Statistical Predictions Based on Accelerated Degradation Data and Spatial Count Data

Yuanyuan Duan

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Yili Hong, Chair
Jie Li, Co-Chair
Feng Guo
Inyoung Kim
William H. Woodall

February 19, 2014
Blacksburg, Virginia

Key Words: Coatings, Covariate process, Clusters, Divide-Recombine, Environmental conditions, Lifetime prediction, Lyme disease, Kernel smoothing, Photodegradation, Usage history, UV exposure, Random effects, Reliability, Spatio-temporal.
Statistical Predictions Based on Accelerated Degradation Data and Spatial Count Data

Abstract

This dissertation aims to develop methods for statistical predictions based on various types of data from different areas. We focus on applications from reliability and spatial epidemiology. Chapter 1 gives a general introduction of statistical predictions. Chapters 2 and 3 investigate the photodegradation of an organic coating, which is mainly caused by ultraviolet (UV) radiation but also affected by environmental factors, including temperature and humidity. In Chapter 2, we identify a physically motivated nonlinear mixed-effects model, including the effects of environmental variables, to describe the degradation path. Unit-to-unit variabilities are modeled as random effects. The maximum likelihood approach is used to estimate parameters based on the accelerated test data from laboratory. The developed model is then extended to allow for time-varying covariates and is used to predict outdoor degradation where the explanatory variables are time-varying.

Chapter 3 introduces a class of models for analyzing degradation data with dynamic covariate information. We use a general path model with random effects to describe the degradation paths and a vector time series model to describe the covariate process. Shape restricted splines are used to estimate the effects of dynamic covariates on the degradation process. The unknown parameters of these models are estimated by using the maximum likelihood method. Algorithms for computing the estimated lifetime distribution are also described. The proposed methods are applied to predict the photodegradation path of an organic coating in a complicated dynamic environment.

Chapter 4 investigates the Lyme disease emergency in Virginia at census tract level. Based on areal (census tract level) count data of Lyme disease cases in Virginia from 1998 to 2011, we analyze the spatial patterns of the disease using statistical smoothing techniques. We also use the space and space-time scan statistics to reveal the presence of clusters in the spatial and spatial/temporal distribution of Lyme disease.

Chapter 5 builds a predictive model for Lyme disease based on historical data and environmental/demographical information of each census tract. We propose a Divide-Recombine method to take advantage of parallel computing. We compare prediction results through simulation studies, which show our method can provide comparable fitting and predicting accuracy but can achieve much more computational efficiency. We also apply the proposed method to analyze Virginia Lyme disease spatio-temporal data. Our method makes large-scale spatio-temporal predictions possible. Chapter 6 gives a general review on the contributions of this dissertation, and discusses directions for future research.
To my family.
Acknowledgments

I would like to express my deepest gratitude to my advisors, Drs. Yili Hong and Jie Li, for their invaluable support, inspiration, suggestions, and guidance throughout the research, especially for their patience and editing skills in the completion of this dissertation. I would like to thank my committee members, Drs. Feng Guo, Inyoung Kim, and William H. Woodall, for serving on my committee, and for the feedback, direction, and assistance when I need it.

I would also like to extend my appreciation to the faculty and staff in the Department of Statistics at Virginia Tech. Special thanks goes to Dr. Birch, Director of Graduate Programs, for always being so helpful.

My thanks also go to all the talented graduate students in Dr. Hong’s research group, Khaled Bedair, Caleb King, Zhibing Xu, Yimeng Xie, and Miao Yuan for your insightful discussion during the group meetings and editorial helps.

A special gratitude and love goes to my family for their love and continued support.

I would like to acknowledge Advanced Research Computing at Virginia Tech for providing computational resources (URL: http://www.arc.vt.edu). This dissertation was partially supported by the National Science Foundation under grants CMMI-1068933 and BCS-1122876 to Virginia Tech.
Contents

List of Figures x

List of Tables xiii

1 General Introduction 1

1.1 Background ................................. 1

1.1.1 General Statistical Prediction .................. 1

1.1.2 Reliability .................................. 2

1.1.3 Spatial Epidemiology ......................... 4

1.2 Motivation ................................... 5

1.2.1 Link Field and Laboratory Result .............. 6

1.2.2 Incorporate Dynamic Covariates ................ 6

1.2.3 Spatio-Temporal Prediction for Epidemiology Data 7

1.3 Overview .................................... 7

Bibliography 8

2 Service Life Prediction Based on Accelerated Degradation Test Data from Laboratory and Field 11

2.1 Introduction ................................ 12

2.1.1 Background ................................ 12

2.1.2 Related Literature .......................... 13

2.1.3 Overview .................................. 14

2.2 Photodegradation Data ........................ 15

2.2.1 Indoor Experiments and Data .................. 15

2.2.2 Outdoor Experiments and Data ................ 18

2.2.3 Notation for Data ............................ 20

2.2.4 Data Cleaning ............................... 20
2.3 Models for Photodegradation Paths ........................................ 21
  2.3.1 The Physical Model ...................................................... 21
  2.3.2 The Statistical Model ................................................... 22
2.4 Modeling of Indoor Data ...................................................... 22
  2.4.1 Initial Analysis of Indoor Data ....................................... 22
  2.4.2 Effects of Explanatory Variables .................................... 23
  2.4.3 The Combined Model ................................................... 27
2.5 The Prediction Model for Outdoor Data .................................. 28
  2.5.1 The Cumulative Damage Model for Outdoor Degradation .......... 28
  2.5.2 Outdoor Predictions .................................................... 31
2.6 Conclusions and Areas for Future Research .............................. 32

Bibliography 36

3 Statistical Methods for Degradation Data with Dynamic Covariates Information and an Application to Outdoor Weathering Data 39
  3.1 Introduction ................................................................. 40
    3.1.1 Background and Motivation ........................................ 40
    3.1.2 Related Literature .................................................... 41
    3.1.3 Overview ............................................................... 42
  3.2 Data .............................................................................. 42
    3.2.1 NIST Outdoor Weathering Data .................................... 42
    3.2.2 Notation for the Data ................................................ 44
  3.3 Model for a Degradation Path ............................................. 46
    3.3.1 General Path Model .................................................... 46
    3.3.2 Modeling Degradation Path with Dynamic Covariates .......... 46
    3.3.3 Functional Forms for Covariate-Effect Function \( f(\cdot) \) .......... 47
  3.4 Degradation Path Model Parameter Estimation ......................... 50
    3.4.1 Parameter Estimation .................................................. 50
    3.4.2 Estimation for Weathering Data .................................... 53
  3.5 Simulation Study ............................................................. 56
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5.1 Simulation Model and Setup</td>
<td>56</td>
</tr>
<tr>
<td>3.5.2 Simulation Results</td>
<td>57</td>
</tr>
<tr>
<td>3.6 Model for Multivariate Covariate Process</td>
<td>59</td>
</tr>
<tr>
<td>3.6.1 General Strategy for Covariates Modeling</td>
<td>59</td>
</tr>
<tr>
<td>3.6.2 Parametric Models for Covariates for Outdoor Weathering Data</td>
<td>59</td>
</tr>
<tr>
<td>3.6.3 Parameter Estimation</td>
<td>61</td>
</tr>
<tr>
<td>3.7 Failure-time Distribution Estimation</td>
<td>64</td>
</tr>
<tr>
<td>3.7.1 Failure-time Distribution for the Population</td>
<td>64</td>
</tr>
<tr>
<td>3.7.2 Application to the Epoxy Degradation Example</td>
<td>66</td>
</tr>
<tr>
<td>3.7.3 Distribution of Remaining Life for Individual Units</td>
<td>67</td>
</tr>
<tr>
<td>3.8 Concluding Remarks and Areas for Future Research</td>
<td>69</td>
</tr>
<tr>
<td>3.A Nonparametric Residual Bootstrap with Adjustment</td>
<td>72</td>
</tr>
<tr>
<td>3.B Graphical Check of Degradation Data Model Fit</td>
<td>74</td>
</tr>
<tr>
<td>3.C Selection of Knots and Spline Orders</td>
<td>74</td>
</tr>
<tr>
<td>3.D Graphical Check of Covariate Process Model Fits</td>
<td>75</td>
</tr>
</tbody>
</table>

Bibliography

4 Spatial and Temporal Emergence Pattern of Lyme Disease in Virginia

Bibliography
5 Statistical Prediction for Virginia Lyme Disease Emergence Based on Spatio-temporal Count Data

5.1 Introduction

5.1.1 Motivation

5.1.2 Existing Methods

5.1.3 Overview

5.2 Data and Model

5.2.1 Spatial Region, Time Domain, and Data

5.2.2 Spatio-Temporal Model

5.3 The Proposed Divide-Recombine Method

5.3.1 Region Division

5.3.2 Divide-Recombine Procedure

5.3.3 A Simple Illustration

5.4 Simulation Study

5.4.1 Simulation Setup

5.4.2 Models for Comparison

5.4.3 Simulation Results

5.5 Extension of the Model to Include Spatio-temporal Interactions

5.5.1 Model with Spatio-temporal Interactions

5.5.2 Divide-Recombine Method for Model with Interaction

5.6 Application to Lyme Disease Count Data

5.6.1 Data

5.6.2 Region Division

5.6.3 Model

5.6.4 Results

5.7 Conclusions and Areas for Future Research

Bibliography

6 General Conclusions

6.1 Conclusions
# List of Figures

2.1 Plot for indoor light lamp spectral irradiance. .......................... 16
2.2 Illustration of FTIR peaks. ................................................. 17
2.3 Degradation paths for specimens with 10%, 40%, 60% and 100% ND filter, BP filter 353 nm, temperature 35°C, and RH 0%. ...................... 18
2.4 Plots of the degradation path for an outdoor unit (a), the temperature (b) as a function of time, the RH (c), and the perspective plot for the UV exposure (d) as a function of time and wavelength. ......................... 19
2.5 Plots of categorical effects of (a) UV, (b) ND filter, (c) temperature, and (d) RH. 24
2.6 Plots of measured (points) and fitted value (lines) by functional form damage model. .................................................. 29
2.7 Plots of measured (points) and predicted value (solid lines) and 95% prediction interval (dashed lines) by cumulative damage model. .................. 33
2.8 Plots of averaged outdoor degradation measurements and predicted values by cumulative damage model. ................................. 34
2.9 Plots of predicted values adjusted by random effects estimated from the 5th to 10th data points. ............................................. 35
3.1 Plot of nine representative degradation paths and dynamic covariate information (the dots show the daily values, connected by line segments). ...... 45
3.2 Splines bases for the covariate effects of UV dosage, temperature, and RH. 50
3.3 Estimated covariate-effect functions for UV dosage, temperature, and RH, and the corresponding approximate 95% pointwise CIs. ............... 55
3.4 Plot of degradation measurements and fitted degradation path for the nine representative specimens shown in Figure 3.1. ............................. 55
3.5 Spline bases (left) and the UV effect function (right) used in simulation. ... 57
3.6 Estimated MSE of the parameter estimators and CP of CI procedure based on 600 repeats for $\beta_0, \sigma_1$, and $\sigma_\varepsilon$. .............................. 58
3.7 Estimated MSE for UV effect function estimator (left) and CP for pointwise CI for UV effect function (right) based on 600 repeats. ............................... 58
3.8 The fitted mean and variance structures for UV dosage, temperature and RH. 63
3.9 The estimated cdf and corresponding 95% pointwise CIs for a population with units starting randomly from day 161 to day 190. ................................. 68
3.10 Estimated expected number of failures and corresponding 95% pointwise CIs versus the observed number of failures as a function of time for the 36 specimens in the outdoor weathering data. ........................................... 68
3.11 The estimated conditional cdf of remaining life and corresponding 95% pointwise CIs for unit G18-10. ................................................................. 70
3.12 Plot of studentized residuals and Q-Q plot of the residuals for the degradation data. ................................................................. 74
3.13 Plot of the AIC values for different number of interior knots under different degrees of smoothness. ................................................. 75
3.14 Sensitivity analysis on the spline orders and the number of interior spline knots on the estimation of covariate-effect functions. .............................. 76
3.15 The plot of ACF for the residuals from the models for UV, temperature, and RH covariate data. ................................................................. 76
3.16 Q-Q plots for the residuals for the UV, temperature, and RH covariate data. 77
4.1 Plots of raw Lyme disease rates from 1998 to 2005. ................................. 87
4.2 Plots of raw Lyme disease rates from 2006 to 2011. ................................. 88
4.3 Plots of smoothed Lyme disease rates from 1998 to 2005. ......................... 91
4.4 Plots of smoothed Lyme disease rates from 2006 to 2011. ......................... 92
4.5 Lyme disease clusters from 1998 to 2005. ........................................... 94
4.6 Lyme disease clusters from 2006 to 2011. ........................................... 95
4.7 Plots of Lyme disease prospective clusters. ........................................... 96
4.8 Plots of Lyme disease retrospective clusters. ........................................... 96
5.1 A simple illustration for the DR procedure. ........................................... 116
5.2 Example of simulated spatial Poisson data (the color shows the counts). ... 117
5.3 Plot for the computing time (in seconds) for different sample sizes and methods, when $\tau_s^2 = 10$ (both axes are in log scale). 120
5.4 Plots of Lyme disease counts from 2003 to 2010. 126
5.5 Level III Ecoregions of Virginia (image source: www.hort.purdue.edu). 128
5.6 Plots of 2011 Virginia Lyme disease counts and predicted mean. 130
5.7 Plots of 2011 Virginia Lyme disease counts and predicted mean. 130
5.8 Plots of 2011 northeast Virginia Lyme disease counts and predicted mean. 131
List of Tables

2.1 Indoor experiment setups. ................................................. 15
2.2 Available temperature (°C) and RH (%) combinations (✓: available; ∅: not available). ......................................................... 16
2.3 Parameter estimates of the categorical effect model. .................. 25
2.4 Parameter estimates of the functional form effects model. ............ 28
3.1 Parameter estimates and approximate 95% CIs for \( \beta_0, \sigma_0, \sigma_1, \rho \) and \( \sigma_\varepsilon \) in the degradation path model. .................................................. 54
3.2 Parameter estimates and corresponding 95% bootstrap CIs for the parameters of the covariate process model in (3.9). ......................... 64
4.1 Lyme disease primary cluster summary by years. .......................... 97
4.2 Lyme disease space-time cluster analysis (prospective) summary, showing the number of census tracts, the average % of population, the % of area, the relative risk, the expected cases and observed cases within each cluster. .......... 98
4.3 Lyme disease space-time cluster analysis (retrospective) summary, showing the number of census tracts, the average % of population, the % of area, the relative risk, the expected cases and observed cases within each cluster. .......... 99
5.1 Computing time in seconds for different sample sizes and methods, when \( \tau_s^2 = 10.120 \)
5.2 Estimation and prediction accuracy when \( N = 100, 400, \tau_s^2 = 10, 1, 0.5, \text{ and } \tau_t^2 = 10. \) ......................................................... 121
5.3 Estimation and prediction accuracy when \( N = 784, 1,296, \tau_s^2 = 10, 1, 0.5, \text{ and } \tau_t^2 = 10. \) ......................................................... 122
5.4 Virginia Lyme disease yearly count data. ................................... 125
5.5 Virginia Lyme disease explanatory variables. .............................. 127
Chapter 1  General Introduction

1.1  Background

1.1.1  General Statistical Prediction

Employing statistics frequently serves two purposes: description and prediction. Statistical prediction is based on generalizability; if observed data can provide enough information to reveal a certain pattern after analysis, we hope we can make some statement about unobserved quantity (Aitchison and Sculthorpe, 1965). The statement may serve as one of the two important goals: inference making or decision guidance. Inference type prediction is to suggest intervals in which the future outcome possibly falls. Decision type prediction provides a potential guide for making decisions from a decision space consisting of subsets of an outcomes space.

Prediction can be challenging because of the uncertainty. An important source of uncertainty is the uncertainty in parameter values. Let the observed data be $X = x$ and let $X$ follow a distribution $F(x; \theta)$, where $\theta$ is the unknown parameter vector. We want to predict an unobserved (future) value of some random variable $Y$. That is, we want to obtain the distribution of $Y|X = x$, which is the conditional predictive distribution $F(y; x, \theta)$. Because $\theta$ is unknown, we need to estimate $\theta$ from data $X = x$ and the estimated value of $\theta$ also has uncertainty. We need to account for this variation in our predictive distribution (Lawless and Fredette, 2005).

In reliability studies, predictions can give the risk of failing for products and systems. It can be used to evaluate the need for back-up products/systems, the need for extreme environmental protection systems, and also to help conduct further criticality analyses. Uncertainty is important in reliability prediction because the prediction of risks for future events has uncertainty. In epidemiology studies, one major objective is to predict, by pointing out a set
of conditions or factors that are related to the disease (Aycock and Russell, 1943). Based on the prediction, resources and treatments can then be optimally utilized. Uncertainty is also important in epidemic prediction. Elderd et al. (2006) showed that the uncertainty in parameter value estimates had important consequences for uncertainty in the predicted number of deaths of population under two vaccination strategies. All predictions can eventually be compared with the actuality. In this way prediction can be used to validate the success of our statistical model (Aitchison and Dunsmore, 1980).

1.1.2 Reliability

Reliability studies involve collecting data on failure times of products and systems. Reliability data mainly come from laboratory experiments, field tests, and warranty records. Reliability assessment for modern products and systems in the field in a timely manner can be challenging. Modern products are usually designed to last a long time. Usually only a few or no failures occur during field tracking studies, making statistical inference challenging.

1.1.2.1 Degradation Data in Reliability

For these modern products, degradation data provide an important resource for obtaining reliability information. Many failures can be linked to an underlying degradation process of the parts of a product/system. For example, fatigue is a major failure mechanism of glass, metals, and ceramics. Wearout of rubber tires and deterioration of paints will gradually cause loss of functionality. Corrosion/oxidation/tarnishing will cause the failure of protective coatings. When it is available, degradation data can provide more quantitative measures for each individual unit before the unit fails. Although not all systems provide degradation data, there are many that do. Power output from a satellite transmitter, power from voltage inverters, rechargeable battery capacity, light intensity in a light-emitting diode (LED), and tire wear are all good examples of degradation data that can be used in quantitative analysis.
1.1.2.2 Accelerated Degradation Test

To speed up the degradation process and obtain reliability information in a timely manner, accelerated degradation tests are commonly used. It has been proved to be useful to predict failure time in engineering and mechanical applications. Many examples are available in Nelson (1990). The analysis at accelerated levels can provide more reliable and precise reliability estimates and a firmer basis for extrapolation at normal use conditions (Meeker et al., 1998). Common accelerating variables include the use rate or aging rate of a product, exposure intensity, voltage stress, temperature, humidity, etc. (Escobar and Meeker, 2006). Multiple acceleration variables are also used. Meeker and Escobar (1998) illustrated acceleration models led by both physical and statistical considerations.

1.1.2.3 Dynamic Covariates

For products and systems in the field, the degradation process usually involves dynamic environmental covariates. For example, degradation of LED power output is mainly due to usage that can be time-varying. Corrosion of crude oil transportation pipeline is subject to outdoor environment and characteristics of the compounds flowing in the pipeline. Degradation of photovoltaic arrays can be caused by both the time-varying usage and outdoor environments. Modern sensor technology allows many systems to collect dynamic data, including products/systems usage and environmental variables. For example, sensors can dynamically record oil pressure and oil/water temperature in locomotive engines (Hahn and Doganaksoy, 2008). A power inverter in solar panel arrays can collect and transmit information on the output of power, the ambient temperature, and humidity every few seconds. The degradation rate is affected by the dynamic covariates information but it is challenging to incorporate such information into the degradation process.
1.1.2.4 Cumulative Damage Model

The cumulative damage model is used to describe failure time data with time-varying covariates. Nelson (1990) used the cumulative damage model to analyze life test data with stepwise varying stress. Bagdonavičius and Nikulin (2001), Vaca-Trigo and Meeker (2009), Hong and Meeker (2010), and Hong and Meeker (2011) also applied this model to failure time data with dynamic covariates. In this model, the amount of cumulative damage at given environmental covariates history is the accumulation of damage per unit time. The cumulative damage model can also be particularly useful for degradation data analysis with dynamic covariates.

1.1.3 Spatial Epidemiology

Epidemiology is the study of associations between disease outcomes and factors that cause, reduce or prevent disease. Spatial statistics provide useful tools for epidemiological research. An early example given in practice is Snow (1855). It was found that the clustering of London’s cholera death rate was among residents using drinking water from a particular well, thus identifying that this disease was transmitted through contaminated water. Following Snow (1855), statistical methods are widely used in spatial epidemiology analysis to detect patterns, identify clusters, assess potential exposures, etc.

1.1.3.1 Areal Data

When data are observed in polygonal units on a map, it is often called areal data. These units are usually defined by administrative divisions, postcodes, census tracts, etc. Areal data, such as population, case counts of a certain disease, are associated with regions rather than particular points (Sahu and Mardia, 2005). Areal data are more common in epidemiology studies in the U. S. because precise location of disease cases is usually under strict confidentiality regulations (Carlin and Banerjee, 2003).
1.1.3.2 Spatio-Temporal Autocorrelation

The analysis of epidemiology data is challenging. It is especially hard when there is spatial or temporal autocorrelation. That is, there exists a systematic pattern of the response variable or a certain explanatory variable for data collected in each spatial location or in a time interval. Large autocorrelation will reduce the effective sample size (Cressie, 1993) and will make it hard to distinguish between a spatial/temporal trend and spatial/temporal autocorrelation. The spatial autocorrelation on areal data depends on the choice of areal units. For example, crop yields may have stronger autocorrelation at the county level than at the state level.

1.1.3.3 Spatial CAR Model

The conditional auto-regressive (CAR) model can be used to model spatial autocorrelation. For example, areal count data are usually modeled by spatial Poisson regression. In particular, let $Y_i$ denote the number of cases of the disease of interest in area $i$, and let $P_i$ denote the population in area $i$, where $i = 1, \ldots, N$. The spatial Poisson regression model is

$$Y_i|\upsilon_i \sim \text{Poisson}(P_i\mu_i), \quad \text{and} \quad \log(\mu_i) = X_i\beta + v_i,$$

where $X_i$ is the covariate matrix and $\beta$ is the regression coefficient vector. Here $v_i$ is the random effect capturing spatial autocorrelation and follows a conditional autoregressive (CAR) with distribution $v_i|\upsilon_{i^*} \sim N(\bar{\upsilon}_{i^*}, \sigma^2_i)$ (Carlin and Banerjee, 2003). Here, $i^*$ are neighbors of $i$, $\bar{\upsilon}_{i^*}$ is the conditional mean, and $\sigma^2_i$ is the conditional variance. The CAR can also be extended to the spatial temporal CAR model for spatio-temporal data.

1.2 Motivation

The main goals of this dissertation are to develop statistical prediction procedures that can be used to link field and laboratory test data, deal with field data with dynamic covariate information, and deal with data of spatial/temporal dependence. The research is motivated
by case study data. The methodologies that we develop, however, are general and can be applied to many other situations.

1.2.1 Link Field and Laboratory Result

It will be practically useful if we can link indoor experiment data to field performance data. Gu et al. (2009) described three approaches to link indoor accelerated degradation data with outdoor exposure data for an epoxy coating system. They showed that the reliability-based predictive model could make successful linkage and prediction. More details of this reliability-based predictive model can be found in Vaca-Trigo and Meeker (2009). However, the accuracy of the predictions still needs improvement. The individual sample variability was not considered and there was no quantification of prediction uncertainty. Moreover, functional forms of environmental effects still need further study.

To improve Vaca-Trigo and Meeker (2009), we extensively investigate the functional forms of environmental effects based on indoor data. We use the cumulative damage model and estimate parameters from the indoor study to predict the outdoor degradation, incorporating unit-to-unit temporal dynamic information. We identify a physically motivated model with individual random components to describe unit-to-unit variability in the degradation data. The developed method is applied to accelerated laboratory test data for a specific coating, in which environmental variables were controlled over time. We also extend the model estimated by indoor data to outdoor field test data, where the environmental variables were uncontrolled.

1.2.2 Incorporate Dynamic Covariates

Motivated by the increasing availability of dynamic covariate information being acquired by systems operated in the field and the need to predict future performance of these systems, we develop a class of models and methods for using such data. We use flexible general path models with individual random components to describe unit-to-unit variability in the degradation data. We develop algorithms to estimate the failure-time distribution induced by the underlying degradation model. We use the NIST outdoor weathering data as our
motivating example and illustrations. The developed methods, however, can have broad applications for many products used in highly variable environments and/or subject to time-varying usages.

1.2.3 Spatio-Temporal Prediction for Epidemiology Data

Epidemiology data with spatial and/or temporal autocorrelation is challenging. We use areal (census tract level or county level) count data of Lyme disease as an illustrative example. We build an spatio-temporal model to incorporate both spatial and temporal autocorrelation and develop a procedure for predictions. We also considered principle component analysis to incorporate the social-demographical and environmental information.

