Integrating grouped and ungrouped data: the point process case

by

Irma del Consuelo Hernández Magallanes

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David R. Brillinger, Chair
Professor Deborah A. Nolan
Professor George G. Judge
Dr. Haiganoush K. Preisler

Fall 2010

Integrating grouped and ungrouped data: the point process case

Abstract

Integrating grouped and ungrouped data: the point process case

by

Irma del Consuelo Hernández Magallanes

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor David R. Brillinger, Chair

Grouped data is a topic that goes back to the end of the nineteenth century at least. Kulldorff [34] refers to grouping as a special case of a more general kind of procedure, called partial grouping. A partially grouped sample refers to the case where available information is associated with a collection of disjoint sets partitioning a domain $\mathcal{S}$. The sample space is divided into non-overlapping sets as in $\mathcal{S} = \mathcal{B} \cup \mathbf{\Omega}$ where $\mathbf{\Omega} = \left(\bigcup_{j=1}^{J} \Omega_j \right)$ and the Ω_j's are themselves non-overlapping. In some of these sets only the counts of observations are recorded (grouped data) while the individual values of the observations falling in the other sets are recorded (ungrouped data). This thesis focuses on spatially partially grouped data.

This work is motivated by an interest in modeling the locations and times of wildfire occurrences that happened in the Continental United States in the period from 1986 to 1996. The data cover fires that occurred in federal and non-federal lands. The federal data consisted of each fire's point location (latitude and longitude) $\tau_i = (x_i, y_i) \in \mathbb{R}^2$ where $i = 1, \ldots, I$ while the non-federal fires were aggregated by county $\Omega_j \in \mathbb{R}^2$, $N(\Omega_j) = n_j$ where $j = 1, \ldots, J$.

Wildfires occurrences can be considered as a point process in $\mathcal{S}$. Brillinger, Preisler and Benoit [5] approximate a point process by a binary process. We propose integrating the two levels of aggregate data, points and counts, by modeling the fires as a binary $(0, 1)$ process on space Y_l $l = 1, \ldots, L$. The sample space is partitioned into small pixels arranged in a regular two-dimensional grid. Each pixel either has a fire or not. The numbers of fires in each non-overlapping set are assumed to be independent and to follow a Binomial distribution.

Under the assumption that the wildfire rate is a smooth varying function of space we propose a spatial smoothing method for partially grouped data. This smoother is based on local regression using the binary process to approximate the partial grouped data.

Based on the binary-valued approximation a logit model is used with the the National Fire Danger Rating System fuel model as explanatory variables. The estimated probabilities are included in a map with the associated uncertainty levels.

Gracias a Dios por bendecirme de tantas maneras.
A mi esposo, Erick.
A mis padres, Irma y Juan.
A mis hermanos, Jéssica y Jorge.

Contents

Acknowledgments

First and foremost, I am so deeply grateful to my advisor, Professor David R. Brillinger. He has been a mentor in a plethora of areas, from research and teaching and most important in life in general. I thank him for patiently teaching me how to do research and introducing me to the field of environmental statistics. Professor Brillinger has cheered me up in difficult times either in research or personally. But also he has had the right words to point out the improvement areas. I really appreciate his dedication; he always provided me with prompt, accurate and thorough feedback. This dissertation would simply not have been possible without his guidance.

I am very grateful to Professor Deborah Nolan for her advices and support during my studies. Since the first day in the Ph.D. she has been very approachable and interested in assuring my success in the program. I was very lucky to GSI for her in the "Data Visualization and its Role in the Practice of Statistics: An undergraduate Summer Program in Statistics" in UCLA. It was a lot of fun and also it was great to realize how much passion she put in her work and how the teaching of Statistics is fundamental in the future of this discipline. I have benefited by her comments related to new research that has been done in the display and exploration of spatial data.

I would like to express my gratitude to Professor George Judge who agreed to be part of my committee and was always available for meeting with me. In one of those meetings, using an example of U.S. voters, he approached in an interesting way the problem of estimating parameters when data is partially grouped. I feel excited about pursuing this suggestion!

I am also very grateful to Dr. Haiganoush Preisler, Statistical Scientist in the USDA Forest Service Pacific Southwest Research Station. Dr Haiganoush provides me with great insight and valuable ideas for the modeling of wildfire occurrences. But above all I consider her a good role model of how research and family work great together! I look forward to keep collaborating with her.

The data used for this work was provided by Robert E. Burgan from the USDA Forest Service, Rocky Mountain Research Station. We want to acknowledge the huge amount of work done by the Rocky Mountain Research Station for cleaning and putting together the information.

I am indebted to Conacyt and UC Mexus for the Graduate Fellowship that funded throughout my Ph.D. studies. I thank the research support received from my advisors grants', NSF and also the ASA Section on Business and Economics Statistics.

I am very grateful to the Statistics Department where I have found a propitious environment to grow as a researcher and as a human being. The Professors and Students makes the Department a rich and diverse academic community. I am indebted to the Staff members who have always been very friendly and has helped me diligently solving any problem.

I specially want to thank Steve Drekic, Professor in the Department of Statistics and Actuarial Science at the University of Waterloo, for guiding my Master's Thesis and following up my academic career. I owe special words to Tapen Sinha, Professor in the Actuarial Science Department at the Instituto Tecnológico Autónomo de México (ITAM). Tapen has been a mentor and friend since undergraduate school, he guided my Undergraduate's Thesis and introduced me to the idea of pursuing graduate school. He has always been there for giving me wise advices and he has always believe in my resilience to thrive through the Ph.D. I am also indebted to my professors at ITAM who encouraged and support my dream of pursuing a Ph.D.: Enrique de Alba, Carlos Bosch and Guillermo Grabinsky.

I also want to thank Dr. Manuel Galán and Dr. Jorge de la Vega at Banco de México (BANXICO), who supported my application for pursuing my studies abroad while working with them. I also want to thank Maestra Cristina Higashida for providing the data related to coins and banknotes demand that I analyzed during my studies.

I feel so blessed by having such good friends, that have fulfilled my life with joy. As you can imagine I am running out of time to submit my dissertation but each of you are in my heart and thoughts. I owe you part of this milestone. Liliana, Raúl, Paulo, Sofía and Luciana. Maddox: Cristy and Marce. Laila and Vero. IBM: Rita, Maru, Marianita, Enrique and Fer. ITAM: Marisol and Juan Pablo. BANXICO: Cris and Alfredo, Luisito and Lili, Xóchitl and Alberto. Berkeley: Frances and Shankar, Angie, Charlotte and Josh, Margaret, Cathy, Pil, Victoria. Newbies: Daisy, Moorea, Oleg, Harry, Allan, Jing. IEOR: Michele and James, Aude and Ryan, Heather and Todd, Nina and John, Barak and Nilli. Mexcal: Yara, Isa, Ivan, Hartmut, Diana and Paulina.

I owe my deepest gratitude to my loving family: my parents, Irma and Juan who has always shown me their love in all my endeavors and have taught me to be a good person. My sister, Jessi, it has been so hard to be physical away from you but you are always so close to my heart. My brother in law Jorge who has always being with us and makes my sister so happy. I also want to thank my grandma Mama Chelo for her example and love and also my uncle Fernando. My parents in law, Mary and Pablo that has given me a wonderful husband.

Este párrafo quiero que se único y especial porque está dedicado a mi esposo, Erick. Fuente de amor y fortaleza, no me canso de darle gracias a Dios por haber encontrado un hombre tan maravilloso. LLenas mi vida de alegrías y orgullo. Empezamos una aventura juntos hace 6 años y seguimos divirtiéndonos y sorprendiéndonos ante tantas vivencias. Confío que Dios nos tiene deparadas muchas sorpresas más.

Chapter 1

Introduction

This chapter describes the structure of the thesis and sets the scene.

Combination of data or sometimes called “combination of observations” is an old term for the numerical analysis of data as stated by Cox. He considered that “combination of data” encompasses the whole of the statistical analysis of data, with an implied emphasis on condensation and summarization. In this thesis we are mainly interested in combining data when they are reported in different levels of aggregation. In particular, we are interested in merging two levels of aggregate data namely individual points and aggregate counts in areas.

The work is motivated by the interest in modeling the locations and times of wildfire occurrences that happened in the Continental United States in the period from 1986 to 1997. The data are provided by the US Forest Service, pulling together information organized by different institutions. The data cover fires that occurred in federal and non-federal lands. The federal data consist of each fire’s point location (latitude and longitude) $\tau_i = (x_i, y_i) \in \mathbb{R}^2$ where $i = 1, \ldots, I$ while the non-federal fires are aggregated by county $\Omega_j \in \mathbb{R}^2$, $N(\Omega_j) = n_j$ where $j = 1, \ldots, J$ for a given time year t. In Fig 1.1, the regions where the grouped data are colored in yellow while the red and blue dots are the fires’ point locations. This type of set up, when both type of data are present $I, J \geq 1$, is what Kulldorf [34] called a partially grouped sample.

One of the strategies for handling data from different sources is to divide the data into simpler subsections, analyze these separately and then in a second stage of the analysis, to merge the conclusions from the component analysis. In this work, we look for developing a joint analysis of the grouped and ungrouped data and provide a global analysis of the conditions of the wildfire occurrences in the Continental United States.

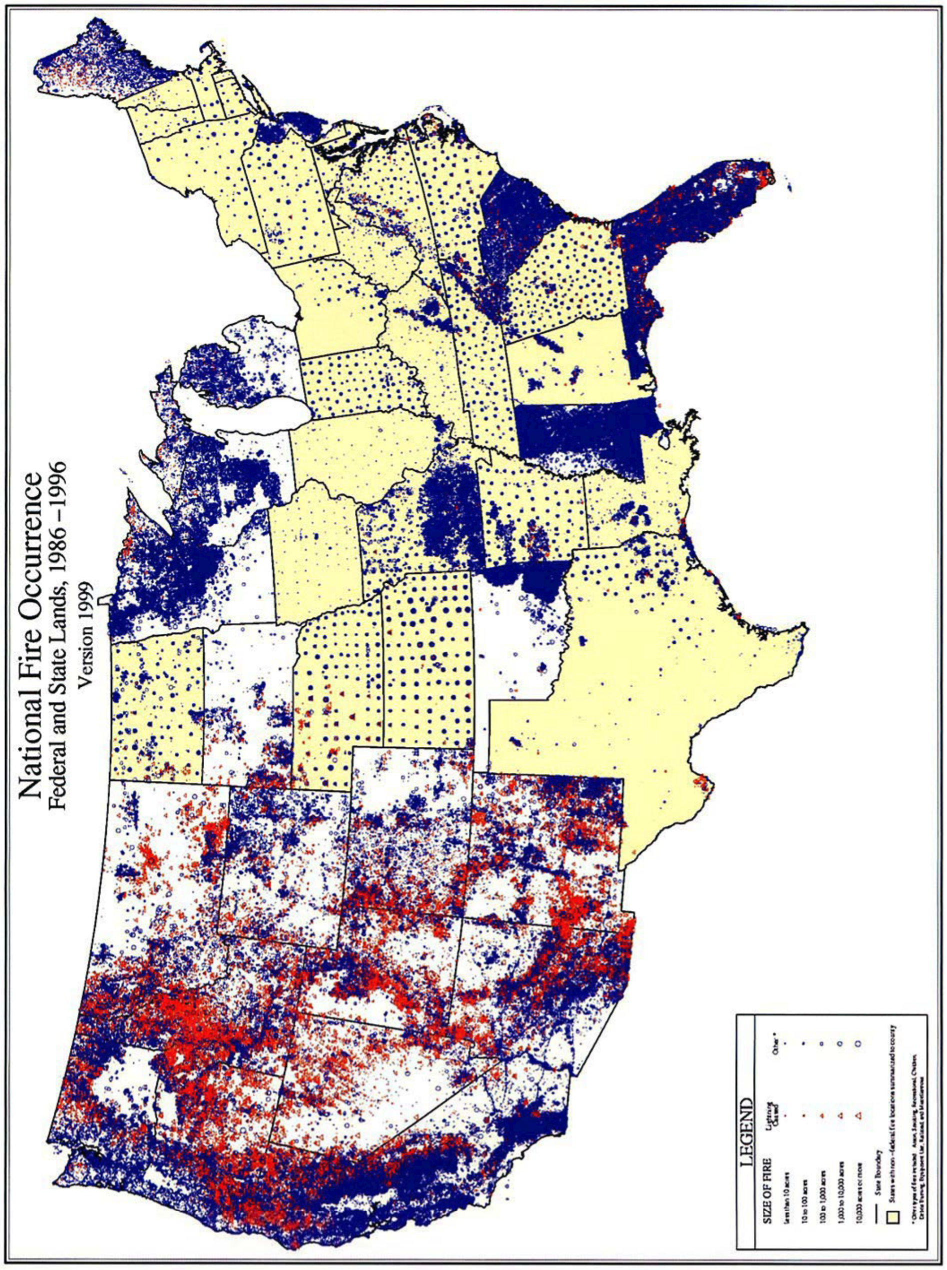
We develop statistical methods that merges grouped and ungrouped data and will lead us to do a risk assessment of wildfires when the data available are partially grouped following

Kulldorff's definition. Brillinger [4] proposes that a risk analysis includes (i) estimation of probabilities, (ii) determination of the distribution of damage and (iii) preparation of products such as formulas, graphics and hazard risk maps. In particular, as a result of this work we generate a risk map based on a spatial smoothing method that incorporates the two levels of aggregate data. We also propose a statistical model that will merge the counts and spatial point process data. Based on this model we estimate probabilities and risk measures of interest.

At first, we work on developing a robust smoothed estimate of the intensity rate of the wildfire locations point process in the Continental United States. This smoother combines the two levels of aggregate data by approximating them with a binary-valued process. Using the binary-valued process, the total number of events in non-overlapping regions $N(\Omega_j)$ are modeled as Binomial and the spatial point process $\tau_i = (x_i, y_i)$ by Bernoulli random variables. As the wildfire intensity rate is assumed to vary smoothly in space, a locally weighted likelihood analysis is employed. This has been shown to be a pertinent estimation technique [1, 2]. By including weights, the local variation and the grouping effect are considered in the model. The maps of these smoothers provide a useful tool for exploratory data analysis as they are one of the effective ways to convey the behaviour of the spread of wildfires [16].

For planning purposes, the Forest Service managers require estimates of the probability of a wildfire occurrence at a particular location and time. Using the binary-valued approximation to a partially grouped sample, a logit transform is employed to forecast locations of future wildfire occurrences in the Continental United States at time t. This model is appropriate for including explanatory variables that have been found to be of importance in the study of wildfire namely fuels, season, topography, weather conditions [6]. We fit our model with a locally weighted likelihood approach and include the fuels [6] and time explanatory variables. We estimate the model by the iteratively reweighted least squares (IRWLS) implemented in the R function glm(). We provide pertinent standard errors.

In the rest of this chapter we provide a description of the data used in this project. Also some notations, definitions and concepts are set down. The rest of the document is structured as follows: chapter 2 includes a literature review of the statistical methods for working with grouped univariate and bivariate data. We also include the extension of the statistical estimators to the partially grouped case. In chapter 3 different smoothing methods for grouped data are included and explored in the wildfire data. We propose a smoother that merges grouped and ungrouped data using a binary process approximation to the point process case. The smoother is based on a weighted likelihood approach where the weights provide the grouped effect and borrow information from the neighbouring counties. In Chapter 4 a weighted logit regression is used to model the wildfires under the partially grouped scheme. We used as an explanatory variable the fuel model referred in [6]. We proposed using two different smoothing parameters , one for the regions where the actual positions and other


National Fire Occurrence
Federal and State Lands, 1986 –1996
Version 1999
LEGEND
SIZE OF FIRE
Lightning Caused
Other *
less than 10 acres
10 to 100 acres
100 to 1,000 acres
1,000 to 10,000 acres
10,000 acres or more
State Boundary
States with non-federal fire locations summarized to county
* Other types of fire include: Arson, Smoking, Recreational, Children, Debris Burning, Equipment Use, Railroad, and Miscellaneous

USDA Forest Service General Technical Report RMRS-GTR-87. 2002. 39

Figure 1.1: The fire data locations are grouped by county in the yellow regions while the red and blue dots are individual fire points locations. The map is included in the "Development of Coarse-Scale Spatial Data for Wildland Fire and Fuel Management" [47]

for the regions where data is grouped. A model assessment and a robust estimation are also included. In Chapter 5 we propose a risk analysis integrated by three parts: a) a smoother that is used as an exploratory data analysis to detect wildfire patterns and higher risk regions; b) estimates of the probability wildfire occurrence in the proximity of location (x, y) and the uncertainty associated to these estimates. In Chapter 6 we include the contributions and future work.

1.1 The data

The National Fire Occurrence database is a database of natural and human-caused fire occurrences for the years 1986-1996. It includes federal data from the USDA Forest Service and four Department of Interior (DOI) agencies: Bureau of Land Management (BLM), Bureau of Indian Affairs (BIA), National Park Service (NPS), and U.S. Fish and Wildlife Service (FWS). It also includes non-federal data from all conterminous states except Nevada.

Federal Fire Occurrence Database - The USDA Forest Service and the Department of the Interior Agencies submitted fire occurrence data in latitude and longitude coordinates. The project “Development of Coarse-Scale Spatial Data for Wildland Fire and Fuel Management” generated GIS coverage from the latitude-longitude coordinates.

Non-federal Fire Database Non-federal fire records were received from all lower 48 except Nevada, which are composed primarily of federal land. The quality and completeness of the data received varied by state and the specifications can be looked at the reference [47].

1.1.1 The pilot study

To begin, we carry out a study to predict wildfire risk for the year 1990 at a pilot group of 5 contiguous states: Minnesota (MN), North Dakota (ND), South Dakota (SD), Wisconsin (WI) and Iowa (IA). We work with five states in order to develop computer routines and insights more easily. In the months to come we will turn to the case of the whole Continental Unites States. The general setting of our problem is provided as Figure 1.2. We define the sample space $\mathcal{S} \in \mathbb{R}^2$ as the region formed by the union of these 5 states plus their adjacent states Table 1.1. The sample space $\mathcal{S}$ is divided into non-overlapping sets as in $\mathcal{S} = \mathcal{B} \cup \mathbf{\Omega}$. $\mathcal{B}$ includes the federal regions while $\mathbf{\Omega} = \left(\bigcup_{j=1}^{J} \Omega_j \right)$ includes the non-federal regions. The Ω_j's stand for the counties (which are also non-overlapping).

One of the reasons for selecting this pilot group is that the counties are seen to be laid out quite regularly. Also this sample set contains a variety of federal and non-federal regions.

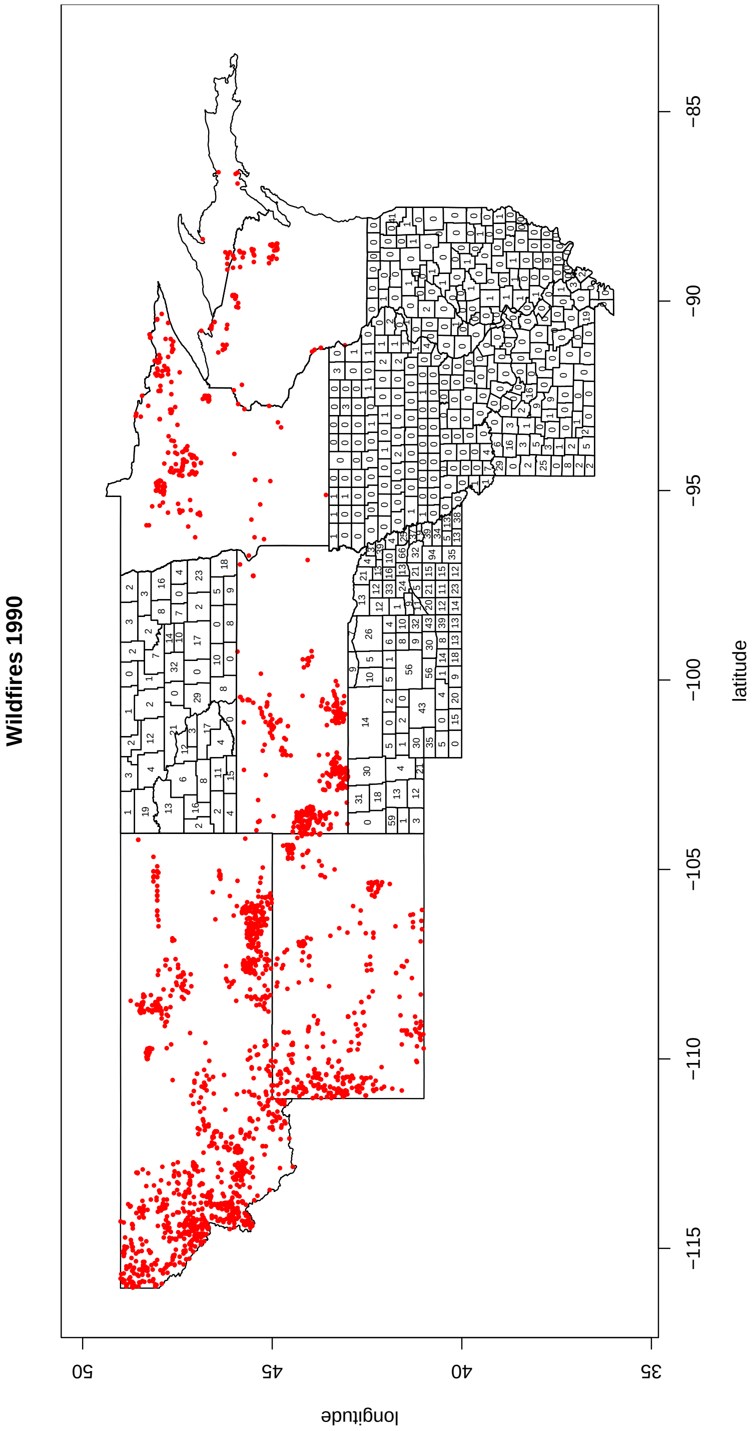


Figure 1.2: Wildfire occurrences recorded in 1990 for the sample space $\mathcal{S}$. The data grouped by county correspond to the states where 1 is reported in the third column of Table 1.1, $\mathbf{\Omega} = \left(\bigcup_{j=1}^{J} \Omega_j\right)$ with $J = 462$.

No.	state	observations	counties
1	Illinois	1	102
2	Iowa	1	99
3	Michigan	0	88
4	Minnesota	0	88
5	Missouri	1	115
6	Montana	0	56
7	Nebraska	1	93
8	North Dakota	1	53
9	South Dakota	0	66
10	Wisconsin	0	77
11	Wyoming	0	23

Table 1.1: The 11 states included in the sample space $\mathcal{S}$. The third column is 0 when data is ungrouped and 1 when grouped. The fourth column shows the number of counties per state.

1.2 Exploratory data analysis

In Figure 1.2 the federal regions that are shown to be more affected by wildfire occurrences in 1990 are: a) west regions of Montana and Wyoming b) north-west part of South Dakota and c) north-east part of Minnesota. We jitter the data to have a better display of the fires and for the estimations done in the following chapters. In non-federal regions the data is displayed numerically and it is difficult to identify those areas which are more prone to wildfire occurrence.

The choropleth map in Figure 1.3 plot the square root of the total number of fires for each county, in both the ungrouped and grouped cases appear. It can be seen that there are two counties in south-west South Dakota that seem to have a larger number of fires. Also the adjacent regions to Lake Superior are more affected by wildfire occurrences, as happens also in the west side of Montana. As this map does not include variables like the area of each county, it could mislead the interpretation of wildfire occurrences.

The density plots of Figures 1.4, 1.5 and the boxplots of Figure 1.6 are for the square root of the total number of fires per county and are divided in two classes of states: the ungrouped and grouped cases. In both cases, the distribution of the total number of fires seems to be skewed to the right. Combining these plots with the choropleth map, we note that the distributions for states with a larger number of fires are less skewed to the right and have less outliers.

The area covered by each fire is available in km^2 and the boxplots in Figure 1.7 provide

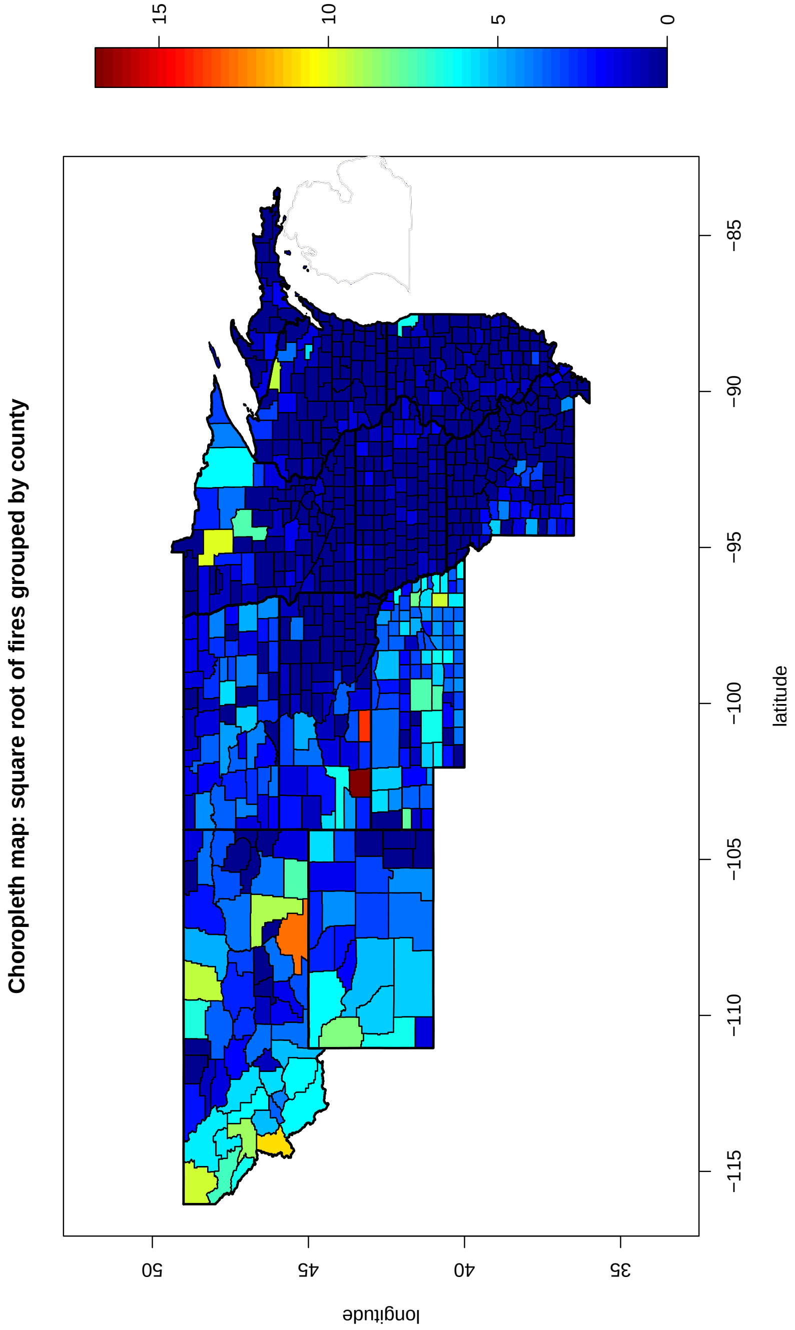


Figure 1.3: Choropleth map of the square root of the total number of federal and non-federal fires grouped by county in 1990, for the states included in the pilot group.

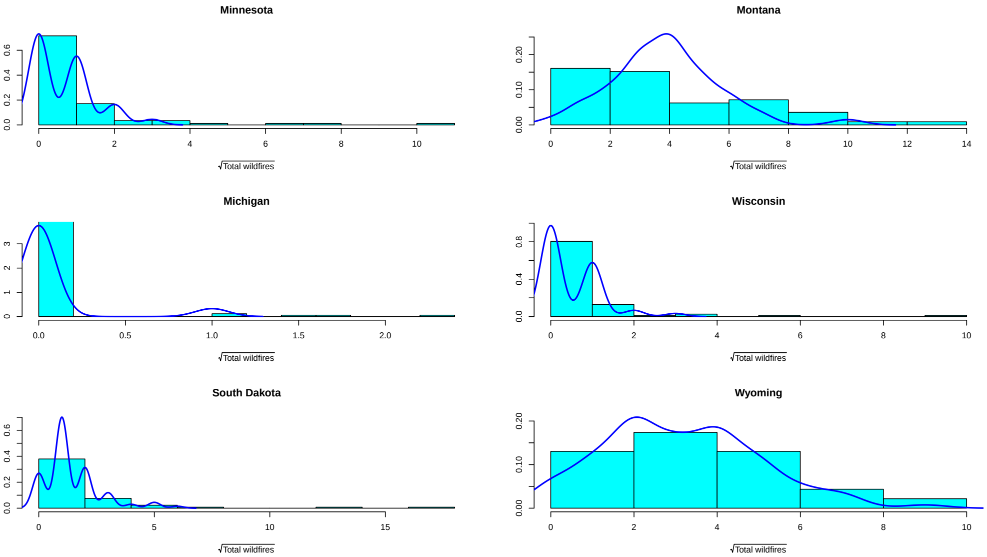


Figure 1.4: Kernel density plots for the square root of the total wildfire occurrences in federal lands for 1990. They are superimposed on histograms.

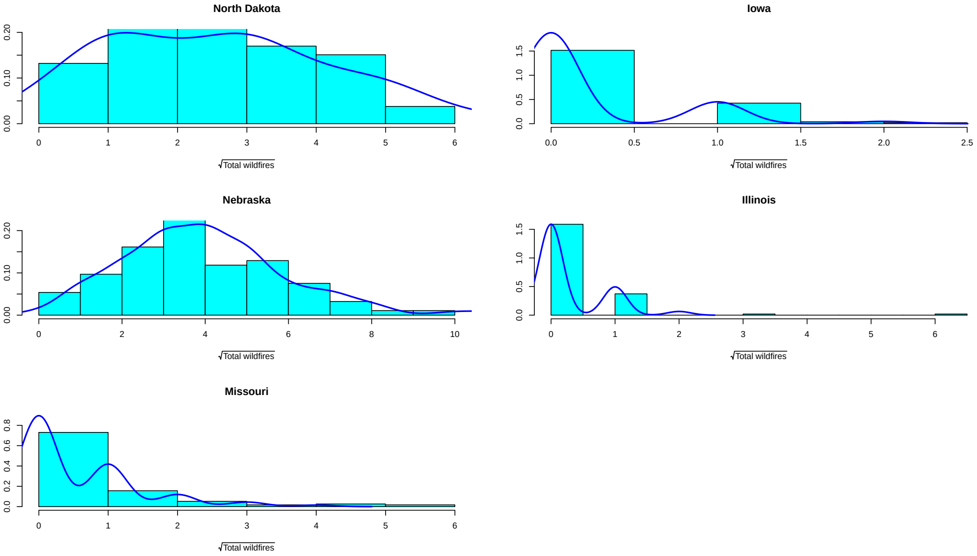


Figure 1.5: Kernel density plots for the square root of the total wildfire occurrences in non-federal lands 1990. Superimposed on a histogram are the kernel density estimates with a Gaussian kernel.

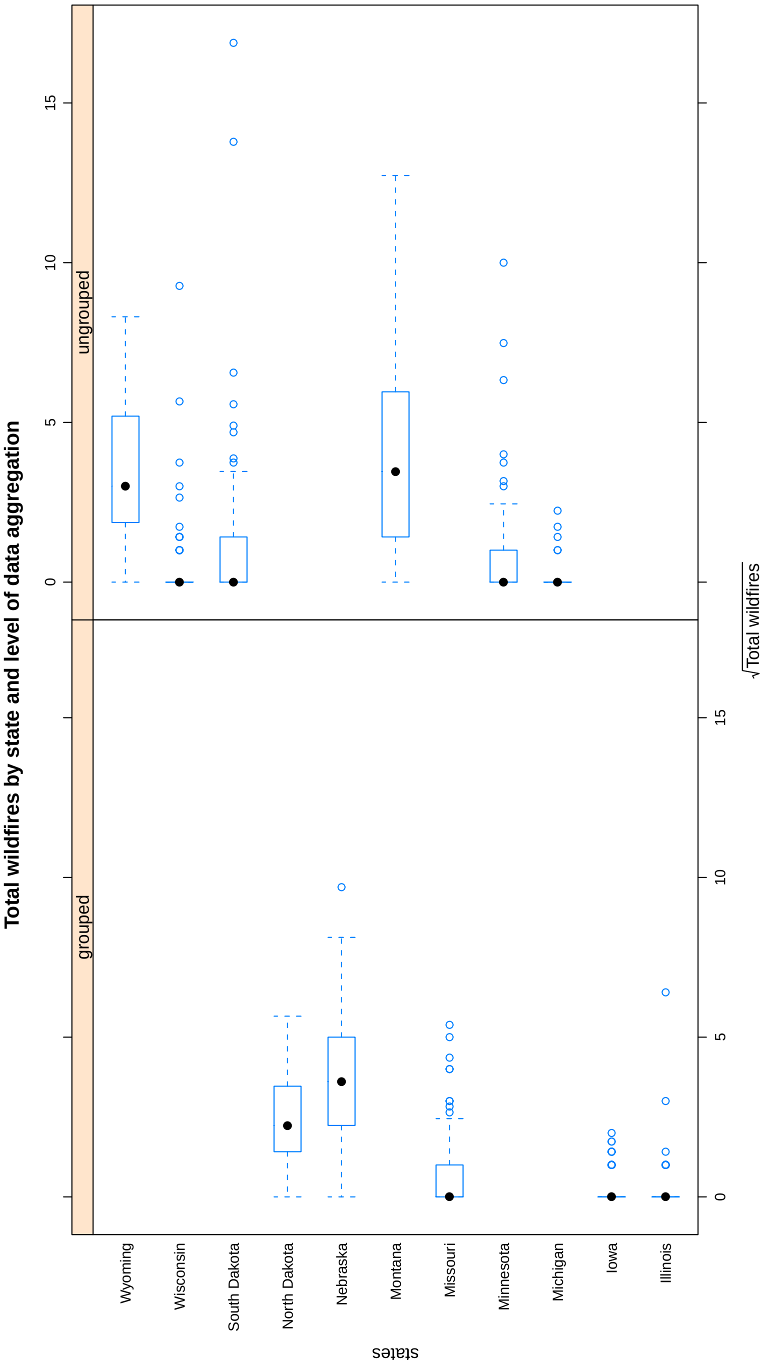


Figure 1.6: Boxplots of the the square root of the total wildfire occurrences both for federal and for non-federal regions in 1990.

their logarithmic transformation. It seems that the majority of fires have a size of less than 1 km^2. The fires reported in the state regions seems to be larger than in the federal regions. Illinois and Iowa do not report the area.

1.3 Background

We will employ a probability space $(\mathcal{S}, \mathcal{F}, \mathbb{P})$ where $\mathcal{S}$ is the sample space, $\mathcal{F}$ is a σ-algebra and $\mathbb{P}$ is a probability measure. The real line $(-\infty, \infty)$ is represented by $\mathbb{R}$ and the n-dimensional real Euclidean space by $\mathbb{R}^n$. The points of $\mathbb{R}^n$ will be represented by an ordered set of real-valued coordinates $\mathbf{y} = (y_1, \ldots, y_n)$.

Random Variable (r.v.) A real valued function $X(\cdot)$ defined on the space $(\mathcal{S}, \mathcal{F}, \mathbb{P})$ is called a random variable if the set $\{s : X(s) < x\} \in \mathcal{F}$ $\forall x \in \mathbb{R}$.

A **bivariate Poisson process** $A \subseteq \mathbb{R}^2$ with intensity $\lambda(x, y|\theta)$, $N(A)$, satisfies [49]:

1. For each Borel set $A \subseteq \mathbb{R}^2$, $N(A)$ is Poisson distributed with parameter

$$\int\int_A \lambda(u, v|\theta) du dv;$$

2. If $A_1, A_2, \ldots, A_k, \ldots$ are disjoint Borel subsets of $\mathbb{R}^2$

$$N\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} N(A_k)$$

In the context of the thesis, $N(A)$ is the count of how many fires points lie in $A \in \mathcal{F}$.

