ABSTRACT

HUBERMAN, DAVID. Advances In Spatial Statistics For Ecological and Environmental Data. (Under the direction of Brian J. Reich and Krishna Pacifici.)

Spatial statistical analysis is a fundamental tool in environmental and ecological research. The increasing availability of previously inaccessible datasets in these fields present new opportunities for statistical advances. In this dissertation, we develop novel statistical techniques which are applied to modern ecological and environmental datasets such as opportunistic ecological survey data, numerical hurricane model output, and spatially-distributed microbiome data.

In Chapter 2, we propose a species distribution analysis method designed to improve spatial estimation of environmental effects on species occupancy by utilizing opportunistic data gathered by citizen scientists. The method borrows concepts from mediation analysis in order to effectively integrate the more common designed survey data with the newer opportunistic data source. Our mediated method is compared to a default spatial occupancy modeling method and shows the capacity to efficiently incorporate large quantities of opportunistic data and improve environmental effect estimation.

In Chapter 3, we develop a flexible conditional density estimation technique incorporating machine learning algorithms to simultaneously estimate the full conditional distribution of the response variable. Conditional density estimation can provide additional insight such as the skewness or kurtosis of the process of interest which could better inform environmental decision-making and policy. The model and subsequent approximation for computation can be understood through an inhomogenous Poisson process (IPP) framework. We demonstrate this technique on a variety of simulated distributions versus other machine learning-based conditional density estimation techniques and then apply it to the problem of short-term forecasting of tropical cyclone intensities. Our method compares favorably to its competitors and can be used on a wide variety of ecological problems. We also list some potential areas of future work to fully explore and unlock the method’s capabilities.

In Chapter 4, we extend the conditional density estimation technique from Chapter 3 to a spatial point pattern analysis context. We develop this technique for use in performing forensic point geolocation on ecological datasets of fungal samples. Again, the IPP framework is applied for model development and computation. Additionally, we introduce user-defined functions in the neural network input layer to capture spatial dependence between the point locations in a discrete approximation to a Cox process.
model. We compare our method to an established forensic geolocation approach in a simulation study and then a real dataset of fungal samples across the continental United States of America. We finish with a discussion of possible future research and challenges related to our approach.
Advances In Spatial Statistics For Ecological and Environmental Data

by
David Huberman

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2020

APPROVED BY:

Jacqueline Hughes-Oliver

Kevin Gross

Brian J. Reich
Co-chair of Advisory Committee

Krishna Pacifici
Co-chair of Advisory Committee
DEDICATION

To my parents, Harry and Roselee; and my sister, Rachel.
Originally from San Francisco, California, David Huberman is the son of Harry Huberman and Roselee Greenholtz. David graduated from Lowell High School in 2008, and University of California Los Angeles in 2012 with a degree in statistics. He continued his education at North Carolina State University and received a masters in statistics in 2017.
I would first and foremost like to thank my advisor Dr. Brian Reich. Dr. Reich provided me with excellent guidance throughout my graduate school career, and his expertise and resourcefulness has been crucial in getting me to this point. There were many times when I thought my projects had hit a dead end only for Dr. Reich to come up with an idea that gave them new life. His patience and optimism made my graduate school experience less stressful than it might have been otherwise (although it was certainly stressful enough regardless). I am very grateful I ended up with Dr. Reich as my adviser. I would also like to thank my co-advisor Dr. Krishna Pacifici for all of his help throughout graduate school and the invaluable ecological expertise he brought to my projects. Unfortunately, he had a severe lack of expertise on research matters pertaining to Steph Curry’s greatness, but other than that he was a great co-advisor. I would also like to thank Dr. Kevin Gross and Dr. Jacqueline Hughes-Oliver for serving on my committee and taking the time to provide feedback on my work. I’d additionally like to thank Dr. Hughes-Oliver for her help during both of my stints serving as her TA.

I wish to thank all the faculty and stuff at North Carolina State University that have helped me along the way. I would like to acknowledge by name Lanakila Alexander, Alison McCoy, Dr. Jason Osborne, Dr. David Dickey, Dr. Marie Davidian, Dr. Dennis Boos, Terry Byron, and Dr. Jaime Collazo for their help. I would also like to thank the building maintenance staff who kept the BOM and SAS and Hunt Library buildings clean and made it possible for me to conduct my research. Additionally, thank you to Dr. Vivian Lew for putting the idea of going to graduate school in my head back at UCLA and to Corey West for providing me with a lot of guidance in DC, I owe you both a lot.

I would like to thank the graduate students in the program who made it welcoming and provided me assistance, advice, and friendship along the way. Thank you to everyone from my cohort for your help and support, especially Munir Winkel, Lilian Cheung, Stephanie ”Real 5” Chen, Amanda Brucker, Jacob Rhyne, Milo Page, Jirapat Samranvedhya, Arnab Hazra and Arkaprava Roy. Thank you to the students from other cohorts who also impacted my time in the program including Susheela Singh, Kevin Gunn, Andrew Giffin, Khuzaima Hameed, Yeng Saanchi, Benjamin Hu, Matt Miller, Anthony Weishampel, Marschall Furman, Alex Long, Katherine Allen, and Kara Martinez.

Last but not least, I would like to thank my family and friends for their support throughout the program. Without them I would not be here. My parents, sister, and Bobby always helped me in any way that was needed with no questions asked. I appreciate all of my Brandeis buddies for keeping in touch after all these years. I thank all of my
friends who kept me motivated throughout a long grind, with a special nod to Nicolas Molina who was with me every step of the way from the Covel Commons computer labs to now (built different). Lastly, I would like to thank my grandparents for surviving. I hope I will make you proud.
## Table of Contents

### List of Tables

- Chapter 1 Introduction
  - 1.1 Estimating the drivers of species distributions with opportunistic data using mediation analysis
  - 1.2 Nonparametric conditional density estimation in a deep learning framework for short-term forecasting
  - 1.3 Geolocation for spatial point pattern data in a deep learning framework
  - 1.4 Overview

- Chapter 2 Estimating The Drivers Of Species Distributions With Opportunistic Data Using Mediation Analysis
  - 2.1 Introduction
  - 2.2 Methods
  - 2.3 Simulation
  - 2.4 Case study
  - 2.5 Discussion
  - 2.6 Acknowledgements

- Chapter 3 Nonparametric Conditional Density Estimation In A Deep Learning Framework For Short-Term Forecasting
  - 3.1 Introduction
  - 3.2 Methods
  - 3.3 Computing
  - 3.4 Simulation study
  - 3.5 Application to tropical cyclone intensity forecasting

### List of Figures

- Chapter 1 Introduction
- Chapter 2 Estimating The Drivers Of Species Distributions With Opportunistic Data Using Mediation Analysis
- Chapter 3 Nonparametric Conditional Density Estimation In A Deep Learning Framework For Short-Term Forecasting

---

**Chapter 1 Introduction**

1. **Estimating the drivers of species distributions with opportunistic data using mediation analysis**
   - 1.1 Presence-absence and presence-only data for occupancy modeling
   - 1.2 Occupancy modeling
   - 1.3 Opportunistic data for occupancy modeling

2. **Nonparametric conditional density estimation in a deep learning framework for short-term forecasting**
   - 2.1 Kernel estimation of the joint and marginal densities
   - 2.2 Bayesian nonparametric mixture models for conditional density estimation
   - 2.3 Machine learning-based density estimation methods

3. **Geolocation for spatial point pattern data in a deep learning framework**
   - 3.1 Poisson process models
   - 3.2 Cox process models
   - 3.3 Gibbs process models

4. **Overview**

---

**Chapter 2 Estimating The Drivers Of Species Distributions With Opportunistic Data Using Mediation Analysis**

1. **Introduction**
2. **Methods**
   - 2.2.1 Mediation analysis review
   - 2.2.2 Mediation for spatial occupancy covariate modeling
3. **Simulation**
4. **Case study**
5. **Discussion**
6. **Acknowledgements**

---

**Chapter 3 Nonparametric Conditional Density Estimation In A Deep Learning Framework For Short-Term Forecasting**

1. **Introduction**
2. **Methods**
   - 3.2.1 Logistic transformation
   - 3.2.2 Polynomial regression model
   - 3.2.3 Deep learning model
3. **Computing**
4. **Simulation study**
5. **Application to tropical cyclone intensity forecasting**
| Table 2.1 | The average MSE, 95% coverage, and relative MSE results for the unmediated and mediated methods on data generated to have a correlation of $\approx 0.8$ on average between $Y_0(s)$ and $\theta(s)$. $p$ and $\kappa$ refer the detection and range parameter values used to generate the data, respectively. $n$ refers to the number of subsetted locations of the original 1,000 that retain $Y_1(s)$ data. | 24 |
| Table 2.2 | The average MSE, 95% coverage, and relative MSE results for the unmediated and mediated methods on data generated to have a correlation of $\approx 0$ on average between $Y_0(s)$ and $\theta(s)$. $p$ and $\kappa$ refer the detection and range parameter values used to generate the data, respectively. $n$ refers to the number of subsetted locations of the original 1,000 that retain $Y_1(s)$ data. | 25 |
| Table 2.3 | The posterior mean and standard deviation (SD) for each parameter under the unmediated and mediated methods for the red-eyed vireo analysis. | 30 |
| Table 3.1 | A table of the average computation times (in minutes) and associated standard errors for evaluating the data from Model 1 across all dataset sizes. The average computation times were recorded for the polynomial MCC approximation ($K = 1$), deep learning MCC approximation ($K = 1$), and deep learning IPP approximation ($K = 10$) conditional distribution estimation methods. The average computation times and standard errors for the deep learning IPP approximation method for 1000 and 4000 observations were calculated using only the first 5 datasets due to the burdensome evaluation time. The average computation times and standard errors for the remaining scenarios were calculated over all 100 datasets. The polynomial MCC approximation method was evaluated using penalized logistic regression, whereas the deep learning IPP approximation methods were evaluated using stochastic gradient descent. | 45 |
| Table 3.2 | The 5-fold mean CRPS values for each lag time for the QRF, DDR, polynomial MCC approximation ($K = 1$), and deep learning IPP approximation ($K = 20$) conditional distribution estimation methods. | 47 |
Table 4.1  A table of the average MAE and accompanying standard deviations for 1,133 microbial dust sample location predictions estimated using the Bayesian Discriminant Analysis, Radial Basis Deep Learning, and Coordinate Deep Learning methods across 5 folds. The deep learning methods are evaluated using each combination of either 30 or 50 nodes feeding into the hidden and output layers and 2 or 5 controls. Bayesian Discriminant Analysis was run using a sequence of candidate bandwidth values from 10 km to 100,000 km separated by increments of 10 km. The methods are evaluated on a grid of 6,266 candidate locations covering the continental USA. The geolocated distances from the true locations are calculated using the great circle distance in kilometers.

Table C.1  A table of the average negative log likelihood values for each of the methods under each of the scenarios along with their associated standard deviations. For each observation, a probability distribution across the spatial domain is predicted conditional on the OTU covariates, with the probability at the true spatial location used to calculate the contribution to the negative log likelihood.

Table C.2  A table of the average negative log likelihood and accompanying standard deviations for 1,133 microbial dust sample location predictions estimated using the Bayesian Discriminant Analysis, Radial Basis Deep Learning, and Coordinate Deep Learning methods across 5 folds. The deep learning methods are evaluated using each combination of either 30 or 50 nodes feeding into the hidden and output layers and 2 or 5 controls. Bayesian Discriminant Analysis was run using a sequence of candidate bandwidth values from 10 km to 100,000 km separated by increments of 10 km. The methods are evaluated on a grid of 6,266 candidate locations covering the continental USA. The geolocated distances from the true locations are calculated using the great circle distance in kilometers.
LIST OF FIGURES

Figure 2.1 Plots of $\theta(s)$ values, $Z(s)$ values, $Y_0(s)$ values when $\text{cor}(Y_0(s), \theta(s)) \approx 0.8$, and $Y_1(s)$ counts at each location for data generated with $\kappa = 1.5$ and $p = 0.75$. ........................................... 23

Figure 2.2 1 by 1 degree aggregated grids of eBird red-eyed vireo occupancy data, BBS red-eyed vireo occupancy proportions, and deciduous forest land cover proportions for 2008. eBird data were restricted to information to the East of the westernmost BBS grid site. Deciduous forest land cover data were restricted to locations with eBird data. eBird data is aggregated within grid cells and years and calculated as the logged proportion of aggregated sightings divided by aggregated effort hours plus 1 as $Y_0(s) = \log(C_0(s)/N_0(s)) + 1$. BBS data is also aggregated within grid cells and years and is divided by the corresponding number of aggregated survey stops as $Y_1t(s)/N_t(s)$. ........................................... 29

Figure 3.1 A boxplot of the distribution of CRPS divergences for each model and dataset size across 100 datasets for the QRF, DDR, Deep Learning $K = 1$, Polynomial $K = 1$, and Deep Learning $K = 10$ conditional distribution estimation methods. The y-axis scale is not synchronized across scenarios and dataset sizes. ................. 43

Figure 3.2 Deep learning IPP approximation ($K = 20$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with all available data. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $Pr(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. ....................................................... 48

Figure 3.3 Polynomial MCC approximation ($K = 1$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with all available data. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $Pr(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. ....................................................... 49
Figure 4.1  A boxplot of the distribution of MAEs for each number of observations and OTU covariates across 100 datasets for the Radial Basis Deep Learning, Coordinate Deep Learning, and Bayesian Discriminant Analysis estimation methods. The deep learning methods were run for both $K = 2$ and $K = 5$ controls. .......................... 60

Figure 4.2  A map of 1,133 microbial dust sample locations and corresponding location predictions by the Bayesian Discriminant Analysis method. The method is evaluated on an evenly spaced grid covering the continental USA. .......................... 63

Figure 4.3  A map of 1,133 microbial dust sample locations and corresponding location predictions by the Coordinate Deep Learning method. The method is evaluated on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 50 nodes feeding into the hidden and output layers, 5 controls, an initial learning rate of 4, and a ridge penalty of 0.00005. .......................... 64

Figure 4.4  A map of 1,133 microbial dust sample locations and corresponding location predictions by the Radial Basis Deep Learning method. The method is evaluated on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 30 nodes feeding into the hidden and output layers, 5 controls, an initial learning rate of 3, and a ridge penalty of 0.0001. .......................... 65

Figure B.1  Deep learning IPP approximation ($K = 20$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 1 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $\Pr(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. .......................... 93
Figure B.2  Polynomial MCC approximation ($K = 1$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 1 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. Pr($Y = y|X$) refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$.

Figure B.3  Deep learning IPP approximation ($K = 20$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 2 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. Pr($Y = y|X$) refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$.

Figure B.4  Polynomial MCC approximation ($K = 1$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 2 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. Pr($Y = y|X$) refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 
Figure B.5 Deep learning IPP approximation ($K = 20$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 3 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $\Pr(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 

Figure B.6 Polynomial MCC approximation ($K = 1$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with with observations from fold 3 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $\Pr(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 

Figure B.7 Deep learning IPP approximation ($K = 20$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 4 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $\Pr(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 
Figure B.8  Polynomial MCC approximation ($K = 1$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 4 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $Pr(Y = y | X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y | X$ refers to the conditional response value $Y$ given the covariate $X$.

Figure B.9  Deep learning IPP approximation ($K = 20$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 5 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $Pr(Y = y | X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y | X$ refers to the conditional response value $Y$ given the covariate $X$.

Figure B.10  Polynomial MCC approximation ($K = 1$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 5 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $Pr(Y = y | X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y | X$ refers to the conditional response value $Y$ given the covariate $X$. 

xv
Figure C.1 The predicted PDF plot for a microbial fungal sample from North Carolina using the Radial Basis Deep Learning method with the Bayesian Discriminant Analysis predicted location and true location overlaid. The plotted locations contain the largest PDF values and account for around 95% of the total PDF probability mass. The microbial fungal sample was purposely selected as an example of a location densely surrounded by other microbial fungal samples. The predicted PDF locations are on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 30 nodes for the hidden and output layers, 5 controls, an initial learning rate of 3, and a ridge penalty of 0.0001.

Figure C.2 The predicted PDF plot for a microbial fungal sample from North Carolina using the Coordinate Deep Learning method with the Bayesian Discriminant Analysis predicted location and true location overlaid. The plotted locations contain the largest PDF values and account for around 95% of the total PDF probability mass. The microbial fungal sample was purposely selected as an example of a location densely surrounded by other microbial fungal samples. The predicted PDF locations are on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 50 nodes feeding into the hidden and output layers, 5 controls, an initial learning rate of 4, and a ridge penalty of 0.00005.

Figure C.3 The predicted PDF plot for a microbial fungal sample from Idaho using the Radial Basis Deep Learning method with the Bayesian Discriminant Analysis predicted location and true location overlaid. The plotted locations contain the largest PDF values and account for around 95% of the total PDF probability mass. The microbial fungal sample was purposely selected as an example of a location sparsely surrounded by other microbial fungal samples. The predicted PDF locations are on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 30 nodes for the hidden and output layers, 5 controls, an initial learning rate of 3, and a ridge penalty of 0.0001.
Figure C.4  The predicted PDF plot for a microbial fungal sample from Idaho using the Coordinate Deep Learning method with the Bayesian Discriminant Analysis predicted location and true location overlaid. The plotted locations contain the largest PDF values and account for around 95% of the total PDF probability mass. The microbial fungal sample was purposely selected as an example of a location sparsely surrounded by other microbial fungal samples. The predicted PDF locations are on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 50 nodes for the hidden and output layers, 5 controls, an initial learning rate of 4, and a ridge penalty of 0.00005.
Chapter 1

Introduction

Statistical analysis is a fundamental tool in environmental and ecological research. Spatial statistical analysis is particularly important to the study of ecological and environmental phenomena as these processes are often influenced by their geographic positions, directions, and/or distances from other spatially referenced objects. Advances in these areas have provided insight into natural processes such as the distributions of species, the intensities of tropical cyclones, or the origins of crime scene materials. In this dissertation, we consider novel spatial statistical methods to address emerging research areas of environmental and ecological interest.

In Chapter 2, we propose a method designed to improve covariate estimation for occupancy modeling by utilizing opportunistic data gathered by citizen scientists. In Chapters 3 and 4, we propose a new conditional density estimation approach and then extend this approach to model point pattern data with spatial dependence. The methods in Chapters 3 and 4 are supported by a point process modeling framework. The main concepts needed to understand Chapters 2, 3, and 4 are reviewed in the remainder of this chapter. We begin by explaining the ecological concept of occupancy, the associated data types, and the occupancy modeling approach. We continue with a review of conditional density estimation, point pattern data, and point process modeling. We conclude by outlining the structure of the rest of the dissertation.

1.1 Estimating the drivers of species distributions with opportunistic data using mediation analysis

Two key ecological descriptors of the status of a species are abundance and occupancy. For a randomly selected site in a given area of interest, abundance refers to a given
species’ population size in the site whereas occupancy refers to a given species’ probability of inhabiting the site. Abundance is a more informative descriptor for determining the distribution of a species compared to occupancy, however the additional insight can come at a greatly increased resource cost to obtain appropriate and accurate count data. Many species distribution modeling strategies have been developed with the intention to estimate species occupancy as a proxy for the overall species distribution rather than abundance due to this limitation.

1.1.1 Presence-absence and presence-only data for occupancy modeling

Presence-absence data designs naturally allow for inference about a species’ probability of occupancy as they are well suited for use in occupancy modeling. In a presence-absence design, the surveyors select random sites within an area of interest and record whether or not they detected the target species. A presence is recorded if they detect the species, otherwise they record an absence. A surveyor may see the species directly to confirm presence, or they may find secondary indicators of the species’ recent presence. The absence may be due to the species not inhabiting the region, or the species may have gone undetected by the surveyors while present. Repeated surveys for presence or absence at each selected site are conducted successively which allows occupancy models to account for the detectability of the species.

Presence-only data are another data type potentially used for occupancy. These data give no information on where the target species is absent within the region of interest, only where it is present. The utility of presence-only data in estimating abundance and occupancy is still being explored. Some researchers argue that occupancy is not identifiable using presence-only data without strong assumptions (Ward et al. 2009). Royle et al. (2012) refute this interpretation and argues that this view is due to a specific model often used for analysis and not an inherent feature of the data type. Hastie and Fithian (2013) argue that this refutation is of little practical use as the amount of data needed to draw conclusions would be ”awfully large”. Additionally, the deficiencies of presence-only designs might be exacerbated in certain ecological scenarios. Brotons et al. (2004) found that wider-ranging species were more sensitive to absence data, suggesting recording absences might be extra important for modeling these types of species distributions. Still, these data can be significantly less resource-intensive to collect than presence-absence data and are often used as the data format for citizen scientist (or opportunistic) datasets.
1.1.2 Occupancy modeling

MacKenzie et al. (2002)’s closed, single season occupancy model is the basis of modern spatial occupancy modeling when imperfect detection is assumed (referred to as the "static model" hereafter). The static model is dependent on the assumptions that (1) the occupancy status of a site is closed or unchanging during the survey period, (2) the probability of occupancy is equal across all sites, (3) species detection in each survey at a given site is independent, and (4) the detection histories at each location are independent. Additionally, this model assumes there are no false positive detections where the species is recorded as present when it is really absent. The model allows for incorporation of covariates for occupancy or detection estimation.

Royle and Link (2006) extend the static model to allow for false positive detection errors. Limitations of this method for dealing with false positives that affect parameter identifiability and estimation have been documented (Fitzpatrick et al. 2009). For instance, estimates of the parameters in the model can be correlated, and false positive vs true positive heterogeneity in detection probabilities cannot be distinguished. Miller et al. (2011) address these issues for two or more detection types. They also present a model for two (or more) detection methods where the surveyed sites are visited again using the second method. Another extension of the static model is for heterogeneous detection that cannot all be captured by covariate information (Royle 2006). Royle and Nichols (2003) suggest modeling detection as a function of site abundance. Stanley and Royle (2005) further recommend a method to account for different detection period lengths.

MacKenzie et al. (2003) develop another major occupancy model by explicitly extending the static model to a spatiotemporal context. This dynamic model allows for an open population across multiple seasons. Changes in occupancy are assumed to be a first-order Markovian process, or dependent only on the previous season’s occupancy status. This model depends on the assumptions that (1) occupancy status does not change within seasons, but may change between seasons, (2) all parameters are constant or have modeled heterogeneity across sites in a season, (3) species detection in each survey at a given site is independent, and (4) the detection histories at each location are independent. Both static and dynamic occupancy modeling can be extended to cases where occupancy has more than two states, assumed to be ordinal in nature (MacKenzie et al. 2009). Various reparametrizations of occupancy and detection have also been proposed (Nichols et al. 2007, Royle and Link 2005).