1.3 Overview

The rest of this dissertation is organized as follows. In Chapter 2, a non-linear mixed model is used to study the indoor data. The effects of the explanatory variables are extensively studied. We then extend the model from indoor data to outdoor data, allowing for dynamic covariates. In Chapter 3, we use a linear mixed model with nonparametric functional forms of environment effects, and apply it to the outdoor epoxy coating degradation data. Shape-restricted splines are used to estimate the effects of dynamic covariates on the degradation process. Chapter 3 is based on a paper that has been tentatively accepted by *Technometrics* (Hong et al., 2014). In Chapter 4, we analyze the spatial patterns and clustering of the disease. Chapter 4 is based on a paper that is under revision for *American Journal of Tropical Medicine and Hygiene* (Li et al., 2014). In Chapter 5, we develop a Divide-Recombine method to predict future counts of Lyme disease cases in Virginia. In Chapter 6, we give a general review on the contributions of this dissertation, as well as discuss directions for future research.
Bibliography


Chapter 2  Service Life Prediction Based on Accelerated Degradation Test Data from Laboratory and Field

Abstract

Service life prediction is of great interest in coating industry. Photodegradation caused by ultraviolet (UV) radiation is the primary cause of failure for coatings, as well as many other products made from organic materials exposed to sunlight. Other environmental factors including temperature and humidity also have effects on the degradation process. Identifying a physically motivated model that can adequately describe the degradation path is an important step in service life prediction of such coatings. The model should also incorporate the effects of explanatory variables such as, the UV spectrum, UV intensity, temperature and humidity. In this chapter, we use a nonlinear mixed model to describe the data because the degradation path is nonlinear function of time. Random effects are also used to account for unit-to-unit variability. The parameters in the model are estimated by maximum likelihood approach. The functional forms of the explanatory variables are extensively studied. The developed method is applied to accelerated laboratory test data for a specific coating, in which spectral UV wavelength and intensity, temperature, and relative humidity are controlled over time. We also extend our model to allow for time-varying covariates and apply the method to outdoor test data where the explanatory variables are uncontrolled.

Key Words: Coatings, Photodegradation, Random effects, Reliability, UV exposure.
2.1 Introduction

2.1.1 Background

Service life prediction by reliability-based methods has been of great interest. Its importance has been increasingly realized in many fields, such as engineering and material sciences. Examples include the prediction of the service life of automobile tires (MacIsaac and Feve, 2009) and the lifetime of coating/paints with acceptable gloss (Martin et al., 1996). Photodegradation caused by ultraviolet (UV) radiation is the primary cause of failure for paints/coatings, as well as many other products made from organic materials exposed to sunlight. Environmental variables including temperature and humidity can also affect degradation rates. Service life prediction of such paints and coatings is difficult because the intensity and the frequency spectrum of UV radiation from the sun are highly variable (both temporally and spatially), and because there is often little understanding of how environmental variables affect chemical degradation processes.

To study the mechanism of degradation processes, both indoor and outdoor experiments were conducted at the U. S. National Institute of Standards and Technology (NIST) over the past years to generate necessary experimental data. Laboratory weathering tests were conducted on the NIST SPHERE (Simulated Photodegradation via High Energy Radiant Exposure), a device in which spectral UV wavelength, spectral intensity, temperature, and relative humidity (RH) can be precisely and accurately controlled over time. Outdoor exposure experiments were conducted on the roof of a NIST laboratory. The chemical degradation of the epoxy coatings exposed was measured every few days by Fourier transform infrared (FTIR) transmission. Longitudinal information on ambient temperature, RH, and the solar spectrum for outdoor specimens were carefully recorded at 12 minute intervals for several months to a year.

In this chapter, a physically motivated model with functional forms of the covariate effects is used to model the data. Random effects are also used to account for variability of individ-
ual specimens. The results facilitate the understanding of the photodegradation mechanism. Based on the indoor data, we build our service life prediction model and use the outdoor data to validate the proposed prediction method.

2.1.2 Related Literature

Traditional failure-time tests can require a long time to obtain enough failures because modern products are designed to last a long time. To overcome the time constraints of failure-time tests, degradation data which can provide quantitative measurements and thus more information than failure data are usually applied. Lu and Meeker (1993) and Hong et al. (2010) gave examples of models and analyses of degradation data. To speed up the degradation process and help gathering information in a more timely manner, accelerated degradation tests are commonly used (Nelson, 1990). The analysis at accelerated stress levels can provide more credible and precise reliability estimates and a firmer basis for extrapolation at normal use conditions (Meeker et al., 1998). Accelerating variables include the use rate or aging rate of product, exposure intensity, voltage stress, temperature, humidity etc. (Escobar and Meeker, 2006). Combinations of these accelerating variables are usually used. Meeker and Escobar (1998) illustrated acceleration models led by both physical and statistical considerations.

For products and systems in the field, the degradation process usually involves dynamic environmental covariates. Dynamic data collection is also becoming much easier with modern sensor technology, motivating the modeling of the effect of dynamic covariates. For example, Hahn and Doganaksoy (2008) described sensors recording dynamic covariates such as oil pressure, oil/water temperature in locomotive engines. Spurgeon et al. (2005) described an automatic system that can monitor dissolved gas in the insulating oil in high-voltage power transformers. The degradation is affected by dynamic information but it is challenging to incorporate such information into the degradation process. Cumulative damage model has been used for time-to-event data with time varying covariates (e.g., Nelson 1990, Subramanian et al. 1995, Bagdonavičius and Nikulin 2001, Vaca-Trigo and Meeker 2009, Hong and Meeker 2010, and Hong and Meeker 2011), where the amount of cumulative damage for the
given environmental covariates history is the accumulation of damage per unit time.

It takes a long time to observe the performance history of new products/systems. For example, many new coatings are springing up rapidly at present, to build performance history for a new coating, we need extensive outdoor exposure, which could take decades (Martin et al., 1996). Besides, the outdoor exposure conditions are complex, due to the joint effects of multiple covariates. While indoor experimental data is short-term, the effect of single covariate can be modeled and the experimental condition can be well controlled. Thus, it will be practically useful if the indoor experimental data and outdoor performance data can be linked. Gu et al. (2009) described three approaches to link indoor accelerated degradation data with outdoor exposure data for an epoxy coating system. A reliability-based predictive model was described which made linkage between indoor data and outdoor data. More details of this reliability-based predictive model are also available in Vaca-Trigo and Meeker (2009). In summary, this predictive model used a nonlinear model for indoor data and a cumulative damage model for outdoor weathering data. The accuracy of prediction, however, still needs improvement. For example, sample to sample variability was not considered and quantification of prediction uncertainty was lacking. In this chapter, we propose a nonlinear mixed-effects model with carefully modeling of environment effects to provide a stronger prediction model. We also quantify the uncertainty by providing prediction interval.

2.1.3 Overview

The rest of this chapter is organized as follows. In section 2.2, we introduce the indoor and outdoor experiment settings and notation for data. In section 2.3, we describe the nonlinear mixed model and defines total effective dosage. In section 2.4, we provide estimates of categorical effects of environmental variables and estimates functional forms of these conditions based on indoor data. In section 2.5, we use parameters estimated from indoor data and cumulative damage effect model to predict for outdoor experiment. Section 2.6 contains conclusions and some discussion.
Table 2.1: Indoor experiment setups.

<table>
<thead>
<tr>
<th>BP filter</th>
<th>306 nm (±3), 326 nm (±6), 353 nm (±21), 452 nm(±79)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND filter</td>
<td>10%, 40%, 60%, 100%</td>
</tr>
<tr>
<td>Temperature</td>
<td>25°C, 35°C, 45°C, 55°C</td>
</tr>
<tr>
<td>RH</td>
<td>0, 25, 50, 75</td>
</tr>
</tbody>
</table>

2.2 Photodegradation Data

2.2.1 Indoor Experiments and Data

The UV light source for indoor experiments is from high intensity light lamps. The maximum total intensity achievable is equivalent to 22 suns. The spectral irradiance of light lamps is denoted by Lamp(λ), where λ is the wavelength. The spectral irradiance of the indoor light lamps is illustrated in Figure 2.1. The irradiance is defined as the power of electromagnetic radiation per unit area incident on a surface. To study the effect of UV spectrum (λ) and UV intensity, the spectral irradiance of the lamps is modified and controlled by bandpass (BP) filters and neutral density (ND) filters. A BP filter only passes UV with wavelengths within a given range. For example, the 306 nanometer (nm) BP filter has a nominal center wavelength of 306 nm and full-width-half maximum values of ±3 nm. The four BP filters used in the experiments have nominal center wavelength of 306 nm, 326 nm, 353 nm, and 452 nm. The ND filter controls the intensity of the UV exposure. For example, a 10% ND filter passes 10% of the exposure. The four ND filters used in the experiments are 10%, 40%, 60% and 100%.

The indoor experiments also have controlled environmental factors such as temperature, and RH. Table 2.1 gives a summary of the experimental factors for the levels of the experiment setups. The temperature levels are 25°C, 35°C, 45°C, and 55°C. The RH levels are 0%, 25%, 50%, and 75%. There are 319 specimens in total for the indoor experiments. Note that not all combinations of the four experimental factors have been run in the experiments. Table 2.2 shows available temperature and humidity combinations.

The damage can not be measured directly but can be measured by FTIR transmission.
Figure 2.1: Plot for indoor light lamp spectral irradiance.

Table 2.2: Available temperature (°C) and RH (%) combinations (✓: available; Ø: not available).

<table>
<thead>
<tr>
<th>Temp</th>
<th>RH</th>
<th>0</th>
<th>25</th>
<th>50</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td></td>
<td>✓</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
</tr>
<tr>
<td>35</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Ø</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td>Ø</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>55</td>
<td></td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>✓</td>
</tr>
</tbody>
</table>
Specific structures of compounds absorb the energy of the infrared at different wavelengths, which causes peaks in the FTIR spectra. The location of the peak usually corresponds to a unique structure and thus can be used to identify the relative concentration of different compounds. The height of a peak is proportional to the concentration of a particular compound or structure. Figure 2.2 gives an illustration of FTIR peaks. Our current modeling focuses on damage wave number 1250 cm$^{-1}$, which corresponds to C-O stretching of aryl ether. Other peaks being used as potentially useful responses include 1510 cm$^{-1}$ (benzene ring mass loss), 1658 cm$^{-1}$ (oxidation products), and 2925 cm$^{-1}$ (CH mass loss) (Bellinger and Verdu, 1984, 1985; Rabek, 1995; Kelleher and Gesner, 1969). As an example of the typical degradation collected in indoor experiments, Figure 2.3 displays degradation paths at damage wave number 1250 cm$^{-1}$ for specimens with 10%, 40%, 60% and 100% ND filters, BP filter 353 nm, temperature 35°C, and RH 0%.
2.2.2 Outdoor Experiments and Data

The UV exposure for the outdoor specimens is from the sun. The UV spectral irradiance, temperature, and RH are, of course, uncontrolled, but are recorded at a 12-minute interval. There are 55 specimens in total for the outdoor experiments. The measurements of degradation are taken every three to four days as the indoor specimens. We still focus on damage wave number $1250 \text{ cm}^{-1}$. As an example of the outdoor data, Figure 2.4 gives the degradation path for an outdoor unit and plots for dynamic environmental variables.

Note that the measurements of chemical degradation and environmental information of outdoor experiments are at different frequencies. The measurement of degradation usually takes every three or four days. The spectral irradiance, temperature, and RH are recorded at 12-minute intervals.
Figure 2.4: Plots of the degradation path for an outdoor unit (a), the temperature (b) as a function of time, the RH (c), and the perspective plot for the UV exposure (d) as a function of time and wavelength.
2.2.3 Notation for Data

The number of incident photons, defined as dose, is denoted by $E_i(\tau_{ik}, \lambda)$, for specimen $i$ at time $\tau_{ik}$ from wavelength $\lambda$ after BP and ND filters, $i = 1, \ldots, n$, $k = 1, \ldots, m_i$. Here $n$ is number of specimens, $m_i$ is the number of time points where the dose is recorded.

The number of incident photons absorbed by a specimen at time $\tau_{ik}$, defined as dosage, is denoted by $D(\tau_{ik}, \lambda)$, where $D(\tau_{ik}, \lambda) = E_i(\tau_{ik}, \lambda)\{1 - \exp[-A(\lambda)]\}$, and $A(\lambda)$ is the spectral absorbance of specimen at specified wavelength $\lambda$. The total number of photons absorbed by a specimen at all wavelengths, defined as total dosage, is denoted by $D_{\text{total}} = \int_{\tau} \int_{\lambda} D(\tau, \lambda)d\lambda d\tau$. We also define $D_{\text{total}}(\lambda) = \int_{\tau} D(\tau, \lambda)d\tau$ as total number of photons absorbed by a specimen at a specific wavelength $\lambda$. The environmental temperature and RH for specimen $i$ at time $\tau_{ik}$ are denoted by $\text{Temp}_i(\tau_{ik})$ and $\text{RH}_i(\tau_{ik})$, respectively. The ND filter level for specimen $i$ at time $\tau_{ik}$ is denoted by $\text{ND}_i(\tau_{ik})$.

The degradation measurement at time $t_{ij}$ is denoted by $Y_i(t_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, M_i$, which is the change of peak’s height before/after exposure measured by FTIR. Here $M_i$ is the number of time points where the degradation measurements are taken and $Y_i(t_{ij})$ measures the damage to a particular compound structure in the coatings, the damage wave number interested is 1250 cm$^{-1}$. Let $t_i = t_{iM_i}$ be the last observation time.

2.2.4 Data Cleaning

The data are preprocessed before the data analysis. During the indoor experiments, the failure of an integrated circuit chip caused one of the chambers to be overheated for a period of time. Thus we dropped data of samples from this chamber (NSS-0011). The degradation path of another sample (NSS-0008-16) shows 2 segments instead of a continuous curve. This might be caused by a problem with this sample, so we also dropped this sample in the data analysis. The degradation also shows a more complicated pattern after damage is higher than 0.6, this behavior might be caused by a change in the degradation mechanism in the specimen. When fitting our model using indoor experimental data, we use data above $-0.6$ because $-0.6$ is far
beyond the definition of failure.

For the outdoor experiments, temperature and/or RH data of some time points (e.g., time $t$, day $i$) are missing, we use the average temperature/RH value at the same time $t$ of each day within 14 days interval (day $i - 7$ to day $i + 7$, if available).

### 2.3 Models for Photodegradation Paths

#### 2.3.1 The Physical Model

The total effective dosage is defined as

$$ S(t) = \int_{0}^{t} \int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} D(\tau, \lambda)\phi(\lambda)d\lambda d\tau. \quad (2.1) $$

Here the function $\phi(\lambda)$ is the quasi-quantum yield function allowing for the fact that a shorter wavelength causes a larger damage. The wavelengths that are of interest are between $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$.

To allow for the environmental effects, consider the following model

$$ S(t) = \int_{0}^{t} f[\text{Temp}(\tau)] g[\text{RH}(\tau)] d[\text{ND}(\tau)] \int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} D(\tau, \lambda)\phi(\lambda)d\lambda d\tau. \quad (2.2) $$

Here $f[\text{Temp}(\tau)]$, $g[\text{RH}(\tau)]$ and $d[\text{ND}(\tau)]$ are the instantaneous acceleration factors due to temperature, humidity effect, and density, respectively. Also, Temp($\tau$) and RH($\tau$) are the environmental temperature and relative humidity (RH) at time $\tau$, respectively. In practice, D($\tau, \lambda$) is recorded but $\phi(\lambda)$ is unknown and needs to be estimated.

Little is known, however, about the functional form of $\phi(\lambda)$ in chemistry literature. Vaca-Trigo and Meeker (2009) used a simple log linear relationship. The estimation of $\phi(\lambda)$ from experimental data can help us understanding material properties and how the UV exposure affect the degradation process for different wavelengths. Environmental factors such as temperature and humidity can accelerate the degradation process. The Arrhenius relationship is widely used to describe the acceleration effect of temperature. The way relative humidity...
and neutral density filter affects the degradation process, however, is unknown. That is, the functional forms of $g$ and $d$ are unknown and need to be estimated from experimental data.

2.3.2 The Statistical Model

In the general path model, the degradation measurement of unit $i$ at time $t_{ij}$ is modeled as

$$Y_i(t_{ij}) = D_i(t_{ij}) + \epsilon_i(t_{ij}),$$

(2.3)

where $D_i(t_{ij})$ is the actual degradation path and $\epsilon_i(t_{ij})$ is the corresponding measurement error. The photodegradation is mainly caused by the effective dosage $S(t)$. By kinetic theory, the degradation path can be well approximated by the following parametric model

$$D_i(t_{ij}) = \frac{\alpha \exp(v_i)}{1 + \exp(-z)},$$

(2.4)

where $z = \{\log[S(t_{ij})] - \mu\}/\sigma$, $\mu$ and $\sigma$ are the parameters describing the location and steepness of the damage curve. The quantity $\alpha$ is the maximum degradation damage when total effective dosage goes to infinity, and $v_i$ is individual random effects for the degradation path, which are caused by uncontrolled and/or unobservable factors.

2.4 Modeling of Indoor Data

2.4.1 Initial Analysis of Indoor Data

The degradation path for damage number 1250 is decreasing. For indoor specimens, the experimental factors are controlled. Thus Temp($\tau$), RH($\tau$), ND($\tau$), and $E(\tau, \lambda)$ are assumed to be time-invariant. Note that $E_i(\lambda) = \text{Lamp}(\lambda)\text{Filter}(\lambda, BP_i, ND_i)$, where Lamp($\lambda$) is the spectral irradiance for the light lamp and Filter($\lambda, BP_i, ND_i$) is a function of BP filter $BP_i$.
and ND filter ND\textsubscript{i}. Thus

\[ z = \frac{\log[S(t_{ij})] - \mu}{\sigma} = \frac{\log[b(BP\textsubscript{i})] + \log[f(Temp\textsubscript{i})] + \log[g(RH\textsubscript{i})] + \log[d(ND\textsubscript{i})] - \mu}{\sigma}, \quad (2.5) \]

where

\[ b(BP\textsubscript{i}) = \int_{\lambda_{min}}^{\lambda_{max}} \int_{0}^{t_{i}} \text{Lamp}(\lambda)\text{Filter}(\lambda, BP\textsubscript{i}, ND\textsubscript{i}) \{1 - \exp[-A(\lambda)]\} \phi(\lambda) \, d\lambda \, d\tau \quad (2.6) \]

is the UV exposure effect when the BP filter is BP\textsubscript{i} and the ND filter is ND\textsubscript{i}. Assume \( \phi(\lambda) \) is constant in a certain small range of \( \lambda \), denoted by \( \bar{\phi}(\lambda) \), which is the average quasi-quantum yield function of this \( \lambda \) range, then \( b(BP\textsubscript{i}) = D_{\text{total}} \bar{\phi}(\lambda) \).

Model (2.5) is called as the “categorical effect model” because the effects of the four experimental factors are modeled as categorical variables. The constraint \( f(35) = g(25) = d(10) = 1 \) is used to make the parameters to be estimable. The estimates of parameters in model (2.5) are obtained by maximum likelihood (ML) estimation method. It is implemented by using the \texttt{nlme} function in R. Degradation paths in a small wavelength interval share similar steepness [i.e., 306 nm (±3)]. We assume \( \sigma \) is mainly determined by wavelength. Thus we denote it as \( \sigma_{\lambda} \), which is a categorical variable based on 4 wavelength intervals.

Table 2.3 lists the ML estimates of the fixed-effect parameters in model (2.5). Although the categorical effect model only provides the estimates of UV, temperature, RH effects at limited points, it is useful for choosing the functional forms of \( \phi(\lambda) \), \( d(ND) \), \( f(Temp) \), \( g(RH) \), and \( \sigma_{\lambda} \) in next modeling stage.

### 2.4.2 Effects of Explanatory Variables

Here we interpret the effects of different explanatory variables.
Figure 2.5: Plots of categorical effects of (a) UV, (b) ND filter, (c) temperature, and (d) RH.
Table 2.3: Parameter estimates of the categorical effect model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Std. Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>-0.6810</td>
<td>0.0130</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[\phi(306)] - \mu )</td>
<td>-6.5620</td>
<td>0.0755</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[\phi(326)] - \mu )</td>
<td>-7.0844</td>
<td>0.0361</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[\phi(353)] - \mu )</td>
<td>-9.0275</td>
<td>0.0323</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[\phi(452)] - \mu )</td>
<td>-10.1087</td>
<td>0.0350</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[d(40)] )</td>
<td>-0.7939</td>
<td>0.0201</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[d(60)] )</td>
<td>-1.0553</td>
<td>0.0199</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[d(100)] )</td>
<td>-1.3082</td>
<td>0.0200</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[f(25)] )</td>
<td>-0.1963</td>
<td>0.0092</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[f(45)] )</td>
<td>0.1973</td>
<td>0.0247</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[f(55)] )</td>
<td>-0.8193</td>
<td>0.0357</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[g(0)] )</td>
<td>0.8749</td>
<td>0.0231</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[g(50)] )</td>
<td>-0.3707</td>
<td>0.0255</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \log[g(75)] )</td>
<td>0.2287</td>
<td>0.0240</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \sigma(306) )</td>
<td>1.5591</td>
<td>0.0149</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \sigma(326) )</td>
<td>1.2336</td>
<td>0.0074</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \sigma(353) )</td>
<td>1.0443</td>
<td>0.0057</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>( \sigma(452) )</td>
<td>0.8416</td>
<td>0.0054</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

2.4.2.1 BP Filter Effect

To obtain a simple estimate of \( \phi(\lambda) \), one can first assume \( \phi(\lambda) \) is constant in a certain range of \( \lambda \). Because \( \int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} D(\tau, \lambda) d\lambda \) is known, one can obtain the estimate of \( \phi(\lambda) \) in certain intervals by the categorical effect model. For example,

\[
\bar{\phi}(306) = \frac{b(306)}{\int_{\tau}^{309} D(\tau, \lambda) d\lambda d\tau} = \frac{b(306)}{D_{\text{total}}}
\]

for \( 303 \leq \lambda \leq 309 \). Similarly, one can obtain the estimates of \( \phi(\lambda) \) for other BP filters, \( 320 \leq \lambda \leq 332, 332 \leq \lambda \leq 374, \) and \( 373 \leq \lambda \leq 531 \). Figure 2.5(a) shows the simple estimate of \( \phi(\lambda) \). The results suggest that for shorter wavelength, the damage is much stronger than longer wavelength. The curve suggests an exponential relationship.

We only have \( D_{\text{total}} \) recorded, that is the total dosage over 4 wavelength ranges. To estimate \( \phi(\lambda) \) more precisely, we need to estimate \( D_{\text{total}}(\lambda) \) for each wavelength, not just
for the 4 wavelength intervals. Note that Lamp(\(\lambda\))Filter(\(\lambda\), BP\(_i\), ND\(_i\)) follows the pattern as in Figure 2.1. The approximate trapezoid area under each \(\lambda\) is denoted as Area\(_{\lambda}\), the integration of Area\(_{\lambda}\) of the 4 wavelength ranges are denoted as Area\(_{\lambda}\), where \(\lambda\) is 306, 326, 353 or 452nm. We define the proportion of area under \(\lambda\) to its corresponding wavelength range as \(P(\lambda) = \frac{\text{Area}_{\lambda}}{\text{Area}_{\lambda}}\). For small intervals such as 303 \(\leq \lambda \leq 309\) and 320 \(\leq \lambda \leq 332\), one can assume \(\{1 - \exp[-A(\lambda)]\}\) is a constant. Thus for 303 \(\leq \lambda \leq 309\) and 320 \(\leq \lambda \leq 332\), \(D_{\text{total}}(\lambda) = \text{Lamp}(\lambda)\text{Filter}(\lambda, \text{BP}_i, \text{ND}_i)\{1 - \exp[-A(\lambda)]\}\) could be approximated by \(D_{\text{total}} \times P(\lambda)\). For 373 \(\leq \lambda \leq 531\), \(\{1 - \exp[-A(\lambda)]\}\) is very small, we assume \(D_{\text{total}}(\lambda) = \text{Lamp}(\lambda)\text{Filter}(\lambda, \text{BP}_i, \text{ND}_i)\{1 - \exp[-A(\lambda)]\}\) is constant over the range. The function \(\phi(\lambda)\) is modeled as an exponential function that \(\phi(\lambda) = \exp(\beta_0 + \beta_\lambda \lambda)\), where \(\beta_0\) and \(\beta_\lambda\) are parameters that describe this exponential relationship. For 332 \(\leq \lambda \leq 374\), lamp spectra curve is too complicated to do trapezoid approximation and \(\{1 - \exp[-A(\lambda)]\}\) is not small enough, so we still assume \(\phi(353)\) is constant and treat it categorically.

The curve of categorical estimates of \(\sigma_\lambda\) suggests an exponential relationship with a lower bound. Thus we use the functional form \(\sigma_\lambda = \sigma_0 + \exp(\sigma_1 + \sigma_2 \lambda)\) to model the effect of UV spectrum on \(\sigma\).

### 2.4.2.2 ND Filter Effect

Figure 2.5(b) shows the effects of ND filter. A power law relationship \(d(\text{ND}_i) = \text{ND}_i^p\) gives a perfect fit of this curve. Note that the Filter(\(\lambda\), BP\(_i\), ND\(_i\)) already includes the effect of ND filter as ND\(_i\) with power one. If the reciprocity law stands, \(p\) should be equal to 1 – 1 = 0.

