: Phase 4**

**Figure 2.5** Estimated and observed (at sampling location) concentrations of atmospheric H$_2$S using six different methods during sampling phase 4. The methods in the left column use a constant mean offset and concentration estimates are not clearly related to CAFO location. On the right column, the LUR offset is used, illustrating the contribution of each CAFO and the effect of wind direction. The edge field is comprised of the 5 southernmost sampling locations.
2.3.6 Comparison of Methods

A validation was performed to compare the estimation performance of the different methods (Figure 2.6). Using the LUR model as opposed to the CM model without geostatistical estimation reduced the relative MSE by 62%, as indicated by the coefficient of variation. Switching from the CM offset to the LUR offset reduced the relative MSE by 24% for the kriging methods and by 25% for the BME methods. Within methods which used the CM offset, K-CM reduced the MSE by 54% relative to CM and BME-CM reduced the MSE by 54% relative to CM. Within the methods that used the LUR offset, K-LUR reduced the MSE by 8% relative to LUR and the BME-LUR method reduced the MSE by 9% relative to LUR. Our best-performing estimation method, BME-LUR, reduced the MSE by 65% relative to the worst-performing method, CM.

**Comparison of mean squared error for methods estimating H$_2$S**

![Comparison of Mean Squared Error](image)

*Figure 2.6* A comparison of mean squared error (MSE) for the estimation of H$_2$S using the different methods. Using the LUR model causes a relative reduction in the MSE of approximately 62%. Geostatistical estimation makes large reductions in MSE even with the constant mean offset, but the combination of distributional (soft data) and the LUR offset provides the most accurate H$_2$S estimates.
2.4 Discussion

The H$_2$S data reported in our analysis is original data that is part of a sampling plan for characterizing the STRF of this compound in an area with a high density of hog CAFOs by using passive samplers that record average H$_2$S concentrations over space and three active single point monitors that record the values every 15 minutes. Due to the limitations of the active monitors (price and power supply), the passive samplers are a cost-effective way to gauge the variability across space. In this manuscript, we tested the hypothesis of whether industrial hog CAFOs are producing off-site atmospheric H$_2$S and then quantify the exposure to the community using a LUR model and geostatistical estimation.

The passive samplers provide us with the time-averaged concentration of H$_2$S for a given spatial location. This misses the temporal variation of H$_2$S which is characterized by long periods of very low values with short-term spikes in concentration. Periods of higher hourly concentration are correlated with residents’ reports of odors and eye and throat irritation (Schinasi et al. 2011; Wing et al. 2008). However, the passive samplers do a better job of capturing the spatial variation and help illustrate the influence of the CAFOs. While there has not been a lot of research to date on the low-level chronic exposures from CAFOs, it is important to consider these impacts on the general population and H$_2$S can be considered a possible marker for a mixture of contaminants affecting residents.

Our estimated sampler detection limit of 0.22 ppb, which was calculated as three times the mean of the lab and field blanks was very close to the detection limit of 0.20 ppb reported in a lab analysis (Fujita et al. 2009). Likewise, the coefficient ($k$) of our MEM was 0.05 – again very close to their reported value of 0.04. Thus, we used the detection limit and measurement error estimated from our study since the small deviations could be associated with field
conditions. We utilized these error estimates in our BME soft data since using distributional data rather than a hardened value has been shown to improve estimation performance when analyzing the space/time distributions of other environmental contaminants, utilizing land use regression (Messier et al. 2012).

One type of error which probably had an impact on our results was the publicly-listed location of our H₂S sources in the CAFO shapefile constructed by NCDENR. Each facility has a spatial point of latitude and longitude coordinates but does not give information about the spatial extent of the operation, the location of the waste pit, or the exact locales of the spray fields on which the waste is emptied. Even with the limitations on source locations, our LUR model performed well. The 62% reduction in MSE relative to a constant concentration of H₂S across space and time, gives a strong indication that the CAFOs are the source of atmospheric H₂S in the area.

There are some caveats regarding time-varying components in our model. The H₂S data was collected from January-June 2007 and since we did not have an entire year of data, we did not attempt to estimate any seasonal changes in the concentration. It is likely that there is an effect based on time of year, but this could be confounded with changes in other weather variables which we did incorporate into the model. The LUR model quantifies the effects that weather variables had on the data that we measured. However, the responding regression coefficients should not be extrapolated to other situations or other time periods.

Wind direction was used to calculate an hourly weight for every CAFO for a given spatial location. There was also available data regarding wind intensity, but it was not used in our LUR model since the overall effect would be very difficult to capture in a model as an increase in wind velocity should increase the range of influence of a single CAFO, but at the same time
disperse the H$_2$S more rapidly. Therefore, we ignored the wind intensity and expect the prevailing direction regardless of intensity to be the most important wind factor.

Other studies have investigated H$_2$S on-site and at select off-site locations where most measurements were made within 40m of an operation (Pavilonis et al. 2013; Thorne et al. 2009). Our closest sampler was 100m from a CAFO as listed by the NCDENR database and our results are therefore intended to model community-level exposures of H$_2$S for residents living >100m from a CAFO. Indeed, the concentrations that we found at the closest distances to a CAFO (between 100m to 200m) were in line with those from other studies. Our global model estimates that the contribution of one CAFO to the community living at a distance $D$>100m from that CAFO is approximately 2.64 ppb multiplied by an exponentially decaying factor equal to $\exp(-3 \frac{D}{1.0 \text{ km}})$. For example, this contribution corresponds to an increase in the long-term average concentration of H$_2$S of about 2 ppb for a resident living 100m from a CAFO, a substantial increase in chronic exposure that may represent the impacts of high peaks of concentrations over short durations. Our exponential model should not be used for distances less than 100m (dashed lines, Figure 2.2) because at these short distances our model estimates values that are much less than what was observed by others near the CAFO operation indicating that our model would greatly underestimate occupational exposures at the confinement buildings.

Temperature, relative humidity, and precipitation were positively associated with H$_2$S concentration and a combination of these variables could possibly indicate the expected seasonal effect when H$_2$S increases during the warmer months. The positive influence of pressure on H$_2$S concentration is expected as it produces more stable weather conditions where the compound is not quickly being dispersed.
We believe our community-based study is one of the first to demonstrate the off-site migration of H$_2$S from hog CAFOs and to develop a space/time model for estimating H$_2$S in an area affected by a large number of operations. While there have been a few instances where H$_2$S levels have been measured at CAFOs, we have been able to estimate a clear exponential decay relationship that exists away from the operation, affecting the surrounding community. While our study area has a high density of these facilities, we developed a sampling plan so that we could reasonably estimate the contribution from one facility by choosing a location where we could get particularly close to the waste pit and place samplers at increasing distances without getting close to another operation. While we are not able to account for the proximity to spray fields in this manner, we are able to demonstrate the off-site migration of H$_2$S. Samplers at other space/time locations allowed us to build a global land use regression model that can estimate the average concentration of atmospheric H$_2$S in this area with a high density of hog CAFOs. The model shows that while proximity to CAFOs is the most important predictor of H$_2$S, other time-varying factors such as changes in weather are important explanatory variables.

We created a separate LUR model for our edge field in order to quantify the H$_2$S contribution from one facility as well as to help evaluate the performance of our global LUR model. The edge model achieved a higher $R^2$ (0.80 vs. 0.62) and a shorter optimal spatial range (0.7 km vs. 1.0 km). While this could simply be due to characteristics of the particular operation, we would expect to find a lower coefficient of determination and longer spatial range when space/time sampling locations are affected by multiple operations. Given that the isolated and global models are reasonably close on these two measures, we have high confidence that our global LUR model is performing well.
Industrial hog operations produce many pollutants that are relevant to human health and quality of life, including ammonia, volatile organic compounds, and endotoxin (Cole et al. 2000). Because many of these are ubiquitous in the ambient environment, H$_2$S is a specific marker of hog operation emissions where other industrial sources of H$_2$S such as petrochemical plants, paper mills, and asphalt plants, are not present, such as in our study area. A growing literature clearly documents acute impacts of the mixture, however chronic exposures, which have received little attention, may also impact human health. A prior study found that industrial hog operation neighbors reported hog odor inside their homes on 12.5% of study days, however indoor H$_2$S concentrations were not measured (Wing et al. 2008). This indoor exposure, which represents intrusion of industrial hog operation air pollutants into homes, suggests that the spatial estimates of outdoor concentration reported here may indicate a similar spatial pattern of indoor concentrations that would also depend on housing structures and ventilation.

Community-based participatory research was an important component of our H$_2$S data collection. While we could have placed devices along roads and in other locations that are publically accessible, this convenience sampling would have made it difficult to get reasonably close to the confinement buildings (Wilson and Serre 2007). The edge field, where we were able to get close to the confinement buildings and waste pit, was privately-owned land. These samples included the minimum distances to the CAFO locations and were important for estimating H$_2$S concentrations at short distances for the purpose of our LUR model. Also, by sampling in locations near where someone is living, we are able to obtain exposures estimates where someone is actually being exposed.

There have been many issues resulting from the rapid adoption of the industrial model for livestock production, but if we make the assumption that this method is here for the foreseeable
future, a change in the way waste is dealt with in North Carolina should be a high priority. There are state regulations regarding when and how waste pits can be emptied, but storing hog waste in open pits and spraying it on nearby fields threatens public health in high-density production areas.
REFERENCES

Agency for Toxic Substances and Disease Registry (2006) Toxicological Profile for Hydrogen Sulfide, Atlanta, GA


North Carolina Department of Environment and Natural Resources (2007) Hog CAFO Inventory, Raleigh, NC


Thu K (Ed.) (1996) Understanding the Impacts of Large-Scale Swine Production: Proceedings from an Interdisciplinary Scientific Workshop, The University of Iowa, Iowa City, IA


CHAPTER 3

SOURCES THAT MOVE IN SPACE AND TIME – THE BAYESIAN UNIFORM MODEL EXTENSION OF BME (BUMBME) FOR DELINEATING CORE AREAS OF SYPHILIS AND GONORRHEA

Overview

Infections are not normally viewed as pollutants in the same sense as an air, water, or soil contaminant, but they share many of the same characteristics. Sexually transmitted infections (STIs), such as syphilis and gonorrhea, are unwanted and have a negative effect on human health. STIs, in particular, have many of the properties of point sources since they often perpetuate themselves through core areas of infection. These areas have persistently higher rates that can develop into outbreaks or spread to nearby sexual networks. In order to delineate areas of concern on a fine space/time scale, Bayesian data using the long term incidence rate to inform the short term incidence rate was created at the census block group level and compared to other geostatistical disease mapping methods. In terms of public health, finding areas of concern with spatial precision and in a timely manner can improve individual health and decrease costs of intervention programs.
3.1 Introduction

Sexually transmitted infections (STIs) are unique from an intervention standpoint because they require close physical contact in order to move between individuals, increasing the length of time it takes for the infection to spread compared to agents that are transmitted through the air and on surfaces. Therefore, along with proper education and access to medical care, a targeted public health campaign should be able to significantly reduce rates.