Nichols et al. (2008) relax the closure-within-period assumption of occupancy models by involving multiple detection methods to allow for occupancy estimation at multiple
scales. These multiple method studies might be used when various species with the same genus or various sizes of the same species are studied separately with distinct designs and detection methods. The multiple scales for occupancy estimation allow a modification of the assumption of a closed population, as the assumption is made for the larger scale but members of the target species may not be absent in the immediate sample site for every survey. Before this, Weir et al. (2005) modeled a function of a day of year variable for seasonal variation in detection probability to account for changes within a closed season. Previous work with multiple detection methods used separate analyses for each detection method (Bailey et al. 2004). Nichols et al. (2008) develop a model that could use site data from all related detection methods to improve inference about study-specific detection probabilities and estimate distinct occupancy probabilities at different surveys. Kendall et al. (2013) more explicitly relax the closure assumption by allowing for staggered arrival and departure times.

Spatial autocorrelation (SAC) can overstate the precision of estimates if unaccounted for or produce biased estimates and overinflation when incorrectly accounted for, and is an important consideration in occupancy modeling. Additionally, autocorrelation can be leveraged to decrease the number of surveys performed at a site by borrowing information from other sites through the spatial relationship (Johnson et al. 2013). A Bayesian framework is convenient to model SAC as it can naturally model spatial random effects. Mattsson et al. (2013) incorporate SAC into occupancy modeling by mixing the static model with autologistic methods for spatial covariates. This method, along with accounting for heterogenous detection and species-specific variables, improved occupancy estimates compared to not accounting for SAC. However, Mattsson et al. (2013) do note that landcover effects were more important than SAC for modeling species distributions in their study. Johnson et al. (2013) extend the incorporation of SAC to larger datasets using spatially correlated random effects. There had been some prior studies which accounted for SAC into occupancy modeling beforehand, but most were limited to smaller datasets due to computational time demands (Sargeant et al. 2005, Magoun et al. 2007). They replace the standard logit link function with a mathematically similar but computationally quicker and more flexible probit link function. Bled et al. (2011) consider SAC in colonization processes for the dynamic model.

1.1.3 Opportunistic data for occupancy modeling

Occupancy analysis commonly uses designed-survey data obtained by a professionally trained team of surveyors. However, the massive increase in the availability and abun-
dance of citizen-collected opportunistic data in recent decades has led to investigations of its quality and capacity for integration with occupancy modeling.

Opportunistic data has helped uncover previously unavailable insight into ecological processes. Volunteers can report sightings from other countries and/or easily collect data from their private property, often producing thorough spatial and temporal data coverage (Dickinson et al. 2010). Dunn and Winkler (1999) use opportunistic avian data to show large-scale trends of birds breeding earlier in the spring due to permanent increases in temperatures. Depending on the rigidity of the data input process, public participation data can be recorded without a specific research objective in mind. Dickinson et al. (2010) explain how opportunistic data collected in this manner has utility as a monitor for noticing unanticipated events. For example, Dhondt et al. (1998) found that house finches with conjunctivitis were being observed with more frequency in Washington D.C., and were then able to track the spread of this disease in the population. Opportunistic data analysis can help locate rare species, even in cases where they were previously thought to be extinct (Losey et al. 2007). Hochachka et al. (1999) used Project FeederWatch volunteer data to identify boom-and-bust bird migration trends that could have been otherwise difficult to uncover.

There are significant quality concerns with opportunistic datasets. Volunteers are often untrained and may have a comparatively worse ability to detect species than a professional researcher. Jiguet (2009) finds that observers become better at detecting birds as they gain more experience. If volunteers make classification decisions about the observation, they could incorrectly identify certain features. Delaney et al. (2008) note that education has an effect on volunteers’ abilities to determine the species and gender of crab. Volunteers can submit purposefully inaccurate data and more frequently falsely detect absent species. With no protocols in place, there is no guarantee that the data are spatially or temporally representative. Highly biased samples that overrepresent rare species and underrepresent common species are possible (Dickinson et al. 2010). These issues have caused some scientists to have concerns over the validity of the data gathered (Delaney et al. 2008). Dickinson et al. (2010) suggest that opportunistic data are better suited for surveillance and exploratory analysis than targeted research. Some scientists believe this is a smarter approach for all ecological analysis due to the complexity of ecological systems masking many processes (Bonney et al. 2009).

Researchers have recently attempted to incorporate opportunistic data into an occupancy modeling framework to extend its use past monitoring purposes. By controlling for detection bias in opportunistic data, occupancy modeling may control for observation and reporting bias in the data as well (Strien et al.). Dorazio (2014) develops a site
occupancy framework that incorporates opportunistic presence-only data with designed-survey count data to estimate abundance by viewing the presence-only data as from a thinned spatially inhomogeneous Poisson process (Warton et al. 2010). Renner et al. (2015) develop a similar approach for combining these two data sources for differing multiple species and explicitly outline the extension to occupancy estimation using either count or presence-absence data. Pacifici et al. (2017) extend this thinking to propose methods of integrating opportunistic data and designed-survey data while explicitly accounting for spatial autocorrelation in occupancy trends. Pacifici et al. (2017) propose treating opportunistic data as a covariate or to influence random effects through correlation with the designed-survey data.

1.2 Nonparametric conditional density estimation in a deep learning framework for short-term forecasting

In Chapters 3 and 4 we develop methods which estimate the full conditional density of the response variable given the covariates, referred to as conditional density estimation. Conditional density estimation is a useful approach to understanding the behavior of the response variable because it provides information on distinguishing features such as skewness, kurtosis, or multimodality. This can be advantageous compared to estimation methods that assume a parametric conditional response distribution if the assumed distribution is misspecified. There are multiple different conditional density estimation approaches that have been theorized, with some of the more prominent approaches briefly reviewed below.

1.2.1 Kernel estimation of the joint and marginal densities

Perhaps the oldest conditional response distribution estimation method is to instead consider the joint distribution of the response and covariates divided by the marginal distribution of the response. A natural method for obtaining these distributional estimates is through non-parametric kernel density estimation, first presented by Rosenblatt (1969). Hyndman et al. (1996) modify the standard kernel density estimator to obtain a smoother with better bias properties. Hall et al. (1999) propose to use an adjusted Nadaraya-Watson estimator for the kernel estimation. Ichimura and Fukuda (2010) develop an algorithm to speed up least squares cross-validation bandwidth selection for
low-dimensional data with many observations. Kernel density estimation methods suffer from intractability when the covariate dimension increases, with much research dedicated to addressing this issue. Hall et al. (2004) introduce a cross-validation method to eliminate irrelevant covariates while optimally smoothing relevant covariates. Hall et al. (2005) and Fan et al. (2009) each suggest approximating the conditional density function via a lower-dimensional distribution of the response conditional on a unit vector multiplied by the covariates. Each of these two papers proposes different criterion for determining the unit vector.

1.2.2 Bayesian nonparametric mixture models for conditional density estimation

Bayesian nonparametric modeling is a popular approach for conditional density estimation. This approach has the capacity to represent any distribution for a sufficient number of mixing distributions. We discuss some of the primary sets of approaches under this methodological umbrella.

Finite mixture models

Finite mixture models are a conditional density estimation technique which considers the conditional target distribution to be a weighted mixture of a set number of parametric distributions. This approach directly models the desired conditional distribution rather than the joint and marginal distributions as with the aforementioned kernel density estimation approaches. Escobar and West (1995) are among the first to describe conditional density estimation in these terms, describing Bayesian inference on mixture models generally but especially using Gaussian distribution mixtures. Gaussian mixture models are a popular choice for this approach to conditional density estimation (Gilardi et al. 2002, Song et al. 2004, Rojas et al. 2005, Fahey et al. 2007). Finite mixture models involve stronger assumptions on the underlying distribution than the kernel density estimation methods, but also require fewer data observations and have better interpretability. This approach has the capacity to represent any distribution as long as the number of distributions being mixed is sufficiently large, although the determination of sufficiently large also depends on the complexity of the distribution being estimated. Peng et al. (1996) propose influencing the mixing proportions and component densities through the covariates and using a Bayesian Markov Chain Monte Carlo (MCMC). Wood et al. (2002) and Geweke and Keane (2007) propose different ways for the covariates to influence the mixing proportions that lead to more efficient Bayesian MCMC implementation algorithms.
FMMs require certain parameter specifications such as the mixing proportion values or number of densities that can be non-trivial to determine.

**Infinite mixture models**

Compared to finite mixture models, infinite mixture models use a countably infinite number of mixing distributions to estimate the target density. An obvious utilization of this method is to directly estimate the conditional density. The density estimation problem is framed as a nonparametric regression which utilizes an infinite-mixture model with an infinite set of mixture weights and a process mixing distribution prior that depends on the covariates. MacEachern (2000) discuss dependent Dirichlet processes (DDPs) as a method to model realizations of random Dirichlet processes that are dependent. The DDP prior is utilized for Bayesian nonparametric regression in various contexts (De Iorio et al. 2004, Dunson et al. 2007, Griffin and Steel 2010). Other process priors have also been considered. Jara and Hanson (2011) propose use of a Polya Tree prior model and induce dependence through different definitions of the splitting probabilities. Rodriguez and Dunson (2011) propose the use of a probit stick-breaking process prior. Tokdar et al. (2010) forego these priors and develop a model using logistic Gaussian processes and subspace projection. Ma (2012) use a two-stage extension of the Polya Tree prior which introduces randomization into the partitioning procedure and results in posterior conjugacy for quicker MCMC computation. Still, Bayesian non-parametric density estimation analysis can become computationally burdensome as data complexity increases. Variable selection techniques have been proposed for certain process prior choices (Chung and Dunson 2009, Kundu and Dunson 2014).

An alternative infinite mixture model approach for conditional density estimation is to model the joint distribution of the covariates and response as an infinite sum of mixture distributions. This approach has commonality with the kernel density estimation approach in that the posterior joint and marginal densities are used to obtain the posterior conditional density. Müller et al. (1996) first posit a model using Dirichlet process mixtures of normal distributions. The implementation requires continuous variables and is only computationally feasible for a small number of covariates. Shahbaba and Neal (2009) extend the idea to enhance computational efficiency, allow for other variable types, and improve model flexibility. Other updates of this Bayesian approach using Dirichlet process mixture models to increase computational efficiency have also been formulated (Park and Dunson 2010, Taddy and Kottas 2010, Hannah et al. 2011). A disadvantage of this method is that it does not directly estimate the conditional density, and also can be slow in terms of computational performance as the dimensions of the problem increase.
1.2.3 Machine learning-based density estimation methods

One of the more broadly defined and accessible approaches to conditional density estimation involves the leveraging of well-known machine learning techniques toward the goal of full conditional distribution prediction. Machine learning techniques are often developed or extended to account for data scenarios with a large number of covariates, a useful advantage for many conditional distribution estimation applications. For instance, Sugiyama et al. (2010) develop a method to directly estimate the ratio of the joint to the marginal distributions through a least squares approach, which Shiga et al. (2015) extend to integrate feature selection capabilities. Efroymovich (2010) use an orthogonal (and in particular Fourier) series density estimator with shrinkage coefficients to reduce the dimension of the problem. Izbicki and Lee (2016) adopts the orthogonal series density estimator framework with a basis that adapts to the geometric features of the data. Izbicki et al. (2017) exploit the orthogonal series property that the expansion coefficients can be estimated through sample means in order to obtain them through a series of univariate regressions. Dalmasso et al. (2020) modify this approach by simultaneously estimating the expansion coefficients through a deep learning approach.

A foundational machine learning-based approach to conditional density estimation is quantile regression forests (QRF), proposed by Meinshausen (2006). Meinshausen explains that by noting all observations in each leaf of a regression tree in a random forest, the random forest can be used to calculate the full conditional distribution as weighted sum of sample quantiles across trees. QRF extensions have been developed including for bias correction or high-dimensional data (Tung et al. 2014). Hothorn and Zeileis (2017) propose a forest of ”transformation trees” which are responsive to changes in the distributional variance and not just mean. Their transformation forest method outperforms QRF in their simulation but can be computationally burdensome. Pospisil and Lee (2018) proposes a similar approach to QRF except with an alternate loss function suited for conditional density estimation.

Deep distribution regression (DDR) is a recent, state-of-the-art deep network learning-based conditional distribution technique (Li et al. 2019). Li et al. (2019) use cutpoints to discretize the response space and apply a multi-class classification method (such as a neural network) on the resulting bins. Li et al. (2019) also give an approach which accounts for bin ordering by applying a binary classification model for each cutpoint and jointly estimating the conditional cumulative distribution function. The classification model can be chosen as a deep learning neural network which allows for complex conditional distributions to be accurately modeled.
1.3 Geolocation for spatial point pattern data in a deep learning framework

In Chapter 4 we extend the conditional density estimation technique proposed in Chapter 3 to apply to spatial point pattern data for a forensic geolocation analysis using ecological fungal data. To do so, we consider our method in the context of point process models. Point process models are the major research area for point pattern analysis. Point pattern data is considered to be a realization of a latent governing process in the point process model framework. The goal for point process model analysis is to properly model the associated intensity function which controls the expected number of point occurrences in the given subdomain. These models are superior to summary statistic computation methods for many analyses goals given their flexibility to account for covariate effects. There are many point process model classes, we focus on three of the most common.

1.3.1 Poisson process models

Poisson process models are arguably the fundamental point process model for point pattern data analysis. A Poisson process assumes that the number of points in any subregion of the spatial domain is Poisson distributed with an intensity function to govern the expected rate of points. A key property of the Poisson process is that the intensity function only contains fixed effects, which ensures process realization points are independent from each other (Møller and Waagepetersen 2003). Generally, these models can be classified into homogenous and inhomogenous Poisson processes. We briefly discuss properties and computational considerations for both.

The homogenous Poisson process (HPP) is the most basic point processes, distinguished by the intensity function being constant across the spatial domain (as well as the aforementioned independence between realized points) (Diggle 2013, Cressie 2015). The HPP model is also known as the complete spatial randomness (CSR) model, and is often used as a null reference model to determine the utility of using more complex process models.

The inhomogenous Poisson process (IPP) differs from the HPP by allowing the intensity function to vary across the spatial domain (Diggle 2013, Cressie 2015). The IPP is a common point process modeling option because the form of the IPP is conducive to maximum likelihood estimation of its intensity parameters. The main difficulty with the IPP model is that the integral in the model denominator can be intractable for a spatially-varying intensity function, but an approximation of the integral can be made
(Berman and Turner 1992, Fithian and Hastie 2013). In the special case where the intensity function has a log-linear structure, a logistic regression-based technique can be used (Baddeley et al. 2015). In fact, Fithian and Hastie (2013) show that the log-linear IPP model and the naive logistic regression (as well as the maximum entropy model) produce equivalent inference if the IPP is correctly specified. For these reasons, the IPP (as well as the HPP) are the most computationally convenient point process models.

### 1.3.2 Cox process models

Another prominent PPM is the Cox process which generalizes the IPP by allowing for a stochastic intensity function (Moller and Waagepetersen 2003). Unlike IPPs, Cox processes model spatial clustering between points through random effects. The log Gaussian Cox process (LGCP) is a commonly used Cox process where the log of the intensity function is defined as a Gaussian process (Møller et al. 1998).

The random effects in Cox process models make ML estimation difficult. Some efficient non-Bayesian methods have been developed for Cox process model evaluation including a composite likelihood approach based on the second-order intensity function of the Cox process and a weighted estimating equation estimator (Guan 2006, Guan and Shen 2010). However, the majority of research in this area has focused on Bayesian approaches for evaluating the Cox process model. Adams et al. (2009) propose a sigmoidal Gaussian Cox process prior for the intensity function with a thinning mechanism to allow for an exact Markov Chain Monte Carlo (MCMC) approach, later adapted to structured point processes by Gunter et al. (2014). Kirichenko and Van Zanten (2015) determine a tuning procedure for this approach to improve estimation of the smoothness of the intensity and achieve optimal convergence. Lloyd et al. (2015) propose a quadratic Gaussian Cox process prior with a variational Bayesian estimation technique which scales to large data well, further modified by John and Hensman (2018). Flaxman et al. (2017) present a frequentist approach using the same square root link function as Lloyd et al. (2015). Kottas and Sansó (2007) use a mixture kernel prior on the intensity with a multi-step inference scheme involving a Gibbs sampler for implementation. Another popular Bayesian approach for models with latent Gaussian structures is integrated nested Laplace approximation (INLA) (Rue et al. 2009). This method is generally faster than MCMC although possibly less accurate (Taylor and Diggle 2014). Beyond INLA, a variety of Bayesian techniques are suitable for evaluating LGCPs such as the Metropolis-Adjusted Langevin algorithm (MALA), Gaussian predictive process approximation, pseudo-marginal MCMC, elliptical slice sampling, and Hamiltonian Monte Carlo.

1.3.3 Gibbs process models

Gibbs process models are a third well known process model which can account for spatial repulsion between points through the explicit modeling of interaction terms (Moller and Waagepetersen 2003). Gibbs process models are difficult to fit with ML estimation just like Cox process models, because the Gibbs likelihood contains an intractable normalizing constant. A frequentist approach is to use maximum Poisson pseudo-likelihood estimation although this method has difficulties in terms of accuracy (Besag 1978, Huang and Ogata 1999, Baddeley and Turner 2000). Baddeley (2017) later extends this idea with a local version of the pseudo-likelihood approach involving a local weighting kernel. Baddeley et al. (2014) develop a computationally superior and less biased evaluation method using logistic regression score functions instead of the pseudo-likelihood score functions, later adapted for a variational Bayesian context by Rajala (2014). These methods work best for a moderate number of covariates, so high dimensional Gibbs process models benefit from feature selection methods (Baddeley et al. 2015, Rajala et al. 2017, Ba and Coeurjolly 2020). For an MCMC approach, the Gibbs process model normalizing constant produces what Murray et al. (2012) refer to as a doubly intractable distribution which makes the Metropolis-Hastings ratio difficult to calculate. Møller et al. (2006) present an auxiliary variable-based solution for MCMC, later modified into a simpler exchange algorithm by Murray et al. (2012). Both of these algorithms require perfect simulation from the target distribution to cancel out the normalizing constant, which can be computationally difficult or impossible. The double sampling algorithm of Liang (2010) addresses this issue and approximates the exchange algorithm without using perfect sampling. Shirota and Gelfand (2017) develop an approximate Bayesian computation (ABC) technique for Gibbs process models which allows for easier simulation of point process realizations and permits use of parallelization to improve computational speed.
1.4 Overview

The rest of the dissertation is organized in the following way. In Chapter 2, we propose a method to improve covariate estimation for spatially distributed presence-absence data via mediation analysis. Our method is compared to an analogous method which does not employ mediation analysis in a simulation and case study. We end the chapter discussing further thoughts regarding the proposed method. In Chapter 3, we develop a nonparametric conditional density estimation method leveraging machine learning models through a logistic transformation. We exploit our method’s relationship with an inhomogenous Poisson process model to propose a computationally convenient approximation. Our method is compared with other machine learning-based conditional density estimation techniques in a simulation and case study. We end the chapter discussing areas for further exploration regarding the proposed method. In Chapter 4, the conditional density estimation method developed for Chapter 3 is augmented to be compatible with spatial point pattern data for the purposes of geolocation analysis. Model extensions to account for spatial dependence between locations are discussed along with a computational approximation. Our method is compared to a standard geolocation method in a simulation and case study. We end the chapter discussing possible methodological modifications and challenges.
Chapter 2

Estimating The Drivers Of Species Distributions With Opportunistic Data Using Mediation Analysis

The contents of this chapter are essentially a reproduction of Huberman et al. (2020b), with minor changes made to suit the dissertation format.

2.1 Introduction

As discussed in Chapter 1, a primary ecological species distribution analysis tool is occupancy modeling, a method which estimates species occupancy from presence/absence data while accounting for imperfect species detection (MacKenzie et al. 2002). Occupancy modeling and analysis traditionally assumes designed-survey data collected by a trained surveyor team. Occupancy modeling data analysis is predicated on the surveyors following the given ecological design, although methodological extensions have been developed to accommodate assumption violations like false positive species presence recordings (Miller et al. 2011). This method of data collection can require significant effort, but typically results in a high quality dataset that correlates well with the underlying spatial process as long as a rigorous probabilistic sample design is followed. Recent technological advances have led to a massive increase in the availability and abundance of citizen-collected opportunistic data. The opportunistic data collection process is significantly less resource-intensive than traditional ecological designed-survey data collection, resulting in a spatiotemporal data coverage that is likely unattainable through the traditional designs. However, the increased coverage of opportunistic data can come at the
price of reduced or uncertain data quality (Delaney et al. 2008, Jiguet 2009, Dickinson et al. 2010).

Effective integration of opportunistic data with existing methods for designed-survey data is essential for unlocking previously unavailable insights on species distributions. Multiple integration methods have been developed which assume opportunistic data to be in a presence-only format where absences are not recorded, briefly reviewed in Chapter 1 (Dorazio 2014, Renner et al. 2015).

Integration methods have also been developed for opportunistic data sources that can reasonably be considered presence-absence data (Pacifici et al. 2017, Miller et al. 2019). Pacifici et al. (2017) proposed several occupancy modeling techniques for data integration depending on the presumed quality of the data. In a scenario where the opportunistic data are believed to be of comparable quality to the designed-survey data, Pacifici et al. (2017) used a shared model where both data sources directly and equally influence occupancy estimation. If the opportunistic data is expected to be less correlated with occupancy than the designed-survey data, Pacifici et al. (2017) suggested a correlation model or a covariate model. In the correlation model, the opportunistic data influence occupancy estimation indirectly through a shared covariance matrix with the designed-survey data. In the covariate model, the opportunistic data are inserted into the design matrix as a covariate.

In this paper, we build on the covariate model described in Pacifici et al. (2017) and propose a method for improving covariate effect estimation with mediation (or path) analysis concepts. In mediation analysis, the relationship between a predictor and outcome variable is theorized to be partially or fully mediated by an intervening variable. The total effect of the predictor on the outcome can be decomposed into a direct effect of the predictor on the outcome and an indirect effect of the predictor on the mediating variable which itself affects the outcome. In our context, we consider the total effect of the ecological predictor of interest on the designed-survey outcome and allow for the opportunistic data to function as the mediator. We exploit the total effect decomposition into direct and indirect effects to incorporate additional opportunistic data. The direct and indirect effects are directly estimated and combined in place of the total effect estimate.