### 2.4.2.3 Temperature Effect

Figure 2.5(c) shows the effects of temperature. The Arrhenius relationship is widely used to describe the acceleration effect of temperature (Meeker et al., 1998). In particular, the Arrhenius relationship is

\[
f(\text{Temp}_i) = \gamma_0 \exp \left( \frac{-E_a}{\text{TempK}_i} \right)
\]  

(2.7)
where $\text{TempK}_i$ is the Kelvin temperature computed as Celsius temperature plus 273.15. The temperature effect should be monotonic increasing according to the Arrhenius relationship. The categorical estimates follow the trend and can fit a straight line except for the samples under temperature 55℃. Because we have samples under temperature 55℃ at only one RH level, this might be caused by interaction of temperature and humidity, because water release is known to affect the degradation. However, we can not estimate interactions because not all combinations of the four experiment factors are done in indoor tests. Thus we remove data from samples under temperature 55℃ for following parameter estimation.

2.4.2.4 RH Effect

There is no literature about the functional form of humidity effect. Figure 2.5(d) shows the effects of RH. The effect is increasing first and then decreasing, suggesting a concave relationship. Thus we use the quadratic relationship

$$\log[g(RH)] = -\beta_{RH}(RH - rh_0)^2$$

(2.8)

as the functional form of RH effects. Here $rh_0$ is the x intercept of the quadratic formula.

2.4.3 The Combined Model

Combining all the functional forms of each effect, we have model

$$\mathcal{D}_i(t_{ij}) = \frac{\alpha \exp(v_i)}{1 + \exp(-z)},$$

(2.9)

where

$$z = \eta_0 + \log(D_{total}) + \log \left[ \frac{\int_{\lambda_{\min}}^{\lambda_{\max}} P(\lambda) \exp(\beta_\lambda \lambda) d\lambda}{(\lambda_{\max} - \lambda_{\min})} \right] + p(\log[ND_i]) - \left[ \frac{\Delta E_{R}}{\text{TempK}_i} \right] - \beta_{RH} [RH_i - rh_0]^2$$

$$\sigma_0 + \exp(\sigma_1 + \sigma_2 \lambda)$$
Table 2.4: Parameter estimates of the functional form effects model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Std.Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>-0.6191</td>
<td>0.0101</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$\beta_\lambda$</td>
<td>-0.0297</td>
<td>0.00026</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$p$</td>
<td>-0.5606</td>
<td>0.0078</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$E_R$</td>
<td>1945.6234</td>
<td>75.8383</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$\beta_{RH}$</td>
<td>-0.0005</td>
<td>0.00001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$rh_0$</td>
<td>45.4761</td>
<td>0.2875</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$\eta_0$</td>
<td>13.4159</td>
<td>0.2539</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$\sigma_0$</td>
<td>0.8019</td>
<td>0.0066</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>7.6784</td>
<td>0.1876</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>-0.0260</td>
<td>0.0006</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

where $D_{total} \times P(\lambda) \times \exp(\beta_\lambda \lambda) \approx D_{total}(\lambda) \exp(\beta_\lambda \lambda)$ contains the information of total effective dosage, denoted by $S^*$. Note the actual total effective dosage (without environmental effects) $S$ is $D_{total}(\lambda) \exp(\beta_0) \exp(\beta_\lambda \lambda)$. Here $\eta_0$ is a constant which need to be estimated. It contains the information of $\mu$ and $\beta_0$. Table 2.4 lists the ML estimates of the parameters in model (2.9). The maximum degradation damage when total effective dosage goes to infinity is -0.6191, not considering random effects. For ND filter effect, the power $p$ is estimated to be $-0.5606$, not 0. Thus there is evidence that the reciprocity law does not hold. Combined with ND effect in $\text{Filter}(\lambda, BP_i, ND_i)$, $1 - 0.5606 = 0.4394$, ND$^{0.4394}$ is the overall effect of ND filters. This could also be seen in Figure 2.3. When all other conditions are the same, larger ND filter effect causes more damage, but the difference of damage between 4 ND levels are not as large as the difference between 4 ND levels. Here $\beta_\lambda < 0$ indicates that shorter wavelength has larger damage. Figure 2.6 shows examples of fitted indoor data using model (2.9), our model works well for indoor data, especially when damage is not too large.

### 2.5 The Prediction Model for Outdoor Data

#### 2.5.1 The Cumulative Damage Model for Outdoor Degradation

The outdoor environmental data are taken every 12 minutes and $D(\tau, \lambda)$ from 300 to 532nm are also recorded every 12 minutes, with a 2nm resolution. Wavelengths are recorded as
Figure 2.6: Plots of measured (points) and fitted value (lines) by functional form damage model.
300, 302, · · · , 532nm. Based on the resolution, we assume the corresponding wavelength intervals to be 299 ≤ \( \lambda \) ≤ 301, 302 ≤ \( \lambda \) ≤ 303, · · · , 531 ≤ \( \lambda \) ≤ 533, each with 2nm length. For computing convenience, we set 60 minutes instead of 12 minutes as unit time, and define unit effective dosage as

\[
\Delta S^*(\tau, \lambda) = \int_\tau^{\tau+60\text{min}} \int_{\lambda-1\text{nm}}^{\lambda+1\text{nm}} D(\tau, \lambda) \exp(\beta_\lambda \lambda) d\lambda d\tau.
\]  

(2.10)

The effective dosage of all wavelengths at time \( \tau \) is \( S^*(\tau) = \int_\lambda \Delta S^*(\tau, \lambda) d\lambda \). The total effective dosage of all wavelengths from the beginning to time \( t \) is \( S^*(t) = \int_\tau S^*(\tau) d\tau \). Temperature and RH are averaged over every 60 minutes. We assume there is no ND filter effect from earth atmosphere and set ND to be 100%.

By the cumulative damage model, the slope of the degradation curve at time \( \tau \) is a function of total effective Dosage \( S^*(t) \) and other environmental effects. That is,

\[
g'(\tau, \lambda) = \frac{dD[\tau, S^*(t), \text{Temp}(\tau), \text{RH}(\tau)]}{\Delta S^*(\tau, \lambda)} = \frac{1}{\Delta S^*(\tau, \lambda) \sigma_\lambda} \times \frac{\alpha \exp(z)}{(1 + \exp(z))^2}
\]

(2.11)

where

\[
z = \frac{\log (S^*|\beta_\lambda|) + \eta_0 + p[\log(ND_i)] - \left( \frac{E_i}{\text{temp} K} \right) - \beta_{RH} (\text{RH} - \text{rh}_0)^2}{\sigma_0 + \exp\left( \sigma_1 + \sigma_2 \lambda \right)}
\]

(2.12)

The unit damage, \( \Delta D(\tau, \lambda) \) is damage at time \( \tau \) caused by a certain 2nm wavelength interval (\( \lambda - 1, \lambda + 1 \)). In particular, \( \Delta D(\tau, \lambda) = g'(\tau, \lambda) \Delta S^*(\tau, \lambda) \). The additivity law is assumed, such that the damage is summed up from each wavelength interval in every 60-minute time slice.

Then \( \Delta D(\tau) \) denotes damage at time \( \tau \) from all wavelengths, \( \Delta D(\tau) = \sum_{\text{range of } \lambda} \Delta D(\tau, \lambda) \). The cumulative damage \( D(t) \) from time 0 to \( t \) from all wavelengths is \( D(t) = \sum_0^t \Delta D(\tau) \).
2.5.2 Outdoor Predictions

Figure 2.7 shows the comparison of measured and predicted values of some outdoor samples using our cumulative damage model, the random effect \( v_i \) is set to zero for all outdoor samples. Prediction intervals are calculated and calibrated following the procedure described in Hong et al. (2009), using the Lawless and Fredette’s predictive distribution (Lawless and Fredette, 2005):

1. Simulate \( B \) bootstrap samples \( \hat{\theta}_{ib} \sim N(\hat{\theta}, \Sigma) \) and \( \hat{\mathbf{v}}_{ib} \sim N(0, \sigma^2_v), b = 1, \ldots, B \). We use \( B = 50,000 \).

2. Compute the bootstrap version of the data \( \mathbf{D}^{*}_{ib}, b = 1, \ldots, B \) by proposed cumulative damage model \( F(\mathbf{D}|\hat{\theta}, \hat{\mathbf{v}}) \).

3. Compute \( U^*_{ib} = F(\mathbf{D}^{*}_{ib}|\hat{\theta}_{ib}, \hat{\mathbf{v}}_{ib}), b = 1, \ldots, B \).

4. Let \( u^l_i, u^u_i \) be the lower and upper \( \alpha/2 \) sample quantiles of \( U^*_{ib} \), respectively. Solve \( F(\mathbf{D}^{l}_{i}|\hat{\theta}, \hat{\mathbf{v}}) = u^l_i, F(\mathbf{D}^{u}_{i}|\hat{\theta}, \hat{\mathbf{v}}) = u^u_i \) for \( (\mathbf{D}^{l}_{i}, \mathbf{D}^{u}_{i}) \), which is the \( 100(1 - \alpha)\% \) calibrated prediction interval.

For some samples, their predicted values fit the measured values well, while for others the predicted values are either bigger or smaller than the measured values, these correspond well to the distribution of random effects. Most of the measured data points are within the calibrated prediction intervals, except some small values which may be due to the measurement error. Because the random effects are normally distributed with mean 0, the average predicted values should be closed to the averaged measured values for all outdoor samples. Figure 2.7 shows the average of predicted and measured damage for all outdoor samples, the average predicted values correspond well to average measured values.

It is also found that the random effects tend to be similar in the same outdoor group, for example, 4 samples from G1 outdoor group all have predictions larger than the measured values, 4 samples from G16OUT group all have predictions smaller than the measured values,
and 4 samples from G4 group all have predictions close to the measured values. These suggest the random effects should be subject to some group conditions such as seasonal effect. The random effects could be estimated by a non-intercept linear function. Figure 2.9 shows examples we estimated random effect \( \exp(v_i) \) using the 5th to 10th data points by a least square fit of \( y_{ij} = \exp(v_i) \hat{y}_{ij} \), then \( \hat{y}_{ij} \) is adjusted by multiplying a factor \( \exp(\hat{v}_i) \). The first 4 data points are not used because their damage values are usually so small that we could not omit the measurement error. The blue lines indicate the predicted value after adjusting, for most samples, the adjusted predicted values match the measured value better than un-adjusted. This suggests a possibility that we could use earlier data points to make future prediction adjusted by random effects.

### 2.6 Conclusions and Areas for Future Research

This chapter serves as an improvement and extension of the predictive model in Vaca-Trigo and Meeker (2009). We develop a physical motivated model with random effects to model damage data, extensively study the functional forms of these environmental effects and estimate parameters from indoor data, then uses cumulative damage model to incorporate these estimated parameters and unit-to-unit temporal dynamic information to predict the outdoor degradation path. We also apply an algorithm to calculate the prediction intervals. The unit-to-unit random effect can be calculated by some early data points. In this way we are possible to make further prediction adjusted by random effects if we could have earlier degradation data. In the future, it would be interesting to consider how other effects, such as seasonal effect, affect the random effects.
Figure 2.7: Plots of measured (points) and predicted value (solid lines) and 95% prediction interval (dashed lines) by cumulative damage model.
Figure 2.8: Plots of averaged outdoor degradation measurements and predicted values by cumulative damage model.
Figure 2.9: Plots of predicted values adjusted by random effects estimated from the 5th to 10th data points.
Bibliography


Chapter 3  Statistical Methods for Degradation Data with Dynamic Covariates Information and an Application to Outdoor Weathering Data

Abstract

Degradation data provide a useful resource for obtaining reliability information for some highly reliable products and systems. In addition to product/system degradation measurements, it is common nowadays to dynamically record product/system usage as well as other life-affecting environmental variables such as load, amount of use, temperature, and humidity. We refer to these variables as dynamic covariate information. In this chapter, we introduce a class of models for analyzing degradation data with dynamic covariate information. We use a general path model with individual random effects to describe degradation paths and a vector time series model to describe the covariate process. Shape restricted splines are used to estimate the effects of dynamic covariates on the degradation process. The unknown parameters in the degradation data model and the covariate process model are estimated by using maximum likelihood. We also describe algorithms for computing an estimate of the lifetime distribution induced by the proposed degradation path model. The proposed methods are illustrated with an application for predicting the life of an organic coating in a complicated dynamic environment (i.e., changing UV spectrum and intensity, temperature, and humidity).

Key Words: Covariate process, Environmental conditions, Lifetime prediction, Organic coatings, System health monitoring, Usage history.
3.1 Introduction

3.1.1 Background and Motivation

For products and systems with high reliability, it is challenging to do field reliability assessment in a timely manner based only on limited lifetime data. When available, degradation data provide a useful resource for obtaining reliability information because there are degradation measurements for each individual unit in the field before the individual unit fails. For products with degradation driven by usage and environmental conditions, information about these variables can be important for modeling the degradation process. For example, the degradation of organic coatings is primarily driven by Ultraviolet (UV) exposure, while temperature and humidity are other important factors. There are many other more examples where degradation is driven by usage and environmental variables such as, the loss of light output from an LED array, the decrease of power output of photovoltaic arrays, the corrosion in an oil transportation pipeline, the vibration from a bearing in a wind turbine, and the loss of color and gloss of an automobile coating.

Developments in technology allow many systems to collect and transmit massive amount of information. It is common nowadays to dynamically record product/system usage and load as well as other environmental variables such as temperature and humidity, which we refer to as dynamic covariate information. For example, even a small device like a power inverter that is used in solar panel arrays can gather and transmit information on the output of power, the ambient temperature and humidity every few seconds. The availability of such large-scale dynamic data creates many opportunities and challenges.

Dynamic covariate data contain rich information that can be useful for modeling and predicting product reliability. One can expect those units which are heavily used and are used under the most extreme environments to fail sooner than those units with lighter usage under normal environmental conditions. Thus it is attractive to incorporate dynamic covariate information into degradation modeling and data analysis, especially when predictions are
required for individual units.

Although not all systems will provide degradation data, there are many that will. Examples include power output from a satellite transmitter, power from solar cells, power from voltage inverters, light output from an LED array, number of paper jams per week in a printer/copier, rechargeable battery capacity, etc. The main goal of this chapter is to develop general models for analyzing degradation data and dynamic covariate information for a fleet of products. Based on the degradation data model, one can obtain estimates for the remaining lifetime distribution and predictions for the product population and for individual units in the field. We use data from an outdoor weathering experiment to illustrate the models and methods.

3.1.2 Related Literature

In literature, general path models are commonly used to analyze degradation data (e.g., Lu and Meeker 1993). For a specified failure definition, the cumulative distribution function (cdf) of the lifetime distribution is induced by the parametric model for the degradation paths. Stochastic models are another class of models to analyze degradation data (e.g., Lawless and Crowder 2004). The stochastic model approach assumes that the data are generated from a stochastic process, such as a Wiener process, a gamma process, or an inverse Gaussian process. By the properties of the assumed underlying stochastic process, the cdf of the lifetime distribution can be obtained. Details on parameter estimation for various degradation models are available in Chapter 13 of Bagdonavičius and Nikulin (2001a). Singpurwalla (1995) considered both univariate and multivariate survival models under dynamic environments.

For degradation data analysis, covariate information and the modeling of covariates are available in several settings such as accelerated repeated-measures degradation tests (e.g., Meeker et al. 1998, and Bagdonavičius and Nikulin 2001b), accelerated destructive degradation tests (e.g., Escobar et al. 2003), and degradation-test experimental designs (e.g., Joseph and Yu 2006, and Park and Padgett 2006). Bagdonavičius et al. (2010) described a stochastic degradation model with time varying covariates. The time-varying covariates are incorporated into the induced cdf of the degradation process through a cumulative damage model (e.g., see
Little work has been done in degradation data modeling that also considers unit-to-unit or temporal variability for covariates. The modeling of the effect of the dynamics on degradation can provide valuable information in several areas. For example, degradation information is important in the area of system health monitoring or condition-based maintenance, where dynamic covariate information is available for continuously monitored systems. Thus, general models for analyzing degradation with dynamic covariate information need to be developed.

### 3.1.3 Overview

The rest of the chapter is organized as follows. Section 3.2 introduces a motivating example from the National Institute of Standards and Technology (NIST) outdoor weathering of epoxy coating experiments. Section 3.2 also introduces the data structure and notation for degradation data with dynamic covariates. Sections 3.3 proposes a general additive model for incorporating dynamic covariate information into the degradation path model. Section 3.4 develops parameter estimation procedures. Section 3.5 uses simulation to validate the inference procedure. Section 3.6 describes parametric models for a multivariate covariate process and the corresponding procedures for parameter estimation. Section 3.7 develops procedures for failure-time distribution estimation based on the parametric models given in Sections 3.3 and 3.6. Section 3.8 contains some concluding remarks and describes possible areas for future research.

### 3.2 Data

#### 3.2.1 NIST Outdoor Weathering Data

The illustrative application is from the NIST outdoor weathering data. The data were collected in a study of the service life of organic coatings in outdoor environments. Outdoor weathering experiments were carried out in Gaithersburg, MD, from 2002 through 2006. There were 36 specimens placed in outdoor environmental chambers on the roof of a building on the
NIST campus, starting at different times over a period of approximately five years. The outdoor temperature, humidity, and Ultraviolet (UV) spectrum and intensity were recorded over this period of time automatically by sensors. See Gu et al. (2009) for more details. In this application, all units were exposed at the same location. There could be various exposure patterns in different applications, as discussed in Section 3.7.1.

A degradation measurement is proportional to the damage to the coating and it was measured periodically for each specimen using Fourier transform infrared spectroscopy (FTIR). The degradation measurements were taken at intervals of several days. For illustration, we consider the degradation that occurs at damage number $1250 \text{ cm}^{-1}$ on the FTIR spectrum, which is attributed to C-O stretching of aryl ether. Figure 3.1a shows nine representative degradation paths from nine specimens, started at different times of the year and on different years. A large part of the variability in these data is due to the varying amount of UV exposure during the nine different periods of time. The time scale of the degradation measurement, denoted by $t$, is the time in days since the start of exposure for a specimen.

The samples used for experiments were specially fabricated using a model epoxy that was known to degrade rapidly. Also, the samples were very thin. In this manner, similar to an accelerated test, experimental information would be obtained on individual units in a timely manner (i.e., in months instead of in years). Although the specimens are different from coatings used in real applications, the degradation mechanism is the same.

For the dynamic covariate information, Figures 3.1b, 3.1c, and 3.1d show the daily values of the UV dosage, temperature, and relative humidity (RH), respectively. The time scale for those covariates, denoted by $\tau$, is the time in days since 01 January 2002. Although the covariates were recorded at much finer resolutions, we aggregated them into daily values for convenience of modeling. Scientifically, an appropriate summary of UV exposure is the UV dosage which is proportional to the number of photons absorbed into the degrading material. The daily UV dosage at day $\tau$ is computed by $\int_{\tau}^{\tau+1} \int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} E(\zeta, \lambda) \left[1 - e^{A(\lambda)}\right] d\lambda d\zeta$ where the spectral irradiance $E(\zeta, \lambda)$ is the dose (proportional to the number of photons hitting the surface) at time $\zeta$ from sun light with wavelength $\lambda$, $[1 - e^{A(\lambda)}]$ is the absorbance rate for
different wavelengths, and $\lambda_{\text{min}} = 300$ nm and $\lambda_{\text{max}} = 532$ nm give the wavelength limits. Wave lengths above 532 nm are not harmful and wave lengths below 300 nm are generally filtered by atmospheric ozone. The UV spectrum and intensity values were recorded at 12-minute intervals but were aggregated into daily dosage values. For the period between day 598 and day 805, the covariate information is not available because there was no experimental data being collected during that period.

As it can be seen in Figure 3.1, the environmental covariates show seasonal patterns and also different degrees of variability during different time periods. For example, the UV dosage shows more variability during summer time than during winter time. Due to different starting times, each specimen has its own profile of dynamic covariate information, resulting in different rates of degradation, as can be seen from Figure 3.1a. For example, those units started in summers initially degraded much more rapidly than those started in winters.

### 3.2.2 Notation for the Data

Here we introduce some notation for the degradation data model and the dynamic covariate model. Let $X(t) = [X_1(t), \ldots, X_p(t)]'$ be the usage/environmental information at time $t$, where $p$ is the number of covariates. Let $\mathbf{X}(t) = \{X(s) : 0 \leq s \leq t\}$ be the history of the covariate process, which records the dynamic information from time 0 to time $t$.

Suppose there are $n$ units/specimens in the field. For unit $i$, denote the degradation measurements at time $t_{ij}$ by $y_i(t_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, and $n_i$ is the number of time points where degradation measurements were taken. The value of covariate $l$ for unit $i$ at the time $s$ is denoted by $x_{il}(s)$. The history of the covariate process for unit $i$ is denoted by $\mathbf{x}_i(t_{in_i}) = \{x_i(s) : 0 \leq s \leq t_{in_i}\}$ which records the dynamic information from time 0 to time $t_{in_i}$ for unit $i$. Here $x_i(s) = [x_{i1}(s), \ldots, x_{ip}(s)]'$. 
Figure 3.1: Plot of nine representative degradation paths and dynamic covariate information (the dots show the daily values, connected by line segments).
3.3 Model for a Degradation Path

3.3.1 General Path Model

Let $D(t); t > 0$ be the actual degradation path and let

$$y(t) = D(t) + \epsilon(t)$$  \hspace{1cm} (3.1)

be the degradation measurement at time $t$. The degradation model implies a degradation path $D(t)$ for each unit in the population. When the degradation level $D(t)$ reaches the failure-definition level $D_f$, a soft failure occurs and we say that the unit has failed. The first crossing time is denoted by $t_D$ and

$$t_D = \min\{t : D(t) \text{ reaches } D_f\}.$$  

The failure-time random variable $T$ is defined as the collection of the failure times $t_D$ for all of the units in the population. The cdf of $T$ is $F(t) = \Pr(T \leq t)$. The estimate of $F(t)$ is used for the reliability prediction for the population, which is obtained by using both the degradation measurements and the dynamic covariate information.

3.3.2 Modeling Degradation Path with Dynamic Covariates

Here we introduce a general additive model to incorporate dynamic covariate information into the degradation path model. In particular, the observed degradation path, conditional on the dynamic covariate information, is modeled as

$$y_i(t_{ij}) = D[t_{ij}; x_i(t_{ij})] + R(t_{ij}; w_i) + \epsilon_i(t_{ij}).$$  \hspace{1cm} (3.2)

The corresponding model for the actual degradation path is $D[t_{ij}; x_i(t_{ij})] + R(t_{ij}; w_i)$. The first component $D[t_{ij}; x_i(t_{ij})] = \beta_0 + \sum_{l=1}^{p} \int_{0}^{t_{ij}} f_l[x_i(\tau); \beta_l]d\tau$ incorporates the dynamic covariates into the degradation path through a covariate-effect function $f(\cdot)$. Here $\beta_0$ is the initial level of degradation, and $\beta_l$ denotes the parameter(s) in covariate-effect function $f_l(\cdot)$, $l = 1, \ldots, p$. 

The coefficient vector for the initial degradation and covariate effects is denoted by $\beta = (\beta_0, \beta_1', \ldots, \beta_p')'$. For covariate $l$, the function $f_l(x_il(\tau); \beta_l)$ represents the effect of $x_il(\tau)$ at time $\tau$ on the degradation process. Thus, $\int_0^t f_l(x_i(\tau); \beta_l) \, d\tau$ is the cumulative effect of $x_i$ on the degradation process up to time $t$.

The cumulative damage model is motivated by the cumulative damage model for the accelerated failure time model in Nelson (2001, Chapter 10). For certain degradation mechanisms (e.g., wearout, crack growth, and the decomposition of chemical structures), the assumption of cumulative effects is appropriate. In the motivating application of this chapter, the environmental variables cause the loss of certain chemical structures in the coating, which reduces the concentration of certain chemical compounds over the time. Thus, the assumption of the cumulative effects is appropriate for the application.

The second component $R(t; w_i)$ is a monotone function of $t$. An individual random effect $w_i$ is used to account for unit-to-unit variability caused by unobservable factors. A simple but useful form of $R(t_{ij}; w_i)$ is $R(t_{ij}; w_i) = w_{0i} + w_{1i}t_{ij}$ where $w_i = (w_{0i}, w_{1i})'$, which has nice interpretation. In particular, $w_{0i}$ and $w_{1i}$ can be interpreted as individual random effects for the initial degradation and the time trend, respectively. A linear additive term also makes parameter estimation convenient. The random effect $w_i$ is modeled by a bivariate normal distribution $N(0, \Sigma_w)$. Let $\sigma_w$ be a general notation for the unique parameters in $\Sigma_w$. The third component $\epsilon_i(t_{ij})$ in (3.2) is the noise term. In literature, for example, Meeker and Escobar (1998), the $\epsilon_i(t_{ij})$’s are often modeled to be independent and identically distributed as $N(0, \sigma_\epsilon^2)$.