1.4 Maximum likelihood estimation

Referring to the notation employed by Davison [11], the likelihood function for θ, based on the data y, is defined to be

$$L(\theta) = f(y|\theta), \qquad \theta \in \Theta$$

where f is the density of probability mass function of the data, regarded as a function of θ for fixed y. When $\mathbf{y} = (y_1, \ldots, y_n)$ is a collection of independent observations the likelihood is

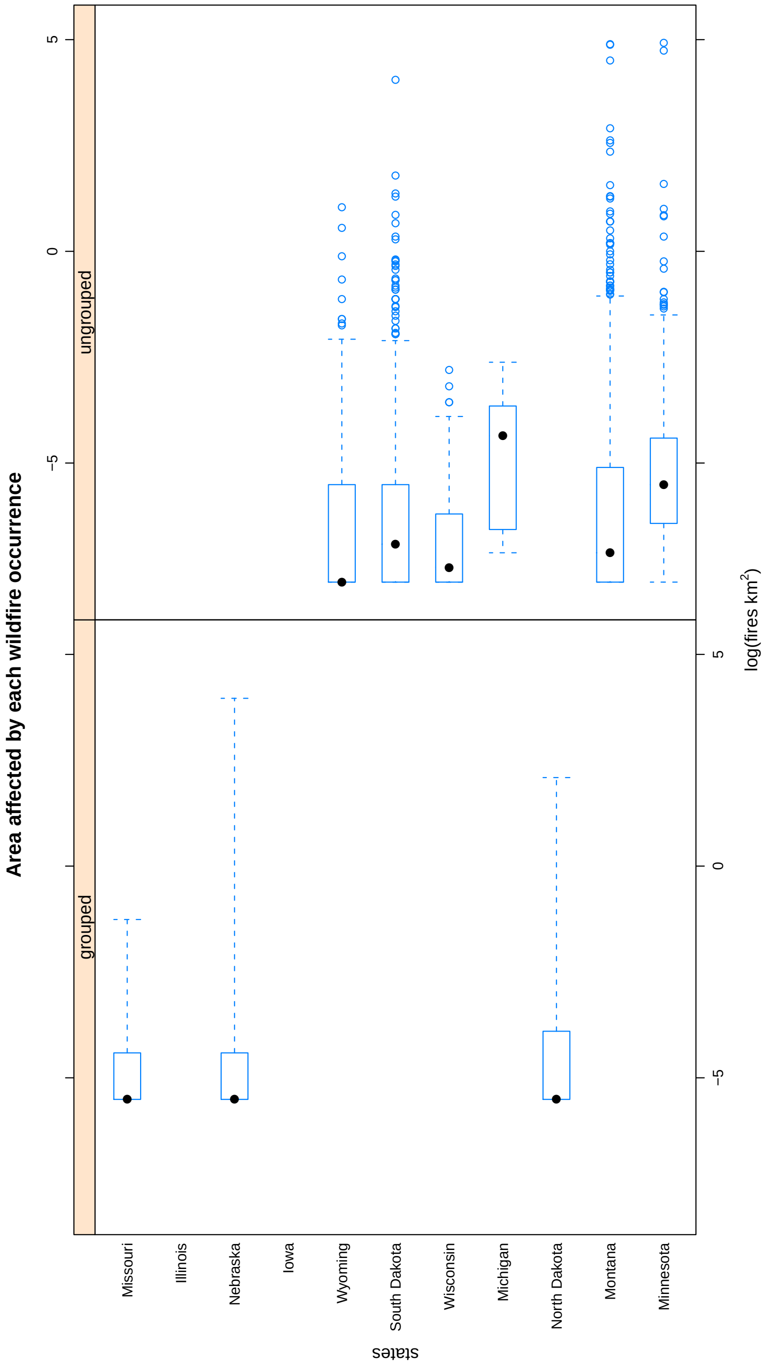


Figure 1.7: Boxplots of the logarithmic transformation of the fire's area in km^2 for both federal and non-federal regions in 1990.

$$L(\theta) = f(\mathbf{y}|\theta) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{f}(\mathbf{y_i}|\theta) \tag{1.1}$$

The log-likelihood is

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(y_i|\theta)$$

1.4.1 Large sample distribution

Next we provide some properties of maximum likelihood estimators under regularity conditions. Suppose that we have a random sample $Y_1, \ldots, Y_n$ from a density $f(y|\theta)$ that satisfies the regularity conditions [11]:

1. the true value θ_0 of θ is interior to the parameter space Θ, which has finite dimension, d, and is compact;
2. the densities defined by any two different values of θ are distinct;
3. there is a neighborhood $\mathcal{N}$ of θ_0 within which the first three derivatives of the log-likelihood with respect to θ exist almost surely, and for $r, s, t = 1, \ldots, p$ the quantity $n^{-1}E\left[|\partial^3\ell(\theta)/\partial\ell_r\partial\ell_s\partial\ell_t|\right]$ is uniformly bounded for $\theta \in \mathcal{N}$ and
4. within $\mathcal{N}$, the Fisher information matrix $I(\theta)$ defined by (1.2) is finite and positive definite.

$$\mathbf{I}\left(\theta\right)_{rs} = E\left[\frac{\partial\ell(\theta)}{\partial\theta_r}\frac{\partial\ell(\theta)}{\partial\theta_s}\right] = E\left[-\frac{\partial^2\ell(\theta)}{\partial\theta_r\theta_s}\right], \qquad r, s = 1, \ldots, p. \tag{1.2}$$

Consistency of the m.l.e. [8]: let $Y_1, \ldots, Y_n$ be i.i.d $f(y|\theta)$ and let Equation (1.1) provide the likelihood function. Suppose θ_0 is the true value of θ. Let $\hat{\theta}$ denote the m.l.e. of θ_0. Let $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions (1-4) on $f(y|\theta)$ and hence for $L(\theta)$, then for every $\epsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n\to\infty} Pr\left(|\tau(\hat{\theta}) - \tau(\theta_0)|\right) \geq \epsilon) = 0$$

Asymptotic efficiency of the m.l.e. [8]: let $Y_1, \ldots, Y_n$ be i.i.d $f(y|\theta)$, let $\hat{\theta}$ denote the m.l.e. of θ, and let $\tau(\theta)$ be a continuous function of θ_0. Under the regularity conditions on $f(y|\theta)$ and hence, $L(\theta)$,

$$\sqrt{n}\left[\tau(\hat{\theta}) - \tau(\theta_0)\right] \to N\left(0, v(\theta_0)\right)$$

in distribution, where

$$v(\theta_0) = \frac{[\tau'(\theta_0)]}{I(\theta_0)}$$

When $\boldsymbol{\theta} \in \mathbb{R}^d$

$$\tau'(\boldsymbol{\theta}_0)^t \mathbf{I}(\boldsymbol{\theta}_0)^{-1} \tau'(\boldsymbol{\theta}_0) \tag{1.3}$$

Chapter 2

Grouped and partially grouped data: models and methods

This chapter 2 includes a literature review of the statistical methods for working with grouped univariate and bivariate data. We also include the extension of the statistical estimators to the partially grouped case.

Grouped data is a topic that goes back to the end of the nineteenth century at least. Haitovsky [23] (2006) presented a substantial historical review of the different approaches that have been used with grouped data. He suggested that the literature on grouping data can be divided into four broad categories:

1. Concerns relationships between moments or (other parameters) estimated from the same data before and after grouping.
2. Making inferences from samples of grouped data ("grouped samples") about the ungrouped population.
3. Optimal grouping, for example when the researcher has control over the grouping of the original data.
4. Estimation methods requiring grouping, either to improve an estimator's properties or as an estimation device.

In this chapter we review some works done in the first two categories. They can be extended to the more general case of partially grouped data.

In the first part of this chapter we review the moment corrections. These attempt to relate directly the same moments calculated from the parent and the grouped populations respectively. Sheppard's corrections are included in the later. We also include the corrections needed for the m.l.e. and for regression coefficients estimated from grouped data. The m.l.e. of a partially grouped sample is estimated using the EM algorithm. We also review linear regression with partially grouped data. We are particularly interested in spatial data, but the univariate case provides suggestions as to the cases we work with. In the second part of the chapter we include the extensions of the Sheppard's and the m.l.e. corrections to the bivariate case. We also develop the m.l.e. for partially grouped data using the EM algorithm for that case.

Haitovsky defined data grouping as the process by which a real-valued random variable, X, with a given distribution function $F(x|\theta)$ (continuous or discrete) is condensed into a discrete distribution function. For example by defining

$$p_j = \int_{c_{j-1}}^{c_j} dF(x|\theta), \qquad j = 1, \ldots, J,$$

where $X \in [c_0, c_J]$ is partitioned by $c_0 < c < \ldots < c_J$ into J disjoint and exhaustive groups. The c_j's are termed interval limits or boundaries and $\Omega_j = (c_{j-1}, c_j]$ the j^{th} interval or group. Grouping transforms a given distribution function, continuous or discrete, into a multinomial one. Given a sample of data, the number of cases falling into the j^{th} group will be denoted by n_j, the j^{th} group frequency. We write $\sum_{j=1}^{J} n_j = n$. We might think of the observations as occurring at the interval midpoints, the interval means or some other representative point of an interval. The population interval mean is defined as the conditional expectation of X given $X \in \Omega_j$ for example, $\bar{x}_j = \dfrac{\int_{c_{j-1}}^{c_j} dF(x|\theta)}{\sum_{j=1}^{J} p_j}$. The interval widths $c_j - c_{j-1}$ may differ.

2.1 Moment corrections for grouped data

In the early statistical work on grouped data, analyses were concentrated in the first category. The interest was to derive the relationships between sample moments calculated from data before and after grouping. In particular, Sheppard [48] in 1898 proposed corrections for estimating product moments when data were grouped into equispaced intervals. M.G. Kendall [33] reviewed the properties of the frequency function $f(x)$ under which Sheppard's corrections apply. Based on Kendall's paper we include a sketch of how Sheppard's corrections are developed.

Define $f(x)$ in the range $x = c_0$ to $x = c_J$, where respectively c_0 and c_J may be $\pm\infty$. Without loss of generality $f(x)$ maybe consider as equal to zero outside this range, and still be written:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

It is assumed that $f(x)$ is continuous in the range c_0 to c_J and that if the range is infinite $f(x)$ converges uniformly to zero as $|x| \to \infty$.

Suppose the range is in fact divided into intervals of width h, and that we are given only the counts of cases falling within those intervals. The probability of an observation in the j^{th} interval, centred at x_j, will then be given by

$$p_j = \int_{-\frac{h}{2}}^{\frac{h}{2}} f(x_j + \epsilon)d\epsilon, \qquad j = 1, \ldots, J,$$

The raw moment of order s about the origin can be defined as

$$\begin{aligned} \bar{\mu}'_s &= \sum_{s=-\infty}^{\infty} x_j^s p_j \\ &= \sum_{s=-\infty}^{\infty} x_j^s \int_{-\frac{h}{2}}^{\frac{h}{2}} (x_j + \epsilon)d\epsilon \end{aligned}$$

Here the bar over the μ denotes that the moment is raw, and the dash that it is taken about 0. The s^{th} moment of the distribution itself, if it exists, is given by:

$$\mu'_s = \int_{-\infty}^{\infty} x^s f(x)dx$$

We may put:

$$\frac{1}{h}\int_{\infty}^{\infty} F(x)dx = \sum_{j=1}^{\infty} F(x_j) - R$$

where R is a remainder term, which is thus true in general provided that the integral and the sum exist. Assuming that the conditions needed for R to be neglected and

$$F(x) = x^s \int_{-\frac{h}{2}}^{\frac{h}{2}} f(x_j + \epsilon) d\epsilon$$

Then,

$$\begin{aligned} \bar{\mu}'_s &= \frac{1}{h} \int_{-\infty}^{\infty} x^s dx \int_{-\frac{h}{2}}^{\frac{h}{2}} f(x+\epsilon) d\epsilon \\ &= \frac{1}{h} \int_{-\infty}^{\infty} \int_{-\frac{h}{2}}^{\frac{h}{2}} x^s f(x+\epsilon) d\epsilon dx \end{aligned}$$

provided that this multiple integral exists. If in addition it is absolutely convergent, we may substitute x for $x + \epsilon$ and integrate with respect to ϵ. We shall then have:

$$\begin{aligned} \bar{\mu}'_s &= \frac{1}{h} \int_{-\infty}^{\infty} \int_{-\frac{h}{2}}^{\frac{h}{2}} (x-\epsilon)^s f(x) d\epsilon dx \\ &= \frac{1}{h} \int_{-\infty}^{\infty} f(x) dx \frac{\left(x + \frac{h}{2}\right)^{s+1} - \left(x - \frac{h}{2}\right)^{s+1}}{s+1} \\ &= \frac{1}{h} \int_{-\infty}^{\infty} f(x) dx \sum_{k=0}^{\left[\frac{s}{2}\right]} 2 \left(\frac{h}{2}\right)^{2k+1} \binom{s+1}{2k} x^{s-2k} \\ &= \sum_{k=0}^{\left[\frac{s}{2}\right]} \left(\frac{h}{2}\right)^{2k} \binom{s}{2k} \frac{1}{2k+1} \mu'_{s-2k} \end{aligned} \tag{2.1}$$

where $\left[\frac{s}{2}\right]$ is the integral part of $\frac{s}{2}$. The raw moments in terms of the actual moments is given in Equation (2.1).

There are two assumptions made in reaching equation: first the absolute convergence, and second that R may be neglected. Sheppard's correction are valid when the range of the distribution function is finite, that there is high-order contact at the terminals of the range and that the remainder term in the Euler-Maclaurin expansion used in deriving the corrections is negligible. Further details on the conditions for Sheppard's correction can be found in reference [33].

Not much later Pearson [42] (1902) introduced the analytical solutions to the general problem of moment computations for grouped distributions under varying assumptions about distributional form, including distribution functions that have infinite ordinates and any slopes at either end of the terminals for example J- and U-shaped distributions.

Heijtan and Rubin [26] refers to coarse data as a kind of incomplete data. Coarse data is when one observes only a subset of the complete-data sample space in which the true, unobservable data lie. They considered grouping as a special case of coarsening data, where the coarsening is known and non-stochastic (e.g. the intervals are set in advance).

2.2 Maximum likelihood estimate for partially grouped data

Kulldorff (1961) generalized the concept of grouped data to partially grouped data. The available sample is partially grouped if its information is associated with a set of disjoint and exhaustive groups partitioning the variable $X \in [c_0, c_J]$, such that in each group either none or all of the individual observations in the region are recorded. The individual values of the observations are denoted by y_i, $i = 1, \ldots, I$.

Consider the pdf or pmf $f(\cdot|\theta)$ depending on the parameter θ. Let the sample space be partitioned into J disjoint and exhaustive groups of widths $c_j - c_{j-1}$. A random sample of size n is now drawn and the numbers falling in each group obtained. Let n_j be the number in the interval $(c_{j-1}, c_j]$ and let $n = \sum_{j=1}^{J} n_j$. Suppose there are also statistically independent observations y_i, $i = 1, \ldots, I$ sampled from $f(\cdot|\theta)$.

We are interested in obtaining the maximum likelihood estimator (m.l.e.) of θ using both the grouped and the non-grouped observations (partial grouping). Write

$$p_j(\theta) = \int_{c_{j-1}}^{c_j} f(y|\theta)dy \qquad j = 1, \ldots, J$$

for the probability of obtaining a value in the j^{th} subinterval $(c_{j-1}, c_j]$. Now the overall likelihood function has the following form:

$$L(\theta) = \prod_{j=1}^{J} p_j(\theta)^{n_j} \prod_{i=1}^{I} f(y_i|\theta) \tag{2.2}$$

The likelihood function depends on the observed counts and points $(n_1, \ldots, n_J, y_1, \ldots, y_I)$. The m.l.e. maximizes (2.2) or, equivalently, the log-likelihood function (2.3) below. Conditions to assure its existence were given in subsection 1.4.1.

$$\ell\,(\theta) \;=\; \sum_{j=1}^{J} n_j \log p_j(\theta) + \sum_{i=1}^{I} \log f(y_i|\theta) \tag{2.3}$$

Let θ_0 be the m.l.e. of θ calculated on the basis of non-grouping using a particular vector $\mathbf{v}_j = \{v_{j1}, \ldots, v_{jn_j}\}$ for $j = 1, \ldots, J$, as the observed values of the variable $N(\Omega_j)$ and the observations $y_1, \ldots, y_I$. The estimator θ_0 ignores grouping and it has been corrected in different ways. In particular, Lindley [35] and Tallis [52] found an "exact" formula for correcting the m.l.e. under grouped data, by using Taylor's Theorem and the midpoints of the intervals, $(c_{j-1} + c_j)/2$. Tallis extended the results for unequal grouping and to the multivariate case. These corrections coincide, under certain conditions, to Sheppard's corrections.

We will introduce an approximation later. It might be possible that no exact formula is found but we might get a large sample distribution.

2.3 Expectation-Maximization (EM) algorithm

Observations that are grouped fall into the category of incomplete data as defined by Dempster, Laird and Rubin [13]. They proposed an approach to iterative computation of the m.l.e.'s when the observations can be viewed as incomplete data. Since each iteration of the algorithm consists of an expectation step (E-step) followed by a maximization step (M-step), it has been called the *EM* algorithm. Dempster and Rubin [14] showed how Sheppard's corrections arise through the application of the EM agorithm.

McLachlan [39] develops the EM algorithm when grouped data are presented. We are interested in developing the case where we have grouped data for some regions and also individual observations for other regions. Recalling our general scheme, let a region of interest be partitioned into J sets Ω_j $j = 1, \ldots, J$. A random sample of size $n = \sum_{j=1}^{J} n_j$ is now drawn from a population with frequency function $f(\cdot|\theta)$ and the numbers falling in each Ω_j obtained. Let n_j be the number of observations falling into Ω_j. Let's suppose we also have I independent observations sampled from the same frequency function $f(\cdot|\theta)$. These observations are not grouped and are located outside the union of Ω_j. The ungrouped observations Y_i are statistically independent of the grouped data, $N(\Omega_j)$.

For given $n = \sum_{j=1}^{J} n_j$ the observed data $(n_1, \ldots, n_J)^T$ has a multinomial distribution, consisting of n draws on J categories with probabilities $\dfrac{p_j(\theta)}{p(\theta)}$ for $j = 1, \ldots, J$ where:

$$p_j(\theta) = \int_{c_{j-1}}^{c_j} f(y|\theta)dy \qquad p(\theta) = \sum_{j=1}^{J} p_j(\theta)$$

and θ is the vector of unknown parameters. Thus the log-likelihood is given by:

$$\ell\left(\theta\right) \;\; = \;\; \sum_{j=1}^{J} n_j \log p_j(\theta) + \sum_{i=1}^{I} \log f(y_i|\theta)$$

This problem can be addressed within the EM framework by introducing the vectors

$$\mathbf{v}_j = (v_{j1}, \ldots, v_{jn_j})^T \qquad j = 1, \ldots, J$$

as missing values. The vector $\mathbf{v}_j$ contains the n_j unobserved individual observations in the jth region Ω_j, $j = 1, \ldots, J$.

It follows that the complete data log-likelihood function for θ is given by:

$$\ell_c\left(\theta\right) \;\; = \;\; \sum_{j=1}^{J} \sum_{l=1}^{n_j} \log f(v_{jl}|\theta) + \sum_{i=1}^{I} \log f(y_i|\theta) \tag{2.4}$$

The E-step, on the $(k+1)$th iteration, is effected by first taking the expectation of Equation (2.4) conditional on $(n_1, \ldots, n_J; y_1, \ldots, y_I)$ leading to:

$$Q(\theta, \theta^{(k)}) = \sum_{j=1}^{J} n_j^{(k)} Q_j(\theta, \theta^{(k)}) + \sum_{i=1}^{I} \log f(y_i|\theta^{(k)})$$

where $Q_j(\theta, \theta^{(k)}) = E_{\theta^{(k)}}\left[\log f(v|\theta)|v \in \Omega_j\right]$ and $n_j^{(k)} = n_j$.

Next the M-step. On the $(k+1)th$ iteration, differentiate $Q(\theta, \theta^{(k)})$ with respect to θ. Suppose that $\theta^{(k+1)}$ is a root of the equation

$$\frac{\partial}{\partial\theta}Q(\theta,\theta^{(k)}) = \sum_{j=1}^{J} n_j^{(k)} \frac{\partial}{\partial\theta} Q_j(\theta,\theta^{(k)}) + \sum_{i=1}^{I} \frac{\partial}{\partial\theta} \log f(y_i|\theta)$$

where

$$\frac{\partial}{\partial\theta}Q_j(\theta,\theta^{(k)}) = E_{\theta^{(k)}}\left[\log f(v|\theta)|v \in \Omega_j\right]$$

These E and M steps are developed for partially grouped observations under the assumption that the pdf is Normal with parameters $\boldsymbol{\theta} = (\mu, \sigma^2)$. The E-step is then

$$\ell_c(\theta) \quad = \quad -\frac{(N+1)}{2}\left[\log\sigma^2 + \log 2\pi\right] - \frac{1}{2\sigma^2}\left[\sum_{j=1}^{J} n_j E_{\theta^{(k)}}\left[(v-\mu)^2|v \in \Omega_j\right] + \sum_{i=1}^{I}\frac{(y_i-\mu)^2}{2\sigma^2}\right]$$

The value $\boldsymbol{\theta}^{(k+1)}$ of $\boldsymbol{\theta}$ that maximizes Equation (2.4) are

$$\mu^{(k+1)} = \frac{\sum_{j=1}^{J} n_j E_{\theta^{(k)}}\left[v|v \in \Omega_j\right] + \sum_{i=1}^{I} Y_i}{\sum_{j=1}^{J} n_j + I}$$

$$\sigma^{(k+1)^2} = \frac{\sum_{j=1}^{J} n_j E_{\theta^{(k)}}\left[\left(v-\mu^{(k+1)}\right)^2|v \in \Omega_j\right] + \sum_{i=1}^{I}\left(Y_i - \mu^{(k+1)}\right)}{\sum_{j=1}^{J} n_j + I}$$

2.4 Regression Models

Suppose there are observations $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_J, Y_J)$ where Y_j are independent and the $\mathbf{x}_j$'s fixed. The distribution of the response Y_j for the j^{th} case is assumed to depend on $\mathbf{x}_j$. In general, $\mathbf{x}_j$ is nonrandom and called a covariate or explanatory variable. The variate Y_j is assumed random and referred to as the response variable. Supposing $\mathbb{E}(Y) = \mu(\mathbf{x})$ one models the data via:

$$Y_j = \mu(\mathbf{x}_j) + \epsilon_j \qquad j = 1, \ldots, J \tag{2.5}$$

The function $\mu(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ may be unknown. In (2.5) the ϵ_j may be assumed i.i.d. variates with distribution function $F(\epsilon|\theta)$. One might assume $\mu(\mathbf{x}_j) = g(\boldsymbol{\beta}, \mathbf{x}_j)$ where g is known except for a vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$. A historically common choice of $g(\boldsymbol{\beta}, \mathbf{x}_j)$ is $g(\boldsymbol{\beta}, \mathbf{x}_j) = \sum_{l=0}^{p} \beta_l x_{jl}$. The model (2.5) then becomes:

$$Y_j = \mathbf{x}_j^T \boldsymbol{\beta} + \epsilon_j \qquad j = 1, \ldots, J$$

This is the model of multiple regression. It is common to assume the ϵ_j to be $IN(0, \sigma^2)$. The model may then be written as:

$$Y_j \sim N(\mathbf{x}_j^T \boldsymbol{\beta}, \sigma^2) \qquad j = \ldots, J$$

The m.l.e. for this model is obtained by maximizing the following likelihood function:

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= f(y_1, \ldots, y_n|\boldsymbol{\beta}, \sigma^2) \\ &= \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2}{2\sigma^2} \right\} \end{aligned}$$

as a function of $\boldsymbol{\beta}$ and σ.

Grouped data in the regression scheme, may involve grouping the dependent variable, the explanatory variables or both. Consider the case where both sides of the equation, are grouped in corresponding ways. Further consider that the data consist of both grouped and ungrouped observations. Let i index ungrouped and j grouped data with $i = 1, \ldots, I$ and $j = 1, \ldots, J$, y_j and $\mathbf{x}_j$ are aggregate. The log-likelihood function is the following:

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{I} \log f(y_i|\theta) + \sum_{j=1}^{J} n_j \log \int_{c_{j-1}}^{c_j} f(y|\theta) dy \qquad (2.6) \\ &= -\frac{I}{2} \log \sigma^2 - \frac{\sum_{i=1}^{I} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2} + \sum_{j=1}^{J} n_j \log \frac{1}{\sqrt{2\pi\sigma^2}} \int_{c_{j-1}}^{c_j} \exp\left\{ -\frac{\left(y - \mathbf{x}_j^T \boldsymbol{\beta}\right)^2}{2\sigma^2} \right\} dy \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$.

Burridge [7] refers to this expression. He develops a reparametrization that has a log-likelihood function which is concave with respect to the transformed parameters:

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \frac{1}{\sigma^I}\prod_{i=1}^{I} f\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\middle|\,\boldsymbol{\theta}\right)\prod_{j=1}^{J}\int_{u_j}^{v_j} f(\epsilon)d\epsilon \\
\ell\left(\boldsymbol{\theta}\right) &= I\log\phi + \sum_{i=1}^{I}\log f\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\middle|\,\boldsymbol{\theta}\right) + \sum_{j=1}^{J}\log\int_{u_j}^{v_j} f(\epsilon)d\epsilon \qquad (2.7)
\end{aligned}
$$

with $u_j = (c_{j-1} - \mathbf{x}_j^T\boldsymbol{\beta})/\sigma$ and $v_j = (c_j - \mathbf{x}_j^T\boldsymbol{\beta})/\sigma$. Using parametrization:

$$\phi = 1/\sigma \qquad \alpha = \mathbf{x}_j^T\boldsymbol{\beta}/\sigma$$

It applies to regression models with densities strictly log-concave i.e. standard normal, logistic and extreme value. This same parametrization applies to grouped data. This condition implies the existence and uniqueness of maximum likelihood estimates and also makes the Newton-Raphson procedure faster and quite insensitive to the choice of starting value.

2.5 Bivariate grouped and partially grouped data

Grouped data in $\mathbb{R}^d$, $d = 2$ follows the definition on the line. It is the process by which a bivariate variable $\mathbf{X}$ with a given distribution function $F(\mathbf{x}|\theta)$ (continuous or discrete) is condensed into a discrete distribution function for example

$$P_j = \int\int_{\Omega_j} dF(\mathbf{x}|\theta)d\mathbf{x}, \qquad j = 1, \ldots, J. \qquad (2.8)$$

where the sample space, $\mathcal{S}$, is partitioned into J disjoint and exhaustive groups Ω_j $j = 1, \ldots, J$. Grouping transforms a given distribution function, continuous or discrete, into a multinomial one. We can think that the observations occurred at the centroid or some other representative point of the region. The centroid of Ω_j is defined as the conditional expectation of $\mathbf{X}$ given $\mathbf{X} \in \Omega_j$ for example, $\bar{x}_j = \int\int_{\Omega_j} dF(\mathbf{x}|\theta)/\sum_{j=1}^{J} p_j$. The regions might be non-equal Ω_j.

2.5.1 Maximum likelihood estimation: the 2-dimensional case

Consider the frequency function $f(\mathbf{x}|\theta)$ defined in $\mathbb{R}^2$ depending on the parameter θ. Consider that we have J non-overlapping regions Ω_j. A random sample of size n is drawn from a population with frequency function $f(\cdot|\theta)$ and the numbers falling in each region counted. Suppose we also have $\mathbf{Y}_i$ $i = 1, \ldots, I$ independent observations sampled from the same frequency function $f(\cdot|\theta)$. These observations are not grouped and are located outside the J

regions. We are interested in obtaining the maximum likelihood estimate of $\hat{\theta}$ from the data set of both the grouped and the ungrouped observations.

Define P_j, the probability of obtaining a value in the j^{th} region, as in Equation (2.8). The likelihood function follows:

$$L(\theta) = \prod_{j=1}^{J} P_j(\theta)^{n_j} \prod_{i=1}^{I} f(\mathbf{y}_i, \theta) \tag{2.9}$$

and it will be maximized with respect to θ. The log-likelihood and the partial derivative are as follow:

$$\begin{aligned}
\ell(\theta) &= \sum_{j=1}^{J} n_j \log \int\int_{\Omega_j} f(\mathbf{x}|\theta) d\mathbf{x} + \sum_{i=1}^{I} \log f(\mathbf{y}_i|\theta) \\
\frac{\partial}{\partial\theta}\ell(\theta) &= \sum_{j=1}^{J} n_j \frac{\partial}{\partial\theta}\left[\log \int\int_{\Omega_j} f(\mathbf{x}, \theta) d\mathbf{x}\right] + \sum_{i=1}^{I} \frac{\partial}{\partial\theta} \log f(\mathbf{y}_i|\theta) \\
&= \sum_{j=1}^{J} \frac{n_j}{\int\int_{\Omega_j} f(\mathbf{x}|\theta) d\mathbf{x}} \left[\frac{\partial}{\partial\theta} \int\int_{\Omega_j} f(\mathbf{x}|\theta) d\mathbf{x}\right] + \sum_{i=1}^{I} \frac{1}{f(\mathbf{y}_i|\theta)} \frac{\partial}{\partial\theta} f(\mathbf{y}_i|\theta)
\end{aligned}$$

2.5.2 The bivariate normal distribution

Consider a bivariate normal $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and the variance-covariance matrix is as follow:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Suppose that $\sigma_1\sigma_2 \neq 0$ and $|\rho| < 1$. Then

$$\begin{aligned}
f_X(y_1, y_2) &= \frac{1}{2\pi\sqrt{\det \boldsymbol{\Sigma}}} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})\right] \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{(\frac{y_1-\mu_1}{\sigma_1})^2 - 2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2}\right.\right. \\
&\qquad \left.\left. + (\frac{y_2-\mu_2}{\sigma_2})^2\right\}\right]
\end{aligned}$$

provided $\rho \neq \pm 1$. Recalling the problem of interest and writing $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we can find an expression for the log-likelihood function:

$$\begin{aligned}\ell\left(\theta\right) &= \sum_{i=1}^{m} \frac{N_i}{2\pi\sqrt{\det \boldsymbol{\Sigma}}} \log \int\int_{A_i} -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) d\mathbf{x} \\ &\quad + \sum_{j=1}^{p} \left\{ -\frac{1}{2}\mathbf{y}_j^T \boldsymbol{\Sigma}^{-1}\mathbf{y}_j + \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)^T \mathbf{y}_j \right\} - \frac{p}{2}\left(\log|\det \boldsymbol{\Sigma}| + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) - p\log\pi\end{aligned}$$

and the m.l.e. follows:

$$\hat{\mu} = \frac{1}{I}\sum_{i=1}^{I} x_i \tag{2.10}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{I}\sum_{i=1}^{I}(x_i - \hat{\mu})(x_i - \hat{\mu})^T \tag{2.11}$$

Chapter 3

Smoothing and likelihood analysis

3.1 Introduction

One purpose of this thesis is to estimate a smooth intensity function of the distribution of wildfires in the continental USA, by integrating two levels of aggregate data. The data is reported as a spatial point process in federal lands and as counts in non-federal lands. In this chapter, a robust smoother that integrates these two type of data and borrows strength from the neighbor regions is proposed. The smoother is based on a locally weighted likelihood approach.

To carry out a pilot study we selected five states: Minnesota (MN), North Dakota (ND), South Dakota (SD), Wisconsin (WI) and Iowa (IA). The general setting of our problem is provided as Fig. 3.1. There is a sample space $\mathcal{S} \in \mathbb{R}^2$, which is divided into non-overlapping sets as in $\mathcal{S} = \mathcal{B} \cup \mathbf{\Omega}$, $\mathbf{\Omega} = \left(\bigcup_{j=1}^{J} \Omega_j\right)$ where the Ω_j's are themselves non-overlapping. In the region $\mathcal{B}$ the individual locations are denoted by $\tau_i = (x_i, y_i)$, $i = 1, \ldots, I$ with (x, y) corresponding to location. For the regions Ω_j the data are n_j, $j = 1, \ldots, J$ with $n_j = N(\Omega_j)$.

Fig. 3.1 displays the fires that happened in 1990 in the states selected for the pilot study. The locations of the fires are represented as dots, while the total counts per county are reported numerically. In Minnesota and Wisconsin, the fires are located principally in the northern regions, which are covered by spruce fir, oak and aspen firch as shown in the forest cover types map Fig. 3.2. The wildfires in South Dakota took place in the south-west boundary area, which is covered by pines. The stream and water bodies map at bottom of Fig. 3.2, indicates that the wildfires in Minnesota are located close to the lakes region. While the higher frequency of events in North Dakota are located along the Missouri River and through the south-east boundary.

A second way to explore the wildfires occurred in 1990 is using a choropleth map. It is a

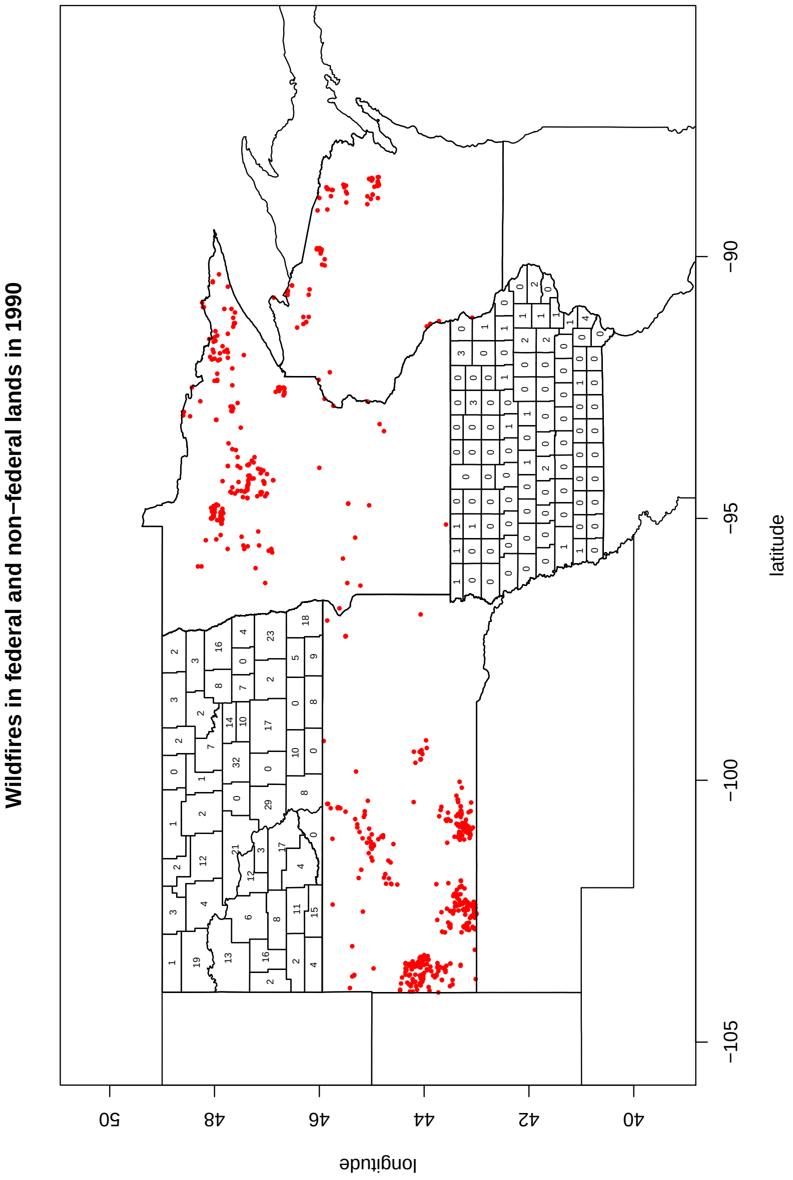


Figure 3.1: Wildfires recorded in 1990 for the sample space $\mathcal{S}$, the region formed by the union of the states: MN, ND, WI, SD and IA. The data grouped by county corresponds to ND and IA, there are $J = 152$ counties $\mathbf{\Omega} = \left(\bigcup_{j=1}^{J} \Omega_j\right)$.

