UNIVERSITY OF CALIFORNIA

Los Angeles

# Separable Temporal Modeling of Point Processes on Linear Networks & Balancing Data Sufficiency and Privacy

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

by

Medha Uppala

2018

ABSTRACT OF THE DISSERTATION

Separable Temporal Modeling
of Point Processes on Linear Networks
&
Balancing Data Sufficiency and Privacy

by

Medha Uppala
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2018
Professor Mark S. Handcock, Chair

The first part of the dissertation focusses on spatial and temporal modeling of point processes on linear networks. Point processes on/near linear networks can simply be defined as point events occurring on or near line segment network structures embedded in a certain space. A separable modeling framework is presented that fits a *formation* and a *dissolution* model of point processes on linear networks over time. Two major applications of the separable temporal model are spider web building activity in brick mortar lines and wildfire ignition origins near road networks.

The second part of the dissertation focusses on analyses of large energy databases, specifically the Energy Atlas database. The main motivation of this part is to explore and understand the issues of balancing necessary data resolution while maintaining consumer privacy. The issue of data resolution and its importance are explored by first tackling a specific policy objective. This is achieved by applying a longitudinal quantile regression model to parcel-level monthly energy consumption in the Westwood neighborhood; the model results aid in fulfilling efficiency goals outlined in the California Senate Bill 350. Then the issue of record privacy is explored through a review of current privacy methods, implementation, data ownership, and concluded with avenues of future research.

The dissertation of Medha Uppala is approved.

Ying Nian Wu

Donatello Telesca

Frederic P. Schoenberg

Mark S. Handcock, Committee Chair

University of California, Los Angeles

2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# ACKNOWLEDGMENTS

I would first like to thank my advisor Mark Handcock for being a patient teacher and mentor. I am also grateful for all the support from the faculty and staff of the Statistics Department. I am thankful for my parents for always allowing me the freedom to do what I want.

I am grateful for Rick Schoenberg's guidance in understanding the wildfire application and finding the right database through his colleague Haiganosh K. Preisler at the Forest Service. I would also like to thank another colleague and former UCLA Statistics alumnus, Kevin Nichols for sharing wildfire data with me. Alexandra Syphard and Jon Keeley have kindly shared their paper data with me and that has been very helpful in pursuing the wildfire application.

Stephanie Pincetl, the head of the Energy Atlas project, has been extremely gracious by allowing me access to the Energy Atlas database, but more importantly, cultivating a discursive environment that probes the right problems; this has allowed me to explore the field of data privacy in a more sensible way. Hannah Gustafson and Dan Cheng of the Energy Atlas team have always been patient in answering all my simplistic questions and guiding me through the database.

Finally, I am thankful for Sasha Voss and Mark Handcock's curiosity that lead them to collect and manually digitize the spider web data almost 2 decades ago, without which the applications in this dissertation would not have been possible.

# VITA

| | |
|---|---|
| 2008–2012 | B.A. Cum Laude (Economics), Boston University |
| 2014–2015 | Statistics Research Assistant, UCLA School of Law |
| 2014–2017 | Teaching Fellow, Department of Statistics, UCLA |
| 2015–2017 | Graduate Consultant, Statistical Consulting Center, UCLA |
| 2016 | Academic Mentor, Institute for Pure and Applied Mathematics RIPS |
| 2016–2017 | Treasurer and Statistics Representative, Mathematics and Physical Sciences Council |
| 2016–2017 | Lead Representative, Statistics Graduate Student Association |
| 2016–2017 | Teaching Assistant Coordinator, Department of Statistics, UCLA |
| 2017 | Research Assistant, UCLA School of Public Affairs |

**Part I**

# Linear Networks

# CHAPTER 1

# Introduction

The main objective of Part 1 of the dissertation is to develop statistical methodology to analyze point events occurring on or near linear networks. While the methodology is being developed on a dataset of spider web positions on mortar lines (see Figure 1.1), another objective is to apply and assess the methodology and its transferability to other areas. Two such areas that could gain from such analysis are street crimes and human related wildfires. Part 1 of the dissertation will demonstrate this methodology and outline the process of applying it to human related wildfires. The ultimate goal is to establish strong statistical methodology that can tackle new issues arising at the common crux of point process and network theories.

## 1.1 Background

While point process theory is well established, spatial analysis of events along networks is not extensive. A large portion of this analysis has been spearheaded by Okabe and collaborators, with applications to boutique openings along Tokyo streets and Acacia tree populations along road networks. The only textbook that addresses this area titled "Spatial Analysis along Networks: Statistical and Computational Method" by Okabe and Sugihara (2012), introduces multiple examples and expands on computational methods appropriate to network processes.

Baddeley's work extends the point process theory to network processes with applications to street crimes in Chicago and spine formation on dendrite networks of neurons. Ang et al. (2012) expand on the (empirical) 'network $K$ function' introduced by Okabe and Yamada,

to incorporate network generality resulting in a more interpretable network $K$ function. The *geometrically corrected* empirical $K$ function weights the shortest path distance between a pair of points $(x_i, x_j)$, with the reciprocal of the number of points on the network situated at the same distance from $x_i$ as $x_j$ is. The corrected network $K$ function is

$$\hat{K}_L(r) = \frac{|L|}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1\{d_L(x_i, x_j) \leq r\}}{m(x_i, d_L(x_i, x_j))} \tag{1.1}$$

where $n$ is the total number of points in the process, $|L|$ is the total length of the network, $d_L(x_i, x_j)$ denotes the shortest path distance and $m(x_i, d_L(x_i, x_j))$ is the number of points on $L$ lying at the exact distance $d_L(x_i, x_j)$ away from point $x_i$. This reciprocal weighting is similar to Ripley's (1977) isotropic edge correction as it considers the number of neighbors on the ball of radius $d_L(x_i, x_j)$ centered on $x_i$. These results are further extended to an 'inhomogeneous' network $K$ function and the pair correlation function. The corrected network $K$ function and the inhomogeneous versions are fitted to a single quadrant of spider data; both reveal only insignificant clustering at larger scales, attributable to spatial inhomogeneity. The authors of Ang et al. (2012) conclude that the spider data does not show evidence of clustering and are consistent with complete spatial randomness. Note that the spider data used for this paper did not include covariate information on the spiderwebs, possibly losing some information. More details about the spider data and covariates are provided in Section 1.3.

Ang et al. (2012) also apply the methods to street crime data from an area close to the University of Chicago. There are 116 points on a linear network of length 9.5km; the covariates on the crime were ignored for this application. The inhomogeneous network $K$ function showed no evidence of clustering; however, the authors believe there is confounding of inhomogeneity and clustering at short distances and suggest further study.