### 3.3.3 Functional Forms for Covariate-Effect Function $f(\cdot)$

Two alternative approaches are available for choosing the functional form for the covariate-effect function $f(\cdot)$. The first approach is based on models motivated by physical, chemical, and engineering knowledge. For example, if there is dynamic information on temperature, the Arrhenius relationship (e.g., page 472 of Meeker and Escobar 1998) can sometimes be used to model the effect of temperature on the rate of a degradation process.
When there is not sufficient knowledge about the form of \( f(\cdot) \) from physical/engineering knowledge or when such models do not fit the data well, an alternative is to use nonparametric methods. For this approach, the function \( f(\cdot) \) is estimated as a linear combination of spline bases. Because most physical variables have a particular relationship with the degradation process (e.g., the degradation rate is increasing as the temperature is increasing), we apply shape restrictions on \( f(\cdot) \). To obtain functional forms for \( f(\cdot) \) with different shape restrictions (e.g., monotonic increasing, decreasing, or convex), we use shaped-restricted splines described in Meyer (2008).

Here we give a brief introduction to bases for shaped-restricted splines. More details can be found in, for example, Ramsay (1988) and Meyer (2008). Consider a general covariate \( z \) with \( n \) values \( \{z_1, \ldots, z_n\} \). The range of \( z \) is denoted by \([z_{\min}, z_{\max}]\). Let \( z = (z_1, \ldots, z_n)' \).

For regression splines of order \( h \), choose \( b \) locations \( d_{h+1}, \ldots, d_{h+b} \), and define knots \( z_{\min} = d_1 = \cdots = d_h < d_{h+1} < \cdots < d_{h+b} < d_{h+b+1} = \cdots = d_{2h+b} = z_{\max} \). The M-spline basis of order \( h \), denoted by \( M^{(h)}_q(z) \), is positive on \((d_q, d_{q+h})\), zero elsewhere, and has the normalization \( \int M^{(h)}_q(u)du = 1 \), for \( q = 1, \ldots, b+h \).

Note that there are \( h+b \) M-spline bases of order \( h \), which are given recursively as follows. Order 1 M-splines are the piecewise constant \( M^{(1)}_q(z) = \mathbb{1}_{(d_q \leq z < d_{q+1})}(d_{q+1} - d_q)^{-1} \) for \( q = 1, \ldots, b+1 \) where \( \mathbb{1}_{(\cdot)} \) is an indicator function. Order \( h \) M-splines are computed recursively by
\[
M^{(h)}_q(z) = \frac{h[(z - d_q)M^{(h-1)}_q(z) + (d_{q+h}-z)M^{(h-1)}_{q+1}(z)]}{(h-1)(d_{q+h}-d_q)} \mathbb{1}_{(d_q \leq z < d_{q+h})}
\]
for \( q = 1, \ldots, b+h \). The I-splines are \( \tilde{I}^{(h)}_q(z) = \int_{z_{\min}}^z M^{(h)}_q(u)du \), \( q = 1, \ldots, b+h \), for \( z \in [z_{\min}, z_{\max}] \). Note that the I-spline bases are monotone increasing functions of \( z \). The I-splines are integrated to obtain C-splines, \( \tilde{C}^{(h)}_q(z) = \int_{z_{\min}}^z \tilde{I}^{(h)}_q(u)du \), \( q = 1, \ldots, b+h \), for \( z \in [z_{\min}, z_{\max}] \).

Let \( \tilde{I}^{(h)}_q(z) = [I^{(h)}_q(z_1), \ldots, I^{(h)}_q(z_n)]' \), \( \tilde{C}^{(h)}_q(z) = [C^{(h)}_q(z_1), \ldots, C^{(h)}_q(z_n)]' \), and \( 1 \) be a vector of ones with length \( n \). To remove the dependency on the constant spline, the I-splines are
regularized by

$$I_q^{(h)}(z) = \tilde{I}_q^{(h)}(z) - P_{\{1\}][\tilde{I}_q^{(h)}(z)], \quad q = 1, \cdots, b + h$$  \hspace{1cm} (3.3)

where $P_{\{1\}][\tilde{I}_q^{(h)}(z)]$ is the projection of $\tilde{I}_q^{(h)}(z)$ onto the linear space spanned by 1. The projection of a vector $v$ onto a linear space is obtained by finding a vector $v_0$ in the linear space that minimizes the distance between $v_0$ and $v$. Similarly, to remove the dependency of the constant and identity splines, the C-splines are regularized by

$$C_q^{(h)}(z) = \tilde{C}_q^{(h)}(z) - P_{\{1,z\}][\tilde{C}_q^{(h)}(z)], \quad q = 1, \cdots, b + h$$  \hspace{1cm} (3.4)

where $P_{\{1,z\}][\tilde{C}_q^{(h)}(z)]$ is the projection of $\tilde{C}_q^{(h)}(z)$ onto the linear space spanned by 1 and $z$.

A monotone function is estimated by a linear combination of the basis functions (I-splines) and a constant function. To constrain the estimate to be monotone increasing, the coefficients of the basis functions must be nonnegative (the coefficient of the constant function is not constrained). A convex regression function is estimated using linear combinations of the basis functions (C-splines) with nonnegative coefficients, plus an unrestricted linear combination of the constant function and the identity function $g(x) = x$.

For the outdoor weathering data, the degradation path of the FTIR damage number 1250 cm$^{-1}$ is monotone decreasing. Higher UV dosage or temperature tends to cause larger damage rates. Thus the effects of UV dosage and temperature are constrained to be monotone decreasing in UV dosage and temperature. The effect of RH is constrained to be concave, based on a graphical analysis of the indoor weathering data in Gu et al. (2009) where the RH was controlled (along with other experimental variables) at different levels for serval groups of test units. Figure 3.2 shows the spline bases for the covariate effects of UV dosage, temperature, and RH. The monotone decreasing splines bases are obtained by using $-I_q^{(h)}(\cdot)$. The concave spline bases are obtained by using $-C_q^{(h)}(\cdot)$. Note that the spline bases in Figure 3.2a and 3.2b are different because their computations are based on covariates with different ranges.
3.4 Degradation Path Model Parameter Estimation

3.4.1 Parameter Estimation

Because the degradation data and dynamic covariate processes are observed at discrete points in time, the discrete-data version of model (3.2) is

\[
y_i(t_{ij}) = \beta_0 + \sum_{l=1}^{p} \sum_{\tau_{ik} \leq t_{ij}} f_l[x_l(\tau_{ik}); \beta_l](\tau_{ik} - \tau_{i,k-1}) + R(t_{ij}; w_i) + \epsilon_i(t_{ij}).
\] (3.5)

Here \(D[t_{ij}; x_i(t_{ij})] = \beta_0 + \sum_{l=1}^{p} \sum_{\tau_{ik} \leq t_{ij}} f_l[x_l(\tau_{ik}); \beta_l](\tau_{ik} - \tau_{i,k-1})\) is still used to denote the discrete-data version of the model, and \(\tau_{ik}\) are the time points where the covariate process was recorded for unit \(k\), with convention that \(\tau_{i0} = 0\). Let \(\theta_D = \{\beta, \sigma_w, \sigma_\varepsilon\}\) be the collection of unknown parameters. The maximum likelihood (ML) method is used for parameter estimation. Given the observed covariate process history, the likelihood is

\[
L(\theta_D|\text{Covariate History}) = \prod_{i=1}^{n} \int_{w_i} \left[ \prod_{t_{ij} \leq t_{im}} \frac{1}{\sigma_\varepsilon} \phi \left( \frac{B[y_i(t_{ij}); x_i(t_{ij}), w_i]}{\sigma_\varepsilon} \right) g_{w_i}(w_i; \sigma_w) \right] dw_i
\] (3.6)
where $B[y_i(t_{ij}); \mathbf{x}_i(t_{ij}), w_i] = y_i(t_{ij}) - D[t_{ij}; \mathbf{x}_i(t_{ij})] - R(t_{ij}; w_i)$, $\phi(\cdot)$ is the probability density function (pdf) of a $N(0, 1)$ distribution, and $g_{w_i}(\cdot)$ is the pdf of a $N(0, \Sigma_w)$ distribution. The ML estimate $\hat{\theta}_D$ is obtained by finding the value of $\theta_D$ that maximizes (3.6). Note that we are conditioning on the entire covariate history when we model the degradation process.

The maximization of (3.6), in general, is non-trivial because numerical methods such as quadrature (e.g., Liu and Pierce 1994) are needed to evaluate the integral in the likelihood function. When shape-restricted splines are used and the random component is modeled as a linear function of $w_i$, the model in (3.5) is a linear mixed-effects model with constraints on the parameters. The estimation of the unknown parameters can be done by using the procedure in Davidov and Rosen (2011). For computational efficiency, the constrained quadratic programming used in Davidov and Rosen (2011) can be replaced by the mixed primal-dual bases algorithm used in Fraser and Massam (1989) to solve the generalized least squares problem under constraints.

When shape-restricted splines are used, that is $f_l[x_{il}(\tau_{ik}); \beta_l] = \sum_{q=1}^{Q_l} B_{lq}[x_{il}(\tau_{ik})] \beta_{lq}$, and the random component is specified to be $R(t_{ij}; w_i) = w_{0i} + w_{1i} t_{ij}$, the model in (3.5) can be represented by

$$y_i(t_{ij}) = \beta_0 + \sum_{l=1}^{p} \sum_{\tau_{ik} \leq t_{ij}} \sum_{q=1}^{Q_l} B_{lq}[x_{il}(\tau_{ik})] \beta_{lq}(\tau_{ik} - \tau_{i,k-1}) + w_{0i} + w_{1i} t_{ij} + \varepsilon_i(t_{ij})$$

$$= \beta_0 + \sum_{l=1}^{p} \sum_{q=1}^{Q_l} G_{lq}(t_{ij}) \beta_{lq} + w_{0i} + w_{1i} t_{ij} + \varepsilon_i(t_{ij}). \quad (3.7)$$

Here $B_{lq}(\cdot)$’s are spline bases and $\beta_{lq}$’s are spline coefficients, $G_{lq}(t_{ij}) = \sum_{\tau_{ik} \leq t_{ij}} B_{lq}[x_{il}(\tau_{ik})](\tau_{ik} - \tau_{i,k-1})$, and $Q_l$ is the number of spline bases for covariate $l$. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$,

$$\mathbf{X}_i = \begin{bmatrix} 1 & G_{11}(t_{i1}) & \cdots & G_{1Q_1}(t_{i1}) & \cdots & G_{1p1}(t_{i1}) & \cdots & G_{1pQ_p}(t_{i1}) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & G_{11}(t_{in_i}) & \cdots & G_{1Q_1}(t_{in_i}) & \cdots & G_{1p1}(t_{in_i}) & \cdots & G_{1pQ_p}(t_{in_i}) \end{bmatrix}, \quad \mathbf{Z}_i = \begin{bmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix}.$$
and $\varepsilon_i = [\varepsilon_i(t_{i1}), \ldots, \varepsilon_i(t_{im_i})]'$. Using this notation, the model in (3.7) can be expressed as

$$y_i = X_i\beta + Z_iw_i + \varepsilon_i.$$  

Note that the variance-covariance matrix of $y_i$ is $\Sigma_i = Z_i\Sigma_w Z_i' + \sigma^2_i I_{n_i}$ where

$$\Sigma_w = \begin{bmatrix} \sigma_0^2 & \rho \sigma_0 \sigma_1 \\ \rho \sigma_0 \sigma_1 & \sigma_1^2 \end{bmatrix}$$

and $I_{n_i}$ is an $n_i \times n_i$ matrix. Let $\sigma_w = (\sigma_0, \sigma_1, \rho)'$. Some components of $\beta$ are constrained to be nonnegative to produce a shape-restricted covariate-effect function. Without loss of generality, let $\beta = (\beta_u, \beta_c)'$ where $\beta_u$ and $\beta_c$ represent unconstrained and constrained parameters, respectively.

The estimation algorithm is as follows.

**Algorithm 1:**

1. Obtain initial values of $\sigma_w$ and $\sigma_\varepsilon$, which can be done by fitting the unconstrained linear mixed-effects model.

2. Compute $\Sigma_i = Z_i\Sigma_w Z_i' + \sigma^2_i I_i$.

3. With $\Sigma_i$ computed in step 2, use the mixed primal-dual bases algorithm to obtain the estimate of $\beta$ by minimizing $\sum_{i=1}^n (y_i - X_i\beta)' \Sigma_i^{-1} (y_i - X_i\beta)$ subject to the constraints that the elements of $\beta_c$ are greater than or equal to 0.

4. Fit a linear mixed-effects model $\hat{\varepsilon}_i = Z_iw_i + \varepsilon_i$ with $\hat{\varepsilon}_i = y_i - X_i\hat{\beta}$ to obtain updated estimates of $\sigma_w$ and $\sigma_\varepsilon$.

5. Repeat steps 2 to 4 until convergence.

Making inferences based on the constrained ML estimator is not straightforward. Some elements of the ML estimate vector may be on the boundary of the parameter space. Although
asymptotic theory is available for constrained ML estimators (e.g., Self and Liang 1987), the bootstrap method provides a flexible and easy-to-implement alternative. The bootstrap inference procedure is also somewhat robust to model departures. After being adjusted by a method that is similar to Carpenter et al. (2003), the residuals and estimated random effects are sampled with replacement to construct a bootstrap version of the data. As pointed out by Morris (2002), direct resampling (i.e., without appropriate adjustment) of residuals and estimated random effects will result in confidence intervals (CIs) that are too narrow. The estimation procedure in Algorithm 1 is then applied to the bootstrapped data to obtain bootstrapped values of parameter estimates. The resampling process is repeated a large number of the times and bias-corrected CIs are constructed based on the bootstrapped parameter estimates (Efron and Tibshirani 1993). The details of the bootstrap algorithm and confidence interval constructions are described in Appendix 3.A. The performance of Algorithm 1 and the bootstrap CI procedure will be investigated through simulations in Section 3.5.

3.4.2 Estimation for Weathering Data

For the outdoor weathering application, the spline bases shown in Figure 3.2 were used to estimate the effect of UV dosage, temperature, and RH. The parameter estimates are obtained by using Algorithm 1. The estimates and CIs for $\beta_0, \sigma_0, \sigma_1, \rho,$ and $\sigma_\varepsilon$ are shown in Table 3.1. Figure 3.3 shows the estimated effect functions for UV dosage, temperature and RH, and the corresponding approximate 95% pointwise CIs. Figure 3.3a shows that larger UV dosages lead to more damage. The UV dosage has a large effect of the damage rate, relative to temperature and RH. Figure 3.4 shows the plot of degradation measurements and fitted degradation path for the nine representative specimens shown in Figure 3.1. The figure shows that the general path model fits the degradation data well. We also checked the Q-Q plot of the residuals (see Figure 3.12 in the appendix) and the plot indicates that the normal distribution assumption holds well.

For the selection of spline orders/degrees, a spline order/degree of two or three is generally sufficient to ensure enough smoothness. We used the AIC criterion (e.g., Eilers and Marx 1996)
Table 3.1: Parameter estimates and approximate 95% CIs for $\beta_0$, $\sigma_0$, $\sigma_1$, $\rho$ and $\sigma_\varepsilon$ in the degradation path model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Bootstrap CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>-0.04164</td>
<td>0.00386</td>
<td>-0.04946 -0.03429</td>
</tr>
<tr>
<td>$\sigma_0$</td>
<td>0.02160</td>
<td>0.00318</td>
<td>0.01540 0.02775</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>0.00067</td>
<td>0.00010</td>
<td>0.00049 0.00087</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-0.47793</td>
<td>0.15143</td>
<td>-0.68934 -0.06847</td>
</tr>
<tr>
<td>$\sigma_\varepsilon$</td>
<td>0.01769</td>
<td>0.00048</td>
<td>0.01688 0.01880</td>
</tr>
</tbody>
</table>

To select the spline orders and knots, we chose the knot locations based on equal sample quantiles, which places the knots in places where there is a sufficient amount of covariate information. For example, for four knots ($b = 4$), the locations were chosen as the 0.2, 0.4, 0.6, and 0.8 sample quantiles of the covariate values. For splines with different orders, we plotted the AIC values as a function of the number of spline knots (see Figure 3.13 in the appendix). Based on the results, we choose order three splines with four knots (i.e., $h = 3$ and $b = 4$) for the weathering data. An alternative approach is to use the asymptotically optimal number for $b$, in which $b \approx c^{1/(2h+1)}$ (e.g., Meyer 2008). Here the sample size $c$ is taken to be the number of time points where the covariates are recorded. For the weathering data, there are 676 time points. With order 3 polynomial pieces (i.e., $h = 3$), the number of knots is $b = 3$ for the covariates in the weather data, which is very close to the selection by the AIC criterion.

We also did a sensitivity analysis on spline orders and the number of interior knots (see Figure 3.14 in the appendix). The figure show that the estimated covariate-effect functions are insensitive to different choices of spline orders and the number of knots. This observation is consistent with the comments in Meyer (2008), which points out the estimated functions are relatively insensitive the number of knots, largely due to the shape restrictions.
Figure 3.3: Estimated covariate-effect functions for UV dosage, temperature, and RH, and the corresponding approximate 95% pointwise CIs.

Figure 3.4: Plot of degradation measurements and fitted degradation path for the nine representative specimens shown in Figure 3.1.
3.5 Simulation Study

The estimation procedure in Section 3.4.1 involves random effects and point estimates that may be on a constraint boundary. In this section, we conduct a simulation to study the performance of the estimation procedure. Specifically, we investigate the mean square error (MSE) of point estimators and coverage probability (CP) of the bias-corrected bootstrap CI procedure.

3.5.1 Simulation Model and Setup

To simplify the setting while still obtaining the insights, we consider UV as the only covariate in the model and all units are exposed to the same UV profile for 200 days. We use the first 200 day UV information from the weathering data as the UV exposure profile for the units in the simulation. The main reason for simplifying the setting is the computing time. Although Algorithm 1 is reasonably fast, the evaluation of the CP of the bootstrap CI procedure is computationally demanding. We use the following model as the true model in the simulation,

\[ y_i(t_{ij}) = \beta_0 + \int_0^{t_{ij}} f[x_i(\tau); \beta] d\tau + w_{1i} t_{ij} + \varepsilon_i(t_{ij}), \quad (3.8) \]

where \( f[x_i(\tau_{ik}); \beta] = \sum_{q=1}^5 B_q[x_i(\tau_{ik})] \beta_q \) is the UV effect function, \( B_q[x_i(\tau_{ik})]'s \) are the spline bases, \( w_{1i} \sim N(0, \sigma_1^2) \), and \( \varepsilon_i(t_{ij}) \sim N(0, \sigma_\varepsilon^2) \). For computing time consideration, we use order two splines with two interior knots and also use a simplified random component. The true values of the parameters are set to \( \beta_0 = -0.03, \beta = (0.005, 0.003, 0, 0.006, 0)' \), \( \sigma_1 = 0.001 \), and \( \sigma_\varepsilon = 0.02 \) to mimic the setting of the weather data. Figure 3.5 shows the spline bases (the constant basis is not shown) and the UV effect function used in simulation. The UV effect function is obtained as a linear combination of \( B_q[x_i(\tau_{ik})]'s \) with coefficient \( \beta \). To create a non-standard situation, two elements of \( \beta \) are set to zero, which are on the boundary of the constrained parameter space.

In the simulation study, we consider six scenarios. The number of experimental units
are chosen to be \( n = 20, 50, \) and 100. For each sample size, the number of degradation measurement points are chosen to be \( m = 25 \) and 50. For each scenario, a dataset is simulated and **Algorithms 1** and **4** (in Appendix 3.A) are used to obtain the point estimates and CIs, respectively. A repeat of 600 datasets is used to obtain the estimates of the MSE of point estimators and CP of CI procedures.

### 3.5.2 Simulation Results

Figure 3.6 shows the estimated MSE of the parameter estimators and CP of CI procedure based on 600 repeats for \( \beta_0, \sigma_1, \) and \( \sigma_\varepsilon \). Figure 3.7 shows the estimated MSE for the UV effect function estimator and CP for pointwise CI for UV effect function based on 600 repeats for the six scenarios. Based on the simulation results, we find that the MSE of the point estimators of the parameters and the UV effect function generally decrease as \( n \) and \( m \) increase. The CP of the CI procedures converge to the nominal 95% level as \( n \) and \( m \) increase. The CP of CI procedure for \( \sigma_1 \) is poor when \( n = 20 \) but it improves when \( n \) increases. The CP of the pointwise CI procedure for the UV effect function is around 90% when \( n = 20 \) and it is around 95% when \( n = 100 \). Overall, the simulation study shows that the performance of estimation and bootstrap CI procedure are good.
Figure 3.6: Estimated MSE of the parameter estimators and CP of CI procedure based on 600 repeats for $\beta_0$, $\sigma_1$, and $\sigma_\varepsilon$.

Figure 3.7: Estimated MSE for UV effect function estimator (left) and CP for pointwise CI for UV effect function (right) based on 600 repeats.
3.6 Model for Multivariate Covariate Process

3.6.1 General Strategy for Covariates Modeling

In order to predict a degradation path into the future, it is necessary to have a parametric model that can adequately predict the covariate process. In general, the following parametric structure for \( X(t) \) can be used for each individual unit,

\[
X(t) = m(t; \eta) + a(t)
\]

where \( m(t; \eta) \) is the mean function with parameter \( \eta \) and some components of \( \eta \) can be random to allow for population unit-to-unit (or time period-to-time period in our application) variability for the covariate process. Depending on the application, the parametric form for \( m(t; \eta) \) can be specified. For example, the environmental temperature of an individual unit can be modeled as

\[
X(t) = \text{Trend}(t) + \text{Seasonal}(t) + a(t)
\]

where \( \text{Trend}(t) \) is the long term trend and \( \text{Seasonal}(t) \) is a seasonal periodic term. The error term \( a(t) \) is assumed to be a stationary process. In some applications, \( a(t) \) for different values of \( t \) can be modeled as independently and identically distributed with \( \mathcal{N}(0, \Sigma_a) \) where \( \Sigma_a \) is the covariance matrix. The vector autoregressive (VAR) time series models in Reinsel (2003) can be used if more complicated structures are needed for modeling \( a(t) \).

3.6.2 Parametric Models for Covariates for Outdoor Weathering Data

For each application, a special modeling effort will be needed to capture the unique features in the covariate process. Here we present the modeling of the environmental variables in the outdoor weathering data. Let \( x_1(\tau), x_2(\tau) \) and \( x_3(\tau) \) be the values of UV dosage, temperature, and RH at time \( \tau \), respectively. For these three variables, there is no significant time trend but the seasonal effect is evident. For time series with a seasonal component, combinations of sine
and cosine functions are commonly used to capture the seasonal component (e.g., Campbell and Diebold 2005).

Based on initial analysis of the covariates in the weathering data, a single sine function is adequate to describe the mean structure of \(x_1(\tau), x_2(\tau)\) and \(x_3(\tau)\). There is also a seasonal pattern in the process variance. For example, there is more variability in UV dosage during the summer months than during the winter months. Thus a seasonal component is also added to the variance structure. In particular, the multivariate time series is modeled by

\[
\begin{bmatrix}
  x_1(\tau) \\
  x_2(\tau) \\
  x_3(\tau)
\end{bmatrix}
= \begin{bmatrix}
  \mu_1 + \kappa_1 \sin \left[ \frac{2\pi}{365}(\tau - \eta_1) \right] \\
  \mu_2 + \kappa_2 \sin \left[ \frac{2\pi}{365}(\tau - \eta_2) \right] \\
  \mu_3 + \kappa_3 \sin \left[ \frac{2\pi}{365}(\tau - \eta_3) \right]
\end{bmatrix}
+ \begin{bmatrix}
  (1 + \nu_1 \{1 + \sin \left[ \frac{2\pi}{365}(\tau - \varsigma_1) \right]\}) \varepsilon_1(\tau) \\
  (1 + \nu_2 \{1 + \sin \left[ \frac{2\pi}{365}(\tau - \varsigma_2) \right]\}) \varepsilon_2(\tau) \\
  \varepsilon_3(\tau)
\end{bmatrix}. \tag{3.9}
\]

The \(\sin(\cdot)\) function with a period of 365 days is used to capture the seasonal pattern in the covariates. For the UV dosage and temperature, extra terms are used to capture the nonhomogeneity of variance over time. A likelihood ratio test suggested that the seasonal pattern is not important in the RH variance component. Thus a constant variance component is used for RH.

To further capture the autocorrelation within each covariate and the correlation among different covariates, a VAR model with lag two [i.e., VAR(2)] is used. In particular, the error term is modeled by

\[
\begin{bmatrix}
  \varepsilon_1(\tau) \\
  \varepsilon_2(\tau) \\
  \varepsilon_3(\tau)
\end{bmatrix}
= \Phi_1 \begin{bmatrix}
  \varepsilon_1(\tau - 1) \\
  \varepsilon_2(\tau - 1) \\
  \varepsilon_3(\tau - 1)
\end{bmatrix}
+ \Phi_2 \begin{bmatrix}
  \varepsilon_1(\tau - 2) \\
  \varepsilon_2(\tau - 2) \\
  \varepsilon_3(\tau - 2)
\end{bmatrix}
+ \begin{bmatrix}
  \varepsilon_1(\tau) \\
  \varepsilon_2(\tau) \\
  \varepsilon_3(\tau)
\end{bmatrix}. \tag{3.10}
\]

where \(\Phi_1\) and \(\Phi_2\) are matrices of regression coefficients, and \([\varepsilon_1(\tau), \varepsilon_2(\tau), \varepsilon_3(\tau)]' \sim N(0, \Sigma_e)\)
are multivariate normal error terms that are independent over time. Here $\Sigma_e$ is the covariance matrix for the error terms. For the weathering example, the model fitting suggested that this VAR(2) model is adequate. For the weathering data, all covariate information was collected at the same place. When the units are exposed at different locations, additional parameters may be needed to describe the covariate processes.