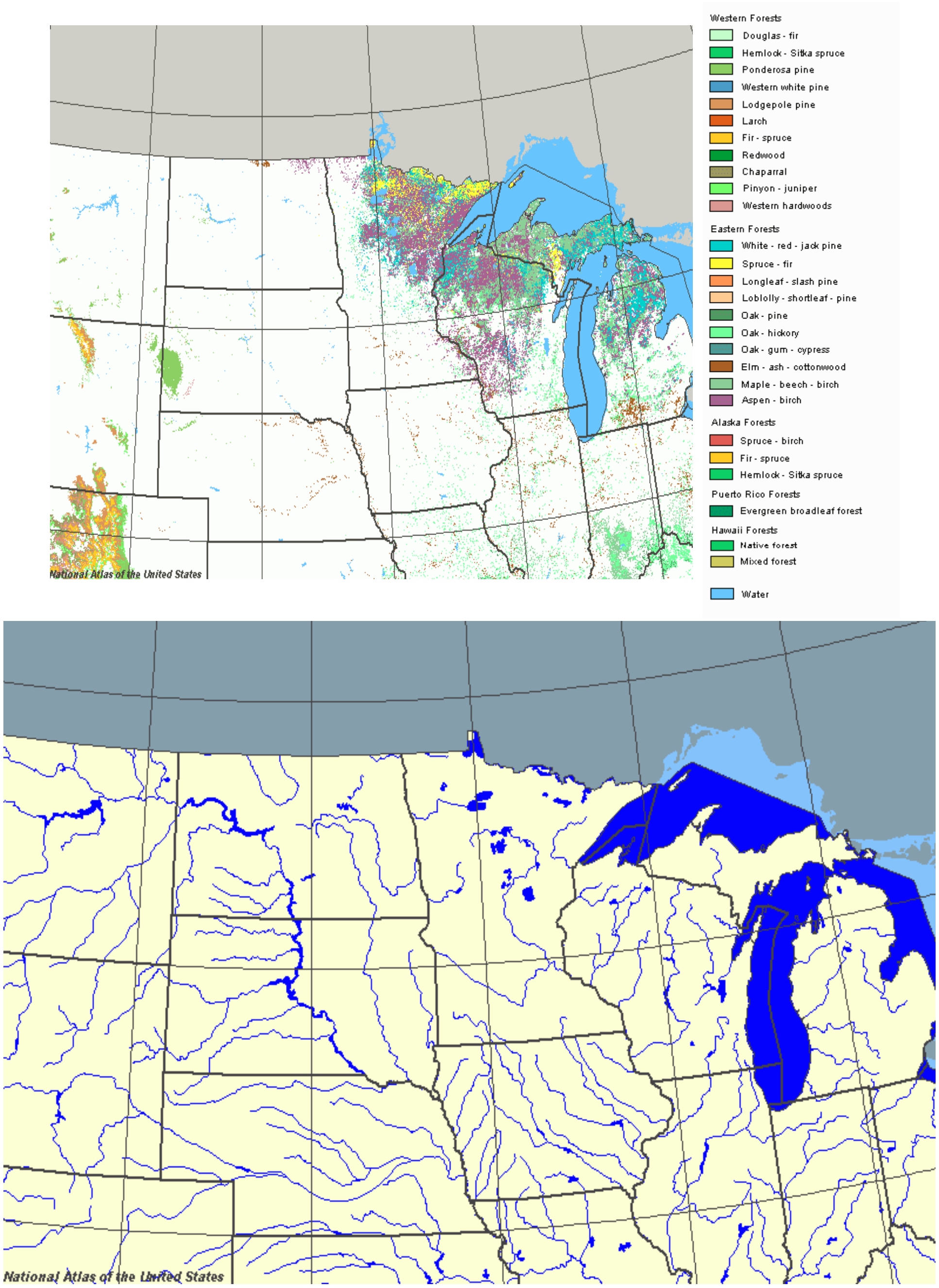


Figure 3.2: Top: Forest cover types. Bottom: Streams and water bodies. Map key in the Appendix. Source: www.nationalatlas.gov.

thematic map in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map. The Fig. 3.5 displays the choropleth map of the total wildfires per county, the point observations were aggregated to a county level. It shows that the counties on the southern west region of South Dakota are the ones with the highest number of wildfires. The north central region of Minnesota and the Vilas county in Wisconsin also experienced a higher number of wildfires.

One of the limitations of a choropleth map is that makes very hard to detect patterns. It could also lead to a bias interpretation because of the absence in the analysis of prognostic factors like the area (km^2's), vegetation and fuel conditions. The following section reviews smoothing methods for grouped data that overcome some of the limitations of the chloropleth maps.

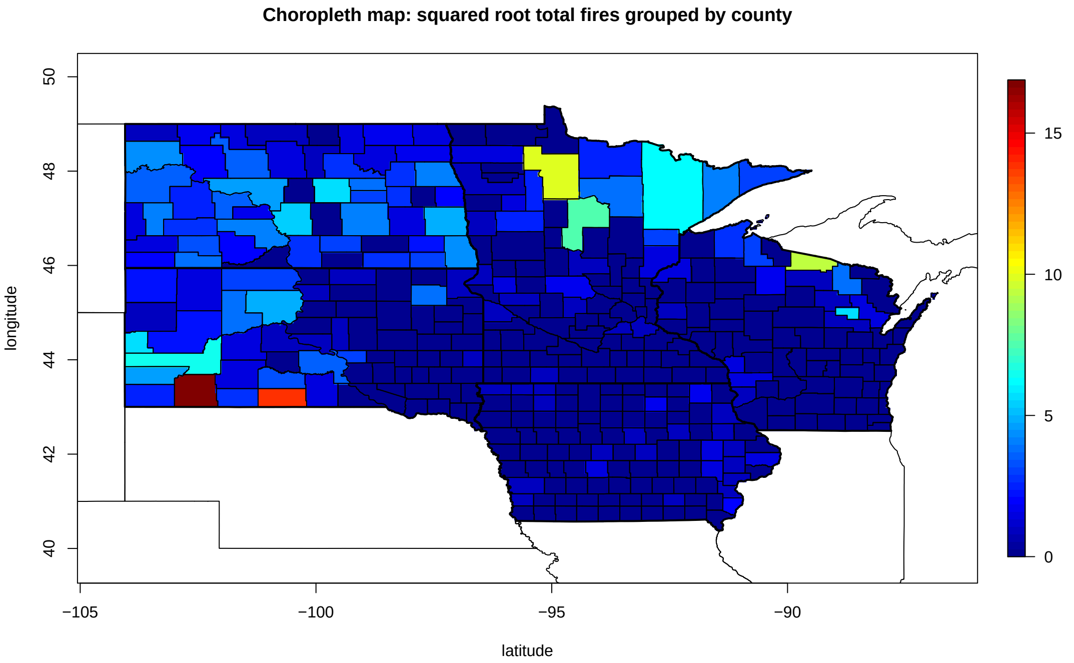


Figure 3.3: Squared root of total wildfires per county for 1990 visualized in a choropleth map $n_j = N(\Omega_j)$ $j = 1, \ldots, J$ where $J = 383$.

3.2 Smoothing spatially grouped data

Geographic data are often reported as counts or averages over some regions in space like: income, diseases, unemployment, etc. When data are reported this way, but the basic phenomenon is a point process such as: disease or fire locations, one has the change of support problem (COS). In the literature [19], it is also referred as spatial misalignment.

Some authors have studied phenomena (i.e. diseases, births) that suggest an intensity function that varies smoothly over the space and the data is reported grouped or averaged [53, 15, 2, 16, 40]. These works allude to the case of just one level of data aggregation, either counts or averages. Although we are interested in integrating two types of data aggregation, it is pertinent to include a review of these works as we present a generalization of their problem. A more extensive review is provided in the following references [20, 19].

Tobler is one of the earlier references in the problem of obtaining a smooth map of a geographical distribution under the constraint that the original data is grouped in a discrete collection of regions. He was interested in knowing how a population density, a continuous quantity, varies over the USA. The interest in obtaining a smooth population density was based on the fact that geographically, there is a mutual influence of places, in this case between states.

Tobler was interested in describing the population density in an area $\mathbf{\Omega}$ given the average population density in each of J non-overlapping and contiguous subareas $\mathbf{\Omega} = \bigcup_{j=1}^{J} \Omega_j$. One of the examples included in that paper was the population of the 48 contiguous states in the United States. The problem was to obtain a smooth, non-negative function λ with the volume matching property:

$$\frac{1}{|\Omega_j|} \iint\limits_{\Omega_j} \lambda(x, y) dx dy = z_j \qquad j = 1, \ldots, J \tag{3.1}$$

where $|\Omega_j|$ is the area of Ω_j, and z_j is the observed average value of λ over Ω_j. Tobler suggested seeking λ to minimize

$$J_1^{\Omega}(\lambda) = \iint\limits_{\Omega_j} \left[\left(\frac{\partial \lambda(x, y)}{\partial x} \right)^2 + \left(\frac{\partial \lambda(x, y)}{\partial y} \right)^2 \right] dx dy \tag{3.2}$$

subject to

$$\frac{1}{|\Omega_j|} \iint\limits_{\Omega_j} \lambda(x, y) dx dy = z_j \qquad j = 1, \ldots, J$$

and

$$\lambda(x,y) \geq 0 \qquad (x,y) \in \Omega$$

Wahba [54] extended Tobler's results for the standarized age adjusted female lung cancer rates in Wisconsin. The data covered the period 1970-1975 and was grouped by county for the 72 counties. She suggested to minimize

$$J_k(\lambda) = \sum_{v=0}^{k} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \binom{k}{v} \left(\frac{\partial^k \lambda(x,y)}{\partial x_1^v \partial x_2^{k-v}} \right)^2 dxdy \tag{3.3}$$

instead of J_1^{Ω}, subject to

$$\frac{1}{|\Omega_j|} \iint\limits_{\Omega_j} \lambda(x,y)dxdy = z_j \qquad j = 1, \ldots, J$$

By minimizing Equation 3.3 instead of Equation 3.2 ensures that the solution $\hat{\lambda}$ exists uniquely provided certain conditions. These conditions can be consulted at [55]. Wahba generalized these results to the volume smoothing problem, find $\lambda \in \chi$:

$$\frac{1}{J} \sum_{j=1}^{J} w_j \left(z_j - \frac{1}{|\Omega_j|} \iint\limits_{\Omega_j} \lambda(x,y)dxdy \right)^2 + h \sum_{v=0}^{k} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \binom{k}{v} \left(\frac{\partial^k \lambda(x,y)}{\partial x^v \partial y^{k-v}} \right)^2 dxdy$$

where w_j is an adequate weight and h is a smooth parameter. Wahba took χ to be the Sobolev space of square integrable functions on Ω with seminorm

$$J_2^{\Omega}(h) = \sum_{v=0}^{k} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \binom{k}{v} \left(\frac{\partial^k \lambda(x,y)}{\partial x^v \partial y^{k-v}} \right)^2 dxdy < \infty$$

where $J_2^{\Omega}(h)$ is a semi-norm on χ with null space the dimensional space $\binom{k+1}{2}$ spanned by the polynomials $(1, x, y, \ldots)$ of total degree $\leq k-1$.

Brillinger [2] worked with the analysis and display of spatially aggregate data over geographic regions for phenomena that are felt to vary smoothly in space. He provided a spatial locally weighted analysis of the births taking place to women aged 25 to 29 in the years 1986 and 1987 for the province of Saskatchewan at the census division level. The effect of the census division Ω_j on the location (x, y) is included in the estimation of the model through the following type of weights:

$$w_j(x,y) = \frac{1}{|\Omega_j|} \iint_{\Omega_j} \mathcal{W}(x-u, y-v)dudv \tag{3.4}$$

with $\mathcal{W}$ the biweight,

$$\mathcal{W}(x,y) = \begin{cases} (1-u^2)^2 & |u| \leq 1, \\ \quad\quad 0 & e.o.c. \end{cases}$$

where $u = \frac{\sqrt{x^2+y^2}}{h}$ for some bandwidth $h > 0$. This approach provided an alternative to the empirical Bayes methods that have been proposed for similar problems.

Eddy and Mockus [16] wanted to estimate and display the incidence rate of mumps as a smoothly varying function of space and time. They find that a smooth function (of the incidence rate) is the best way to convey the behaviour of the spread of the disease. They were interested in the estimation of a smoothly varying scalar function λ with a three dimensional argument $\lambda(x, y, t)$. They thought of x and y as indicating spatial coordinates and t as indicating a time coordinate. First we will focus on the spatial domain of their study.

Their data consist on the monthly averaged incidence rates of mumps reported for each state in USA for the period 1968-1988. Their problem is to estimate a function $\lambda(x, y)$ (just the spatial domain) given sets $\{\Omega_j\}$, states, and data $\{z_j\}$, average rate, for $j = 1, \ldots, J$. The relationship between λ and data is given by

$$z_j = \iint\limits_{\Omega_j} \lambda(x,y)dG(x,y) \qquad j = 1, \ldots, J. \tag{3.5}$$

where $\Omega_j \subset \mathbf{\Omega}$ and $G(x, y)$ is the population distribution.

They propose to estimate $\lambda(x, y)$ as in Equation **??** as follow:

- Choose a set of points and values for every Ω_j, $(x_{ij}, y_{ij}) \in \Omega_j$ and $i = 1, \ldots, k_j$ respectively for $j = 1, \ldots, J$.
- The number of points k_j in Ω_j are taken to be proportional to the area $|\Omega_j|$.
- The points (x_{ij}, y_{ij}) are distributed in Ω_j so that they repel each other and the boundary of Ω_j.

- The values of μ at (x_{ij}, y_{ij}) are assumed constant for each Ω_j, $n_{ij} = n_j$.
- Use the estimator

$$\hat{\mu}(x, y) = \frac{\sum_i \mathcal{W}(x - x_{ij}, y - y_{ij}) n_{ij}}{\sum_i \mathcal{W}(x - x_{ij}, y - y_{ij})} \tag{3.6}$$

where $\mathcal{W}(x - x_{ij}, y - y_{ij}) = \exp\{-\lambda ||s - s_{ij}||^2\}$, $s = (x, y)$, $s_{ij} = (x_{ij}, y_{ij})$ and λ is the smoothing parameter.

3.3 Locally weighted likelihood analysis: the Poisson case

Brillinger [2] points that locally weighted likelihood analysis is a pertinent estimation technique for nonelementary distributions varying in space. Stanislawis proposed the notion of a weighted likelihood for the nonparametric kernel estimation of a regression function. She considered data be of the form $(\mathbf{u}_l, Y_l)$ $l = 1, \ldots, L$ where $\mathbf{u}_l = (u_l, v_l) \in [0, 1]^2$ are lattice points and the Y_l are independent random variables from a family of distributions with parameter $\theta_l = \theta(u_l, v_l)$, with $\theta(\cdot)$ having a continuous partial derivative of order $k \geq 2$. The goal was to arrive at a non-parametric estimate $\hat{\theta}(x, y)$ for a fixed point $(x, y) \in [0, 1]^2$. She considered the estimator $\hat{\theta}$ that maximizes the weighted log-likelihood function:

$$\ell_w(\theta) = \sum_{l=1}^{L} \mathcal{W}\left(\frac{x - u_l}{h}, \frac{y - v_l}{h}\right) \log f(Y_l|\theta) \tag{3.7}$$

The number of wildfires has been modeled as a Poisson distribution by different authors [12, 43, 36, 37]. At first we consider that the total number of wildfires per county are distributed as a Poisson distribution. Under the assumption that the intensity function of the wildfire point process is smooth, a locally weighted likelihood estimate is used. This fitting procedure gives flexibility to include the grouping effect through the use of appropriate weights that borrow information from the neighbor regions.

There will be some improvements over the chloropleth map, by modeling the total number of wildfires by county as Poisson $\lambda(u, v|\theta)$ and using a weighted likelihood fit. Some of the improvements are the inclusion in the model of the influence of the neighbour's wildfires in the location (x, y) through weights of the form Equation 3.4. Secondly the intensity function varies smoothly over space and conveys better the trend of the wildfires.

As a result of including the neighbour states to the pilot group (MN, ND, WI, SD, IA), the extended sample space is called $\mathcal{S}$ as shown in Fig. 3.4. Assume f follows a Poisson distribution and the weighted log-likelihood function follows:

$$\ell_w(\theta) = -\sum_{j=1}^{J} w_j(x,y)\left[\iint\limits_{\Omega_j} \lambda(u,v|\theta)dudv + n_j \log \iint\limits_{\Omega_j} \lambda(u,v|\theta)dudv\right] \tag{3.8}$$

$\theta = \theta(x,y)$ where (x,y) is a point location in the pilot group and $J = 860$.

If we assume that the intensity rate is constant over the sample space $\lambda(u,v|\theta) = \lambda$, an homogeneous Poisson process, the λ that maximizes at position (x,y) the weighted log-likelihood function Equation 3.8 has the following form:

$$\hat{\lambda}(x,y) = \frac{\sum_{j=1}^{J} w_j(x,y)n_j}{\sum_{j=1}^{J} w_j(x,y)|\Omega_j|} \tag{3.9}$$

where $|\Omega_j|$ is the area of j county in km^2. It is an extension of the kernel estimator studied by [18, 41, 46]:

$$\hat{\lambda}(x,y) = \sum_{l=1}^{L} Y_l \mathcal{W}\left(\frac{x-u_l}{h}, \frac{y-v_l}{h}\right) / \sum_{l=1}^{L} \mathcal{W}\left(\frac{x-u_l}{h}, \frac{y-v_l}{h}\right) \tag{3.10}$$

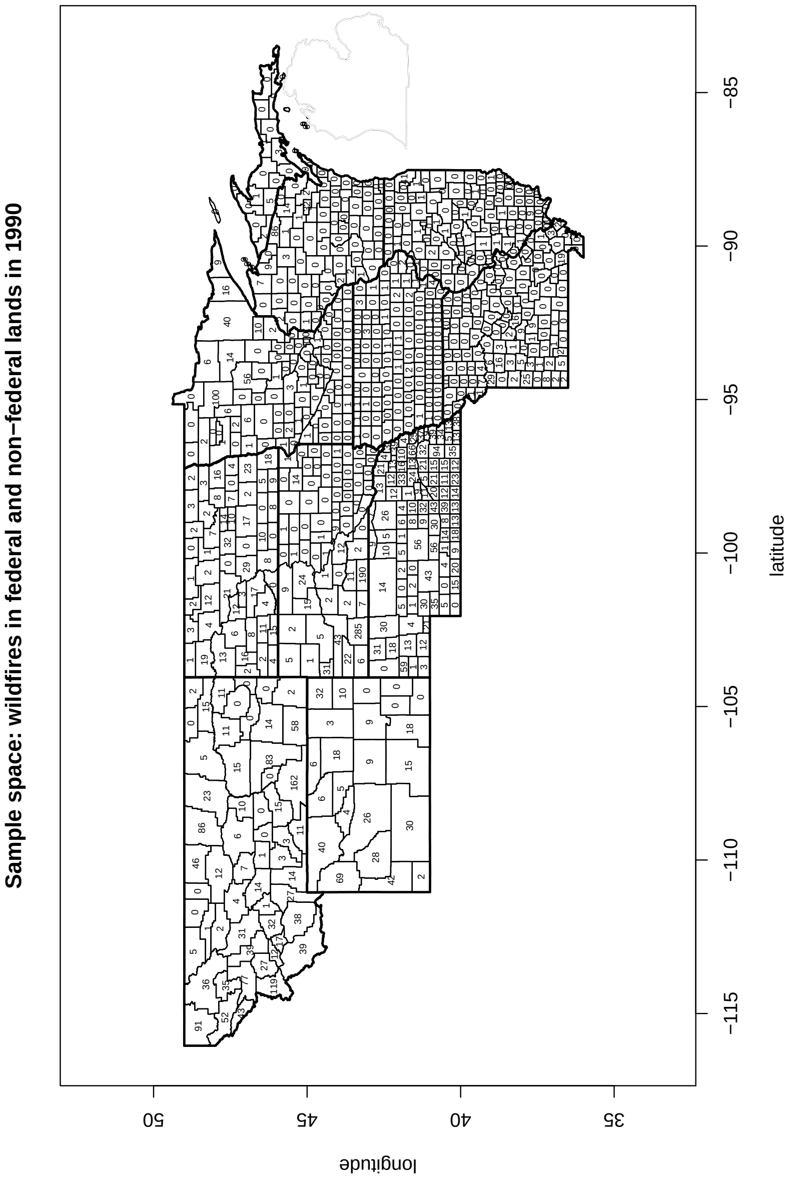

Sample space: wildfires in federal and non−federal lands in 1990
latitude
longitude
−85
−90
−95
−100
−105
−110
−115
50
45
40
35

Figure 3.4: The pilot group states plus their neighbor states with their corresponding total wildfires per county in 1990.

We use two different weight functions in the smoothed estimate of the intensity function in Equation 3.16, a naive weight and a biweight function. Other weight functions are explored by different authors [54, 30]

1. A naive weight function is first used in Equation 3.16

$$w_j(x,y) = \begin{cases} \frac{1}{|\Omega_j|} & (x,y) \in \Omega_j, \\ 0 & e.o.c. \end{cases}$$

   This weight corresponds to $\mathcal{W}(\cdot)$ a delta function in Equation 3.4. The estimate of the smoothed intensity function is shown in the bottom part of Fig. 3.5.

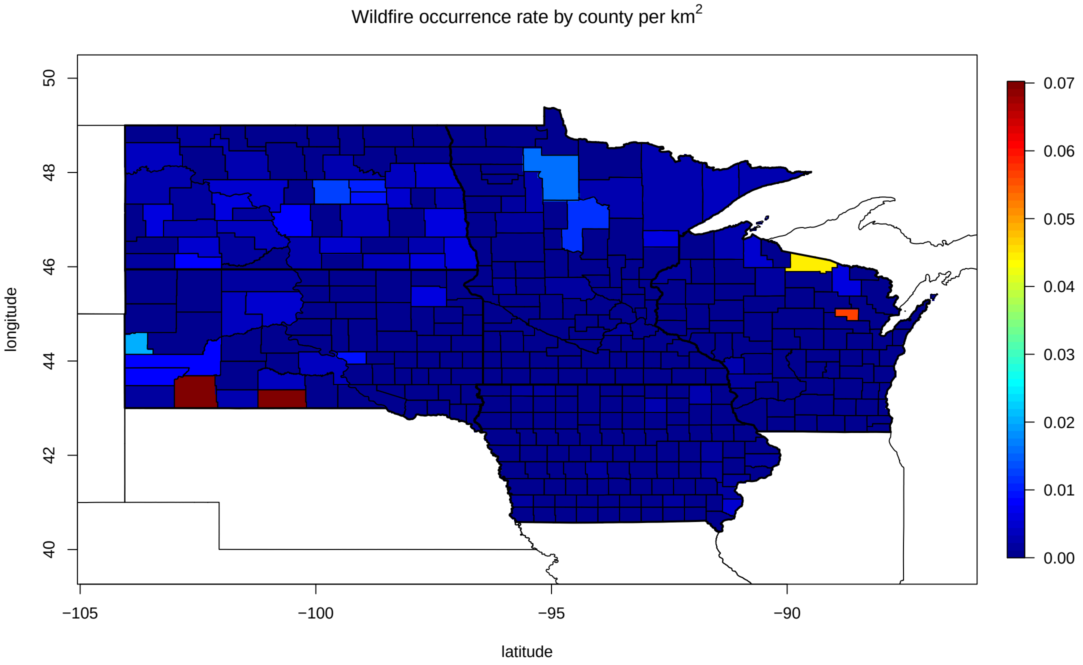


Figure 3.5: Wildfire occurrence rate by county per km^2. It is equivalent to smooth the spatially grouped data by county with the weight function in Equation 1.

2. A second analysis is done with the biweight function (also called bisquare) for 4 different bandwidths, we will name h_2 the bandwidths associated to grouped data $h_2 = (.5, 1, 1.5, 2)$. The weights $w_j(x,y)$ follow Brillinger's approach Equation 3.4 and $\mathcal{W}$ is the biweight function Equation 3.2. The weight function for the county Sioux in North Dakota for four different bandwidths are included in figures 3.6, 3.7 and 3.8. The four values of h_2 illustrated in that figure correspond to no smoothing, a small

amount of smoothing, a moderate amount and a significant amount of smoothing.

The weight functions are shown in three different displays: a grid color-scale rectangles also known as image, contour and perspective plots. For small bandwidths the weight function is constant over the region Ω_j, at the boundaries the value is zero and it doesn't extend to the contiguous counties. As the bandwidth increases, the effect of the fires occurred at a particular county extends to larger regions. The census division are artificial boundaries (with some exceptions) that doesn't define different properties that influence wildfire occurrences.

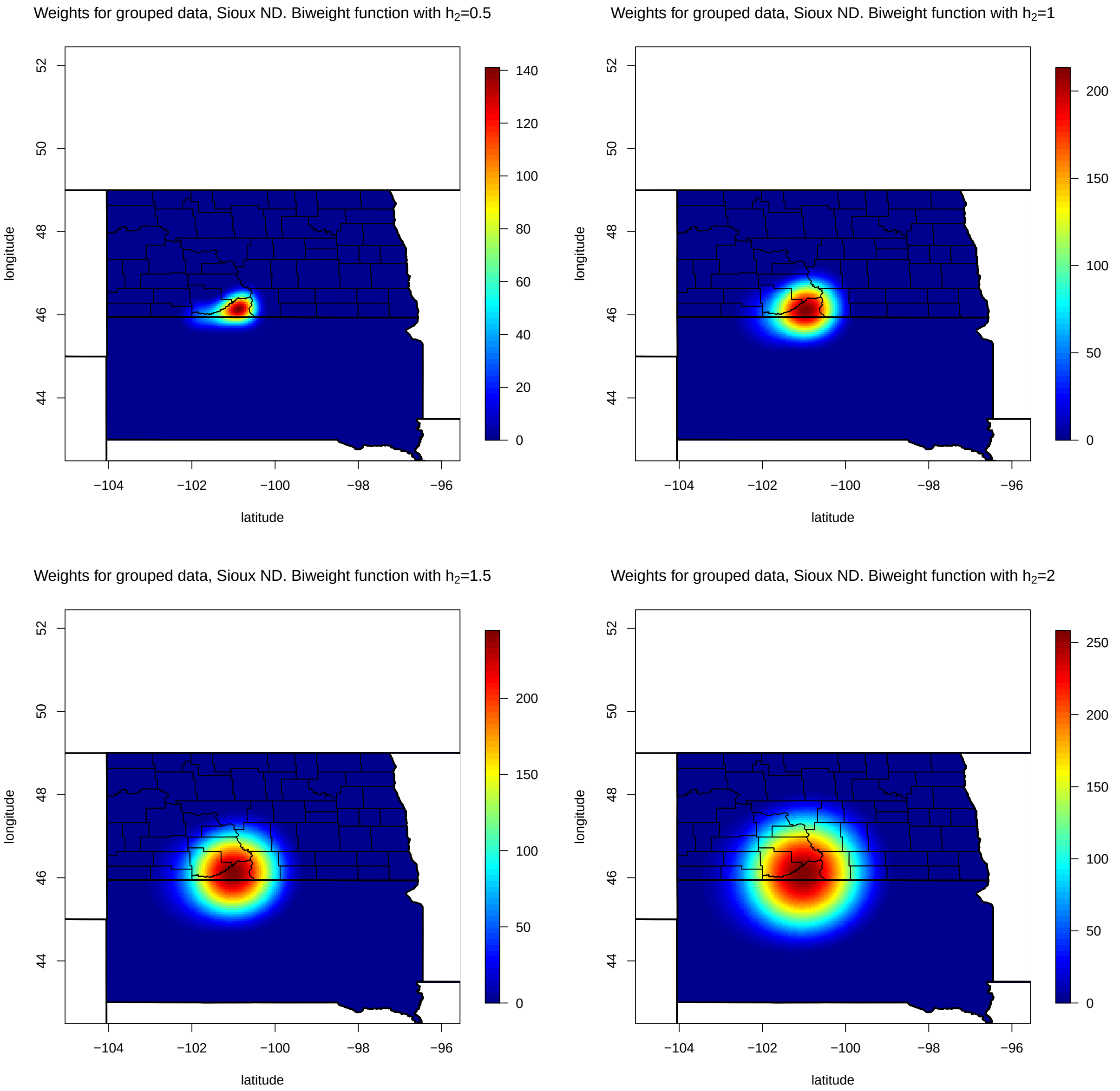
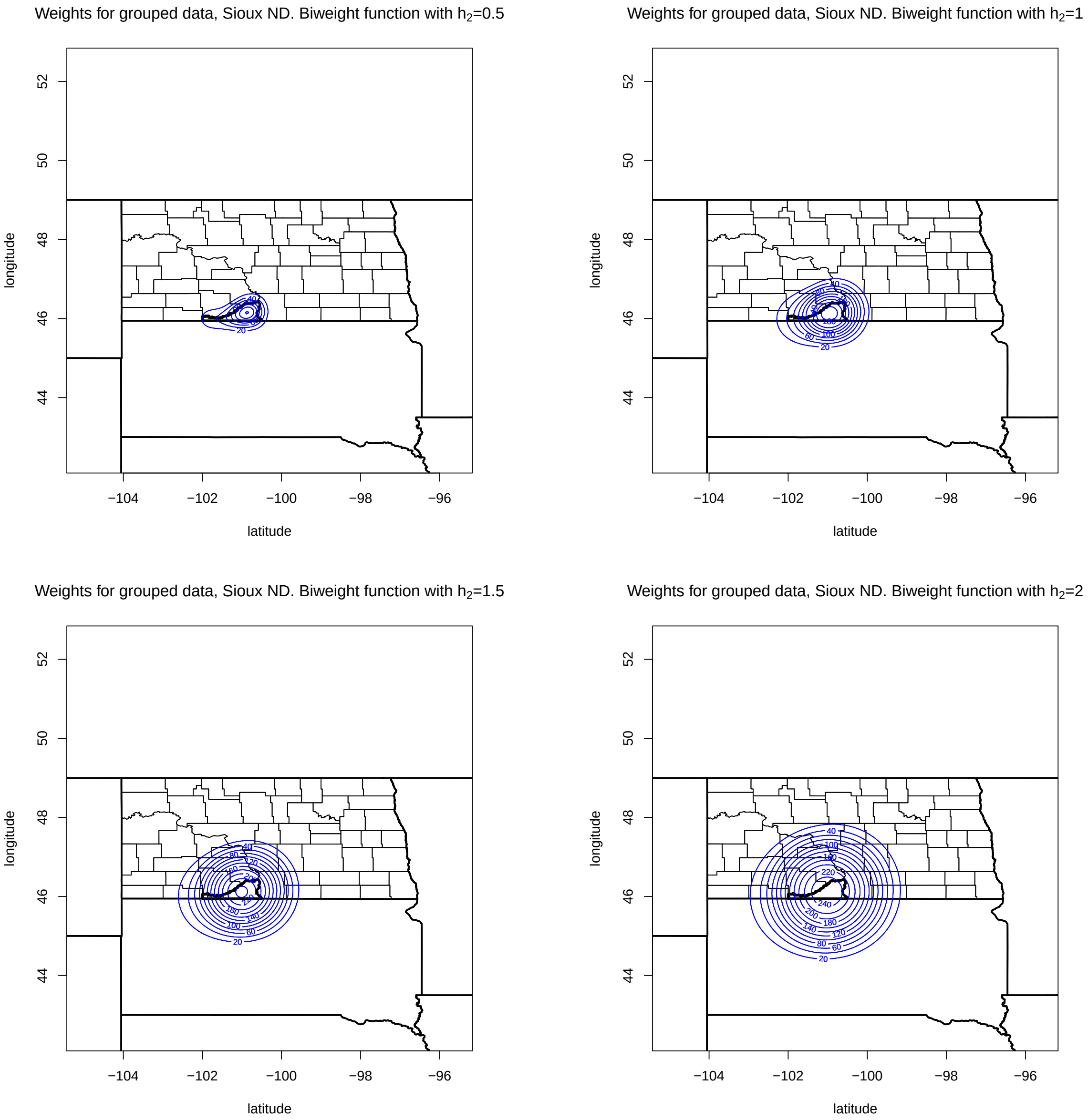

Weights for grouped data, Sioux ND. Biweight function with h_2=0.5
Weights for grouped data, Sioux ND. Biweight function with h_2=1
Weights for grouped data, Sioux ND. Biweight function with h_2=1.5
Weights for grouped data, Sioux ND. Biweight function with h_2=2
longitude
latitude
52
50
48
46
44
−104
−102
−100
−98
−96
0
20
40
60
80
100
120
140
50
150
200
250

Figure 3.6: Image plots of the weight function $w_j(x, y)$ propose in definition Equation 3.4 for the data that correspond to Sioux county in North Dakota. The sequence of plots correspond to the following bandwidths $h_2 = (.5, 1, 1.5, 2)$ which goes from no smoothing to a significant amount of smoothing.



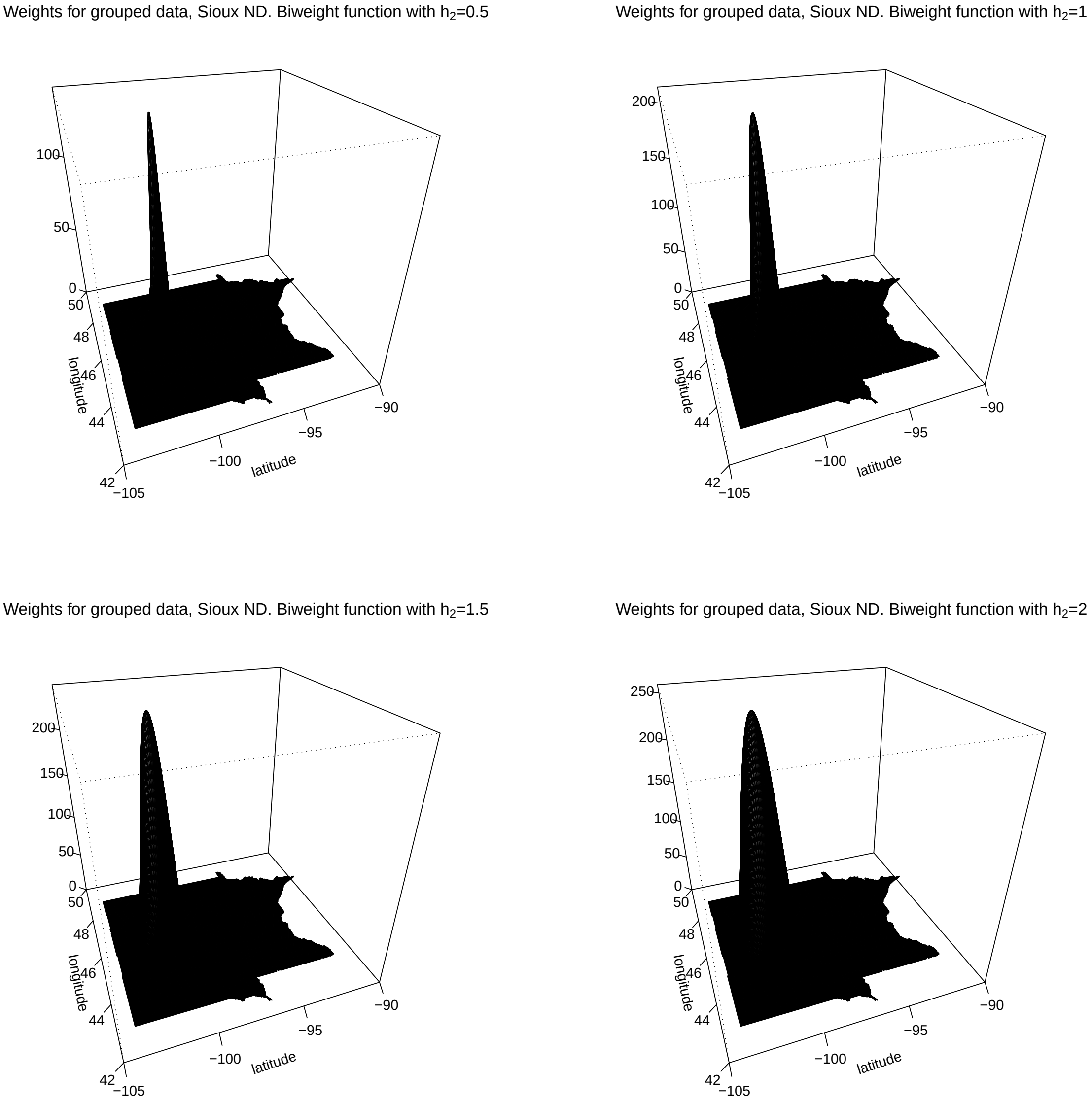
Weights for grouped data, Sioux ND. Biweight function with h_2=0.5

Weights for grouped data, Sioux ND. Biweight function with h_2=1

Weights for grouped data, Sioux ND. Biweight function with h_2=1.5

Weights for grouped data, Sioux ND. Biweight function with h_2=2

Figure 3.7: Contour plots of the weight function $w_j(x, y)$ propose in definition Equation 3.4 for the data that correspond to Sioux county in North Dakota. The sequence of plots correspond to the following bandwidths $h_2 = (.5, 1, 1.5, 2)$ which goes from no smoothing to a significant amount of smoothing.


Weights for grouped data, Sioux ND. Biweight function with h_2=0.5

Weights for grouped data, Sioux ND. Biweight function with h_2=1

Weights for grouped data, Sioux ND. Biweight function with h_2=1.5

Weights for grouped data, Sioux ND. Biweight function with h_2=2

Figure 3.8: Perspective plots of the weight function $w_j(x, y)$ propose in definition Equation 3.4 for the data that correspond to Sioux county in North Dakota. The sequence of plots correspond to the following bandwidths $h_2 = (.5, 1, 1.5, 2)$ which goes from no smoothing to a significant amount of smoothing.

The figure 3.9 shows the smoothed estimates for the intensity function in Equation 3.16 with the weight functions for four different bandwidths as shown in Figs. 3.6, 3.7 and 3.8. In the first two plots, by applying none or very little smoothing, we can detect two spots with higher intensity with respect to the rest of the territory of interest. These two spots are located in the south-east border of South Dakota. When using a wider bandwidth like in the third and fourth plots, it seems that the wildfire occurrences in north Minnesota are related to the conditions next to the Superior Lake. Also the fires occurred in Wisconsin seems to be related with the Superior Lake. Using a locally weighted likelihood analysis, it is also possible to identify an interrelation between states, for example the fires occurred in the border of South Dakota are related to the ones in north Nebraska. Also it seems that there is a relation between the fires in the border between Nebraska and Iowa.