Syphilis is caused by *Treponema pallidum* bacteria and those that have been exposed to this infection can eventually develop a genital ulcer. Untreated primary syphilis can lead to secondary syphilis which has moderate to severe disease symptoms including rash, fever, and even dementia. Gonorrhea, which is caused by exposure to *Neisseria gonorrhoeae*, has symptoms of disease that include urethritis, cervicitis, and prostatitis. Additionally, these infections can elevate the rate of HIV transmission between persons (Cohen 1998; Fleming and Wasserheit 1999).

Although rates of syphilis and gonorrhea have decreased over the years, they still remain higher than the targets set by the Centers for Disease Control, whom plans to eliminate syphilis in the United States (CDC 2006; CDC 2010). Higher STI rates are most problematic in the southeastern part of the United States, including North Carolina, where these infections are prevalent in both urban and rural areas, with some of the highest rates along the I-95 corridor that bisects eastern North Carolina (CDC 2010). This area is typically poorer, less educated, and has a higher percentage of minorities compared to the rest of the state. It also has less access to health care which could be helping to perpetuate the infections.

In order to make significant improvements in the current incidence rate, high resolution space/time maps are necessary to delineate core areas with maximum spatial accuracy and to
detect outbreaks in a timely manner. Core areas are locations with persistently high rates and include high risk individuals whom create a core of infection. A targeted intervention in these places can reduce rates by helping to eliminate the high rate in the core as well as prevent potential new cases in surrounding areas affected by the core (Rothenberg 1983). Likewise, using resources to stop outbreaks quickly can lower the chance that the infections will move to other sexual networks which in turn can spread the STI through space.

Routine surveillance is necessary for accomplishing these goals and has been standard for several STIs for many years (Eng and Butler 1997). All states monitor syphilis and gonorrhea incidence using case reporting and many states mandate reporting of positive tests. While good data has been collected on reported STI cases for some time, calculating accurate incidence rates are hindered by the fact that the system relies on individuals getting tested in a timely manner.

Simply using case data to create an incidence rate, assumed to be measured without error, is problematic. Aggregating the data over space and/or time can help stabilize the incidence rate, but the reduction in resolution decreases the ability to identify target areas where intervention is needed. Alternatively, summing cases over small space/time windows means that the incidence rate can be unstable due to a smaller denominator and one additional incident case can create a significant difference in the rate; an issue commonly referred to as the small number problem (Jones and Kirby 1980; Kennedy 1989).

As disease mapping has become more popular with the rise of GIS software, new methods are being implemented to address the small number problem. One such approach is Poisson kriging (PK) which uses a distribution for the rate at each space/time location rather than a point value (Oliver et al. 1998). The variance of the distribution is calculated as a normal approximation of the Poisson and increases with an increase in the risk and decreases with an
increase in the population used to calculate the incidence rate. The fact that this method is based on the Poisson distribution makes it desirable since this distribution has often been applied to model disease (Goovaerts 2005). This method works well when the risk can be well-estimated, but STIs rates are more likely to be based on the local historical rate and risk factors that are distributed heterogeneously. The Poisson kriging method is also constrained by the assumption that the incidence rate is normally distributed.

Therefore we implement two approaches that do not rely on an estimate of the risk to calculate the variance and do not use a Gaussian distribution. The first approach is a Uniform Model Extension for Bayesian Maximum Entropy (UMBME) which applies a uniform interval around the observed incidence rate where the variance is based on the population at risk. This method accounts for sampling variability and has recently been used to model HIV (Hampton et al. 2011). We then expand upon this idea by creating a Bayesian datum for each space/time location where the uniform interval is used to update a prior distribution that is based on the historical incidence rate. We refer to this novel approach as the Bayesian Uniform Model Extension for Bayesian Maximum Entropy (BUMBME).

We apply and compare four methods: space/time simple kriging (SK), PK, UMBME, and BUMBME for mapping syphilis and gonorrhea in the state of North Carolina. In order to compare the estimation performance of these methods in the light of the small number problem, we introduce the concept of an asymptotic validation. Given that we expect the incidence rate to be unstable for a small space/time window with a low population, we require that the population is sufficiently large for a space/time datum to be removed and re-estimated; evaluating how each method performs as we increase the population size needed to be included in the analysis.
Syphilis and gonorrhea cover a range of rates, with syphilis averaging 7 cases per 100,000 and gonorrhea averaging approximately 199 cases per 100,000 (NCDHHS 2007). Additionally, syphilis is more prone to misclassification bias that results from accurately estimating the infection date. These data were aggregated at the census block group (CBG) level for each 6 month incidence period, a high mapping resolution where the small number problem needs to be adequately addressed. The contrasting qualities of these STIs can help assess the advantages to different smoothing techniques and illustrate the importance of using an asymptotic metric for evaluating geostatistical estimation methods affected by the small number problem.

3.2 Methods

3.2.1 Datasets

Reported syphilis and gonorrhea cases in the state of North Carolina were obtained from databases maintained by the state’s Department of Health and Human Services. Syphilis cases were available from 1999 through 2010 and gonorrhea cases were available from 2005 through 2010. A gonorrhea case was included in our analysis if its location could be geocoded and it had a diagnosis date associated with it. A syphilis case was included if it could be geocoded, had a diagnosis date, and was categorized as primary, secondary, or early latent syphilis.

Cases were geocoded at the North Carolina State Health Department using ArcGIS 9.3 and matched to three geographic locators. The primary locator was created by the North Carolina Emergency Response System and contains point locations for North Carolina households. The secondary locator was created by the North Carolina Department of Transportation and contains street-level geographic data. The tertiary locator was created using ESRI’s 2006 Street Map
shapefile and is primarily used for locating residences with outdated street names, prisons and military bases. We used post office addresses when PO Boxes were provided. Military addresses could not be geocoded due to missing street addresses. The geocoded cases were then aggregated to the centroid of their CBG, which we believe was the smallest available census unit that could still provide a sufficient amount of anonymity. The block group boundaries were obtained from shapefiles produced by the U.S. Census Bureau and they correspond to those delineated by the 2000 Census.

The date of infection for gonorrhea was assumed to be close to the reported date of diagnosis. The infection date for syphilis cases was backdated from the reported diagnosis date by 45, 90, and 183 days for the stages of primary, secondary, and early latent syphilis, respectively. Late stage syphilis cases were removed from our analysis since there was too much uncertainty regarding the proper infection period.

3.2.2 Latent Rate and Asymptotic Validation

The incidence rate is defined as the number of new cases in the at-risk population divided by the time duration to enumerate the new cases. We seek to map the latent incidence rate (Eq. 3.1) for syphilis and gonorrhea based on the rate observed from the state database. Analogous to Hampton et al. 2011, we define the latent incidence rate for a CBG with centroid at location $s_i$ for duration $T$ (0.5 years), centered at time $t_j$ as:

$$ X_{ij} = \lim_{n_{ij} \to \infty} \frac{Y_{ij}}{n_{ij}/T} \quad (3.1) $$

, where $Y_{ij}$ is the number of new cases of disease and $n_{ij}$ is the population at risk at time $t_j$. The at-risk population for each CBG was determined to be 80% of the CBG population ages 15-60. The
observed incidence rate for CBG $i$ for time $t_j$ is defined as $R_{ij} = Y_{ij} / n_{ij}T$. The difference between the observed rate and the latent rate can be expressed as:

$$R_{ij} = X_{ij} + \varepsilon_{ij} \quad (3.2)$$

, where $\varepsilon_{ij}$ denotes the measurement error due to sampling variability in CBG $i$ for time $j$.

In this study, we compare four geostatistical approaches to modeling the sampling variability of syphilis and gonorrhea in North Carolina, each yielding an estimate of the latent rate which we refer to as $\hat{X}_{ij}^{(q)}$ for a given geostatistical method $q$. Using Eq. 3.1, the performance of method $q$ can theoretically be quantified by calculating the mean squared error (MSE) where

$$MSE^{(q)} = \lim_{n_{ij} \to \infty} \frac{1}{M} \sum_{i,j=1}^{M} (\hat{X}_{ij}^{(q)} - R_{ij})^2 \quad \text{with} \quad M \text{ referring to the number of space/time estimation locations.}$$

To practically evaluate a method, we introduce the concept of an asymptotic validation analysis as an approximation to the theoretical MSE by only using locations $ij$ such that $n_{ij}>n$ to calculate the MSE:

$$MSE_{R,n}^{(q)} = \frac{1}{M'} \sum_{i,j:M' \quad n_{ij}>n} (\hat{X}_{ij}^{(q)} - R_{ij})^2 \quad (3.3)$$

, where $M'$ is the number of space/time locations that satisfy $n_{ij}>n$. As $n$ increases, the observed incidence rate approaches the latent rate.

### 3.2.3 Simple Kriging, Poisson Kriging, UMBME

Rolling 6 month observed incidence rate data were created for each month at each CBG centroid for both syphilis and gonorrhea. Four methods incorporated these space/time incidence rate data to produce a geostatistical estimate of the incidence rate across North Carolina from January 2005 through January 2010 for syphilis and July 2008 through January 2011 for gonorrhea. We utilized the BME computational library for producing the space/time geostatistical estimation due to its rigorous nonlinear mathematical framework that includes the
ability to incorporate both Gaussian and non-Gaussian data (Christakos 1990; Serre and Christakos 1999; Christakos 2000; Christakos et al. 2002; Christakos et al. 2005). BME is a two-stage process which uses maximum entropy to organize the general knowledge ($G$) of the STRF (such as space/time mean trend and covariance functions) to compute a prior probability density function (PDF) of the space/time process. The prior PDF is then updated by the site-specific knowledge ($S$) which can include hard (measured without error) and soft (characterized by a PDF) data to produce a posterior PDF that characterizes the STRF at any space/time point. When only hard data is available, BME produces the simple space/time kriging estimator.

In our first method, we assume that the measurement error $\varepsilon_{ij}$ is 0 and the observed rate $R_{ij}$ is equal to the latent rate $X_{ij}$. In this case, BME reduces to the space/time simple kriging (SK) method of linear geostatistics, which produces the best linear unbiased estimator at an unmeasured location from available measured data that are weighted ($\lambda_{ijk}$) based on the kriging system of equations (Eq. 3.4).

$$\hat{X}_{ij}^{SK} = \sum_{k=1}^{K} \lambda_{ijk} R_k \quad (3.4)$$

We refer to this as the hard data method, because it assumes that the BME site-specific data is measured without error and thus is equal to the true rate.

While SK is the most basic and common mapping approach, we use Poisson kriging (PK), extended to space/time, as a current disease mapping method which accounts for the small number problem by smoothing the observed rate based on the risk. Poisson kriging, assumes that the observed number of cases $Y_{ij}$ follows a Poisson distribution so that $Y_{ij} \mid X_{ij}, n_{ij} \sim \text{Poisson}(n_{ij}TX_{ij})$ where the parameter $n_{ij}TX_{ij}$ is the expected number of cases for a given space/time location (Goovaerts 2005; Goovaerts and Gebreab 2008). The measurement error $\varepsilon_{ij}$ is assumed to be Gaussian distributed with a variance $\sigma_{ij}^2 = m_j^*/n_{ij}T$ where $m_j^*$ is the
population-weighted mean of the observed rates for time $j$. Space/time Poisson kriging solves a system of equations equivalent to Eq. 3.4 and is implemented in the BME package as Gaussian-distributed soft data.