Baddeley et al. (2014a) apply the corrected network $K$ function to spines on dendrite networks of neurons resulting in an extension to multitype points. The data used in the paper is of an observed spine pattern on dendrite network of a rat neuron. At a specific resolution, the spines can be categorized into 3 types based on their size. It is necessary to study the spatial distribution of spines to understand normal function and disease processes; and since

the spines occur only on a dendrite network, event analysis along *linear* networks is a natural choice. Fitting the inhomogeneous network $K$ function did not reveal significant clustering but the authors present numerous caveats to interpreting these results in a biological context. The *in vivo* dendrite network occurs in a 3D space but the observed data was a response of a designed experiment; the *in vitro* neurons are almost flat and the third dimension was ignored in this case. The authors also express a need for more explicit models that can handle such data structures.

## 1.2 Motivation

Thus far, the existing literature presents computational and descriptive methods complementary to point process literature adapted to linear networks. There has been little work that consciously integrates network structures into spatial point pattern analysis. Given the advent of this new data structure, the methodology presented in the dissertation under Chapter 2 is an effort to adapt and integrate spatial analysis and network theory, and produce a compounded modeling framework. The motivation is to harness insights from spatial geometry imposed by network structures to better understand and predict event occurrence.

The spiderweb data provides the locations of webs made by the small *Oecobius navus* in the crevices of mortar lines on brick walls. The indented mortar lines provide the perfect structural support for web building, while the overall network of lines delineate and guide web formation over the whole wall. Initially pursued as a curious application, the spiderweb data has been used as a guide to build a spatial and temporal model of point processes on linear networks (Chapter 2). One can see that this example is analogous to many structural processes observed in multiple fields; examples include events along stream networks, events along Interstate high-tension wires, water and gas pipeline networks, etc.

The two applications that have caught attention that can largely benefit from such modeling, with extensions to prediction, are street crimes and human related wildfires originating on/near road and trail networks. Mapping and predicting 'hot spots' for street crimes and human related wildfire origins could greatly aid police and firefighter efforts on ground.

Crimes in cities follow a categorical, spatial pattern depending on the type of target and crime; the distribution of residential and commercial areas naturally influence crime occurrence and spatial temporal modeling of crimes over city road networks is the logical next step (Hering and Bair (2014)).

In case of wildfires, it has been recorded that more than 95% of California wildfires have human related origins; as most wildfire origins are in non-urban settings, we want to map and understand their interaction with the extensive road and trail network that exists in most park and forest areas of California. With the right and sufficient covariate information, mapping wildfire origins with respect to road and trial networks could reveal high intensity regions that require greater fire management. The methodology presented below is a step towards achieving greater insight into such areas that require a rigorous amalgamation of both point process and spatial network theories. The model application to wildfire origins is presented in Chapter 3.

## 1.3 Spider Data

This data was identified by entomologist Sasha Voss of University of Western Australia and manually digitized by Mark Handcock. The dataset records the *Oecobius navus* spider web positions in the mortar lines of brick walls. The dataset consists of weekly recordings from two different sites ($D$ & $F$), each with two fixed quadrants ($Q1$ & $Q2$), observed over a six week period from May to July 1999. The quadrants were defined to be $1125mm^2$ in size. Each weekly recording captured the position of the webs (all containing spiders, indicated by WA or WJ) and the position of individual spiders (indicated by A or J) to the closest millimeter. The covariate information on the spider was whether it is an Adult (A) or a Juvenile (J) individual (see Figure 1.1). Site $D$ houses a few extra quadrants titled $R1, R2, R3$ and $R4$, recorded over 5 days, that only provide locations with no covariate data; it is likely that these locations are of both webs and individual spiders.

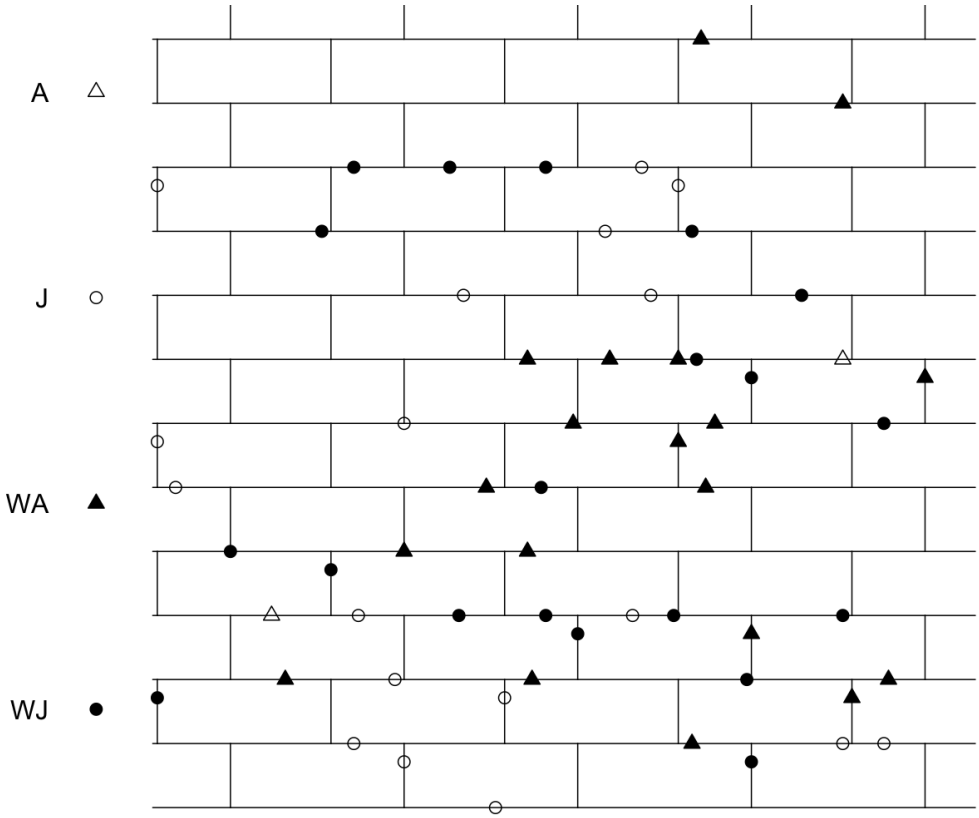


Figure 1.1: Site D, Quadrant Q2 on 10th June 1999.

### 1.3.1 Background on *Oecobius navus*

While the *Oecobius navus* is not extensively studied, the available behavioral traits could aid model specification. This cosmopolitan dweller grows up to 4mm in diameter (see Figure 1.2) and prefers to build webs on heat retaining substrates. These spiders are not known to live communally with each web generally housing only a single spider. Each web also generally contains one or more egg sacs. A disturbance of the web causes the spider to escape in a straight line fashion, often encroaching and displacing another spider from its web. As a result, we cannot attribute webs and egg sacs to specific individuals.

### 1.3.2 Randomization Tests

Permutation tests that map the nearest neighbor covariate distributions were conducted over all the quadrants. After identifying and noting the contingency table of covariates on


Figure 1.2: Scale of *Oecobius navus* shown on an American penny.

the first nearest neighbor from all points, the marks (covariates WJ, WA, J and A) on the points are permuted $n$ times, using a distribution proportional to the original marks. The set of covariates on the first nearest neighbor for each of the $n$ permuted point processes are recorded and plotted to be compared with the original contingency table of nearest neighbor covariates. Overall, the tests do not reveal any significant results with respect to the nearest neighbor covariates. However, there are a few quadrants that revealed significantly higher values of juvenile webs neighboring other juvenile webs. This might be attributable to the fact that there is almost always an egg sac in every web.