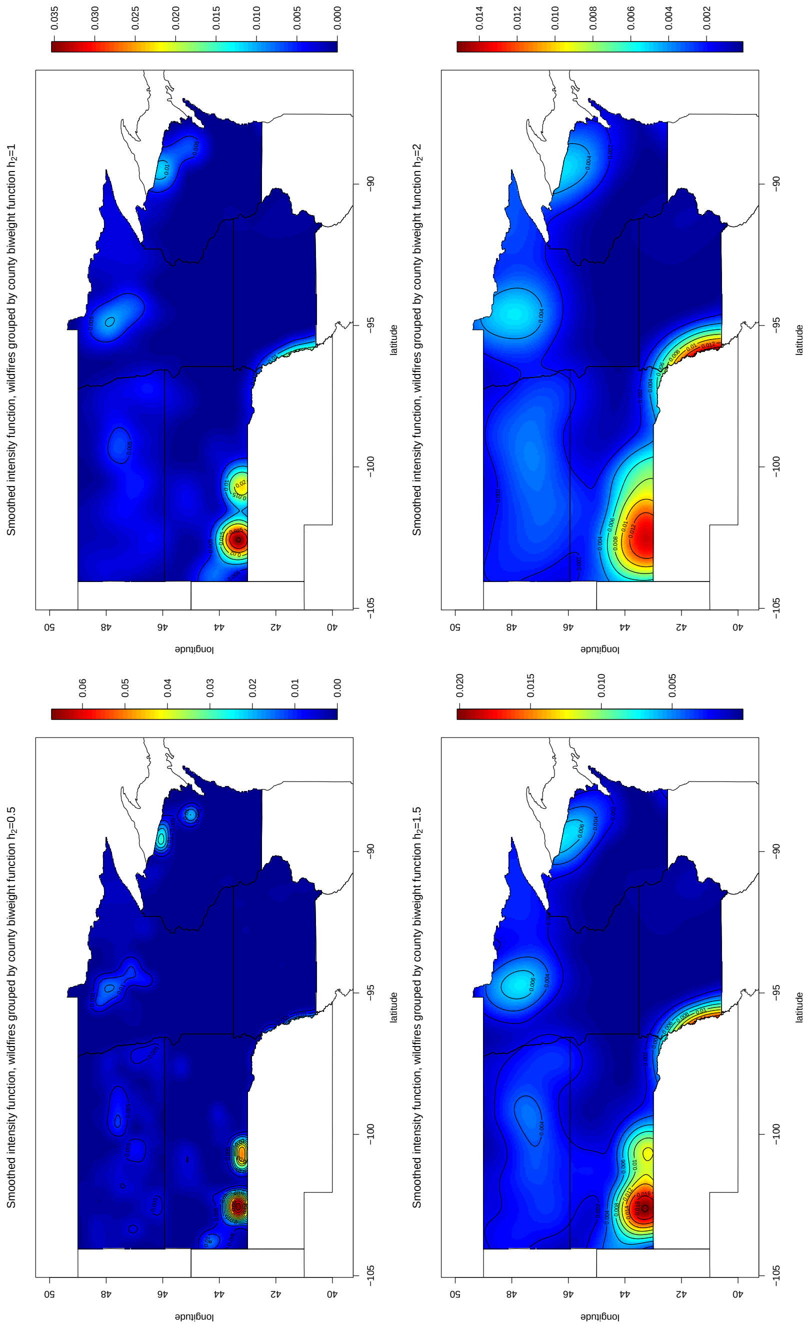

Smoothed intensity function, wildfires grouped by county biweight function h2=1
Smoothed intensity function, wildfires grouped by county biweight function h2=2
Smoothed intensity function, wildfires grouped by county biweight function h2=0.5
Smoothed intensity function, wildfires grouped by county biweight function h2=1.5
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.020
0.015
0.010
0.005
latitude
longitude
−90
−95
−100
−105
50
48
46
44
42
40

Figure 3.9: Smoothed estimate of the intensity function when data is grouped by county and assume to follow a Poisson distribution. These estimates are based on Equation 3.16 for four different bandwidths $h_2 = (.5, 1, 1.5, 2)$. Each corresponds to different amount of smoothing as shown in Figs. 3.6, 3.7, 3.8.

3.4 Partially grouped data

In the previous section, we assumed that the number of wildfires per county can be modeled by a Poisson distribution with intensity function $\lambda(x, y|\theta)$. Under the assumption of a smoothed intensity rate, we estimated it at the position (x, y), $\hat{\lambda}(x, y)$, for different bandwidths $h_2 = (.5, 1, 1.5, 2)$. This approach neglected the point locations in the federal regions and incorporated them to the model by grouping them by county.

In this section, we present a new model that incorporates the point locations and the counts. We propose using a binary-valued process to approximate the wildfire point process to merge the grouped and ungrouped data. Asymptotically this model also assumes that the number of wildfires in the sample space $\mathcal{S}$ can be modeled as a Poisson process. Using the binary approximation we arrive to a smoothed intensity function, that turns to be robust as will be shown in section 4.6.

First we will include the results obtained by Brillinger, Preisler and Benoit [5] to approach a spatial-temporal point process by a binary process. For now we work with the spatial case.

3.4.1 Approximation of a point process by a binary-valued process

Supposing that the space domain is broken up into pixels $(x, x + dx] \times (y, y + dy]$. Consider the spatial point process, N, with intensity function assumed to exist and defined by

$$\lambda(x, y) = Prob\{dN(x, y) = 1\}/dxdy$$

where $dN(x, y) = N(dx, dy)$ counts the number of fires in the pixel. Supposing that λ contains a parameter θ the log-likelihood function will be written.

$$\ell(\theta) = \int_x \int_y \log\left(\lambda(x, y)|\theta\right) dN(x, y) - \int_x \int_y \lambda(x, y|\theta) dxdy \tag{3.11}$$

Snyder and Miller [49]. Covariates can be included in λ.

We use one of the practical approaches to using the log-likelihood in practice Equation 3.11. Replace the spatial point process $N(dx, dy)$, by 0-1 valued process $Y_{x,y}$ on a lattice with $Y_{x,y} = 1$ if there is a fire in the corresponding pixel and by 0 otherwise for $(x, y) \in \mathbb{R}$.

Suppose then that

$$Prob\{Y_{x,y} = 1\} = \lambda_{x,y} \tag{3.12}$$

A Bernoulli approximation to the log-likelihood 3.11 is now

$$\sum_{x,y} Y_{x,y} \log\left(\lambda_{xy}\right) + \sum_{x,y} (1 - Y_{x,y}) \log\left(\lambda_{x,y}\right)$$

To simplify the notation for the moment, index the pixels by l rather than (x, y). In what follows we assumed that Y_l are independent given the covariates, the grouping effect and the influence of adjacent fires. The grouping effect and the influence of adjacent fires are included in the model by the weight functions employed in the fitting procedure.

$$\sum_{l} Y_l \log\left(\lambda_l\right) + \sum_{l} (1 - Y_l) \log\left(\lambda_l\right)$$

3.4.2 Approximation of a partially grouped sample by a binary-valued process

The problem of interest is to obtain a robust smoothed estimate of the intensity function for the wildfire point process for a partially grouped sample. The log-likelihood function for the point locations $\tau_i = (x_i, y_i)$ $i = 1, \ldots, I$ and the total number of fires by county $N(\Omega_j) = n_j$ $j = 1, \ldots, J$ is

$$\begin{aligned} \ell(\theta) &= \iint\limits_{\mathcal{B}} \log\left(\lambda(x, y)|\theta\right) dN(x, y) - \iint\limits_{\mathcal{B}} \lambda(x, y|\theta) dxdy \\ &+ \sum_{j=1}^{J} n_j \log \iint\limits_{\Omega_j} \left(\lambda(x, y)|\theta\right) dxdy - \sum_{j=1}^{J} n_j \iint\limits_{\Omega_j} \lambda(x, y|\theta) dxdy \end{aligned}$$

This log-likelihood function is approximated by replacing the partially grouped sample by a binary-valued process Y_l, on a lattice with $Y_l = 1$ if there is a fire in the corresponding pixel and by 0 otherwise. We propose the following binomial approximation to the log-likelihood function 3.13:

$$\begin{aligned} \ell(\theta) &= \sum_{l=1}^{N_{\mathcal{B}}} Y_l \log\left(\lambda_l\right) + \sum_{l=1}^{N_{\mathcal{B}}} (1 - Y_l) \log\left(1 - \lambda_l\right) \\ &+ \sum_{j=1}^{J} \log \binom{N_j}{n_j} + \sum_{j=1}^{J} n_j \log\left(\rho_j\right) + \sum_{j=1}^{J} (N_j - n_j) \log\left(1 - \rho_j\right) \end{aligned} \tag{3.13}$$

The fires occurred in federal regions are modeled as $Y_l \sim Bin(1, \lambda_l)$ and the counts of fires occurred in non-federal regions are modeled as $N(\Omega_j) \sim Bin(N_j, \rho_j)$. $N_{\mathcal{B}}$ is the collection of pixels in the federal regions and N_j is the collection of pixels in the j-county. While λ_l follows the definition in Equation 3.12, ρ_j is the probability of a fire in a pixel of the collection of pixels in Ω_j.

3.5 Smoothing spatially partially grouped data

In this section we present a smoothed estimate or sometimes called smoother, of the intensity function of the wildfire point process when the observations are partially grouped. The smoother is based on a locally weighted likelihood estimation (subsection 3.3) for the Binomial approximation to the log-likelihood function Equation 3.13. The weighted log-likelihood function is then:

$$\begin{aligned}\ell_w\left(\boldsymbol{\lambda},\boldsymbol{\rho}\right) &= \sum_{l=1}^{N_\mathcal{B}} \mathcal{W}(x-u_l,y-v_l)\left[Y_l\log\left(\lambda_l\right)+\left(1-Y_l\right)\log\left(1-\lambda_l\right)\right] \\ &+ \sum_{j=1}^{J} w_j(x,y)\left[n_j\log\left(\rho_j\right)+\left(N_j-n_j\right)\log\left(1-\rho_j\right)\right] \end{aligned} \tag{3.14}$$

and it is maximized at $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1(x,y),\ldots,\hat{\lambda}_{N_\mathcal{B}}(x,y))$ and $\hat{\boldsymbol{\rho}} = (\hat{\rho}_1(x,y),\ldots,\hat{\rho}_J(x,y))$.

We construct a smoother that assumes the rate of occurrence of a wildfire in the sample space $\mathcal{S}$ is constant. This assumption implies that $\lambda_1 = \ldots = \lambda_{N_\mathcal{B}} = \rho_1 = \ldots = \rho_j = \pi$ where $\pi \in \mathbb{R}$. Under the assumption of a constant rate, generally and for the Poisson case, given the weighting technique employed it might be assumed local constancy. The number of fires in the sample space follows an homogeneous Poisson process, where $E\left[Y_{(x,y)}\right] = Prob\{Y_{(x,y)} = 1\} = \pi$. The weighted log-likelihood function is then:

$$\begin{aligned}\ell_w\left(\pi\right) &= \sum_{l=1}^{N_\mathcal{B}} \mathcal{W}(x-u_l,y-v_l)\left[Y_l\log\pi+\left(1-Y_l\right)\log\left(1-\pi\right)\right] \\ &+ \sum_{j=1}^{J} w_j(x,y)\left[n_j\log\pi+\left(N_j-n_j\right)\log\left(1-\pi\right)\right] \\ \frac{\partial\ell_w\left(\pi\right)}{\partial\pi} &= \sum_{l=1}^{N_\mathcal{B}} \mathcal{W}(x-u_l,y-v_l)\left[\frac{Y_l}{\pi}-\frac{(1-Y_l)}{1-\pi}\right]+\sum_{j=1}^{J} w_j(x,y)\left[\frac{n_j}{\pi}-\frac{(N_j-n_j)}{1-\pi}\right] \end{aligned} \tag{3.15}$$

where (u_l, v_l) are grid points. It is maximized at

$$\hat{\pi}_1(x,y) = \frac{\displaystyle\sum_{l=1}^{N_\mathcal{B}}\mathcal{W}(x-u_l,y-v_l)Y_l+\sum_{j=1}^{J}w_j(x,y)n_j}{\displaystyle\sum_{l=1}^{N_\mathcal{B}}\mathcal{W}(x-u_l,y-v_l)+\sum_{j=1}^{J}w_j(x,y)N_j} \tag{3.16}$$

for $\forall\ (x, y) \in \mathcal{S}$. This estimate is also an extension of the kernel estimator Equation 3.10.

The weights included in the log-likelihood function incorporate to the model the local variation and the influence of nearby fires in the position (x, y). The weights $w_j(x, y)$ as defined in Equation 3.4 represent the influence of census division j on the location (x, y), like risks caused by humans or environmental conditions. These weights are shown in Fig. 3.6 for four different bandwidths h_2. The weights associated to the ungrouped data $\mathcal{W}$, integrate the effect of the fire point locations. The latter follows the idea of a kernel estimate of the intensity function of a point process for a bandwidth h_1.

Figure 3.10 shows the estimates using the partially grouped sample with the same bandwidths for the grouped and ungrouped data $h_1 = h_2$. In particular, we can see that there is a higher probability of a wildfire occurrence in two regions of Minnesota, in the north and north-east regions. The north region is located where the Red Lake and also the Mississippi River crosses this area. The north-east region is connected to the Lake Superior. In North Dakota there is a spot which can be related to the lakes that cover this area.

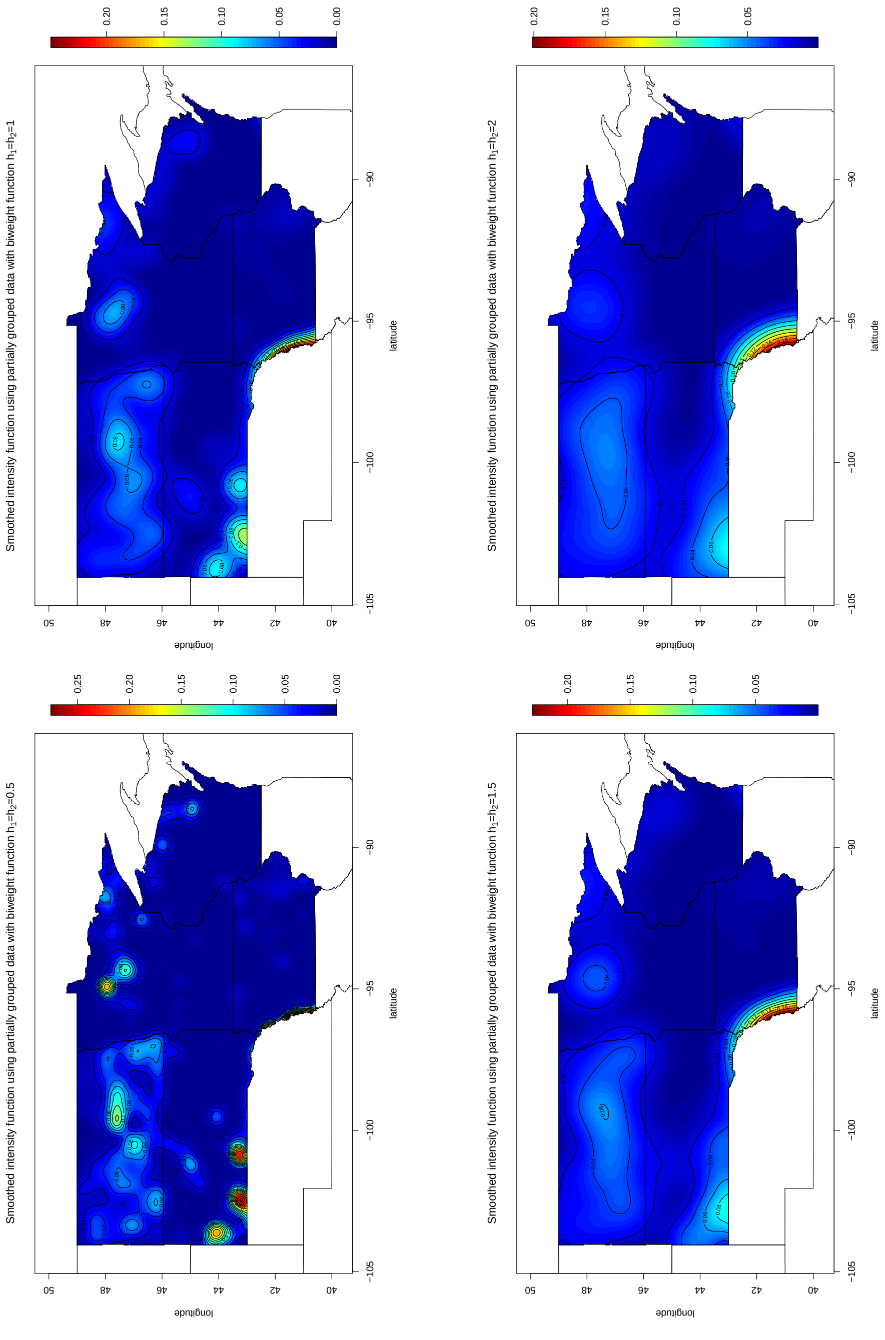

Smoothed intensity function using partially grouped data with biweight function h1=h2=1
Smoothed intensity function using partially grouped data with biweight function h1=h2=2
Smoothed intensity function using partially grouped data with biweight function h1=h2=0.5
Smoothed intensity function using partially grouped data with biweight function h1=h2=1.5
latitude
longitude
−105
−100
−95
−90
50
48
46
44
42
40
0.25
0.20
0.15
0.10
0.05
0.00

Figure 3.10: Smoothed estimate of the intensity function $\hat{\pi}(x, y)$ Equation 3.16 when the observations are reported partially grouped. The biweight function is used and the bandwidths are assumed equal for the ungrouped and grouped data $h_1 = h_2$. The four different estimates show different levels of smoothing that goes from very small smoothing to significant smoothed $h1 = h2 = (.5, 1, 1.5, 2)$.

3.6 Selection of the bandwidth

When implementing a semi-parametric or non-parametric model two things must be chosen: the weight function $\mathcal{W}$ and the bandwidth h. In our analysis, we use the biweight function based on its good performance and generally high efficiency over a wide range of distributions [31]. The bandwidth controls the size of the neighborhood, that is, the degree of smoothing imposed on the noisy observations. The bandwidth for each dimension h (latitude and longitude) are the same as shown in Equation (3.2).

We propose using two different bandwidths, $h = h_1$ for the ungrouped data and $h = h_2$ for the grouped data. By using a semi-parametric approach we are looking for borrow information from the neighbour regions. In a first analysis, we explore four different bandwidths to obtain different levels of smoothing. Figs. 3.11, 3.12, 3.13 and 3.14 include the smoother for the intensity function under partially grouped data Equation (3.16) for the different pair combinations of h_1 and h_2.

When combining data with different levels of aggregation we consider that the bandwidths selection has two main purposes: a) detecting local trends within each region when the data is either grouped or ungrouped b) borrowing strength from the neighbor's observations, either they are given in the same level of data aggregation or they are different. As a preliminary analysis we select the bandwidth for grouped data $h_2 = 1.5$ because the weight function extends to the neighbor counties as shown Fig. 3.8. We go with a wider bandwidth for the ungrouped data $h_1 = 2$. A following study will be to do a cross-validation analysis and look for the best combination of bandwidths.

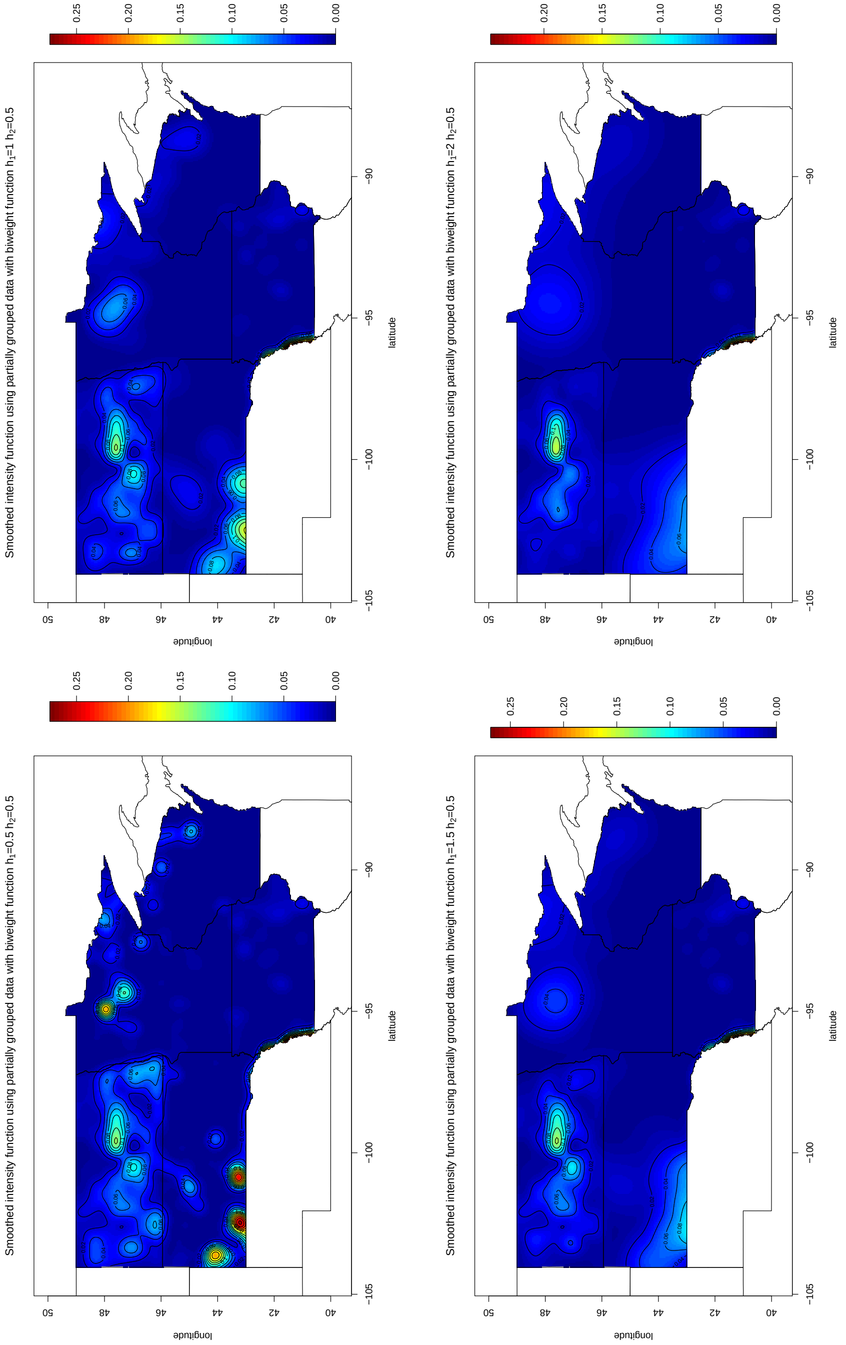

Smoothed intensity function using partially grouped data with biweight function h1=1 h2=0.5
Smoothed intensity function using partially grouped data with biweight function h1=2 h2=0.5
Smoothed intensity function using partially grouped data with biweight function h1=0.5 h2=0.5
Smoothed intensity function using partially grouped data with biweight function h1=1.5 h2=0.5
latitude
longitude
−105
−100
−95
−90
40
42
44
46
48
50
0.00
0.05
0.10
0.15
0.20
0.25

Figure 3.11: Smoothed estimate of the intensity function $\hat{\pi}_1(x, y)$ in Equation (3.16) when the observations are reported partially grouped. The biweight function is used with combinations of very little smoothing for grouped data ($h_2 = .5$) with a sequence of smoothness for ungrouped data $h_1 = (.5, 1, 1.5, 2)$.

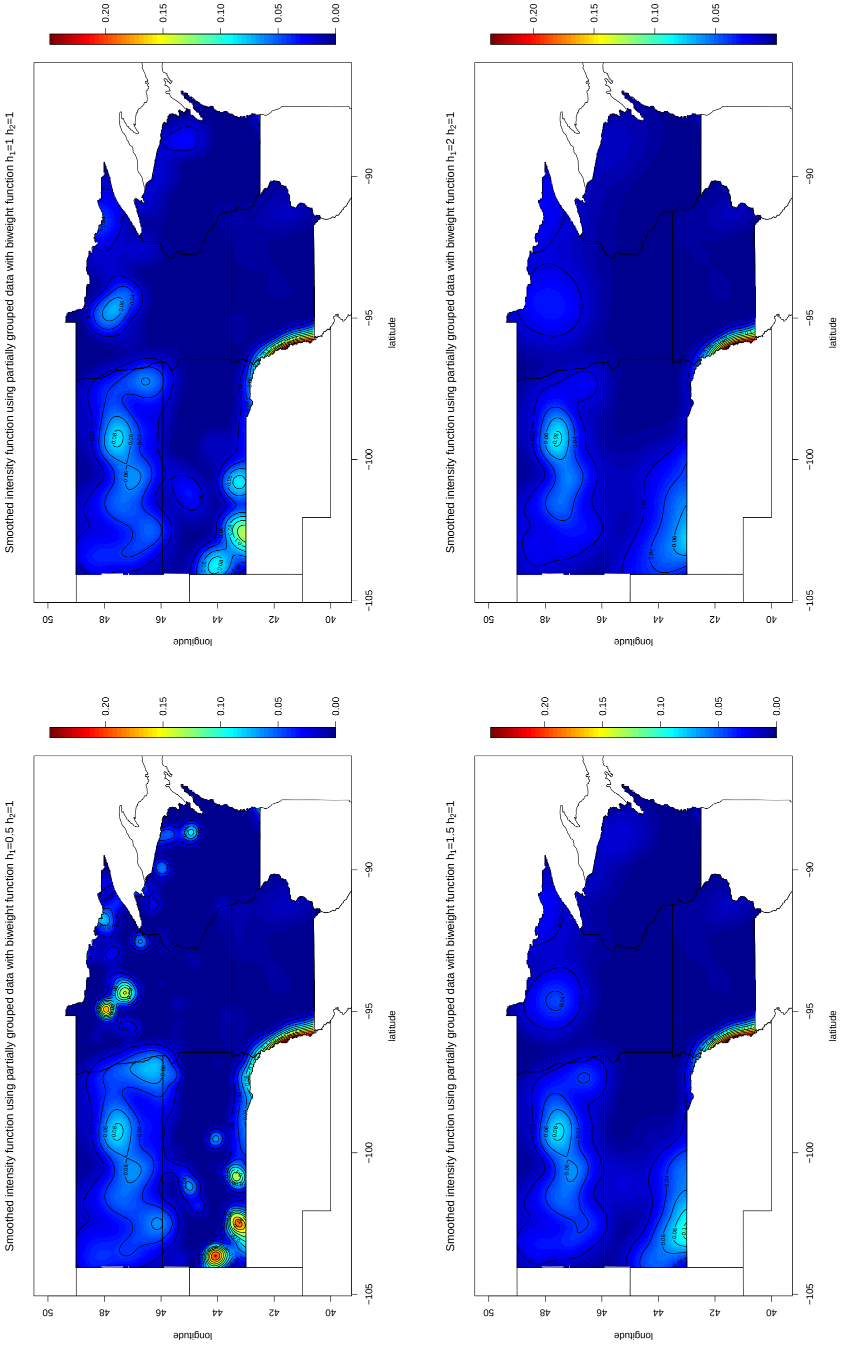

Smoothed intensity function using partially grouped data with biweight function h1=1 h2=1
Smoothed intensity function using partially grouped data with biweight function h1=2 h2=1
Smoothed intensity function using partially grouped data with biweight function h1=0.5 h2=1
Smoothed intensity function using partially grouped data with biweight function h1=1.5 h2=1
longitude
latitude
50
48
46
44
42
40
−105
−100
−95
−90
0.20
0.15
0.10
0.05
0.00

Figure 3.12: Smoothed estimate of the intensity function $\hat{\pi}_1(x, y)$ in Equation (3.16) when the observations are reported partially grouped. The biweight function is used with combinations of little smoothing for grouped data ($h_2 = 1$) with a sequence of smoothness for ungrouped data $h_1 = (.5, 1, 1.5, 2)$.

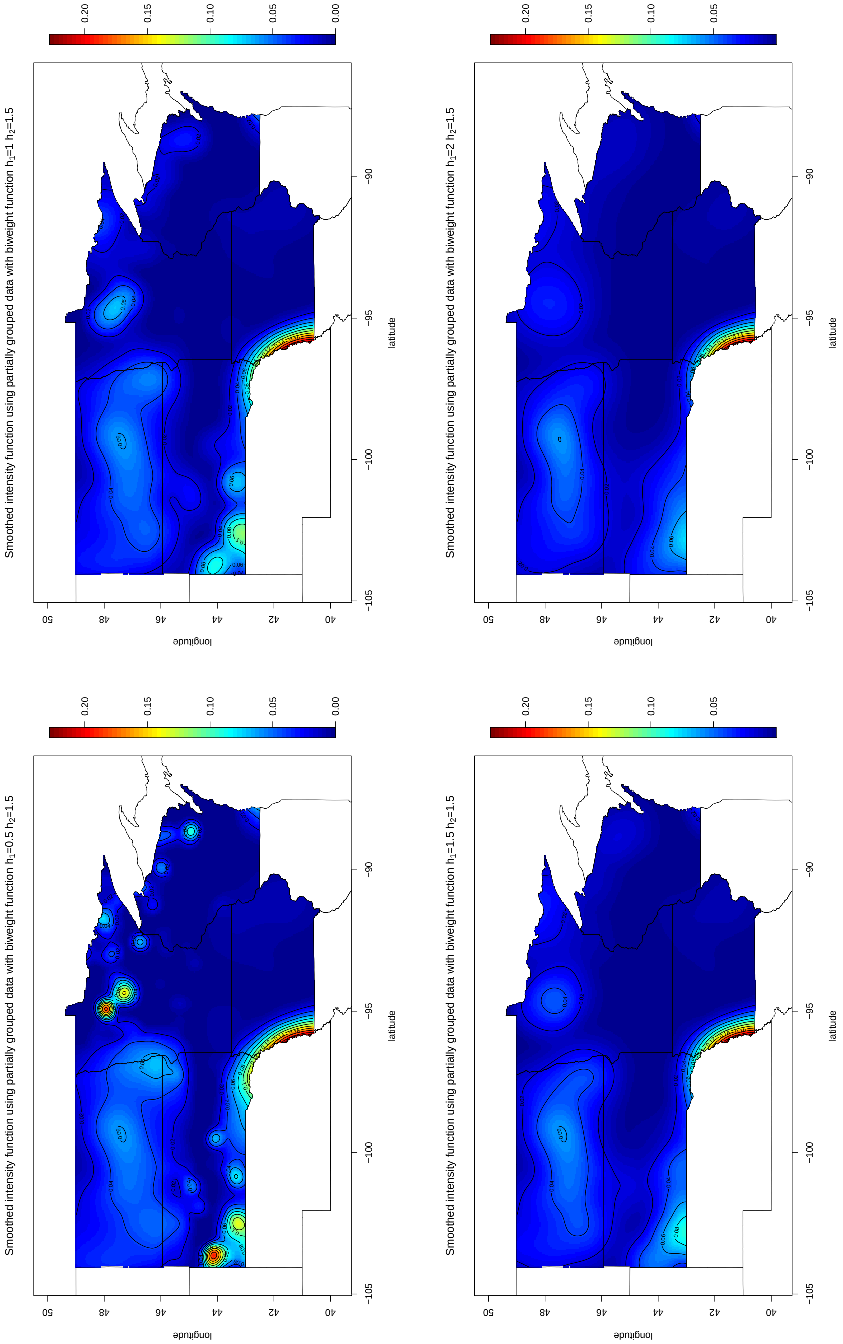

Smoothed intensity function using partially grouped data with biweight function h1=1 h2=1.5
Smoothed intensity function using partially grouped data with biweight function h1=2 h2=1.5
Smoothed intensity function using partially grouped data with biweight function h1=0.5 h2=1.5
Smoothed intensity function using partially grouped data with biweight function h1=1.5 h2=1.5
0.20
0.15
0.10
0.05
0.00
longitude
latitude
50
48
46
44
42
40
−90
−95
−100
−105

Figure 3.13: Smoothed estimate of the intensity function $\hat{\pi}_1(x, y)$ in Equation (3.16) when the observations are reported partially grouped. The biweight function is used with combinations of some smoothing for grouped data ($h_2 = 1.5$) with a sequence of smoothness for ungrouped data $h_1 = (.5, 1, 1.5, 2)$.

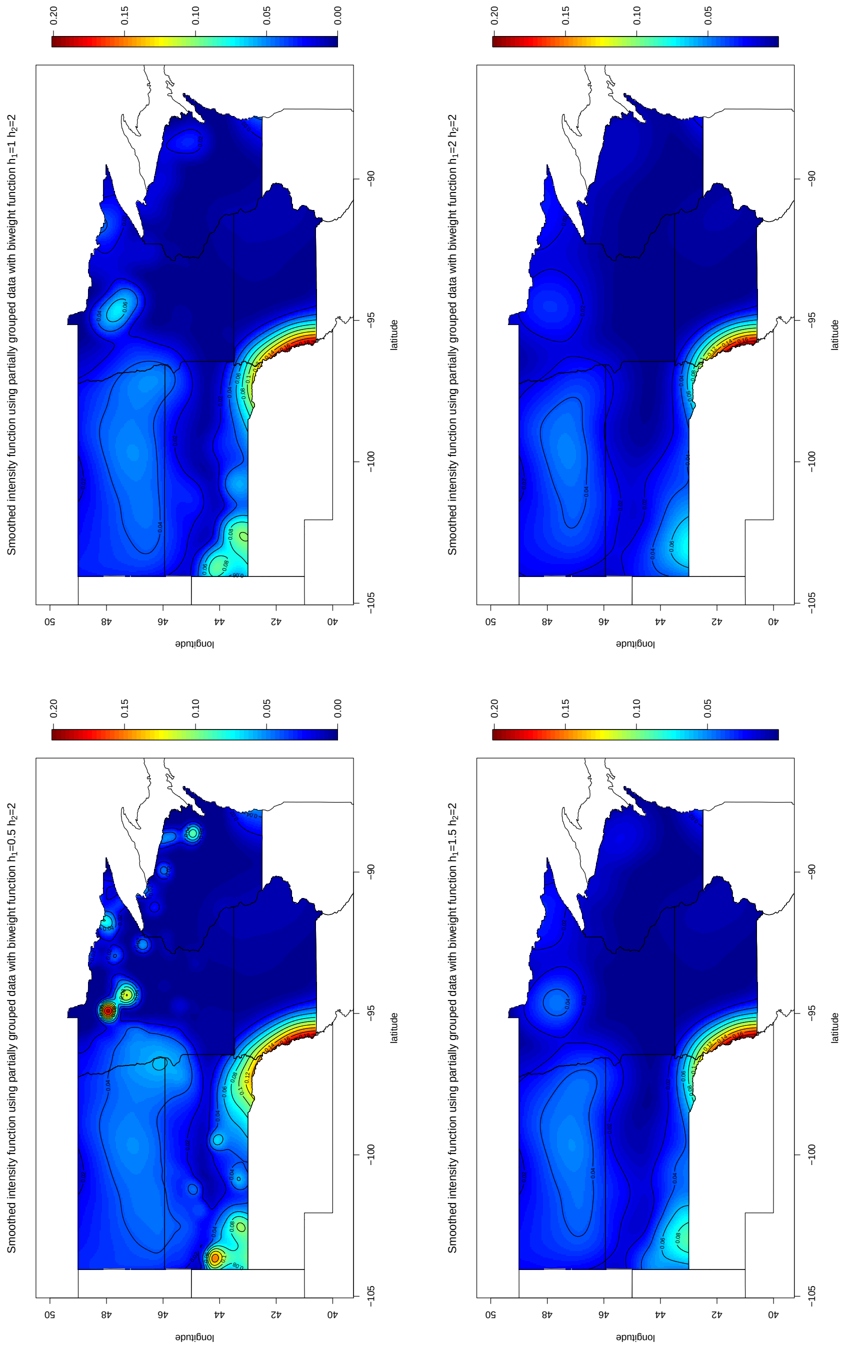

Smoothed intensity function using partially grouped data with biweight function h_1=1 h_2=2
Smoothed intensity function using partially grouped data with biweight function h_1=2 h_2=2
Smoothed intensity function using partially grouped data with biweight function h_1=0.5 h_2=2
Smoothed intensity function using partially grouped data with biweight function h_1=1.5 h_2=2
0.20
0.15
0.10
0.05
0.00
latitude
−105
−100
−95
−90
longitude
50
48
46
44
42
40

Figure 3.14: Smoothed estimate of the intensity function $\hat{\pi}_1(x, y)$ in Equation (3.16) when the observations are reported partially grouped. The biweight function is used with combinations of significant smoothing for grouped data ($h_2 = 2$) with a sequence of smoothness for ungrouped data $h_1 = (.5, 1, 1.5, 2)$.