The third method, UMBME, assumes that since the observed rate can only increase in increments of $1/n_{ij}$, an alternative model for the measurement error $\varepsilon_{ij}$ is a bounded uniform distribution (Eq. 3.5) which represents the sampling variability (Hampton et al. 2011). The UMBME method treats each observation as a representative sample of the latent rate that has a measure of variability dependent on the size of the population $n_{ij}$ without making the assumption that the risk of disease is the same over the entire study area.

$$X_{ij} \mid R_{ij}, n_{ij} \sim \text{Uniform} \left( R_{ij} - \frac{0.5}{n_{ij}T}, R_{ij} + \frac{0.5}{n_{ij}T} \right) \quad (3.5)$$

### 3.2.4 BUMBME

The fourth method, which we introduce here, expands on the concept of UMBME by using it to derive a likelihood for a Bayesian soft datum (Eq. 3.6), and thus we refer to it as BUMBME. We seek to map the latent rate based on the observed incidence rates and their corresponding populations as well as the observed long-term incidence rate. We refer to the observed long-term incidence rate as $W_{ij}$ for CBG $i$ which ends at time $t_j-3$ months (ends at the beginning of the corresponding observed incidence rate) over duration $T_L$. The observed long-term incidence rate is calculated as $W_{ij} = \frac{U_{ij}}{n_{ij}T_L}$ where $U_{ij}$ is the number of cases over the longer incidence duration period and $T_L$ is 36 months for gonorrhea and 60 months for syphilis. Using Bayes’ Theorem, we can calculate the latent rate distribution $f_S(X_{ij} \mid R_{ij}, W_{ij}, n_{ij})$ based on the observed incidence rate, the observed long-term incidence rate, and the population according to Eq. 3.6. In the Bayesian context, the distribution $f_S(X_{ij} \mid W_{ij}, n_{ij})$ is referred to as a prior.
distribution of the latent rate, \( f_S(\mathbf{R}_{ij} | \mathbf{X}_{ij}, W_{ij}, n_{ij}) \) is a likelihood function, and \( A = \int d\mathbf{x}_{ij} f_S(\mathbf{R}_{ij} | \mathbf{X}_{ij}, W_{ij}, n_{ij}) \) is the normalization constant.

\[
    f_S(\mathbf{R}_{ij} | \mathbf{X}_{ij}, W_{ij}, n_{ij}) = A^{-1} f_S(\mathbf{R}_{ij} | \mathbf{X}_{ij}, W_{ij}, n_{ij}) f_S(\mathbf{X}_{ij} | W_{ij}, n_{ij}) \tag{3.6}
\]

We model the prior distribution for the latent rate \( f_S(\mathbf{X}_{ij} | W_{ij}, n_{ij}) \) by plotting each space/time location’s observed incidence rate against its corresponding observed long-term incidence. For a given space/time location \((i'j')\), the mean and variance (Eq. 3.7) of the distribution is first obtained non-parametrically, plotting the average long-term incidence rate against an average observed incidence rate that is weighted based on the population at risk for a moving window of +/- 10 cases per 100,000 person-years, i.e.

\[
    \mathbb{E}[\mathbf{X}_{i'j'} | \mathbf{W}_{i'j'}, n_{i'j'}] = \frac{\sum_{i,j} n_{ij} R_{ij}}{\sum_{i,j} n_{ij}}, \text{ and } \tag{3.7a}
\]

\[
    \text{Var}[\mathbf{X}_{i'j'} | \mathbf{W}_{i'j'}, n_{i'j'}] = \frac{\sum_{i,j} n_{ij} (R_{ij} - \mathbb{E}[\mathbf{X}_{i'j'} | \mathbf{W}_{i'j'}, n_{i'j'}])^2}{\sum_{i,j} n_{ij}} \tag{3.7b}
\]

where the summation is for all \(ij\) that are such that \(W_{ij}\) is within +/- 10 cases per 100,000 person-year of \(W_{i'j'}\).

For long-term incidence rate values that lie beyond the extent of the non-parametric model, a simple linear regression was fit to the plot of observed long-term incidence rates against observed incidence rates to obtain an estimate of the mean and variance. We use the mean and variance from Eq. 3.7 to parameterize the Bayesian prior with a lognormal distribution.

Next we construct the likelihood function \( f_S(\mathbf{R}_{ij} | \mathbf{X}_{ij}, W_{ij}, n_{ij}) \) with an argument similar to Eq. 3.5 based on the uncertainty due to sampling error (Eq. 3.8). Since, the observed incidence rate is independent of the long-term incidence given the latent rate and population, the likelihood function can be reduced to \( f_S(\mathbf{R}_{ij} | \mathbf{X}_{ij}, n_{ij}) \). Given the parameters of the Uniform distribution, the observed incidence rate \( R_{ij} \) is within \(1/n_{ij}\) of \(X_{ij}\) so that:
\[ R_{ij} - \frac{0.5}{n_i} \leq X_{ij} \leq R_{ij} + \frac{0.5}{n_i} \]  

(3.8)

By simply adding and subtracting terms to Eq. 3.8, we get that:

\[ X_{ij} - \frac{0.5}{n_iT} \leq R_{ij} \leq X_{ij} + \frac{0.5}{n_iT} \]  

(3.9)

Thus, we use a Uniform distribution for the likelihood function, so that

\[ R_{ij} \mid X_{ij}, n_{ij} \sim \text{Uniform} \left( X_{ij} - \frac{0.5}{n_iT}, X_{ij} + \frac{0.5}{n_iT} \right) \]  

(3.10)

The Bayesian derived posterior distribution from Eq. 3.6 is non-Gaussian, again requiring BME to produce the estimates.

### 3.2.5 Space/Time Offset and Covariance

We define the offset removed data \( X'_{ij} = X_{ij} - o_{ij} \) by removing a space/time offset \((o_{ij})\) from each data point \( ij \), so that the site-specific datum \( f_s(X_{ij}) \) for each method is transformed to \( f_s(X'_{ij}) = f_s(X_{ij}) - o_{ij} \). The space/time offsets for SK, PK, and UMBME are the same since each site-specific datum is centered on the observed value whereas the BUMBME offset is based on the mean of the BUMBME posterior distribution from Eq. 3.6. We then model the variability and uncertainty associated with \( X'_{ij} \) using a homogeneous/stationary STRF for which the set of observed values \( X'_{ij} \) represents one realization. A space/time covariance was fit to the offset-removed data for each method by minimizing the difference between the experimental covariance and a theoretical function that was assumed to be space/time separable with a short-range exponential spatial component, a long-range exponential spatial component, a short-range Gaussian temporal component whose range was set to 6-months (due to the rolling 6-month window) and a long-range exponential temporal component. The experimental covariance was calculated using the mean of the distribution of each space/time incidence rate datum.
The four geostatistical methods were compared qualitatively according to their attributes such as creating a realistic map of incidence rates on such a fine space/time scale and the amount of smoothing during both stable and outbreak infection periods. In addition, we evaluated their performance quantitatively on the asymptotic validation as defined in Eq. 3.3.

3.3 Results

3.3.1 Summary

In general, observed syphilis rates were the highest in the eastern part of North Carolina and overall rates declined from 2000 until about 2006 before increasing again. There does not appear to be a general change in statewide gonorrhea rates over time and the locations with the highest gonorrhea rates seem to maintain these high rates. Urban areas as well as sections of the rural northeast and rural southeast have the highest observed incidence of gonorrhea. As with syphilis, the mountainous areas in the western part of North Carolina have very low observed rates compared to the rest of the state.

3.3.2 Model for BUMBME Prior Distribution

For BUMBME, a nonparametric model was developed to create a prior distribution of the current incidence rate based on a CBG’s long term incidence rate – 5 years for syphilis and 3 years for gonorrhea (Figure 3.1). As expected there was a positive relationship with the two rates for both STIs and variability of the model increased with increasing long-term incidence rate. For syphilis, the nonparametric 10 person-year moving window could estimate short-term incidence rates up to a long-term incidence rate of 200 cases per 100,000 person-years. The gonorrhea 10 person-year moving window could estimate up until a long-term incidence rate of
about 2500 cases per 100,000 person-years. The mean and variance from the respective models were used to characterize the BUMBME prior using a lognormal distribution as described in section 3.2.4.

**Syphilis and gonorrhea long-term vs. short term rate model**

![Syphilis Long-Term v Short-Term Rate Model](image1)

![Gonorrhea Long-Term v Short-Term Rate Model](image2)

**Figure 3.1** A nonparametric exponentially weighted moving window was used to model (a) the 6-month (short-term) syphilis incidence rate for a census block group based on the 60-month (long-term) incidence rate (b) 6-month (short-term) gonorrhea incidence rate for a census block group based on the 36-month (long-term) incidence rate.
3.3.3 Space/Time Estimation

A theoretical space/time separable covariance model was fit to the experimental covariance after removing a space/time offset for each data point as described in section 3.2.5. Geostatistical estimates of the syphilis and gonorrhea incidence rate were created using the SK, PK, UMBME, and BUMBME methods. State maps that show the 6-month incidence rate for syphilis were produced monthly from January 2005 through January 2010. Predictably, the hard data produced the roughest map, with high rates and low rates in close proximity since the method makes the assumption that the observed rate is a good representation of the true rate. PK and UMBME produced a smoother, more realistic illustration of the true rate. The BUMBME method provided increased smoothing over PK and UMBME and highlights some different areas of concern which have had higher incidence rates in the past. We produced statewide gonorrhea maps from July 2008 through January 2011 for each method as well. While the smoothing differences among methods are less apparent than syphilis due to the greater prevalence of gonorrhea, BUMBME again increases incidence rate estimates in locations that have had higher historical rates. For the purposes of illustrating the distribution of syphilis and gonorrhea in North Carolina we show their respective estimated incidence rates using BUMBME from January 1, 2009 to July 1, 2009 in Figure 3.2.
Figure 3.2 The six month incidence rate for January 1, 2009 to July 1, 2009 is estimated with the space/time BUMBME method for (a) syphilis and (b) gonorrhea showing that these STIs disproportionately affect the rural, eastern part of North Carolina.

3.3.4 Asymptotic Validation

In order to test which data type provides better estimates of syphilis incidence, the asymptotic validation was performed by removing each space/time point and re-estimating based on the remaining points (Figure 3.3) for those locations with a sufficiently large population. The estimation was compared to the observed rate to quantify how well each data type reproduced the actual data. As the population increases, the observed rate approaches the latent rate.
MSE reduction relative to simple kriging for syphilis and gonorrhea using asymptotic validation

(a)

(b)

Figure 3.3 A validation was performed which removed each space/time data point and re-estimated the incidence rate at the location based on the remaining data. The difference between the observed and estimated values was used to calculate a mean squared error for each method. This procedure was repeated so that a block group must have a minimum population to be included in the validation for (a) syphilis and (b) gonorrhea.