### 3.6.3 Parameter Estimation

The estimation of the parameters in models (3.9) and (3.10) is done in two steps. In the first step, ML estimation is used to remove the seasonal trends in the mean and variance structures. Then the residuals are obtained. For the first stage, the working variance-covariance structure of $[\varepsilon_1(\tau), \varepsilon_2(\tau), \varepsilon_3(\tau)]'$ is taken to be $\sigma^2_e I_3$. The main goal of model (3.9) is to remove the mean structure and the simplified variance structure still provides consistent estimator for the mean structure. The ML method is used to obtain the parameter estimates for the model in (3.9). One needs to program the likelihood function and then use an optimization algorithm (e.g., `optim()` function in R 2013) to maximize it. The convergence of the optimization algorithm is checked by trying different starting values and carefully examining of the contour plots of the loglikelihood functions. The estimates of the parameters and corresponding bootstrap CIs for the covariate process model in (3.9) are listed in Table 3.2. We used a simple percentile bootstrap approach to obtain CIs. Figure 3.8 shows the fitted mean structure, estimated error terms, and the estimated standard deviation (SD) of the error term for UV dosage, temperature, and RH. The figure shows that model (3.9) adequately fits the mean and variance structure for UV dosage, temperature and RH data.

In the second step, ML estimation is used to fit the VAR model to the residuals. The computing of the parameter estimates uses multivariate least squares (e.g., Lütkepohl 2005, Chapter 3), which is computationally efficient. The estimates of $\Phi_1$, $\Phi_2$ and $\Sigma_e$ are as follows (the subscripts are the corresponding standard errors),
\[
\Phi_1 = \begin{pmatrix}
0.582_{0.041} & 0.020_{0.034} & 0.020_{0.011} \\
0.095_{0.061} & 0.634_{0.051} & 0.018_{0.016} \\
-0.070_{0.197} & -0.046_{0.166} & 0.594_{0.054}
\end{pmatrix},
\]

\[
\Phi_2 = \begin{pmatrix}
-0.109_{0.040} & -0.019_{0.034} & -0.013_{0.011} \\
-0.106_{0.061} & 0.030_{0.051} & 0.015_{0.017} \\
0.388_{0.198} & -0.108_{0.166} & -0.112_{0.054}
\end{pmatrix},
\]

and

\[
\Sigma_e = \begin{pmatrix}
8.870_{1.342} & 4.081_{0.725} & -20.073_{2.374} \\
4.081_{0.725} & 19.178_{2.663} & -43.636_{4.320} \\
-20.073_{2.374} & -43.636_{4.320} & 200.960_{13.834}
\end{pmatrix}.
\]

The standard errors of the parameter estimates of \(\Phi_1, \Phi_2\) and \(\Sigma\) are also obtained by using the bootstrap method. The bootstrap is carried out by sampling the estimated error term \([e_1(\tau), e_2(\tau), e_3(\tau)]'\) with replacement and then using the parametric models in (3.9) and (3.10) to obtain a bootstrap version of the covariate data. The bootstrap version of the parameter estimates is obtained by using a two-step approach. The above process is repeated a large number of times (e.g., 10,000) to obtain the bootstrap distribution of parameter estimators.

We also examined the autocorrelation function (ACF) of the estimated \([e_1(\tau), e_2(\tau), e_3(\tau)]'\) to check the assumption of the VAR model (see Figure 3.15 in the appendix). The ACF functions of the time series residuals showed no evidence of autocorrelation. We also used Q-Q plot to check the normal assumption on \([e_1(\tau), e_2(\tau), e_3(\tau)]\) (Figure 3.16 in the appendix). The plots show that the VAR(2) model provides an adequate description of the residuals.
Figure 3.8: The fitted mean and variance structures for UV dosage, temperature and RH.
Table 3.2: Parameter estimates and corresponding 95% bootstrap CIs for the parameters of the covariate process model in (3.9).

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Bootstrap CI Lower</th>
<th>95% Bootstrap CI Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>UV Dosage</td>
<td>$\mu_1$</td>
<td>24.71</td>
<td>0.58</td>
<td>23.55</td>
<td>25.83</td>
</tr>
<tr>
<td></td>
<td>$\kappa_1$</td>
<td>18.95</td>
<td>0.71</td>
<td>17.54</td>
<td>20.32</td>
</tr>
<tr>
<td></td>
<td>$\eta_1$</td>
<td>79.24</td>
<td>2.09</td>
<td>75.25</td>
<td>83.40</td>
</tr>
<tr>
<td></td>
<td>$\varsigma_1$</td>
<td>77.69</td>
<td>3.67</td>
<td>70.44</td>
<td>84.86</td>
</tr>
<tr>
<td></td>
<td>$\nu_1$</td>
<td>1.80</td>
<td>0.24</td>
<td>1.37</td>
<td>2.29</td>
</tr>
<tr>
<td>Temperature</td>
<td>$\mu_2$</td>
<td>25.05</td>
<td>0.52</td>
<td>24.04</td>
<td>26.05</td>
</tr>
<tr>
<td></td>
<td>$\kappa_2$</td>
<td>16.54</td>
<td>0.69</td>
<td>15.17</td>
<td>17.91</td>
</tr>
<tr>
<td></td>
<td>$\eta_2$</td>
<td>103.19</td>
<td>2.65</td>
<td>97.92</td>
<td>108.27</td>
</tr>
<tr>
<td></td>
<td>$\varsigma_1$</td>
<td>33.53</td>
<td>13.73</td>
<td>6.96</td>
<td>60.66</td>
</tr>
<tr>
<td></td>
<td>$\nu_2$</td>
<td>0.31</td>
<td>0.09</td>
<td>0.14</td>
<td>0.49</td>
</tr>
<tr>
<td>RH</td>
<td>$\mu_3$</td>
<td>40.01</td>
<td>1.02</td>
<td>38.06</td>
<td>42.06</td>
</tr>
<tr>
<td></td>
<td>$\kappa_3$</td>
<td>-4.73</td>
<td>1.52</td>
<td>-7.38</td>
<td>-0.54</td>
</tr>
<tr>
<td></td>
<td>$\eta_3$</td>
<td>39.00</td>
<td>18.8</td>
<td>6.71</td>
<td>80.23</td>
</tr>
</tbody>
</table>

3.7 Failure-time Distribution Estimation

3.7.1 Failure-time Distribution for the Population

The failure-time distribution provides the reliability information for an unobserved population. We use $\theta_D$ and $\theta_X$ to denote the unknown parameters in the degradation model and covariate process model, respectively. Let $\theta = \{\theta_D, \theta_X\}$. The model for the actual path is $D[t; X(\infty)] + R(t; w)$. Given the covariate process $X(\infty) = \mathbf{x}(\infty)$ and the individual random effect $w$, the degradation path is deterministic. The first crossing (failure) time $t_D$ for a particular unit can be obtained. That is

$$t_D = \min\{t : D[t; \mathbf{x}(\infty)] + R(t; w) = D_f \}. \quad (3.11)$$

Thus the first crossing time $t_D$ is a function of $D_f$, $\mathbf{x}(\infty)$, and $w$. Numerical methods are often needed to solve for $t_D$ from (3.11). Because $X(\infty)$ and $w$ are random, the first crossing time (i.e., the failure-time of the product), denoted by $T$, is a random variable. The cdf of
\[ T = T[D_f, X(\infty), w] \] is

\[ F(t; \theta) = E_{X(\infty)} E_w \Pr \{ T[D_f, X(\infty), w] \leq t \}, \quad t > 0. \] (3.12)

In most situations, there is no explicit form for \( F(t; \theta) \) and it has to be computed by using numerical methods or Monte Carlo simulation.

Substituting \( \hat{\theta} \) into \( F(t; \theta) \) in (3.12), one obtains an estimate of the cdf. Because an explicit form for \( F(t; \hat{\theta}) \) is, in general, not available, a simulation approach is used to evaluate \( F(t; \hat{\theta}) \).

The following algorithm is used for computing \( F(t; \hat{\theta}) \).

**Algorithm 2:**

1. Simulate the covariate process with the parameter equal to \( \hat{\theta}_X \).

2. Simulate the random effect \( w \) from \( N(0, \Sigma_w) \) with the parameter equal to \( \hat{\theta}_D \).

3. Compute the simulated degradation path \( D[t; X(\infty)] + R(t; w) \) with the simulated covariate process and random effect.

4. Given the simulated degradation path, compute the failure-time \( t_D \) by solving (3.11).

5. Repeat the above steps 1 to 4 \( B \) times (e.g., \( B = 10,000 \)) to obtain the simulated failure-times \( t_D^b, b = 1, \ldots, B \) where \( B \) is chosen large enough to provide sufficient precision.

The estimate of \( F(t; \theta) \) is obtained by \( F(t; \hat{\theta}) = B^{-1} \sum_{b=1}^R 1_{(t_D^b \leq t)} \).

The CIs for the cdf can be obtained as follows. Because the bootstrap parameters \( \hat{\theta} \) are obtained in previous sections, one needs to repeat Algorithm 2 for each set of bootstrap version of parameter estimates. The pointwise CIs for the cdf are then obtained as the sample quantiles of the bootstrap estimates of \( F(t; \theta) \). If the focus is on an individual unit, Algorithm 2 can also be used to obtain the estimates for the cdf of an individual unit.

The manner in which our model and inference procedures would be used in applications will depend on relationship between the units of interest and the processes generating the
covariates that will affect the degradation processes. Some example scenarios include the following:

1. In some applications all units in the population will be subject to different realizations of a covariate processes that could be adequately modeled as independent (but not identically distributed) from unit to unit (location to location). For example, when considering the loss of light output from LED lights in households, the usage history (the covariate) are different but can be considered to be independent from household to household.

2. Similar to the groups of units in the weathering experiment, one might be interested in estimating the failure time distribution of a population of units that are all subject to the same realization from the covariate processes. For example, when considering the power output decrease of solar panels installed at one location (e.g., one power plant), the environmental variables such as temperature and humidity can be considered to be the same for each unit in the field.

3. Similar to the overall weathering experiment, one might be interested in groups of units put into service at various points in time (known as staggered entry) so that the groups are subject to different parts of the same covariate process. The illustrative outdoor weathering example in this chapter falls into this category.

### 3.7.2 Application to the Epoxy Degradation Example

To illustrate the use of our methods, we use the outdoor weathering setting and assume that there is a hypothetical population with infinite size and that units randomly enter service, according to a uniform distribution, between day 161 and day 190. Each unit has its own independent realization of the covariate processes, from the observed processes in the experiment. Figure 3.9 shows the estimated cdf and the corresponding 95% pointwise CIs for this hypothetical population. Most of the units in the population fail between 50 days to 150 days
after they are put into service. Similar results can be obtained for the cdf of an individual unit (e.g., a unit started at day 161).

For the NIST outdoor weathering data, we also checked how well the failure-time model fits the observed failure times. For the weathering data, we use a failure threshold $D_f = -0.4$. Generally, this would be chosen to be the level of degradation at which the performance of the coating would not be acceptable (e.g., the level at which there would be customer perceivable loss of gloss or color). There were 36 units that were put into experiments at different times from 2002 to 2006 (i.e., in a staggered entry pattern). According to this failure definition, there were 17 failures and the other 19 units had survived.

Figure 3.10 shows the estimated expected number of failures and corresponding 95% point-wise CIs versus the observed number of failures as a function of time for the 36 specimens in the outdoor weathering data. The dots show the observed cumulative number of failures as a function of time. The estimated expected number of events is computed based on the estimated degradation path and covariate process models. For the periods from day 0 to day 597 and from day 806 to day 1153, the covariate processes for the weathering data had already been observed. We treat the covariates as fixed when we compute the estimated expected number of failures. For those periods that are between day 598 and day 805 and after day 1153, as shown in Figure 3.1, the covariate information is missing. The covariate process for those periods, however, are needed to compute the estimated expected number of events. We used multiple realizations of the covariate processes simulated from the fitted model and the results were averaged for those periods. The results in Figure 3.10 show that the estimated expected number of failures agree with the observed number of failures well except that there is an abrupt jump around day 50 in the observed number of failures.

3.7.3 Distribution of Remaining Life for Individual Units

Given the observed degradation path and the covariate process up to time $t_{in}$ for individual $i$, the distribution of the remaining failure-time is needed in some applications. In particular, the conditional distribution of remaining life for individual with $X_i(t_{in}) = x_i(t_{in})$ and $T > t_{in}$,
Figure 3.9: The estimated cdf and corresponding 95% pointwise CIs for a population with units starting randomly from day 161 to day 190.

Figure 3.10: Estimated expected number of failures and corresponding 95% pointwise CIs versus the observed number of failures as a function of time for the 36 specimens in the outdoor weathering data.
\[ \rho_i(s; \theta) = \mathbb{E}_{X_i(\infty)|X_i(t_{ini})=x_i(t_{ini})} \mathbb{E}_w \Pr \{ T[\mathcal{D}_f, X(\infty), w] \leq t_{ini} + s | T > t_{ini} \}, s > 0. \]

Here \( \rho_i(s; \theta) \) gives, conditional on \( X_i(t_{ini}) = x_i(t_{ini}) \) and \( T > t_{ini} \), the probability of failure before time \( s \). Similar algorithms can be used to evaluate the conditional distribution \( \rho_i(s; \bar{\theta}) \) for individual \( i \) and the corresponding pointwise CIs. The difference is that the future degradation path and the covariate process are conditional on \( x_i(t_{ini}) \) and the degradation measurements that have been observed up to time \( t_{ini} \).

For illustration, we consider the specimen labeled “G18-10” with age of 158 days at the end of exposure. The observed degradation level for G18-10 had not reached \( \mathcal{D}_f = -0.4 \) after 158 days. Conditional on the observed degradation path and the covariate history for this unit, we can compute the estimated cdf of remaining life for this unit. Figure 3.11 shows the estimated conditional cdf for unit G18-10 and the corresponding 95% pointwise CIs. The results in Figure 3.11 show that the remaining life of this unit is roughly with the range of 20 days to 50 days.

### 3.8 Concluding Remarks and Areas for Future Research

Motivated by the increasing availability of dynamic covariate information being acquired by systems operating in the field and the needs to predict future performance of these systems, we develop a class of models and methods for using such data. We illustrate these methods with the NIST outdoor weathering data. We use flexible general path models with individual random components to describe unit-to-unit variability in the degradation data. Parametric models are used to model the covariate processes. We develop algorithms to compute an estimate for the failure-time distribution induced by the underlying degradation model. We use the outdoor weathering data to illustrate the modeling process and the estimation of the failure-time distribution functions.

Although the NIST outdoor weathering data was our motivating example, the methods
Figure 3.11: The estimated conditional cdf of remaining life and corresponding 95% pointwise CIs for unit G18-10.

developed in this chapter can have broad applications for many products used in highly variable environments and/or subject to time-varying usages. For example, the degradation of LED power output is mainly due to usage that can be time varying. The corrosion of crude oil transportation pipeline is subject to the outdoor environments and the characteristics of the compounds flowing in the pipeline. Damage done to structures in aircraft will depend on the number of takeoff-landing cycles and other stresses encountered during operation. Also, the degradation of photovoltaic arrays can be caused by both the time-varying usage and the outdoor environments. Thus there are tremendous opportunities to apply the method developed in the chapter.

For weathering applications, in the past decades, many solar UV monitoring sites have been established at different geographical locations within the United States and worldwide (e.g., as described in Kaetzel 2001). The solar UV spectrum and intensity, as well as temperature and relative humidity are recorded at high resolution. Such information can be used for prediction for the lifetime of products that are subject to degradation. When the covariate information is from different locations, spatial correlations may need to be considered for the model to
predict the covariates for a population. Spatial data modeling techniques can be applied. In other applications, where the product is not exposed to sunlight or other weather variables, our models can still be used to model degradation as a function of other variables like load or amount of use.

The covariate-effect function based on splines are only defined over the range of the data. As with all other nonparametric methods, extrapolation will be challenging. In the weathering data, there is enough historical data to cover the needed range of covariates. The shape-restricted spline, however, allows one to do extrapolation to some extent. Suppose, for example, that the possible range of a covariate is \([z_1, z_2]\) but the data only cover a range \([z_1 + \delta, z_2 - \delta]\) for a small positive \(\delta\). If the covariate-effect function is constrained to be monotone increasing, one can construct spline basis with \(z_1\) and \(z_2\) as boundary knots with other interior knots placed between \([z_1 + \delta, z_2 - \delta]\). Then estimate of the covariate-effect function can still be obtained over the range \([z_1, z_2]\). In this case, \(\delta\) should be relatively small compared to the entire data range. Extrapolation based on monotone splines would be an interesting topic to investigate in future research.

The additive model for degradation paths proposed in this chapter is equivalent to a linear degradation path when the covariates are time invariant. In the future, it will be useful to consider a degradation path \(D(t) = g\{D[t; x(t)] + R(t; w)\}\) where \(g(\cdot)\) is a nonlinear function that depends on some unknown parameters. The estimation for such a model will, however, be challenging. We did not consider interactions between covariates in the model and it would be interesting to consider interaction effects in future research. It is, however, practically challenging to impose shape restriction so that the covariate-effect functions have physical meaning. Also, there needs to be enough data points to estimate the interaction effects of covariates.

There is a significant amount of research that has been done in the area of functional regression (e.g., Yao et al. 2005, Müller and Zhang 2005, and Liu and Müller 2008). There has been little work to consider shape restrictions on covariate-effect functions in the degradation setting. We believe shape restrictions are useful when we already have physical knowl-
edge/preliminary information on the shape of a covariate-effect function. It would be interesting to investigate the application of function linear regression methods to degradation data with time-varying covariates which also incorporate physical and engineering knowledge.

Appendix

3.A Nonparametric Residual Bootstrap with Adjustment

In this appendix, we describe the nonparametric residual bootstrap method for the linear mixed-effects model used in Section 3.4.1. The method is similar to Carpenter et al. (2003) but we customize it for the model in this chapter. Resampling the estimated random effects and residuals without appropriate adjustments will result in bias (Carpenter et al. 2003). The steps of the bootstrap algorithm are described as follows.

Algorithm 3:

1. Compute the parametric estimates for model (3.7) using Algorithm 1 and then obtain the residuals $\tilde{\epsilon}_i = [\hat{\epsilon}_i(t_{i1}), \ldots, \hat{\epsilon}_i(t_{im})]'$ and estimated random effects $\tilde{w}_i = (\tilde{w}_{0i}, \tilde{w}_{1i})'$, $i = 1, \ldots, n$.

2. Adjust the estimated random effects by a linear transformation. The idea is to match the estimated approximate variance-covariance matrix for $\tilde{w}_i$ to the estimated variance-covariance of $w_i$ (i.e., $\tilde{\Sigma}_w$). The estimated variance-covariance matrix of $\tilde{w}_i$ can be approximated by $A = \tilde{\Sigma}_w Z_i (\tilde{\Sigma}_i)^{-1} Z_i \tilde{\Sigma}_w$. The adjusted estimated random effect is obtained by $\tilde{\Sigma}_w^{1/2} A^{-1/2} \tilde{w}_i$, $i = 1, \ldots, n$, where $\tilde{\Sigma}_w^{1/2}$ is the square root matrix of $\tilde{\Sigma}$ and $A^{-1/2}$ is the inverse of the square root matrix of $A$.

3. Adjust the residuals by a linear transformation. The idea is to match the estimated approximate variance-covariance matrix for $\tilde{\epsilon}_i$ to the estimated variance-covariance of $\epsilon_i$ (i.e., $\hat{\sigma}^2_\epsilon I_{n_i}$). The estimated variance-covariance matrix of $\tilde{\epsilon}_i$ can be approximated by $B = (I_{n_i} - Z_i \tilde{\Sigma}_w Z_i (\tilde{\Sigma}_i)^{-1}) \tilde{\Sigma}_i (I_{n_i} - Z_i \tilde{\Sigma}_w Z_i (\tilde{\Sigma}_i)^{-1})'$. The adjusted residuals are obtained by
\[ \tilde{e}_i B^{-1/2} \tilde{e}_i, \ i = 1, \ldots, n, \] where \( B^{-1/2} \) is the inverse of the square root matrix of \( B \).

4. Sample with replacement from the adjusted residuals and estimated random effects to obtain \( \varepsilon_i^*(t_{ij}), w_i^* \). Note that for random effects, we sample the vector with replacement instead of individual elements.

5. Compute the bootstrap version of the data by

\[ y_i^*(t_{ij}) = \hat{\beta}_0 + \sum_{l=1}^{p} \sum_{q=1}^{Q_l} G_{lq}(t_{ij}) \hat{\beta}_{lq} + w_{0i}^* + w_{1i}^* t_{ij} + \varepsilon_i^*(t_{ij}). \]

6. Re-fit the model to the bootstrap version of the data using Algorithm 1 to obtain the bootstrap estimates of model parameters.

7. Repeat steps 3 to 5 \( B \) times to obtain \( B \) sets of bootstrap parameter estimates for inference. The number of repeats was \( B = 10,000 \) for the results in Section 3.4.1.

Let \( \theta \) be any unknown quantity that is of interest and \( \hat{\theta} \) be the estimate. We use the following bias-corrected bootstrap CI procedure to obtain a CI for \( \theta \).

**Algorithm 4:**

1. Sort the \( B \) bootstrap estimates \( \hat{\theta}^1, \ldots, \hat{\theta}^B \) in increasing order and obtain \( \hat{\theta}^{*(b)}, b = 1, \ldots, B \).

2. The lower and upper bounds of the approximate 100(1 - \( \alpha \))% CI for \( \theta \) is

\[ \left[ \hat{\theta}^{*(l)}, \hat{\theta}^{*(u)} \right] \]

where \( l = B \Phi_{nor} (2z_q + z_{\alpha/2}) \) and \( u = B \Phi_{nor} (2z_q + z_{1-\alpha/2}) \). Here \( z_p = \Phi_{nor}^{-1}(p) \) is the \( p \) quantile of the standard normal distribution, \( q \) is the proportion of the \( B \) values of \( \hat{\theta}^* \) that are less than \( \hat{\theta} \), and \( \lceil \cdot \rceil \) is the round function.
3.B Graphical Check of Degradation Data Model Fit

Figure 3.12 shows the studentized residuals for the degradation model. The plot shows that the constant variance assumption holds reasonably well. We also use the Q-Q plot to check the normality assumption, as shown in Figure 3.12b. The plot also shows that the normality assumption holds well. These graphical checks indicate that overall the model assumptions hold reasonably well. Those figures are referenced in Section 3.4.2 of the chapter.

3.C Selection of Knots and Spline Orders

We use the AIC criterion (e.g., Eilers and Marx 1996) to do a systematic selection of the number of knots on our spline-based model. Figure 3.13 shows the plot of the AIC values for different numbers of interior knots under different degrees of smoothness. For each spline order, we plot the AIC values as a function of the number of the spline knots. Based on the results in the figure, order three splines with four knots are used to fit the degradation paths in the chapter. The AIC value is very close to that of order two splines with four knots. The
model fitting results are also very close to that of order two splines with four knots.

We also conduct a sensitivity analysis on the spline orders and the number of knots. Figure 3.14 shows a sensitivity analysis to study the effects that the spline order and the number of interior splines knots have on the estimation of the covariate-effect functions. The results show robustness on the spline orders and the number of knots. This is consistent with the conclusion in Meyer (2008), in which it is pointed out the estimated functions are relatively insensitive to the number of knots, largely due to the shape restrictions. These figures are referenced in Section 3.4.2 of the chapter.

3.D Graphical Check of Covariate Process Model Fits

We also examine the autocorrelation function (ACF) of the residuals (estimated \([e_1(\tau), e_2(\tau), e_3(\tau)]^T\)) to check the assumption of the vector auto-regressive (VAR) model. Figure 3.15 shows the plot of ACF for the residuals from the models for ultra-violet (UV), temperature, and relative humidity (RH) covariate data. The plot shows no evidence of autocorrelation.
Figure 3.14: Sensitivity analysis on the spline orders and the number of interior spline knots on the estimation of covariate-effect functions.

Figure 3.15: The plot of ACF for the residuals from the models for UV, temperature, and RH covariate data.