3.7 Weights Implementation

In this section we describe how to implement the weights used in the locally weighted likelihood analysis Equation (3.15). We are interested in estimating these weights for the region formed by Minnesota, North Dakota, Wisconsin, South Dakota and Iowa. We define a "plot" grid (x, y) over the 5 states region and estimate both the grouped and ungrouped weights at each of them. The weights used for grouped data are

$$w_j(x, y) = \frac{1}{|\Omega_j|} \iint_{\Omega_j} \mathcal{W}(x - u, y - v) du dv \tag{3.17}$$

and will be numerically approximated as follow:

- Generate a "working" grid over the county Ω_j lengthening the sides at both ends by δ. The estimations are done with $\delta = 1$ and it was chosen based on how far the fires that occurred in Ω_j can spread into adjacent regions. The grid is defined by two parameters east-west and north-south, $\delta_u = 3.4$ km and $\delta_v = 3.9$ respectively.
- The "plot grid" is defined by the parameters $\delta_x = 4.3$km (East-West) and $\delta_y = 4.9$km (North-South). The "working grid" is finer than the "plot" grid.
- For each (x, y) we estimate $\mathcal{W}(x - u_l, y - v_l)$ for (u_l, v_l) in the county of interest Ω_j.
- The weight $w_j(x, y)$ associated to position (x, y) that shows the effect of county Ω_j, is the sum over the set of pixels in the j-county.

$$w_j(x, y) = \sum_{l=1}^{N_j} W(x - u_l, y - v_l) \tag{3.18}$$

where N_j is the collection of pixels in the j-county.

Chapter 4

Generalized linear model for partially grouped data

In this chapter we propose a model that merges two levels of aggregate data (individual points and aggregate counts in areas) to forecast the risk of wildfire occurrences in the Continental US. The model approximates a partially grouped sample by a binary-valued process. The point observations $\tau_i = (x_i, y_i)$ are approximated by the Bernoulli distribution while the total number of events in non-overlapping regions, $N(\Omega_j)$, by the binomial distribution. We use a a generalized linear model with the logit link function to include explanatory variables that have been found to be of importance in the study of wildfire (fuels, season, topography, weather conditions) [6]. They can be used as lead variables for estimating the risk of wildfire occurrences. We use a locally weighted likelihood analysis to fit the model and to include the grouping effect.

At first we include a review of the wildfire occurrence data set and the notation that will be used in the rest of this chapter. In the second section it is explained how the partially grouped observations is approximated with a binary-valued process, assuming the wildfire point process is simple and the binary-valued random variables are equi-distributed. Finally, we implement a logit model to forecast the risk of wildfire occurrence at a given time including the fuel category as an explanatory variable.

4.1 Introduction

Our pilot study predicts wildfire risk for the year 1990 at 5 contiguous states: Minnesota (MN), North Dakota (ND), South Dakota (SD), Wisconsin (WI) and Iowa (IA). The general setting of our problem is provided as Figure 4.1. We define the sample space $\mathcal{S} \in \mathbb{R}^2$ as the region formed by the union of the 5 states just mentioned plus their adjacent states. The sample space $\mathcal{S}$ is divided into non-overlapping sets as in $\mathcal{S} = \mathcal{B} \cup \boldsymbol{\Omega}$. $\mathcal{B}$ includes the federal regions while $\mathbf{\Omega} = \left(\bigcup_{j=1}^{J} \Omega_j\right)$ include the non-federal regions. The Ω_j's stand for the counties which are also non-overlapping.

The data cover fires that occurred in both federal and non-federal lands. The federal data consist of fires' point locations (latitude and longitude) $\tau_i = (x_i, y_i) \in \mathbb{R}^2$ $i = 1, \ldots, I$ while the non-federal data are aggregated by county, j, and the number in Ω_j are represented by $N(\Omega_j) = n_j$ $j = 1, \ldots, J$.

In the figure, the fires locations are represented as dots, while the total counts per county are reported by the n_j. The set up with some data grouped and other ungrouped was named a partially grouped sample by Kulldorff [34]. It occurs when both type of data are present $I, J \geq 1$. A partially grouped sample refers to the case where available information is associated with a collection of disjoint sets partitioning a domain, $\mathcal{S}$. It is based on the choice for each set of whether its contents are points or a count. The set may be infinitesimally small as in the case of a point. While Kulldorff dealt with real-valued data, we will consider the points to lie in either space or space-time.

4.2 Binary-valued random variables

We model a partially grouped sample with a binary-valued process Y_l. To begin superimpose a lattice over the sample space $\mathcal{S}$. Let $N_{\mathcal{S}}$ denote the collection of pixels in $\mathcal{S}$. By superimposing a grid over the set $\mathcal{S}$ of concern, we can create a useful approximation of the original problem. When its location is known explicitly we assign a fire to its nearest lattice point, as shown at Figure 4.2.

The accuracy of this approximation depends on how fine a grid has been chosen, and just where the original points lie. The grid is defined by two parameters east-west and north-south, $\delta_u > 0$ and $\delta_v > 0$ respectively, see Figure 4.2. For estimation purposes, the grids' parameters are approximately $\delta_u = 3.4$ km and $\delta_v = 3.9$ km and they define the finest grid that let us use some efficient routines already implemented in the library(spatstat) in R.

Now let Y_l denotes the number of fires assigned to the l-th lattice point $l = 1, \ldots, N_{\mathcal{S}}$. In Figure 4.3 we display for Minnesota, a grid gray-scale rectangles also known as image where the lighter scale correspond to lattice pixels where at least one wildfire occurred. It is equivalent to being $Y_l > 0$ for the pixel defined by the l-th lattice point. One sees that the events are concentrated in northern Minnesota region, where the lighter gray pixels are located. For display purposes, this figure uses a coarser grid than the one used for carrying out the estimations.

If we assume that the wildfire point process is simple, that is the points are separated, then

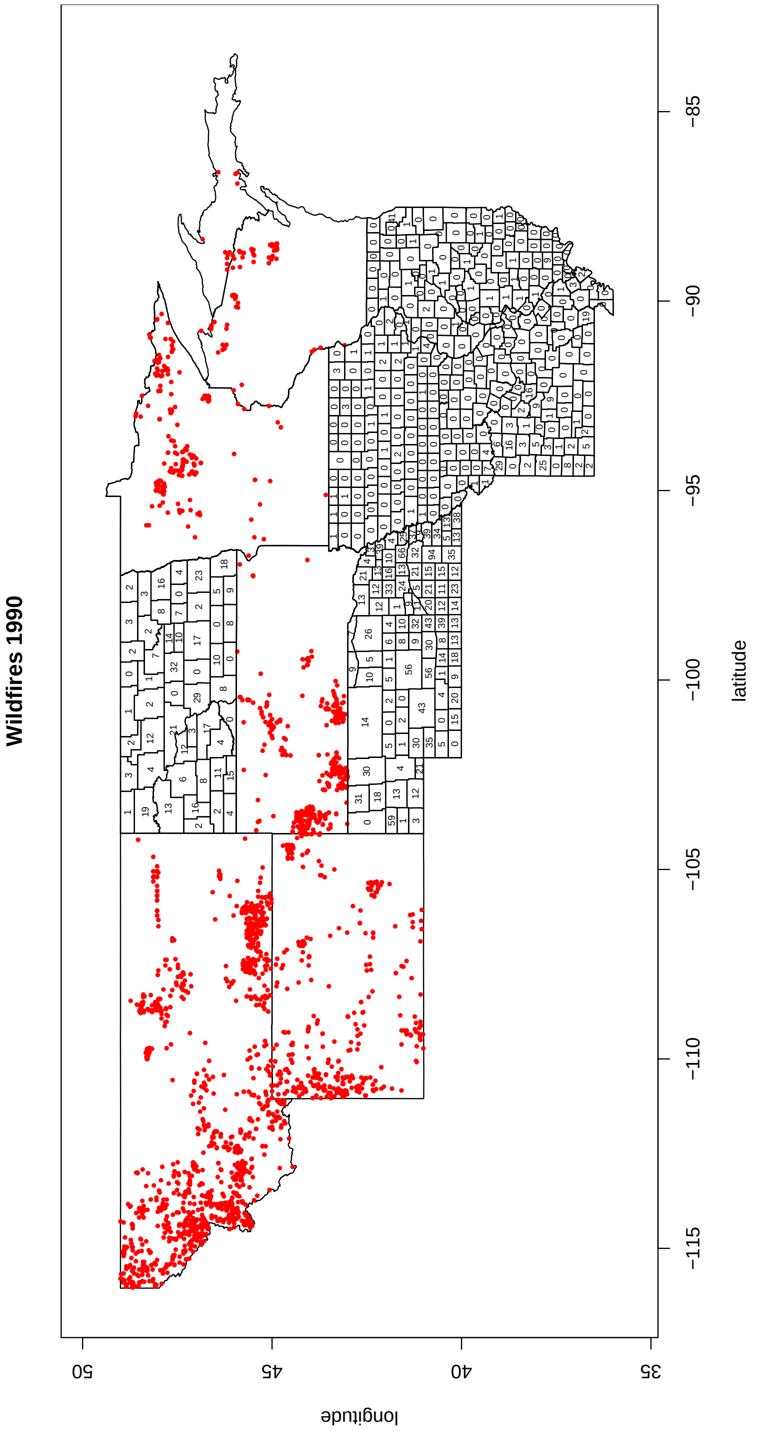


Figure 4.1: Wildfire occurrences recorded 1990 for the sample space $\mathcal{S}$. The data grouped into county corresponds to the states where where 1 is reported in the third column of table 1.1, $\mathbf{\Omega} = \left(\bigcup_{j=1}^{J} \Omega_j\right)$ with $J = 462$.

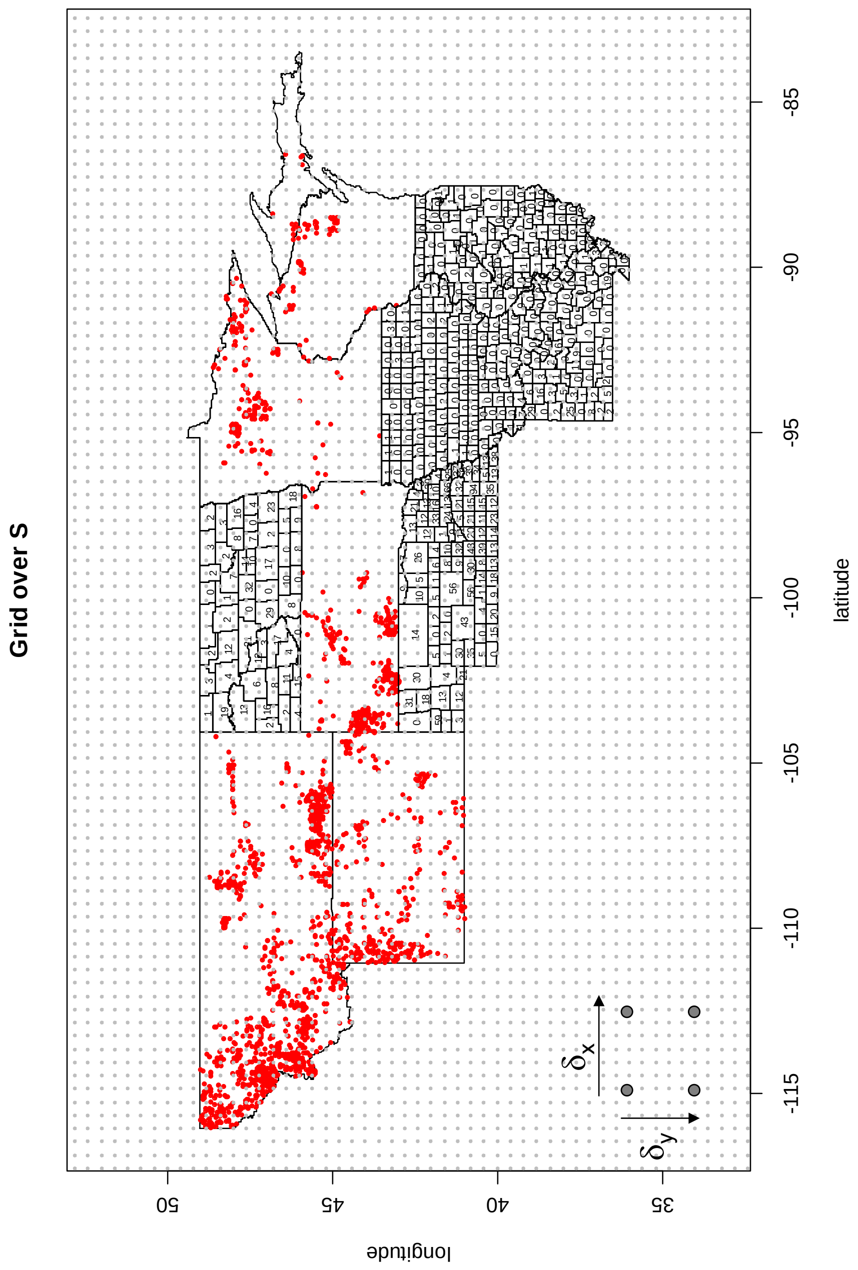


Figure 4.2: Lattice over the sample space $\mathcal{S}$ with parameters $\delta_u > 0$ and $\delta_v > 0$. The set of pixels in $\mathcal{S}$, $N_{\mathcal{S}}$, are plotted with light grey points. Each wildfire location is assigned to the closest lattice point.

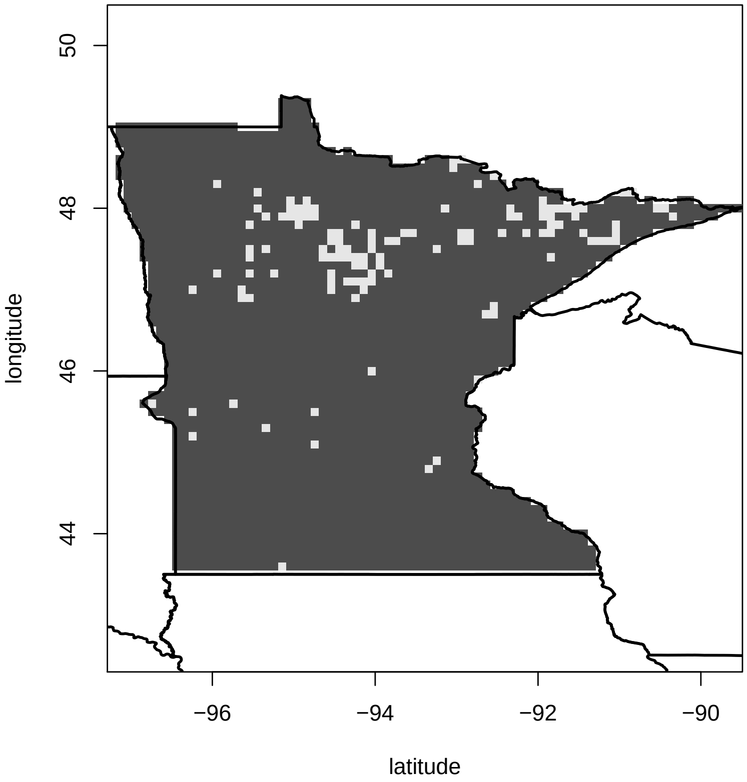


Figure 4.3: Grid gray-scale squares over Minnesota where the lighter scale correspond to lattice points with at least one wildfire assigned to these points, $Y_l > 0$.

no two points of the process are coincident. This going to a limit, the number of fires that can happen in each infinitesimal pixel is either 1 or 0. The Y_l are then binary random variables. We are following the idea of approximating a point process by a binary-valued process as proposed by Brillinger, Preisler and Benoit [5] for example.

Next we model the grouped data in region Ω_j with Binomial random variables, $N(\Omega_j) \sim Bin(N_j, \lambda_j)$ where N_j is the total number of grid points in Ω_j and n_j the correspondent number of pixels where a wildfire occurred. The actual positions of the wildfire occurrences within Ω_j are unknown and the binary random variables Y_l are unavailable but we can express the total number of fires in Ω_j as:

$$N(\Omega_j) = \sum_{l=1}^{N_j} Y_l = n_j \qquad j = 1, \ldots, J \tag{4.1}$$

We look to estimate for any point (x, y), the probability of occurrence of a wildfire near it. We are interested in estimating the expected value of the binary random variable Y at position (x, y), which will be name $Y_{x,y}$. Using Equation (4.1) the conditional expected value of $Y_{x,y}$ given $N(\Omega_j) = n_j$ may be evaluated as

$$\begin{aligned} E\left[Y_{x,y}\Big|\sum_{l=1}^{N_j} Y_l = n_j\right] &= Pr\left\{Y_{x,y} = 1\Big|\sum_{l=1}^{N_j} Y_l = n_j\right\} \qquad \text{since} \qquad Y_{x,y} = 0, 1 \\ &= \frac{n_j}{N_j} \end{aligned} \tag{4.2}$$

assuming equidistribution. The assumption of equidistribution implies that the pixels are of equal areas, δ_u and δ_v are fixed, this can be relaxed later. In Equation (4.2) we can see that the expected value of Y at position (x, y) is proportional to the expected number of events in Ω_j.

We further consider that the occurrence probability of a fire is a smoothly varying function of space. To estimate it, we introduce a locally weighted estimate of $Y_{x,y}$ namely:

$$\sum_{l=1}^{N_j} \mathcal{W}(x - u_l, y - v_l) Y_l \tag{4.3}$$

where (u_l, v_l) are the grid points. $\mathcal{W}$ is an appropriate weight function. If we assume equidistribution, the weighted conditional expected value of $Y_{x,y}$ given $N(\Omega_j) = n_j$ is:

$$\begin{aligned} E\left[\sum_{l=1}^{N_j} \mathcal{W}(x-u_l, y-v_l) Y_l \Big| \sum_{l=1}^{N_j} Y_l = n_j\right] &= \sum_{l=1}^{N_j} \mathcal{W}(x-u_l, y-v_l) E\left[Y_l \Big| \sum_{l=1}^{N_j} Y_l = n_j\right] \\ &= \frac{n_j}{N_j} \sum_{l=1}^{N_j} \mathcal{W}(x-u_l, y-v_l) \qquad (4.4) \end{aligned}$$

This is seen to be a smoothed estimator of Equation (4.2).

When the data are grouped into county, the weights used in the estimate Equation (4.4), are meant to take into account the effect of the county division j on the location (x, y). For the analysis of the wildfire occurrences, the county division is in many cases an artificial boundary that does not define the limits of where the conditions actually change. We define $\mathcal{W}$ to be the biweight function in Equation (3.2). It will provide a smoothing effect of the fires happened in the region j on the location (x, y).

In Figure 4.4 we show the behaviour of the selected weights for grouped data:

$$w_j(x, y) = \frac{1}{|\Omega_j|} \iint\limits_{\Omega_j} \mathcal{W}(x-u, y-v) du dv \qquad (4.5)$$

where $\mathcal{W}$ is the biweight function. At the top, the weight function for the interval $[-40, 40]$ is shown. The equidistributed bumps that form the weight function as explained in section 3.7 are also plot. In the bottom of that same figure, we can see the weight function for the county McLean in North Dakota.

4.3 Generalized linear models for binary data

In order to extend our model to include covariates or explanatory variables we use the structure of the generalized linear models (g.l.m.). In this section we give an overview of g.l.m. for binary data but a complete reference can be found in [38]. The g.l.m. extend the

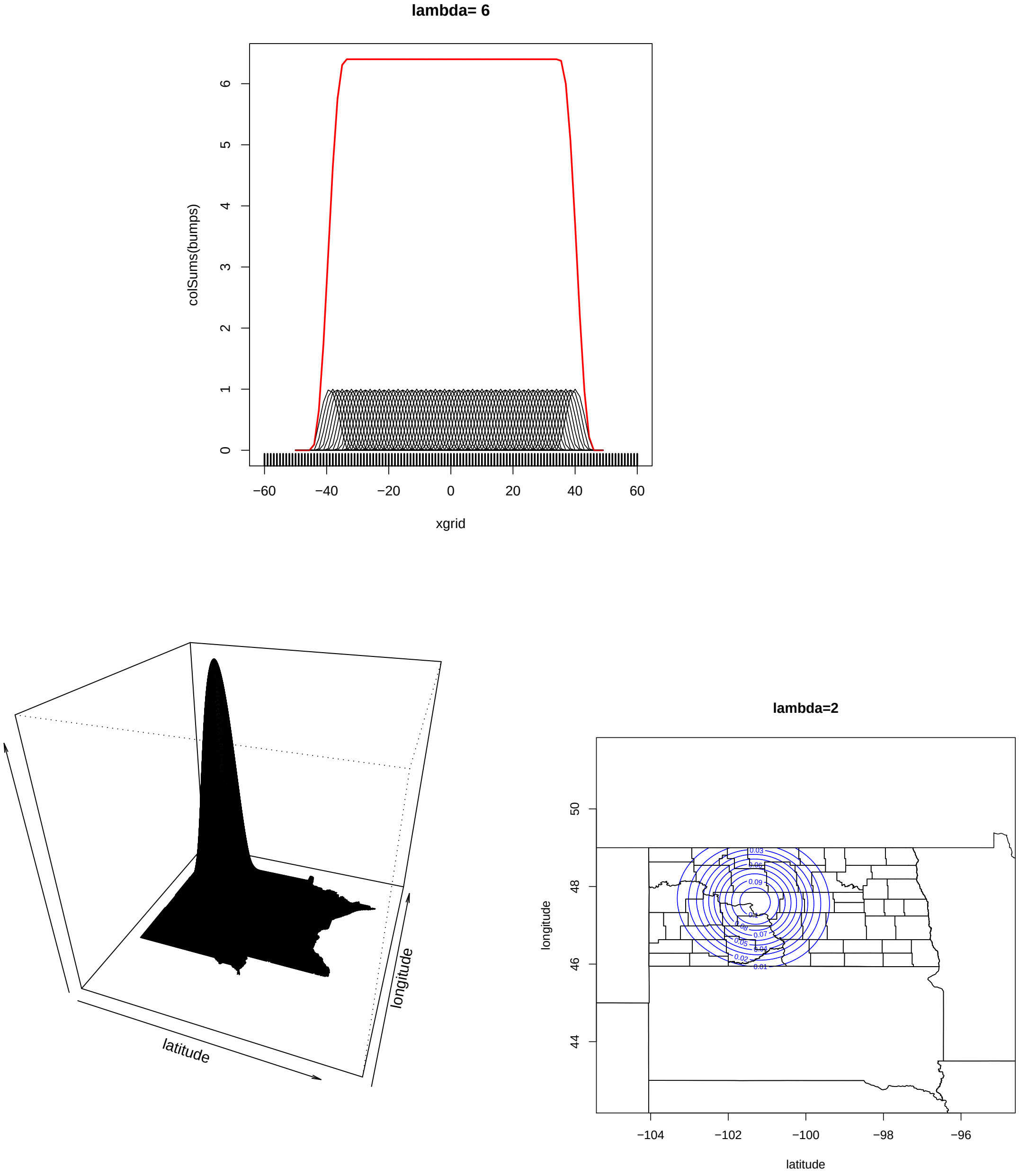


Figure 4.4: Top: weight function for the interval $[-40, 40]$ using the one variable case of Equation (3.4) with bandwidth $h = 6$. Bottom: Perspective and contour plot for the weight function for McLean North Dakota with bandwidth $h = 2$ using Equation (3.4).

linear model to include non-normal distributions. The covariates $\mathbf{x_1}, \ldots, \mathbf{x_p}$ produce a linear predictor $\boldsymbol{\eta}$ given by

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

with unknown coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and a design matrix $\mathbf{X}$. The link between the random components and covariates can be written as

$$\boldsymbol{\eta} = g\left(\boldsymbol{\mu}\right) \qquad \boldsymbol{\mu} = g^{-1}\left(\boldsymbol{\eta}\right)$$

where $\boldsymbol{\mu}$ is the mean of the random component of the g.l.m. and $g(\cdot)$ is the link function. The link function may be any monotone differentiable function.

Consider the J independent Binomial random variables $N(\Omega_1), \ldots, N(\Omega_J)$. For g.l.m. the relationship between the vector $\boldsymbol{\mu}$ and the covariates as summarized by a design matrix $\mathbf{X}$ of order $J \times p$, that leads to the following form:

$$\eta_j = g(\mu_j) = \sum_{r=1}^{p} x_{jr}\beta_r \qquad j = 1, \ldots, J \tag{4.6}$$

If $N(\Omega_j) \sim Bin(N_j, \rho_j)$ and the link function is the logit, we have:

$$\eta_j = \log\left(\frac{\rho_j}{1-\rho_j}\right) = \sum_{r=1}^{p} x_{jr}\beta_r \qquad j = 1, \ldots, J \tag{4.7}$$

The weighted log-likelihood function is now

$$\ell_w(\boldsymbol{\rho}) \;\; = \;\; \sum_{j=1}^{J} w_j(x, y)\left[n_j \log\left(\rho_j\right) + \left(N_j - n_j\right)\log\left(1 - \rho_j\right)\right] \tag{4.8}$$

Substituting the logit link in the weighted log-likelihood function, one can write:

$$\begin{aligned}
\ell_w(\boldsymbol{\rho}) &= \sum_{j=1}^{J} w_j(x,y)\left[n_j \log\left(\frac{\rho_j}{1-\rho_j}\right) + N_j \log(1-\rho_j)\right] && (4.9)\\
&\textit{and} \\
\ell_w(\boldsymbol{\beta}) &= \sum_{j=1}^{J} w_j(x,y)\left[n_j \sum_{r=1}^{p} x_{jr}\beta_r - N_j \log\left(1+\exp\sum_{r=1}^{p} x_{jr}\beta_r\right)\right] && (4.10)
\end{aligned}$$

where $N(\Omega_j) = n_j$. In the following subsection we describe a parameter estimation procedure for this model.

4.3.1 Parameter estimation

The maximum likelihood estimates of the parameters $\boldsymbol{\beta}$ that appear in Equation (4.9) can be obtained by iterative weighted least squares (IRWLS). In order to establish the equations needed for using IRWLS we need to derive the weighted log-likelihood equations for the parameter $\boldsymbol{\beta}$. First the derivative of the Equation (4.8) to ρ_j is

$$\frac{\partial \ell_w(\boldsymbol{\rho})}{\partial \rho_j} = \sum_{j=1}^{J} w_j(x,y)\frac{n_j - N_j\rho_j}{\rho_j(1-\rho_j)}$$

Using the chain rule, the derivative with respect to β_r is

$$\frac{\partial \ell_w(\boldsymbol{\beta})}{\partial \beta_r} = \sum_{j=1}^{J} w_j(x,y)\frac{n_j - N_j\rho_j}{\rho_j(1-\rho_j)}\frac{\partial \rho_j}{\partial \beta_r}$$

For convenience express $\partial\rho_j/\partial\beta_r$ as a product:

$$\frac{\partial \rho_j}{\partial \beta_r} = \frac{\partial \rho_j}{\partial \eta_j}\frac{\partial \eta_j}{\partial \beta_r} = \frac{\partial \rho_j}{\partial \eta_j}x_{jr}$$

Thus the derivative with respect to β_r or the score function is

$$U_j = \frac{\partial \ell_w(\boldsymbol{\beta})}{\partial \beta_r} = \sum_{j=1}^{J} w_j(x,y) \frac{n_j - N_j \rho_j}{\rho_j(1-\rho_j)} \frac{\partial \rho_j}{\partial \eta_j} x_{jr} \tag{4.11}$$

The variance-covariance matrix of the U_j's has terms

$$\mathbf{I}(\boldsymbol{\beta})_{rs} = E\left(U_r U_s\right)$$

which form the information matrix or Fisher information $\mathbf{I}$. Using the Equation 4.11

$$\begin{aligned}
\mathbf{I}(\boldsymbol{\beta})_{rs} &= \sum_{j=1}^{J} w_j(x,y) \frac{N_j}{\rho_j(1-\rho_j)} \frac{\partial \rho_j}{\partial \beta_r} \frac{\partial \rho_j}{\partial \beta_s} \\
&= \sum_{j=1}^{J} w_j(x,y) N_j \left(\frac{\partial \rho_j}{\partial \eta_j}\right)^2 \frac{1}{\rho_j(1-\rho_j)} x_{jr} x_{js} \\
&= \{\mathbf{X}^T \mathbf{W} \mathbf{X}\}_{rs}
\end{aligned}$$

where $\mathbf{W}$ is a diagonal matrix of entries given by

$$\mathbf{W} = \text{diag}\left\{ w_j(x,y) N_j \left(\frac{\partial \rho_j}{\partial \eta_j}\right)^2 \frac{1}{\rho_j(1-\rho_j)} \right\}$$

The estimating equation for the method of scoring generalizes to

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + \left[\mathbf{I}^{(m-1)}\right]^{-1} \mathbf{U}^{(m-1)} \tag{4.12}$$

where $\boldsymbol{\beta}^{(m)}$ is the vector of estimates of parameters $\beta_1, \ldots, \beta_p$ at the *mth* iteration. In Equation 4.12, $\left[\mathbf{I}^{(m-1)}\right]^{-1}$ is the inverse of the information matrix with elements given by 4.12 and $\mathbf{U}^{(m-1)}$ is the vector of elements given by Equation 4.11, all evaluated at $\boldsymbol{\beta}^{(m-1)}$. If both sides of equation are multiplied by $\mathbf{I}^{(m-1)}$ it is obtained

$$\mathbf{I}^{(m-1)}\boldsymbol{\beta}^{(m)} = \mathbf{I}^{(m-1)}\boldsymbol{\beta}^{(m-1)} + \mathbf{U}^{(m-1)} \tag{4.13}$$

The right-hand side of Equation 4.13 can be written as

$$\mathbf{X}^T\mathbf{W}\mathbf{z} \tag{4.14}$$

where $\mathbf{z}$ has elements

$$z_j = \sum_{r=1}^{p} x_{jr}\beta_r^{(m-1)} + \left(\frac{n_j}{N_j} - \hat{\rho}_j\right)\frac{\partial \eta_j}{\partial \rho_j}$$

all quantities being are evaluated at $\hat{\boldsymbol{\beta}}^{(m-1)}$.

Hence the iterative equation 4.13, can be written as

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}^{(m)} = \mathbf{X}^T\mathbf{W}\mathbf{z} \tag{4.15}$$

and we begin using some initial approximations $\boldsymbol{\beta}^{(0)}$.

The Equation 4.15 is in the same form as the normal equations for a linear model obtained by weighted least squares, except that it has to be solved iteratively because, in general, $\mathbf{z}$ and $\mathbf{W}$ depend on $\boldsymbol{\beta}$. Thus for g.l.m., maximum likelihood estimators are obtained by an iterative weighted least squares procedure.

The weighted maximum likelihood estimates can be obtained by including the pertinent weights in the R function glm(). Regularly the glm() function uses the following weights:

$$w_0{}^{-1} = \left(\left. \frac{\partial \eta_j}{\partial \rho_j}^2 \right|_{\hat{\eta}_0} V_0 \right)$$

where V_0 is the variance of the proportions. In particular, to include the grouping effect in our estimators we will instead use the following weights w_0' in the glm() function:

$$w_0'^{-1} = w_j(x,y)^{-1} \left(\left. \frac{\partial \eta_j}{\partial \rho_j}^2 \right|_{\hat{\eta}_0} V_0 \right)$$

4.4 Logit model for partially grouped data

Under the assumption of a fine grid, i.e. small δ_u, δ_v, and a point process $\tau_i = (x_i, y_i)$ with isolated points, the events in $\mathcal{B}$ can be approximated by Bernoulli random variables Y_l $l = 1, \ldots, N_{\mathcal{B}}$. Then $Y_l \sim Bin(1, \lambda_l)$, where λ_l is the probability of the occurrence of a fire in the pixel corresponding to the grid point (u_l, v_l). In the same way, the events grouped into county has the binomial distribution $N(\Omega_j) \sim Bin(N_j, \rho_j)$ $j = 1, \ldots, J$ where ρ_j is the probability of occurrence of a fire in a pixel of the collection of pixels in Ω_j. The random variables Y_l and $N(\Omega_j)$ are independent, conditional on the covariates, the grouping effect and the influence of neighbour wildfire occurrences. The latter is included in the model by appropriate weights in the likelihood function.

The locally weighted log-likelihood function combining the grouped and ungrouped data is:

$$\begin{aligned} \ell_w(\boldsymbol{\lambda}, \boldsymbol{\rho}) &= \sum_{l=1}^{N_{\mathcal{B}}} \mathcal{W}(x - u_l, y - v_l) \left[Y_l \log(\lambda_l) + (1 - Y_l) \log(1 - \lambda_l) \right] \\ &+ \sum_{j=1}^{J} w_j(x, y) \left[n_j \log(\rho_j) + (N_j - n_j) \log(1 - \rho_j) \right] \end{aligned}$$

In the federal regions, the explanatories $\mathbf{x_1}, \ldots, \mathbf{x_p}$ for the point observations are provided in the basic entity, points. The fires in non-federal regions are reported grouped into county and the respective explanatories are used at the county level. The explanatories for the county level are called $\mathbf{x'_1}, \ldots, \mathbf{x'_{p'}}$. Therefore, we define the relationship between the probabilities and the covariates for the two levels of aggregation, points Y_l and counts $N(\Omega_j)$:

$$\eta_l = \log\left(\frac{\lambda_l}{1 - \lambda_l}\right) = \sum_{r=1}^{p} x_{lr} \beta_r \qquad l = 1, \ldots, N_{\mathcal{B}} \tag{4.16}$$

$$\eta_j = \log\left(\frac{\rho_j}{1 - \rho_j}\right) = \sum_{r=1}^{p} x'_{jr} \beta_r \qquad j = 1, \ldots, J \tag{4.17}$$

where the link is the logit function.

Substituting into the locally weighted log-likelihood function Equation (3.14)

$$\begin{aligned}
\ell_w\left(\boldsymbol{\lambda},\boldsymbol{\rho}\right) &= \sum_{l=1}^{N_{\mathcal{B}}}\mathcal{W}(x-u_l,y-v_l)\left[Y_l\log\left(\frac{\lambda_l}{1-\lambda_l}\right)+Y_l\log\left(1-\lambda_l\right)\right] \\
&+ \sum_{j=1}^{J}w_j(x,y)\left[n_j\log\left(\frac{\rho_j}{1-\rho_j}\right)+N_j\log\left(1-\rho_j\right)\right] \qquad (4.18)\\
\textit{and} \\
\ell_w\left(\boldsymbol{\beta}\right) &= \sum_{l=1}^{N_{\mathcal{B}}}\mathcal{W}(x-u_l,y-v_l)\left[Y_l\sum_{r=1}^{p}x_{lr}\beta_r-Y_l\log\left(1+\exp\sum_{r=1}^{p}x_{lr}\beta_r\right)\right] \\
&+ \sum_{j=1}^{J}w_j(x,y)\left[n_j\sum_{r=1}^{p}x'_{jr}\beta_r-N_j\log\left(1+\exp\sum_{r=1}^{p}x'_{jr}\beta_r\right)\right] \qquad (4.19)
\end{aligned}$$

Using IRWLS, as described in the Subsection 4.3.1, it can be maximized for

$$\hat{\boldsymbol{\beta}}=(\hat{\beta}_1(x,y),\ldots,\hat{\beta}_p(x,y))$$

.

4.4.1 Smoother based on a logit model

In this subsection we introduce an elementary logit model with only a constant as an exploratory variable. The linear predictor follows:

$$\eta_l=\eta_j=\beta_1 \qquad (4.20)$$

assuming that the probability of a wildfire occurrence in the sample space is constant. This assumption is relaxed when we used a weighted likelihood approach so we are assuming the probability of a wildfire occurrence is constant in the support of the weight function. Substituting the linear predictor Equation (4.20) in the weighted log-likelihood function Equation (4.18), we can predict the probability of a wildfire occurrence in the proximity of the position (x,y). Under this assumption the predicted probability is named $\hat{\pi}_2(x,y)$

$$\hat{\pi}_2(x,y)=\frac{\exp\hat{\beta}_1}{1+\exp\hat{\beta}_1} \qquad (4.21)$$

where $\hat{\beta}_1 = \hat{\beta}_1(x, y)$ and

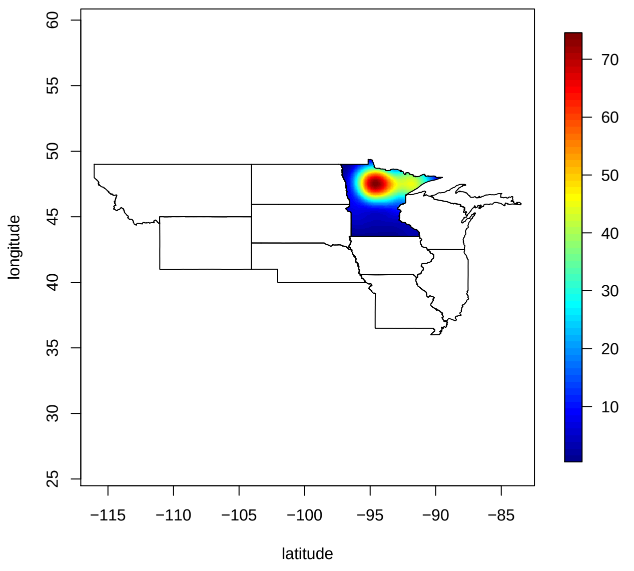
$$\hat{\beta}_1(x, y) = \log \left(\frac{\sum_{l=1}^{N_{\mathcal{B}}} \mathcal{W}(x - u_l, y - v_l) Y_l + \sum_{j=1}^{J} w_j(x, y) n_j}{\sum_{j=1}^{J} w_j(x, y) \left[N_j - n_j \right]} \right)$$

Either maximizing the weighted log-likelihood function or using the glm() function we obtain the same expression. The estimates in Equation (4.21) are plotted in the bottom Figure 4.6. These estimates combine two different weight functions, one for the ungrouped data and other for the grouped data. Figure 4.5 shows both weight functions using the biweight with two different bandwidths, h_1 for the ungrouped and h_2 for the grouped data. The federal and non-federal regions differ in the degree of smoothing that is applied.