Ang et al. (2012)'s inhomogeneous corrected network $K$ function mentioned above was also fit on the spider quadrants using the *spatstat* package in R. As we can see in Figure 1.3, the $K$ function reveals varying levels of clustering, dispersion and even complete randomness with respect to a poisson distribution depending on the quadrant. Figure 1.3 presents four such quadrants and their inhomogeneous network $K$ functions.

## 1.4 Point Processes on Linear Networks

A *point process* is a set of unordered points $\mathbf{x} = x_1, \ldots, x_i, \ldots, x_n$, such that $n \geq 0$ and $x_i \in W$ where $W$ is a spatial window in $d$-dimensional space $\mathbb{R}^d$, $d \geq 1$. The window W is assumed to have finite positive volume $|W|$.

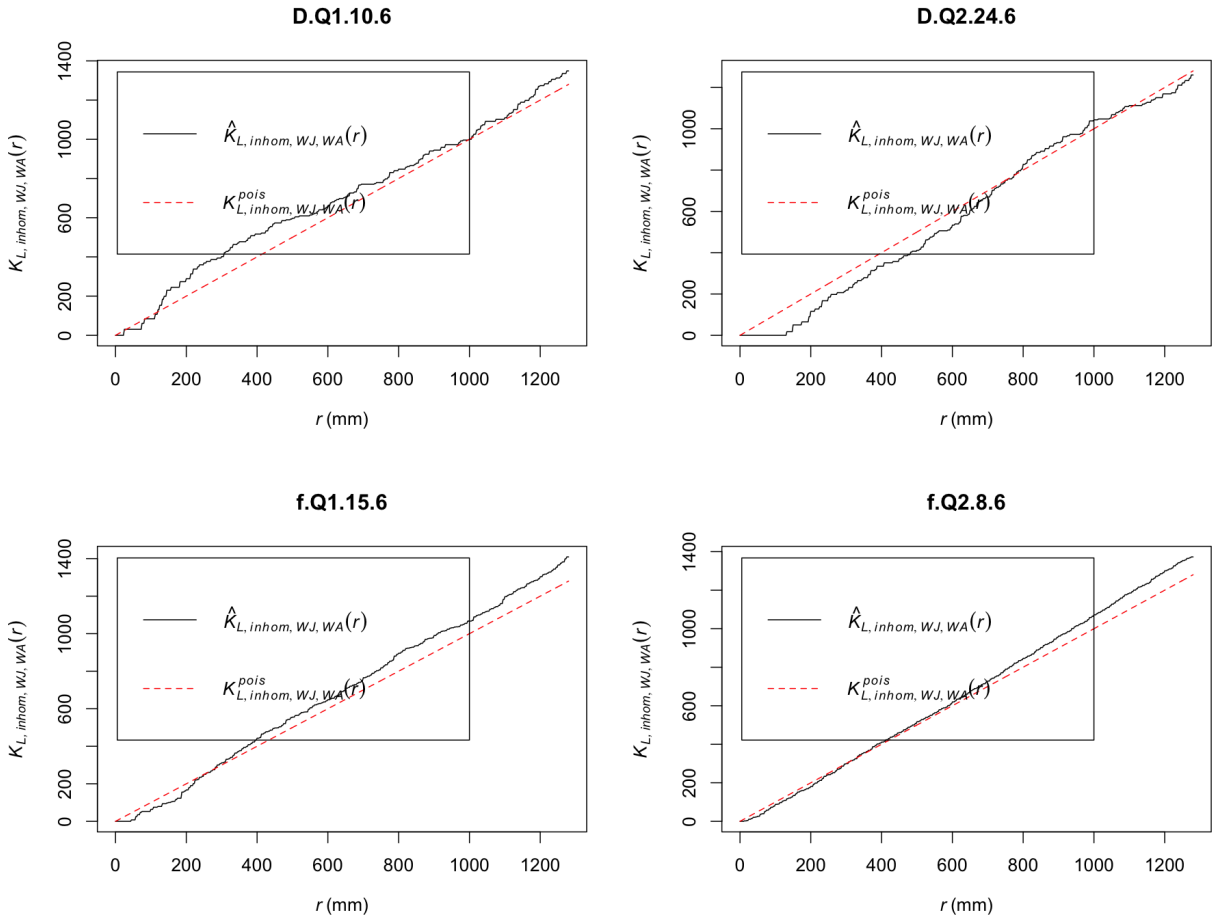


Figure 1.3: Inhomogeneous network $K$ function on 4 unique quadrants

A *line segment* in a plane with endpoints $u$ and $v$ is written as $[u,v] = \{tu + (1-t)v : 0 \leq t \leq 1\}$. A *linear network* $L$ is defined as a union $L = \bigcup_{i=1}^{n} \ell_i$ of a finite collection of line segments $\ell_1, \ldots, \ell_n$ in the plane. The total length of all the line segments in $L$ is denoted by $|L|$. Another way to represent a *linear network* $L$ is by a set of vertices $v_1, \ldots, v_n$ and edges $e_{ij} = [v_i, v_j]$, such that the intersection of two different edges occurs at only one point.

A *point process on a linear network* is a point process that occurs *on* or *near* a linear network, generally in a 2-dimensional space. The occurrence of the point process on/near the linear network could be due to the structural limitations of the point generating process such as road accidents on a street network, or due to a contextual relationship between the point process and the environment it occurs in. For instance, the occurrence of human

settlements near and around river networks demonstrates the synergic interaction between a point process and a pre-existing linear network structure. While the majority of the examples illustrated in this dissertation are physical and social events occurring on a plane, the occurrence of such data structures is not limited to these fields. The occurrence of spines on a neuron dendrite network is one such case.

### 1.4.1 Inhomogeneous Intensity function

The first established step in understanding a point process (on a linear network or otherwise) is by estimating an intensity function (first moment measure) for the points. The intensity function $\lambda(u)$ provides the probability of a point occurring at infinitesimal space; otherwise it is the expected infinitesimal rate of the point in the observed window $W$. Given the covariate information on the spider and web positions, the ideal case is to estimate an inhomogeneous intensity function $\lambda_\theta(u)$ that depends on the covariates to capture the 'spatial trend'. Ideally, this can be done through formulating a likelihood function and estimating the parameters through maximum likelihood. However, formulating and evaluating a likelihood function comprising of individual intensities of the points $P(\mathbf{X}|\Theta)$, where $\mathbf{X} = (x_1, \ldots, x_n)$ is highly intractable. A convenient approximation to this is the pseudolikelihood function that is a product of conditional intensities of each point. The conditional intensity assumes point independence given all other points in the process.