Our new BUMBME data creation method outperformed the other methods for syphilis estimation and decreased the MSE by up to 25% at CBGs with the largest populations. The other
smoothing techniques PK and UMBME also performed better than SK for syphilis estimation where low rates particularly suffer from the small number problem. There was not much difference between PK and UMBME even though they obtain their variance in different ways. A smaller difference was observable among the estimation methods in the gonorrhea estimation, which has higher and more stable rates. However using the asymptotic validation, BUMBME appears to perform the best with a reduction in MSE relative to SK as high as 12%.

3.3.5 Comparison of Methods in Outbreak Locations

The difference between the space/time estimation methods is also evident by plotting the geostatistical estimate at a given census block group over time. By construct, the SK estimate must equal the observed rate, which leads to a noisy time series that jumps from time point to time point, providing evidence that there is likely a difference between the observed and latent rates. By contrast the PK, UMBME, and BUMBME methods provide different levels of smoothing. When only a few cases are added from one time step to the next, BUMBME provides more smoothing than PK and UMBME (Figure 3.4a) since it accounts for the long term incidence which changes progressively. However when cases start to escalate over successive time steps, by construct the UMBME and BUMBME estimates must lie within +/- (0.5 / nij) of the observed rate while that constraint does not exist for PK, and as a result PK estimates tend to be smoother while UMBME and BMEBME tend to better capture the corresponding sharp rise in incidence rate (Figure 3.4b).
Figure 3.4 For a fixed CBG, time series of the observed rates and estimated rates for (a) syphilis and (b) gonorrhea over time.
3.4 Discussion

In the last several years as GIS software has become more widely used, there has been increased interest in mapping infections and diseases so that rates can be correlated with spatially-dependent explanatory variables and interventions can be targeted to be most effective. Often researchers map the observed incidence rate which may not provide the most accurate representation of the unobservable latent rate. Using the observed rate is particularly problematic when the disease has a low prevalence and the population at risk is relatively small because small changes in the number of incident cases may lead to significant changes in the observed incidence rate, which are spurious and may mask more relevant underlying space/time infection patterns.

Incidence rates of syphilis and gonorrhea in North Carolina provide an ideal dataset for which to compare space/time mapping methods. The two infections cover a range of prevalence rates, as there were 6.3/100,000 primary and secondary syphilis cases/person and 150.4/100,000 gonorrhea cases/person in the state during the year 2009 (CDC 2010). Some of the most problematic infection areas include the rural, eastern part of the state which means there is often a smaller population at risk, increasing the importance of accounting for the small number problem. There is also reason to believe that syphilis and gonorrhea behave differently in space and time, with gonorrhea being more auto-correlated.

We compared four geostatistical methods, attempting to produce the unobservable latent rate on a fine space/time scale. These methods rely on different assumptions. SK assumes that the observed rate and latent rate are identical, PK smoothes the local observed rate based on the overall risk of infection, UMBME assumes that the observed rate is a sample of the latent rate whose error depends on the population at risk, and BUMBME relies on the assumptions of
UMBME buts weights the distribution based on the relationship between the long-term and short-term incidence rate.

In SK, the observed rates at the centroids of the densely-packed CBGs must equal the estimate and this method predictably produces a noisy map where high and low rates often appear in close proximity in space and time. We do not believe that this is a very realistic depiction of the true incidence rate when the small number problem is an issue. Also, the rural census block groups tend to be much larger in size in order to contain a similar population to other block groups. With the data treated as hard, rural block groups have more visual influence on the maps since the distance between centroids is greater. Additionally, CBGs with the smallest populations and less stable rates are given the same amount of weight, providing poor reliability for mapping the true incidence. This does not mean that the simple space/time kriging method produces uninformative results. These maps are quick to compute and do a good job of highlighting locations with a high observed rate. The over-influence that it gives to rural areas might actually be somewhat beneficial since these are places that often do not have the medical and staff resources needed on site to handle an outbreak and attention is therefore needed for these places.

The space/time Poisson kriging method was adapted from spatial Poisson kriging which traditionally has been used to estimate a disease in space rather than space/time. Rather than estimating the risk using the total cases divided by the total population, we make the assumption that the risk for a time period $j$ can be well approximated for dividing the number of incident cases in time period $j$ divided by the product of the duration of time period $j$ and the total population at center of time period $j$. This is a natural extension of spatial to space/time Poisson kriging. Adding the temporal component does make the estimate of risk a little less stable, but
we believe that it also makes it a little more dynamic thereby allowing to better study how the geographical pattern of disease outbreaks evolve over time.

While the Poisson kriging method is a useful tool to smooth out unreliable rates observed on small populations, it is based on assumptions that cannot be verified, which limits the utility of the maps produced, and predisposes the type of smoothing performed. The premise of PK is that the number of incident cases $Y_{ij}$ observed over a duration $T$ centered at time $t_j$ for a cohort of $n_{ij}$ individuals residing at location $s_i$ is Poisson distributed, i.e. $Y_{ij} \mid X_{ij}, n_{ij} \sim \text{Poisson}(n_{ij}TX_{ij})$.

Since the Poisson distribution is derived as a limiting case of the binomial distribution, it relies on two assumptions that are unrealistic in the context of the disease mapping of STIs: that the cohort is closed, and that the probability $TX_{ij}$ that an individual contracts the disease over the duration $T$ is the same for all persons in that closed cohort. In truth the population residing in a CBG does not constitute a closed cohort because many individuals enter and leave that cohort, each carrying probabilities of acquiring the disease that are significantly different than $TX_{ij}$.

Furthermore for STIs there is usually a high variability and clustering in individual risks within a cohort. Therefore the disease risk $X_{ij}$ estimated by PK is not a quantity grounded on theoretical correctness, rather the PK estimator should be simply examined in terms of its smoothing property. The smoothing property of the PK estimator is such that there is no hard limit put on the difference between an observed rate $R_{ij}$ and the corresponding PK estimate $X_{ij}$ (i.e. the smoothing is unbounded). As a result PK allows more smoothing than UMBME and BUMBME when there are large changes in the number of observed cases over a short spatial distance or time durations, which can be problematic from a disease surveillance point of view, because PK may over-smooth the sharp rise preceding disease outbreaks (Figure 3.4b).
This argument led to the formation of the UMBME soft datum that is constructed as a uniform interval around the observed rate, which in effect puts a hard limit to the amount of smoothing allowed since the estimated rate must lie within an interval of +/- 0.5/(n_iT) centered at the observed rate. This corresponds to an interval of length 1/(n_iT) which exactly corresponds to the smallest measurable increment of the observed rate. By using this soft datum the estimator is just as capable as PK to smooth out unreliable rates observed on small populations. However this soft datum also restricts the smoothing allowed so that the estimated rate does not deviate from the observed rate more than what can be attributed to the smallest measurable increment. As a result the smoothing properties of this model are that it allows (on average) similar smoothing as PK, but that smoothing is restricted when there is a large change in the number of observed cases over a short space/time distance. This can be beneficial for disease surveillance as the uniform model allows us to smooth out isolated spurious changes in rates but retains changes when they occur in short proximity, such as during the ramp up preceding a disease outbreak.

The idea for the Bayesian soft data was developed as a hybrid between the basic ideas behind the PK and UMBME methods. The prior distribution for the current incidence rate is based on the long-term incidence of a CBG, which can be thought of as a localized approximation of the risk. This prior gets updated by the uniform interval around the observed rate. We think that this method benefits from the fact that STIs often have a core area and they are more localized than other infections.

Visually, the maps produced using the BUMBME method were distinctly different from the other methods in two main ways. First, BUMBME showed a higher estimated incidence in locations that had recorded an elevated number of cases in the past due to using the long term incidence rate to predict the short term incidence rate. We believe that this reflects Rothenberg’s
idea of a core of infection that persists over time. BUMBME typically estimated rates that were slightly above the other methods for syphilis since the vast majority of locations record 0 syphilis cases for a given 6-month period and BUMBME is the only method that has a mean above 0 in these places.

Since comparing methods based on smoothing properties is somewhat subjective, we sought a way to evaluate how they compare in reproducing the unobservable latent rate. Therefore, we introduced the concept of an asymptotic validation analysis that relies on the simple fact that as the population at risk increases, the observed rate approaches the latent rate. We think that the results of our asymptotic validation reveal several key points. First the prevalence of the infection is important. Syphilis, which has a small number of cases benefits from each of the smoothing methods when compared to SK, as evidenced by the relative reduction in MSE that increases asymptotically. The relative reduction in MSE compared to SK disappears for UMBME and PK in the asymptotic validation analysis of gonorrhea. This infection, which has a much higher rate than syphilis, is more stable in the population which benefits SK which relies on the rates of neighbors without smoothing. BUMBME still performs best in this analysis, something that we attribute to the additional information that is gained by letting the long-term incidence rate influence the short-term incidence rate. Thus by accounting for core areas, BUMBME is producing more realistic estimates that go beyond simply addressing the small number problem.

We chose a 6-month incidence period based on the fact that a subsequent goal of the research is the ability to determine outbreaks in real time. We examined three regions which were known to have syphilis outbreaks during the time period when the data were available. The 3-month incidence period did not do a very good job illustrating the ramp-up of the outbreaks as
the time period was too short to remove the noise from the data. Both the 6-month and 12-month incidence periods were able to show the continual increase in incidence rates that we expected from these outbreaks, so 6-months was chosen since it was the finest time scale. This did not seem to be much of an issue with the gonorrhea data due to a higher prevalence and thus more cases. Since a 6-month incidence was used for syphilis, we maintained this period for gonorrhea as a basis of comparison.

One of the limitations of this work is the uncertainty surrounding the date of infection, specifically for syphilis. While we account for sampling error, we do not address the misclassification of a case into the incorrect incidence period. The syphilis cases are back-dated by stage to obtain a rough estimate of date of infection. We could account for this uncertainty by increasing the interval of the likelihood function (possibly based on stage of infection), which would also increase smoothing. Rather than using a uniform interval, one with a trapezoidal shape could center the weight of the interval on the most likely date of infection, but with hard endpoints that would not be achieved by assuming date of infection has a Gaussian distribution.