The smoother used in the previous chapter in Equation (3.9) can be compared with the estimate in this section, Equation (4.21). The difference is that the first neglected the point observations and considered both the federal and non-federal fires grouped into county. It assumed that the total number of fires per county had a Poisson distribution. The second uses a binary-valued process to integrate the two levels of aggregate data, points and counts. In the limit, both cases consider that the total number of fires by county has a Poisson distribution with expected values $\left(\sum_{(x,y) \in \Omega_j} \pi_1(x, y) \right) |\Omega_j|$ and $\left(\sum_{(x,y) \in \Omega_j} \pi_2(x, y) \right) |\Omega_j|$ respectively.

Figure 4.6 compares both smoothers, the first two plots show the estimate with data grouped into county for two different bandwidths. The third shows the estimate with the logit model. A clear difference between the two smoothers is that using data grouped into county show a relationship between the regions of high fire probability in North Dakota and South Dakota. This relationship is not detected with the second smoother, in the contrary it detects two independent regions with high probability, one in central North Dakota and other in northwest South-Dakota. The first smoother detects an area with high probability of wildfire occurrence in Wisconsin but this does not happen with the smoother based on the logit model.

Some of the explanatory variables that have been used in the modeling of wildfire occurrences [44] are: fuel category, elevation, wind speed, state of weather, topography, vegetation. Including some explanatories in the model that can be reasonably forecasted will provide us



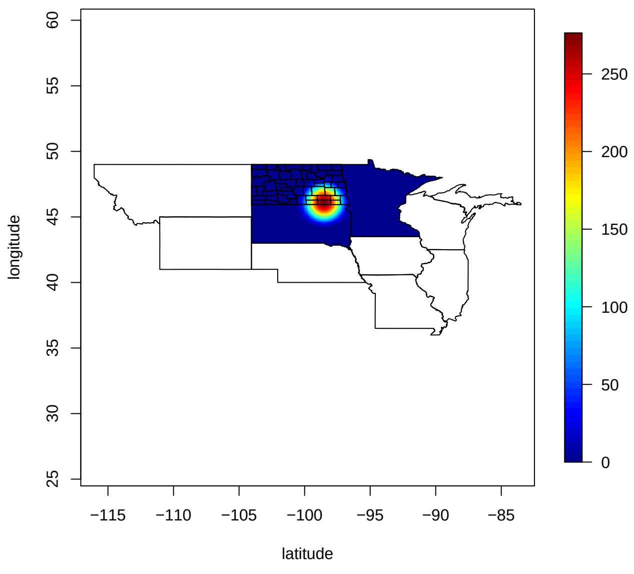


Figure 4.5: Weight function associated to the point (x, y) in the log-likelihood function. Top: the weight function for ungrouped data is based on the binary process $\sum_{l=1} \mathcal{W}(x-u_l, y-v_l)Y_l$, where the l index runs over the pixels in MN. Bottom: weight function used for grouped data $w_j = \iint_{\Omega_j} \mathcal{W}(x-u, y-v)dudv$ in particular Ω_j is Dickey county, North Dakota. In both cases $\mathcal{W}$ is the biweight function with bandwidths $h_1 = 2$ and $h_2 = 1.5$ respectively.

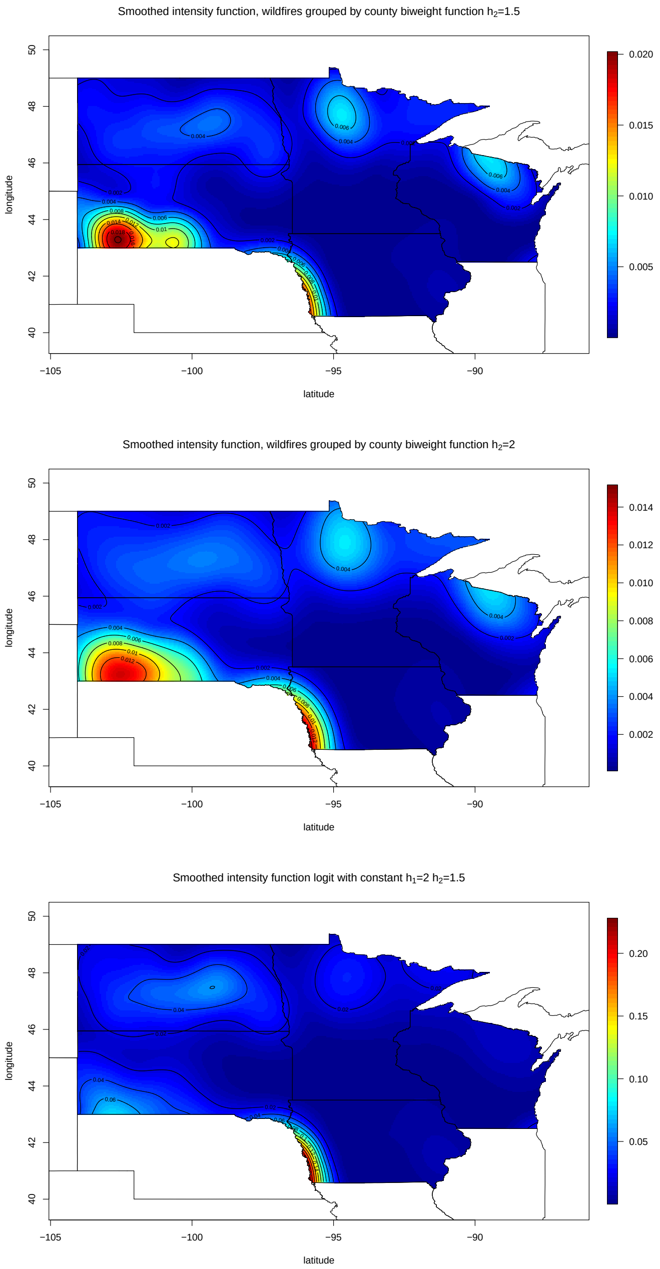

Smoothed intensity function, wildfires grouped by county biweight function h_2=1.5
longitude
latitude
50
48
46
44
42
40
−105
−100
−95
−90
0.020
0.015
0.010
0.005
Smoothed intensity function, wildfires grouped by county biweight function h_2=2
longitude
latitude
0.014
0.012
0.010
0.008
0.006
0.004
0.002
Smoothed intensity function logit with constant h_1=2 h_2=1.5
longitude
latitude
0.20
0.15
0.10
0.05

Figure 4.6: Smoothed estimate of the intensity function with data grouped into county. The total number of fires in a county is assumed to have a Poisson distribution. A weighted likelihood fit with the biweight function using the bandwidths a) top: $h_2 = 1.5$ b) middle: $h_2 = 2$. The bottom provides a smoothed estimate of the probability of occurrence of a wildfire. This is the partially grouped sample, the logit link with only a constant β_1 included in the linear predictor. A weighted likelihood fitting procedure is used with the biweight function and the two bandwidths $h_1 = 2$ and $h_2 = 1.5$.

with useful predictions. A first analysis to follow uses the fuel conditions as the explanatory variable.

4.5 Fuel model as explanatory variable

The National Fire Danger Rating System (NFDRS) provides the fuel conditions over the Continental US. It is estimated for a 1 km grid by Burgan et al. [6]. The NFDRS Fuel Map is Figure 4.7 and the characteristics for each fuel model are provided in Table 4.1.

National Fire Danger Rating fuel models [51] have been mapped across the lower 48 states at 1 km resolution [6]. The map was derived from a combination of satellite imagery used to create a land cover database for the conterminous U.S., and ground data sampled from across the U.S.

To construct the fuel model map, ground sample data was obtained from 2560 plots of 1 km^2 scattered randomly across the U.S., and an NFDR fuel model was assigned to each plot. This data, combined with the landcover class data and an Omernick Ecoregion map, permitted counting the number of times a fuel model was assigned to each sampled landcover class within each ecoregion. The preliminary map was then revised by having at least one person from each Forest Service Region go to the Fire Sciences Laboratory to review the map. This person was well versed in the location and extent of vegetation in the region, and what NFDR fuel models best represent it [51].

4.5.1 Model implementation

We are interested in predicting the probability of a wildfire occurrence in a pilot group $s = \{MN, ND, WI, SD, IA\}$. Our sample space, $\mathcal{S}$, includes these states plus their neighbours. The information from the adjacent regions is included in the prediction of the probability of a wildfire occurrence at (x, y) through a weighted likelihood analysis. The weighted likelihood fit uses weight functions that assign more influence to those observations closer to (x, y) and less influence to those further from it. Therefore, to predict the probability of wildfire occurrence for a particular state s, the logit model just includes the information from the neighbour states as the weights for the observations further away can be neglected.

The linear predictors for the logit link function based on Equation 4.16 including the fuel categories are:

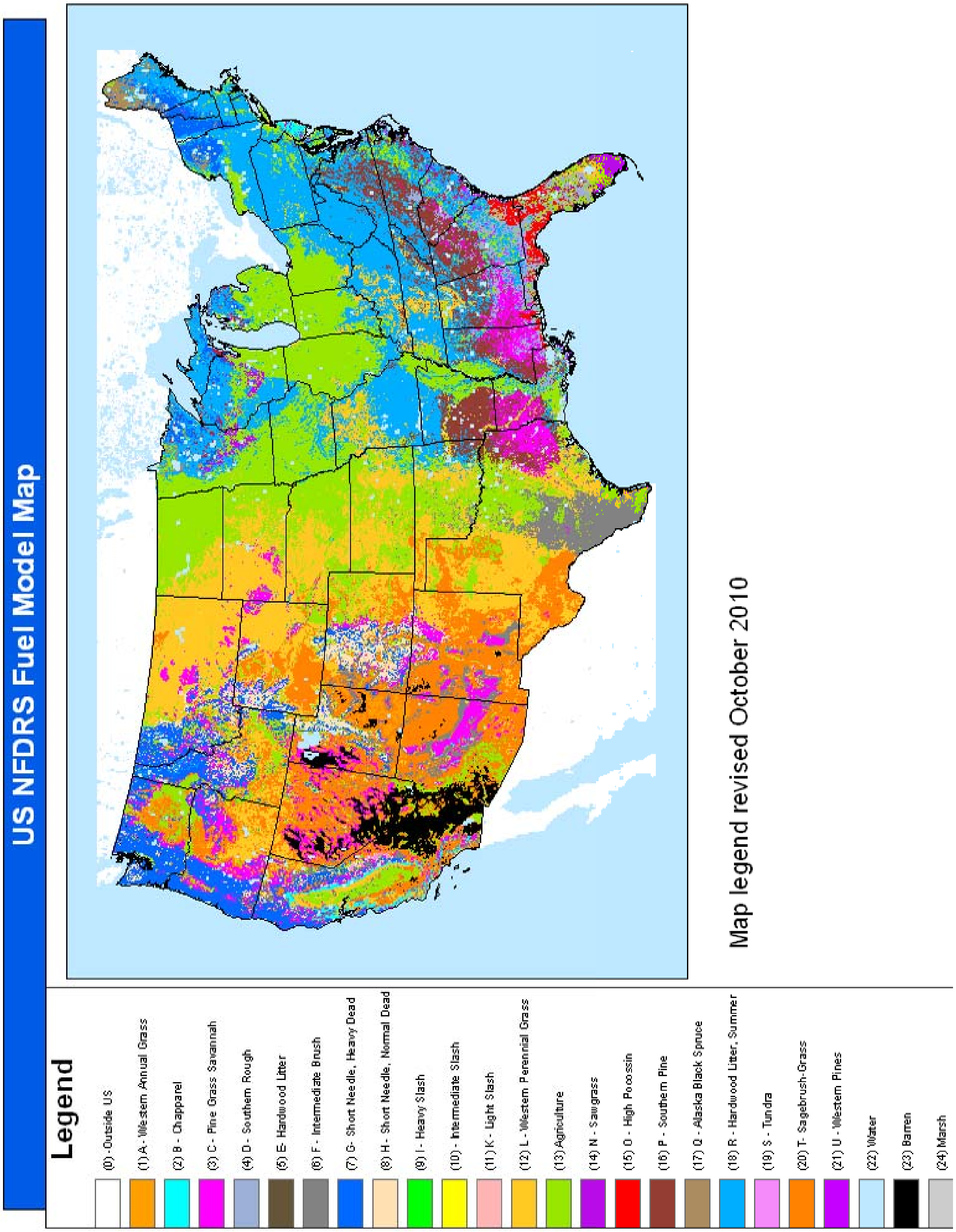


Figure 4.7: US NFDRS Fuel Model Map

NFDR Fuel	Fuel Load (T/Hectare)		Extinction	
Model	Live	Dead	Moisture (%)	Vegetation Represented
A	0.67	0.45	15	Western annual grasses
B	25.78	17.93	15	California mixed chaparral
C	2.91	3.14	20	Pine grass savanna
D	8.41	6.73	30	Southern rough
E	——	——	——	Hardwoods (winter)
F	20.18	13.45	15	Intermediate brush
G	29.14	21.30	25	Short needle conifers with heavy dead load
H	67.3	10.09	20	Short needle conifers with normal dead load
I	——	——	——	Heavy logging slash
J	——	——	——	Intermediate logging slash
K	——	——	——	Light logging slash
L	1.12	0.56	15	Western perennial grasses
M	——	——	——	Agricultural land
N	4.48	6.73	25	Saw grass or other thick stemmed grasses
O	20.18	17.93	30	High pocosin
P	3.36	4.48	30	Southern pine plantation
Q	12.33	14.57	25	Alaskan black spruce
R	2.24	3.36	25	Hardwoods (summer)
S	3.36	3.36	25	Alpine tundra
T	6.73	3.36	15	Sagebrush-grass mixture
U	2.24	7.85	20	Western long-needle conifer
V	——	——	——	Water
W	——	——	——	Barren
X	——	——	——	Marsh

Table 4.1: Fuel loadings and extinction moisture used in calculating the Fire Potential Index.

$$\eta_l = \beta_1^{(s)} + \sum_{k=2}^{K} I_{lk}^{(s)} \beta_k^{(s)} \qquad l = 1, \ldots, N_{\mathcal{B}} \tag{4.22}$$

$$\eta_j = \beta_1^{(s)} + \sum_{k=2}^{K} I_{jk}^{(s)'} \beta_k^{(s)} \qquad j = 1, \ldots, J \tag{4.23}$$

where $(x, y) \in s$, $\eta = \eta(x, y)$ and $\boldsymbol{\beta} = \boldsymbol{\beta}(x, y)$. As the fuel model is a categorical explanatory variable, $I_k^{(s)}$ above is an indicator function for the k fuel category, $I_k^{(s)} = 1$ if the k category is present in the s state or its neighbours. Table 4.5.1 shows the fuel categories that covers each state and their neighbours.

For each (x, y) the model is defined by the following matrix expression

NFDR Fuel Model	MN	ND	WI	SD	IA
A					
B					
C	+	+		+	+
D					
E					
F				+	
G	+	+	+	+	+
H	+	+		+	+
I					
J					
K					
L	+	+	+	+	+
M	+	+	+	+	+
N					
O	+	+	+	+	+
P					
Q	+	+	+	+	+
R	+	+	+	+	+
S		+		+	
T	+	+		+	+
U	+	+	+	+	+
V	+	+	+	+	+
W		+		+	
X					

Table 4.2: NFDRS fuel categories that covered MN, ND, WI, SD and IA and their respective neighbours. The fuel categories are described in Table 4.1.

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_{N_{\mathcal{B}}} \\ n_1 \\ \dots \\ n_J \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & I_{11} & \cdots & I_{1K} \\ \dots & \dots & \dots & \dots \\ 1 & I_{N_{\mathcal{B}}1} & \cdots & I_{N_{\mathcal{B}}K} \\ 1 & I'_{11} & \cdots & I'_{1K} \\ \dots & \dots & \dots & \dots \\ 1 & I'_{J1} & \cdots & I'_{1K} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \dots \\ \beta_K \end{pmatrix}$$

for simplicity the index (s) is omitted in the indicator functions and $\mathbf{X}$ is the design matrix. The indicator functions for the federal regions is $I_{lk} = 1$ when the k fuel category covers the pixel correspondent to the l pixel and zero otherwise. The indicator function for the j county is $I'_{Jk} = 1$ when the model occupies the larger area of the county's territory and zero otherwise. This assumption seems viable as the NFDR map shows that the majority of the county's territory is covered by one or two fuel categories.

The probability of a wildfire occurrence is estimated for dots of a grid (x, y) covering the five states of interest. The estimated probability with, fuel category as an explanatory variable, is denoted $\hat{\pi}_3$:

$$\hat{\pi}_3(x, y) \quad = \quad \frac{\exp\{\hat{\beta}_1^{(s)} + \sum_{k=2}^{K} I_k \hat{\beta}_k^{(s)}\}}{1 + \exp\{\hat{\beta}_1^{(s)} + \sum_{k=2}^{K} I_k \hat{\beta}_k^{(s)}\}} \qquad (x, y) \in s \tag{4.24}$$

where $\hat{\boldsymbol{\beta}}^{(s)} = (\hat{\beta}_1^{(s)}(x, y), \dots, \hat{\beta}_K^{(s)}(x, y))$ and $I_k = I_k(x, y)$. The indicator functions $I_k(x, y) = 1$ when the k fuel category covers the pixel corresponding to the grid point (x, y). We omit the first fuel category $(k = 1)$ to avoid that the design matrix will not have full rank and the parameters are not identifiable. For each state, as we omit the first fuel category the effect of the other fuel categories is measured relative to the omitted category.

The map of estimated probabilities is Figure 4.8. One sees: a) In North Dakota there is a strip with a higher probability of wildfire occurrence. This region is located closed to the Lake Sakakawea. b) In the southwestern region of South Dakota there is a higher probability of fire and it seems to be located in the Black Hills National Forest. The causes of the high risk region in the north-central need to be explored. c) In Minnesota, the regions with higher probability can be associated with some lakes. d) In Iowa the western border seems to experience a higher risk of fire. It could be associated to the the Missouri River or a highway that runs parallel to the border. The road may be from human caused fires.

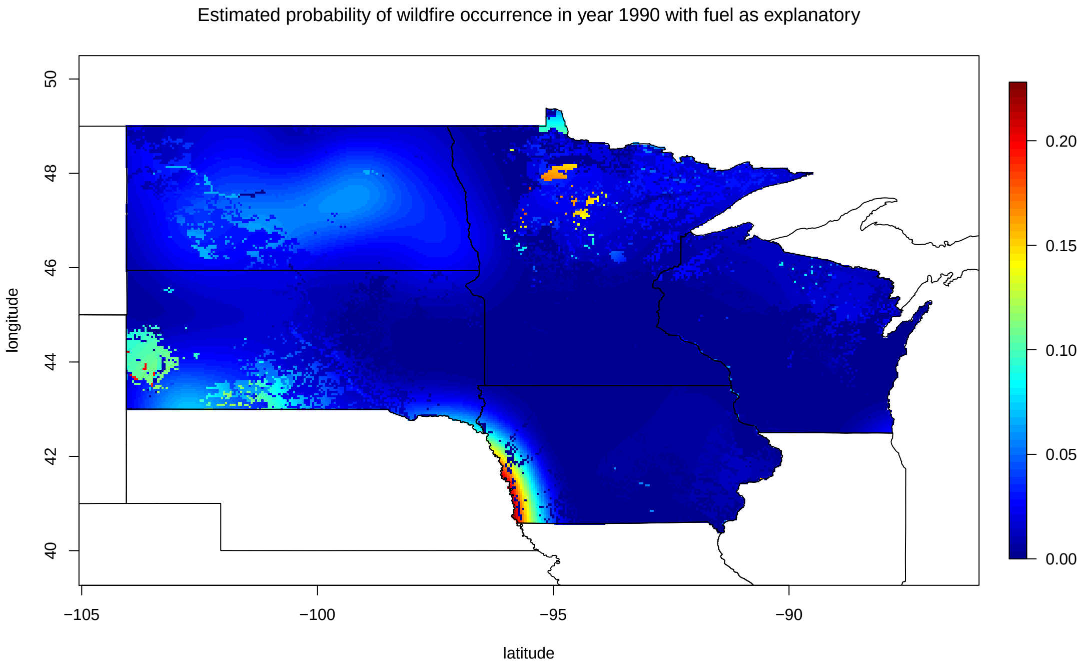

Estimated probability of wildfire occurrence in year 1990 with fuel as explanatory
longitude
latitude
50
48
46
44
42
40
−105
−100
−95
−90
0.20
0.15
0.10
0.05
0.00

Figure 4.8: Probability map using the estimator in Equation 4.18 including the NFDRS fuel model as explanatory. The biweight function is used and two different bandwidths $h_1 = 2$ and $h_2 = 1.5$ for ungrouped and grouped data respectively.

4.6 Model Assessment

One way to assess a model is through the analysis of the residuals. As the observations in non-federal regions are not provided in the basic unit, points, we need to form residuals with the data grouped into county. Brillinger and Preisler [5] use Le Cam and Hodges' Poisson Binomial distribution as an approximation to the distribution of the total number of wildfire occurrences.

Le Cam and Hodges [29] begin by considering a large number, say n, of events that might occur. The chance, say p, that any specified one of these events will occur, is small. Assuming that the events are independent, N has exactly the binomial distribution, say $Bin(n, p)$. If we now let $n \to \infty$ and $p \to 0$ so that $np \to \lambda$ where λ is fixed and $0 < \lambda < \infty$, it is shown that $Bin(n, p)$ tends to the Poisson distribution $P(\lambda)$ with expectation λ. They consider this explanation as often not satisfactory because the various trials cannot in many applications reasonably be regarded as equally likely to succeed. Let p_i denote the success probability of the ith trial, $i = 1, 2, \ldots, n$. Then N has the distribution called "Poisson binomial". Starting from this more realistic model they prove that the Poisson approximation will be good provided only α is small, whether n is small or large, and whatever value $\sum_{i=1}^{n} p_i$ may have.

In our case, we recall the binary-valued random variable $Y_{x,y}$ defined on a lattice, with $Y_{x,y} = 1$ if there is a fire in the corresponding pixel and by 0 otherwise. We assume that these random variables are independent given the covariates, the grouping effect and the influence of neighbour wildfire occurrences. $N(\Omega_j)$ will now be assumed to have the distribution called the "Poisson Binomial". Using Le Cam and Hodges result it can be shown that $N(\Omega_j)$ has in the limit the Poisson distribution with parameter $\lambda_j = \sum_{(x,y)\in\Omega_J} \pi(x, y)$ provided the $\alpha = \max\{\pi(x, y)\}_{(x,y)\in\Omega_j}$ is small, whether the number of pixels in Ω_j, N_j is small or large and whatever value $\sum_{(x,y)\in\Omega_J} \pi(x, y)$ may have.

Using any distribution the expected number of wildfires in the region Ω_j, $E[N(\Omega_j)]$ is

$$E\left[N(\Omega_j)\right] = \sum_{(x,y)\in\Omega_j} \pi(x, y)$$

The errors are

$$\epsilon_j = n_j - E[N(\Omega_j)] \qquad j = 1, \ldots, 860$$

In Figure 4.9 we include the boxplots and density estimates of the county standarized residuals grouped by state. The standarized residual for the j county has the following expression

$$\hat{\epsilon}_j/\sqrt{v\hat{a}r(\hat{\boldsymbol{\epsilon}}_s)} \tag{4.25}$$

where $\sqrt{v\hat{a}r(\hat{\boldsymbol{\epsilon}}_s)}$ stands for the standard errors of the county residuals in a particular state. It seems that our model is not modeling well a number of large observations. The boxplots for the federal lands (MN, WI and SD) suggests that another kind of distribution with heavy tail might need to be considered. As there are several outliers, we are interested in our estimates to be resistant. For Iowa where the data is reported grouped by county, the residuals show that our model is underestimating the number of fires per county. Mapping the counties with large residuals could give spatial intuition about explanatory variables that are not included in the model.

4.7 Robust locally weighted regression

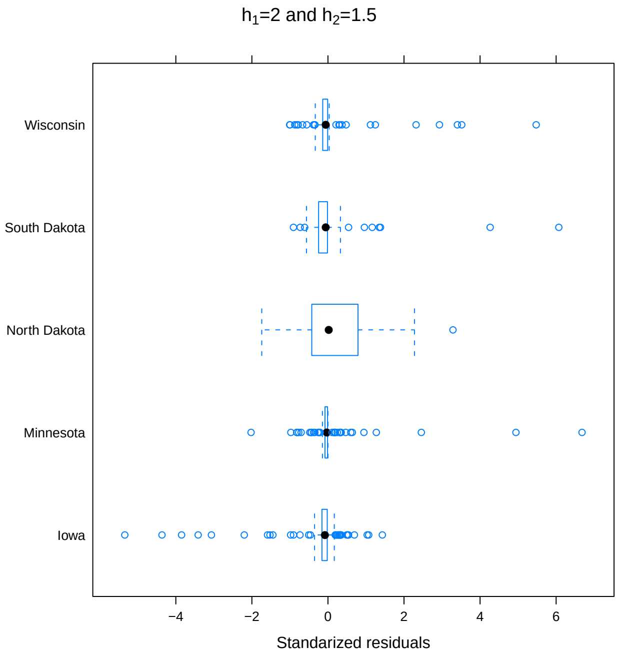
The estimated parameters are obtained by maximizing the weighted log-likelihood function in Equation 4.9. These parameters under the regularity conditions defined in the subsection 1.4.1 have the asymptotic properties of the m.l.e., consistency and efficiency.

The analysis of the residuals shows that the model might not follow the assumption of normality. This may lead us our estimates are robust? Mallows (1979) suggested performing both robust and standard methods and to compare the results. If the differences are minor, either set can be presented. If the differences are not, one must consider why not, and the robust guides to next steps.

We use a robust fitting procedure that guards against deviant points distorting the smoothed points. Through an iterative downweighting procedure similar to the one proposed by Cleveland [9], we use weights that depend on the discrepancy between the data and the fitted model. The method consists on smoothing a partially grouped sample formed by $\{(u_l, v_l, Y_l)_{l=1,\dots,N_{\mathcal{B}}}, (n_j)_{j=1,\dots,J}\}$, in which the predicted value at (x, y) is the value of a logit model fit to binary data using weighted least squares. The weight for (u_l, v_l) is large if (u_l, v_l) is far from (x, y) and small if it is close. The position of the n_j fires are not available then we use the weights $w_j(x, y)$ in Equation 3.4 to include the group effect, it is large if Ω_j is far from (x, y) and small if it is not.

The robust locally weighted regression consists on the following sequence of operations. For each (x, y):

1. Compute the estimates $\hat{\beta}_r^{(s)}(x, y)$, of the parameters in the logit model for the independent variables $\{(Y_l)_{l=1,\dots,N_{\mathcal{B}}}, (n_j)_{j=1,\dots,J}\}$. It is fit by IWLS with the weights



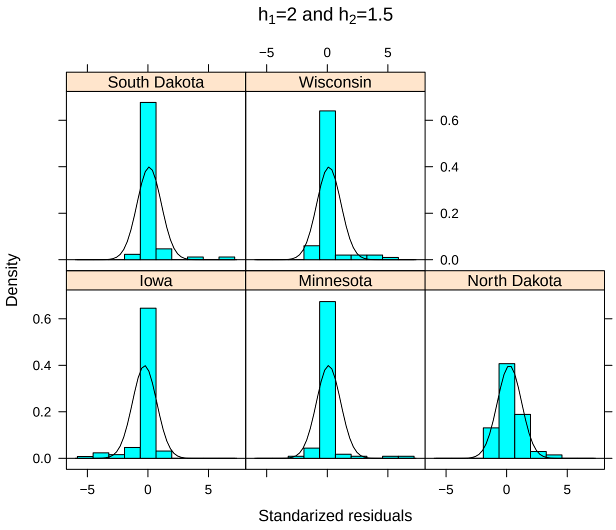


Figure 4.9: Standarized residuals corresponding to the logit model with the NFDRS fuel model as explanatory. In the top, the boxplots are for the standarized residuals grouped by state. In the bottom, kernel density plots for the standarized residuals by state. Superimposed on a histogram are the kernel density estimates with a Gaussian kernel.

$\mathcal{W}(x - u_l, y - v_l)$ $l = 1, \ldots, N_{\mathcal{B}}$ and $w_j(x, y)$ $j = 1, \ldots, J$ for the ungrouped and grouped data respectively. The weight function is "tricube":

$$\mathcal{W}_t(x, y) = \begin{cases} (1 - |u|^3)^3 & |u| \leq 1, \\ 0 & |u|. \end{cases}$$

In this case, the $\hat{\beta}_r^{(s)}(x, y)$ are the values of β that maximize the weighted log-likelihood function Equation 4.20. The fitted values are $\{(\hat{\pi}_l)_{l=1,\ldots,N_{\mathcal{B}}}, (\hat{\pi}_j)_{j=1,\ldots,J}\}$.

2. Define $\{(\hat{\epsilon}_l)_{l=1,\ldots,N_{\mathcal{B}}}, (\hat{\epsilon}_j)_{j=1,\ldots,J}\}$ be the residuals in the final iteration of the IWLS fit. Let $M = median\,((\hat{\epsilon}_l)_{l=1,\ldots,N_{\mathcal{B}}}, (\hat{\epsilon}_j)_{j=1,\ldots,J})$ be the median of the residuals. Define robustness weights by

$$\begin{aligned} w_l^{(m)} &= \mathcal{W}(\hat{\epsilon}_l/6M) \qquad l = 1, \ldots, N_{\mathcal{B}} \\ w_j^{(m)} &= \mathcal{W}(\hat{\epsilon}_j/6M) \qquad j = 1, \ldots, J \end{aligned}$$

where $\mathcal{W}$ is the biweight function defined in the Equation 3.2 and (m) refers to the m iteration.

3. Compute the new $\hat{\pi}(x, y)$ by fitting the logit model using the weighted least squares with weights $w^{(m)}\mathcal{W}_l$ $l = 1, \ldots, N_{\mathcal{B}}$ and $w^m w_j$ $j = 1, \ldots, J$.

4. Repeatedly carry out steps 2 and 3 until convergence. The final $\hat{\pi}(x, y)$ provides a robust locally weighted regression predicted value.

The robust predicted values are included in Figs. 4.10 and 4.11, they are the result of 2 iterations of the downweighting procedure. The first two figures use the same bandwidths for the ungrouped and the grouped data while the third figure applies different amount of smoothing to the ungrouped and grouped data. In the three cases, the central region of North Dakota experiences a high risk of wildfire occurrence before and after the robust procedure. Also, the high risk region at the border of South Dakota with Nebraska and Iowa, experiences a contraction in its area when plotted.

Green [21] establishes that these estimates are also resistant as they are the result of multiplying the prior weights used by glm() with the weights $w^{(m)}$ obtained in the iterative procedure.

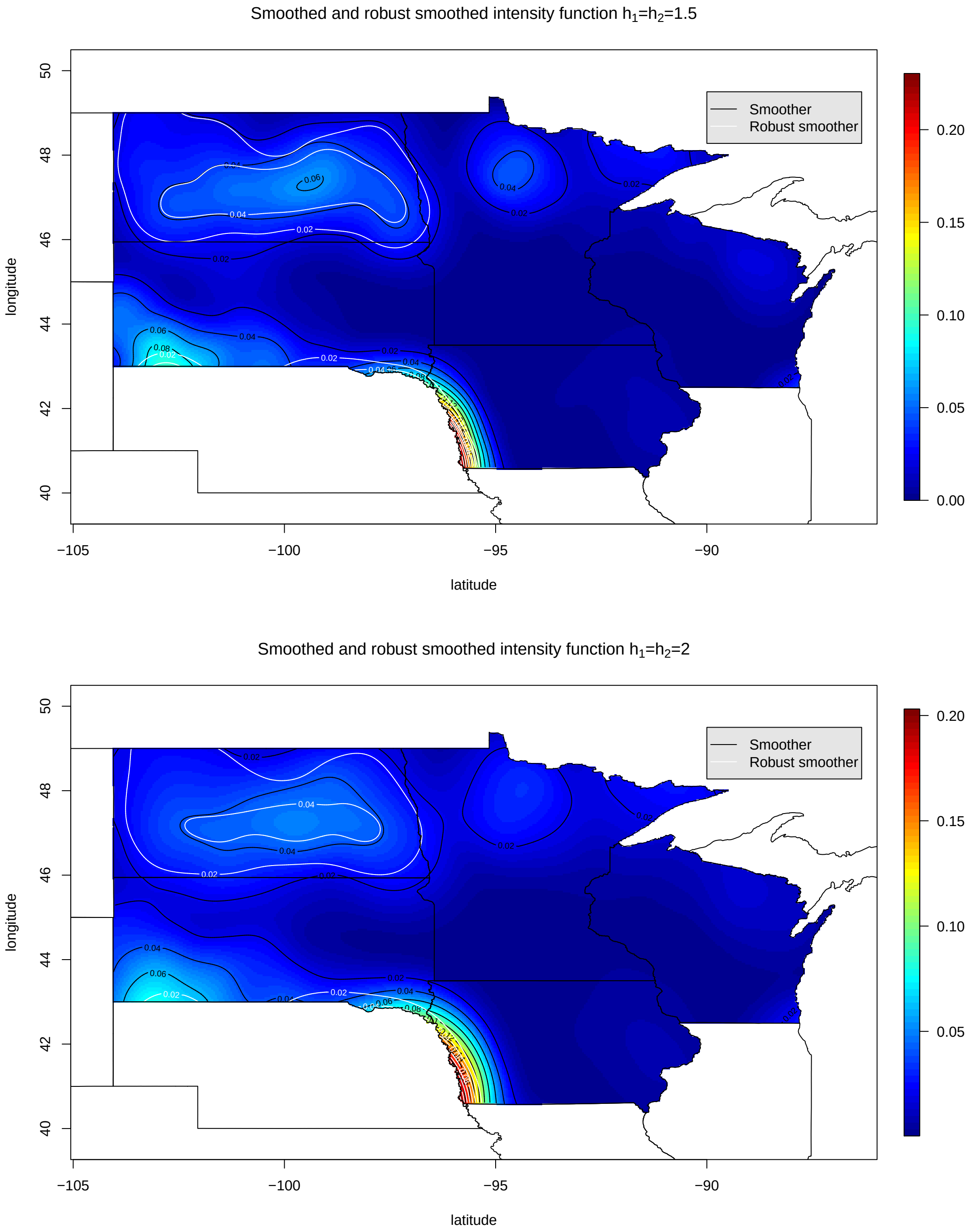


Figure 4.10: The black contour plots are the estimates of the probability of wildfire occurrences in Equation 4.21 by a logit model with just a constant as a regressor. The bandwidths are assumed equal $h_1 = h_2$. The white contour plots are the robust estimates using the iterative downweighting procedure with 2 iterations as explained in section 4.7.

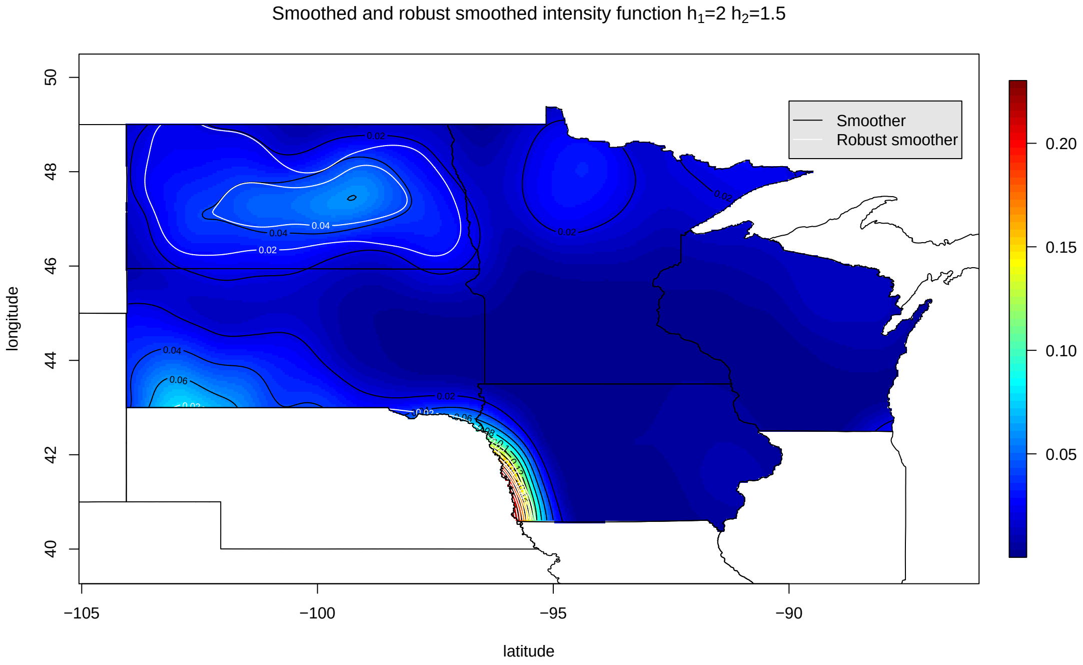


Figure 4.11: The black contour plots are the estimates of the probability of wildfire occurrences given Equation 4.21 by a logit model with just a constant as a regressor employing the bandwidths $h_1 = 2$ and $h_2 = 1.5$. The white contour plots are the robust estimates using the iterative downweighting procedure with 2 iterations as explained in section 4.7.