The first step in formulating the pseudolikehood function is to estimate the (Papangelou) conditional intensity at every location $u \in W$ as:

$$\lambda(u; \mathbf{x}) = \frac{f(\mathbf{x} \cup \{u\})}{f(\mathbf{x} \setminus \{u\})} \tag{1.2}$$

Using the covariate information on web locations, we can use a log-linear model to estimate this conditional intensity as:

$$\lambda_\theta(u; \mathbf{x}) = \exp\left(\theta^T S(u; \mathbf{x})\right) \tag{1.3}$$

where $S(u; \mathbf{x})$ is a matrix of spatial and network covariates defined at each point $u$ in $W$, while $\theta \in \Theta$ is the vector of parameters to be estimated. The natural covariates appropriate

for the spider data include euclidean and geodesic distance to the nearest neighbor, covariates on the nearest neighbor, the mortar line network etc.

### 1.4.2 Inference

Given the conditional intensity (1.2), the pseudolikelihood of a point process as defined by Besag (1977) is given by:

$$PL(\Theta; \mathbf{x}) = (\prod_{x_i \in W} \lambda_\theta(x_i; \mathbf{x})) \exp(-\int_W \lambda_\theta(u; \mathbf{x}) du) \tag{1.4}$$

Substituting our log-linear conditional intensity (1.3) into (1.4) gives us the following pseudolikelihood:

$$PL(\Theta; \mathbf{x}) = (\prod_{x_i \in W} \exp(\theta^T S(x_i; \mathbf{x})) \exp(-\int_W \exp(\theta^T S(u; \mathbf{x})) du) \tag{1.5}$$

The numerical solution to the above expression needs iterative algorithms. The estimation method used by Baddeley and Turner (2000) in R's *spatstat* package is an adaptation of a technique introduced by Berman and Turner (1992). The estimation method, with the help of a quadrature rule, discretizes the integral calculation to a weighted poisson likelihood. We can approximate the integral in (1.4) using any quadrature rule,

$$\int_W \lambda_\theta(u; \mathbf{x}) du \approx \sum_{j=1}^{m} \lambda_\theta(u_j; \mathbf{x}) w_j \tag{1.6}$$

where $u_j$, $j = 1, \ldots, m$, are points in $W$ and $w_j > 0$ are quadrature weights summing to $|W|$. This produces an approximation to the log-pseudolikelihood of (1.4) as:

$$\log PL(\Theta; \mathbf{x}) \approx \sum_{i=1}^{n} \log \lambda_\theta(x_i; \mathbf{x}) - \sum_{j=1}^{m} \lambda_\theta(u_j; \mathbf{x}) w_j \tag{1.7}$$

An observation by Berman and Turner (1992) notes that if the list of points $\{u_j, j = 1, \ldots, m\}$ includes all the data points $\{x_i, i = 1, \ldots, n\}$, then (1.7) can be rewritten as

$$\log PL(\Theta; \mathbf{x}) \approx \sum_{j=1}^{m} (y_j \log \lambda_j - \lambda_j) w_j, \tag{1.8}$$

where $\lambda_j = \lambda_\theta(u_j)$ and $y_j = z_j/w_j$, and

$$z_j = \begin{cases} 1 & \text{if } u_j \text{ is a data point, } u_j \in \mathbf{x} \\ 0 & \text{if } u_j \text{ is a dummy point, } u_j \notin \mathbf{x} \end{cases} \tag{1.9}$$

The right side of (1.8) for a fixed $\mathbf{x}$ is formally the log-likelihood of independent Poisson variables $Y_k$ with mean $\lambda_k$, with weights $w_k$. As a result, a GLM software capable of handling weighted likelihoods where the weights do not add up to one, and accepts non-integer response values can be used to estimate the pseudolikehood. However, as the GLM setup assumes independent poisson variables, which is not the case here, the standard errors produced by the software are invalid. As an alternative, Moving block bootstrap standard errors are presented.

### 1.4.3 Moving Block Bootstrap for Spatial Data

As the Papangelou conditional intensity of spider occurrence is spatially dependent, one cannot draw random samples from the spider dataset to produce bootstrap samples. As an alternative, the Moving Block Bootstrap (MBB), introduced independently by Kunsch (1989) and Liu and Singh (1992), is employed here to produce bootstrapped standard errors. The idea of applying moving block bootstrap to spatial data was also introduced by Hall (1985).

Assume that $X_1, X_2, \ldots$ is a sequence of stationary random variables, and let $\mathbf{X_n} = (X_1, \ldots, X_n)$ denote the observations. In the spider case, $\mathbf{X_n}$ represents the complete set of data and dummy points from the wall quadrant. Given the data, blocks of size $l$ are defined as $B_j = (X_j, \ldots, X_{j+l-1})$, $j = 1, \ldots, N$ where $N = n - l + 1$ denotes all the possible overlapping blocks in $\mathbf{X_n}$. The bootstrap blocks are obtained by selecting a random sample of $b$ blocks from the full set of overlapping blocks $\{B_1, \ldots, B_N\}$.

For the standard errors in Table 1.1, the moving block bootstrap method was implemented with block size of about 13,000 data points. The entire set of data and dummy points over a single quadrant is around 14,000 data pints. This results in a set of about 1001 overlapping blocks. Given that the data and dummy points are spatially ordered, 1000 samples of blocks

were drawn, and modeled through the GLM software to obtain the empirical distribution of each of the model coefficients. The sample standard deviation of this empirical distribution provides the moving block bootstrap standard errors of the model coefficients. The 2.5% and 97.5% quantile values of the empirical distributions provides the MBB 95% confidence interval, as seen in Table 1.1.

## 1.5 Application to Spider Data

An example of this log-linear conditional intensity function for quadrant $Q2$ in site $D$, is shown in Table 1.1. The intensity is estimated on the locations of both the spiders and the webs. This was a deliberate choice to understand how the occurrence of spiders affects placement of webs around the quadrant. In general, the presence of juvenile spiders (compared to adult spiders) increases the intensity of spider *and* web presence. The same increase is seen with adult and juvenile occupied webs compared to just adult spider presence. This lends to the idea that juvenile spiders, and webs are relatively more common and stable than the adult spiders; this supports the already known fact that the *Oecobius Navus* spiders are highly mobile. The intensity of webs and juvenile spiders is higher compared to adult spiders over all the three models (Table 1.1); however, beside the first model (with both Euclidean and Geodesic distances), the estimates in the last two models are right on the edge of the MBB 95% confidence interval. This could indicate the absence of strong regressors in these models.

The distance measures (Euclidean and Geodesic) provide a slightly different story. Table 1.1 provides three different intensity estimates with, and without the geodesic, and euclidean distance measures. While both the euclidean and geodesic distances to the closest point negatively affect spider and web occurrence, the geodesic distance effect is much larger. Given the construction of the mortar linear network, web and spider occurrence is heavily dependent on the shortest path distance to the next closest spider or web. Higher the geodesic distance to the next closest point on the quadrant, lower the conditional intensity of a web or spider occurrence. This indicates a certain degree of grouping or radius of activity for the

spiders and web occurrence; a radius of activity that occurs along the mortar line network.