Using the optimal method for estimating the incidence rates for these STIs is not just an academic exercise in statistical estimation. Since sexual contact is necessary to transmit STIs between individuals, the infections progress over months rather than days to weeks. This period provides enough time to analyze data and conduct a targeted intervention. Individuals often are not aware they have contracted an STI, so the efficient use of public health resources can slow the spread before it becomes a fully developed outbreak. There is a higher transmission rate of HIV from individuals with other STIs, particularly those with gonorrhea, providing synergistic motivation for bringing rates under control as quickly as possible. Overall we have shown that BUMBME is a useful tool to smooth out unreliable rates caused by the small number problem,
create maps that more realistically account for core areas, and better capture the ramp up
preceding outbreaks. Thus, we think that BUMBME has the potential in future works to be an
extremely useful outbreak detection tool, especially with infections which suffer from small
number estimate problems.
Centers for Disease Control and Prevention (2006) The national plan to eliminate syphilis from the United States, Atlanta, GA

Centers for Disease Control and Prevention (2010) Sexually Transmitted Disease Surveillance 2009, Atlanta, GA


Fleming DT, Wasserheit JN (1999) From epidemiological synergy to public health policy and practice: the contribution of other sexually transmitted diseases to sexual transmission of HIV infection. Sex Transm Infect. 75(1):3-17


Hampton KH, Serre ML, Gesink DC, Pilcher CD, Miller WC (2011) Adjusting for sampling variability in sparse data: geostatistical approaches to disease mapping. International Journal of Health Geographics 10(54)


Overview

The BUMBME method described in Chapter 3 increased smoothing of the incidence rate on a fine space/time scale in most situations, but given that estimates must lie within +/- 0.5/(n_ijT) of the observed rate, this method smooths less when the rate increases rapidly, which should make it ideal for detecting outbreaks of syphilis and gonorrhea. In this chapter, the BUMBME outbreak detection algorithm is introduced and compared to SaTScan - a commonly used method for detecting emerging clusters of infection. Bringing outbreaks of STIs under control quickly and efficiently is important to stop the spread to other sexual networks and a necessary component for reducing the overall rate of STI infections.
4.1 Introduction

There have been significant developments in outbreak detection algorithms over the last decade. This can be attributed to an increase in funding to combat bioterrorism following the attacks on September 11, 2001 as well as the dramatic increase in the use of geographic information systems (GIS) software. The most popular techniques can often be dichotomized into syndromic surveillance or scan statistics (cluster detection algorithms) (Pfeiffer 2008; Hohle 2009). These methods can further be categorized as temporal, spatial, or spatiotemporal (Tsui et al. 2011).

Syndromic surveillance which often utilizes CUSUM methods, works well for identifying human health responses that are identifiable soon after exposure. Adapted from quality control algorithms to monitor hospitals and emergency rooms for symptoms that could be related to an event such as an outbreak of smallpox, these methods can be easily extended to public health department monitoring of common infections where there are a sizable number of symptomatic cases and they are diagnosed in a timely matter (Farrington et al. 1996; Rogerson 1997; Hutwagner et al. 1997; Reis and Mandl 2003). Low resolution monitoring such as hospital locations or the county level are often sufficient since these types of infections can be quickly transmitted and the risk in the population is fairly homogeneous.

Scan statistics (or cluster detection algorithms) search for spatial windows that have a higher number of cases than would be expected. They have become popular in epidemiology, especially with the development of the free, downloadable SaTScan software package. During the past several years, SaTScan has advanced to find clusters in space/time, incorporating several probability models, and identify elliptically-shaped clusters. When used prospectively, SaTScan can locate emerging clusters or outbreaks (Kulldorff and Nagarwalla 1995; Kulldorff 1997;

The North Carolina Department of Health and Human Services has maintained a database of syphilis infections dating back to 1999 and gonorrhea infections since 2005. North Carolina, along with other southeastern states, has higher rates of these and other sexually transmitted infections (STIs), compared to the rest of the United States. Syphilis is caused by *Treponema pallidum* bacteria and those that have been exposed to this infection can eventually develop a genital ulcer. Untreated primary syphilis can lead to secondary syphilis which has moderate to severe disease symptoms including rash, fever, and even dementia. Gonorrhea, which is caused by exposure to *Neisseria gonorrhoeae*, has symptoms of disease that include urethritis, cervicitis, and prostatitis. Additionally, these infections can elevate the rate of HIV transmission between persons (Cohen 1998; Fleming and Wasserheit 1999). The state is interested in lowering syphilis and gonorrhea rates by targeted intervention which includes identifying core areas of infection (clusters) and stopping outbreaks at an early stage.

Syndromic surveillance and cluster detection have several deficiencies with regard to outbreak detection of gonorrhea and syphilis on a fine space/time scale. Syndromic surveillance is a temporal method that can be implemented at multiple spatial locations, but does not account for spatial dependence. STI risks are heterogeneously distributed through the population and they move through sexual networks which have a spatial component, so accounting for this dependence is important. Also, this method needs to have symptoms or case counts which are robust so that a large count can be confidently flagged as an abnormality. STI symptoms often do not manifest themselves quickly and STI case counts are often very low at the fine scale of the CBG level, even when cases are aggregated over 6 months, so using a CUSUM method would likely flag some locations any time a case was reported. Aggregating cases to a higher
geographic level reduces our ability to pinpoint where outbreaks are occurring. Since risk is distributed heterogeneously, spatial accuracy is of greater importance.

The cluster detection method, which compares the cases observed within a space/time window to those outside the window, is dependent upon the number of cases across the study area. While this makes saTScan useful for identifying clusters or core areas of STIs, the heterogeneity of cases makes it difficult to identify outbreaks that commonly occur in close proximity to the cores. Also, saTScan is restricted to certain shapes so that the method does not suffer from selection biases. Obviously, this causes a loss in spatial resolution. Like syndromic surveillance, the saTScan method does not address the small number problem which occurs when the population is not large enough to maintain a stable rate. As the desire for higher spatial resolution increases, the small number problem becomes more of a barrier to mapping the unobservable latent incidence rate (Jones and Kirby 1980; Kennedy 1989).

STIs require close physical contact in order to move between individuals and they are associated with a number of risk factors – two attributes which make them heterogeneously distributed across space. They often have a persistent core that propagates the infection so that finding a cluster and finding an outbreak are often not one and the same. Also, there is evidence that sexual networks are not always compact in space but rather can move from one area to another through a bridge individual. When examining STIs at the census block group (CBG) level, one additional case can make a significant difference in the estimated rate and the observed rate is not always a good estimate of the unobservable latent rate (Hampton et al. 2011). These are problems which call for a more dynamic spatiotemporal outbreak detection method.

In order to estimate the latent rate, we developed the Bayesian Uniform Method extension of Bayesian Maximum Entropy (BUMBME) which estimates a prior distribution of
the 6-month incidence rate based on the long-term incidence rate. BUMBME has been shown to improve the estimate of the latent rate compared to other disease mapping methods through an asymptotic validation. By assuming that the observed incidence rate is a sampling of the latent rate, BUMBME places hard bounds on the latent rate estimate so that it does not over-smooth the data when an outbreak is occurring (Chapter 3). In this Chapter, we develop an algorithm that identifies significant changes in the space/time BUMBME posterior over time in order to indicate where outbreaks were occurring. We compare this new BUMBME space/time geostatistical outbreak detection method to space/time prospective SaTScan based on sensitivity and specificity for detecting outbreaks of syphilis and gonorrhea in North Carolina.

4.2 Methods

4.2.1 Datasets

Reported syphilis and gonorrhea cases in the state of North Carolina were obtained from databases maintained by the state’s Department of Health and Human Services. Syphilis cases were available from 1999 through 2010 and gonorrhea cases were available from 2005 through 2011. A case was included in our analysis if its location could be geocoded and it had a diagnosis date associated with it. Additionally, syphilis cases were backdated from the reported diagnosis date by 45, 90, and 183 days for the stages of primary, secondary, and early latent syphilis, respectively, to place the case in its most likely incidence period. Late stage syphilis cases were removed from our analysis since there was too much uncertainty regarding the proper incidence time period.

Cases were geocoded at the North Carolina State Health Department using ArcGIS 9.3 and matched to three geographic locators. The primary locator was created by the North Carolina
Emergency Response System and contains point locations for North Carolina households. The secondary locator was created by the North Carolina Department of Transportation and contains street-level geographic data. The tertiary locator was created using ESRI’s 2006 Street Map shapefile and is primarily used for locating residences with outdated street names, prisons and military bases. We used post office addresses when PO Boxes were provided. Military addresses could not be geocoded due to missing street addresses. We refer to the cases used in our analysis as reported incident cases.

The geocoded cases were then aggregated to the centroid of their CBG, which we believe was the smallest available census unit that could still provide a sufficient amount of anonymity. The block group boundaries were obtained from shapefiles produced by the U.S. Census Bureau and they correspond to those delineated by the 2000 Census.

4.2.2 Defining Outbreaks

Outbreak boundaries were delineated for syphilis and gonorrhea separately, by retrospectively identifying space/time clusters of cases, using SaTScan 9.3. Cases were aggregated monthly and parameters were set so that the spatial area of the SaTScan cluster could not contain more than 5% of the state’s population and the time period could not be longer than 24 months. Additionally, clusters were not allowed to overlap in space or time. A discrete Poisson probability model was used for our SaTScan analysis of the incidence rate and an identified space/time cluster was deemed to be significant if p<0.10. In order for a space/time location to be labeled as an outbreak location, it had to be within the first 6 months of a retrospective SaTScan STI outbreak boundary and be in the top 1% of all observed space/time incidence rates (≥ 45 cases/100,000 person-years for syphilis and ≥535 cases/100,000 person-
years for gonorrhea). We define the first six months as the “ramp-up” period of an outbreak, when rates are increasing and the outbreak is growing.

4.2.3 Identifying Outbreaks

The procedure described above yielded CBG centroids that could be dichotomized as “outbreak location” or “not an outbreak location” at each month for both syphilis and gonorrhea. We attempted to correctly identify the outbreak and non-outbreak locations using space/time prospective SaTScan, a widely-used software program for finding case clusters, and our new method for finding outbreaks which utilizes the rate of change in the posterior distribution produced by geostatistical estimation of the incidence rate using the BUMBME method.

We ran prospective saTScan for each month, from January 1, 2005 to January 1, 2010 for syphilis and July 1, 2008 to January 1, 2011 for gonorrhea. A reported incidence case was included in the prospective SaTScan analysis if its incidence date was prior to the date being evaluated. The spatial parameter of space/time prospective SaTScan was set to cover a spatial area no larger than 5% of the state’s population. The length of the time period for a prospective cluster was allowed to be unlimited, but due to the prospective nature of the analysis, a significant cluster (p<0.10) must include the most recent time period where data are available. Our analysis yielded a collection of spatial areas for each month that SaTScan identified as prospective clusters. A CBG centroid inside one of these clusters was labeled as an outbreak location identified by prospective SaTScan.

Our new approach consisted of two parts: (1) producing the geostatistical estimates and (2) analyzing the rate of change with regards to the expectation of the estimation posteriors to identify outbreaks. Six month incidence rate data were created for each month at each CBG
centroid for both syphilis and gonorrhea. This was done by summing the respective number of syphilis or gonorrhea cases within the CBG during the previous 6 months and dividing by the block group’s estimated population at risk for the corresponding year. In order to estimate the CBG population for each year, we interpolated the value by using the block group’s 2000 census population and its 2007 estimate, extrapolating the rate of change during this 7 year period to years beyond 2007.

We utilized the BME computational library for producing the space/time geostatistical estimation due to its rigorous nonlinear mathematical framework that includes the ability to incorporate both Gaussian and non-Gaussian data. BME is a two-stage process which uses maximum entropy to organize the general knowledge \( G \) of the STRF (such as space/time mean trend and covariance functions) to compute a prior probability density function (PDF) of the space/time process. The prior PDF is then updated by the site-specific knowledge \( S \) which can include hard (measured without error) and soft (characterized by a PDF) data to produce a posterior PDF that characterizes the STRF at any space/time point. When only hard data is available, BME produces the simple space/time kriging estimator.