Two key features of our method are its relative speed and ability to incorporate large quantities of opportunistic data. One of the indirect effect components can be estimated by linearly regressing the opportunistic data on the predictor variable. This is computationally convenient compared to a spatial analysis that would require the inversion of a non-diagonal covariance matrix, explained more in Section 4.2 (Gelfand et al. 2010).
This also allows for the inclusion of additional opportunistic data from locations where no designed-survey data is available.

After a basic review of mediation analysis, we expand on the details of our method and its desirable features. Following this, we highlight the utility of our method with a simulation study and then apply it to real world designed-survey and opportunistic datasets.

2.2 Methods

2.2.1 Mediation analysis review

Mediation (or path) analysis is an extension of regression that attempts to quantify the causal relationship between a group of variables. For notational simplicity we will only consider one independent variable, however multiple covariates could be included if desired. In this framework, the expected effect of increasing independent variable $X$ by one unit on outcome variable $Y_1$ is referred to as the total effect (Hayes 2009). This effect might be direct, or it might occur indirectly through a mediator variable (Hayes 2009). Formally, the direct effect represents the change in $Y_1$ when $X$ is increased by one unit, holding the mediator value $Y_0$ fixed (Alwin and Hauser 1975). The indirect effect represents the change in $Y_1$ when $X$ is held fixed and $Y_0$ is adjusted by how much it is expected to change if $X$ had been increased by one unit (Alwin and Hauser 1975).

For concreteness, suppose we are interested in the effect of silvicultural practices on songbird distribution. We could straightforwardly regress the presence/absence of the songbird on different forest management regimes to get an estimate of the effect of silviculture. Alternatively, we might also consider an intervening variable such as forest type which mediates the relationship between forest management and songbird distribution. In other words, perhaps some of the total effect of silviculture on songbird distribution is due to its direct causal relationship, while the rest of the total effect is due to an indirect relationship where silviculture affects forest type which in turn affects songbird distribution. The decomposition of the total effect into direct and indirect effects is the key concept leveraged in our method.

In a linear regression context, the total effect can be decomposed into the sum of direct and indirect effects (Alwin and Hauser 1975). The regression equation for capturing the total effect of $X$ on $Y_1$ is

$$Y_1 = X\beta + \epsilon_1$$

(2.1)
where \( X = \begin{pmatrix} 1 & X \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 & \beta_1 \end{pmatrix}^T \), and \( \epsilon_1 \) represents a mean-zero normal error.

We take the expectation of \( Y_1 \) and obtain \( E[Y_1] = X\beta \), where \( \beta_1 \) represents the total effect of \( X \) on \( Y_1 \). \( \delta \) represents the effect of the mediating variable \( Y_0 \) on \( Y_1 \), given the effect of \( X \). The regression equations representing the total effect decomposition through the mediator are

\[
Y_0 = X\eta + \epsilon_2, \tag{2.2}
\]
\[
Y_1|Y_0 = X\tilde{\beta} + Y_0\delta + \epsilon_3 \tag{2.3}
\]

where \( \eta = \begin{pmatrix} \eta_0 & \eta_1 \end{pmatrix}^T \), \( \tilde{\beta} = \begin{pmatrix} \tilde{\beta}_0 & \tilde{\beta}_1 \end{pmatrix}^T \), and \( \epsilon_2 \) and \( \epsilon_3 \) are independent mean-zero normal errors. Using these equations, the expectation of \( Y_1 \) can be calculated as

\[
E[Y_1] = E[E[Y_1|Y_0]] = E[X\tilde{\beta} + Y_0\delta] = X\tilde{\beta} + X\eta\delta = X(\tilde{\beta} + \eta\delta). \tag{2.4}
\]

Thus, the total effect \( \beta \) can be equivalently written as \( \beta = \tilde{\beta} + \eta\delta \), where \( \bar{\beta}_1 + \eta_1\delta \) represents the decomposed total effect of \( X \) on \( Y_1 \). \( \tilde{\beta} \) and \( \eta\delta \) represent the direct and indirect effects, respectively (Alwin and Hauser 1975). Kenny (2018) notes that the equivalence of \( \beta \) and \( \tilde{\beta} + \eta\delta \) is guaranteed in a multiple regression context if the same cases are used across the analysis (as well as the same covariates). Kenny (2018) also states that the total effect and its decomposition may only be approximately equivalent in certain regression settings such as multilevel modeling or structural equation modeling with latent variables. For these scenarios, Kenny (2018) argues for computing \( \tilde{\beta}, \eta \), and \( \delta \) to estimate \( \beta = \bar{\beta} + \eta\delta \) over computing \( \beta \) directly.

Occupancy models are often constructed hierarchically to accommodate the assumed latent occurrence process (Royle and Dorazio 2008). Moreover, opportunistic data will likely be available at significantly more locations than designed-survey data are. In many cases, it would be a suboptimal use of available information to restrict these data to only locations where designed-survey data are available. For these reasons, the implication from Kenny (2018) is that estimating \( \beta = \bar{\beta} + \eta\delta \) through estimating \( \tilde{\beta}, \eta \), and \( \delta \) might be more useful for occupancy modeling inference than estimating \( \beta \) in an unmediated fashion as commonly done.
2.2.2 Mediation for spatial occupancy covariate modeling

Let \( s \) represent a spatial location. For this problem, \( X(s), Y_1(s), \) and \( Y_0(s) \) are functions of spatial location \( s \) that represent a habitat variable of interest, designed-survey species binary/count data, and opportunistic species binary/count data, respectively. \( X(s), Y_1(s) \) and \( Y_0(s) \) are the spatially analogous versions of \( X, Y_1 \) and \( Y_0 \) introduced in 2.2.1, with \( X(s) \) being a \( 1 \times p \) vector of spatial covariates. For simplicity, we assume \( p = 2 \), an intercept and predictor variable of interest. We note that the method outlined in this paper assumes \( Y_1(s) \) is the number of positive observations in a fixed number of trials and \( Y_0(s) \) follows a log-normal distribution. Opportunistic data is likely to be recorded as positive count data, which is incompatible with our normality assumption. However, a log transform of the opportunistic data would have valid support for a normal distribution (from \(-\infty \) to \( \infty \)) and be able to retain the linear relationship. Thus, our resolution is to assume the log transformed opportunistic data is normally distributed. The mediation method could be adapted and applied for other types of data as long as the linear relationships of these variables are retained.

Unmediated method

This method estimates the coefficient in Eq. 2.1 for the effect of \( X(s) \) on \( Y_1(s) \). We define a spatial occupancy model analogous to Eq. 2.1. A probit model is selected to reflect the assumed presence/absence format of \( Y_1(s) \). A location must contain both \( Y_1(s) \) and \( X(s) \) information to be eligible for analysis with the unmediated method’s spatial occupancy model.

The designed-survey data \( Y_1(s) \sim \text{Binomial}(N(s), pO(s)) \) is conditional on the number of samples \( N(s) \), detection probability \( p \), and \( O(s) \), the true occurrence state process. To model the true occurrence state, Johnson et al. (2013) suggest introducing \( Z(s) \), a latent, continuous, and normally distributed analog of \( O(s) \) so that

\[
O(s) = \begin{cases} 
1 & Z(s) > 0 \\
0 & Z(s) \leq 0.
\end{cases}
\]  

(2.5)

Note that the mediated independent effect calculations done in 2.2.1 assume a linear relationship between \( Y_1(s) \) and \( X(s) \). However, \( Y_1(s) \) is binomially distributed, making a direct linear link with \( X(s) \) impossible. Instead, this method considers the latent variable \( Z(s) \) in the place of \( Y_1(s) \) to maintain the structure in Eq. 2.1.
We then model $Z(s)$ as a conditional Gaussian process with $X(s)\beta$ and spatial random effect $\theta(s)$ entered into the mean function and $1 - \lambda$ as the variance. Recall from 2.2.1 that $\beta = (\beta_0, \beta_1)$ is a 2 by 1 coefficient vector with $\beta_0$ and $\beta_1$ representing the intercept and coefficient, respectively. $\theta(s)$ is modeled as a Gaussian process with $E[\theta(s)] = 0$ and $\text{Cov}[\theta(s_1), \theta(s_2)] = \lambda f(h)$ where $h = ||s_1 - s_2||$, $\lambda \in (0, 1)$, and $f$ is an exponential correlation function.

In summary, the unmediated model is

$$Y_1(s)|O(s), p \sim \text{Binomial}(N(s), pO(s)),$$

where $O(s) = 1(Z(s) > 0)$, \hspace{1cm} (2.6)

$$Z(s)|\theta(s), \lambda \sim \text{Normal}(\theta(s) + X(s)\beta, 1 - \lambda).$$ \hspace{1cm} (2.7)

The variance in the conditional $Z(s)|\theta(s)$ distribution is set as $1 - \lambda$ so that the variance of $Z(s)$ marginal of $\theta(s)$ is one and thus $\beta$ and $\lambda$ are identified. To understand this, first recall that a normal variable can be standardized by subtracting its mean and dividing by its standard deviation so that the transformed variable follows the standard normal distribution. Then, the probability that a site is occupied (i.e., $Z(s) > 0$) is

$$\Pr(Z(s) > 0) = \Pr\left(\frac{Z(s) - E[Z(s)]}{\sqrt{\text{Var}[Z(s)]}} > \frac{-E[Z(s)]}{\sqrt{\text{Var}[Z(s)]}}\right) = \Phi\left(\frac{X(s)\beta}{\sigma}\right)$$ \hspace{1cm} (2.8)

where $\sigma$ is the standard deviation of $Z(s)$ and $\Phi$ is the standard normal cumulative density function. The standard deviation will change the interpretation of $\beta$ if $\sigma \neq 1$, so we define the unmediated model to avoid this.

**Mediated method**

The mediated estimate is obtained in two separate steps, as we need to estimate coefficients from both Eq. 2.2 and 2.3. $Y_0(s)$ is taken to be the logarithm of the observed event rate (i.e., the number of observations divided by the observation effort). Let $\eta = (\eta_0, \eta_1)$ be the coefficient vector capturing the effect of $X(s)$ on $Y_0(s)$ with $\eta_0$ and $\eta_1$ representing the intercept and coefficient of interest, respectively. For the coefficient in Eq. 2.2, we model $Y_0(s)|X(s)$ as

$$Y_0(s)|X(s) \sim \text{Normal}(X(s)\eta, \sigma_0^2).$$ \hspace{1cm} (2.9)
The linear effect of $X(s)$ on $Y_0(s)$ is quantified by $\eta$, which can be estimated using ordinary least squares (OLS) estimation. The linear relationship between $Y_0(s)$ and $X(s)$ ensures that $\beta$ can be decomposed using $\eta$ similarly to the decomposition described in Eq. 2.4. In this step, we ignore spatial correlation in $Y_0(s)|X(s)$ for estimating $\eta$. This is justified as we assume the opportunistic dataset is large, so the least squares estimator of $\eta$ is unbiased and consistent (under regularity conditions) even in the presence of spatial dependence. We would expect most ecological opportunistic datasets to be sufficiently large for use with our method, as less resource-intensive data collection is a key feature of the data type.

The spatial occupancy model for the mediated method is similar to the model in 2.2.2 with adjustments made to incorporate $Y_0(s)$. Recall from 2.2.1 that $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1)$ is a 2 by 1 coefficient vector with $\tilde{\beta}_0$ and $\tilde{\beta}_1$ representing the intercept and coefficient, respectively. Then, the model is

$$Y_1(s)|O(s), p \sim \text{Binomial}(N(s), pO(s)), \text{ where } O(s) = \mathbb{1}(Z(s) > 0), \quad (2.10)$$

$$Z(s)|Y_0(s), \theta(s), \lambda \sim \text{Normal}(\theta(s) + X(s)\tilde{\beta} + Y_0(s)\delta, 1 - \lambda). \quad (2.11)$$

As in 2.2.2, $\theta(s)$ is a mean-zero Gaussian process with $\text{Cov}[\theta(s_1), \theta(s_2)] = \lambda f(h)$ where $h = ||s_1 - s_2||$ and $f$ is an exponential correlation function.

The direct and indirect effects for the mediated method are given by the $\tilde{\beta}$ and $\eta\delta$ vectors, respectively. Summing these two effects gives the total effect $\tilde{\beta} + \eta\delta$. This vector contains the mediated intercept and coefficient effects.

As mentioned in Section 4.1, the mediated method allows for the usage of more location data than the unmediated method. The spatial occupancy model for the mediated method requires locations with $Y_1(s)$, $Y_0(s)$, and $X(s)$ information, similar to the data requirements for the unmediated method’s spatial occupancy model. However, the mediated method regression step can also incorporate locations which only have $Y_0(s)$ and $X(s)$ information. The regression component thus allows for the use of locations that were not eligible for analysis in the unmediated method because they did not contain designed-survey information. Designed-survey data may be more difficult to obtain at many locations than opportunistic and covariate information. This allows the mediated method to potentially analyze significantly more data than the unmediated method. Furthermore, as explained in the Introduction, the linear regression step is a more convenient way of bringing in additional locations with opportunistic and covariate information than
fitting another spatial model. Recall that \( p \) gives the number of predictor variables and \( n \) gives the number of observations. The computational complexity for the linear regression step is \( O(p^2 n) \), compared to \( O(n^3) \) for a standard non-diagonal covariance matrix inversion. Consequently, the linear regression step is comparatively quicker as long as \( p \) is less than \( n \). For many applications, we would expect this to hold true.

Recall in Eq. 2.8 we explained that \( \text{Var}[Z(s)] \) needs to be a scalar to avoid identifiability issues. It can be shown for the mediated method that \( \text{Var}[Z(s)] = 1 + \delta^2 \sigma_0^2 \) using the law of total variance. Therefore, we must apply a correction to the intercept and coefficient estimates to remove the influence of \( \delta^2 \sigma_0^2 \) on \( \text{Var}[Z(s)] \). Let \( \hat{\sigma}_0^2 \) be the residual standard error estimate of \( \sigma_0^2 \) calculated in the linear regression step in Eq. 2.9. Then for the mediated method, we take Eq. 2.8 with \( \sigma = \sqrt{1 + \delta^2 \sigma_0^2} \) and obtain

\[
\text{Pr}(Z(s) > 0) = \Phi \left( \frac{\hat{\beta}_0 + \hat{\eta}_0 \delta}{\sqrt{1 + \delta^2 \hat{\sigma}_0^2}} + X(s) \frac{\hat{\beta}_1 + \hat{\eta}_1 \delta}{\sqrt{1 + \delta^2 \hat{\sigma}_0^2}} \right) = \Phi \left( \frac{\hat{\beta}_0^* + X(s) \hat{\beta}_1^*}{\sqrt{1 + \delta^2 \hat{\sigma}_0^2}} \right) \tag{2.12}
\]

where \( \hat{\beta}_0^* \) and \( \hat{\beta}_1^* \) represent the corrected intercept and coefficient estimates for the mediated method, respectively. These estimates can be corrected after the models have been evaluated and the parameter estimates have been obtained.

### 2.3 Simulation

We conduct a simulation study to compare the performance of the unmediated and mediated methods from 2.2.2. Data analysis in this project was conducted using a 2014 MacBook Pro with a 2.2 GHz Intel Core i7 processor and 16 GB 1600 MHz DDR3 Memory. The statistical software used was RStudio v1.1447 with R v3.5.0.

We spatially generated 1,000 data points on an evenly-spaced 40 × 25 grid with grid spacing one unit. Habitat data \( X(s) \) are generated for these locations as realizations of a mean-zero Gaussian process with a variance equal to one and an exponential correlation function with range equal to one unit so that correlation between orthogonally adjacent observations is \( e^{-1} \approx 0.368 \). Then, designed-survey and opportunistic data are generated
as

\[
Y_1(s)|O(s), p \sim \text{Binomial}(N(s), pO(s)), \quad \text{where } O(s) = 1(Z(s) > 0) \tag{2.13}
\]

\[
Y_0(s)|\theta(s) \sim \text{Normal}(a + b\theta(s), \sigma^2)
\]

\[
Z(s)|\theta(s), \lambda \sim \text{Normal}(\theta(s) + X(s)\beta, 1 - \lambda)
\]

\[
\theta \sim \text{MVN}(0, \lambda \Sigma(\kappa))
\]

where \(\theta = (\theta(s_1), \ldots, \theta(s_n))\) is the vector of random spatial effects, \(a\) and \(b\) are constants, \(\sigma^2 > 0\) is the conditional variance for the mediating variable, and all other terms are as previously defined. Recall from 2.2.2 that \(Y_0(s)\) is the logarithm of a count variable. The log transform is necessary to obtain the appropriate support bounds for a normal distribution. \(Y_0(s)\) is normally distributed to maintain the linear relationship between the variables needed for our mediation decomposition. The rest of the data generation process follows the model described in 2.2.2. For all data in the simulation, we set \(\beta_0 = 0, \beta_1 = 0.5, N(s) = 5, \lambda = 0.5, \nu = 0.5, a = 0,\) and \(b = 1.\) Let \(\kappa\) the spatial range parameter in the exponential correlation function, and recall that \(p\) represents the detection probability. We randomly generate \(X(s)\) and \(Y_1(s)\) data for each grid location for all possible combinations of the simulation settings

1. \(\rho \in \{0, 0.8\},\)
2. \(p \in \{0.25, 0.75\},\) and
3. \(\kappa \in \{0.5, 1.5\}.\)

The parameter \(\rho\) controls the dependence between the true latent process and the auxiliary data, with \(\rho = 0\) giving \(\text{cor}(Y_0(s), \theta(s)) \approx 0\) which makes \(Y_0(s)\) irrelevant, and \(\rho = 0.8\) giving \(\text{cor}(Y_0(s), \theta(s)) \approx 0.8.\) Figure 2.1 gives an example spatial representation of data generated with the model and settings \(p = 0.75\) and \(\kappa = 0.5.\) We repeat this data generation 300 times for each simulation setting. For each generated dataset, a random subset of either \(n = 50\) or \(n = 200\) grid locations are selected and the designed-survey data from the unselected sites is removed to create sixteen simulation settings. The opportunistic data is retained for all sites.

Bayesian MCMC techniques are implemented on the subsetted data to evaluate the spatial occupancy models for both methods and obtain the posterior distributions of the parameters. The additional linear regression step in the mediated method is evaluated
through OLS estimation, using all 1,000 data locations. The coefficient estimate for the mediated method is corrected as in 2.2.2. For details on the prior distributions, posterior updates, and other model details, refer to A.1. Code for the simulation and case study can be found at https://github.ncsu.edu/dbhuberm/MediatedOccupancyAnalysis.

The average effect estimate mean squared error (MSE), coverage, relative MSE, and relative MSE standard error are calculated for both methods using all 300 datasets for each setting. The MSE is calculated as

\[
\text{MSE} = \frac{\sum_{i=1}^{300} (\hat{\beta}_i - \beta_i)^2}{n}
\]  

(2.14)

where \(i\) indexes the 300 datasets used for the given setting, \(\beta\) represents the true effect of \(X(s)\) on \(Y_1(s)\), and \(\hat{\beta}\) represents the posterior-mean estimator for the unmediated or mediated method. Coverage is calculated as the percentage of simulations for which the true effect is included in the equal-tailed 95% posterior interval. The relative average MSE is calculated as the mediated average MSE divided by the unmediated average MSE (multiplied by a factor of 100). A relative average MSE below 100 indicates the

Figure 2.1: Plots of \(\theta(s)\) values, \(Z(s)\) values, \(Y_0(s)\) values when \(\text{cor}(Y_0(s), \theta(s)) \approx 0.8\), and \(Y_1(s)\) counts at each location for data generated with \(\kappa = 1.5\) and \(p = 0.75\).
mediated average MSE is smaller than the unmediated average MSE. The associated standard error is calculated using bootstrap sampling.

Table 2.1: The average MSE, 95% coverage, and relative MSE results for the unmediated and mediated methods on data generated to have a correlation of $\approx 0.8$ on average between $Y_0(s)$ and $\theta(s)$. $p$ and $\kappa$ refer to the detection and range parameter values used to generate the data, respectively. $n$ refers to the number of subsetted locations of the original 1,000 that retain $Y_1(s)$ data.

<table>
<thead>
<tr>
<th>Settings</th>
<th>Unmediated</th>
<th>Mediated</th>
<th>Relative Avg MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$ $\kappa$ $n$</td>
<td>Avg MSE</td>
<td>Coverage</td>
<td>Avg MSE</td>
</tr>
<tr>
<td>0.25 0.5 200</td>
<td>0.025</td>
<td>95.7%</td>
<td>0.023</td>
</tr>
<tr>
<td>0.25 0.5 50</td>
<td>0.120</td>
<td>95.3%</td>
<td>0.090</td>
</tr>
<tr>
<td>0.25 1.5 200</td>
<td>0.027</td>
<td>93.7%</td>
<td>0.021</td>
</tr>
<tr>
<td>0.25 1.5 50</td>
<td>0.121</td>
<td>93.0%</td>
<td>0.084</td>
</tr>
<tr>
<td>0.75 0.5 200</td>
<td>0.011</td>
<td>94.7%</td>
<td>0.010</td>
</tr>
<tr>
<td>0.75 0.5 50</td>
<td>0.048</td>
<td>95.7%</td>
<td>0.045</td>
</tr>
<tr>
<td>0.75 1.5 200</td>
<td>0.010</td>
<td>96.0%</td>
<td>0.009</td>
</tr>
<tr>
<td>0.75 1.5 50</td>
<td>0.047</td>
<td>95.7%</td>
<td>0.040</td>
</tr>
</tbody>
</table>

Table 2.1 compares model performance in terms of average MSE and coverage across settings when $\rho = 0.8$ and thus $\text{cor}(Y_0(s), \theta(s)) \approx 0.8$. Across all of these simulation settings, the mediated method produces relative MSEs below 100. All but two of these settings are more than a standard error below 100, showing the utility of the mediated method in estimating the effect of $X(s)$ when the mediating variable $Y_0(s)$ is strongly correlated with the true spatial process. Both methods result in around 95% coverage for all of the settings, and there do not appear to be any trends in coverage percentage related to the simulation setting parameters.

Altering $p$, $\kappa$ or $n$ all appear to influence the relative average MSEs. Generally, the relative average MSE decreases as $n$ is decreased from 200 to 50, indicating that the mediated method’s value increases when there are comparatively fewer locations with $Y_1(s)$ information. Holding $p$ and $n$ constant, setting $\kappa = 1.5$ decreases the relative average MSE compared to setting $\kappa = 0.5$. This is likely due to the increased utility of extra spatial information. Lastly, decreasing $p$ from 0.75 to 0.25 appears to improve the relative average MSE, though this did not occur when $\kappa = 0.5$ and $n = 200$.