Figure 3.16 shows the Q-Q plots for the residuals for the UV, temperature, and RH covariate data. The results show the normality assumption holds well. Overall, these figures show that the VAR(2) model provides an adequate description of the residuals. These figures are referenced in Section 3.6.3 of the paper.
Figure 3.16: Q-Q plots for the residuals for the UV, temperature, and RH covariate data.
Bibliography


Chapter 4  Spatial and Temporal Emergence Pattern of Lyme Disease in Virginia

Abstract

The emergence of infectious diseases over the past several decades has highlighted the need to better understand and prepare for epidemics and the spread of endemic diseases into new areas where they may become established. As these diseases expand their geographic range, cases are recorded over multiple time periods, making the analysis and prediction more complicated. In this study, based on areal (census tract level) count data of Lyme disease cases in Virginia from 1998 to 2011, we analyzed the spatial patterns of the disease using statistical smoothing analysis. We also used the space and space-time scan statistics to reveal the presence of clusters in the spatial or spatial/temporal distribution of Lyme disease. Our results confirm and quantify the continued emergence of Lyme disease to the south and west in states along the eastern coast of the United States, and highlight areas where education and surveillance needs are highest.

Key Words: Kernel smoothing, Clusters, Space and Space-time scan, Lyme disease.
4.1 Introduction

Lyme disease has become the most common vector-borne disease in the United States. The disease is caused by the bacterium *Borrelia burgdorferi* and is transmitted to humans by the tick vector, *Ixodes scapularis* or *Ixodes pacificus*. Patients suffering from Lyme disease can have acute symptoms such as fever, headache, swollen glands, fatigue and skin rash during early stage of infection. If not treated in the early stage of infection, patients could also develop severe and chronic symptoms that can affect multiple areas of the body, such as: transient heart block, facial palsy, shooting pains or numbness in the extremities, memory problems, or arthritis in major joints such as the knees. The infection of Lyme disease has become a significant public health burden in the U.S. In fact, a 1998 study estimated the treatment cost of Lyme disease in the U.S. was about $2.5 billion over a five year period (Maes et al., 1998).

Although Lyme disease was initially endemic to New England and a few other northeastern states, over the last several decades it has expanded its range southward, westward and northward from its initial endemic range in the eastern United States. Since the first identification of the illness in 1975 in the town of Lyme, CT (Steere, 1982), cases of the disease have been reported in the northeastern United States, the upper Midwest and California (Waller et al., 2007; Kitron and Kazmierczak, 1997; Eisen et al., 2006). There were 248,000 reported cases from 1992 to 2006 and a number of states (Bacon et al., 2008) including Virginia, have experienced an expansion in the endemic range of Lyme disease in recent years. As can be seen from Figure 4.1, there was a steady increase in Lyme cases in Virginia from the late 1990s to 2006, but the number of cases more than tripled in the five year period from 2007 to 2011 compared to the previous five year period. According to the Virginia Department of Health (VDH), the increase is not due solely to reporting or diagnostic changes, but rather is the result of increased cases as it has emerged southward into Virginia from neighboring endemic states. Up until the early 2000s Virginia was on the front line of the Lyme disease range expansion. Rapid yet geographically-localized suburban/periurban development and the southward expansion of the disease into Virginia makes the state a valuable model for
addressing important scientific questions regarding the role of environmental conditions and anthropogenic habitat change on Lyme disease emergence at multiple spatial scales. First, we need to quantify and visualize Lyme disease in Virginia and confirm its emergence, which is the main purpose of this chapter.

Spatial statistics, including spatial smoothing and cluster analysis, are often used to analyze and create better visualization of spatial patterns of disease. One typical goal of spatial pattern visualization, which usually utilizes disease rates, is to provide insight into the geographic variation in disease risk (Waller and Gotway, 2003). A challenge exists, however, because we often have spatial regions that have either small numbers of cases, or a small population. This is usually referred as the small number problem (Waller and Gotway, 2003). The true spatial pattern of disease risk may be obscured by this issue because the rates of some regions might be better estimated than others merely because there are more data. Spatial smoothing is a statistical approach to reduce the noise associated with estimation of the rates in different geographic regions (Waller and Gotway, 2003). Another goal of spatial pattern visualization is to detect whether there are unusual aggregations of disease cases in certain geographic regions, i.e., disease clusters, in order to provide a clearer picture of where scarce public health resources should be targeted. In this broader context, the goal of this chapter is to quantify spatial and temporal emergence patterns of human cases of Lyme disease in Virginia between 1998 and 2011 to determine where human case clustering occurs. Understanding spatial and temporal emergence patterns will pinpoint priority areas for the education of physicians, so they are aware that their patients could be afflicted by the disease, and the general public, who can employ preventative measures to avoid contracting the illness. Additionally, an understanding of diffusion patterns can aid in the creation of predictive models for future spread.
4.2 Materials and Methods

4.2.1 Study Area

Statewide surveillance for Lyme disease in Virginia started in 1989 when it became a reportable disease (Seukep et al., 2013). Isolated Lyme disease cases were first seen to be most prevalent on the Eastern Shore of Virginia but subsequently became prevalent in the northern Virginia counties adjacent to Maryland. Northern Virginia is one of the most populated regions in the state. In recent decades, Virginia has seen the disease spread southward and westward into the less developed and less densely populated central and southern regions (Figure 4.2). Meanwhile, across the state, suburbanization has brought residential land use into landscapes that were formerly characterized by agricultural land, mixed with small woodlots and other forested land. The resulting disturbance of this pattern has created a fragmented mix of land cover types that form habitat favoring white-tailed deer and the white-footed mouse, both key species in tick reproduction and Lyme disease transmission cycle. While there is a limitation in treating Virginia as a closed system, the focus of this chapter is on the emergence of Lyme disease from Virginia’s northern counties toward the south, rather than on the initial spread of Lyme disease into the state.

4.2.2 Lyme Disease Data

The Virginia Department of Health receives all Lyme disease positive laboratory reports for patients from commercial testing laboratories, and these reports as well as Lyme disease case reports provided to VDH by physicians are investigated as part of Virginia’s Lyme disease case surveillance. Lyme case classification at VDH is based on the National Surveillance Case Definition for Lyme disease (CDC National Notifiable Diseases Surveillance Definitions) (dis, 2013). Lyme disease cases that were counted after 1998 were mapped by patient address for geographic analysis at VDH. We obtained the VDH case coordinate data and aggregated these latitude and longitude data to census tracts for analysis. The reason for choosing the
census tract as the unit of analysis includes: 1) availability of population data as well as other demographic information for each census tract, facilitating standardization of the raw count data; 2) census tracts boundaries are based, in most cases, on permanent features such as roads, rivers, or railroads, which constitute potential barriers to the movement of mice or deer in comparison to arbitrary political boundaries, such as zip code boundaries; and 3) a previous study found that road-bounded polygons of varying sizes (Jackson et al., 2006b), which are similar to census tracts, produced stronger models in a study of Lyme disease and forest-edge habitats than a study design using grid cells of standard sizes.

The time frame we considered in this chapter is from 1998 to 2011, excepting 1999 and 2002, which were removed from the analysis because we do not have case data at the census tract level for those years. There were 6,714 cases of Lyme disease identified through the study period (1998 to 2011). The cases identified in the last five years of the study period, i.e., from 2007 to 2011, account for 74% of the total cases. Most disease pattern studies involve only a four or five year period (Jackson et al., 2006a; Chaput et al., 2002) but investigating a time period of more than a decade allows us not only to detect the spatial pattern, but also the temporal pattern, of a disease’s emergence. Population information for each census tract is available using 2010 census data that can be downloaded from the Census Bureau. For all other years, we have intercensal estimates at the county level (the 2000 census tract definitions differ from those of 2010, hence the 2000 census tract population cannot be used directly). A projection is made to obtain the population in each census tract for each year, i.e., the population for census tract \( i \) in year \( j \) is determined by the product of proportion of tract \( i \)s population in its county in 2010 census data and the countys population in year \( j \) based on the intercensal estimates.

Due to changes to the criteria for case reporting made to the 2008 National Lyme disease Surveillance Case Definition, higher standards of clinical and laboratory evidence were required for case reporting starting in 2008. Therefore, we screened the case data reported prior to 2008 to remove potentially spurious points. As a remedy to inaccuracies, we identified tracts that had cases reported before 2008 but not after 2008, and the cases in those tracts before
2008 were considered probable misdiagnoses. There were 44 cases found and removed from analysis, resulting in a total of 6,714 cases used in the study.

4.3 Visualization

4.3.1 Spatial Smoothing

Disease rates reflect the chance for contracting the disease in a certain area. Hence, a disease rate map is particularly useful for public health practitioners to determine the high risk areas. However, a map of raw rates may obscure the spatial pattern in disease risk, particularly if the rates are based on populations of varying sizes (Waller and Gotway, 2003). For example, suppose there are two regions A and B, where region A has 100 cases and 1,000 people while region B has only 1 case and 10 people. Both regions will result in a disease rate of 1/10. However, we must notice that the variability in the estimated rates in different regions, which depends on the populations, is quite different. In our study, based on Virginia’s 2010 census data, the maximum population for a census tract is 22,060 while the minimum is only 1. Disease rates based on larger populations are better estimated, while those based on smaller populations might be elevated artificially simply because the sample size is not large enough. Visually, a map of raw disease rates usually could not clearly convey the information about disease spread patterns due to the high variability that exists in disease estimation (Figures 4.1 and 4.2).

To address this issue, we use a locally weighted average approach that utilizes a kernel function to smooth the rate for each census tract by averaging the values associated with neighboring regions (Waller and Gotway, 2003). The aggregated case number for census tract $i$ during year $t$ is denoted by $N(s_i, t)$, where $i = 1, \ldots, n$, $t = 1, \ldots, T$, $n$ is the number of census tracts in Virginia (per 2010 definition) and $T$ is the number of years where the data at census tract level are available. The population for census tract $i$ at year $t$ is denoted by $E_{it}$. Further, let $d_{ij}$ be the distance between the centroid of census tract $i$ and $j$. The raw disease rate for census tract $i$ during year $t$ is denoted by $\lambda(s_i, t)$, which is, by definition, estimated
Figure 4.1: Plots of raw Lyme disease rates from 1998 to 2005.
Figure 4.2: Plots of raw Lyme disease rates from 2006 to 2011.
by \( N(s_i, t)/E_{it} \). The locally weighted estimates of the disease rate \( \hat{\lambda}(s_i, t) \) for census tract \( i \) during year \( t \) is obtained using the following formula,

\[
\hat{\lambda}(s_i, t) = \frac{\sum_{j=1}^{n} K_b(s_i, s_j)N(s_j, t)}{\sum_{j=1}^{n} K_b(s_i, s_j)E_{jt}},
\]

where \( K_b(s_i, s_j) \) is a bi-variate kernel function with bandwidth parameter \( b \), which assigns more weights to observations that are closer to location \( s \). While there are several choices for the kernel function, we found the bi-variate Gaussian function works very well through cross-validation. The bandwidth parameter determines how many neighboring regions would be included in the locally weighted averaging. The choice of the bandwidth parameter is determined for each year, respectively, using cross-validation (Waller and Gotway, 2003). The implementation of the spatial smoothing is done using R (R, 2013).

4.3.2 Spatial and Temporal Clustering

It is important to know if the cases are distributed randomly over space/time, or if we could identify spatial/temporal clusters for statistical significance in order to better understand a disease’s emergence pattern. To locate disease clusters, we used a local analysis method using spatial scan statistics (Kulldorff, 1997, 1998). Using a circular window with a continuously adjusted radius, this spatial statistical technique can detect probable size-varied geographic clusters as the center of the window moves over the study area. A helpful review and generalization of the likelihood ratio test statistic is available in Fraker et al. (2008). It has been used to detect potential geographic clusters of various human and equine disease (Hjalmars et al., 1996; Viel et al., 2000; Lian et al., 2007; Huang et al., 2009). We applied the Kulldorff’s spatial clustering method to investigate the spatial pattern of the Virginia human Lyme disease case data in each year of the study period.

To incorporate the time dimension into the analysis, a space-time scan statistic was used where the scan window changes to a cylinder with an additional time dimension (Kulldorff et al., 1998; Kulldorff, 2001; Sonesson, 2007). With a varying time period for the cylinder,
the space-time scan statistic can identify significant space-time clusters, such as those found in an equine West Nile epidemic study (Lian et al., 2007). Woodall et al. (2008) provides a detailed review and performance evaluation of prospective scan methods including Kulldorff’s method. We applied the Kulldorff’s space-time scan statistic to the Virginia Lyme case data which included the whole study period, in both prospective and retrospective ways. Both the spatial and space-time scan statistics are implemented in the publicly available software SaTScan (www.satscan.org).

4.3.3 Results

Figures 4.3 and 4.4 show the smoothed rate from 1998 to 2011. Year 1998 to 2005 were during the period where the case number was still low but steadily increasing. Year 2006 was before the number of cases increased dramatically. Since 2006, the disease count has been high and hit a peak in 2010. Compared to the raw rate maps shown in Figures 4.1 and 4.2, the smoothed maps show more continuity in the disease rate, and hence are more valuable in identifying major trends in emergence. At the beginning of our investigation period, Lyme disease cases were focused in the northern Virginia area. Then the main disease occurrence area shifted southward to less developed areas. By the end of the investigation period, the disease had spread to the southern boundary of the state. The eastern shore of Virginia on the Delmarva Peninsula has a moderate rate in all time periods, but we have to notice that the population in that area is very low so the rate might be elevated artificially.

Using the maximum spatial cluster size of ≤50% of the total population, the statistically significant spatial clusters (p-value < 0.05) are identified for each year. Results are shown in Figures 4.5 and 4.6. The primary cluster is the one that is least likely to have occurred by chance by maximum likelihood estimation. The secondary clusters are those that are in rank order after the primary cluster, by their likelihood ratio test statistic. For example, in 2003, the primary cluster was the northern Virginia area or semi-rural counties west of Washington, D.C. The cluster included 78 census tracts and contains only 3.78% of the state’s population. In 2006, the primary cluster was the same general area but somewhat expanded and shifted
Figure 4.3: Plots of smoothed Lyme disease rates from 1998 to 2005.
Figure 4.4: Plots of smoothed Lyme disease rates from 2006 to 2011.
a bit towards more populated tracts. This cluster includes 332 census tracts and contains 17.67% of the state’s population. It had a relative risk (RR) of 11.75, with 252 observed cases vs. 62 expected cases. The disease continued to expand toward the southwestern part of the state since then. In 2010, the primary cluster includes 426 census tracts, which accounts for 16.67% of the state’s area, and contains 24.86% of the population. It had an RR of 6.38, with 836 observed cases vs. 306 expected cases. Cluster information is summarized in Table 1. Notice that from 2006 to 2010, the percentage of population in the primary cluster has not increased. This effect is due to the fact that the primary cluster shifted from a more populated area to a less populated area in some years (Figure 4.4). However, the area of the primary cluster keeps growing (Table 4.1), expanding into the less populated western part of the state.

The space-time scan statistic reveals that the primary prospective cluster (Figure 4.7) covers the period from 2008 to 2011, and is located in the northern Virginia area with 426 census tracts and an average of 24.69% population. It had an RR of 6.71, with 2748 observed cases while the expected cases was 673. The primary retrospective cluster (Figure 4.8) covers the period from 2007 to 2010, which is also located in the northern Virginia area with 268 census tracts and 15.34% of the state’s population. It had a RR of 8.6, with 2336 observed cases while the expected cases was 415.

4.4 Discussion

Our study found that the spatial distribution of Lyme disease in Virginia is not random, but rather clustered, which is consistent with a previous study of Lyme disease’s emergence in New York (Glavanakov et al., 2001). Clusters were also observed in other tick-borne disease such as Human Granulocytic Ehrlichiosis (HGE) (Chaput et al., 2002). Clusters are of great importance to public health practitioners, as they point out which areas suffered most from the disease in the past. For example, physicians in the primary cluster area should be well educated to recognize that patients in this area are under a high risk of Lyme disease infection.
Figure 4.5: Lyme disease clusters from 1998 to 2005.
Figure 4.6: Lyme disease clusters from 2006 to 2011.
Figure 4.7: Plots of Lyme disease prospective clusters.

Figure 4.8: Plots of Lyme disease retrospective clusters.
Table 4.1: Lyme disease primary cluster summary by years.

<table>
<thead>
<tr>
<th>Year</th>
<th>Annualized Rate (Per 10,000 Population)</th>
<th>Primary Cluster Location</th>
<th>No. of Census Tracts Included in the Primary Cluster</th>
<th>% of population of the primary cluster</th>
<th>% of area of the primary cluster</th>
<th>Relative Risk of the Primary Cluster</th>
<th>Expected Cases of the Primary Cluster</th>
<th>Observed Cases of the Primary Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998</td>
<td>10</td>
<td>Eastern Shore</td>
<td>4</td>
<td>0.26</td>
<td>0.87</td>
<td>38.41</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2000</td>
<td>20</td>
<td>Northern VA</td>
<td>55</td>
<td>2.10</td>
<td>1.82</td>
<td>30.49</td>
<td>3</td>
<td>57</td>
</tr>
<tr>
<td>2001</td>
<td>21</td>
<td>Northern VA</td>
<td>270</td>
<td>14.03</td>
<td>4.18</td>
<td>14.26</td>
<td>21</td>
<td>107</td>
</tr>
<tr>
<td>2003</td>
<td>25</td>
<td>Northern VA</td>
<td>78</td>
<td>3.78</td>
<td>2.77</td>
<td>20.98</td>
<td>7</td>
<td>84</td>
</tr>
<tr>
<td>2004</td>
<td>28</td>
<td>Northern VA</td>
<td>59</td>
<td>2.88</td>
<td>1.77</td>
<td>17.59</td>
<td>6</td>
<td>72</td>
</tr>
<tr>
<td>2005</td>
<td>35</td>
<td>Northern VA</td>
<td>124</td>
<td>6.86</td>
<td>3.22</td>
<td>15.92</td>
<td>18</td>
<td>143</td>
</tr>
<tr>
<td>2006</td>
<td>46</td>
<td>Northern VA</td>
<td>332</td>
<td>17.67</td>
<td>3.81</td>
<td>11.75</td>
<td>62</td>
<td>252</td>
</tr>
<tr>
<td>2007</td>
<td>123</td>
<td>Northern VA</td>
<td>279</td>
<td>15.67</td>
<td>7.91</td>
<td>10.11</td>
<td>148</td>
<td>618</td>
</tr>
<tr>
<td>2008</td>
<td>119</td>
<td>Northern VA</td>
<td>253</td>
<td>14.43</td>
<td>7.05</td>
<td>9.74</td>
<td>135</td>
<td>580</td>
</tr>
<tr>
<td>2009</td>
<td>110</td>
<td>Northern VA</td>
<td>328</td>
<td>18.63</td>
<td>8.97</td>
<td>6.57</td>
<td>162</td>
<td>521</td>
</tr>
<tr>
<td>2010</td>
<td>154</td>
<td>Northern VA</td>
<td>426</td>
<td>24.86</td>
<td>16.67</td>
<td>6.38</td>
<td>306</td>
<td>836</td>
</tr>
<tr>
<td>2011</td>
<td>123</td>
<td>Northern VA</td>
<td>295</td>
<td>17.47</td>
<td>20.12</td>
<td>6.96</td>
<td>172</td>
<td>585</td>
</tr>
</tbody>
</table>
Table 4.2: Lyme disease space-time cluster analysis (prospective) summary, showing the number of census tracts, the average % of population, the % of area, the relative risk, the expected cases and observed cases within each cluster.

<table>
<thead>
<tr>
<th>Cluster Location</th>
<th>Cluster Indicator</th>
<th>Time Frame of the Cluster</th>
<th>No. of Census Tracts</th>
<th>Average % of Population</th>
<th>% of Area</th>
<th>Relative Risk</th>
<th>Expected Cases</th>
<th>Observed Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern VA</td>
<td>Primary</td>
<td>2008/1/1 to 2011/12/31</td>
<td>426</td>
<td>24.69</td>
<td>16.67</td>
<td>6.71</td>
<td>673</td>
<td>2748</td>
</tr>
<tr>
<td>Southwestern VA</td>
<td>Secondary</td>
<td>2008/1/1 to 2011/12/31</td>
<td>21</td>
<td>1.35</td>
<td>3.07</td>
<td>4.34</td>
<td>37</td>
<td>157</td>
</tr>
<tr>
<td>Eastern Shore</td>
<td>Secondary</td>
<td>2008/1/1 to 2011/12/31</td>
<td>12</td>
<td>0.58</td>
<td>2.55</td>
<td>4.35</td>
<td>16</td>
<td>68</td>
</tr>
<tr>
<td>Central VA</td>
<td>Secondary</td>
<td>2010/1/1 to 2011/12/31</td>
<td>97</td>
<td>5.18</td>
<td>15.43</td>
<td>1.7</td>
<td>71</td>
<td>120</td>
</tr>
</tbody>
</table>
Table 4.3: Lyme disease space-time cluster analysis (retrospective) summary, showing the number of census tracts, the average % of population, the % of area, the relative risk, the expected cases and observed cases within each cluster.

<table>
<thead>
<tr>
<th>Cluster Location</th>
<th>Cluster Indicator</th>
<th>Time Frame of the Cluster</th>
<th>No. of Census Tracts</th>
<th>Average % of Population</th>
<th>% of Area</th>
<th>Relative Risk</th>
<th>Expected Cases</th>
<th>Observed Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern VA</td>
<td>Primary</td>
<td>2007/1/1 to 2010/12/31</td>
<td>268</td>
<td>15.34</td>
<td>7.85</td>
<td>8.60</td>
<td>415</td>
<td>2336</td>
</tr>
<tr>
<td>Central VA</td>
<td>Secondary</td>
<td>2010/1/1 to 2011/12/31</td>
<td>173</td>
<td>10.09</td>
<td>18.7</td>
<td>3.39</td>
<td>139</td>
<td>446</td>
</tr>
<tr>
<td>Southwestern VA</td>
<td>Secondary</td>
<td>2008/1/1 to 2011/12/31</td>
<td>21</td>
<td>1.35</td>
<td>3.07</td>
<td>4.34</td>
<td>37</td>
<td>157</td>
</tr>
<tr>
<td>Eastern Shore</td>
<td>Secondary</td>
<td>2003/1/1 to 2005/12/31</td>
<td>11</td>
<td>0.60</td>
<td>2.11</td>
<td>5.42</td>
<td>12</td>
<td>62</td>
</tr>
</tbody>
</table>
Especially, newly emerging clusters should be given additional special attention as they usually indicate the direction in which the disease is spreading. Specifically, Lyme disease in Virginia is clearly spreading towards southwestern part of the state as we can see the primary cluster is expanding in this direction. Based on our results, the southwestern part of Virginia is highly likely to join this primary cluster in the next few years. Actions from public health practitioners and the general public are much needed in order to prevent additional cases in this area as well as rapidly diagnose and treat those that do result. We also need to notice the limitation of our results, that is, the spatial analyses are based on locations where the cases were reported which are not necessarily where the infection was acquired.

Most of Virginia is abundant in tracts of both forested and herbaceous land cover which can provide an ideal habitat for animals such as white-tailed deer and the white-footed mouse. And as these animals play important roles in tick reproduction and Lyme disease transmission, it is not unexpected that Lyme disease would eventually become more prevalent in Virginia. One of our next steps in the research is to utilize environmental data and population data to build a model between Lyme disease occurrence, and environmental and demographic variables (Seukep et al., 2013). This will help us to identify the factors that affect Lyme disease infection risk, explain the pattern of Lyme disease spread and hopefully provide better guidance to public health practitioners.
Bibliography


Chapter 5  Statistical Prediction for Virginia Lyme Disease Emergence Based on Spatio-temporal Count Data

Abstract

The increasing demand for modeling spatio-temporal data is computationally challenging due to the large scale spatial and temporal dimensions involved. The traditional Markov Chain Monte Carlo (MCMC) method suffers from slow convergence and is computationally expensive. The Integrated Nested Laplace Approximation (INLA) has been proposed as an alternative to speed up the computation process by avoiding the extensive sampling process required by MCMC. However, even with INLA, handling large-scale spatio-temporal prediction datasets remains difficult, if not infeasible, in many cases. This chapter proposes a new Divide-Recombine (DR) prediction method for dealing with spatio-temporal data. A large spatial region is divided into smaller subregions and then INLA is applied to fit a spatio-temporal model to each subregion. To recover the spatial dependence, an iterative procedure has been developed to recombine the model fitting and prediction results. In particular, the new method utilizes a model offset term to make adjustments for each subregion using information from neighboring subregions. Stable estimation/prediction results are obtained after several updating iterations. Simulations are used to validate the accuracy of the new method in model fitting and prediction. The method is then applied to the areal (census tract level) count data for Lyme disease cases in Virginia from 2003 to 2010.

Key Words: Divide-Recombine, INLA, Intrinsic CAR, Markov random field, MCMC, Spatio-temporal.
5.1 Introduction

5.1.1 Motivation

There is an increasing demand for more efficient modeling of spatio-temporal data for the many applications that incorporate location-and-time-aware sensors in everyday devices such as mobile devices and GPS sensors. Examples of the use of spatio-temporal data are abundant, including disease outbreaks in specific areas in a certain time period, economic monitoring of countries over periods of several years, daily air pollution measurements for cities for a month, long-term weather predictions for farmers, satellite images of parts of the sky, and brain imaging for different illnesses. The application that is the focus of this chapter is the prediction of future Lyme disease counts for Virginia based on historical count data.