Chapter 5

Risk analysis

A basic goal of this work is to provide to the US Forest Service with a risk analysis of the wildfire occurrences in the Continental US, both in federal and non-federal regions. Risk may be defined as the probability of some hazardous event or catastrophe, the chance that something bad will occur [4]. A formal risk analysis includes (i) estimation of probabilities, (ii) determination of the distribution of damage and (iii) preparation of products for the uses such as formulas, graphics and hazard risk maps.

We propose a risk analysis with two parts: a) a smoother for a partially grouped sample. It may be used as an the exploratory tool in the data analysis. b) Predicted probabilities for a given time t of wildfire occurrence as a function of location with fuel category as an explanatory variable. Also the uncertainties and the uncertainty associated to these predictions.

5.1 Smoother

Under the assumption that the wildfire occurrence point process has a smooth intensity function we propose a smoother that integrates the ungrouped and grouped wildfire data. It is compared with a smoother based only on grouped data by county. Both smoothers are based on a weighted and estimating equations likelihood analysis. Their structure is linear, the smoothed values being linear combinations of the responses Equation (3.16,4.21).

Kafadar and Morris [32] consider that process of smoothing geographical data aims to:

1. reveal the relationship between the response and location (longitude, latitude), which may suggest a functional model that describes the connection between response and the environment;

2. magnify the underlying trend by reducing the variance in the smoothed values, at the expense of some bias;

3. reduce attention to unusual values or outliers;

4. reveal patterns in the residuals once the smoothed trend has been removed;

5. minimize the undesirable consequences of aggregated data.

We are interested in a smoother that elicits wildfire patterns and higher risk regions, that are not perceptible when the data are partially grouped. In the top of Fig. 5.1 the wildfire occurrences for 1990 are plotted grouped by county and as locations as available. This display makes it difficult to detect trend or surprises. The middle and bottom plots of the same figure, present the ways proposed for integrating the two levels of data aggregation.

The plot in the middle shows the smoother based on data aggregated by county and the bottom one shows the smoother based on the binary-valued approximation to the partially grouped sample. The weights used in their weighted likelihood estimation include the corresponding area in, scaling the number of fires by the area exposed to risk. Both smoothers identify a higher risk of wildfire occurrence in: a) central North Dakota; b) north Minnesota and c) the boundary between South Dakota, Nebraska and Iowa.

Their findings differ in the relationship between the wildfire occurrences in North and South Dakota. The first smoother uncovers a link between these regions the other does not. Possible reasons for this difference is that when data is grouped by county, the fire occurrences are distributed uniformly over the whole county with the biweight function Fig. 4.5 while in the second case the fires' locations for South Dakota are included. A second reason could be that this effect may vary if a different bandwidth is used. Also the robust analysis included in section 4.7, shows evidence that a further analysis is needed to verify their differences.

5.2 Probability estimates and measures of uncertainty

We use a logit model to predict the probability of wildfire occurrence in the region formed by the pilot group (MN, ND, WI, SD and IA). The model includes the NFDRS fuel classification as an explanatory variable. The predictions are done for a grid with pixels having sides of lengths $\delta_x = 4.4$ km and $\delta_y = 4.9$ km.

The predicted probability for the pixel that corresponds to the location (x, y) can be expressed as follow:

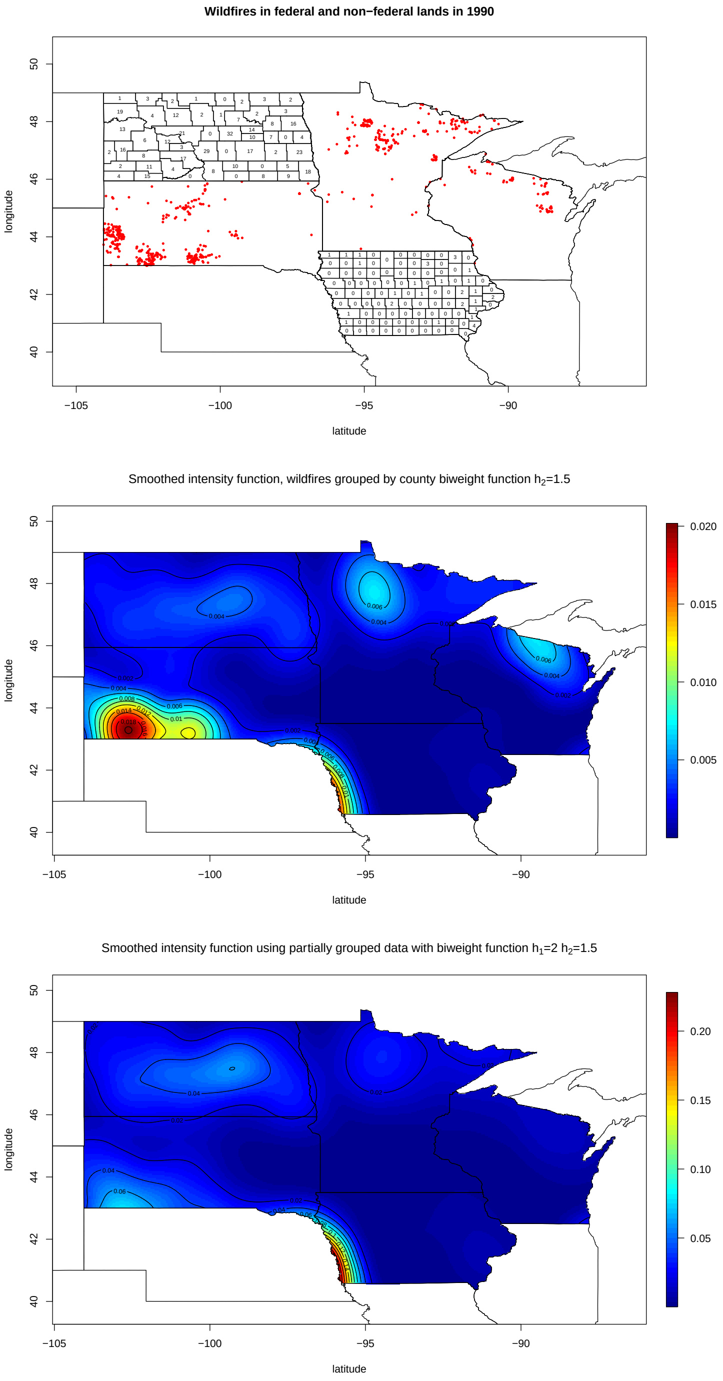

Wildfires in federal and non−federal lands in 1990
longitude
latitude
50
48
46
44
42
40
−105
−100
−95
−90
Smoothed intensity function, wildfires grouped by county biweight function h_2=1.5
0.020
0.015
0.010
0.005
0.004
0.006
0.01
Smoothed intensity function using partially grouped data with biweight function h_1=2 h_2=1.5
0.20
0.15
0.10
0.05
0.04
0.06

Figure 5.1: Top: wildfire occurrences in 1990 displayed as points or county totals. Middle: smoother with data grouped by county Equation (3.16), with biweight and moderate smoothing ($h_2 = 1.5$). Bottom: smoother based on the binary-valued approximation in Equation (4.21), employing biweight function and different bandwidths for the ungrouped ($h_1 = 2$) and the grouped data ($h_2 = 1.5$).

$$\hat{\pi}_3(x,y) \quad = \quad \frac{\exp\{\hat{\beta}_1^{(s)} + \sum_{k=1}^{K} I_k \hat{\beta}_k^{(s)}\}}{1 + \exp\{\hat{\beta}_1^{(s)} + \sum_{k=1}^{K} I_k \hat{\beta}_k^{(s)}\}} \qquad (x,y) \in s \tag{5.1}$$

where $s = \{MN, ND, WI, SD, IA\}$ and the index k runs over the fuel categories existing in the state. The estimated parameters $\hat{\boldsymbol{\beta}}^{(s)} = (\hat{\beta}_1^{(s)}(x,y), \ldots, \hat{\beta}_K^{(s)}(x,y))$ are obtained by maximizing the weighted log-likelihood function Equation (4.18) or with the glm() function including the respective weights shown in Equation (4.16).

The estimated pointwise probabilities are displayed in Fig. 5.2 and findings are: a) in North Dakota there is a strip with a higher probability of wildfire occurrence. This region is located closed to the Lake Sakakawea; b) in the south-west region of South Dakota there is a higher probability of fire and it seems to be located in the Black Hills National Forest. It could be explored for the causes of the high risk region in the north-central part; c) in Minnesota, the regions with higher probability can be associated with the fuel category of hardwood litter, located in the central north region of the state. In this region, there are also some hydrological sources like the North Lake and the Lake of the Woods; d) in Iowa the west border appears to experience a higher risk of fire, it can be associated with the Missouri River or a highway that runs parallel to the border.

In the next section, we include associated confidence intervals. Those provide a measure of the precision with which inferences can be made with our model. We also test if the NFDRS fuel categories are related to the number of wildfire occurrences in the Continental US.

5.2.1 Uncertainty estimates

1. The Fisher information matrix defined in Equation (1.2) for the logit model in Equation (4.18) with p covariates is as follows:

$$\mathbf{I}(\boldsymbol{\beta}) = \begin{pmatrix} \sum_{l=1}^{N_{\mathcal{B}}} \lambda_l (1-\lambda_l) + \sum_{j=1}^{J} N_j \rho_j (1-\rho_j) & \cdots & \sum_{l=1}^{N_{\mathcal{B}}} x_p \lambda_l (1-\lambda_l) + \sum_{j=1}^{J} N_j x_p \rho_j (1-\rho_j) \\ \vdots & \ddots & \vdots \\ \sum_{l=1}^{N_{\mathcal{B}}} x_p \lambda_l (1-\lambda_l) + \sum_{j=1}^{J} N_j x_p \rho_j (1-\rho_j) & \cdots & \sum_{l=1}^{N_{\mathcal{B}}} x_p^2 \lambda_l (1-\lambda_l) + \sum_{j=1}^{J} x_p^2 N_j \rho_j (1-\rho_j) \end{pmatrix}$$

where λ_l is the probability of a wildfire occurrence at the pixel corresponding to the l grid point, $N_{\mathcal{B}}$ is the total grid points in the federal region, ρ_j is the probability of

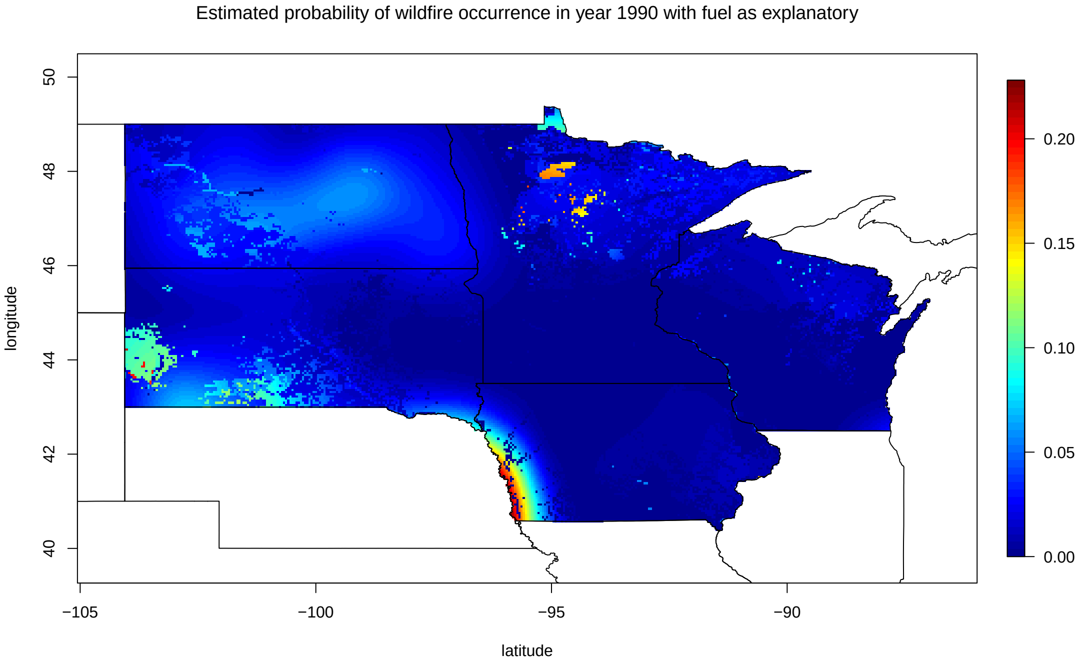

Estimated probability of wildfire occurrence in year 1990 with fuel as explanatory
longitude
50
48
46
44
42
40
−105
−100
−95
−90
latitude
0.20
0.15
0.10
0.05
0.00

Figure 5.2: Probability map using the estimator in Equation (4.18). The model including the NFDRS fuel data. The biweight function with the bandwidths $h_1 = 2$ and $h_2 = 1.5$ for ungrouped and grouped data respectively was employed.

wildfire occurrence in a pixel in the j county and x_p is the p covariate in the design matrix. The maximum likelihood estimates are obtained by IRWL as described in Subsection 4.3.1. For the final iteration, $\hat{\mathbf{I}}(\hat{\boldsymbol{\beta}})^{-1}$ is the estimated variance-covariance matrix for $\boldsymbol{\beta}$.

Under the regularity conditions included in Subsection 1.4.1, we can use the large sample approximate distribution of the m.l.e. for the rth parameter $\hat{\beta}_r \rightarrow N(\beta_r, \mathbf{I}(\boldsymbol{\beta})^{-1}_{r\hat{r}})$ in distribution. For a vector of parameters to be estimated, dividing the estimate of $\hat{\beta}_r$ $r = 1, \ldots, p$ by its standard error is a common practice for examining the significance of the rth explanatory variable in the wildfire occurrence. We use the t-statistic

$$\hat{\beta}_r / \sqrt{v\hat{a}r(\hat{\beta}_r)} \qquad \sim \qquad t_{J+N_B-p}$$

where $\hat{\beta}_r = \hat{\beta}_r(x, y)$, $v\hat{a}r(\hat{\beta}_r)$ is the element in the position (r, r) of the inverse of the observed information matrix $\hat{\mathbf{I}}\left(\hat{\boldsymbol{\beta}}\right)^{-1}$.

In Fig. 5.3, bottom, we plot the estimated t-statistic for the elementary model where the linear predictor just includes a constant term β_1. The values varies from -9.87 to -1.40 and we can consider that the number of fires experiences a spatial variation in the pilot group area.

Before exploring the estimated t-statistics for the model with fuel characteristics, Table 4.5.1 shows the fuel categories that cover the pilot group. The indicator functions included in the linear predictor Eqn. (4.21) correspond to each fuel category. For each state, as we omit the first fuel category the effect of the other fuel categories is measured relative to the omitted category.

The estimated t-statistics for each parameter $\hat{\beta}_1^{(s)}, \ldots, \hat{\beta}_K^{(s)}$ are displayed independently for each state in Figs. 5.4, 5.5, 5.6, 5.7 and 5.8. Our findings were: a) in Wisconsin the constant term is statistically significant . b) in South Dakota the effect of the western Perennial grass in the wildfire occurrence is not statistically significant relative to the pine grass savannah that is the Black Hills National Forest. It is interesting that in the rest of the regions the fuel categories are not statistically significant probably another kind of information related to the vegetation is needed. Also it must be kept in mind that the NFDRS fuel map is intended for use to assess fire danger across the continental U.S., not for fire behavior assessment at any specific site. The purpose of

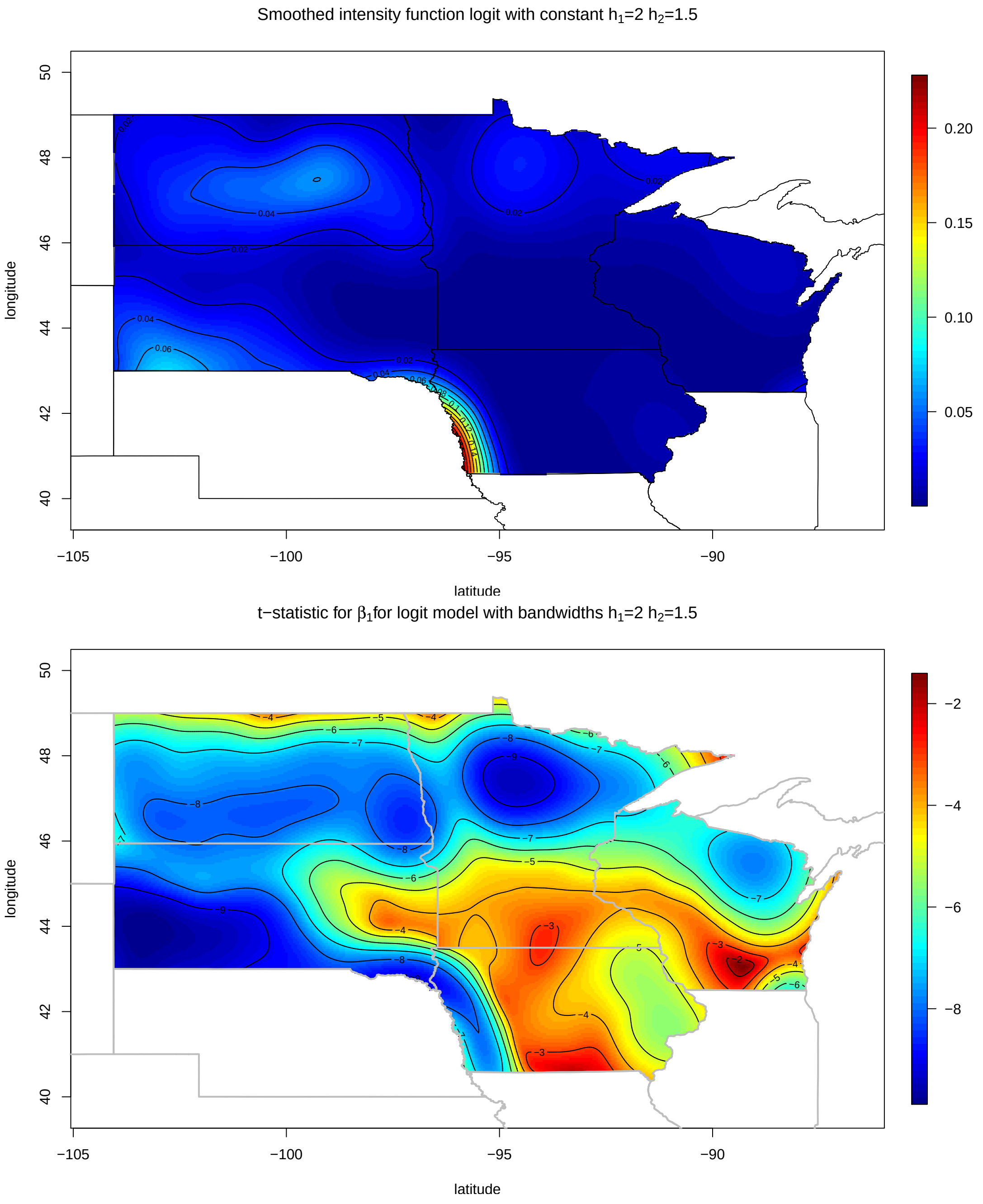

Smoothed intensity function logit with constant h_1=2 h_2=1.5
longitude
latitude
50
48
46
44
42
40
−105
−100
−95
−90
0.20
0.15
0.10
0.05
0.04
0.06
0.02
t−statistic for $\beta_1$for logit model with bandwidths h_1=2 h_2=1.5
longitude
latitude
50
48
46
44
42
40
−105
−100
−95
−90
−2
−4
−6
−8
−3
−5
−7
−9

Figure 5.3: Top: probability map using the estimator in Eqn. (4.21) including only a constant in the linear predictor with the biweight function and bandwidths $h_1 = 2$ and $h_2 = 1.5$ for ungrouped and grouped data respectively. Bottom: estimated t-statistic for the elementary model with just only a constant in Eqn. (4.21). The values of the t-statistic varies from -9.87 to -1.40.

NFDR Fuel Model	MN	ND	WI	SD	IA
A					
B					
C	+	+		+	+
D					
E					
F				+	
G	+	+	+	+	+
H	+	+		+	+
I					
J					
K					
L	+	+	+	+	+
M	+	+	+	+	+
N					
O	+	+	+	+	+
P					
Q	+	+	+	+	+
R	+	+	+	+	+
S		+		+	
T	+	+		+	+
U	+	+	+	+	+
V	+	+	+	+	+
W		+		+	
X					

Table 5.1: The five states that constitute the pilot group: MN, ND, WI, SD and IA with the respective NFDRS fuel categories that cover their neighbours and them. The fuel categories are described in Table 4.1.

the map must be remembered rating fire danger across large geographic areas.

The difficulty of using the t-statistic individually to examine the significance of the rth explanatory variable is that the distributions are marginal. So about 5% of the null values will appear significant at the 5% level. That is why we include a analysis of deviance (ANODEV) for the two nested models, elementary model (only a constant) and the one with fuel categories:

$$\Theta_0 \subset \Theta$$

where Θ_0 stands for the parameters under the null (only a constant) with $K-1$ restrictions and Θ stands for the parameters in the model with the fuel categories. The test statistic is the difference in deviances

$$2\left[\ell_w(\tilde{\Theta}) - \ell_w(\hat{\Theta}_0)\right] - 2\left[\ell_w(\tilde{\Theta}) - \ell_w(\hat{\Theta})\right] \tag{5.2}$$

where $\tilde{\Theta}$ is the m.l.e for the saturated model (one parameter for each observation). Under the null hypothesis the difference in deviances has a χ^2 distribution with $K-1$ degrees of freedom.

2. The standard error for the estimated probability at position (x, y) is obtained by the delta method,

$$Var[g(\hat{\boldsymbol{\beta}})] \to g'(\hat{\boldsymbol{\beta}})^T var(\hat{\boldsymbol{\beta}}) g'(\hat{\boldsymbol{\beta}}))$$

in particular for the logit link function

$$g(\hat{\boldsymbol{\beta}}) = \frac{\exp{(x^t\hat{\boldsymbol{\beta}})}}{1 + \exp{(x^t\hat{\boldsymbol{\beta}})}}$$

Figures 5.11 and 5.12 show the approximate 95% confidence intervals for the logit models, the elementary (only a constant) and the one with the fuel categories respectively.

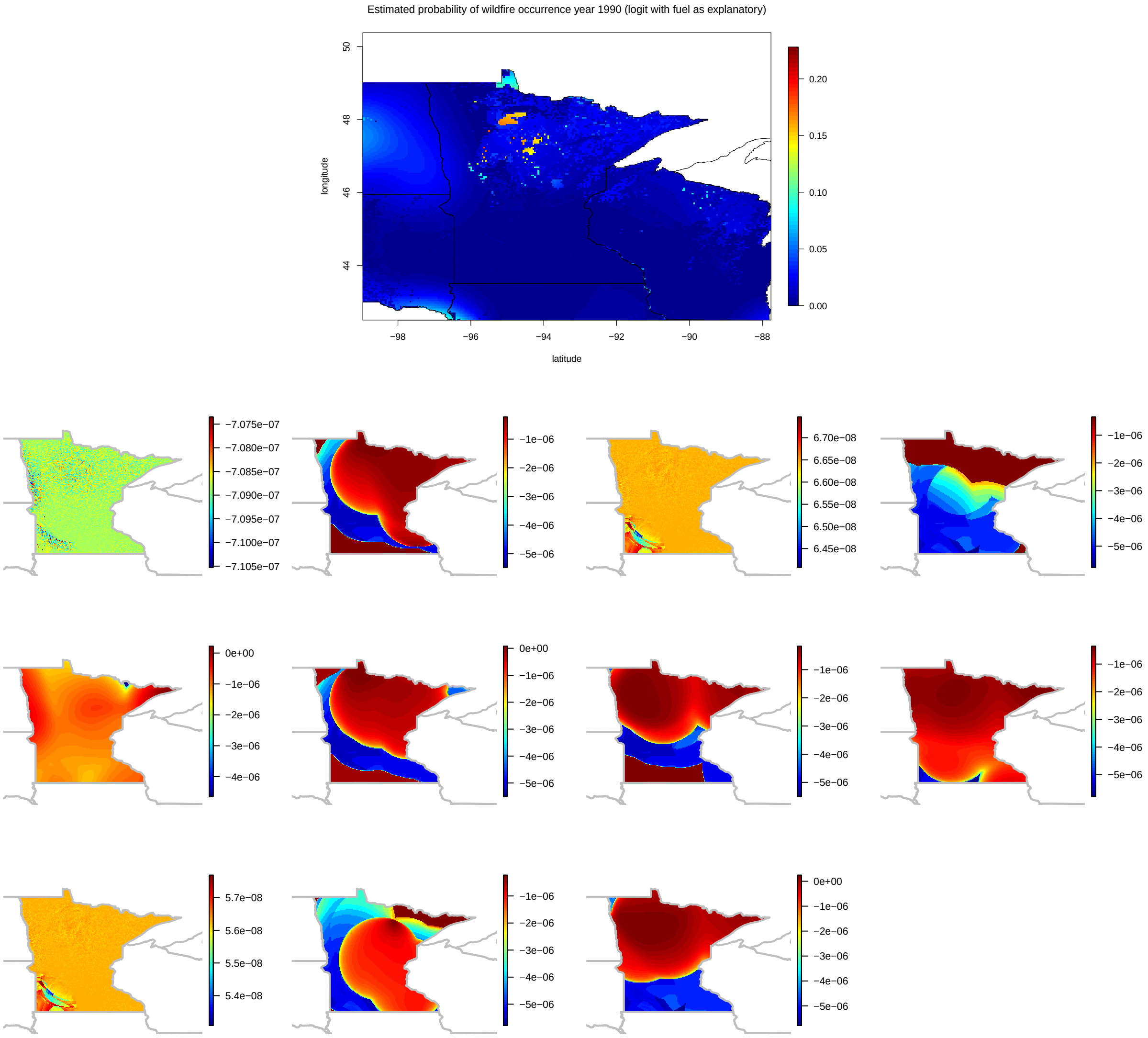


Figure 5.4: Top: probability map for Minnesota using Eqn. (4.18) with the fuel categories shown in Table 4.5.1. The biweight function with bandwidths $h_1 = 2$ and $h_2 = 1.5$ for ungrouped and grouped data respectively. Bottom: estimated t-statistic for $\hat{\beta}_1^{(s)}, \ldots, \hat{\beta}_{11}^{(s)}$ where $s = MN$, Minnesota and the omitted category is the pine grass savannah. The values of the t-statistic vary from -5.78 e-06 to 2.44 e-07.

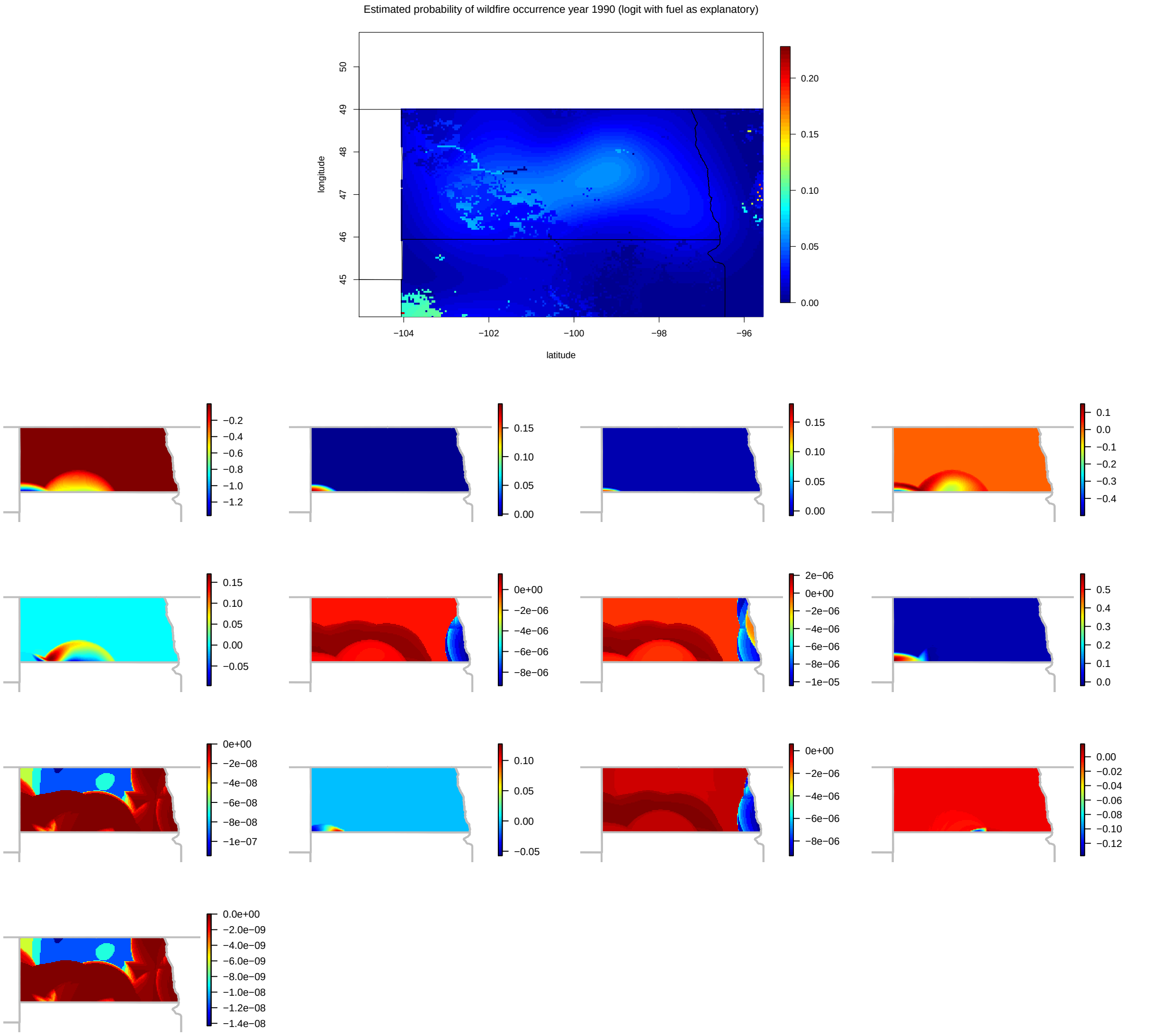


Figure 5.5: Top: probability map for North Dakota using Eqn. (4.18) with the fuel categories shown in Table 4.5.1. The biweight function with bandwidths $h_1 = 2$ and $h_2 = 1.5$ for ungrouped and grouped data respectively. Bottom: estimated t-statistic for $\hat{\beta}_1^{(s)}, \ldots, \hat{\beta}_{11}^{(s)}$ where $s = ND$ and the omitted category is the pine grass savannah. The values of the t-statistic vary from -1.36 to 0.58.

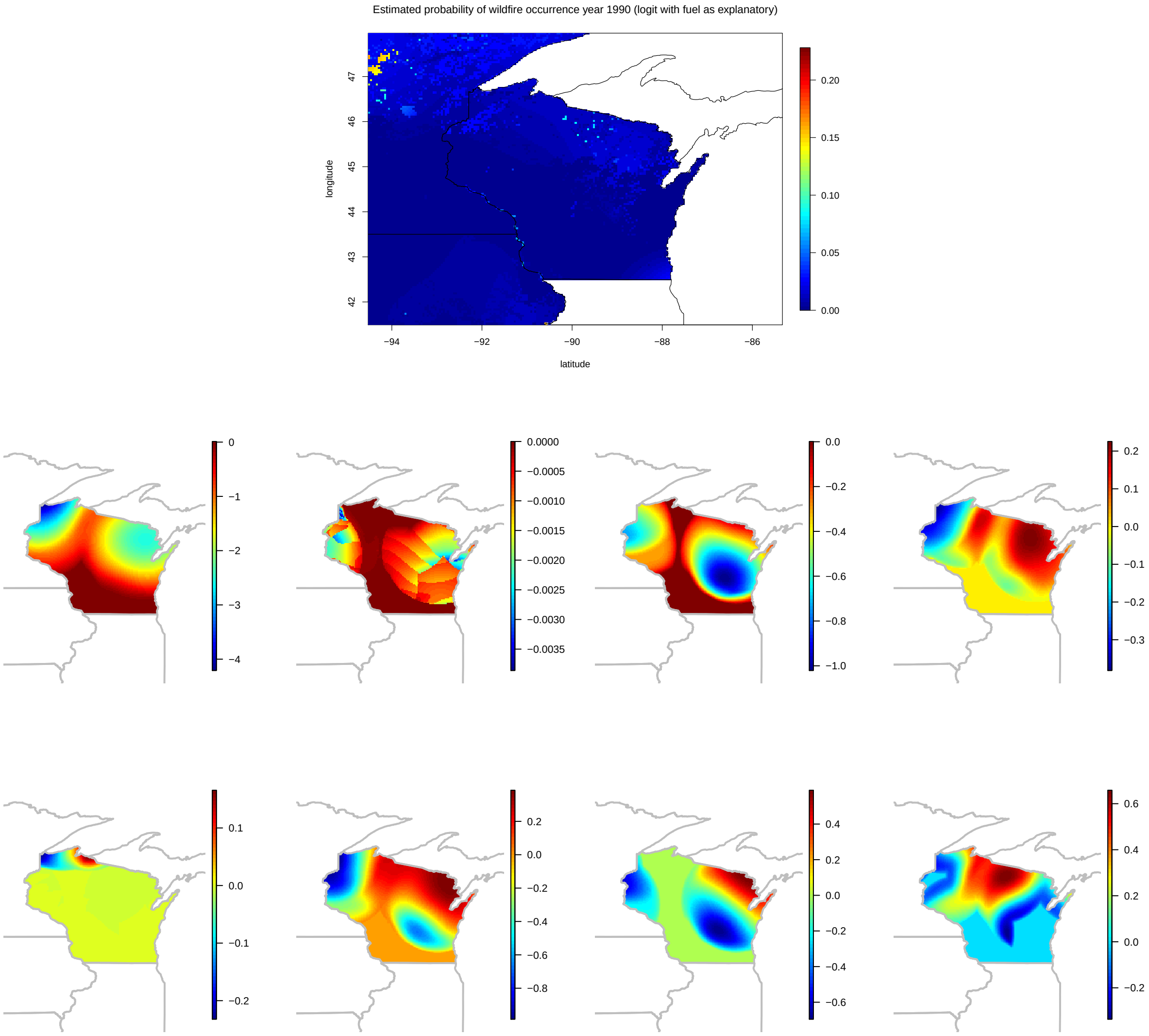


Figure 5.6: Top: probability map for Wisconsin using Eqn. (4.18) with the fuel categories shown in Table 4.5.1. The biweight function with bandwidths $h_1 = 2$ and $h_2 = 1.5$ for ungrouped and grouped data respectively. Bottom: estimated t-statistic for $\hat{\beta}_1^{(s)}, \ldots, \hat{\beta}_{11}^{(s)}$ where $s = WI$, Wisconsin and the omitted category is the conifers. The t-statistic vary from -4.20 to 0.65.