Table 2.2 compares model performance in terms of average MSE and coverage across settings when $\rho = 0$ and thus $\text{cor}(Y_0(s), \theta(s)) \approx 0$. As expected, the mediated method does not improve compared to the unmediated method when the auxiliary data is uncor-
Table 2.2: The average MSE, 95% coverage, and relative MSE results for the unmediated and mediated methods on data generated to have a correlation of \( \approx 0 \) on average between \( Y_0(s) \) and \( \theta(s) \). \( p \) and \( \kappa \) refer the detection and range parameter values used to generate the data, respectively. \( n \) refers to the number of subsetted locations of the original 1,000 that retain \( Y_1(s) \) data.

<table>
<thead>
<tr>
<th>Settings</th>
<th>Unmediated</th>
<th>Mediated</th>
<th>Relative Avg MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>( \kappa )</td>
<td>( n )</td>
<td>( \text{Avg MSE} )</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>200</td>
<td>0.025</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
<td>50</td>
<td>0.120</td>
</tr>
<tr>
<td>0.25</td>
<td>1.5</td>
<td>200</td>
<td>0.027</td>
</tr>
<tr>
<td>0.25</td>
<td>1.5</td>
<td>50</td>
<td>0.121</td>
</tr>
<tr>
<td>0.75</td>
<td>0.5</td>
<td>200</td>
<td>0.011</td>
</tr>
<tr>
<td>0.75</td>
<td>0.5</td>
<td>50</td>
<td>0.048</td>
</tr>
<tr>
<td>0.75</td>
<td>1.5</td>
<td>200</td>
<td>0.010</td>
</tr>
<tr>
<td>0.75</td>
<td>1.5</td>
<td>50</td>
<td>0.047</td>
</tr>
</tbody>
</table>

lated with the true process. The relative average MSEs increase compared to the strong correlation case for all of these settings. All relative average MSE estimates are within one standard error of 100 except the setting with \( p = 0.25, \kappa = 0.5, \) and \( n = 200 \), which is at least one standard error above. However, these results confirm that the mediated method still produces valid statistical inference in this case (i.e., coverage near 95%). As with most other opportunistic data fusion methods, the quality of the supplementary dataset should be considered before deciding to use this method.

2.4 Case study

To provide an example of this method’s usage in ecological statistics, we consider the species distribution of the red-eyed vireo (\( Vireo olivaceus \)) across the continental United States from 2008 to 2013. The objective is to assess the relationship between deciduous forest land cover change and red-eyed vireo occupancy.

We chose the red-eyed vireo because the species exhibits a number of attributes shared by many North American landbirds. For example, they are migratory, insectivorous, and use a wide range of plant communities. The red-eyed vireo breeds in the eastern and northern United States and much of southern Canada and winters in South America. It is a common and widespread species that uses a wide variety of forest habitats to nest and glean insects from foliage. Counts of red-eyed vireos have been increasing by 0.8% annually since 1966 and by 1.0% annually since 2003, with some localized declines in
The designed-survey dataset used in our case study is presence-absence data from the North American Breeding Bird Survey (BBS) (Sauer et al. 2015). BBS data arises from observers who sample a set of pre-specified North American routes with 50 stops. At each stop on a given route, the observers survey the surrounding area. For this case study, the surveys at each stop within a route are considered independent. Each route is associated with a unique latitude and longitude, so that the presence/absence surveys from all 50 stops are considered to be surveys from the same location. For each survey route, a naive occupancy proportion between 0 and 1 can be calculated by dividing the number of individuals recorded by the number of stops.

The opportunistic dataset used in this study is from the Cornell Lab Of Ornithology’s eBird database (eBird) (Sullivan et al. 2009). Citizen scientists record and submit records from birds that they have seen or heard electronically. Here we only used recordings that came from a checklist provided by eBirds. Therefore, we can consider this data as presence/absence because if no positive detection was recorded for a species then it is safe to assume it was not detected. We could not make that determination without the extra user-provided information that all species on the full checklist were surveyed for, due to the possibility that the user did not survey the location for that species. In addition, we only used eBird data where the user indicated their effort (in hours). We standardized the eBird data by effort to give the average number of bird presence recordings per unit effort (see Eq. 2.15). Only stationary count eBird data from March through July were included to filter for data quality.

The land cover covariate data were obtained using the National Land Cover Database (NLCD) created by the Multi-Resolution Land Characteristics consortium (MRLC) (Vogelmann et al. 2001). The raster state datasets used in this paper were downloaded from https://viewer.nationalmap.gov/basic/. The NLCD contains high-resolution land cover image data across the United States of America for years 2006 and 2011. Each pixel represents a coordinate with land cover type information. The site land cover data for years 2008 through 2010 and 2012 through 2013 was obtained through linear interpolation, using the 2006 and 2011 site land cover data as endpoints. All grid sites contained land cover information.

We generate a 1 by 1 degree rectangular grid over the mainland United States spanning from -125 to -67 degrees longitude and 24.5 to 49.5 degrees latitude, restricted to grid locations containing land area. We further restrict the longitude and latitude range of this grid to match the longitude and latitude range of our BBS red-eyed vireo data, resulting in a grid spanning from -99 to -67 degrees longitude and 24.5 to 49.5 degrees.
latitude. This restriction was made because we focused on the Eastern half of the United States.

BBS, eBird, and covariate data are assigned to their respective grid cells and aggregated for analysis. The number of BBS species recordings and number of BBS survey stops are aggregated from all BBS routes within the given grid cell. The aggregated BBS variable is calculated as the proportion of species recordings to survey stops. For eBird, the species recordings and effort hours are aggregated within the given grid cell and the aggregated eBird variable is calculated as

\[ Y_0(s) = \log \left( \frac{C_0(s)}{N_0(s)} + 1 \right) \]  

(2.15)

where \( C_0(s) \) and \( N_0(s) \) represent the sum of reported eBird species presence recordings and the sum of reported eBird effort hours at grid site \( s \), respectively.

The aggregated landcover proportion represents the percentage of the given grid cell covered by deciduous forests. Due to data size issues for most states, the land cover raster data was inputted for each state individually and pre-aggregated at a small scale (squares of 50 horizontal and vertical pixels) in order to be feasible to convert to non-raster data for analysis. The raster data for Connecticut, District of Columbia, Delaware, Maryland, Massachusetts, New Hampshire, New Jersey, Rhode Island, and Vermont were small enough to be converted with no pre-aggregation. The data for each state are projected to latitude and longitude coordinates using the World Geodetic System (WGS) 84 projection. The latitudes and longitude extents for each state are determined to ensure overlap of locations from different state datasets is negligible. Finally, the land cover information is mapped to the specified grid. The aggregated land cover proportion is calculated as the number of pixels in the given grid cell representing deciduous forest land cover divided by the number of land cover pixels in the given grid cell overall.

There were 463 grid locations for each year with at least one red-eyed vireo occupancy data source, resulting in 2,778 total grid locations over 6 years. There was BBS information on red-eyed vireo occupancy in 1,884 or 68% of these grid locations, compared to eBird red-eyed vireo occupancy data in 2,597 or 93% of the grid locations. The size of the grid was selected to limit the number of grid locations which contained BBS but not eBird information. Our method assumes that all locations with BBS information also contain eBird information, so this information needs to be interpolated if not available. For a 1 by 1 degree grid, 60 of the locations with BBS information did not also have
eBird information, or around 2% of the total grid locations. We chose this grid size as a balance between number of observations and keeping this percentage small, because the number of locations with BBS but no eBird information only increased at higher resolutions. eBird proportion values for these 2% of grid locations with incomplete information were interpolated through simple Kriging, with Kriged eBird proportion values below 0 truncated to 0 (Cressie 1990). Figure 2.2 displays the aggregated naive (and cleaned) BBS red-eyed vireo occupancy, aggregated naive eBird red-eyed vireo occupancy, and aggregated deciduous forest land cover proportions data for 2008, respectively.

The mediated and unmediated method introduced in this paper were modified to account for possible temporal correlation across seasons (years). The spatial occupancy model was modified to be a spatiotemporal occupancy model as

\[
Y_{1t} | O_t, p \sim \text{Binomial}(N_t, O_t p), \text{ where } O_t = 1(Z_t > 0),
\]

\[
Z_t | \theta_t, \beta \sim \text{MVN}(\theta_t + X_t \beta, (1 - \lambda)I_n),
\]

\[
\theta_1 \sim \text{MVN}(0, \lambda \Sigma(\kappa)),
\]

\[
\theta_t | \theta_{t-1} \sim \text{MVN}(\rho \theta_{t-1}, (1 - \rho^2) \lambda \Sigma(\kappa)).
\]

where \(Y_0, \delta\) is additively inserted into the conditional mean for \(Z_t\) when using the mediated method, as in the single season version. For the mediated method, the linear regression step introduced in Eq. 2.9 is conducted on eBird and land cover data from all seasons simultaneously. The corrected intercept and coefficient estimates described in Eq. 2.12 are calculated using these all seasons parameter estimates. The formulas for these two steps remain unchanged.

As in the simulation, the spatiotemporal models are evaluated using MCMC methods. For \(\rho\), a random sample is drawn from 1,000 evenly-spaced candidates between -1 and 1. The candidate probabilities are weighted by the associated standardized likelihood values. The other priors are similar to the priors given in A.1 but generalized for a multi-season model.

Table 2.3 gives the posterior mean and standard deviation for each parameter under the unmediated and mediated methods. The eBird red-eyed vireo coefficient in the mediated method was only 0.18 (with a standard deviation of 0.84), suggesting a weak relationship with the BBS red-eyed vireo data. The unmediated method estimate of the effect of deciduous forest land cover on red-eyed vireo occupancy is 2.34 (with a standard deviation of 0.30).
Figure 2.2: 1 by 1 degree aggregated grids of eBird red-eyed vireo occupancy data, BBS red-eyed vireo occupancy proportions, and deciduous forest land cover proportions for 2008. eBird data were restricted to information to the East of the westernmost BBS grid site. Deciduous forest land cover data were restricted to locations with eBird data. eBird data is aggregated within grid cells and years and calculated as the logged proportion of aggregated sightings divided by aggregated effort hours plus 1 as \( Y_0(s) = \log\left(\frac{C_0(s)}{N_0(s)} + 1\right) \). BBS data is also aggregated within grid cells and years and is divided by the corresponding number of aggregated survey stops as \( Y_1(s)/N_t(s) \).
Table 2.3: The posterior mean and standard deviation (SD) for each parameter under the unmediated and mediated methods for the red-eyed vireo analysis.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unmediated Method</th>
<th>Mediated Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Partial Sill</td>
<td>0.84</td>
<td>0.15</td>
</tr>
<tr>
<td>Range</td>
<td>0.88</td>
<td>0.25</td>
</tr>
<tr>
<td>Detection</td>
<td>0.25</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Temporal Correlation</td>
<td>0.96</td>
<td>0.02</td>
</tr>
<tr>
<td>Unmediated/Mediated Intercept</td>
<td>1.06</td>
<td>0.12</td>
</tr>
<tr>
<td>Unmediated/Mediated Landcover Coefficient</td>
<td>2.34</td>
<td>0.30</td>
</tr>
<tr>
<td>Mediated eBird Coefficient</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

For the mediated method, the linear regression coefficient estimate was $\hat{\eta}_1 \approx 0.01$ while the variance estimate was $\hat{\sigma}_0^2 \approx 0.0002$. Thus, the mediated method effect of deciduous forest land cover on red-eyed vireo occupancy is calculated as

$$
\frac{\hat{\beta}_1 + \hat{\eta}_1 \hat{\delta}}{\sqrt{1 + \hat{\delta}^2 \hat{\sigma}_0^2}} \approx \frac{2.29 + (0.01)(0.18)}{\sqrt{1 + (0.18)^2(0.0002)}}
$$

$$
\approx 2.29
$$

with a posterior standard deviation of 0.65. These are identical to the estimates of $\hat{\beta}_1$ and SD($\hat{\beta}_1$) given for the mediated method in Table 2.3. We expect that opportunistic data more highly correlated with the designed-survey data would have highlighted the mediated method’s capacity for insight. On the other hand, this result demonstrates the method’s conservativeness when using lower correlated opportunistic data. The mediated method point estimate of the deciduous forest land cover effect on red-eyed vireo occupancy is similar, with the main penalty for including lower correlated opportunistic data being a larger standard error. In cases where the opportunistic data has a low correlation with the designed-survey data, the supplementary incorporation of opportunistic data ensures the mediated estimate will not deviate too sharply from the unmediated estimate. If the opportunistic dataset is known to be well correlated with the designed-survey dataset and computational speed is not a concern, other more aggressive methods might be considered. Even then, many other methods would only allow for the incorporation of opportunistic data where designed-survey already exists, so this method might...
still be preferable.

2.5 Discussion

The mediated method proposed in this paper attempts to improve upon estimates of a predictor variable’s effect on species occupancy from designed-survey data. This method utilizes a secondary data source, likely opportunistic data, in a supporting capacity as a mediating variable. The simulation study showed that when the opportunistic data source is well correlated with the designed data source, our mediated method produces smaller MSEs and higher coverage relative to the unmediated method. The mediated and unmediated methods performed similarly when the opportunistic data source was not well correlated with the designed data source in the simulation study.

The mediated method did not provide significant improvement when applied to red-eyed vireo occupancy data compared to the unmediated method. Both methods indicated that deciduous forest land cover had an effect on red-eyed vireo occupancy. The two methods did not produce notably different estimates, however, the unmediated method obtained higher precision. This difference in precision is unsurprising. Every parameter estimator has some amount of associated variation (or uncertainty). The unmediated method estimates a single parameter, whereas the mediated method estimates a linear combination of multiple parameters. The mediated method estimator involves more sources of variation and consequently has a higher level of associated uncertainty. The mediated method result suggests the supplementary inclusion of the opportunistic data source buffers against poor estimation when the data is not well correlated. Furthermore, the mediated method is computationally convenient and constructed to allow for opportunistic data usage beyond the locations with designed-survey data. Therefore, we consider it a low risk, flexible, and relatively expeditious occupancy modeling option for multiple data sources.

Additionally, we introduce a spatiotemporal version of the method for the case study with an order 1 autoregressive (AR1) correlation structure. The data was highly temporally correlated, with the autocorrelation parameter being estimated close to one. Temporal variation was not made a focus in this paper, but it could be worth more thoroughly examining the effectiveness of the mediated method in a spatiotemporal context with various temporal correlation strengths.

Our proposed method can theoretically be applied to grid data or point data. However, the spatial occupancy model step in the mediated method requires that any location with designed-survey data must also have opportunistic and covariate data. This is less
likely if data are collected for points rather than a grid. We would not expect that these three data sources to be sampled at the exact same locations, in which case we would need to map them to a common grid to perform the mediated analysis. It is possible that the designed-survey and covariate information are obtained from the same data source at the same locations, in which case unmediated analysis on point data is feasible. Otherwise, the unmediated method would also require data gridding.

Regarding the grid resolution selection, we expect that the highest possible grid resolution is preferable as this results in the most data points. At high grid resolutions, it is likelier that some grid locations have BBS but not eBird information, in which case the eBird information must be interpolated. We would like to minimize the amount of interpolated data, a preference in conflict with using the highest possible grid resolution. An alternative approach could include reconciling the difference in spatial resolution (change of support; Pacifici et al. (2019)) before conducting mediation analysis, but this would require fitting a much more complicated model. For our case study, we selected a grid size that resulted in around 2% of locations requiring eBird interpolation. In general, we anticipate the theoretically superior spatiotemporal coverage of opportunistic data will allow for the selection of relatively high grid resolutions with minimal or no need to interpolate opportunistic data values. Still, it might be worthwhile to analyze results from different grid resolutions and come up with a rule of thumb for grid selection that balances these two concerns.

2.6 Acknowledgements

We thank the Cornell Lab of Ornithology for making available the eBird data and the US Geological Survey’s Patuxent Wildlife Research Center (along with cooperating researchers) for making available the North American Breeding Bird Survey data on the red-eyed vireo. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.
Chapter 3

Nonparametric Conditional Density Estimation In A Deep Learning Framework For Short-Term Forecasting

The contents of this chapter have been submitted to Springer’s Environmental and Ecological Statistics journal and are awaiting their review.

3.1 Introduction

Short-term forecasting of environmental processes has many applications including solar and wind power generation, ambient air pollution, and extreme weather events. In this paper, we combine numerical model output with statistical methods to forecast hurricane wind intensity. Rather than providing a single value as the point prediction, we model the entire uncertainty distribution of the response given the numerical model forecast. This conditional distribution regression provides a comprehensive assessment of uncertainty, including the forecast distribution’s spread, skewness and tail probabilities.

To provide a flexible prediction model, we incorporate supervised machine learning methods, which have become a popular tool for statistical analysis in the last few decades. Methods such as random forest regression, neural networks, and linear regression can be employed using state-of-the-art statistical software to clarify complicated relationships between covariates and target variables. Generally, machine learning predictive modeling has been developed for making point predictions such as the conditional mean or me-
Accompanying prediction interval techniques provide uncertainty quantification. This differs from conditional density estimation, a technique which estimates the full distribution of the target variable given the covariates. In some applications, conditional density estimation is preferred. For instance, an estimate of a tropical cyclone’s maximum wind speed conditional on the sea surface temperature can provide information not available from a conditional mean estimate. A certain sea surface temperature might result in a strongly positively skewed maximum wind speed distribution, giving a better idea of the worst case scenario under these conditions.

Various approaches have been developed to estimate the distribution of the target variable conditional on the covariates. One technique is to estimate the joint distribution of the target variable and covariates as well as the joint distribution of the covariates and divide the former by the latter. Kernel density estimation of these two densities is a common approach, first proposed by Rosenblatt (1969). Hyndman et al. (1996) modify the standard kernel density estimator to obtain a smoother with better bias properties. Hall et al. (1999) propose to use an adjusted Nadaraya-Watson estimator for the kernel estimation. These methods suffer from intractability when the covariate dimension increases. The proposed remedies for this issue have been modifications to reduce the covariate space or to develop a density estimator for high-dimensional data (Hall et al. 2004, 2005, Fan et al. 2009).

Bayesian nonparametric mixture modeling is another common conditional density estimation approach. Finite mixture models (FMMs) are a subset of mixture modeling techniques which consider the conditional target distribution to be a mixture of several parametric (often Gaussian) distributions (Escobar and West 1995, Gilardi et al. 2002, Song et al. 2004, Rojas et al. 2005, Fahey et al. 2007). Covariate effects can be introduced in either the mixing proportions and/or densities. Bayesian Markov Chain Monte Carlo (MCMC) methods are often used to fit these models (Peng et al. 1996, Wood et al. 2002, Geweke and Keane 2007). FMMs require certain parameter specifications such as the mixing proportion values or number of densities which can affect their overall inference capabilities.

Infinite mixture models are another common Bayesian nonparametric mixture modeling approach. One class of infinite mixture model techniques attempts to directly estimate the conditional density via an infinite set of mixture weights and a process mixing distribution prior dependent on the covariates. Dunson et al. (2007) develop a Bayesian density regression model using a local, covariate-weighted mixture of DP priors. Trippa et al. (2011) and Jara and Hanson (2011) propose use of a Polya Tree (PT) prior model and induce dependence through different definitions of the splitting probabilities. Tokdar
et al. (2010) forego these priors and develops a model using logistic Gaussian processes and subspace projection. Still, Bayesian non-parametric density estimation analysis can be computationally burdensome as data complexity increases, leading to some variable selection techniques being proposed (Chung and Dunson 2009, Kundu and Dunson 2014). Infinite mixture models for estimating the joint distribution of the response and covariates have also been proposed (Müller et al. 1996, Shahbaba and Neal 2009, Park and Dunson 2010, Taddy and Kottas 2010, Hannah et al. 2011). A disadvantage of this class of techniques is that it does not directly estimate the conditional density, and also can be slow in terms of computational performance as the dimensions of the problem increase.

Machine learning algorithms are another useful and arguably more accessible class of conditional density estimation methods. One approach is to use an orthogonal series density estimator that adapts to the geometric features of the data and reduces the dimension of the problem, with additional improvements later proposed via incorporation of regression and deep learning algorithms (Efromovich 2010, Izbicki and Lee 2016, Izbicki et al. 2017, Dalmasso et al. 2020). Meinshausen (2006) proposes a foundational method of quantile regression forest (QRF). By noting all observations in each leaf, a random forest can be used to calculate the full conditional distribution as a weighted sum of sample quantiles across trees. Multiple conditional density estimation methods using random forests to improve on QRF accuracy and/or speed have been developed (Tung et al. 2014, Hothorn and Zeileis 2017, Pospisil and Lee 2018). Recently, Li et al. (2019) proposed deep distribution regression (DDR) as a deep network learning-based conditional distribution technique. Li et al. (2019) use cutpoints to discretize the response space and apply a multi-class classification method (such as a neural network) on the resulting bins. Li et al. (2019) also give an approach which accounts for bin ordering by applying a binary classification model for each cutpoint and jointly estimating the conditional cumulative distribution function.

Similar to DDR, we consider a conditional density estimation approach that incorporates machine learning algorithms for our short-term tropical cyclone intensity forecasting problem. A logistic transformation is made on the model output layer to obtain an expression of the conditional density function. The flexibility of the model specification allows for algorithms such as polynomial regression or deep learning models to be used. Our method evaluates only a single set of model parameters and simultaneously estimates the full conditional distribution. This information sharing allows our method to forecast well when minimal data is available, and the relatively limited number of parameters needing to be estimated ensure computational speed for the polynomial regression model choice. The gradient calculation can quickly become intractable for complex model
choices, so we incorporate theory from ecological and epidemiological statistics. Fithian and Hastie (2013) review models that can be used to evaluate presence-only survey data, including the inhomogenous Poisson process (IPP) model. We adapt the IPP framework to our data setting to justify a discrete approximation of our forecasting method for computational purposes. We also justify a special case of this method through a matched case-control context to further increase computational efficiency (Jarner et al. 2002).

After a review of the method and some potential model choices, we discuss the computational considerations for its implementation. Following this, the methodological strengths and weaknesses of our method are explored with a simulation and a short-term forecasting application, with the takeaways and next steps summarized in a discussion section.