Lyme disease is an illness caused by the bites of black-legged ticks infected with a bacterial species known as *Borrelia burgdorferi*. The transmission ecology of Lyme disease involves a set of complex interactions between black-legged ticks, *Borrelia burgdorferi*, competent animal hosts, deer, and the environment. Environmental variables such as suburbanization can often result in fragmented and disturbed forest environments that favor the white-footed mouse, which are highly competent reservoirs of *Borrelia burgdorferi* and white-tailed deer, which serve as primary blood sources for female adult tick populations.

Lyme disease is the most common vector-borne disease in the United States. It is currently spreading southward along the East Coast and Virginia is on the front line of this expansion. Lyme disease was first discovered in Lyme, Connecticut in 1975. During the next decade it became prevalent in southern and western areas of New England and New York and by the 1990s it had spread to southern New Jersey, southeastern Pennsylvania, Delaware, northern and coastal Maryland, and the eastern shore of Virginia (Bacon et al., 2008). Since the late 1990s, Virginia has seen a steady increase in human cases. It is most prevalent in northern Virginia, but is spreading south-westward through the state.

The emergence of infectious diseases in recent decades has highlighted the need to better
understand and predict outbreaks. Diseases of this type tend to expand their geographic range and the case counts are usually recorded over multiple time periods. Thus spatio-temporal data are often available. Virginia’s Lyme disease data are provided by the Virginia Department of Health (VDH), who collect their data across different time periods and spatial regions. The objective of the work reported in this chapter was therefore to build a spatio-temporal model for predicting future Lyme counts, taking into account the spatial and temporal dependence of the data.

5.1.2 Existing Methods

When analyzing spatio-temporal data, it is necessary to consider any autocorrelations between both successive time points and nearby areas. When the dataset becomes large, computational efficiency is a major concern. Traditionally, Markov Chain Monte Carlo (MCMC) algorithms are used in spatial and/or spatio-temporal models to estimate parameters. The calculation of likelihood, however, often involves computing the inverse of a large covariance matrix whose complexity is \( O(N^3) \), which may require a substantial amount of computing time. Here, \( N \) is the number of spatial areas. This problem is called the “big \( N \) problem” (Banerjee et al., 2004); MCMC is not computationally feasible when \( N \) is too large. The situation is even worse for spatio-temporal models. When consider spatio-temporal interactions, the dimension of the covariance matrix is \( N \times T \), where \( T \) is the number of time points. For datasets covering a large spatial region over a long time period, the complexity of the computing may well exceed the capacity of current computing systems.

Recently, several methods have been proposed to attempt to solve this computationally challenging problem. These include rank-reducing approximations (Banerjee et al., 2008; Gelfand et al., 2012) and the Integrated nested Laplace approximation (INLA). One interesting rank-reducing method is the knot-based predictive approach, which reduces computing time by using a fixed set of representative “knots”. The set of \( M \) knots has a much lower dimension \( M \ll N \), comparing to the original number of datapoints \( N \) and the dimension of covariance matrix is correspondingly reduced from \( N \times N \) to \( M \times M \). The spatial interpolation at a
single site $s_0$ can be derived by applying point to point kriging to the selected knots. The number of knots is selected by comparing sensitivity and speed performance by implementing different values for $M$. Liang and Kumar (2013) have proposed a new Markov cube kriging method. Their method partitions the spatio-temporal region into spatio-temporal cubes and then performs the interpolation using hierarchical time-space kriging. Kriging thus plays an important role in interpolating from knots to the locations of interest in these methods. These methods have been widely used for point reference data. Although some area-to-area kriging methods have been developed (Goovaerts, 2006), the extension of those methods to areal data is often complicated because it is necessary to consider the different sizes, shapes, and properties of areas during area-to-area kriging. Aggregation over areas also reduces the resolution of the original dataset.

Other rank-reducing methods include the spatial random effect proposed by Cressie and Johannesson (2008) and the spatio-temporal random effects models developed by Cressie et al. (2010). These methods define spatial/spatio-temporal random effects in terms of a set of spatial basis functions whose coefficients then need to be estimated. As the number of basis functions is fixed, the rank is reduced to the number of basis functions. This method has been extensively used for point reference data, especially for satellite data. One application in areal data handling has been for a study of the distribution of SIDS over the counties in North Carolina (Sengupta and Cressie, 2013), where the state was divided into 13 regions and then a 13-dimensional binary $(0, 1)$ basis was utilized. However, when there is a large number of areas involved, the choice of dimension for the basis can be difficult.

The integrated nested Laplace approximation (INLA) was proposed for hierarchical Gaussian Markov random field models in Rue et al. (2009). Instead of the extensive sampling needed for MCMC, INLA directly computes approximations of the marginal posteriors for all parameters and random effects and can thus perform quick and accurate non-sampling based Bayesian inferences. The nested approximation is based on the Laplace approximation (Tierney and Kadane, 1986) and the Gaussian approximation (Rue and Held, 2005). The goal of most spatio-temporal analysis is to predict a future event at a specific location and time.
Performing predictions is relatively straightforward with INLA because the prediction is done as part of the model fitting by treating the desired prediction responses as missing data. INLA uses the sparseness of the precision (inverse covariance) matrix to speed up the calculation. However, this is not always possible where the precision matrix lacks sufficient sparseness, for example when the spatio-temporal interaction is complicated. An R interface for the INLA program (r-INLA package) is available at the website www.r-inla.org.

This chapter presents a new method to deal with large scale spatio-temporal prediction that is based on a combination of the Divide-Recombine (DR) method and INLA. Simulations are used to validate the speed and accuracy of the new method, which is then applied to an areal count dataset containing the Lyme disease count data at census tract level in Virginia for the period from 2003 to 2010.

5.1.3 Overview

The remainder of this chapter is organized as follows. Section 5.2 introduces the spatio-temporal data, along with the covariates and models for describing this type of data. Sections 5.3 develops a DR method for spatio-temporal prediction and provides an algorithm that is suitable for parallel computing. Section 5.4 uses simulations to validate the accuracy and speed of the new method. Section 5.5 extends the DR method to allow for spatio-temporal interactions. Section 5.6 applies the DR method to Lyme disease data for Virginia. The chapter ends with Section 5.7, which contains some concluding remarks and describes possible areas for future research.
5.2 Data and Model

5.2.1 Spatial Region, Time Domain, and Data

Consider a spatial domain $S = \{1, 2, 3, \ldots, N\}$, where the areas are indexed by $i = 1, \ldots, N$, and $N$ is the number of areas. The neighbors of area $i$ are denoted by $N_i$, for $i \in S$. That is,

$$N_i = \{j \in S : j \text{ is a neighbor of } i\}, i \in S.$$

The neighborhood can be defined by adjacency criterion such as sharing a common border or inverse distance. In this chapter, neighborhood is defined in terms of the sharing a common border criterion. The number of neighbors for the $i$th area is denoted as $n_i$. The time points we considered are indexed by $t$, $t = 1, 2, 3, \ldots, T$, which contains a total of $T$ time points. Here, the response variable of interest is the spatio-temporal counts. The counts for area $i$ at time $t$ are denoted by $y_{it}$, $i = 1, \ldots, N$, and $t = 1, \ldots, T$. Covariates may also be available for analysis. The value of covariate $l$ for area $i$ at the time $t$ is denoted by $x_{itl}$, where $i = 1, \ldots, N$, $t = 1, \ldots, T$, $l = 1, 2, 3, \ldots, L$, and $L$ is the number of covariates. The covariates can also be written in a vector form $x_{it}$. That is, $x_{it} = (x_{it1}, x_{it2}, \ldots, x_{itL})'$.

5.2.2 Spatio-Temporal Model

The response variable $y_{it}$ that is of scientific interest is modeled using hierarchical modeling (Jin et al., 2005; Donald et al., 2011). In particular, the Poisson distribution is used to model $y_{it}$ and the mean term $\mu_{it}$ is modeled as

$$\log(\mu_{it}) = \beta_0 + x_{it}'\beta + \upsilon_i + \gamma_t. \quad (5.1)$$

Here $x_{it}$ is the covariates vector, $\beta$ is the corresponding coefficient vector, and $\beta_0$ is the intercept term. The random effect $\upsilon_i$ is a spatially structured component and the random effect $\gamma_t$ is the temporally structured component.
For areal data, autoregressive model is commonly used for the random effect $\upsilon_i$. One common choice is the conditional autoregressive (CAR) model, which was first introduced by Besag (1974). It is used to specify the conditional distribution of each random variable given their neighbors. In low-dimensional space, in order to have substantial spatial dependence the spatial dependency parameter $\rho$ in the CAR model must be close to one (Besag, 1981). After setting $\rho$ to 1, CAR becomes the intrinsic CAR (iCAR) model (Banerjee et al., 2004). Denote the vector of the Gaussian random variables as $\upsilon = (\upsilon_1, \upsilon_2, \cdots, \upsilon_N)'$. The joint density of $\upsilon$ is (Knorr-Held, 2000)

$$
\pi(\upsilon|\tau_s) \propto \exp \left( -\frac{\tau_s^2}{2} \upsilon'Q\upsilon \right).
$$

Here $Q$ is called the spatial precision matrix which reflects the neighborhood structure, and $\tau_s$ is the spatial precision parameter. Taking a weight of 1 for each pair of neighbors, then $Q_{ij} = -1$ if $i$ and $j$ are neighbors, otherwise $Q_{ij} = 0$. The diagonal element of $Q$ is $Q_{ii} = n_i$ where $n_i$ is the number of neighbors of the $i$th area. For example, if we consider a $3 \times 3$ grid spatial region as follows

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>
the corresponding $Q$ matrix is

$$
Q = \begin{pmatrix}
2 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
-1 & 3 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & -1 & 2 & 0 & 0 & -1 & 0 & 0 & 0 \\
-1 & 0 & 0 & 3 & -1 & 0 & -1 & 0 & 0 \\
0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\
0 & 0 & -1 & 0 & -1 & 3 & 0 & 0 & -1 \\
0 & 0 & 0 & -1 & 0 & 0 & 2 & -1 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 3 & -1 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 2
\end{pmatrix}.
$$

Because $Q$ is singular with a deficient rank of $N - 1$ for the iCAR model, a direct inverse matrix is not possible. Thus $Q^{-1}$ will be the pseudo-inverse matrix of $Q$. The CAR model can also be written in conditional distribution form as

$$
\nu_i | \nu_j, j \in N_i \sim N \left( \bar{\nu}_j, \frac{1}{\tau^2 s_{i_j}} \right),
$$

and $\bar{\nu}_j$ is the average of the random effects of all neighbors $N_i$. The Bayesian approach is particularly suitable for iCAR models because it uses the full conditional distribution of $\nu_i$. The iCAR prior is also called pairwise difference prior (Besag and Kooperberg, 1995). Because the rank of $Q$ is $N - 1$, the constraint $\sum_{i=1}^{n} \nu_i = 0$ is needed to ensure the identifiability of the intercept term $\beta_0$ (Carlin and Banerjee, 2003).

The temporal random effect $\gamma_t$ is usually modeled using a time series model, such as the first or second order random walk (RW), first or second order autoregressive (AR), moving average (MA) or autoregressive integrated moving average (ARIMA) models. In the hierarchical Poisson model utilized here, a first order random walk (Clayton, 1996) is applied, denoted by RW(1), to model for the temporal trend $\gamma_t$. By definition, a time series $\gamma_t$ is said to follow an RW(1) if the difference from one observation to the next observation satisfies the following
model,
\[ \gamma_t = \gamma_{t-1} + a_t. \]

Here \( a_t \) is the random noise term that accounting for differences from one observation to the next observation and follows a normal distribution. Denoting the vector form of the temporal random effects as \( \gamma = (\gamma_1, \gamma_2, \cdots, \gamma_T)' \), the joint density of \( \gamma \) is derived as (Knorr-Held, 2000),
\[ \pi(\gamma|\tau) \exp \left( -\frac{\tau_t^2}{2} \gamma'W \gamma \right), \]
where \( W \) is the temporal precision matrix with a deficient rank of \( T - 1 \). For example, consider a time domain with 6 time points. The corresponding \( W \) matrix is
\[
W = \begin{pmatrix}
1 & -1 & 0 & 0 & 0 & 0 \\
-1 & 2 & -1 & 0 & 0 & 0 \\
0 & -1 & 2 & -1 & 0 & 0 \\
0 & 0 & -1 & 2 & -1 & 0 \\
0 & 0 & 0 & -1 & 2 & -1 \\
0 & 0 & 0 & 0 & -1 & 1
\end{pmatrix}.
\]

Just as for \( \nu_i \), \( \gamma_t \) must also be centered around zero so that the intercept term \( \beta_0 \) can be identified. Although RW(1) is one of the simplest time series models, it is widely used in epidemiology, finance, and many other areas. RW(1) is also used in in this chapter to illustrate the DR method. The method described here, however, can also be easily extended to other time series models.

### 5.3 The Proposed Divide-Recombine Method

#### 5.3.1 Region Division

The first step in the new Divide-Recombine (DR) method is to divide the original spatial domain into \( D \) smaller disjoint regions; the original spatial region is denoted as \( S \). The
smaller regions are called subregions and are denoted as $D_d$, with $d = 1, \cdots, D$. That is

$$S = \bigcup_{d=1}^{D} D_d \text{ and } D_d \cap D_{d'} = \emptyset, \forall d, d' \in \{1, \cdots, D\}.$$

For area $i$ which belongs to subregion $D_{d_i}$, the neighbors of $i$ in its subregion $D_{d_i}$ are referred to as subset-neighbors and denoted as

$$\mathcal{M}_i = N_i \cap D_{d_i}, \ i \in D_{d_i}.$$

The number of these subset-neighbors for $i$ is denoted as $m_i$. The neighbors of $i$ in different subregions are called pseudo-neighbors and denoted as

$$\mathcal{M}_i^* = N_i \cap (S \setminus D_{d_i}), \ i \in D_{d_i}.$$

The number of these pseudo-neighbors is $m_i^*$. Note that for the original spatial region $S$, the neighbors for area $i$ consist of the union of $\mathcal{M}_i$ and $\mathcal{M}_i^*$. That is,

$$N_i = \mathcal{M}_i \cup \mathcal{M}_i^*.$$

The total number of neighbors is $n_i = m_i + m_i^*$.

5.3.2 Divide-Recombine Procedure

The DR procedure is described as follows.

DR Procedure:

1. Use INLA to fit model (5.1) to each subregion $D_i, i = 1, \cdots, D$. This is can be done efficiently by using parallel computing. The random effect of area $i$ depends only on its subset-neighbors $\mathcal{M}_i$, and the diagonal element $Q_{ii}$ of the precision matrix is $m_i$. This step gives initial value of $v_i$ and represents the 0th iteration.
2. For area $i$, $i = 1, \cdots, N$, in order to account for the dependence from both subset-neighbors $\mathcal{M}_i$ and pseudo-neighbors $\mathcal{M}_i^*$, it is necessary to calculate an offset term, denoted by $\kappa_i^{(k-1)}$, that is computed using estimated random effects $v_j^{(k-1)}$ from neighbors in $\mathcal{M}_i^*$ from the $(k - 1)^{th}$ iteration. Here, the iterations are indexed by $k = 1, 2, 3, \cdots$.

In particular, the offset term is computed as

$$\kappa_i^{(k-1)} = \frac{\sum_{j^* \in \mathcal{M}_i^*} v_j^{(k-1)}}{m_i + m_i^*}.$$ 

3. The effects of the pseudo-neighbors are incorporated by adding $\kappa_i^{(k-1)}$ to model (5.1) and by adjusting the diagonal elements of the spatial precision matrix $Q$ to $n_i$ for each subregion $\mathcal{D}_i$. That is, for the $k$th iteration, we have

$$v_i | v_{j, j^* \in \mathcal{M}_i}, v_{j^*, j^* \in \mathcal{M}_i^*} \sim N\left(\frac{\sum_{j \in \mathcal{M}_i} v_j}{m_i + m_i^*} + \kappa_i^{(k-1)}, \frac{1}{\tau^2(m_i + m_i^*)}\right).$$

Or equivalently,

$$\log(\mu_{it}) = \kappa_i^{(k-1)} + \beta_0 + \mathbf{x}'_{it} \beta + v_i + \gamma_t.$$  

(5.2)

4. With the updated offset term $\kappa_i^{(k-1)}$, use INLA to fit model (5.2) to each subregion $\mathcal{D}_i$, where $i = 1, \cdots, D$, separately by using parallel computing.

5. Repeat steps 2 to 4 until convergence is achieved. The number of iterations needed for convergence is denoted as $K$.

During iterations, the values of $\kappa_i$’s are checked for convergence. When $\kappa_i$’s are stable, that is $\max_{i \in S} |\kappa_i^{(k)} - \kappa_i^{(k-1)}| \leq 0.0005$, the iterations end. The estimates of random effects and fixed effect parameters will be stable if $\kappa_i$’s are stable. The computing process for this DR Procedure can easily be parallelized for different subregions to save time, and thus is computationally efficient. The time cost is approximately the time of fitting the largest subregion multiplied by $K$, the number of iterations.
5.3.3 A Simple Illustration

To illustrate the principle of the DR procedure, we consider a simple situation. Suppose a spatial region $S = \{1, 2, 3, 4\}$ contains 4 areas indexed by $i = 1, 2, 3, 4$. The four areas are arranged in a $1 \times 4$ array, as shown in Figure 5.1. The spatial random effects $\upsilon_i$, where $i = 1, 2, 3, 4$, follow an iCAR prior. That is, $\upsilon_i | \upsilon_j, j \in N_i \sim N(\bar{\upsilon}_i, \sigma^2/n_i)$, where $\sigma$ is the inverse of $\tau_s$. The spatial region is divided into two subregions: $D_1 = \{1, 2\}$ and $D_3 = \{3, 4\}$, as shown in Figure 5.1.

Area 2 is adjacent to areas 1 and 3. Thus, area 1 is its subset-neighbor and area 3 is its pseudo-neighbor, which is also shown in Figure 5.1. In particular, $N_2 = \{1, 3\}$, $M_2 = \{1\}$ and $M_2^* = \{3\}$. Also, the number of neighbors is $m_2 = 1$, $m_2^* = 1$, and $n_2 = m_2 + m_2^* = 2$.

For the initial iteration, to estimate the random effect of area 2, only information from $D_1$ is used in the following model,

$$\upsilon_2 | \upsilon_1 \sim N\left( \frac{\upsilon_1}{1}, \frac{\sigma^2}{1} \right).$$

Then, to incorporate dependence information from both areas 1 and 3, for the $k$th iteration, the offset term is updated by $\kappa_2^{(k-1)} = \upsilon_3^{(k-1)} / (m_2 + m_2^*) = \upsilon_3^{(k-1)} / 2$. Thus

$$\upsilon_2 | \upsilon_1, \upsilon_3 \sim N\left( \frac{\upsilon_1}{2} + \frac{\upsilon_3^{(k-1)}}{2}, \frac{\sigma^2}{2} \right).$$

Similarly, area 3 is adjacent to areas 2 and 4. Area 4 is its subset-neighbor and area 2 is its pseudo-neighbor. For the initial iteration, the model becomes

$$\upsilon_3 | \upsilon_4 \sim N\left( \frac{\upsilon_4}{1}, \frac{\sigma^2}{1} \right),$$

and for the $k$th iteration the model is

$$\upsilon_3 | \upsilon_4, \upsilon_2 \sim N\left( \frac{\upsilon_4}{2} + \frac{\upsilon_2^{(k-1)}}{2}, \frac{\sigma^2}{2} \right).$$

Areas 1 and 4 have no pseudo-neighbors. Thus they depend solely on their own subset-
neighbors, areas 2 and 3, respectively. During the iterations, the following models are used
\[
v_1 | v_2 \sim N\left(\frac{v_2}{1}, \frac{\sigma^2}{1}\right) \quad \text{and} \quad v_4 | v_3 \sim N\left(\frac{v_3}{1}, \frac{\sigma^2}{1}\right).
\]

5.4 Simulation Study

5.4.1 Simulation Setup

To study the performance of the DR procedure, data \(y_{it}\) can be simulated from the Poisson distribution with the following simplified mean structure,
\[
\log(\mu_{it}) = \beta_0 + v_i + \gamma_t.
\]

For simplicity, set \(\beta_0 = 0\) in the simulation. The data are simulated based on a two-dimensional spatial domain with \(N\) areas and \(T = 10\) time points. The spatial region is specified in terms of a \(\sqrt{N} \times \sqrt{N}\) grid, as shown in Figure 5.2. Ten time points were chosen because that is close to the value in the data for the Lyme disease application that is described later in the chapter. Here, the spatial random effect \(v_i\) follows an iCAR prior and is simulated by using the pseudo-inverse of the spatial precision matrix \(Q\). The temporal random effect \(\gamma_t\) is simulated from the RW(1) prior, using a precision matrix \(W\). The precision parameters
for the iCAR and RW(1) priors are denoted by \( \tau_s \), and \( \tau_t \), respectively. A range of values for \( \tau_s \) are considered in the simulation, \( \tau_s^2 = 0.5, 1 \), and 10. A wide range of sample sizes \( N \) are investigated, namely \( N = 100, 400, 784, 1,296, 4,096, \) and 7,056, with values chosen so that each sample size produces a square grids with a length of side equal to \( \sqrt{N} \). Because the DR procedure is purely a spatial division with no temporal division involved, only one value of \( \tau_t \) is used in the simulation: \( \tau_t^2 = 10 \), producing a variance parameter of 0.1. For each parameter setting, 100 datasets are simulated and the average performance of those 100 datasets is reported. Figure 5.2 shows an example of the simulated data for a certain time point with a spatial sample size \( N = 100 \).

### 5.4.2 Models for Comparison

For each simulated dataset, data from \( t = 1, \cdots, 9 \) were used as training set to build the model to predict the counts at \( t = 10 \). The simulated data for \( t = 10 \) was then compared to the prediction generated from the model. The first step was to build the model using data from the whole spatial region \( S \) for the model in (5.3), referred to here as the “All-Data” method.
The fitting and prediction process was performed using the INLA package in R. The spatial region was then divided into 2, 4 and/or 8, 16, 49, 64, and 144 subregions evenly, based on the size of $N$, and the DR procedure was applied to each subregion. The fitting and prediction process was also performed by using the INLA package.

The computing time and accuracy was compared for the All-Data method, DR procedure, and MCMC method. For the model using the All-Data method, the time was recorded based on one run. For the DR procedure, parallel computing was used such that the computing time of one iteration was approximately the time needed for running one subregion. The offset terms $\kappa$ typically reach convergence after 7-8 iterations, and the model fitting and prediction also produce stable results. Here 10 iterations were performed to make sure of the convergence. Thus, 10 multiplied by the time needed to run one subregion was recorded as the result. For the MCMC method, block-wise implementations of the adaptive Metropolis-Hastings algorithm were utilized for sampling and the spatial random effects were updated in a block of size 10. The total time was recorded for one chain of 5,000 burn-in and 5,000 MCMC samples. The time comparison was performed on a laptop with an Intel Core i5-3210 CPU (2.5 GHz), 4GB Memory and 64-bit Windows 7 operating system.

Fitting and prediction accuracy were also compared. The fitting accuracy was calculated by using the log of the probability of obtaining counts from $t = 1$ to $t = 9$ in the training set given the fitted mean $\hat{\mu}_{it}$, $i = 1, \cdots, N$, $t = 1, \cdots, 9$. The prediction accuracy was calculated using log of probability of getting true counts at $t = 10$ given the predicted mean $\hat{\mu}_{it}$, $i = 1, \cdots, N$, $t = 10$.

5.4.3 Simulation Results

Table 5.1 and Figure 5.3 show the computing times for the methods being compared. As the spatial sample size $N$ increases, the time cost for the All-Data method increases dramatically, especially for $N \geq 1000$. For the DR method, the time cost depends solely on the size of each subregion. The MCMC method is consistently the slowest among all the methods tested for all sample sizes. When sample size $N$ is small, fitting the All-Data method is much faster.
than running 10 iterations using the DR method. However, when \( N \) is large, the benefit of the DR method begins to show. This ranking also depends on how many subregions that are used. When \( N \) is large, the more subregions are divided, the faster the speed achieved. When \( N \) reaches 7,056, the All-Data method costs more than 200 times as much time as the DR method using 144 subregions. Although no datasets larger than 7,056 were tested, the trend in the computing time is clear, as are the benefits to be gained in terms of saving computing time by using the DR method. In practice, spatial datasets tend to be very large, so the DR method has the potential to serve as a valuable new tool.

The accuracy of fitting and prediction achieved by the different simulation settings are shown in Tables 5.2 and 5.3. Here, the accuracy is computed as the average of 100 simulated datasets for each setting. Due to computing time cost concerns, the datasets were limited to sizes at or below \( N = 1,296 \). With smaller \( \tau_s \), the fitting accuracy becomes slightly worse, and the prediction accuracy decreases. Comparing the results of different methods using the same simulation settings, the fitting accuracy of the DR method is not always as good as the fitting achieved by the All-Data method but in a comparable scale. The prediction accuracy for the DR method, however, is often better than or very close to the results for the All-Data method. These results indicate that the new DR method is comparable to the All-Data method in accuracy.