The BUMBME approach was used to produce the geostatistical estimates. This method smoothes the map of the incidence rate by using a distribution of possible incidence rates in place of the observed rate for each space/time location (soft data). The distributions correspond to Bayesian posterior distributions which were derived by using a uniform distribution around the observed rate as a likelihood function to update a lognormal prior distribution for the 6-month incidence rate which comes from a model which bases this on the CBG’s long-term incidence rate. Since a time period is needed for the long-term incidence rate, BUMBME
geostatistical estimates are available for January 2005 to January 2010 for syphilis and July 2008 to January 2011 for gonorrhea.

The geostatistical estimation yields a BME posterior distribution for each 6-month period ending at time $j$ for every CBG centroid $i$ which can be denoted as $f_k(y_{k,ij})$ where the subscript $K$ represents the physical knowledge $K=G \cup S$ within the BME framework. At each centroid, moving forward through time, we compare the expected value of the posterior distribution to that of the prior time period, indicating when it has increased sufficiently and summing these significant increases over time (Eq. 4.1). When $a_{ij} \geq \beta_3$, an alarm is indicated at CBG $i$ for the 6-month period ending at time $j$.

$$a_{ij} = \sum_{t=j-\beta_1}^{j} I \left[ (\hat{E}[f_k(y_{k,ij})] - \hat{E}[f_k(y_{k,ij-1})] > \beta_2 \hat{V}[f_k(y_{k,ij-1})]^{1/2} ) \right] \quad (4.1)$$

The set of parameters $\beta = [\beta_1, \beta_2, \beta_3]$ are estimated by optimizing the function that calculates the sensitivity ($Sn$) and specificity ($Sp$) of our method for identifying outbreaks and the associated cost (Eq. 4.2). Analogous to the definitions used in a test for disease, we define $Sn$ as correctly identifying (sounding an alarm) a space/time outbreak location and $Sp$ is defined as correctly identifying (not sounding the alarm) the space/time locations that are not outbreak locations. The type I error cost ($C_1$) – the cost associated with not correctly identifying locations where an outbreak is actually occurring – and the type II error cost ($C_2$) – the cost of utilizing public health resources when an outbreak is not occurring – are complicated functions which are dependent on infection type, the spatial area an outbreak covers, the number of people and potential cases affected, and the type of intervention, as well as other factors.

$$C = \frac{c_2 \ast N \ast p \ast (1-Sn) + N \ast (1-p) \ast (1- Sp)}{c_1} \quad (4.2)$$
Since the number of space/time locations ($N$) in the study area and the proportion ($p$) of them labeled as outbreak locations are held constant, we use the ratio of costs to obtain a relative cost ($C$) performance for a given set of parameters.

4.3 Results

4.3.1 Outbreaks Identified by SaTScan

SaTScan was run retrospectively with a maximum spatial size of 5% of the state’s population and maximum temporal size of 24 months to find where clusters of syphilis existed. We define the first 6 months of these clusters as the outbreak period to be identified. SaTScan identified 10 statistically significant syphilis clusters in North Carolina between January 1, 2005 and January 1, 2010. There was one 6-month cluster, one 12-month cluster, and the rest were near or at the 24-month maximum in length. These clusters were fairly evenly distributed throughout the study period. The smallest number of CBGs in a cluster was 11 and the largest number of CBGs in a cluster was 242. The diameters of the clusters range in size from 4.78 to 27.70 km.

The same spatial and temporal limits were used to retrospectively identify clusters of gonorrhea. SaTScan found 29 statistically significant clusters of gonorrhea between July 1, 2008 and January 1, 2011. The gonorrhea clusters were also well distributed throughout the study period and lasted for variable lengths of time. The smallest cluster included only 1 CBG and the largest included 254 CBGs. The diameters of these clusters ranged in size from 0 to 91.32km. The most significant gonorrhea cluster was also the largest in terms of CBGs and diameter and lasted from July 1, 2009 to October 31, 2010.
CBGs that had an observed syphilis incidence rate greater than 45 cases/100,000 person-years and an observed gonorrhea incidence rate greater than 535 cases/100,000 person-years within the first six-months of a retrospective SaTScan cluster were deemed outbreak locations. There were 55 space/time outbreak locations for syphilis and 272 space/time outbreak locations for gonorrhea that we sought to identify using prospective SaTScan and our BUMBME geostatistical outbreak detection method.

4.3.2 Outbreak Detection Maps

The BUMBME outbreak method was visually more dynamic than saTScan, delineating the outbreak with a non-circular shape since it does not suffer from the pre-selection bias that saTScan does. Figure 4.1 illustrates how the methods identify the locations with high rates at the beginning of the retrospectively-identified outbreak period. This is particularly apparent with gonorrhea which has a higher prevalence rate and thus a number of contiguous locations experiencing the outbreak simultaneously.
Figure 4.1 The BUMBME outbreak detection method has the ability to delineate the shape of the (a) syphilis and (b) gonorrhea outbreak without restrictions, identifying the high rate areas during the ramp-up period with greater sensitivity and specificity.
4.3.3 ROC Analysis

Receiver operating characteristic (ROC) curves were created for the possible combinations of $\beta_1$, $\beta_2$, and $\beta_3$ to quantify the performance of the outbreak detection methods. As seen in Figure 4.2b, the BUMBME method is clearly superior in terms of sensitivity and specificity for identifying gonorrhea outbreaks. For syphilis there is a large increase in sensitivity with a very slight decrease in specificity, making BUMBME the optimal choice.

The optimal parameters of the model were determined as those which minimized the squared distances to perfect sensitivity and specificity on the ROC curve. These parameters increased sensitivity from 87% to 93% and specificity from 88% to 91% with respect to prospective saTScan for gonorrhea and increased sensitivity from 52% to 87% for syphilis with only a 3% loss in specificity.

**Figure 4.2.** An ROC curve shows that the BUMBME outbreak detection method outperforms prospective saTScan in terms of sensitivity and specificity for detecting outbreaks of (a) syphilis and (b) gonorrhea.
4.4 Discussion

STI outbreak detection at the census block group level is a good example where using simple case counts for the detection of outbreaks is deficient. In North Carolina, gonorrhea averages approximately 200 cases per 100,000 person-years and syphilis is far less common with about 7 cases per 100,000 person-years. With 5,264 census block groups in North Carolina, both infections will have very low case counts at this census level within a 6-month period. There will be many instances where the observed rate is 0 when the unmeasurable latent rate is not and as the CBG population decreases, the small number problem becomes a bigger issue as the observed rate becomes less stable. Also, STIs and their risk factors are distributed heterogeneously across space, in particular, necessitating an outbreak detection method that brings in localized information.

We address these issues by utilizing space/time geostatistical estimates with distributional soft data. The BUMBME method has some nice attributes for mapping STI data with high resolution. First, it does not rely on distributional assumptions made by other methods which are often difficult to verify. The BUMBME method selects a prior distribution for the 6-month incidence rate based on a CBG’s long-term (more stable) incidence rate, which can be thought of as a localized approximation of the risk. This prior is updated based on the assumption that the observed rate is a sample of an underlying unobservable latent rate that has bounds based on the population at risk. This allows the method to smooth small changes in rate while maintaining a geostatistical estimate that is within the bounds of sampling error so that incidence rate estimates are not over-smoothed during outbreaks of infection.

Our estimates were produced at the CBG level due to a balance of privacy protection, spatial resolution, and data stability. The locations of individuals with STIs such as syphilis and
gonorrhea need to be protected for obvious reasons. Each individual’s geocoded location was randomly geomasked within its respective block group at the state health department since block group populations are usually sufficiently large enough to protect an individual. On occasion, cases had to be geomasked to a neighboring block group, introducing some additional misclassification error into the observed incidence rate. If privacy had not been a concern and resolution had been maintained at the census block level, the observed rate would have been more unstable with such low populations. Upscaling to the census tract would have resulted in a large loss in outbreak detection resolution, but more stable observed rates. We are unable to predict how our method might perform under each condition, but the block level is probably impractical from a privacy perspective and the tract level is less useful for a targeted intervention. For those interested in identifying outbreaks among subpopulations rather than the general population, the CBG might not be large enough to protect privacy. However, by using our method at the census tract level and then targeting intervention to areas of the tract with demographics matching the subpopulation might be a way to research these more specific outbreaks.

There is an element of temporal misclassification in the data, particularly for syphilis. We roughly accounted for this misclassification by backdating infection dates based on syphilis stage of disease. However, cases can still be placed in an improper 6-month period, particularly as the infection progresses. We could attempt to account for this misclassification in a future model by extending the bounds of the uniform likelihood function. This would provide additional smoothing which may or may not improve the performance of our BUMBME outbreak detection model. Since STIs are most prevalent in rural areas of North Carolina that have lower
socioeconomic status and less access to health care, it is possible that gonorrhea could also benefit from an increased smoothing of the incidence rate.

While the geostatistical estimation is spatiotemporal, the outbreak detection currently is not. The BUMBME estimate for a given space/time location weights the co-located and surrounding soft data based on the space/time covariance function. When rates in surrounding areas are higher, these will most likely raise the estimate. The outbreak detection only analyzes the rate of change in a given CBG. A more sophisticated version of our method will take the surrounding areas into account before deciding whether or not to sound an alarm. We would expect to lower the cutoff for a given location if the rates around it are increasing which would place the area at a higher risk for an outbreak. Since the BUMBME estimate has already incorporated these data and shifted the posterior distribution, the cutoff also needs to decrease based on the fact that it will be harder to find a significant increase.

Our method, as it is currently implemented, finds an optimal set of parameters for detecting outbreaks. These parameters are the percent increase with regards to the standard deviation in the expectation of the posterior compared to the previous time period, the number of previous months included in the evaluation, and the number of significant increases needed to sound an alarm. This optimal parameter set is estimated using the entire time period and the parameters are applied as constants to obtain the method’s results. The only parameter that is somewhat dynamic is the percentage increase in standard deviation because indicating a significant increase is dependent on the posterior’s variance. We envision that the next iteration of our method will make the parameter values more dependent on space/time location.

Space/time prospective SaTScan was used as our comparison method since it is available as a free software package which has been widely used in epidemiological research to find
clusters of disease. The methodology has evolved since its introduction to include more probability models and elliptically-shaped clusters in space. It still lacks the ability to be truly dynamic in space due to the method’s vulnerability to pre-selection bias. Since we were using SaTScan to help define our outbreaks, this could have impacted our results.

The observed rate was used to indicate the outbreak boundaries and identify the outbreak locations within those boundaries. There has already been a discussion regarding the fact that these observed rates are not always the best proxy for the true rate. We looked at only identifying outbreak locations that also had large block group populations and presumably more stable observed rates, but these resulted in too few locations. We believe that using the observed rate to identify outbreaks, when we know this rate is imperfect, and evaluating our method against prospective SaTScan, which relies on the observed rate, only strengthens the case for our method. We could have simulated true and observed data for our comparison, but creating a simulated observed dataset would have relied on too many assumptions.