3.2 Methods

We are interested in approximating the conditional distribution of response variable \( Y \in \mathbb{R} \) given the covariate information \( X \in \mathbb{R}^p \), denoted \( h(y|X) \). Our method requires a lower and upper bound for the target variable, which we address through a transformation of the response variable onto the unit interval. Suppose we transform \( Y \) through a cumulative distribution function \( G \) as \( Z = G(Y|X) \in [0,1] \). Note that the transformation of \( Y \) into \( Z \) to be on the unit interval is not unique, we could instead determine an upper and lower interval bound for \( Y \) on its original scale.

In this section, we will outline our method for approximating the conditional distribution of the transformed response, \( f(z|X) \), however the conditional density of the original \( h(y|X) \) can be recovered applying the change of variable formula as

\[
  h(y|X) = f(G(y)|X) \left| \frac{\partial G}{\partial y} G(y) \right|.
\]

(3.1)

If \( f(z|X) \) is uniformly distributed, the resulting \( h(y|X) \) distribution will be governed by \( G \). In other words, \( G \) is the base predicted distribution family, as opposed to the uniform distribution if no transformation of \( Y \) is made.

3.2.1 Logistic transformation

Let \( q(z, X) \) be a smooth function over \( z \) and \( X \). The logistic transformation (e.g. Lenk (1988)) relates \( q(z, X) \) to \( f(z|X) \) as
\( f(z|X) = \frac{e^{q(z,X)}}{\int_0^1 e^{q(u,X)} du}. \) (3.2)

Since \( q(z, X) = A(z, X) + B(X) \) gives the same density as \( q(z, X) = A(z, X) \), the main effect terms for \( X \) are removed. As the support of \( q(z, X) \) is arbitrarily flexible, any smooth conditional probability density function \( f(z|X) \) can be modeled with this transformation. In practice, this integral may be intractable. Discrete approximation techniques are discussed in 3.3 after introducing potential model choices.

A smooth underlying \( q \) function allows for the simultaneous estimation of a single set of model parameters. A similar logistic transformation on an underlying model was used in Tokdar et al. (2012) to develop a simultaneous quantile regression estimation method. The information sharing inherent in this approach enabled estimation of multiple quantiles concurrently, improving on previous quantile regression estimation methods.

Another advantage of this method is its flexibility. The only required \( q \) function specification is smoothness, which allows for many non-parametric model possibilities. We consider two such models in this paper which draw from machine learning ideas, a polynomial regression model and a deep learning model. However, our method can easily be applied to other smooth model choices such as an additive model with splines.

### 3.2.2 Polynomial regression model

The Weierstrass Approximation Theorem states that for any continuous real-valued function on a closed interval, there exists a polynomial function that can approximate it arbitrarily well (Weierstrass 1885). The polynomial function is therefore a logical candidate for the smooth function in our method. Let \( B \) be an integer representing the largest polynomial power used for the centered \( Z \) values, with \( b \) representing the given polynomial power. Recall \( j = 1, \ldots, p \) represents the covariate. Also, let \( o = 1, \ldots, O \) index the polynomial degree associated with the covariate terms. We let \( o = 2 \) and give the second-order model as

\[
q(z, X) = \sum_{b=1}^{B} \left[ (z - .5)^b \xi_{b0} + \sum_{j=1}^{p} \sum_{a=1}^{2} (z - .5)^b X_j^a \xi_{bjo} + \sum_{j \neq k} (z - .5)^b X_j X_k \upsilon_{bl} \right] \tag{3.3}
\]

where \( \xi_{b0} \) represents the intercept, \( \xi_{bjo} \) represent the covariate coefficients, and \( \upsilon_{bl} \) represent the \( l \) interaction term coefficients. A higher order model follows this structure in the
obvious way. The terms are centered by subtracting 0.5 to reduce collinearity, and the main effects of $X$ are removed because they do not affect the conditional distribution.

### 3.2.3 Deep learning model

A deep learning model is another natural choice for the underlying smooth function. The universal approximation theorem states that a feed-forward artificial neural network with at least one hidden layer can approximate a continuous function on a compact space arbitrarily well (Hornik et al. 1989). We propose a deep learning model with an input layer, at least one hidden layer, and an output layer. One hidden layer is given here for notational simplicity, but additional layers could be added if desired. Let $\delta$, $\gamma$, and $\beta$ represent the output layer, hidden layer, and input layer parameters, respectively. Let $H$ and $I$ represent the output and hidden layer nodes, respectively. Lastly, let $r = 1, ..., R$ and $t = 1, ..., T$ index the number of neurons in the hidden and output layer, respectively. The model is

$$q(z, X) = \sum_{t=1}^{T} \delta_t f_A(H_t), \quad (3.4)$$

$$H_t = \gamma_{0t} + \sum_{r=1}^{R} \gamma_{tr} f_A(I_r), \quad (3.5)$$

$$I_r = \beta_{0r} + \beta_{1r} (z - 0.5) + \sum_{j=2}^{p+1} \beta_{jr} X_j, \quad (3.6)$$

where $f_A$ is an activation function. Exponential linear unit (ELU) or rectified linear unit (ReLU) are two possible activation function options.

### 3.3 Computing

**Inhomogenous Poisson Process (IPP) Approximation**

The only restriction for $q(z, X)$ is that it is smooth, potentially allowing the model to be highly complex. This model specification flexibility is an appealing feature, but can make the integral in the logistic transformation intractable. We can view our method in an inhomogenous Poisson process (IPP) model framework to justify a discrete logistic transformation which is more computationally feasible. The conditional density in 3.2 has the form of an IPP model with domain on the unit interval $[0, 1]$ and log-intensity
\( q(z, X) \). The justification for this model can be found in B.1.

Fithian and Hastie (2013) describe a discrete approximation of the IPP model, which we can apply to our context. Suppose we have a dataset with \( i = 1, ..., n \) observations. We let \( z_i \) denote the transformed response value for observation \( i \). We can view the univariate random variable \( Z \) conditioned on \( X \) as a location on the unit interval, so we can consider the observed data as realizations of a point process over the unit interval. We follow the IPP approximation literature and propose to approximate the likelihood contribution of observation \( i \) as

\[
f(z_i | X) \approx \frac{e^{q(z_i, X)}}{e^{q(z_i, X)} + \sum_{k=1}^{K} e^{q(z_{ik}^*, X)}}
\]  

(3.7)

for \( k = 1, ..., K \). \( z_{ik}^* \sim \text{Uniform}(0, 1) \) controls are uniquely selected for each observation. Fithian and Hastie (2013) argue that this Monte Carlo approximation to the denominator of 3.2 is accurate for sufficiently large \( K \) in terms of approximating continuous conditional densities. The main effects of \( X \) are removed for this discrete logistic transformation just as they were in 3.2. \( z_{ik}^* \) can instead be selected using a fixed grid across the unit interval, but we expect this choice would require a larger \( K \) unless the data is evenly spread across the response space. This even data spread is the motivation for our transformation of \( Y \) by a CDF function, as a well-defined CDF can render the transformed data roughly uniform across the unit interval.

Another view of 3.7 is that \( z_i \) represents a sample from the location distribution of cases and the \( z_{ik}^* \) represent \( K \) matched samples from the uniform control distribution (Jarner et al. 2002). As mentioned in B.1, even a small \( K \) provides valid information about the \( q \) function. Thus, we can consider either the IPP approximation with large \( K \) to approximate the IPP integrated intensity and the matched case-control approximation where \( K = 1 \). We expect that a larger \( K \) value will induce more accurate parameter estimation, but at an additional computational cost that may not always be feasible.

Let \( \theta \in \mathbb{R}^m \) represent the parameter vector for the chosen \( q \) model, which we can write as \( q(z, X; \theta) \). The negative log likelihood for our model is

\[
\ell(\theta | Z_i, X_i) = \sum_{i=1}^{n} \left\{ -q(z_i, X_i; \theta) + \log \left[ \frac{e^{q(z_i, X_i; \theta)}}{e^{q(z_i, X_i; \theta)} + \sum_{k=1}^{K} e^{q(z_{ik}^*, X_i; \theta)}} \right] \right\} + \mathfrak{Z} ||\theta||^2
\]

(3.8)

where \( \mathfrak{Z} \geq 0 \) is a ridge penalty included to avoid model overfitting. Standard optimiza-
tion methods can be employed with this approximation by minimizing the negative log likelihood objective function. For \( K = 1 \), the method effectively reduces to logistic regression and the polynomial model can be evaluated using penalized logistic regression analysis techniques (Friedman et al. 2010). This technique arrives at a solution extremely quickly, making the polynomial method very accessible for large datasets. For deep learning methods, we perform stochastic gradient descent. Details for these implementation choices can be found in B.2.

Let \( l = 1, ..., L \) index a set of transformed response values. We can predict the conditional distribution at these transformed response values given covariate vector \( X \) and estimated parameter vector \( \theta = \hat{\theta} \) as

\[
f(z_l|X, \hat{\theta}) \approx \frac{e^{q(z_l, X; \theta)}}{\sum_{j=1}^{L} e^{q(z_j, X; \theta)}}.
\] (3.9)

This can be transformed back to the original scale via 3.1. A key advantage of our method is its simultaneous estimation of the model parameters. This structure ensures that we implicitly share information across all of our quantile estimates. For a method like DDR with a multinomial logistic regression classification model, each bin has its own set of parameters to be evaluated (excluding one bin which serves as a reference for the others). If a bin contains few or no observations, then that bin’s parameter estimates may be volatile and unreliable. A large number of cut points may be desired to approximate a continuous distribution estimate, which makes it more likely there are empty or sparsely filled bins. Our method avoids this issue by estimating parameters for only a single model, implicitly assuring information is shared across all quantile estimates. For certain model choices, another benefit of this single set of model parameters is that our method becomes computationally quicker than DDR and even QRF.

3.4 Simulation study

We conduct a simulation study to evaluate our method against the aforementioned DDR and QRF methods (Meinshausen 2006, Li et al. 2019). We compare these three machine learning-based methods in terms of effectiveness in predicting the conditional distribution of the target variable, explained below.

We simulate data from four distributions, first used by Li et al. (2019) for their complicated structures. Model 1 has a linear mean function, but also an error term
that varies with the covariates. The other three models have a nonlinear mean function. Models 2 and 3 are mixture distributions, while Model 4 uses a skew-normal distribution for the errors. Formally, the models are specified as

- **Model 1:** \[ Y = X^T \beta_1 + \exp (X^T \beta_2) \epsilon, \]
  - \( X \sim \text{MVN}(0, I_5), \)
  - \( \beta_1 \sim N(0, I_5), \beta_2 \sim N(0, 0.45 I_5), \epsilon \sim N(0, 1). \)

- **Model 2:** \[ Y = [10 \sin (2\pi X_1 X_2) + 10X_4 + \epsilon_1] \pi_1 + [20 (X_3 - 0.5)^2 + 5X_5 + \epsilon_2] (1 - \pi_1), \]
  - \( X_1, \cdots, X_{10} \text{iid} \sim \text{Uniform}(0, 1), \)
  - \( \pi_1 \sim \text{Bernoulli}(0.5), \epsilon_1 \sim N(0, 2.25), \epsilon_2 \sim N(0, 1). \)

- **Model 3:** \[ Y = [\sin (X_1) + \epsilon_1] \pi_1 + [2 \sin (1.5X_1 + 1) + \epsilon_2] (1 - \pi_1), \]
  - \( X_1 \sim \text{Uniform}(0, 10), \)
  - \( \pi_1 \sim \text{Bernoulli}(0.5), \epsilon_1 \sim N(0, 0.09), \epsilon_2 \sim N(0, 0.64). \)

- **Model 4:** \[ Y = 10 \sin (2\pi X_1 X_2) + 20 (X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon, \]
  - \( X_1, \cdots, X_{10} \text{iid} \sim \text{Uniform}(0, 1), \)
  - \( \epsilon \sim \text{SkewNormal}(0, 1, -5). \)

For each scenario, we simulate 100 datasets of size of 200, 1000, or 4000 observations to explore the relative efficacy of our method for various sample sizes. The datasets are randomly divided into training and testing data using a 75%/25% split. The models are fit using the training data, and then the distribution for each testing dataset observation is determined. For all models, the covariate data was normalized.

To evaluate the accuracy of a distribution estimate, we first calculate the range of the training response data and further extend it by 10%. We then calculate 100 evenly-spaced cut points between the extended range boundaries. For each model, we calculate the empirical CDF value associated with every cut point to get the conditional distribution estimate for every observation. We use the divergence function associated with the continuous ranked probability score (CRPS) to evaluate method performance (Gneiting and Raftery 2007, Krüger et al. 2016). The CRPS divergence is defined as

\[
d_{\text{CRPS}} = \frac{1}{N} \sum_{n=1}^{N} \int_{l}^{u} \left\{ \hat{F}(y|X_n) - F(y|X_n) \right\}^2 dy.
\]
This integral is approximated using 1000 evenly gridded points and the resulting approximation is normalized by the range of the data. For the simulation study, $N$ denotes the number of testing set observations for the given scenario.

We apply a matched case-control (MCC) justified approximation with $K = 1$ randomly selected controls to both the polynomial and deep learning models. Additionally, we apply an inhomogenous Poisson process (IPP) justified approximation with $K = 10$ randomly selected controls to the deep learning model in the simulation scenarios with 200 observations. For the polynomial MCC approximation method, the first-order interaction terms between covariates and squared covariate terms were included in the covariate pool for Models 1, 2, and 4. For Model 3, there was only one covariate variable so no interaction terms were possible. The highest polynomial power used in the model was $B = 3$.

Both deep learning approximations were applied using a model structure with one hidden layer. 30 nodes feed into the hidden and output layers each and the chosen activation function was the exponential linear unit (ELU). For the polynomial and deep learning methods, we select the normal cumulative distribution function (CDF) $\Phi$ to transform $Y$ as $Z = G(Y|X) = \Phi \left( \frac{Y - X\beta}{\sigma} \right)$ and estimate mean coefficient $\beta$ and standard deviation $\sigma$ parameters using ordinary least squares (OLS) regression. This choice ensures that the base distribution prediction for each observation in the testing dataset is Gaussian and centered at the OLS conditional mean. A larger ridge penalty (which lessens the deviance of the parameters from each other) will influence the predicted distribution toward this base distribution.

The polynomial MCC approximation is evaluated using a penalized logistic regression method while the deep learning approximations are evaluated using stochastic gradient descent. For more details on the implementation and evaluation of the models in these two methods, see B.2.

The classification models for the DDR method were constructed using the deep-conditional-distribution-regression Python package found at https://github.com/RLstat/deep-conditional-distribution-regression. The joint binary cross entropy loss objective function was selected due to its superior performance over the multinomial objective function in Li et al. (2019). Models were built with a single hidden layer and a 0% dropout rate. The ELU activation function is selected for the hidden layer, with a softmax activation function applied on the output layer.

The QRF method was utilized with 500 trees were built using the quantregForest package in R. This package predicts the conditional response values associated with inputted quantiles, so 100 evenly-spaced quantiles from .00001 to .99999 were generated.
and the QRF models estimated the cut points associated with these quantiles.

Figure 3.1: A boxplot of the distribution of CRPS divergences for each model and dataset size across 100 datasets for the QRF, DDR, Deep Learning $K=1$, Polynomial $K=1$, and Deep Learning $K=10$ conditional distribution estimation methods. The y-axis scale is not synchronized across scenarios and dataset sizes.

Figure 3.1 gives the simulation results. In general, both deep learning approximation methods performed well compared to DDR. The deep learning MCC approximation model outperformed DDR in terms of median CRPS divergence in 8 of the 12 data scenarios. The polynomial MCC approximation model performed worse against DDR by comparison, only producing a lower median CRPS divergence in 4 of the 12 scenarios.

The deep learning IPP approximation method noticeably improved the CRPS divergence results compared to the deep learning MCC method in all four scenarios with 200 observations. In Models 1, 3, and 4 with 200 observations, this approximation beat both QRF and DDR in terms of median CRPS divergence, suggesting that this deep learning
approximation is more useful than the MCC approximation in situations with a small sample size.

Our deep learning MCC approximation method also outperformed QRF in terms of median CRPS divergence in 7 of the 12 scenarios, although the relative CRPS divergence ranges in Model 1 with 4000 observations suggests our method may not have produced better results in that scenario. QRF fared better than the polynomial MCC approximation model in the majority of scenarios, although the polynomial MCC approximation produced lower median CRPS divergence values across dataset sizes in Model 3.

Our deep learning method performed relatively better in terms of CRPS divergence in Models 1 and 3 compared to Models 2 and 4. Model 1 had a normal distribution structure which may have been advantageous for our method since we used the normal quantile function to transform our data. Model 3 was a mixture distribution as Model 2 was, but only had a single covariate compared to the 10 covariates in Models 2 and 4.

Table 3.1 gives the average computation times for the polynomial and deep learning approximation methods for Model 1. The deep learning IPP approximation computation times for 1000 and 4000 observations were calculated on only 5 datasets, whereas the computation times for the other scenarios were calculated for all 100 datasets. The deep learning MCC and IPP approximations were significantly more computationally burdensome than the polynomial MCC approximation. The deep learning MCC approximation average computation time was over an hour for 4000 observations. On average, the deep learning IPP approximation for 10 controls took roughly five or six times as long to evaluate as the deep learning MCC approximation. Figure 3.1 suggests the deep learning MCC approximation and especially deep learning IPP approximation are preferable to the polynomial MCC approximation for conditional distribution estimation in many data scenarios, however it may not be as readily scalable to larger datasets. On the contrary, the increase in computation time from 200 observations to 4000 observations for the polynomial MCC approximation was negligible. The polynomial MCC approximation is easily applicable to large datasets in data scenarios where the deep learning approximations are computationally unfeasible.

3.5 Application to tropical cyclone intensity forecasting

We apply our method to calibrate short-term tropical cyclone wind intensity forecasts. A conditional distribution estimation approach to this problem could provide additional
Table 3.1: A table of the average computation times (in minutes) and associated standard errors for evaluating the data from Model 1 across all dataset sizes. The average computation times were recorded for the polynomial MCC approximation ($K = 1$), deep learning MCC approximation ($K = 1$), and deep learning IPP approximation ($K = 10$) conditional distribution estimation methods. The average computation times and standard errors for the deep learning IPP approximation method for 1000 and 4000 observations were calculated using only the first 5 datasets due to the burdensome evaluation time. The average computation times and standard errors for the remaining scenarios were calculated over all 100 datasets. The polynomial MCC approximation method was evaluated using penalized logistic regression, whereas the deep learning IPP approximation methods were evaluated using stochastic gradient descent.

<table>
<thead>
<tr>
<th>Dataset Size</th>
<th>Mean Computation Time (Minutes) For Model 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Polynomial $K = 1$</td>
</tr>
<tr>
<td>200</td>
<td>0.0002 (0.0002)</td>
</tr>
<tr>
<td>1000</td>
<td>0.0002 (0.0001)</td>
</tr>
<tr>
<td>4000</td>
<td>0.0005 (0.0001)</td>
</tr>
</tbody>
</table>

context on response distribution features to better inform policy decisions compared to a point estimate approach (Cloud et al. 2019). Our data comes from Hurricane Weather Research and Forecasting (HWRF) Model, developed and maintained by the U.S. Environmental Modeling Center (EMC) (Biswas et al. 2017). HWRF is a deterministic atmosphere-ocean model used for hurricane research and forecasting. The HWRF model includes a forecasted maximum 10-meter wind speed value which is designated as the covariate of interest. The actual maximum 10-meter wind speed value is the response variable. Covariate and response information are recorded up to four times a day for each day a tropical cyclone is active in 6 hour increments. At each time point, forecasted covariate data and response data are given for up to 96 hours into the future by 3 hour increments.

The full dataset contains information from 65 tropical cyclones located around the Atlantic Seaboard between 2013 and 2017. For this application, we focus on lag 3 and lag 6 forecast predictions and subset the overall dataset of 45,639 observations to obtain two smaller datasets of 1,383 observations each for only these lag times. Observations with missing response values were removed. The final lag 3 and lag 6 datasets each had 1,267 observations.
The polynomial regression method was implemented using the MCC approximation with a single control, $K = 1$. The highest polynomial power used in the polynomial model was $B = 3$, and the quadratic covariate term was included in the covariate matrix. The deep learning method was implemented using an IPP approximation with $K = 20$. The deep learning model was built with a single hidden layer, where 15 nodes feed into the hidden and output layers each, and uses an ELU activation function. The polynomial model was evaluated using penalized logistic regression and the deep learning model was evaluated using mini-batch stochastic gradient descent. For both methods, a variety of ridge penalties were considered. A ridge penalty of 0.000001 was selected for the deep learning method for both lags and the polynomial method for lag 3, and a ridge penalty of 0.0005 was selected for the polynomial method for lag 6. For the deep learning method, a variety of initial learning rates were also considered, with the optimally tuned models using an initial learning rate of 1 for both lags. Further details on how these models were fit are given in B.2.

The QRF model was built using 500 trees and evaluated using the quantregForest R package. The DDR method was run using a deep learning classification model and evaluated in Python using the deep-conditional-distribution-regression package. The model had one hidden layer with 15 nodes and a 0% dropout rate to mimic the deep learning approximation model specifications. The joint binary cross entropy loss objective function was selected. The ELU activation function was applied to the hidden layer, with the softmax activation function used for the output layer. As in the simulation study section, we select a normal CDF to transform $Z$ and estimate the CDF parameters using OLS regression.

The tropical cyclones were randomly assigned to one of five folds, and 5-fold cross validation was performed. For each fold, we calculate the CRPS of the testing set to evaluate method performance as the CRPS divergence is unavailable without knowledge of the true distribution (Matheson and Winkler 1976, Hersbach 2000). CRPS is defined as

$$CRPS = \frac{1}{N} \sum_{n=1}^{N} \int_0^u \left\{ \hat{F}(y|X_n) - I(y \geq Y_n) \right\}^2 dy.$$

As with the CRPS divergence evaluation, the integral is approximated using 1000 evenly-gridded points and the resulting approximation is normalized by the range of the data. For this application, $N$ refers to the number of observations in the given testing fold.

Table 3.2 gives the average CRPS across folds and the accompanying standard error.
Table 3.2: The 5-fold mean CRPS values for each lag time for the QRF, DDR, polynomial MCC approximation \((K = 1)\), and deep learning IPP approximation \((K = 20)\) conditional distribution estimation methods.