The geodesic distance estimate in the second model is outside the MBB 95% confidence interval, but all the other distance estimates (euclidean and geodesic) in the other models are not significant or on the edge of the 95% CI. This could indicate that while the geodesic distance measure is affecting web and spider occurrence *more* than the euclidean distance measure, the models might be missing other stronger regressors that capture web and spider intensity. Other regressors could include brick substrate temperature, presence of predators and prey, time of the day, etc. Either way, all three models present a much higher geodesic estimate than an euclidean estimate. This indicates that the *Oecobius Navus* spider species' activity *is* motivated by the mortar line network; conceptually, the indented mortar lines not only provide structure for web building and protection from the elements, but also seem to motivate the location of web building in relation with neighboring webs.

The plotted conditional intensities of the second and third models of Table 1.1 are shown in Figure 1.4. The top 20% of the estimated log intensities is plotted in yellow while the lower 80% of the estimated log intensities is plotted in purple. The model with the geodesic distance measure has produced much more spatially concentrated intensities while the model with the euclidean distance measure has high intensity spread out around the region of spider and web presence. The spatial plot for the model with both the euclidean and geodesic distance measures looks very similar to the model plot with just the geodesic distance measure.

The high intensity concentration in the geodesic model plot was explored further by attempting a few alternative ideas. These included categorizing the geodesic distance into levels and log transforming the right skewed geodesic distances. However, both options once again produced highly concentrated intensity estimates over the linear network. While there is evidence that geodesic distance does affect web presence, range of geodesic spider activity should be further explored through localized network related statistics. This could shed light on the concentrated, localized intensities.

**Intensity of D.Q2.10.6 with Euclidean & Geodesic distances**

| Parameter | Coefficient Estimate | MBB SE | MBB 95% CI |
|---|---|---|---|
| (Intercept) | -5.015 | 4.789 | (-27.3, -4.9) |
| Euclidean dist | -0.032 | 0.057 | (-0.128, 1.22e-14) |
| Geodesic dist | -6.782 | 1.498 | (-6.813, 5.34e-15) |
| Juvenile | 2.202 | 1.314 | (-8.37e-13, 2.101) |
| Web Adult | 2.251 | 1.475 | (-8.37e-13, 2.251) |
| Web Juvenile | 2.298 | 1.330 | (-8.37e-13, 2.299) |

**Intensity of D.Q2.10.6 with Geodesic distance**

| Parameter | Coefficient Estimate | MBB SE | MBB 95% CI |
|---|---|---|---|
| (Intercept) | -5.164 | 5.049 | (-27.3, -5.8) |
| Geodesic dist | -6.757 | 1.599 | (-6.55, 2.49e-15) |
| Juvenile | 2.201 | 1.713 | (-8.79e-13, 2.276) |
| Web Adult | 2.251 | 1.826 | (-8.79e-13, 2.537) |
| Web Juvenile | 2.298 | 1.662 | (-8.79e-13, 2.377) |

**Intensity of D.Q2.10.6 with Euclidean distance**

| Parameter | Coefficient Estimate | MBB SE | MBB 95% CI |
|---|---|---|---|
| (Intercept) | -4.604 | 4.707 | (-27.3, -4.92) |
| Euclidean dist | -0.352 | 0.092 | (-0.4, 4.91e-15) |
| Juvenile | 2.201 | 2.344 | (-9.14e-13, 2.317) |
| Web Adult | 2.249 | 2.437 | (-9.14e-13, 2.642) |
| Web Juvenile | 2.304 | 2.052 | (-9.14e-14, 2.50) |

Table 1.1: Intensity function summaries of D.Q2.10.6 quadrant where MBB stands for Moving Block Bootstrap

## 1.6 Conclusion

The intensity of both the spider and web locations is a simple, first step in setting up point process theory on linear networks. It has revealed that despite the spiders' free movement over the brick wall, web construction and activity are delineated to the linear network and their occurrence is dependent on path distances to the nearest neighbors. This sort of structured data that combines point patterns and linear networks is abundant but barely explored. The next couple Chapters extend the current theory to a temporal setting and

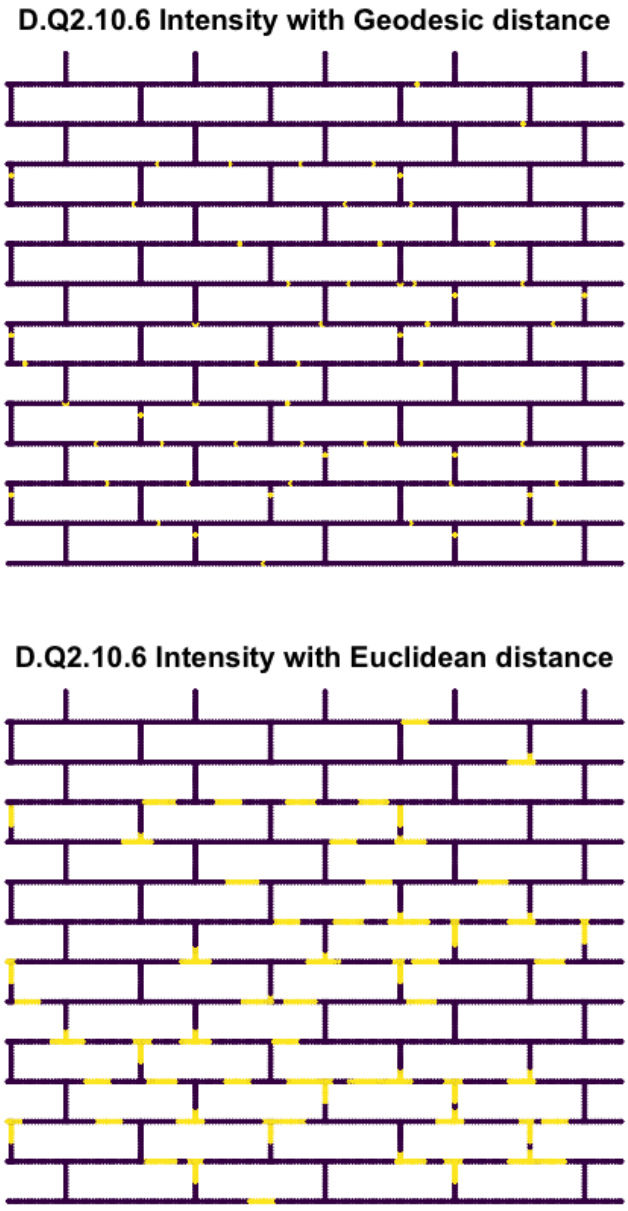


Figure 1.4: Intensity function estimates over the D.Q2.10.6 quadrant with Geodesic and Euclidean distance measures. The Yellow regions indicate the top 20% of the estimated log intensities while the purple indicates the lower 80% of the estimated intensities. The Intensity function with both the Geodesic and Euclidean distance measures is very similar to the one with just the Geodesic distance measure.

apply it to wildfire ignition locations near road networks.