The use of GIS and mapping has increased dramatically over the last several years and space/time methods have become more sophisticated. Traditional methods for detecting outbreaks are most successful when a diagnosis of infection is made and reported in a timely matter so that the observed rate is highly correlated with the true rate. As the prevalence of the infection decreases it becomes more difficult to detect outbreaks with higher spatial resolution. We believe that our method, which relies on the observed rate as only one source of information for detecting an outbreak, can benefit detection when the rate of space/time misclassification increases and/or prevalence decreases. By improving the space/time resolution of detection, we have the ability to target intervention and maximize the benefit to cost ratio of public health disease and infection programs.
REFERENCES


Hampton KH, Serre ML, Gesink DC, Pilcher CD, Miller WC (2011) Adjusting for sampling variability in sparse data: geostatistical approaches to disease mapping. International Journal of Health Geographics 10(54)


CONCLUSIONS

Point sources are ideal targets for intervention strategies, because they often have the potential to make a big impact. This thesis has demonstrated that using modern geostatistical methods as opposed to traditional kriging estimators can improve estimation accuracy of pollution and infectious disease from these point sources.

Using a mass fraction model to estimate PAH from existing PM$_{2.5}$ monitors was a useful proof of concept in the case of the World Trade Center disaster. Developing a more robust mass fraction model could eventually lead to estimation of PAH from the existing PM$_{2.5}$ network of air monitors at minimal additional cost. As air pollution research delves into the components of particulate matter that are harmful to human health, this potentially could be a useful method for conducting a large scale study on health effects from PAH exposure as well as other air pollutants.

It has been documented that industrial hog operations in North Carolina that release large amounts of untreated waste into the environment affect the health and well-being of surrounding residents in a myriad of ways. The land use regression model developed using passive H$_2$S samplers demonstrated that hog CAFOs are indeed the source of this pollutant, which can be a marker for others. Our non-Gaussian error model provided slight improvements over traditional kriging for estimating the 2-week average value of H$_2$S. However, more significant improvements are expected when estimating the instantaneous (15 minute) concentration of H$_2$S with data collected from active sampling devices.
When estimating incidence rates of infection, particularly with high resolution, there are a variety of potential measurement errors that are less commonly seen with environmental pollutant data. These include infected individuals getting tested in a timely manner and correctly estimating the infection date. Thus, an error model for the observed rate is important for accurately estimating the unmeasurable latent rate. Again, taking a modern geostatistical approach showed improvement in estimation. The first implementation of the BUMBME outbreak detection algorithm showed promise for using a geostatistical estimator to detect outbreaks. More complex versions of this basic concept could lead to additional improvements.

Modern geostatistical methods such as Bayesian Maximum Entropy continue to advance the science in exposure estimates, disease mapping, and other variables that have a distribution that changes in space and time. Mapping point sources of pollution and infectious disease ranging from a single source to multiple spatial sources to dynamic sources appear to be an area of environmental science and epidemiology where application of these emerging methodologies is beneficial for improving public health.
APPENDIX A

Mass Fraction Spatiotemporal Geostatistics – Theory

This section provides the mathematical framework for mass fraction spatiotemporal geostatistics. While we refer to PAHs, the framework can be applied to any particle-bound compound.

Space/Time Random Fields

Environmental contaminants are often characterized by high variability and uncertainty in their distributions across space and time. Therefore, a stochastic representation of the contaminant of interest is useful for characterizing this variability and uncertainty. A space/time random field (S/TRF) is one such representation, where the field is denoted by $X(p)$ where $p = (s,t)$ is a space/time location, with the vector $s = (s_1,s_2)$ representing spatial coordinates and $t$ corresponding to time.

The statistical moments of a S/TRF, such as its mean and covariance, provide a general knowledge base ($G$) describing the distribution of the contaminant concentration across space and time. The mean

$$m_x(p) = E[X(p)] \quad (A1)$$

where $E[.]$ is the stochastic expectation operator, describes systematic spatial and temporal trends, while the covariance

$$c_x(p, p') = E[(X(p) - m_x(p))(X(p') - m_x(p'))] \quad (A2)$$

describes space/time dependencies of the contaminant concentrations between points $p$ and $p'$. 

If we let \( x_i \sim X(p_i) \), then a given set of mapping points \( p_{\text{map}} = \{ p_1, ..., p_m \} \) yields a corresponding collection of random variables \( x_{\text{map}} = \{ x_1, ..., x_m \} \) that describes the S/TRF at \( p_{\text{map}} \). Under the general knowledge base, the uncertainty characterizing the S/TRF at the mapping points \( p_{\text{map}} \) is modeled in terms of the joint probability density function (PDF)

\[
f_G(\chi_{\text{map}}, p_{\text{map}}) \, d\chi_{\text{map}} = \text{Prob}[\chi_{\text{map}} < x_{\text{map}} < \chi_{\text{map}} + d\chi_{\text{map}}] \tag{A3}
\]

where \( \chi_{\text{map}} = \{ \chi_1, ..., \chi_m \} \) represent a plausible realization of \( x_{\text{map}} \) at points \( p_{\text{map}} \), \( \text{Prob}[.] \) is the probability operator, and the subscript \( G \) is used to emphasize that the PDF uses the general knowledge \( G \).

In addition to the general knowledge base \( G \) characterizing statistical moments of \( X(p) \), there is also a site-specific knowledge base \( S \) that is obtained independently at specific sites across space and time. The site-specific knowledge base \( S \) includes measured concentrations obtained at monitoring stations, as well as predicted values obtained from regression models using explanatory variables measured at other specific sites. If the contaminant is measured without noticeable error at a set of points \( p_{\text{hard}} \), then we refer to this type of site-specific knowledge as the hard data \( \chi_{\text{hard}} \), which provides a deterministic value for \( x_{\text{hard}} \), i.e. \( \text{Prob}[x_{\text{hard}} = \chi_{\text{hard}}] = 1 \). On the other hand, the soft data at specific points \( p_{\text{soft}} \) refers to values predicted or measured with noticeable error. Hence the vector of random variables \( x_{\text{soft}} \) at points \( p_{\text{soft}} \) is characterized under the site-specific knowledge base \( S \) by the PDF:

\[
f_S(\chi_{\text{soft}}) \, d\chi_{\text{soft}} = \text{Prob}[\chi_{\text{soft}} < x_{\text{soft}} < \chi_{\text{soft}} + d\chi_{\text{soft}}] \tag{A4}
\]

where the subscript \( S \) is used to emphasize that that PDF uses the site-specific knowledge base \( S \). Under these two types of site-specific knowledge, space/time locations where data are available are denoted as \( p_{\text{data}} = \{ p_{\text{hard}}, p_{\text{soft}} \} \).
The S/TRFs used to model PAHs

We introduce several S/TRFs that will be useful to model the spatiotemporal distribution of PAH contaminants. We let \( Z(p) \) be the S/TRF representing the distribution of an individual PAH across space and time. Using monitors to measure atmospheric PAHs is one of the most reliable ways for obtaining data regarding its S/TRF. Unfortunately, ambient PAHs are not typically measured by current monitoring systems, and therefore the measurement data are less dense in the space/time continuum compared to the atmospheric particulate matter to which it is bound. This leads to the particulate matter less than 2.5 microns in size (PM\(_{2.5}\)) as an obvious choice as secondary information about the S/TRF for PAHs.

The monitoring data analyzed in this work as well as some previous works (Serre et al. 2004; Pleil et al. 2004a; Puangthongthub et al. 2007) show that atmospheric PAHs and PM\(_{2.5}\) have a statistical distribution that is approximately log-normal, with the bulk of the data close to 0 and a tail of high values. Therefore a log transformation will be used to normalize both S/TRFs. Hence we let \( Y(p) = \ln(Z(p)) \) represent the S/TRF of log-PAH and \( U(p) = \ln(PM_{2.5}(p)) \) represent the S/TRF of log-PM\(_{2.5}\).

One of the main objectives of this work is to develop a mass fraction approach to integrate data on both PAHs and PM\(_{2.5}\) in the space/time mapping of PAHs. This is done by introducing the log-mass fraction S/TRF defined as \( W(p) = \ln(Z(p)/PM_{2.5}(p)) = Y(p) - U(p) \). This formula shows that given the log-mass fraction S/TRF and the log-PM\(_{2.5}\) S/TRF, the log-PAH S/TRF can simply be obtained from:

\[
Y(p) = W(p) + U(p) \quad (A5)
\]
Finally, we capture known trends in the spatial and temporal distribution of \( Y(p) \) using a deterministic global mean trend function \( m_Y(p) \). For example, the spatial component of \( m_Y(p) \) might be known to have higher values at the WTC site where fires provided a source of PAHs, while the temporal component of \( m_Y(p) \) might be known to have decayed in the months following the collapse of the WTC towers as these fires were controlled. We then denote the S/TRF representing the mean-trend removed log transformed PAH as \( X(p) = Y(p) - m_Y(p) \). The S/TRF \( X(p) \) models the residual variability and uncertainty in log-PAH. General knowledge about this field includes its local mean trend function \( m_X(p) \) as well as its covariance function \( c_X(p,p') \).

**Mass Fraction Approach for Creating Soft Data**

The criteria pollutant PM\(_{2.5}\) has been measured using an accurate federal reference method (FRM) at a substantial number of monitoring stations over the entire United States for several years. In this work we complement these measurements with a set of accurate PAH measurements obtained for a small subset of the PM\(_{2.5}\) samples. Hence, we have a small set of points \( p_{\text{hard}} \) where both log-PAHs and log-PAH/PM\(_{2.5}\) are accurately measured, and a larger set of points \( p_{\text{soft}} \) where only log-PM\(_{2.5}\) is accurately measured. The accurate measurements for log-PAHs and log-PAH/PM\(_{2.5}\) at \( p_{\text{hard}} \) are denoted as \( y_{\text{hard}} \) and \( w_{\text{hard}} \), respectively. The accurate measurements of log-PM\(_{2.5}\) at \( p_{\text{soft}} \) are denoted as \( u_{\text{measured}} \).

The main assumption in the mass fraction approach is that while both PAHs and PM\(_{2.5}\) exhibit high variability across space and time (e.g. high spatial variability due to effects of wind direction, airflow barriers, etc.), it is reasonable to assume that the mass fraction PAH/PM\(_{2.5}\) has a similar statistical distribution across points \( p_{\text{hard}} \) and \( p_{\text{soft}} \), as long as these points are in an air
shed affected by a same fire (because the combustion source for PAHs is similar across that air
shed). Using that assumption we use the measured log-mass fraction data \( w_{\text{hard}} \) at \( p_{\text{hard}} \) to predict its value \( w_{\text{soft}} \) at \( p_{\text{soft}} \), and then we use Eq. A5 to combine the predicted \( w_{\text{soft}} \) with measured log-
PM\(_{2.5}\) data \( u_{\text{measured}} \), which provides the soft data \( y_{\text{soft}} \) for log-PAHs at points \( p_{\text{soft}} \), i.e. \( y_{\text{soft}} = w_{\text{soft}} + u_{\text{measured}} \).