<table>
<thead>
<tr>
<th>Response Lag Time</th>
<th>Method</th>
<th>QRF</th>
<th>DDR</th>
<th>Polynomial</th>
<th>Deep Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Hour Lag</td>
<td></td>
<td>0.0247</td>
<td>0.0288</td>
<td>0.0218</td>
<td>0.0212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0037)</td>
<td>(0.0096)</td>
<td>(0.0023)</td>
<td>(0.0022)</td>
</tr>
<tr>
<td>6 Hour Lag</td>
<td></td>
<td>0.0293</td>
<td>0.0299</td>
<td>0.0267</td>
<td>0.0258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0043)</td>
<td>(0.0070)</td>
<td>(0.0032)</td>
<td>(0.0033)</td>
</tr>
</tbody>
</table>

for each method. Our polynomial and deep learning approximation methods outperform QRF and DDR by these metrics. Additionally, the deep learning IPP approximation slightly outperforms the polynomial MCC approximation in terms of average CRPS. The lag 6 predictions result in a higher average CRPS for each method than the lag 3 predictions, due to the increased difficulty of forecasting further into the future.

As a comparison, the 5-fold mean CRPS value for the conditional Gaussian distribution evaluated via OLS estimation was calculated. The OLS-predicted mean CRPS values were 0.0220 (with standard error 0.0020) and 0.0269 (with standard error 0.0026) for lag 3 and lag 6, respectively. The deep learning IPP approximation and polynomial MCC approximation both outperform these estimated average CRPS values, although the improvement made by the polynomial MCC approximation is very slight.

Figure 3.2 displays the deep learning IPP approximation conditional response distribution predictions for lag 3 and lag 6 when using all of the training data. The lag 3 predicted quantiles for this model look generally unimodal and Gaussian, with some slight left skewness for smaller covariate values. The lag 6 predicted quantiles are also generally unimodal and Gaussian for the larger covariate values, but exhibit some clear non-normality and skewness for the lower covariate values. Both models used for Figure 3.2 were fit using the same ridge penalties and initial learning rates as the models used to calculate the CRPS values for each fold in Table 3.2. The predicted distributions for each individual fold that were used to calculate the mean CRPS results in Table 3.2 are not necessarily equivalently shaped to these plotted predicted distributions. For instance, the deep learning IPP approximation method for lag 6 in the third fold predicts a somewhat bimodal distribution of maximum 10-m wind speed for larger covariate values. For plots of the predicted distributions for each individual fold, see B.3. Overall, these somewhat Gaussian-shaped predicted distribution visualizations are consistent with the CRPS results that suggest the OLS-predicted distribution method performs only slightly
Figure 3.2: Deep learning IPP approximation \((K = 20)\) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with all available data. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. \(\Pr(Y = y | X)\) refers to the relative probability that the maximum 10-meter wind speed \(Y = y\) occurs given the HWRF-forecasted maximum 10-meter wind speed value \(X\). \(Y | X\) refers to the conditional response value \(Y\) given the covariate \(X\).
Figure 3.3: Polynomial MCC approximation \((K = 1)\) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with all available data. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. \(\Pr(Y = y|X)\) refers to the relative probability that the maximum 10-meter wind speed \(Y = y\) occurs given the HWRF-forecasted maximum 10-meter wind speed value \(X\). \(Y|X\) refers to the conditional response value \(Y\) given the covariate \(X\).
worse than the deep learning IPP approximation for $K = 20$ in this application.

Figure 3.3 displays the polynomial MCC approximation conditional response distribution for lag 3 and lag 6 when using all of the training data. Again, the models used to predict these distribution quantiles maintained the same ridge penalties and initial learning rates as the corresponding models used to obtain the polynomial MCC approximation average CRPS in Table 3.2. The larger and smaller covariate values are associated with distributions with sharper peaks, whereas the predicted distributions for the middle covariate values have broader, less symmetric peaks. The lag 3 and lag 6 predicted distributions look more similar here than the lag 3 and lag 6 predicted distributions in Figure 3.2.

### 3.6 Discussion

We propose a flexible conditional distribution estimation method which can incorporate machine learning techniques such as deep learning models or polynomial regressions. We examined the performance of some of these model types for different data distributions in a simulation study, finding that our method implemented with a deep learning model outperformed other conditional distribution estimation methods in multiple scenarios. In a real world application of our method, we found both the deep learning and polynomial model-based methods provided useful insight on tropical cyclone maximum wind speed forecasting compared to other methods, with the deep learning method performing best in terms of the mean 5-fold CRPS performance metric.

Further approximation and/or computational techniques for this method can fully unlock its utility for conditional distribution estimation. We introduced an IPP-based discrete approximation with $K$ controls to make model evaluation feasible, but were limited to selecting a small $K$ and a relatively basic deep learning model structure with one hidden layer and 30 nodes. We expect that our method could substantially improve its predictive accuracy if a more complex deep learning model structure were tractable. Integration of our method with TensorFlow or another deep learning optimization programming language could be helpful in this regard. Perhaps an approximation that reduces the number of observations could also be incorporated to improve methodological accuracy.

Another potential methodological improvement is through the selection of the control values for our case-control based approximation. Fithian and Hastie (2014) describe a local case-control sampling technique meant to address conditional imbalance in addition to the marginal imbalance issue addressed by standard case-control sampling. Per-
haps incorporating this approach or another weighted control selection technique could be adapted to our framework to improve the conditional distribution estimation for a smaller number of controls.

For complicated models with many parameters (from multiple covariates, layers, and/or nodes), the ridge penalty in 3.8 influences the parameters towards 0 so that they deviate less from each other. As a result, \( g(Z|X) \) tends toward a uniform distribution, and \( f(Y|X) \) consequently tends toward the distribution implied by the transformative cumulative distribution function. In both the simulation and application, a Gaussian cumulative distribution function was selected to transform \( Y \) to \( Z \). A conditionally normal response distribution is often assumed in statistics, so this specification is reasonable for many applications. Still, perhaps a more sophisticated optimization algorithm would allow for more deviance between parameter estimates and be less influenced by the choice of CDF.

Additionally, the CDF function parameters were estimated via OLS. OLS requires there to be more observations than covariates in order to obtain a unique parameter solution. This restriction might disallow the inclusion of higher order interactions in the polynomial approximation model when the sample size is small because it would result in more covariates than training observations. In this scenario, the CDF function transformation should not be used. Instead, boundaries should be chosen for \( Y \) and our method can be analogously applied.

### 3.7 Acknowledgements

We’d like to thank Dr. Kevin Gunn for lending equipment and technical assistance to obtain some of the results for the tropical cyclone intensity forecasting application.
Chapter 4

Geolocation For Spatial Point Pattern Data In A Deep Learning Framework

4.1 Introduction

Geolocation, or the prediction of an object’s spatial provenance conditional on its measurable characteristics, is an analytical approach with a broad range of uses from animal tracking to individualizing a user’s website experience (Phillips et al. 2004, Han et al. 2014). An important application is forensic geolocation, a method used to estimate the location origin of articles at a crime scene. Many of these articles retain dust or soil that harbor a multitude of microbial life forms such as fungi which are useful for geolocation purposes due to their abundance and specific habitat locations (Blackwell 2011, Hawksworth and Wiltshire 2011, Talbot et al. 2014). Grantham et al. (2015) conducted the first formal geolocation analysis of dust samples. Grantham et al. (2015) proposed a Bayesian Discriminant Analysis (BDA) method to estimate the probability of each microbial species at each candidate location and then use these estimates to predict the provenance of new microbial samples. Grantham et al. (2019) develop a more sophisticated microbial geolocation tool called DeepSpace by partitioning the spatial domain and fitting a neural network classification model on the partitioned regions.

Spatial point pattern analysis is the study of spatial distributions of points in a defined domain, making it a natural fit for geolocation prediction. Early on, point pattern analyses generally involved non-parametric, descriptive techniques of the given data such as quadrat methods, kernel estimation methods, distance methods, nearest-neighbor dis-
tribution function methods, and K-function methods (Ripley 1977, Diggle 2013, Cressie 2015). Another early method commonly used for Geographical Information Systems (GIS) problems involved fitting a naive logistic regression with pseudo-absences or discretized absences, however this approach has various issues (Baddeley et al. 2010, Warton et al. 2010). The major research area for point pattern analysis in recent decades has been specifying the underlying point process mechanism from which the point pattern data realizations are generated (Moller and Waagepetersen 2003). Specifically, point process modeling (PPM) is preferred over summary statistic computation for its ability to account for the effect of spatial (or non-spatial) covariates. The homogenous Poisson process (HPP) is one of the simplest PPMs and often used as a null model in point pattern analysis (Diggle 2013, Cressie 2015). The HPP assumes that the number of points in any subregion of the spatial domain is Poisson distributed with a constant intensity function. The inhomogenous Poisson process (IPP) is a generalization of this model which allows for a spatially-varying intensity function (Diggle 2013, Cressie 2015). All Poisson processes have the independent scattering property that points are assumed independent (Moller and Waagepetersen 2003). The Cox process model is a generalization of the IPP which considers random effects in the intensity function to model potential spatial clustering (Moller and Waagepetersen 2003). The log Gaussian Cox process (LGCP) is a commonly used Cox process where the log of the intensity function is defined as a Gaussian process (Møller et al. 1998). A Gibbs process is another useful PPM which can explicitly model interaction terms between points and account for spatial regularity (Moller and Waagepetersen 2003).

An application of a point pattern analysis framework to forensic geolocation requires high-dimensional methodological scalability in terms of efficiency and accuracy to analyze the large number of species present in a given dust or soil sample. An IPP is a popular PPM choice because its form allows for maximum likelihood (ML) estimation of the intensity function parameters, although an integral approximation must first be made (Berman and Turner 1992, Fithian and Hastie 2013). Fithian and Hastie (2013) explain the derivation of naive logistic regression or maximum entropy modeling from the log-linear IPP model, although issues with logistic regression estimation occur if the IPP is misspecified. A logistic regression-based technique can equivalently be implemented for a log-linear intensity function, making it the most scalable choice for high-dimensional data with access to well-studied generalized linear model (GLM) variable selection techniques (Baddeley et al. 2015). The random effects in Cox process models make ML estimation difficult, but the hierarchical structure of these models allow for Markov Chain Monte Carlo (MCMC) methods (Illian et al. 2009, Chakraborty et al. 2011). MCMC
does not scale well to high dimensions, so more efficient MCMC algorithms have been proposed for Cox processes and specifically LGCPs (Adams et al. 2009, Gunter et al. 2014). Another popular Bayesian approach for models with latent Gaussian structures is integrated nested Laplace approximation (INLA) (Rue et al. 2009). This method is generally faster than MCMC although possibly less accurate (Taylor and Diggle 2014). Efficient non-Bayesian methods include a composite likelihood approach based on the second-order intensity function of the Cox process and a weighted estimating equation estimator (Guan 2006, Guan and Shen 2010). Variational Bayesian evaluation methods for LGCP models are another computationally efficient approach (Teng et al. 2017, Aglietti et al. 2019). A variety of Bayesian variable selection techniques have been considered for LGCPs (Junior and da Silva 2017). Gibbs process models also are difficult to fit via ML estimation due to the presence of an intractable normalizing constant, with a maximum Poisson pseudo-likelihood generally used in favor of MCMC for speed purposes (Besag 1978, Baddeley and Turner 2000). Baddeley et al. (2014) develop a computationally superior and less biased evaluation method using logistic regression score functions instead of the pseudo-likelihood score functions, later adapted for a variational Bayesian context by Rajala (2014). These methods work best for a moderate number of covariates, so high dimensional Gibbs process models benefit from feature selection methods (Baddeley et al. 2015, Rajala et al. 2017, Ba and Coeurjolly 2020).

A predictive geolocation technique which estimates the full conditional density function of the response location gives a more complete summary of the point pattern behavior. Some common conditional density estimation (CDE) approaches include finite mixture modeling, kernel estimation of the joint and marginal distributions, and Bayesian nonparametric techniques. (Rosenblatt 1956, Escobar and West 1995, Dunson et al. 2007). A set of more accessible CDE methods are grounded in machine learning techniques. Meinshausen (2006) developed a quantile regression forest (QRF) method utilizing the random forest structure to obtain conditional quantile estimates. Li et al. (2019) proposed a univariate deep distribution regression (DDR) for CDE, partitioning the response space multiple times, fitting a classification model onto each resulting partition set, and then combining each partition model’s conditional quantile estimates. Grantham et al. (2019) expanded the ideas in Li et al. (2019) to two-dimensions to apply the technique for forensic geolocation purposes, using a deep learning classification model for its compatibility with high-dimensional dust sample data. The methods proposed by Li et al. (2019) and Grantham et al. (2019) both require a partition of the response domain, which can adversely affect estimation stability or prediction precision if the partitions contain too few or too many response observations, respectively. Huberman et al.
(2020a) introduces a univariate conditional density estimation method to leverage the IPP model structure with an underlying neural network to simultaneously estimate the conditional density quantiles without requiring a domain partition. The deep learning model in Huberman et al. (2020a) generally outperforms both the QRF and DDR methods in multiple simulated settings.

In this chapter, we develop a multivariate nonparametric conditional density estimation method via an IPP model with the intensity function modeled as a neural network to obtain forensic geolocation estimates. We focus on point geolocation rather than the full conditional distribution to mimic other geolocation approaches. We propose two neural network model definitions each designed to capture spatial dependence between the location data, highlighting the method’s flexibility in specification. A computational approximation of our method is described in order to evaluate our model and obtain parameter estimates. We demonstrate the utility of our method with a simulation study and then apply it to a dataset of dust samples across the continental USA to display its comparative potential to an established forensic geolocation technique in BDA. We finish with a discussion of features, challenges, and potential improvements for our method.

4.2 Geolocation models

For observations $i \in \{1, \ldots, n\}$, the response variable is the spatial location $s_i = (s_{i1}, s_{i2})$. The observations are assumed to be confined to the sampling window $s_i \in \mathcal{D} \subset \mathbb{R}^2$, e.g., for our analysis $\mathcal{D}$ is the continental USA. The response is regressed onto the covariates $X_i(u) = \{X_{i1}(u), \ldots, X_{ip}(u)\}$ for spatial location $u = (u_1, u_2) \in \mathcal{D}$. The covariates could be spatially-varying functions such as the elevation at location $u$, or non-spatial variables such as the presence of operational taxonomic unit (OTU) $j$ in sample $i$, in which case $X_{ij}(u) = X_{ij}$. We denote the collection of covariates across the spatial domain as $\mathcal{X}_i = \{X_i(u); u \in \mathcal{D}\}$.

Our goal is to model the conditional distribution of the location given the covariate vector. We then attain the point geolocation estimate as the mode of this distribution. In the context of forensic geolocation example, this is the distribution of the sample’s provenance given its dust microbiome. The response is modeled as a spatially inhomogeneous Poisson process (IPP). Given $\mathcal{X}_i$, the IPP has a spatially-varying intensity function $\lambda(u; \mathcal{X}_i)$ so that the conditional density function of $s_i$ given that a sample is observed is

$$f(s|\mathcal{X}_i) = \frac{\lambda(s; \mathcal{X}_i)}{\int_{\mathcal{D}} \lambda(u; \mathcal{X}_i)du}. \quad (4.1)$$
Rather than specify a parametric model for the intensity function, such as a log-linear function of $X_i$, we model the log-intensity function using a Feed-Forward Neural Network (FFNN) model. This flexible non-parametric regression model allows for non-linear covariate effects, interactions between covariates, and interactions between spatial location and covariates, i.e., spatially-varying effects.

The log-intensity function at $u$ is defined by the covariates at $u$ and user-defined functions of $u$, denoted $C_1(u),...,C_q(u)$ and defined below. The FFNN model with one hidden layer is

$$
\log\{\lambda(u; X_i)\} = \sum_{t=1}^{T} \delta_t f_A(H_{ti})
$$

$$
H_{ti} = \gamma_{0t} + \sum_{r=1}^{R} \gamma_{rt} f_A(I_{ri})
$$

$$
I_{ri} = \beta_{0r} + \sum_{j=1}^{p} \beta_{jr} X_{ij}(u) + \sum_{l=1}^{q} \alpha_{lr} C_l(u),
$$

where $f_A$ is the activation function, $T$ and $R$ are the number of nodes feeding into the output and hidden layers, respectively, and the unknown parameters are $\{\delta_t\}$, $\{\gamma_{rt}\}$, $\{\beta_{jr}\}$ and $\{\alpha_{lr}\}$. We consider two versions of this general model in Sections 4.2.1 and 4.2.2 that vary by their choice of $C_l(u)$.

### 4.2.1 Spatial coordinate deep learning (Coordinate DL)

A straightforward choice for the user-defined functions $C_l(u)$ is to simply use the elements of the spatial coordinates $u = (u_1, u_2)$ as features in the FFNN model. That is, in (4.2), set $q = 2$ and $C_1(u) = u_l$ for $l \in \{1, 2\}$. These elements can optionally be centered to reduce collinearity. It follows from the universal approximation theorem (Hornik et al. 1989) that this relatively simple model with sufficiently large $R$ and $T$ can approximate any true intensity function, $\lambda$. Parametric functions of the spatial coordinates are often used as covariates in spatial point pattern analysis (Gelfand et al. 2010), but using the FFNN model allows for an arbitrarily complex relationship between the coordinates and the intensity function. A disadvantage of this approach is that it by requiring a large number of nodes, computing may be slow and convergence may be difficult to ensure.
4.2.2 Radial basis function deep learning (Radial Basis DL)

The second model we consider is motivated by a log-Gaussian Cox process (Moller and Waagepetersen 2003). Ignoring covariates, the Cox process log intensity function is \( \log\{\lambda(u)\} = \theta(u) \) where \( \theta \) is a Gaussian process. While this provides a flexible model, a Gaussian process leads to computational difficulties for large datasets (Heaton et al. 2019). To improve computation, Gaussian processes are often approximated using a finite basis expansion \( \theta(u) \approx \sum_{l=1}^{q} C_l(u) \alpha_l \) where \( C_l(u) \) are known basis functions and the coefficients \( \alpha = (\alpha_1, ..., \alpha_q)^T \) have a Gaussian distribution. By choosing the basis functions and covariance of \( \alpha \) appropriately, a rich class of models for \( \theta \) is achieved (e.g., Banerjee et al. 2008, Cressie and Johannesson 2008, Nychka et al. 2015). By selecting the \( C_l(u) \) to mimic a low-rank approximation to a Gaussian process, we center our nonparametric FFNN model on a parametric log-Gaussian Cox process.

Radial basis functions with compact support are selected as in Nychka et al. (2015). Similar basis functions were used by Li et al. (2020) as inputs to a FFNN for spatial prediction. The basis functions are determined by a grid of knots \( u^*_1, ..., u^*_q \) that span the sampling window \( D \). The Wendland kernel basis functions are

\[
C_l(u) = \frac{1 - d_l(u)}{3} \left\{ \frac{35d_l(u)^2 + 18d_l(u) + 3}{d_l(u)} \right\} I\{0 \leq d_l(u) \leq 1\} \quad (4.3)
\]

where \( d_l(u) = ||u - u^*_l||/\phi \) is the distance between location \( u \) and knot \( u^*_l \) and \( \phi \) is the kernel bandwidth. Following Nychka et al. (2015), the knots are taken to be a regular grid with grid spacing \( \delta \) and the bandwidth is \( \phi = 2.5\delta \) to balance minimizing basis function overlap and ensuring the \( C_l(u) \) functions provide a complete basis over \( D \).

The inclusion of radial basis functions in the input layer centers the FFNN model on the log-Gaussian Cox process and thus encodes the prior belief that the intensity function is a smooth spatial process. This also demonstrates the flexibility of the proposed model as it includes the log-Gaussian Cox process as a special case. However, the additional inputs complicate optimization and the model includes an additional tuning parameter, the number of basis functions \( q \).

4.3 Computation

Evaluating the integral in (4.1) is infeasible for the FFNN model and an approximation is needed. For related IPP models, Fithian and Hastie (2013) and Huberman et al. (2020a) justify a discrete approximation via a connection with a matched case-control study. This
approximation can be adapted to the models in Section 4.2. We define \( \{s_{i1}^*, ..., s_{iK}^*\} \in \mathcal{D} \) as a set of \( K \) randomly-selected control values. A discrete approximation to (4.1) is

\[
 f(s_i|\mathcal{X}_i) = \frac{\lambda(s_i; \mathcal{X}_i)}{\int \lambda(u; \mathcal{X}_i) du} \approx \frac{\lambda(s_i; \mathcal{X}_i)}{\lambda(s_i; \mathcal{X}_i) + \sum_{k=1}^K \lambda(s_{ik}^*; \mathcal{X}_i)}.
\]

(4.4)

This approximation is most accurate when the number of controls is large, however it is a valid approximation for any choice of \( K \) as justified by the matched case-control explanation in Huberman et al. (2020a). Fithian and Hastie (2013) suggest an alternative method of selecting controls for each observation on a fixed grid spanning the domain, however this alteration would likely require more controls and be more computationally intensive, especially in multiple dimensions.

Let \( \Theta \) be the collection of unknown parameters \( \{\delta_t\}, \{\gamma_{rt}\}, \{\beta_{jr}\} \) and \( \{\alpha_{lr}\} \) and denote the intensity function as \( \lambda_\Theta(u; \mathcal{X}_i) \) to signify its dependence on the parameters. The objective function corresponding to (4.4) is

\[
 \ell(\Theta) = \sum_{i=1}^N \left[ -\log\{\lambda_\Theta(s_i; \mathcal{X}_i)\} + \log \left\{ \lambda_\Theta(s_i; \mathcal{X}_i) + \sum_{k=1}^K \lambda_\Theta(s_{ik}^*; \mathcal{X}_i) \right\} \right] + z||\Theta||^2
\]

(4.5)

where \( z \geq 0 \) is a ridge penalty included to avoid overfitting the model parameters. This function is optimized using stochastic gradient descent. This model optimization technique can require a demanding tuning process to optimize settings such as the ridge penalty or initial learning rate with respect to the chosen model performance metric. More details are given on the tuning process used in this chapter in Sections 4.4 and 4.5 as well as Appendix C.1. Once the model is optimized, the estimated PDF can be approximated for a grid of values within the domain. The conditional location prediction for a given observation can be taken as the mode of the estimated PDF.