5.5 Extension of the Model to Include Spatio-temporal Interactions

5.5.1 Model with Spatio-temporal Interactions

In some applications, the model in (5.1) will further include an interaction term, yielding the following model.

\[
\log(\mu_{it}) = \beta_0 + \mathbf{x}'_{it}\mathbf{\beta} + v_i + \gamma_t + \delta_{it}.
\]  

(5.4)
Table 5.1: Computing time in seconds for different sample sizes and methods, when $\tau_s^2 = 10$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>Model</th>
<th>Time(s)</th>
<th>$N$</th>
<th>Model</th>
<th>Time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>All-Data</td>
<td>5.7</td>
<td>1,296</td>
<td>All-Data</td>
<td>117.7</td>
</tr>
<tr>
<td></td>
<td>2 subregions</td>
<td>35.1</td>
<td></td>
<td>4 subregions</td>
<td>119.6</td>
</tr>
<tr>
<td></td>
<td>4 subregions</td>
<td>21.6</td>
<td></td>
<td>8 subregions</td>
<td>63.1</td>
</tr>
<tr>
<td></td>
<td>MCMC</td>
<td>102.5</td>
<td></td>
<td>16 subregions</td>
<td>43.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MCMC</td>
<td>890.2</td>
</tr>
<tr>
<td>400</td>
<td>All-Data</td>
<td>15.5</td>
<td>4,096</td>
<td>All-Data</td>
<td>1389.3</td>
</tr>
<tr>
<td></td>
<td>2 subregions</td>
<td>66.3</td>
<td></td>
<td>4 subregions</td>
<td>763.7</td>
</tr>
<tr>
<td></td>
<td>4 subregions</td>
<td>32.1</td>
<td></td>
<td>8 subregions</td>
<td>220.0</td>
</tr>
<tr>
<td></td>
<td>8 subregions</td>
<td>22.5</td>
<td></td>
<td>16 subregions</td>
<td>71.3</td>
</tr>
<tr>
<td></td>
<td>MCMC</td>
<td>180.0</td>
<td></td>
<td>64 subregions</td>
<td>24.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MCMC</td>
<td>8510</td>
</tr>
<tr>
<td>784</td>
<td>All-Data</td>
<td>87.5</td>
<td>7,056</td>
<td>All-Data</td>
<td>4463.6</td>
</tr>
<tr>
<td></td>
<td>2 subregions</td>
<td>151.4</td>
<td></td>
<td>8 subregions</td>
<td>793.2</td>
</tr>
<tr>
<td></td>
<td>4 subregions</td>
<td>80.3</td>
<td></td>
<td>16 subregions</td>
<td>183.3</td>
</tr>
<tr>
<td></td>
<td>8 subregions</td>
<td>35.9</td>
<td></td>
<td>49 subregions</td>
<td>60.1</td>
</tr>
<tr>
<td></td>
<td>16 subregions</td>
<td>20.5</td>
<td></td>
<td>144 subregions</td>
<td>21.4</td>
</tr>
<tr>
<td></td>
<td>MCMC</td>
<td>439.0</td>
<td></td>
<td>MCMC</td>
<td>21610</td>
</tr>
</tbody>
</table>

Figure 5.3: Plot for the computing time (in seconds) for different sample sizes and methods, when $\tau_s^2 = 10$ (both axes are in log scale).
Table 5.2: Estimation and prediction accuracy when \( N = 100, 400, \tau_s^2 = 10, 1, 0.5, \) and \( \tau_t^2 = 10. \)

<table>
<thead>
<tr>
<th>( N )</th>
<th>( \tau_s^2 )</th>
<th>( \tau_t^2 )</th>
<th>Model</th>
<th>Fit Accuracy</th>
<th>Predict Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>10</td>
<td>All-Data</td>
<td>-1169.56</td>
<td>-158.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-1177.45</td>
<td>-158.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-1190.17</td>
<td>-156.50</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>10</td>
<td>All-Data</td>
<td>-1156.07</td>
<td>-161.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-1155.66</td>
<td>-161.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-1163.73</td>
<td>-159.89</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>10</td>
<td>All-Data</td>
<td>-1167.03</td>
<td>-207.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-1165.27</td>
<td>-206.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-1170.30</td>
<td>-204.54</td>
</tr>
<tr>
<td>400</td>
<td>10</td>
<td>10</td>
<td>All-Data</td>
<td>-4657.07</td>
<td>-619.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-4660.43</td>
<td>-619.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-4671.71</td>
<td>-617.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8 subregions</td>
<td>-4700.63</td>
<td>-615.47</td>
</tr>
<tr>
<td>400</td>
<td>1</td>
<td>10</td>
<td>All-Data</td>
<td>-4635.87</td>
<td>-687.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-4634.25</td>
<td>-693.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-4630.63</td>
<td>-691.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8 subregions</td>
<td>-4630.33</td>
<td>-686.42</td>
</tr>
<tr>
<td>400</td>
<td>0.5</td>
<td>10</td>
<td>All-Data</td>
<td>-4695.12</td>
<td>-816.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-4691.75</td>
<td>-816.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-4686.10</td>
<td>-815.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8 subregions</td>
<td>-4682.58</td>
<td>-811.77</td>
</tr>
</tbody>
</table>
Table 5.3: Estimation and prediction accuracy when $N = 784, 1,296$, $\tau^2_s = 10, 1, 0.5$, and $\tau^2_t = 10$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\tau^2_s$</th>
<th>$\tau^2_t$</th>
<th>Model</th>
<th>Fit Accuracy</th>
<th>Predict Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>784</td>
<td>10</td>
<td>10</td>
<td>All-Data</td>
<td>-9125.81</td>
<td>-1201.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-9129.25</td>
<td>-1201.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-9134.25</td>
<td>-1201.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8 subregions</td>
<td>-9166.62</td>
<td>-1201.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16 subregions</td>
<td>-9225.99</td>
<td>-1200.89</td>
</tr>
<tr>
<td>784</td>
<td>1</td>
<td>10</td>
<td>All-Data</td>
<td>-9140.48</td>
<td>-1427.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-9137.61</td>
<td>-1427.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-9133.14</td>
<td>-1425.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8 subregions</td>
<td>-9123.39</td>
<td>-1422.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16 subregions</td>
<td>-9121.59</td>
<td>-1415.45</td>
</tr>
<tr>
<td>784</td>
<td>0.5</td>
<td>10</td>
<td>All-Data</td>
<td>-9321.93</td>
<td>-1511.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-9319.12</td>
<td>-1510.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-9312.60</td>
<td>-1510.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8 subregions</td>
<td>-9303.28</td>
<td>-1507.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16 subregions</td>
<td>-9305.07</td>
<td>-1502.00</td>
</tr>
<tr>
<td>1,296</td>
<td>10</td>
<td>10</td>
<td>All-Data</td>
<td>-15169.52</td>
<td>-1862.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-15173.27</td>
<td>-1862.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-15176.85</td>
<td>-1862.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8 subregions</td>
<td>-15198.33</td>
<td>-1861.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16 subregions</td>
<td>-15253.68</td>
<td>-1859.96</td>
</tr>
<tr>
<td>1,296</td>
<td>1</td>
<td>10</td>
<td>All-Data</td>
<td>-15128.32</td>
<td>-2354.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-15126.83</td>
<td>-2354.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-15121.29</td>
<td>-2353.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8 subregions</td>
<td>-15113.62</td>
<td>-2351.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16 subregions</td>
<td>-15104.37</td>
<td>-2346.33</td>
</tr>
<tr>
<td>1,296</td>
<td>0.5</td>
<td>10</td>
<td>All-Data</td>
<td>-15355.11</td>
<td>-2378.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 subregions</td>
<td>-15352.00</td>
<td>-2378.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 subregions</td>
<td>-15346.69</td>
<td>-2379.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8 subregions</td>
<td>-15336.09</td>
<td>-2379.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>16 subregions</td>
<td>-15326.95</td>
<td>-2377.01</td>
</tr>
</tbody>
</table>
where $\delta_{it}$ is a spatio-temporal interaction term. As an illustration, the most complicated interaction case described in (Knorr-Held, 2000) is examined, namely the type IV interaction.

In the type IV interaction, the precision matrix of the interaction, denoted by $R$, is the Kronecker product of the precision matrices of the iCAR and RW(1) models, $R = Q \otimes W$. Because the precision matrix of neither iCAR nor RW(1) is full ranked, the constructed Kronecker product is also singular with a rank of $(N - 1) \times (T - 1)$. A Constraint must be added to take into account this rank deficiency, and Rue and Held (2005) gives a detailed description on how to impose the necessary constraints.

For area $i$, when only the subset-neighbors are considered, the conditional mean of the interaction term $\delta_{it}$ is

$$
\bar{\delta}_{it} = \begin{cases} 
\delta_{i, t+1} + \frac{\delta_{jt}}{m_i^t} - \frac{\delta_{jt+1}}{m_i^t}, & \text{for } t = 1, \\
\delta_{i, t-1} + \frac{\delta_{jt}}{m_i^t} - \frac{\delta_{jt-1}}{m_i^t}, & \text{for } t = T, \\
\frac{1}{2} (\delta_{i, t+1} + \delta_{i, t-1}) + \frac{\delta_{jt}}{m_i^t} - \frac{(\delta_{jt+1} + \delta_{jt-1})}{2m_i^t}, & \text{for } t = 2, \cdots, T - 1.
\end{cases}
$$

As in the definition of the offset term $\kappa_i$, information regarding pseudo-neighbors $j^*$ is incorporated by adding another offset term $\xi_{it}$ to the conditional mean of $\delta_{it}$.

5.5.2 Divide-Recombine Method for Model with Interaction

To fit the spatio-temporal interaction model, the DR procedure is now extended as follows:

**DR Procedure with Interaction:**

1. Use INLA to fit model (5.4) to each subregion $D_i, i = 1, \cdots, D$. This is can be done efficiently by using parallel computing. The random effect of area $i$ depends solely on its subset-neighbors $M_i$, and the diagonal element $Q_{ii}$ of the precision matrix is $m_i$. This step yields the initial value of $\nu_i$ and $\delta_{it}$ and represents the $0^{th}$ iteration.

2. For area $i, i = 1, \cdots, N$, to account for the dependence from both subset-neighbors $M_i$ and pseudo-neighbors $M_i^*$, the offset terms, denoted by $\kappa_i^{(k-1)}$ and $\xi_{it}^{(k-1)}$ are calculated.
using estimated random effects \(v_{j}^{(k-1)}\) and \(\delta_{j}^{(k-1)}\) from neighbors in \(M_i^*\) from the \((k-1)th\)
iteration. Here, the iterations are indexed by \(k = 1, 2, 3, \ldots\). In particular, the offset
term is computed as

\[
\kappa_i^{(k-1)} = \frac{\sum_{j^* \in M_i^*} v_{j^*}^{(k-1)}}{m_i + m_i^*} \quad \text{and,}
\]

\[
\xi_{it}^{(k-1)} = \begin{cases} 
\frac{\sum_{j^* \in M_i^*} \delta_{j^*}^{(k-1)}}{m_i + m_i^*} - \frac{\sum_{j^* \in M_i^*} \delta_{j^*}^{(k-1)}}{m_i + m_i^*}, & \text{for } t = 1, \\
\frac{\sum_{j^* \in M_i^*} \delta_{j^*}^{(k-1)}}{m_i + m_i^*} - \frac{\sum_{j^* \in M_i^*} \delta_{j^*}^{(k-1)}}{m_i + m_i^*}, & \text{for } t = T, \\
\frac{\sum_{j^* \in M_i^*} \delta_{j^*}^{(k-1)}}{m_i + m_i^*} - \frac{\sum_{j^* \in M_i^*} \left( \delta_{j^*}^{(k-1)} + \delta_{j^*}^{(k-1)} \right)}{2(m_i + m_i^*)}, & \text{for } t = 2, \ldots, T - 1.
\end{cases}
\]

3. The effects of the pseudo-neighbors are incorporated by adding \(\kappa_i^{(k-1)}\) and \(\xi_{it}^{(k-1)}\) to
model (5.4) and by adjusting the diagonal elements of spatial precision matrix \(Q\) to \(n_i\)
for each subregion \(D_i\). Recalculate spatio-temporal interaction precision matrix \(R\) using
adjusted \(Q\) by \(R = Q \otimes W\). That is, for \(k^{th}\) iteration, we have

\[
\log(\mu_{it}) = \kappa_i^{(k-1)} + \xi_{it}^{(k-1)} + \beta_0 + x_{it}' \beta + v_i + \gamma_t + \delta_{it}.
\]

(5.7)

4. With the updated offset terms \(\kappa_i^{(k-1)}\) and \(\xi_{it}^{(k-1)}\), INLA can now be used to fit model (5.7)
to each subregion \(D_i, i = 1, \ldots, D\), separately by using parallel computing.

5. Repeat steps 2 to 4 until convergence is reached.

### 5.6 Application to Lyme Disease Count Data

#### 5.6.1 Data

The new method can now be applied to the Lyme disease data for Virginia. The first step
is to aggregate the Lyme disease cases from VDH to census tract level. There are a total
of 1907 census tracts in Virginia and their boundaries are based on geographical features
such as rivers, and roads. Neighborhoods are defined as sharing common boundaries and the
Table 5.4: Virginia Lyme disease yearly count data.

<table>
<thead>
<tr>
<th>Year</th>
<th># of Cases</th>
<th>Year</th>
<th># of Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>122</td>
<td>2005</td>
<td>274</td>
</tr>
<tr>
<td>2000</td>
<td>156</td>
<td>2006</td>
<td>357</td>
</tr>
<tr>
<td>2001</td>
<td>149</td>
<td>2007</td>
<td>959</td>
</tr>
<tr>
<td>2002</td>
<td>259</td>
<td>2008</td>
<td>933</td>
</tr>
<tr>
<td>2003</td>
<td>195</td>
<td>2009</td>
<td>868</td>
</tr>
<tr>
<td>2004</td>
<td>216</td>
<td>2010</td>
<td>1233</td>
</tr>
</tbody>
</table>

latitude and longitude of the centroid of each census tracts are recorded. The time domain being examined here is the 9-year period from 2003 to 2011; the data for the first 8 years will be used to build the model, and then the built model will be used to predict the 2011 data and the results will be compared with the actual dataset. The response variable is the Lyme count, aggregated to census tract level. In general, Virginia has seen a steady increase in human cases since the late 1990s (Table 5.4).

The data were first cleaned by removing questionable cases. The Lyme surveillance case definition changed in January 2008, so some of the cases counted prior to 2008 should be regarded as highly suspect if there have been no cases reported since then. Given the tremendous increase in numbers and geographic spread of Lyme disease that occurred starting in 2007 and the media attention that Lyme disease began to receive at that time, it would be unexpected for counties that had cases up through 2007 to not have any after that date. Therefore, after examining the cases in each county carefully, 26 counties that recorded 0 case from 2008 onwards but reported cases before that were removed and the corresponding count set to 0 in the analysis. The first step in the analysis visualized the spatio-temporal pattern of Lyme disease across the state. Figure 5.4 shows the plots of Lyme counts for each year from 2003 to 2010. Lyme cases occurred mostly in the northeast part in Virginia in 2005, but then the disease began to spread in a southwesterly direction across the state. In some areas it grew rapidly, for example, around Floyd in the New River Valley area.

The United States Census Bureau provides demographic and environmental information at the census tract level. Some demographic and environmental variables have been found to
Figure 5.4: Plots of Lyme disease counts from 2003 to 2010.
be correlated to Lyme disease and/or the abundance of black-legged ticks in previous studies (Bacon et al., 2008; Ostfeld and Keesing, 2000; Glass et al., 1995; Das et al., 2002; Allan et al., 2003; Ostfeld et al., 2006; Frank et al., 1998; Brownstein et al., 2005). These variables were therefore chosen as explanatory variables in this chapter. The Census Bureau does not provide this information for every year, so only the data from 2006 or 2010 were available for this study. Table 5.5 shows the notation, reported year, and other useful information regarding these variables.

There are 41 census tracts with 0 inhabitants or with missing values for the demographic and environmental variables, so these census tracts were deleted, resulting in a total of 1866 census tracts. Principal component analysis was used to process the covariates because of the high level of multicollinearity among these variables. The variables x1-x14 were firstly centered by mean and standardized by their standard deviation. The first 6 components $C_1, \cdots, C_6$ contain most variation, so these 6 components were used in the following analysis.
5.6.2 Region Division

The preliminary study of Lyme cases indicates the presence of spatial autocorrelation. From Figure 5.4, it can be seen that Virginia can be divided into 2 major parts. The northwestern part along the Blue Ridge Mountains tends to have most of the Lyme cases. This pattern corresponds well to the level III ecoregions map of Virginia (Figure 5.5, image source: www.hort.purdue.edu). Specifically, the two major divisions of Virginia can consist of the ecoregions along the Blue Ridge Mountains (64, 66, 67, and 69 on the map) and those in the southeastern Plains portion of the state (45, 63, and 65 on the map). For this study, Virginia was divided into 29 subregions according to these ecoregions. In order to ensure that each subregion contains enough counts to allow for precise estimates of the model parameters, a minimum threshold value of 5 non-zero counts was imposed.

5.6.3 Model

The Lyme disease counts $y_{it}$ are modeled by the Poisson distribution with the mean model $\mu_{it}$

$$\log(\mu_{it}) = \log(P_{it}) + \beta_0 + \beta_1 C_{1i} + \cdots + \beta_6 C_{6i} + v_i + \gamma_t + \delta_{it}.$$  

(5.8)
The model contains a spatially structured component \( v_i \), a temporally structured component \( \gamma_j \), and a spatio-temporal interaction term \( \delta_{it} \). Here, \( P_{it} \) is the population of census tract \( i \) at year \( t \), \( \beta_0 \) is the intercept term, \( C_{1i}, \cdots, C_{6i} \) are the first 6 principle components of census tract \( i \), and \( \beta_1, \cdots, \beta_6 \) are the corresponding coefficients.

5.6.4 Results

The observed counts and the mean counts predicted by DR method for the year 2011 for each census tract are plotted in Figures 5.6 and 5.7. The northeastern part of Virginia, which contains most of the Lyme disease cases, is highlighted in Figure 5.8. After 20 iterations, the log of the probability of obtaining observed counts given fitted mean, is \(-6597.70\), for a total of \(1866 \times 8\) data points. The log of the probability of obtaining the 2011 observed values given the mean predicted by the 2003 to 2010 data, is \(-1305.22\), for a total of 1866 data points. The model’s prediction performance is thus reasonably good. The predicted mean is relatively close to most of the clusters in the observed spatial pattern, especially in areas where the risk is high such as in northern Virginia. In the southeastern part, where Lyme disease is rare (only a few census tracts have non-zero count cases and most areas have no cases), the predicted means for all these areas are close to zero. This is due to the strong emphasis on neighbor effects in the CAR model. If most of the neighbors of area \( i \) have a small random effect, the model will tend to generate a low prediction mean for area \( i \). These prediction results could serve to identify possible areas where education and surveillance needs are highest.

5.7 Conclusions and Areas for Future Research

Spatio-temporal datasets have the potential to be quite large and thus are computationally intensive, making them impossible to analyze directly in some case. This chapter reports the development of a novel method for analyzing large-scale spatio-temporal datasets. The DR method is an effective way to accelerate computing speed while still giving accurate results. Simulations revealed the results of the DR method to be as good or better than those obtained
Figure 5.6: Plots of 2011 Virginia Lyme disease counts and predicted mean.

Figure 5.7: Plots of 2011 Virginia Lyme disease counts and predicted mean.
by direct modeling of the whole dataset but in a fraction of the time that would be required, making DR method a potentially very useful approach for a number of practical applications such as brain imaging and disease mapping. Also, unlike other methods such as knot selected rank-reduction which result in losing some resolution, the DR method uses data from all the available spatial/temporal observations, avoiding the need to sacrifice resolution for speed.

In practice, the choice of how to best divide $S$ into subregions may be based on a number of different considerations. Firstly, if geographical information or background knowledge from a previous study is available, it may be advantageous to divide the subregions according to this information. For example, in the Lyme disease application, the subregions were chosen based on ecoregions. Secondly, it is important to make sure there is enough information in the subregion data to precisely estimate the parameters. In the case of Poisson data, if all of the response counts in a subregion are zero, it will be nearly impossible to obtain good estimates. Lastly, because the computing time depends on the size of the largest subregion, it is necessary to ensure that the largest subregion, whether divided evenly or not, is not too large.

The DR method fits the fixed effects by using a separate set of coefficients for each subregion.
which is not a concern when the purpose is prediction. When the focus is on estimation, however, the DR method can be extended to generate common estimates of coefficient covering the entire spatial region, which may be an interesting topic for future research. In this chapter, an RW(1) time series model was used for the purposes of illustration. However, the offset term for more complicated spatio-temporal interactions may in turn be more complicated itself and so future research should focus on extending the models to incorporate other forms of spatio-temporal interactions. Another interesting future topic is the extension of the adjacency spatial matrix $Q$ to some more complicated spatial weighted matrix forms, such as distance based weight or boundary based weight. Identifying bias in the DR process will also be interesting.
Bibliography


D. Liang and N. Kumar. Time-space kriging to address the spatiotemporal misalignment in the large datasets. *Atmos Environ*, 72:60–69, 2013.


Chapter 6  General Conclusions

This chapter summarizes the major conclusions and contributions of this dissertation and suggests some possible areas for future research.

6.1 Conclusions

Service life prediction is of great interest in the coating industry, especially for analyzing field data with dynamic explanatory variables such as, the UV spectrum, UV intensity, temperature, and humidity. Chapter 2 describes the development of a nonlinear mixed model with functional forms for this type of environmental covariate. Parameters are then estimated from lab data and used to predict degradation in the field using a cumulative degradation model. In Chapter 3, a degradation model that incorporates shape restricted splines is created and used to model the covariate process using a vector auto-regressive time series model. The models are then used to estimate failure time distributions and predict remaining life time. For both Chapters 2 and 3, the bootstrap procedure is utilized to generate prediction intervals.

The emergence of infectious diseases over the past several decades has highlighted the need to better understand and prepare for epidemics and the spread of endemic diseases into new areas. As these diseases expand their geographic range, cases are recorded over multiple time periods, making the analysis and prediction more complicated. In Chapter 4, kernel smoothing and cluster analysis of Lyme disease counts is performed, adjusted by population, to identify geographic areas at high risk. Chapter 5 presents and tests a Poisson model that can be used to predict future incident rates by using the proposed Divide-Recombine method.
6.2 Contributions

The research reported in Chapter 2 serves as an improvement and extension of efforts to link lab experimental data to outdoor field data. A physical motivated model is developed that incorporates random effects to better describe the degradation data. The effects of the explanatory variables, for which little knowledge was known, are extensively studied. After proposing possible functional forms of relative humidity, UV dosage and intensity, these forms are applied to lab experimental data. A cumulative damage model is also utilized to allow for dynamic covariates and the method is then applied to outdoor test data. The functional forms and cumulative nonlinear model developed as results of this work provide more accurate confidence intervals for the estimation of outdoor field degradation based on lab experimental data.

In Chapter 3, a new class of general path models is proposed that uses shape restricted splines to incorporate information of dynamic covariates. The new algorithm developed to estimate this general path model addresses a gap in the research, as there has been little previous work on the shape restrictions affecting covariate-effect functions. New algorithms to compute an estimate for the failure-time distribution and the remaining lifetime induced by the underlying degradation model are also described in this chapter. Although the primary motivation for this research is to develop an application specifically for outdoor degradation data, the new method will find broad application in many other fields.

Chapter 4 reports the first systematic study of the spatial patterns of Lyme disease in Virginia. Statistical smoothing analysis and space/space-time scan statistics are used to reveal the presence of clusters in the spatial/spatio-temporal distribution. These results confirm and quantify the continued emergence of Lyme disease to the south and west in states along the eastern coast of the United States, and highlight areas where education and surveillance needs are highest. This work will contribute to a wider study of the pattern of Lyme disease in the United States.

In Chapter 5, a new Divide-Recombine method is proposed for large-scale spatio-temporal
prediction and offers substantial improvements in computing time when dealing with large datasets. This method offers a promising alternative to other reduced rank methods for spatial analysis. It has the added advantage of not being restricted to the use of only selected data points, unlike other methods; the new method does not sacrifice resolution or produce prediction uncertainty from kriging.

6.3 Areas for Future Work

With regard to the degradation analysis, future research should focus on how best to estimate the random effect for each unit for nonlinear mixed models. The use of nonlinear functions similar to those described in Chapter 2 in the degradation paths to replace additional models that end up being equivalent to a linear degradation path model should also be explored. Interactions between covariates in the model have not been considered, and this could also be an interesting avenue for future research. It is, however, challenging to estimate interaction effects for a nonlinear model and to impose shape restrictions on interactions. There is also the additional constraint of having enough data points to estimate the interaction effects. Another interesting area could be to impose shape restrictions on the application of functional regression to degradation data.

With regard to the Lyme disease study, an RW(1) time series prior is used to illustrate the new DR method. Future research should focus on extending the new DR method to generate common estimates of coefficients for the entire spatial region, even when the focus is not prediction. Also, the model could be extended to include other forms of spatio-temporal interactions, with a more complicated time series prior.