# CHAPTER 2

# Separable Temporal Model on Linear Networks

The spider data presented in Chapter 1 was collected as weekly snapshots, over six weeks. This allows for a natural extension to temporal modeling. The goal of this Chapter is to capture the dynamic action of new and existing points entering and exiting the process. The spider data is apt for such modeling as new webs are built while existing webs persist and disintegrate over time. The primary concept is to estimate separable *formation* and *dissolution* intensity functions that capture the change in the point process over a single time step. Conditioned on time step $t$, the product of the *formation* and *dissolution* models estimates the point process at time step $t+1$. The model can be extended to be conditioned on all the previous time steps, allowing for generalized *formation* and *dissolution* models that could be used for prediction.

The impetus for this separable temporal model was gained from the Separable Temporal Exponential Random Graph model (STERGM) framework for dynamic networks developed by Krivitsky and Handcock (2014). They present separable *formation* and *dissolution* models that estimate tie formation and breakage between nodes over time.

## 2.1 Motivation for the Separable Model

The addition of a linear network to a point process space is a crucial aspect of the Separable Temporal Linear Point Process (STLPPM) model presented in this Chapter. While this model is only relevant for point patterns that occur on or near linear networks, the network's delineation of events is important in supplying new information to understand the point pattern. For example, wildfires that have human-related causes can be modeled with respect

to existing road networks to understand how human movement and activity lead to wildfire ignitions. The density of roads and building structures could determine areas with higher intensity of human-caused wildfire ignitions.

The motivation for a separable modeling framework for point processes is to understand how the underlying social or geological factors propel a point process. Simultaneously, the separable model captures the factors' effects on the point's persistence and dissolution over time. In social/geological processes, some factors might affect event *formation* more than others and similarly with event *dissolution*. While the formation and dissolution processes might interact over the long run, for computational and interpretive advantages, we assume formation and dissolution separability over a time step, conditioned on the previous time step. Given the state of the process in the previous time step, the formation and dissolution processes are conditionally independent allowing us to model them with a different set of sufficient statistics.

For example, given the state of the spider webs in the previous step, the probability of a new web forming is independent of the web dissolution process taking place on the same plane. The formation probability is also *conditionally* independent of all events(webs) around it in the current time step.

## 2.2 Model Specification

Given an observed point process on a linear network $\mathbf{x}^t$ at time step $t$, the *formation point process* $\mathbf{X}^+$ includes the new points along with the existing points at $t$ and the *dissolution point process* $\mathbf{X}^-$ includes only the points that persisted on from time $t$. The realized counterparts of these processes are $\mathbf{x}^+$ and $\mathbf{x}^-$. Given $\mathbf{x}^+$, $\mathbf{x}^-$ and $\mathbf{x}^t$, the point process $\mathbf{x}^{t+1}$ can be evaluated as follows:

$$\mathbf{x}^{t+1} = \mathbf{x}^- \cup (\mathbf{x}^+ \backslash \mathbf{x}^t) \tag{2.1}$$

where $\mathbf{x}^+ = \mathbf{x}^t \cup \mathbf{x}^{t+1}$ and $\mathbf{x}^- = \mathbf{x}^t \cap \mathbf{x}^{t+1}$. If $\mathbf{X}^+$ is conditionally independent of $\mathbf{X}^-$ given $\mathbf{X}^t$ then

$$P(\mathbf{X}^{t+1} = \mathbf{x}^{t+1} | \mathbf{X}^t = \mathbf{x}^t; \boldsymbol{\theta}) = P(\mathbf{X}^+ = \mathbf{x}^+ | \mathbf{X}^t = \mathbf{x}^t; \boldsymbol{\theta}) P(\mathbf{X}^- = \mathbf{x}^- | \mathbf{X}^t = \mathbf{x}^t; \boldsymbol{\theta}) \tag{2.2}$$

Here *separability* is defined as $\mathbf{X}^+$ being conditionally independent of $\mathbf{X}^-$ given $\mathbf{X}^t$, such that the parameter space $\boldsymbol{\theta}$ is a product of the individual parameter spaces $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$. This indicates that the point formation process is (conditionally) independent of the point dissolution process given the point process at the previous time step; the formation and the dissolution processes do not interact, once the point process at the beginning of the time step is observed.

Individually, the web *formation* model is:

$$P(\mathbf{x}^+ | \mathbf{x}^t) = \prod_{u \in W} \lambda(u; \mathbf{x}^+) = \prod_{u \in W} \exp\left\{(\boldsymbol{\theta}^+)^T S^+(u; \mathbf{x}^+, \mathbf{x}^t)\right\} \tag{2.3}$$

and the web *dissolution* model is:

$$P(\mathbf{x}^- | \mathbf{x}^t) = \prod_{u \in W} \lambda(u; \mathbf{x}^-) = \prod_{u \in W} \exp\left\{(\boldsymbol{\theta}^-)^T S^-(u; \mathbf{x}^-, \mathbf{x}^t)\right\} \tag{2.4}$$

where $u$ indicates every point on the whole window $W$ on which the linear point process was observed. The conditional intensities of the *formation* and *dissolution* processes are estimated using log-linear models(as seen in section 1.4.2), with $S^+(u; \mathbf{x}^+, \mathbf{x}^t)$ and $S^-(u; \mathbf{x}^-, \mathbf{x}^t)$ representing the sufficient statistics, based on $\mathbf{X}^t = \mathbf{x}^t$, which attempt to capture the tenacity and degradation of webs over time.

## 2.3 Pseudolikelihood Inference

Using equations (2.3) and (2.4), the conditional intensity for the separable model (2.2) can be written as follows:

$$\begin{aligned}\lambda_\theta(\mathbf{x}^{t+1}|\mathbf{x}^t) &= \prod_{u\in W}\lambda(u;\mathbf{x}^+)\prod_{u\in W}\lambda(u;\mathbf{x}^-) \qquad (2.5)\\ &= \prod_{u\in W}\exp\{(\boldsymbol{\theta}^+)^T S^+(u;\mathbf{x}^+,\mathbf{x}^t)\}\prod_{u\in W}\exp\{(\boldsymbol{\theta}^-)^T S^-(u;\mathbf{x}^-,\mathbf{x}^t)\}\\ &= \exp\{\sum_{u\in W}((\boldsymbol{\theta}^+)^T S^+(u;\mathbf{x}^+,\mathbf{x}^t))\}\exp\{\sum_{u\in W}((\boldsymbol{\theta}^-)^T S^-(u;\mathbf{x}^-,\mathbf{x}^t))\}\\ &= \exp\{\sum_{u\in W}((\boldsymbol{\theta}^+)^T S^+(u;\mathbf{x}^+,\mathbf{x}^t)) + \sum_{u\in W}((\boldsymbol{\theta}^-)^T S^-(u;\mathbf{x}^-,\mathbf{x}^t))\} \qquad (2.6)\end{aligned}$$