### 4.4 Simulation study

We conduct a simulation study to compare the efficacy of the Coordinate DL method, our Radial Basis DL method, and the Bayesian Discriminant Analysis (BDA) of Grantham et al. (2015). The sampling window is the unit square \( \mathcal{D} = [0, 1]^2 \). To simulate forensic geolocation data, the covariate information is assumed to be binary indicators to parallel microbial OTU sample information. Data are generated by first sampling the \( p \) binary
OTU indicators $X_i = (X_{i1}, \ldots, X_{ip})$ from an autoregressive process to induce correlation,

\[ X_{i1} \sim \text{Bernoulli}(0.5), \quad \text{and} \]

\[ X_{ij+1} | X_{ij} \sim \text{Bernoulli}\{0.9X_{ij} + 0.1(1 - X_{ij})\}. \]  

(4.6)  

(4.7)

The responses are then generated from the mixture density

\[ f(s | X_i) = \sum_{h=1}^{V} \pi_h(X_i) \phi(s; \nu_h, \sigma^2 I_2), \]  

(4.8)

where $\pi_h(X_i)$ are mixture probabilities defined below, $\phi(s; \nu, \Sigma)$ is the bivariate normal density function with mean $\nu$ and covariance $\Sigma$, $\nu_h$ are fixed knot locations and $\sigma^2$ is the mixture variance. We set the $\nu_h$ to the grid \{1/4, 1/2, 3/4\}² so that $V = 9$ and set $\sigma = 0.1$. The mixture components follow the multinomial logistic model $\pi_h(X_i) = \eta_{ih} / \left( \sum_{m=1}^{V} \eta_{im} \right)$ for log($\eta_{ih}$) = $X_i \beta_h$. The covariate weights $\beta_h = (\beta_{h1}, \ldots, \beta_{hp})^T$ are assumed to be sparse with a different subset of covariates active for each mixture component. Specifically, we set $\beta_{hj} = (1 - |2h - j|/10)_+$ where $(x)_+ = \max\{x, 0\}$.

Each simulated dataset is comprised of $n$ independent observations, $s_1, \ldots, s_n$. We simulate 100 datasets for $n = 500$ and either $p = 25$ or $p = 100$. For both deep learning methods, the neural network structure has one hidden layer with $R = T = 30$ nodes and the exponential linear unit (ELU) activation function. The deep learning methods are each run for $K = 2$ and $K = 5$ control values. The Radial Basis DL method was fit with the evenly spaced $10 \times 10$ grid \{0, 1/9, 2/9, ..., 1\}² for a total of 100 basis knots. The deep learning models are evaluated using stochastic gradient descent. For more details on the optimization algorithm and selected tuning parameters, see Appendix C.1.

For BDA, we selected bandwidth values for each covariate using the generalized cross-validation heuristic in Grantham et al. (2015). A sequence from 0.01 to 100.01 separated by 0.1 were designated as the candidate bandwidth values. A flat prior on the candidate coordinates is assumed and the log-likelihood function in Grantham et al. (2015) is evaluated for each of the testing coordinate candidates.

Each dataset is separated into training and testing datasets by a 70%/30% split. For each testing set observation and all methods, the predictive density is evaluated on a $100 \times 100$ grid spanning the unit square domain and the coordinate with the highest density is taken as the prediction, $\hat{s}_i$. Methods are compared using the Median Absolute Error (MAE) between the predicted and true location of the sample. The average negative log likelihood values across datasets are also calculated after tuning for MAE performance to capture the distributional prediction capabilities of the methods. For
these results, see Appendix C.2.

Figure 4.1: A boxplot of the distribution of MAEs for each number of observations and OTU covariates across 100 datasets for the Radial Basis Deep Learning, Coordinate Deep Learning, and Bayesian Discriminant Analysis estimation methods. The deep learning methods were run for both $K = 2$ and $K = 5$ controls.

Figure 4.1 gives the distribution of MAEs across 100 datasets for each conditional estimation method in each of the two data scenarios. The deep learning methods generally outperform BDA in both scenarios, although the Coordinate DL approach with 2 controls produces similar or worse MAE values than BDA. Excluding the Coordinate DL method with 2 controls, the deep learning methods produced more precise MAE results across datasets in terms of spread.

Radial Basis DL outperforms Coordinate DL in both simulation scenarios. The Radial Basis DL approach results in lower MAE values for both 2 and 5 controls in comparison to Coordinate DL, while also producing a smaller range of MAE values. Even Radial
Basis DL with 2 controls outperforms Coordinate DL with 5 controls, indicating the former method is superiorly predicting true locations for these simulation scenarios.

The performance advantage for Coordinate DL compared to BDA is more definitive when using 5 controls, whereas Radial Basis DL outperforms BDA for both 2 and 5 controls. The Coordinate DL performance improves as the number of controls are increased from 2 to 5. This same improvement does not occur for the Radial Basis DL method, where the method performs similarly for 2 and 5 controls for each number of OTUs. None of the methods appear significantly worse at predicting 100 OTUs compared to 25 OTUs, although the MAE range does comparatively increase somewhat for Coordinate DL with 100 OTUs.

4.5 Case study

We demonstrate our method for a forensic geolocation application. Our data are a collection of microbial dust samples from various homes across the continental United States (Dunn et al. 2013). Dust samples can contain DNA from various organisms such as bacteria or animals that can help determine the sample’s spatial origin and aid forensics research. DNA can be analyzed via sequencing techniques to determine the various organisms present at the sample. Operational taxonomic units (OTUs) are clusters of similar DNA sequences that are determined based on similarity thresholds. Our goal is to accurately geolocate dust samples conditional on the presence or absence of these OTUs at the sample location.

We compare the efficacy of each spatial deep learning method to BDA as in the simulation study. Our dataset contains 1,133 dust samples and 763 OTUs, along with latitude and longitude coordinates for each sample. These data are a publicly available subset of the full proprietary OTU dataset considered in Grantham et al. (2019). The location coordinates and OTUs are designated as the response variables and spatial covariates, respectively. The covariate information is converted to binary presence/absence indicators for non-zero and zero entries, respectively. This conversion is made to comply with previous analyses (Grantham et al. 2015). The dataset is split into 5 folds of similar size. The competing methods are run five times with a different fold functioning as the testing dataset for each run while the unselected folds comprise the corresponding training dataset.

For all methods, a bounding box around the continental United States is determined as the domain limits and a $100 \times 100$ grid of candidate prediction locations is generated within the bounding box. Only candidate prediction locations falling on the continental
USA are considered, for a final total of 6,266 eligible candidate prediction locations. A location prediction is selected for each testing set fold observation and the MAE is calculated for each fold. The MAE values are averaged across the five folds for each method to obtain the average 5-fold MAE. For this application, the great circle distance (measured in kilometers) is used to calculate the Wendland radial basis functions as well as the distance of the predicted locations to the true locations.

The spatial geolocation deep learning methods were implemented with a single hidden layer and with ELU activation functions. For each deep learning method, models were constructed with either 30 or 50 nodes feeding into the hidden and output layers and optimized using either $K = 2$ or $K = 5$ controls. For the Radial Basis DL method, the knot grid spacing was set using a grid of basis function knots spaced 3 degrees apart. Only the knots located within the continental United States were retained, resulting in 821, 198, and 88 knots respectively for each of the three knot grid spacing settings. Mini-batch stochastic gradient descent was used to evaluate the deep learning models. A variety of initial learning rates and ridge penalty values were considered for each model and the specifications that resulted in the minimum average MAE were selected. More details on the tuning process and the final tuning parameters can be found in Appendix C.1. For the BDA method, a sequence from 10 to 100,000 kilometers separated by increments of 10 kilometers was used as the vector of candidate bandwidth values for each OTU covariate.

As in Section 4.4, the average negative log likelihood values across folds are also calculated after tuning for MAE performance, and can be found in Appendix C.2.

Table 4.1 gives the average 5-fold MAE and associated standard deviations of the 1,133 microbial dust sample location predictions for each of the competing methods under the optimal model settings. Figures 4.3 and 4.4 provide the accompanying visualization of the location predictions for the Coordinate DL and Radial Basis methods using the model settings which resulted in the lowest average MAE, while Figure 4.2 gives the accompanying visualization for BDA. Overall, the deep learning methods outperformed BDA in terms of the average 5-fold MAE for all but one of the model settings.

The top setting for Radial Basis DL was 30 nodes feeding into the hidden and output layers with 5 controls and the top setting for Coordinate DL was 50 nodes feeding into the hidden and output layers with 5 controls. The average MAE decreased as the number of controls increased across all settings. For the Radial Basis DL approach, this contrasts with the result in the simulation setting where increasing the number of controls did not decrease the average MAE. The average MAE decreased as the number of nodes feeding into the hidden and output layers increased from 30 to 50 in three of the four settings,
Table 4.1: A table of the average MAE and accompanying standard deviations for 1,133 microbial dust sample location predictions estimated using the Bayesian Discriminant Analysis, Radial Basis Deep Learning, and Coordinate Deep Learning methods across 5 folds. The deep learning methods are evaluated using each combination of either 30 or 50 nodes feeding into the hidden and output layers and 2 or 5 controls. Bayesian Discriminant Analysis was run using a sequence of candidate bandwidth values from 10 km to 100,000 km separated by increments of 10 km. The methods are evaluated on a grid of 6,266 candidate locations covering the continental USA. The geolocated distances from the true locations are calculated using the great circle distance in kilometers.

<table>
<thead>
<tr>
<th>Method</th>
<th>Neural Network Node Structure</th>
<th>Number Of Controls</th>
<th>Average 5-Fold MAE (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDA</td>
<td>N/A</td>
<td>N/A</td>
<td>409 (33)</td>
</tr>
<tr>
<td>Coordinate DL</td>
<td>30 Nodes In Hidden And Output Layers</td>
<td>2</td>
<td>360 (47)</td>
</tr>
<tr>
<td></td>
<td>50 Nodes In Hidden And Output Layers</td>
<td>5</td>
<td>315 (15)</td>
</tr>
<tr>
<td>Radial Basis DL</td>
<td>30 Nodes In Hidden And Output Layers</td>
<td>2</td>
<td>456 (35)</td>
</tr>
<tr>
<td></td>
<td>50 Nodes In Hidden And Output Layers</td>
<td>5</td>
<td>354 (40)</td>
</tr>
</tbody>
</table>

**Figure 4.2:** A map of 1,133 microbial dust sample locations and corresponding location predictions by the Bayesian Discriminant Analysis method. The method is evaluated on an evenly spaced grid covering the continental USA.
Figure 4.3: A map of 1,133 microbial dust sample locations and corresponding location predictions by the Coordinate Deep Learning method. The method is evaluated on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 50 nodes feeding into the hidden and output layers, 5 controls, an initial learning rate of 4, and a ridge penalty of 0.00005.
Figure 4.4: A map of 1,133 microbial dust sample locations and corresponding location predictions by the Radial Basis Deep Learning method. The method is evaluated on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 30 nodes feeding into the hidden and output layers, 5 controls, an initial learning rate of 3, and a ridge penalty of 0.0001.
although the gains in accuracy were often minimal compared to the gains achieved by increasing the number of controls.

The Coordinate DL method outperformed the Radial Basis DL method across all of the model settings in terms of average 5-fold MAE. The reduction in average MAE was larger for Radial Basis DL when increasing the number of controls compared to Coordinate DL, however the average MAE for Coordinate DL with 2 controls was similar to the average MAE for Radial Basis DL with 5 controls.

The modes of the deep learning model predicted distributions were selected as the point estimates, however the deep learning method can estimate the full conditional distribution. For examples of full conditional distribution predictions obtained from using the deep learning methods as well as a brief discussion on distributional characteristics from these examples, see Appendix C.3.

4.6 Discussion

In this chapter, we propose a pliable geolocation method taking advantage of IPP model properties in order to enable use of a FFNN for estimation without needing to partition the domain. The FFNN input layer is specified to either include radial basis functions or location coordinates as the Radial Basis DL and Coordinate DL methods. A simulation study and forensic geolocation case study are conducted to compare these methods. Results from both studies show our methods outperforming BDA in location prediction for most of the settings that were considered. Our method shows promise as a geolocation prediction technique for spatially correlated point pattern data.

Further investigation is needed to determine the utility of Radial Basis DL compared to Coordinate DL, as neither method consistently outperformed the other for both the simulation and case study. We did not thoroughly explore the methodological options available for the Radial Basis DL method. The density of our single-resolution, evenly spaced grid layout was kept constant for both the simulation and case study. Perhaps altering the grid density could have improved Radial Basis DL performance. Alternatively, perhaps a multi-resolution grid could be specified, as simulation results in Nychka et al. (2015) demonstrate these grids can provide accurate and flexible approximations. Additionally, the single-resolution knot layout could allow for a more spatially-informed penalty such as a conditional autoregressive (CAR) penalty that weights the basis function parameters based on spatial adjacency (Besag et al. 1991). The CAR penalty modification could also be implemented for a multi-resolution grid if the spatial adjacency definition is appropriately determined.
Future work is needed to uncover the full capabilities of our method as well as compare it with the DeepSpace geolocation method in Grantham et al. (2019). Huberman et al. (2020a) made a fairly analogous methodological comparison in the univariate response space between their method and DDR and showed the superiority of their method for multiple simulation scenarios. We would expect a similar takeaway between our method and DeepSpace. As in Huberman et al. (2020a), this chapter serves as a proof of concept but is computationally limited in terms of increasing the complexity of the FFNN. An integration of our model structure with TensorFlow would allow this method to optimally perform in more complex geolocation data scenarios.

We use an IPP model with a log-linear intensity function to take advantage of the relatively straightforward maximum likelihood estimation process. Unfortunately, the IPP does not account for spatial attraction or repulsion between observations as well as point process models with more difficult evaluation algorithms like Cox processes or Gibbs processes. We attempt to discretely approximate an LGCP through the usage of radial basis functions, but perhaps a generalization of our model to allow for random effects could be more effective for geolocation when these spatial dependence patterns are present.

As with most methods involving the evaluation of deep learning models, the tuning process can be difficult. We considered the initial learning rate and the ridge penalty as tuning parameters, which resulted in many potential settings and a tuning process that may not have been sufficiently thorough in terms of exploring potential models. Perhaps the increased number of controls not improving the Radial Basis DL approach in the simulation study was a result of an incomplete tuning process rather than a methodological feature. Similarly, perhaps the Radial Basis DL approach could have outperformed Coordinate DL in the application if more tuning parameters (including the layout of the basis knot grid) were considered. Figure 4.4 shows fewer predicted locations compared to the methods used in Figures 4.3 and 4.2, a possible indicator that the gradient descent algorithm for this method did not converge as well as it did for Coordinate DL.
References


Alex L Rojas, Christopher R Genovese, Christopher J Miller, Robert Nichol, and Larry Wasserman. Conditional density estimation using finite mixture models with an application to astrophysics. 2005.


Arco J Strien et al. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models.


Nguyen Thanh Tung, Joshua Zhexue Huang, Imran Khan, Mark Junjie Li, Graham Williams, et al. Extensions to quantile regression forests for very high-dimensional data. 2014.


Appendix A

Estimating the drivers of species distributions with opportunistic data using mediation analysis

A.1 Occupancy model evaluation details

The occupancy model in 2.2.2 can be considered similar to the model in 2.2.2, with $Y_0\delta$ inserted additively into the mean of $Z(s)$. If considered this way, then the parameter estimation process described below apply to both occupancy models proposed in this paper. We present only the updates for the spatial occupancy model described in 2.2.2 for simplicity. This model can be written as

$$Y_1(s)|O(s), p \sim \text{Binomial}(N(s), pO(s)), \quad \text{where } O(s) = 1(Z(s) > 0), \quad \text{(A.1)}$$

$$Z(s)|\theta(s), X(s), \lambda \sim \text{Normal}(\theta(s) + X(s)\beta, 1 - \lambda), \quad \text{(A.2)}$$

$$\theta \sim \text{MVN}(0, \lambda \Sigma(\kappa)). \quad \text{(A.3)}$$

$\Sigma$ represents the exponential correlation function. The chosen priors are $\beta \sim \text{MVN}(0, \tau^2 I_n)$, $p \sim \text{Uniform}(0, 1)$, logit $\lambda \sim \text{Normal}(e, h)$, and $\log \kappa \sim \text{Normal}(r, w)$.

We select $\tau^2 = 10^2$, $e = r = 0$, and $h = w = 1$ as the hyper-parameters for the priors. $\beta, \theta, \lambda, \kappa, Z, O,$ and $p$ are estimated using Markov Chain Monte Carlo (MCMC) methods. Initial parameter values are chosen and then the parameters are sequentially updated at each iteration by randomly sampling from the full conditional posterior distributions with the other parameters held fixed at their current values. For each run of
the model, there are 3000 iterations with a 500 iteration burn-in threshold. Convergence was evaluated through inspecting trace plots for each parameter.

The full conditional distributions for $p$, $\theta$, and $\beta$ are

\begin{align*}
    p|O(s) & \sim \text{Beta} \left( \sum_{O(s)=1} Y_1(s) + 1, \sum_{O(s)=1} N(s) - \sum_{O(s)=1} Y_1(s) + 1 \right), \quad (A.4) \\
    \theta|\lambda, \Sigma, Z, \beta & \sim \text{MVN} \left( \left[ \frac{1}{1 - \lambda} I_n + \frac{1}{\lambda} \Sigma^{-1} \right]^{-1} \left[ \frac{1}{1 - \lambda} (Z - X\beta) \right], \left[ \frac{1}{1 - \lambda} I_n + \frac{1}{\lambda} \Sigma^{-1} \right]^{-1} \right), \text{ and} \quad (A.5) \\
    \beta|\lambda, Z, \theta & \sim \text{MVN} \left( \left[ \frac{1}{1 - \lambda} X^T X + \frac{1}{\tau^2} I_n \right]^{-1} \left[ \frac{1}{1 - \lambda} X^T (Z - \theta) \right], \left[ \frac{1}{1 - \lambda} X^T X + \frac{1}{\tau^2} I_n \right]^{-1} \right). \quad (A.6)
\end{align*}

$Z$ and $O$ are simultaneously updated. We define

\begin{align*}
    A &= \text{Binomial}(Y_1(s); N(s), p) \left[ 1 - \Phi \left( \frac{-\theta(s) - X(s)\beta}{\sqrt{1 - \lambda}} \right) \right], \text{ and} \quad (A.7) \\
    B &= \text{Binomial}(Y_1(s); N(s), 0) \left[ \Phi \left( \frac{-\theta(s) - X(s)\beta}{\sqrt{1 - \lambda}} \right) \right]. \quad (A.8)
\end{align*}

Then, $O(s)$ is conditionally updated as

\begin{equation}
    O(s)|\lambda, \theta(s), \beta, Z(s) \sim \text{Binomial}(1, \Pr(Z(s) > 0)), \quad (A.9)
\end{equation}

where

\begin{equation}
    \Pr(Z(s) > 0) = \frac{A}{A + B}. \quad (A.10)
\end{equation}
Then, $Z(s)$ is immediately updated as

$$Z(s)|\lambda, \beta, \theta(s), O(s) \sim \begin{cases} \text{Truncated Normal}_{(0, \infty)} \left( \theta(s) + X(s)\beta, 1 - \lambda \right) & O(s) = 1 \\ \text{Truncated Normal}_{(-\infty, 0)} \left( \theta(s) + X(s)\beta, 1 - \lambda \right) & O(s) = 0. \end{cases}$$

(A.11)

The $\lambda$ and $\kappa$ parameters are updated using Metropolis-Hastings sampling with random-walk normal proposal distributions and standard normal priors. The proposal distributions were given for the logit transform of $\lambda$ and the log transform of $\kappa$ as

$$\log x_c \sim \text{Normal}(x, \sigma^2)$$

(A.12)

where $\sigma^2$ represents the tuned variance parameter and $x$ and $x_c$ represent the current and candidate value of the parameter being updated, respectively. The full conditional distribution for $\kappa$ is proportional to

$$\frac{1}{\lambda^{n/2} |\Sigma(\kappa)|^{1/2}} \exp \left( -\frac{1}{2} \theta^T \frac{1}{\lambda} \Sigma(\kappa)^{-1} \theta \right) \phi(x)$$

(A.13)

where $x$ represents the parameter being updated and $\phi$ is the standard normal pdf density function. The full conditional distribution for $\lambda$ is proportional to Eq. A.13 multiplied by

$$\frac{1}{(1 - \lambda)^{n/2}} \exp \left( -\frac{1}{2} (Z - \theta - X\beta)^T \frac{1}{1 - \lambda} (Z - \theta - X\beta) \right).$$

(A.14)
Appendix B

Nonparametric conditional density estimation in a deep learning framework for short-term forecasting

B.1 Case-control sampling justification

We exploit the relationship between an inhomogenous Poisson process (IPP) model and our problem to re-frame case and control samples into presence and background samples from presence-only datasets. Presence-only data is generally used with species distribution modeling where surveyors record all species presences within a pre-specified region along with randomly sampled background data across the region. Fithian and Hastie (2013) lay out how to model presence-only data in a two-dimensional spatial domain using an IPP model. For our method, the domain for cases (and controls) is [0, 1] as previously noted. The IPP requires an intensity function $\lambda$ to be specified which represents the likelihood of a case being present at any location in the given domain. The average intensity function for response value $z$ over the domain is given as

$$\Lambda = \int_0^1 \lambda(z)dz.$$  (B.1)

An IPP model with intensity function $\lambda$ gives the probability distributions for both the total number of cases as well as the locations of those cases. Conditional on the number of cases (governed by a Poisson distribution with mean $\Lambda$), the locations of the
cases are independently and identically distributed as

\[ \Pr(Z = z) = \frac{\lambda(z)}{\Lambda}. \]  \hspace{1cm} (B.2)

If we define \( \lambda(z) = e^{q(z, X)} \), then the IPP model becomes

\[ \Pr(Z = z) = \frac{e^{q(z, X)}}{\int_0^1 e^{q(z, X)} \, dz}. \] \hspace{1cm} (B.3)

We recognize this distribution form in the continuous logistic transformation formula given in 3.2. The cases are independently and identically distributed, so the log likelihood objective function is

\[ \ell(\theta | Z, X) = \sum_{i=1}^{n} \left[ q(z_i, X_i; \theta) - \log \left( \int_0^1 q(z, X_i; \theta) \, dz \right) \right]. \] \hspace{1cm} (B.4)

where \( i = 1, \ldots, n \) and \( \theta \) is the parameter vector for the selected \( q \) model, as before. Fithian and Hastie (2013) discuss how to evaluate the integral in the denominator by approximating it using a finite set of control (background) samples. Let \( k = 1, \ldots, K \) index the control samples, so that \( z_{ik} \) denotes the \( k \)th control value for observation \( i \). Then, the log likelihood objective function with an approximated integral becomes

\[ \ell(\theta | Z, X) = \sum_{i=1}^{n} \left[ q(z_i, X_i; \theta) - \log \left( e^{q(z, X_i; \theta)} + \sum_{k=1}^{K} e^{q(z_{ik}, X_i; \theta)} \right) \right]. \] \hspace{1cm} (B.5)

We recognize this as the log likelihood objective function in 3.8. Fithian and Hastie do not approximate the integral using a \( q \) function with a case plus a set of controls, instead only using the set of controls. However the case is only providing more information about the integral than the controls would on their own so we do not expect this to produce any inconsistencies or biases. They state that the control points can be uniformly sampled from the domain, which implies that selecting as few as \( K = 1 \) controls still can provide valid inference about the target distribution function. Fithian and Hastie also state that the control points can be chosen through weighted sampling, referring to using quadrature weights. The additional ridge penalty component in 3.8 could have been added to the IPP likelihood as well.
Note that for presence-only data, Fithian and Hastie detail possible sampling bias issues that can arise due to imperfect detection of presences during data collection. However, we do not think detectability issues are relevant to our usage of this framework. We make the assumption that the detectability parameter equals 1 in our context, so we do not have to worry about sampling bias in this respect.