A nonparametric method can be used to obtain the predicted values \( w_{\text{soft}} \) of the log-mass
fraction. The nonparametric method has the advantage of not being constrained by model
assumptions. A space/time moving window is an example of a way the log-mass fraction can be
estimated non-parametrically. An estimator \( \hat{E}_S[w_{\text{soft},j}] \) for the expected value \( E_S[w_{\text{soft},j}] \) of the
random variable \( w_{\text{soft},j} \) representing the log-mass fraction at any soft data point \( p_j \in p_{\text{soft}} \) is:

\[
\hat{E}_S[w_{\text{soft},j}] = \frac{\sum_{i=1}^{N_h(p_j)} w_{\text{hard},i}}{N_h(p_j)} \quad (A6)
\]

where \( N_h(p_j) \) is equal to the number of points \( p_i \in p_{\text{hard}} \) such that \( d(p_i,p_j) \leq D \), with \( d(p_i,p_j) \) a
composite space/time distance function (Christakos et al. 2002) and \( D \) a space/time distance
specified such that points \( p_i \) are \( p_j \) are in an air shed affected by the same fire. Similarly, the
estimator \( \hat{V}_S[w_{\text{soft},j}] \) for the variance is:

\[
\hat{V}_S[w_{\text{soft},j}] = \frac{\sum_{i=1}^{N_h(p_j)} (w_{\text{hard},i} - \hat{E}_S[w_{\text{soft},j}])^2}{(N_h(p_j)-1)} \quad (A7)
\]

The log-PAH soft data at point \( p_j \) is \( y_{\text{soft},j} = w_{\text{soft},j} + u_{\text{measured},j} \). Since it is assumed that the
log-PM\(_{2.5}\) measured value, \( u_{\text{measured},j} \), is obtained without error (or at least with a measurement
error variance that is much smaller than the prediction error variance \( V_S[w_{\text{soft},j}] \)), it follows that estimators for the expected value and variance of \( y_{\text{soft},j} \) are given by

\[
\hat{E}_S[y_{\text{soft},j}] = \hat{E}_S[w_{\text{soft},j}] + u_{\text{measured},j} \quad (A8)
\]

114
and
\[
\hat{V}_S[y_{soft,j}] = \hat{V}_S[w_{soft,j}] \quad \text{(A9)}
\]
respectively. These estimators provide knowledge about the mean and variance of log-PAHs. Without knowledge about other statistical moments, the PDF for the soft data that maximizes information entropy is Gaussian. Hence the soft data for log-PAHs may be given by the following site-specific PDF:
\[
f_s(y_{soft,j}) = \left(2\pi \hat{V}_S[w_{soft,j}]\right)^{-1/2} \exp\left(-\frac{(y_{soft,j} - \hat{E}_S[y_{soft,j}])^2}{2\hat{V}_S[y_{soft,j}]}ight) \quad \text{(A10)}
\]
Therefore, the site specific knowledge about log-PAHs can be divided into two categories: the hard data directly measured at points \(p_{\text{hard}}\), and the soft data with PDF given by Eqs. (A6)-(A10) created using a mass fraction approach from the additional log-PM\(_{2.5}\) measurements available at points \(p_{\text{soft}}\). The hard and soft data for the residual log-PAH S/TRF \(X(p)\) are finally obtained from that of \(Y(p)\) using \(X(p) = Y(p) - m_Y(p)\), where the mean trend \(m_Y(p)\) is a deterministic (known) function.

**Bayesian Maximum Entropy Estimation**

For exposure assessment purposes, it is necessary to estimate residual log-PAHs over the entire space/time domain. Given a domain and \(p_{\text{data}} = [p_{\text{hard}}, p_{\text{soft}}]\), space/time estimation points \(p_k\) are chosen to complement \(p_{\text{data}}\). The points \(p_k\) should at least be regularly spaced on a grid so that the entire domain is covered. The number of points in this grid depends on a balance of grid size and computing power. However, additional points can be added to the set \(p_k\) to improve resolution in areas important to a particular study. These might include places with higher population density or those known to have a high exposure gradient. The points \(p_k\) are
space/time locations where log-PAH residuals are estimated based on the data available at $p_{hard}$ and $p_{soft}$, using the covariance structure $c_x(p, p')$. The set $p_k$ will be a subset of $p_{map} = [p_{hard}, p_{soft}, p_k]$. Therefore, the mapping points of the log-PAH S/TRF will include points with hard PAH data, soft PAH data derived from PM$_{2.5}$, and estimation points important to producing a particular exposure map for a space/time domain.

A full description of the BME framework has been described elsewhere (Christakos 1990, 2000; Serre and Christakos 1999; Christakos et al. 2002). By way of summary, given the general and site-specific knowledge bases discussed above, the information about the exposure field of PAHs can be processed under the following three stages of the BME framework.

(i) **Structural (or prior) stage:** The general knowledge base ($G$) is evaluated at all mapping points $p_{map} = [p_{hard}, p_{soft}, p_k]$. The maximization of a Shannon measure of information entropy at $p_{map}$ under constraints derived from the general knowledge $G$ leads to the selection of the prior PDF $f_G(\chi_{map})$.

(ii) **Specificatory (or meta-prior) stage:** The site-specific knowledge base ($S$) is organized into the hard and soft data, $\chi_{hard}$ and $\chi_{soft}$, respectively. The hard data is obtained from direct measurements of PAHs, while the soft data is obtained from PM$_{2.5}$ measurements using the mass fraction approach (Eqs. A6-A10).

(iii) **Integration stage:** The general knowledge base $G$ and site specific knowledge base $S$ are integrated using an operational Bayesian conditionalization rule that leads to the following posterior PDF

$$f_k(\chi) = A^{-1} \int_{-\infty}^{+\infty} f_S(\chi_{soft}) f_G(\chi_{map}) d\chi_{soft} \quad (A11)$$
where the subscript $K$ represents the physical knowledge $K = G \cup S$, and $A$ is a normalization parameter. This posterior PDF defines the statistical distribution of the exposure field at any estimation point.
Appendix B

Radiello Lab Instructions

Hydrogen sulfide (H₂S)

_radiello components to be used_

<table>
<thead>
<tr>
<th>White diffusive body</th>
<th>code</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>supporting plate code</td>
<td>121</td>
<td></td>
</tr>
<tr>
<td>vertical adapter code</td>
<td>122</td>
<td>(optional)</td>
</tr>
<tr>
<td>chemiadsorbing cartridge code</td>
<td>170</td>
<td></td>
</tr>
</tbody>
</table>

**Principle**

The cartridge code 170 is made of microporous polyethylene and impregnated with zinc acetate. Hydrogen sulphide is chemiadsorbed by zinc acetate and transformed into stable zinc sulfide. The sulfide is recovered by extraction with water. In contact with an oxidizing agent as ferric chloride in a strongly acid solution it reacts with the N,N-dimethyl-p-phenylenediammonium ion to yield methylene blue.

\[
\text{N,N-dimethyl-p-phenylenediammonium} \rightarrow \text{methylene blue}
\]

Methylene blue is quantified by visible spectrometry.

**Sampling rate**

Sampling rate \(Q_{298}\) at 298 K (25°C) and 1013 hPa is \(0.096 \pm 0.005\) ng·ppb\(^{-1}\)·min\(^{-1}\) or 69 ml·min\(^{-1}\).

**Effect of temperature, humidity and wind speed**

Sampling rates varies from the value at 298 K on the effect of temperature (in Kelvin) as expressed by the following equation
\[ Q_K = 0.096 \left( \frac{K}{298} \right)^{3.8} \]

where \( Q_K \) is the sampling rate at the temperature \( K \) and \( Q_{298} \) is the reference value at 298 K. This produces a variation of ±5% for 10 °C variation (upwards or downwards) from 25 °C. Sampling rate is invariant with humidity in the range 15-90% and with wind speed between 0.1 and 10 m·s\(^{-1}\).

**Calculations**

Once \( Q_K \) at the sampling temperature has been calculated, the concentration \( C \) is obtained according to the equation:

\[ C = \frac{m}{Q_K \cdot t \cdot 1.000} \]

where \( m \) is the mass of sulphide ion in µg found onto the cartridge and \( t \) is the exposure time in minutes.

**Exposure**

Exposure duration is allowed to vary from 1 hour to 15 days. Sampling is linear from 2,000-50,000,000 ppb·min of H\(_2\)S.

**Limit of detection and uncertainty**

The limit of detection is 30 ppb for 1 hour exposure or 1 ppb for 24 hours exposures. The uncertainty at 2s is 8.7% over the whole allowed exposure range.

**Storage**

The cartridges are stable at least for 12 months before and 6 months after exposure. Do not expose all of the cartridges belonging to the same lot (lot number and expiry date are printed onto the plastic bag): keep at least two of them as blanks.

**Analysis**

**Reactives**

- **sulphuric acid**: slowly add 25 ml of concentrated sulphuric acid to 10 ml water and let the solution cool.
- **amine**: dissolve 6.75 g of N,N-dimethyl-p-phenylenediammonium oxalate in the sulphuric acid solution. Dilute this solution to 1 liter with sulphuric acid - water 1:1 v/v. Kept in a dark bottle and well capped, this solution is stable for at least four weeks. **CAUTION**: this solution is very poisonous.
**ferric chloride**: dissolve 100 g of ferric chloride hexahydrate (FeCl₃·6H₂O) in 40 ml of water

**ferric chloride-amine**: mix 10 ml of ferric chloride solution with 50 ml of amine solution. This solution has to be freshly prepared.

**sulphuric acid for dilution**: slowly dissolve 40 ml of concentrated sulphuric acid in 900 ml of water, let the solution cool and make up to 1,000 ml.

**Procedure**

Add 10 ml of water to the plastic tube containing the cartridge, recap and stir vigorously, preferably by a VORTEX stirrer.

Add 0.5 ml of ferri chloride - amine solution, recap immediately and stir. The tube must be capped immediately in order to avoid that the developed hydrogen sulfide can escape from the tube before reacting. Wait for 30 minutes and measure absorbance at 665 nm using water to zero the spectrophotometer. The colour is stable for several weeks.

Do the same with two or three unexposed cartridges of the same lot and obtain the average blank value, then subtract it to the samples. Be careful to apply the same dilution ratio to the samples and the blanks.

**IMPORTANT:**

Absorbance is linear up to 1,200 absorbance units, corresponding to an exposure value of about 80,000 ppb·min. If higher absorbance values are obtained, dilute the samples with the sulphuric acid for dilution.

**NEVER USE WATER TO DILUTE.**

**Calibration**

Calibration curves may be prepared by sodium sulfide standard solutions, that have to be titrated just before use. As diluted sodium sulfide are very unstable (the sulfide content can diminish as much as the 10% in an hour time) it is strongly recommended to make use of the calibration solution code 171, following the instructions included.

**user tip**

**Code 171 calibration solution** relieves you from the task of preparation and titration of the sodium sulfide solutions.