Given this model specification, the pseudolikelihood function of the separable temporal linear point process model (using Besag's pseudolikelihood definition) is

$$PL(\Theta;\mathbf{x}^{t+1}) = \Big(\prod_{x_i\in W}\lambda_\theta(\mathbf{x}^{t+1}|\mathbf{x}^t)\Big)\exp\Big(-\int_W \lambda_\theta(\mathbf{x}^{t+1}|\mathbf{x}^t)du\Big) \qquad (2.7)$$

$$\log PL(\Theta;\mathbf{x}^{t+1}) = \sum_{i=1}^{n}\log\lambda_\theta(\mathbf{x}_i^{t+1}|\mathbf{x}^t) - \int_W \lambda_\theta(\mathbf{x}^{t+1}|\mathbf{x}^t) \qquad (2.8)$$

The integral in (2.8) can be approximated using a finite sum quadrature rule, as shown in section 1.4.2:

$$\int_W \lambda_\theta(u|\mathbf{x}^t,\boldsymbol{\theta})du \approx \sum_{j=1}^{m}\lambda_\theta(u_j|\mathbf{x}^t,\boldsymbol{\theta})w_j \qquad (2.9)$$

where $u_j, j = 1,\ldots,m$, are points in $W$ and $w_j > 0$ are quadrature weights summing to $|W|$. This produces an approximation to the log-pseudolikelihood in (2.8) as:

$$\log PL(\Theta;\mathbf{x}^{t+1}) \approx \sum_{i=1}^{n}\log\lambda_\theta(\mathbf{x}_i^{t+1}|\mathbf{x}^t) - \sum_{j=1}^{m}\lambda_\theta(u_j|\mathbf{x}^t,\boldsymbol{\theta})w_j \qquad (2.10)$$

Consider a list of points $\{u_j, j = 1,\ldots,m\}$ that include all the data points $\{x_i, i = 1,\ldots,n\}$, then (2.10) can be rewritten as:

$$\log PL(\Theta;\mathbf{x}^{t+1}) \approx \sum_{j=1}^{m}(y_j\log\lambda_j - \lambda_j)w_j, \qquad (2.11)$$

where $\lambda_j = \lambda_\theta(u_j)$ and $y_j = z_j/w_j$, and

$$z_j = \begin{cases} 1 & \text{if } u_j \text{ is a data point, } u_j \in \mathbf{x} \\ 0 & \text{if } u_j \text{ is a dummy point, } u_j \notin \mathbf{x} \end{cases} \tag{2.12}$$

The expression in (2.11) is similar to the log-likelihood of independent poisson variables $Y_k$ with means $\lambda_k$ taken with weights $w_k$. Once again, similar to what was implemneted in Chapter 1, a GLM framework software that is capable of handling weighted likelihoods and non-integer variable was used to fit the STLPPM framework. Separate, non-parametric moving block standard errors are estimated for each of the covariates. Given (six weeks of) data and dummy points of size 100000, and a block size of around 93000, there are about 7000 possible overlapping blocks to sample from. Keeping the data and dummy points spatially and then, temporally consecutive, 1000 samples were drawn and the separable model was rerun on the 1000 blocks. The standard errors, the 2.5%, and 95.5% quantile values of the coefficient empirical distributions are presented in Table 2.1. More details and some alternatives are considered in Section 3.4.2.

## 2.4 Separable Temporal Modeling of the Spider Data

This section presents estimates of the *formation* and *dissolution* models on the second quadrant of site D, referred to as D.Q2 from here onwards. Quadrant D.Q2 was observed over six weeks from 27th May 1999 to 2nd July 1999. As mentioned before, each web is occupied by either an adult or juvenile spider and free standing adult and juvenile spider locations were also recorded. However, since these cosmopolitan spiders are extremely mobile, the model below is estimated on just the web locations.

For both the *formation* and *dissolution* models, first the $\mathbf{x}^+$ and $\mathbf{x}^-$ processes were compiled. The *formation* process is a union of all the points that occurred in time steps $t$ and $t+1$. The *dissolution* process is the intersection of all the points that occurred in time steps $t$ and $t+1$. This way, given the six weeks of data on quadrant D.Q2, five *formation* and *dissolution* processes were compiled.

Aside from the geodesic distance (closest path distance to the next nearest web) as one of the covariates, a *persistence* covariate was introduced in the *dissolution* model of quadrant D.Q2. As the *Oecobius Navus* spider webs survive for periods longer than a week, the weekly quadrant records captured the same web over multiple weeks. However, the web occupants (Adult or Juvenile spiders) changed over the weeks. As a result, the *persistence* covariate captures if a web is new, or has existed from the previous time step as is, or if the web's occupant switched over the time step. These levels are labeled as new, exist and switch in Table 2.1.

Since not much is known about web building by the *Oecobius Navus* spiders, the *formation* model was kept simple with just the geodesic path distance and web's spider factor as the two covariates.

**Formation model Intensity of D.Q2 quadrant**

| Parameter | Coefficient Estimate | MBB SE | MBB 95% CI |
|---|---|---|---|
| (Intercept) | -4.145 | 2.684 | (-27.302, -27.302) |
| Web Juvenile | 0.054 | 0.425 | (-2.52e-12, 2.71e-12) |
| Geodesic Distance | -5.212 | 0.574 | (-1.34e-14, 1.28e-14) |

**Dissolution model Intensity of D.Q2 quadrant**

| Parameter | Coefficient Estimate | MBB SE | MBB 95% CI |
|---|---|---|---|
| (Intercept) | -3.029 | 2.094 | (-27.302, -27.302) |
| Web Adult.switch | -18.198 | 1.559 | (-5.89e-12, 5.95e-12) |
| Web Adult.new | -18.191 | 1.558 | (-5.89e-12, 5.69e-12) |
| Web Juvenile.exist | 0.041 | 0.035 | (-5.59e-12, 6.04e-12) |
| Web Juvenile.switch | -18.191 | 1.558 | (-5.89e-12, 5.89e-12) |
| Web Juvenile.new | -18.206 | 1.560 | (-5.89e-12, 5.80e-12) |
| Geodesic Distance | -5.104 | 0.427 | (-1.08e-14, 9.19e-15) |

Table 2.1: Formation and Dissolution intensities of D.Q2 quadrant where MBB stands for Moving Block Bootstrap

## 2.5 Interpretation

The separable temporal model was set up to understand the formation and dissolution processes of a point pattern separately. The idea is to gain insight into factors that allow a new point event to occur or an existing point event to persist over time.

### 2.5.1 Formation Model

The idea of the formation model is to estimate the intensity of a new event occurring at a point on a window given certain covariates. The formation model intensity on quadrant D.Q2 provides the conditional intensity of a new web forming given the shortest path distance to the next closest web and the spider covariate of a web at that point from the previous time step.

The formation model estimates, seen in Table 2.1 indicate that the geodesic distance negatively effects new web formation over time. The larger the geodesic path distance to the next closest web, the lower the conditional intensity of a new web forming at that point. This grouping behavior was also seen in the simple conditional intensity function estimated over a single quadrant in Chapter 1. This indicates that new web formation occurs within a certain radius of existing webs, and the formation activity is along the network lines; this could be due to juvenile spiders from existing webs moving out to build new webs nearby.