**B.2 Computational details, model tuning, and model specifications**

We use a variety of techniques to determine the optimal set of parameter estimates for the chosen model.

**B.2.1 Polynomial regression model**

For the polynomial regression model, we can manipulate the data in order to evaluate the parameters using penalized logistic regression. Recall the polynomial regression model in 3.3. We can express the polynomial model in 3.3 in matrix form as $q(z, X) = X(z, B)\xi$, where $X(z, B)$ is the full covariate matrix as a function of the transformed response $z$ and the highest polynomial power $B$ and $\xi = [\xi_1, ..., \xi_B]$ is the associated parameter vector. Then, we can rewrite the objective function in 3.8 using the polynomial regression from 3.3 as

$$
\ell(\theta|Z_i, X_i) = \sum_{i=1}^{n} \left\{-\log \frac{e^{X_i(z_i, B)\xi}}{1 + e^{X_i(z_i, B)\xi}}\right\} + \mathcal{S}||\theta||^2 \tag{B.6}
$$

$$
= \sum_{i=1}^{n} \left\{-\log \frac{1}{1 + e^{X_i(z_i, B)\xi}}\right\} + \mathcal{S}||\theta||^2. \tag{B.7}
$$

We recognize the form of this objective function for a logistic regression where every binary outcome equals 1. If we add a small amount of dummy observations where the covariate and outcome values are all 0, we can evaluate our parameter vector using penalized logistic regression. For each run of the polynomial regression case-control method, we used two dummy observations. The GLMNet package in R evaluates these parameters extremely fast, making this method practical and convenient (Friedman et al. 2009).

We ran preliminary simulation runs on a smaller number of datasets for a variety of
ridge penalties to determine the appropriate penalty values to use for each scenario of the simulation and the application. For Scenario 1, 2, and 4 the ridge penalties used were 0.025, 0.05, and 0.05, respectively. For Scenario 3, the ridge penalty used for 200 observations was 0.025, whereas the ridge penalty used for the other two sample size settings was 0.01.

For the application, we considered a variety of ridge penalties to determine which penalty minimized the mean 5-fold CRPS. For lag 3 and lag 6, a ridge penalty of 0.000001 and 0.0005 were chosen, respectively.

B.2.2 Deep learning model

For the deep learning case-control approximation, we cannot manipulate the data like in B.7 and instead rely on gradient descent to evaluate the parameters.

Deep learning models can be difficult to train with a basic gradient descent algorithm. We extend or alter the basic gradient descent algorithm in a multitude of ways designed to improve parameter estimation. Mini-batch gradient descent is used with batch sizes of 50 and an initial step size of 1. For the case study, the batch sizes are approximate as the number of observations is not divisible by 50. The models are run for 600 gradient descent steps for the simulation and 300 gradient descent steps for the case study, respectively. Adaptive moment estimation (ADAM) is implemented along with a step decay that halves the initial step size every 100 steps for the simulation models and 50 steps for the case study models (Kingma and Ba 2014). We apply the batch normalization algorithm described in Ioffe and Szegedy (2015). Model weights are initiated using the He initialization scheme, while the batch normalization shift and scale parameters are initialized to 0 and 1, respectively (He et al. 2015).

The gradient descent algorithm requires the calculation of a gradient vector which contains the first derivative value of the objective function with respect to each parameter. We obtain these values using back-propagation. The individual chain rule components for the gradient vector calculations for a single observation and case/control value are given in B.8. The gradient vector can be calculated by summing the appropriate terms across observations and cases/controls.
\[
\begin{align*}
\frac{\partial \ell}{\partial q(z_i, X_i; \theta)} &= \frac{1}{e^{q(z_i, X_i; \theta)} + e^{q(z^*_i, X_i; \theta)}} \\
\frac{\partial q(z^*_i, X_i; \theta)}{\partial q(z_i, X_i; \theta)} &= \frac{\partial q(z^*_i, X_i; \theta)}{\partial \delta_t} = f_A(H_t) \\
\frac{\partial q(z_i, X_i; \theta)}{\partial f_A(H_t)} &= \frac{\partial q(z^*_i, X_i; \theta)}{\partial f_A(H_t)} = \delta_t \\
\frac{\partial f_A(H_t)}{\partial H_t} &= \begin{cases} 
1 & \text{if } H_t > 0 \\
\alpha \exp H_t & \text{if } H_t \leq 0
\end{cases} \\
\frac{\partial H_t}{\partial f_A(I_r)} &= \gamma_{tr} \\
\frac{\partial H_t}{\partial \gamma_{tr}} &= 1 \\
\frac{\partial H_t}{\partial \gamma_{tr}} &= f_A(I_r) \\
\frac{\partial f_A(I_r)}{\partial I_r} &= \begin{cases} 
1 & \text{if } I_r > 0 \\
\alpha \exp I_r & \text{if } I_r \leq 0
\end{cases} \\
\frac{\partial I_r}{\beta_{or}} &= 1 \\
\frac{\partial I_r}{\beta_{or}} &= z_i \\
\frac{\partial I_r}{\beta_{tr}} &= X_{ji}.
\end{align*}
\]

The components for the batch normalization parameters can be obtained following the steps in (Ioffe and Szegedy 2015).

We ran preliminary simulation runs on a smaller number of datasets for a variety of ridge penalties to determine the appropriate penalty values to use for each scenario of the simulation and the application. For simulation Scenario 1, we used 0.025, 0.002, and 0.00025 as the penalties for 200, 1000, and 4000 observations respectively. For simulation Scenario 2, we used 0.015, 0.0025, and 0.001 as the penalties for 200, 1000, and 4000 observations respectively. For simulation Scenario 3, we used 0.0075, 0.001, and 0.001 as the penalties for 200, 1000, and 4000 observations respectively. For simulation Scenario 4, we used 0.01, 0.001, and 0.0005 as the penalties for 200, 1000, and 4000 observations respectively.
For the application, we considered a variety of ridge penalties and initial learning rates for the deep learning IPP approximation method. For both lags, the tuning settings which optimized the mean 5-fold CRPS were a ridge penalty of 0.000001 and an initial learning rate of 1.

B.3 Predicted distribution plots for individual folds

The predicted distribution plots for a range of covariate values using the full tropical cyclone dataset is given in Section 3.5, however the results in Table 3.2 were produced by obtaining separate models for each testing dataset fold. Here, we present the predicted distribution plots for each individual fold for both methods and both lags for thoroughness. The predicted response distributions plotted below can vary somewhat compared to the overall predicted response distributions fitted using the full tropical cyclone dataset. For instance, the deep learning IPP approximation for lag 3 and fold 3 predicts a somewhat bimodal response distribution given a larger covariate value input.
Figure B.1: Deep learning IPP approximation ($K = 20$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 1 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. Pr($Y = y|X$) refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 
Figure B.2: Polynomial MCC approximation ($K = 1$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 1 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $\Pr(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 

94
Figure B.3: Deep learning IPP approximation ($K = 20$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 2 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $\Pr(Y = y | X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 
Figure B.4: Polynomial MCC approximation ($K = 1$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 2 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. Pr($Y = y | X$) refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y | X$ refers to the conditional response value $Y$ given the covariate $X$. 

96
Figure B.5: Deep learning IPP approximation \((K = 20)\) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 3 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. \(Pr(Y = y|X)\) refers to the relative probability that the maximum 10-meter wind speed \(Y = y\) occurs given the HWRF-forecasted maximum 10-meter wind speed value \(X\). \(Y|X\) refers to the conditional response value \(Y\) given the covariate \(X\).
Figure B.6: Polynomial MCC approximation \((K = 1)\) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with with observations from fold 3 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. \(Pr(Y = y|X)\) refers to the relative probability that the maximum 10-meter wind speed \(Y = y\) occurs given the HWRF-forecasted maximum 10-meter wind speed value \(X\). \(Y|X\) refers to the conditional response value \(Y\) given the covariate \(X\).
Figure B.7: Deep learning IPP approximation ($K = 20$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 4 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $Pr(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 
Figure B.8: Polynomial MCC approximation ($K = 1$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 4 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $\Pr(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 
Figure B.9: Deep learning IPP approximation ($K = 20$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 5 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $\Pr(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 

---

Lag 3 Maximum 10–m Wind Speed Predicted Conditional Distribution

Deep Learning, Fold 5

Approximate Covariate Value

- $X = 20$
- $X = 34$
- $X = 49$
- $X = 64$
- $X = 79$
- $X = 93$
- $X = 108$
- $X = 123$
- $X = 137$
- $X = 152$

---

Lag 6 Maximum 10–m Wind Speed Predicted Conditional Distribution

Deep Learning, Fold 5

Approximate Covariate Value

- $X = 20$
- $X = 35$
- $X = 50$
- $X = 65$
- $X = 80$
- $X = 94$
- $X = 109$
- $X = 124$
- $X = 139$
- $X = 154$
Figure B.10: Polynomial MCC approximation ($K = 1$) conditional maximum 10-meter wind speed distribution predictions for lag 3 and 6 using model constructed with observations from fold 5 as the testing dataset and all other observations as the training dataset. Estimated conditional response probabilities for 100 equally spaced quantiles sequenced between 0.5 and 99.5 are displayed with linear interpolation between quantiles. The 0.5th and 99.5th quantile density function values are rounded to 0. $P(Y = y|X)$ refers to the relative probability that the maximum 10-meter wind speed $Y = y$ occurs given the HWRF-forecasted maximum 10-meter wind speed value $X$. $Y|X$ refers to the conditional response value $Y$ given the covariate $X$. 
Appendix C

Geolocation for spatial point pattern data in a deep learning framework

C.1 Model tuning and model specifications

C.1.1 Simulation

A mini-batch stochastic gradient descent algorithm was used to evaluate the deep learning model. All models were run for 750 steps with a batch size of 50. For each dataset, a set of initial parameter values were obtained using the He initialization scheme described in He et al. (2015), and the algorithm was evaluated. Adaptive moment estimation (ADAM) is incorporated in the parameter updates to individualize learning rates to parameters and improve convergence speed (Kingma and Ba 2014). The batch normalization algorithm described in Ioffe and Szegedy (2015) is applied to improve optimization. Covariate and basis function inputs were normalized using the given training dataset mean and variance. Gradient vector calculations were obtained using backpropagation for each observation and control value and then appropriately summing individual components across observations and controls.

Each deep learning model for each scenario was tuned for a variety of initial learning rates (between 1 and 5) and ridge penalty values (between 0 and .5) by first running the model for a smaller number of datasets and selecting the tuning parameters which optimized MAE performance. A set of the top ridge penalties were determined for each method and models with these ridge penalties and each initial learning rate were fit using all 100 generated datasets. The ridge penalties and initial learning rates which optimized model performance were selected as the final settings. This process was performed using 2 controls, and then the methods were run with their final settings for 5 controls as well.
The negative log likelihood values were calculated after selecting these tuning settings.

For the Coordinate DL method, a ridge penalty of 0.0005 and initial learning rate of 4 were selected for 500 observations and 25 OTUs. A ridge penalty of 0.005 and initial learning rate of 2 were selected for 500 observations and 100 OTUs. For the Radial Basis DL method, a ridge penalty of 0.0005 and initial learning rate of 2 were selected for 500 observations and 25 OTUs. A ridge penalty of 0.0005 and initial learning rate of 4 were selected for 500 observations and 100 OTUs.

C.1.2 Case Study

The algorithm used to evaluate the deep learning model for the case study is similar to the algorithm described above for the simulation. Batch sizes were initially set at 50 for each batch and the remaining unassigned observations were sequentially added to the batches.

As mentioned in 4.5, models were fit using multiple different specifications for the number of nodes and controls. For each of these settings, a variety of initial learning rates (between 1 and 5) and different ridge penalties (between 0 and .5) were considered. The 5-fold dataset was for all of these tuning parameters at each combination of nodes and controls and the tuning parameters which minimized the average MAE were selected. The final tuning parameters for each of the considered models are given in the following list.

- Coordinate DL, 30 Nodes, 2 Controls:
  - Ridge penalty 0.00005 and initial learning rate 3.
- Coordinate DL, 30 Nodes, 5 Controls:
  - Ridge penalty 0.0001 and initial learning rate 5.
- Coordinate DL, 50 Nodes, 2 Controls:
  - Ridge penalty 0.001 and initial learning rate 5.
- Coordinate DL, 50 Nodes, 5 Controls:
  - Ridge penalty 0.00005 and initial learning rate 4.
- Radial Basis DL, 30 Nodes, 2 Controls:
  - Ridge penalty 0.0005 and initial learning rate 2.
• Radial Basis DL, 30 Nodes, 5 Controls:
  – Ridge penalty 0.0001 and initial learning rate 3.
• Radial Basis DL, 50 Nodes, 2 Controls:
  – Ridge penalty 0.0001 and initial learning rate 1.
• Radial Basis DL, 50 Nodes, 5 Controls:
  – Ridge penalty 0.0001 and initial learning rate 2.

C.2 Negative log likelihood estimation results

The methods in this paper were evaluated using average MAE as the performance metric for the purposes of geolocation. For each of the geolocation methods, the mode of the predicted conditional distributions was obtained to use as the point estimates for the MAE calculation.

The average negative log likelihood is an alternative metric which provides insight into the full conditional distribution performance of the method by comparing the fit of the predicted conditional distribution to the true conditional distribution. For each observation, the predicted probability at the grid location closest to the true location is obtained. The negative mean of the logarithm of the probabilities across all testing set observations is then calculated as this performance metric. A smaller value indicates the method predicted higher probabilities near the true locations, whereas a larger value indicates the method predicted lower probabilities near these locations.

We present the average negative log likelihood results for all of the methods used in the simulation and case studies here. The deep learning methods were tuned to minimize average MAE and not the average negative log likelihood, so the results here do not necessarily follow the same patterns that optimally tuned results might.

Tables C.1 give the negative log likelihood information and associated standard deviations for all of the methods and settings employed in Section 4.4. In the simulation study, BDA outperforms the deep learning methods by this metric. For the simulation, Coordinate DL produces smaller negative log likelihood values than Radial Basis DL on average. For three of the four scenarios, increasing the number of controls decreases the negative log likelihood value. An increase in the number of OTUs decreases and increases the average negative log likelihood for Radial Basis DL and Coordinate DL, respectively.
Table C.1: A table of the average negative log likelihood values for each of the methods under each of the scenarios along with their associated standard deviations. For each observation, a probability distribution across the spatial domain is predicted conditional on the OTU covariates, with the probability at the true spatial location used to calculate the contribution to the negative log likelihood.

<table>
<thead>
<tr>
<th>Method</th>
<th>Number Of Controls</th>
<th>Number Of Covariates</th>
<th>25 OTUs</th>
<th>100 OTUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDA</td>
<td>N/A</td>
<td>10.07 (.33)</td>
<td>10.33 (.32)</td>
<td></td>
</tr>
<tr>
<td>Radial Basis DL</td>
<td>2</td>
<td>34.78 (3.62)</td>
<td>28.98 (2.26)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>34.36 (3.66)</td>
<td>30.71 (1.97)</td>
<td></td>
</tr>
<tr>
<td>Coordinate DL</td>
<td>2</td>
<td>14.54 (1.41)</td>
<td>21.13 (2.51)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>13.08 (1.01)</td>
<td>19.30 (2.18)</td>
<td></td>
</tr>
</tbody>
</table>

Table C.2: A table of the average negative log likelihood and accompanying standard deviations for 1,133 microbial dust sample location predictions estimated using the Bayesian Discriminant Analysis, Radial Basis Deep Learning, and Coordinate Deep Learning methods across 5 folds. The deep learning methods are evaluated using each combination of either 30 or 50 nodes feeding into the hidden and output layers and 2 or 5 controls. Bayesian Discriminant Analysis was run using a sequence of candidate bandwidth values from 10 km to 100,000 km separated by increments of 10 km. The methods are evaluated on a grid of 6,266 candidate locations covering the continental USA. The geolocated distances from the true locations are calculated using the great circle distance in kilometers.

<table>
<thead>
<tr>
<th>Method</th>
<th>Neural Network Structure</th>
<th>Number Of Controls</th>
<th>Number Of Controls</th>
<th>Average 5-Fold MAE (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDA</td>
<td>N/A</td>
<td>N/A</td>
<td>25.56 (2.91)</td>
<td></td>
</tr>
<tr>
<td>Radial Basis DL</td>
<td>30 Hidden Layer and</td>
<td>2</td>
<td>17.24 (1.64)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Output Layer Nodes</td>
<td>5</td>
<td>19.74 (3.38)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50 Hidden Layer and</td>
<td>2</td>
<td>20.67 (1.67)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Output Layer Nodes</td>
<td>5</td>
<td>17.95 (1.10)</td>
<td></td>
</tr>
<tr>
<td>Coordinate DL</td>
<td>30 Hidden Layer and</td>
<td>2</td>
<td>18.77 (4.06)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Output Layer Nodes</td>
<td>5</td>
<td>17.04 (2.01)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50 Hidden Layer and</td>
<td>2</td>
<td>12.12 (2.68)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Output Layer Nodes</td>
<td>5</td>
<td>17.73 (2.50)</td>
<td></td>
</tr>
</tbody>
</table>
Table C.2 gives the average negative log likelihood values for all of the methods and settings employed in Section 4.5. The deep learning methods outperform BDA in the case study, in contrast with the results in Table C.1. Coordinate DL results in lower average negative log likelihood values compared to Radial Basis DL, consistent with the simulation study results. Increasing the number of controls resulted in a lower average negative log likelihood in two of the four scenarios. Similarly, increasing the number of nodes resulted in a lower average negative log likelihood in two of the four scenarios.

C.3 Predicted conditional PDF plots for case study

The deep learning methods determine a prediction by selecting the mode location of the full conditional distribution using an evenly spaced grid of candidate location values across the continental USA. It is infeasible to present conditional PDF plots for each microbial fungal sample location, so we present an example of a location sparsely surrounded by other microbial fungal samples on the West Coast and a location densely surrounded by other microbial fungal samples on the East Coast. The plotted locations contain the largest PDF values and account for around 95% of the total PDF probability mass. Both examples are from the first testing location fold.

As noted in 4.5, there were a sizable number of locations relatively optimally predicted by each of the methods. Consequently, these four PDF plots are not intended to be representative examples and the discussed characteristics should not be assumed to apply to every microbial fungal sample in a sparsely or densely populated area of the USA.

Figures C.1 and C.2 give the conditional density predictions for a densely surrounded microbial fungal sample from North Carolina using the optimally tuned Coordinate DL and Radial Basis DL methods. In this example, Radial Basis DL almost perfectly geolocates the sample, with probability mass surrounding the true location. Coordinate DL also geolocates the sample well, with the location falling within the 95% probability mass. The Radial Basis DL predicted PDF has a higher peak and is more compact.

Figures C.3 and C.4 give the conditional density predictions for a sparsely surrounded microbial fungal sample from Idaho using the optimally tuned Coordinate DL and Radial Basis DL methods. 95% of the Coordinate DL predicted probability mass is concentrated in western Oregon, whereas 95% of the Radial Basis DL predicted probability mass is concentrated in New York. The peak PDF value is larger for Radial Basis DL, but its PDF estimate is very inaccurate here. Neither of the example predicted PDF plots for the sparsely surrounded fungal sample contain the true location within the 95% probability region, which also occurs for many other locations not plotted in this section. This
Figure C.1: The predicted PDF plot for a microbial fungal sample from North Carolina using the Radial Basis Deep Learning method with the Bayesian Discriminant Analysis predicted location and true location overlaid. The plotted locations contain the largest PDF values and account for around 95% of the total PDF probability mass. The microbial fungal sample was purposely selected as an example of a location densely surrounded by other microbial fungal samples. The predicted PDF locations are on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 30 nodes for the hidden and output layers, 5 controls, an initial learning rate of 3, and a ridge penalty of 0.0001.
Figure C.2: The predicted PDF plot for a microbial fungal sample from North Carolina using the Coordinate Deep Learning method with the Bayesian Discriminant Analysis predicted location and true location overlaid. The plotted locations contain the largest PDF values and account for around 95% of the total PDF probability mass. The microbial fungal sample was purposely selected as an example of a location densely surrounded by other microbial fungal samples. The predicted PDF locations are on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 50 nodes feeding into the hidden and output layers, 5 controls, an initial learning rate of 4, and a ridge penalty of 0.00005.
Figure C.3: The predicted PDF plot for a microbial fungal sample from Idaho using the Radial Basis Deep Learning method with the Bayesian Discriminant Analysis predicted location and true location overlaid. The plotted locations contain the largest PDF values and account for around 95% of the total PDF probability mass. The microbial fungal sample was purposely selected as an example of a location sparsely surrounded by other microbial fungal samples. The predicted PDF locations are on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 30 nodes for the hidden and output layers, 5 controls, an initial learning rate of 3, and a ridge penalty of 0.0001.
Figure C.4: The predicted PDF plot for a microbial fungal sample from Idaho using the Coordinate Deep Learning method with the Bayesian Discriminant Analysis predicted location and true location overlaid. The plotted locations contain the largest PDF values and account for around 95% of the total PDF probability mass. The microbial fungal sample was purposely selected as an example of a location sparsely surrounded by other microbial fungal samples. The predicted PDF locations are on an evenly spaced grid covering the continental USA. The tuning parameters used for the method producing these predictions were 50 nodes for the hidden and output layers, 5 controls, an initial learning rate of 4, and a ridge penalty of 0.00005.
suggests that our method could improve in terms of prediction accuracy with more data or a more sophisticated deep learning model.

An exploratory analysis of Radial Basis DL for locations in the first testing fold shows that the method poorly predicts many of the West Coast samples as being located in New York but more accurately predicts locations further to the East. This issue might indicate the presence of suboptimal tuning settings, algorithmic convergence, or basis knot grid selections as discussed in Section 4.6 rather than a fundamental issue of the Radial Basis DL method for this data.