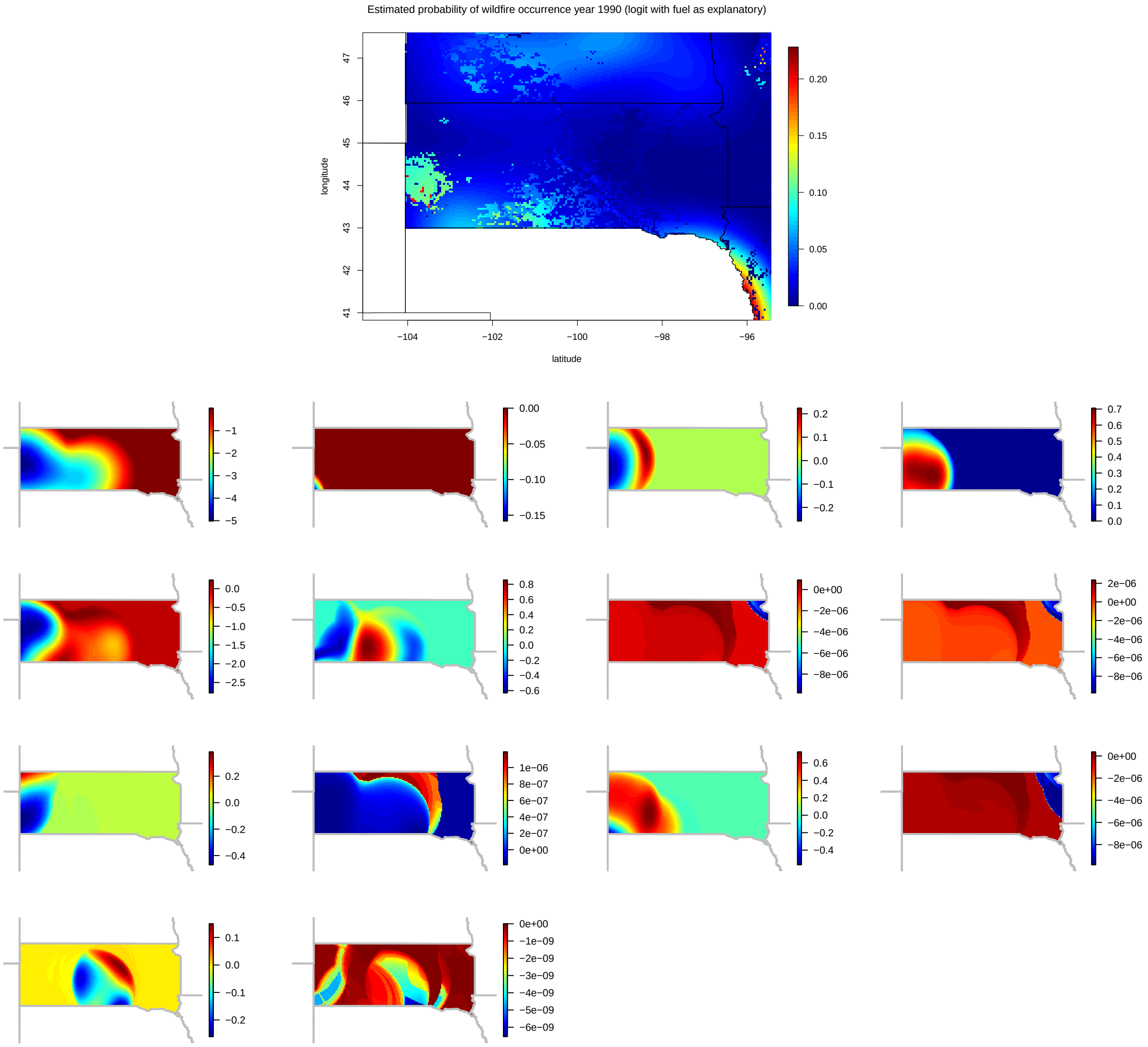


Figure 5.7: Top: probability map for South Dakota using Eqn. (4.18) with the fuel categories shown in Table 4.5.1. The biweight function with bandwidths $h_1 = 2$ and $h_2 = 1.5$ for ungrouped and grouped data respectively. Bottom: estimated t-statistic for $\hat{\beta}_1^{(s)}, \ldots, \hat{\beta}_{11}^{(s)}$ where $s = SD$, South Dakota and the omitted category is the pine grass savannah. The values of the t-statistic vary from -5.01 to -0.85.

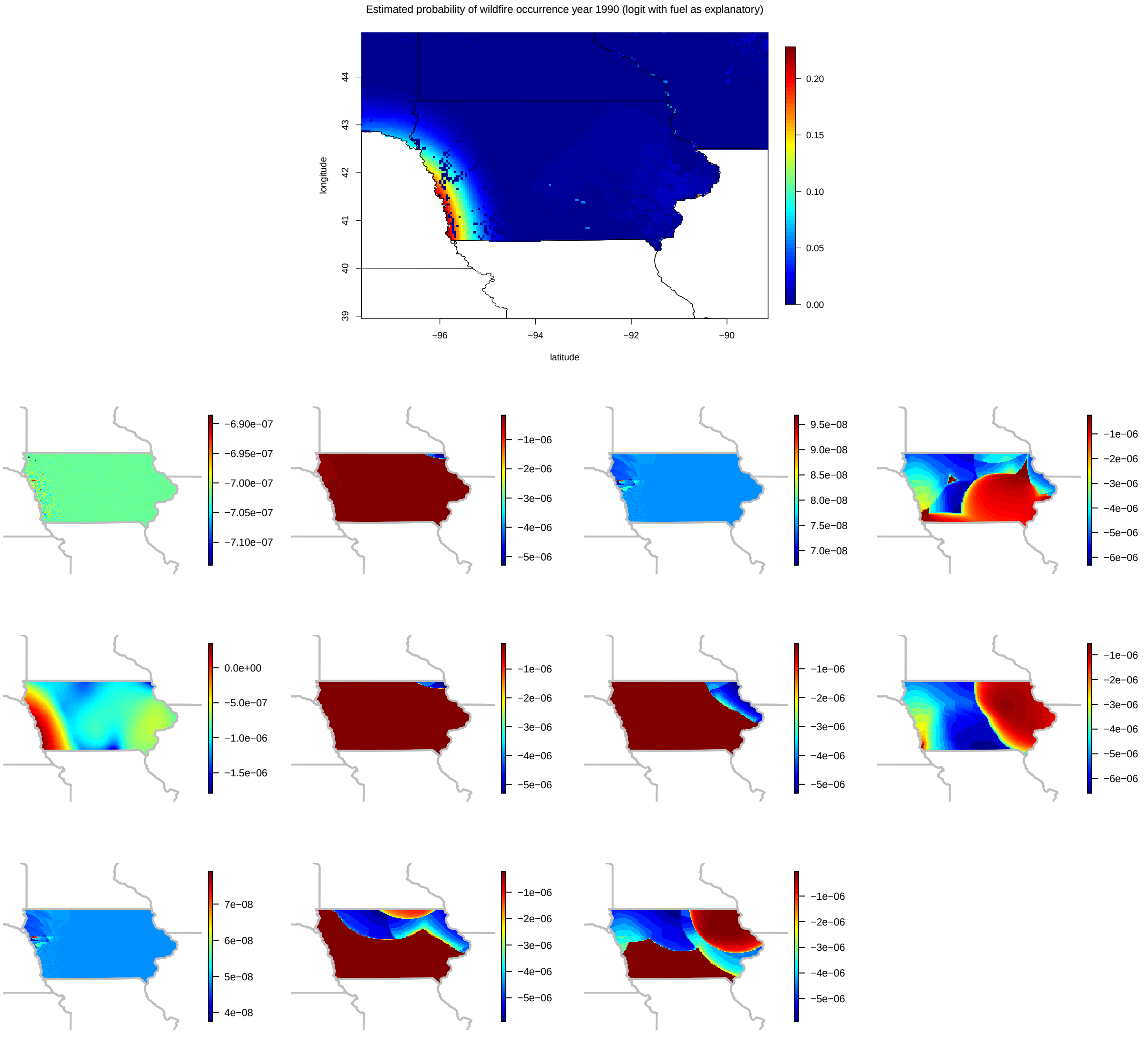


Figure 5.8: Top: probability map for Minnesota using the estimator in Eqn. (4.18) with the fuel categories shown in Table 4.5.1. The biweight function with the bandwidths $h_1 = 2$ and $h_2 = 1.5$ for ungrouped and grouped data respectively. Bottom: estimated t-statistic for $\hat{\beta}_1^{(s)}, \ldots, \hat{\beta}_{11}^{(s)}$ where $s = IA$ and the omitted category is the pine grass savannah. The values of the t-statistic varies from -6.59 e-06 to 3.48 e-07.

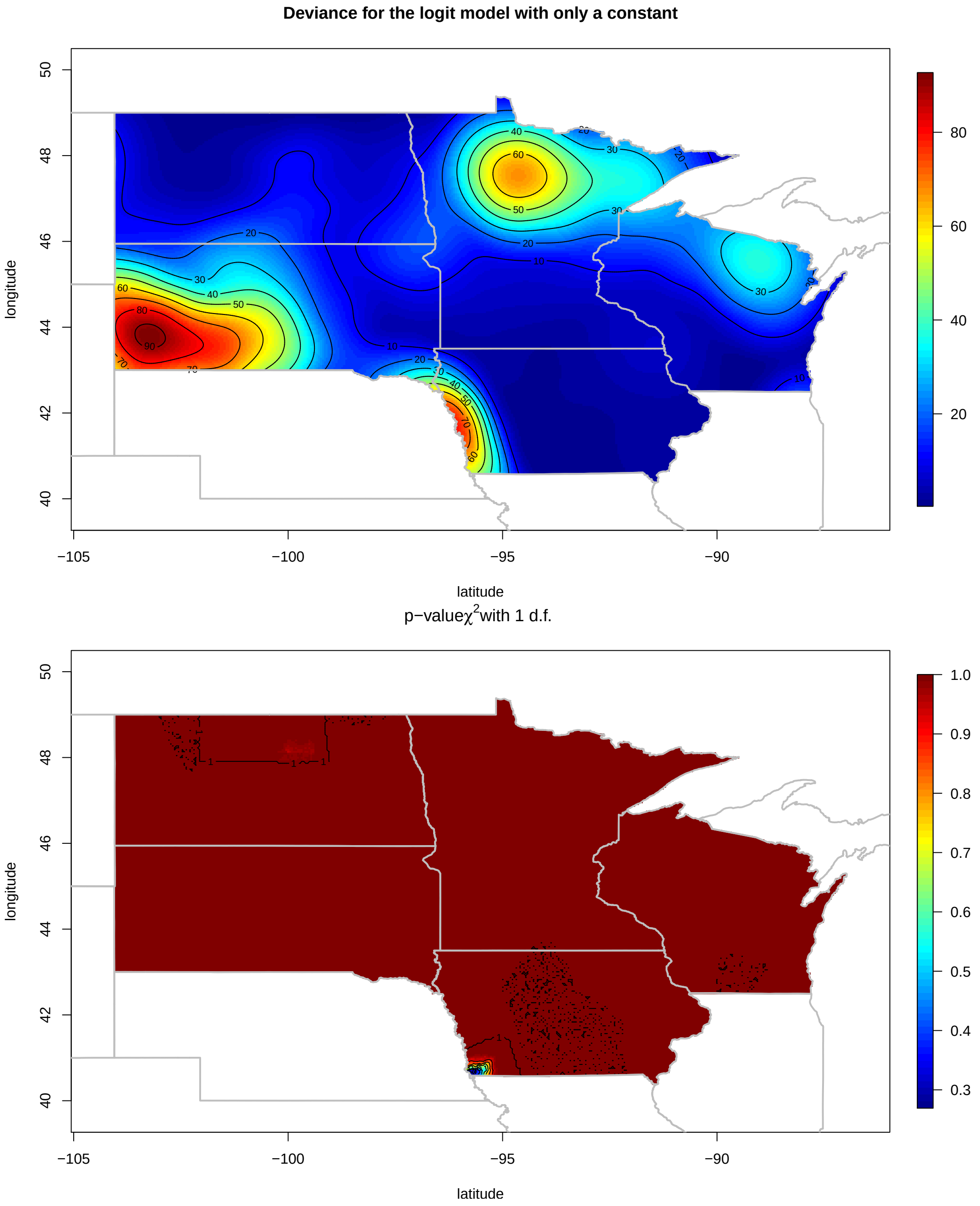


Figure 5.9: Top: deviance for the elementary model with only a constant as explanatory variable. Bottom: the p-value for the χ^2 distribution with 1 degrees of freedom.

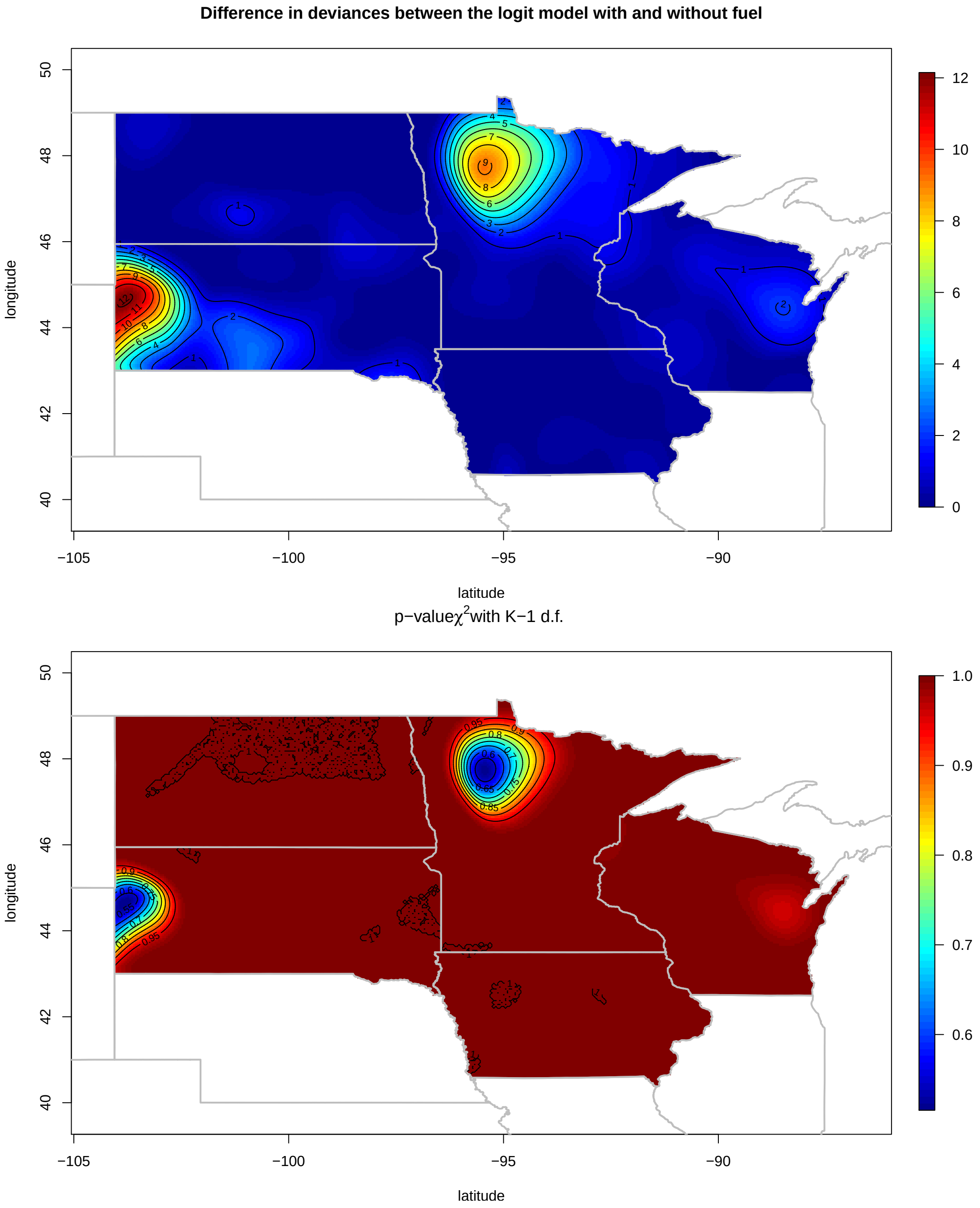
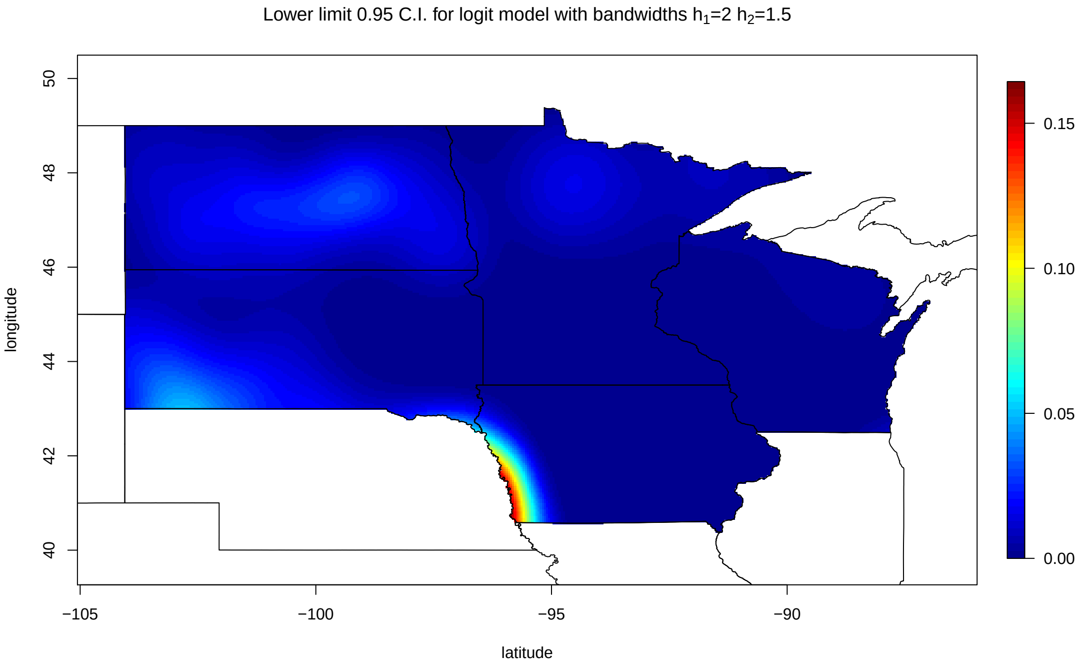


Figure 5.10: Top: difference in deviances in Equation 5.2 of the null when only a constant is included in the logit model and where fuel categories are included as explanatory. Bottom: the p-value for the χ^2 distribution with $K-1$ degrees of freedom.

In both cases, the confidence intervals for the predictions in the boundary between Iowa and Nebraska are narrower. In the elementary model, the uncertainty associated with the predictions in the north central part of Wisconsin is less. The confidence intervals for the logit with fuel model are less extend in northern Minnesota and in south-east region of South Dakota.


Lower limit 0.95 C.I. for logit model with bandwidths h_1=2 h_2=1.5
longitude
latitude
50
48
46
44
42
40
−105
−100
−95
−90
0.15
0.10
0.05
0.00

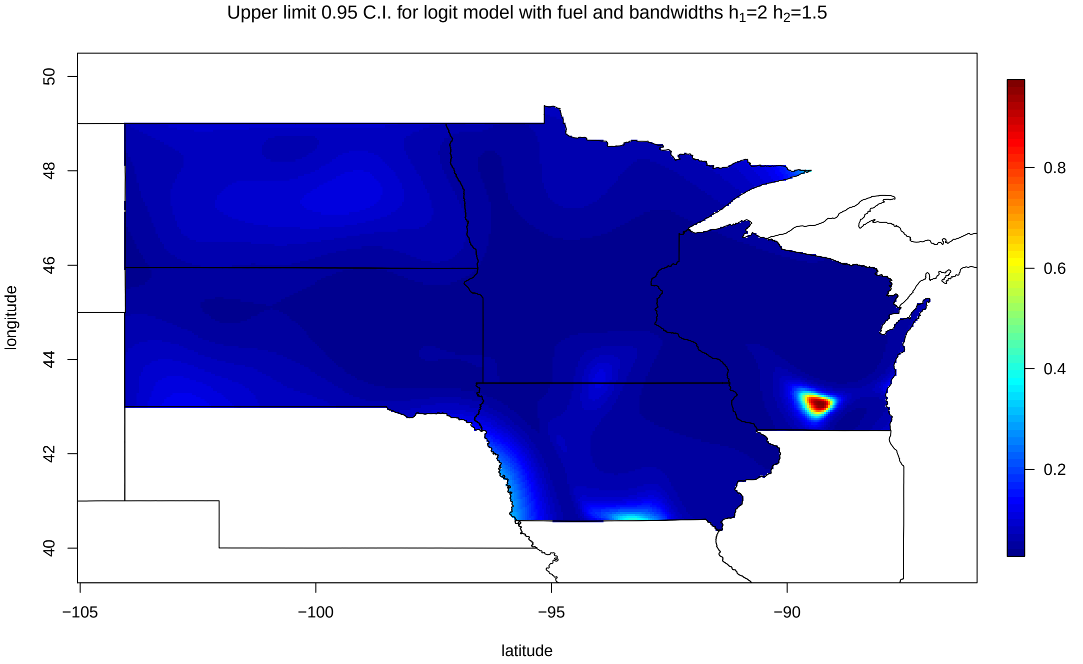
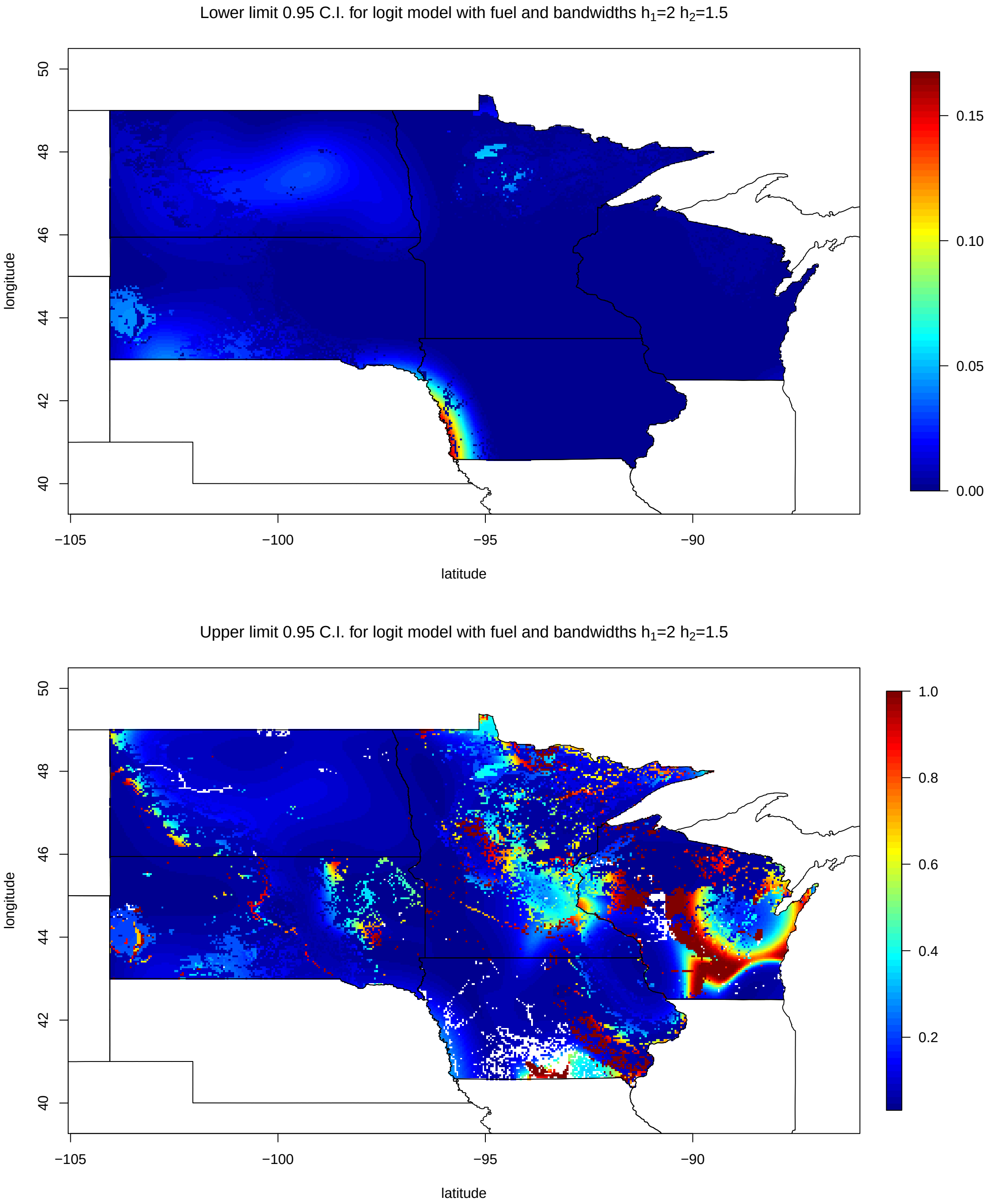

Upper limit 0.95 C.I. for logit model with fuel and bandwidths h_1=2 h_2=1.5
longitude
latitude
50
48
46
44
42
40
−105
−100
−95
−90
0.8
0.6
0.4
0.2

Figure 5.11: A 95% confidence interval for the predicted probability using just a constant in the linear predictor of the logit model. Top: lower limit of the confidence interval. Bottom: upper limit of the confidence interval.



Lower limit 0.95 C.I. for logit model with fuel and bandwidths h_1=2 h_2=1.5

Upper limit 0.95 C.I. for logit model with fuel and bandwidths h_1=2 h_2=1.5

Figure 5.12: A 95% confidence interval for the predicted probability using the NFDRS fuel category as an explanatory variable. Top: lower limit of the confidence interval. Bottom: upper limit of the confidence interval.

Chapter 6

Conclusions and future work

In this chapter we summarize the contributions of this thesis and point out some possibilities for future work.

6.1 Summary

One of the goals of this thesis is the development of statistical methods that provide an integrated analysis of different sets of spatial data. In particular, we examine partially grouped samples where the observations are available in two levels of data aggregation: points and aggregate counts in area. These two types of data refer to the same stochastic process and therefore we develop a joint analysis, as opposed to transforming the observations to the same level of aggregation (for example, by aggregating the points located in the same area).

A contribution of the work is the use of a binary-valued process as a way to approximate a partially grouped sample. Previously, a binary-valued process had been used to model spatial point processes (see, for example, [5]), and we extended this result to the more general case of approximating the combination of point locations and aggregate counts. Based on this approximation we propose a smoother as an exploratory tool in the analysis of ungrouped and grouped data.

We consider the wildfire occurrences process to have a locally constant smoothed intensity function. Based on this assumption we developed a smoother based on a weighted likelihood analysis fit of the binary-valued approximation. The smoother assumes that the intensity rate is constant in the support of the weight function. In our analysis, the weights play two important roles: to include the grouping effect in the model and to capture local variation. The analysis uses two different bandwidths, one for the ungrouped data and other for the grouped data. The bandwidths are assumed to be constant in both coordinates. In the future, the model can be refined to include a cross-validation analysis to select the bandwidths. In addition, it may be worthwhile to explore if having bandwidths by region or state improves our results.

A fundamental part of the risk analysis is to predict the mean rate of wildfire occurrence in 1991 for the pilot group. These estimates are based on a logit model that rest on the binary-valued approximation. We include the NDFRS fuel model as an explanatory variable and notice that its significance on the of wildfire-occurrence probability varies spatially. Future work is underway to incorporate other explanatory variables.

The risk analysis of the wildfire occurrence is thought to be useful for the allocation of resources for the US Forest Service. Therefore, measuring the uncertainty of our predictions is fundamental to evaluate the reliability of our estimates. The estimated standard errors are used to construct the 95% confidence intervals for the predicted probabilities. The visualization of uncertainty measures, like confidence intervals, is a topic that deserves additional study.

The model assessment is based on the residuals analysis using the observations grouped by county and applying the binomial Poisson distribution proposed by Le Cam and Hodges. It will be of interest to propose a model assessment that can be done using the 0-1 approximation. As there are some differences between the estimations done with robust and standard methods, other robust estimates might be explored.

In this thesis, in order to develop computer routines and insights more easily, we focused on the data for one year and five states. Work is underway to use our model to analyze the data for several years and on the whole Continental Unites States.

6.2 Problems for future research

We mention a number of research topics that arise from this thesis:

1. **Spatial-Temporal Analysis**: A natural extension of this work is to include the time domain in our model. The idea of binary approximation can be extended to a spatial-temporal analysis; this is done by dividing the domain into voxels $(x, x+dx] \times (y, y+dy] \times (t, t+dt]$. The corresponding intensity function for the spatial-temporal intensity function will be [5]:

$$\lambda(x,y,t) = Prob\{dN(x,y,t) = 1|H_t\}/dxdydt$$

where $dN(x, y, t) = N(dx, dy, dt)$ counts the number of fires in the voxel where H_t is the history of N process up to and including time t. In particular, predicting the probability of a wildfire occurrence for a day in the year is of interest..

2. **Parameter properties**: Chapter 2 reviews a selection of grouped-data estimation methodologies that can be extended to the more general case of partially grouped data. Chapter 2 also shows how the raw moments can be related to the ones obtained with grouped data like Sheppard's correction [48]. A question that might be of interest to explore is the relationship between the parameters estimated with complete observations and those with partially grouped data, for example using the centroid as the value of observed data for grouped data.

3. **Volatility**: Future research is needed in order to develop methods for modeling volatility with partially grouped data. Here we briefly mention an idea in this direction. For a given position (x, y), a data smoother produces two sequences: the smooth sequence, $S_{x,y}$, and the rough sequence, $\sigma_{x,y}$; and compound smoothers might be obtained by re-roughing. In terms of the volatility of the process, we can use of $\sigma_{x,y}$, the conditional standard deviation. Some preliminary estimates of the volatility in terms of the standard deviation $\hat{\sigma}_{x,y}$ may be:

$$\begin{aligned} Y_{x,y} &= S_{x,y} + \sigma_{x,y}\epsilon_{x,y} \\ \hat{S}_{x,y} &= smooth(Y_{x,y}) \\ \hat{\sigma}_{x,y} &= smooth(|Y_{x,y} - \hat{S}_{x,y}|) \\ \hat{\epsilon}_{x,y} &= \frac{Y_{x,y} - \hat{S}_{x,y}}{\hat{\sigma}_{x,y}} \end{aligned}$$

where $\epsilon_{x,y}$ is white noise.

4. **Weighted Likelihood analysis**: it may be important to consider the boundary effects of the estimates obtained from the weighted likelihood analysis.

Bibliography

[1] David R. Brillinger. Discussion of Stone. *Annals of Statistics*, (5):622–662, 1977.

[2] David R. Brillinger. Spatial temporal modelling of spatially aggregate birth data. *Survey Methodology*, 16(2):255–269, 1990.

[3] David R. Brillinger. *Time Series Data Analysis and Theory.* SIAM, 2001.

[4] David R. Brillinger. Three environmental probabilistic risk problems. *Statistical Science*, 18(4):412–421, 2003.

[5] David R. Brillinger, Haiganoush K. Preisler, and John W. Benoit. Risk assessment: a forest fire example. *The Institute of Mathematical Statistics Lecture NotesMonograph Series. Statistics and Science: A Festschrift for Terry Speed*, 40:177–196, 2003.

[6] Robert E. Burgan, Robert W. Klaver, and Jacqueline M. Klaver. Fuel models and fire potential from satellite and surface observations. *International Journal of Wildland Fire*, 3(8), August 1998.

[7] J Burridge. A note on maximum likelihood estimation for regression models using grouped data. *Journal of the Royal Statistical Society*, 43(1):41–45, 1981.

[8] George Casella and Roger L. Berger. *Statistical Inference.* Duxbury Advanced Series. Duxbury, Pacific Grove, California, 2002.

[9] William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), December 1979.

[10] William S. Cleveland and Clive Loader. Smoothing by local regression: Principles and methods. *AT&T Bell Laboratories*, 1995.

[11] A.C. Davison. *Statistical Models.* Cambridge Series in Statistical and Probabilistic Mathematics, 2003.

[12] P. W. A. Dayanada. Stochastic models for forest fires. *Ecological Modelling*, 3(4), October 1977.

[13] Arthur P. Dempster, N. M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[14] Arthur P. Dempster and Donald B. Rubin. Rounding error in regression: The appropriateness of sheppard's corrections. *Journal of the Royal Statistical Society*, 45(1):51–59, 1983.

[15] Nira Dyn and Grace Wahba. On the estimation of functions of several variables from aggregated data. *SIAM Journal on Mathematical Analysis*, 13(1):134–152, Jan 1982.

[16] William F. Eddy and Audris Mockus. An example of the estimation and display of a smoothly varying function of time and space the incidence of the disease mumps. *Journal of the American Society for Information Science*, 45(9):686–693, 1994.

[17] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, pages 309–368, 1922.

[18] T. Gasser and H. G. Müller. Kernel estimation of regression functions. *Lecture Notes in Mathematics: Smoothing Techniques for Curve Estimation eds. T. Gasser and M. Rosenblatt*, pages 23–68, 1979.

[19] Alan E. Gelfand, Peter J. Diggle, Montserrat Fuentes, and Peter Guttorp. *Handbook of Spatial Statistics*. Handbooks of Modern Statistical Methods. Chapman and Hall, CRC, Boca Raton, Florida, 2010.

[20] Carol A. Gotway and Linda J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458), June 2002.

[21] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):149–192, 1984.

[22] Peter Guttorp, David R. Brillinger, and Frederic Paik Schoenberg. Point processes, spatial. *Encyclopedia of Environmetrics*, 3:1571–1573, 2002.

[23] Y. Haitovsky. Grouped data. *Encyclopedia of Statistical Sciences*, 2006.

[24] Daniel F. Heitjan. Inference from grouped continuous data a review. *Statistical Science*, 4(2):164–179, May 1989.

[25] Daniel F. Heitjan. Regression with bivariate grouped data. *Biometrics*, 47(2):549–562, Jun 1991.

[26] Daniel F. Heitjan and Donald B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, 1991.

[27] Peter J. Huber. Discussion of Stone. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, edited by Lucien Lecam and Jerzy Neyman.*, 1:221–233, July 1965.

[28] N. Inagaki. Asymptotic relations between the likelihood estimating function and the maximum likelihood estimator. *Annals of the Institute of Statistical Mathematics*, 25(1):1–26, July 1973.

[29] J. L. Hodges J.R. and Lucien Le Cam. The poisson approximation to the poisson binomial distribution. *The Annals of Mathematical Statistics*, 31(3), September 1960.

[30] Karen Kafadar. Simultaneous smoothing and adjusting mortality rates in US counties melanoma in white females and white males. *Statistics in Medicine*, (18):3167–3188, 1999.

[31] Karen Kafadar. The influence of John Tukey's work in robust methods for chemometrics and environmetrics. *Chemometrics and intelligent laboratory systems*, (60):127–134, 2002.

[32] Karen Kafadar and Max D. Morris. Nonlinear smoothers in two dimensions for environmental data. *Chemometrics and intelligent laboratory systems*, (60):113–125, 2002.

[33] Maurice G. Kendall. *Kendall's advanced theory of statistics Vol I.* Halsted Press, 1994.

[34] Gunnar Kulldorff. *Estimation from grouped and partially grouped samples.* John Wiley and Sons, Inc., New York, 1961.

[35] D. V. Lindley. Grouping corrections and maximum likelihood equations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 46(1):106, Apr 1949.

[36] D. Mandallaz and R. Ye. Statistical model for the prediction of forest fires. *Report Project Minerve II, ETH Zurich*, 1996.

[37] D. Mandallaz and R. Ye. Prediction of forest fires with poisson models. *Canadian Journal of Forest Research*, 27:1685–1694, 1997.

[38] P. McCullagh and J.A. Nelder. *Generalized Linear Models.* Chapman and Hall CRC Monographs on Statistics and Applied Probability, 1999.

[39] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics. Wiley Interscience, Hoboken, New Jersey, 2008.

[40] Audris Mockus. Estimating dependencies from spatial averages. *Journal of Computational and Graphical Statistics*, 7(4):501–513, Dec 1998.

[41] H. G. Müller. Smooth optimum kernel estimators of densities, regression curves and modes. *The Annals of Statistics*, 12(2):766–774, June 1984.

[42] Karl Pearson. On the systematic fitting of curves to observations and measurements. *Biometrika*, 1(3):266–303, Apr 1902.

[43] M. Poulin-Costello. *People-caused forest fire prediction using poisson and logistic regression.* Master Thesis Department of Mathematics and Statistics, University of Victoria, Victoria, Canada, 1993.

[44] Haiganoush K. Preisler, David R. Brillinger, Robert E Burgan, and J. W. Benoit. Robust locally weighted regression and smoothing scatterplots. *International Journal of Wildland Fire*, 74(13):133–142, 2004.

[45] C. Radhakrishna Rao. *Linear Statistical Inference and Its Applications.* Wiley Series in Probability and Statistics, 2002.

[46] J. Rice. Boundary modification for kernel regression. *Communications in Statistics, Part A Theory and Methods*, 13:893–900, 1984.

[47] Kirsten M. Schmidt, James P. Menakis, Colin C. Hardy, and David L. Bunnell. Development of coarse-scale spatial data for wildland fire and fuel management. *United States Department of Agriculture Forest Service Rocky Mountain Research Station*, General Technical Report RMRS-87, April 2002.

[48] W. F. Sheppard. On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant divisions of a scale. *Missing*, 29(634):353–380, May 1898.

[49] Donald L. Snyder and Michael Miller. *Random Point Processes in Time and Space.* Springer Texts in Electrical Engineering. Springer-Verlag, New York, New York, 1991.

[50] Joan G. Stanislawis. The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84, March 1989.

[51] USFS-WFAS Wildland Fire Assessment System. *NFDRS Fuel Model Map.* 2010. URL: http://www.wfas.net/index.php/nfdrs-fuel-model-static-maps-44.

[52] G. M. Tallis. Approximate maximum likelihood estimates from grouped data. *Technometrics*, 9(4):599–606, Nov 1967.

[53] Waldo R. Tobler. Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530, Sep 1979.

[54] Grace Wahba. Numerical experiments with the thin plate histospline. *Communications in Statistics-Theory and Methods*, 1981.

[55] Grace Wahba. Splines in nonparametric regression. *Encyclopedia of Environmetrics*, 2000.

[56] Herman Wold. Sheppard's correction formulae in several variables. *Skandinavisk Aktuarietidskrift*, 17:248–255, 1934.

[57] Simon N. Wood. *Generalized Additive Models. An Introduction with R.* Texts in Statistical Science. Chapman and Hall CRC, Boca Raton,Florida, 2006.