Another factor affecting new web formation is whether the specific point was previously occupied by an adult or juvenile spider. The conditional intensity of a new web forming is slightly higher for a juvenile spider than an adult spider. It is known that almost every web contains an egg sac (Cobb (1994)), and this could lead to higher number of juvenile spiders moving out to build new webs. The formation model could reveal more with an entomologist's insight; better field knowledge could also aid in choosing more appropriate formation model covariates such as presence of predators and prey etc.

Both the covariate estimates are well outside the Moving block bootstrap 95% confidence intervals, as seen in Table 2.1. However, the current formation model could possibly be

improved upon by adding stronger regressors.

### 2.5.2 Dissolution Model

The idea of the dissolution model is to understand the process of point event persistence rather than estimate the exact time of point event dissolution. Another interpretation of the dissolution model is that it estimates factors that affect the length and lifetime of a point event rather than estimate the lifetime of the point event itself. The conditional intensity of the dissolution model estimates the likelihood of a point event surviving until the next time step.

The coefficient estimates, seen in Table 2.1, reveal that geodesic path distance to the next closest web is still negatively affecting the probability of a web surviving beyond the observed time. This indicates that the farther an existing web is to a nearest web, the lower its likelihood of surviving to the next time step. This reiterates that there is an active radius of web building close to pre-existing webs.

The more revealing covariate is the *persistence* factor with Adult-occupied existing web as the reference level. The likelihood of an adult-occupied web persisting is much higher than for an adult-occupied web that used to be a juvenile-occupied web. Similarly, an existing adult-occupied web has a higher persistence probability than a newly formed web with an adult. This could speak to the fact that similar spider species are known to encroach on fellow spider webs and occupy them (Cobb (1994)).

For the juvenile-occupied webs, the likelihood of persistence for an existing juvenile-occupied web is barely larger than it is for an existing adult-occupied web. Once again, juvenile-occupied webs that switched from adult occupancy have a very low probability of surviving onto the next time step. The same goes for newly formed juvenile-occupied webs, when compared to existing adult-occupied webs.

Generally, the dissolution model indicates that the pre-existing webs are highly likely to persist, with juvenile-occupied webs lasting slightly longer than adult-occupied webs. Webs that experience spider switching have much lower life spans. Finally, majority of the webs

seem to last for the larger part of the six-week span of data collection, indicating a need to expand the time span, and also the length of the time step. Alternative model covariates (such as weather and predators) might also improve insight into what allows a web to persist for longer periods.

Once again, all the covariate estimates are well outside the MBB 95% Confidence intervals, indicating that web occupation and occupant switching are higly indicative of web life span in these spider species.

## 2.6 Conclusion

The separable temporal point process model on linear networks was set up to understand the progression and dissolution of point patterns distinctly occurring on linear networks. While the spider web location data on mortar lines is an interesting application, the model could gain from stronger, more prescient applications. Wildfire ignition locations near road networks is one such application. The model setup of conditional independence between the formation and dissolution processes is an easy first step but depending on further applications, this assumption could be changed.

Moreover, since the separable model is estimated using maximum pseudolikehood, it is lacking in model assessment methods. Some possible options for model assessment are explored in Chapter 3.

# CHAPTER 3

# Applications to Wildfire Origins

## 3.1 Motivation

One of the most pressing, large-scale disasters affecting the state of California and other parts of the country are wildfires. California is specifically prone to high-intensity fires due to its dry summers and vegetation. Southern California also faces high speed winds called the Santa Ana winds, that originate inland and bring hot, dry air with them, which further create critical fire weather conditions. The California Wildland Fire Coordinating Group reports that "approximately 95% of all wildfires in California have human-related origins". A paper by Balch et al. (2017), analyzing 1.5 million government records of wildfires between 1992 to 2012, reports that about 84% of all US wildfires have human-related origins and humans are also extending the length of fire seasons. Human-related origins of wildfires include equipment use, vehicular fire, campfire, debris burning, smoking, children and arson.

Where ever a human-related wildfire originates, an individual needed to have access to the area, most likely through a network of roads, paths and trails. Moreover, there has been a rapid expansion of wildland-urban interface (WUI) in the country, increasing the chances of harmful human activity near fuel rich areas (fuel rich indicates low moisture vegetation). Given the high frequency of human-related ignitions and growing wildland-urban interface (WUI) in the country, it is useful to understand how human density and movement affect the spatial and temporal distributions of wildfire ignitions. The STLPPM model framework introduced above and tested on spider webs is appropriate to model wildfire ignitions with respect to road and trail networks. It would be immensely useful if we could identity wildfire ignition hot spots highly correlated with human foot and vehicular traffic.

## 3.2 Current Literature

To forecast and monitor wildfire ignition hazard, the United States National Fire Danger System (NFDRS), established in 1972, produces several indices based on fuel type, weather, topography and other fire rise conditions. The Fire Spread Component (how fast the fire spreads), the Energy Release Component (the energy produced by the fire), and Ignition Component are three components that contribute to the more commonly used Burning Index (difficulty of control). The components themselves are based on meteorological and environmental factors. The Burning Index is used to predict wildfire hazard on a day to day basis and is extensively used by fire managers to make decisions about resource allocation, preparedness, and warnings. Schoenberg et al. (2007) show that the weather variables, that the Burning Index is composed of, are better predictors of short-term wildfire hazard and activity than the Burning Index itself. These weather variables include wind speed, relative humidity, precipitation and temperature. The analysis in Schoenberg et al. (2007) was conducted on the origin times and *centroid* locations of 592 wildfires in the Los Angeles county recorded between January 1976 and December 2000. The area burnt is added as a mark on each fire. The paper uses daily weather data collected by the 16 RAWS(Remote Automated Weather Stations) located across the Los Angeles county. The RAWS were set up by the United States Forest Service in 1976 and each station collects various weather data and sends daily summaries to a central data center for archival.

The first model in Schoenberg et al. (2007) is a simple spatial, temporal conditional intensity function fit using the daily Burning Index, a spatial background rate estimated by kernel smoothing the 1950-1975 fires and a distribution of wildfire areas of the 1976-2000 fires. This model is compared to a series of alternative models that use the RAWS variables–relative humidity, precipitation, temperature and wind speed–directly in estimating the conditional intensity function. The comparison shows that the simple point process models perform better in predicting wildfires than the Burning Index itself. The paper concludes that while the weather variables are strong indicators of wildfire occurrence, many other variables such as fuel age, fuel moisture, vegetation, land use and other human interaction variables could

improve model performance.

The Xu and Schoenberg (2011) paper is an extension of the above paper, where the authors refine the point process models and include new variables such as burn area, fuel age and wind direction. The wildfire data analyzed here is the same as the one described above. Once again, it is important to note that the papers model the *centroid* locations of the polygon burn area, and *not* the ignition locations of the wildfires. Before model fitting, the daily meteorological variables, from the 16 RAWS stations across the Los Angeles county, are smoothed using kernel methods. Kernel regression was used to temporally and spatially smooth the relative humidity, precipitation, wind speed and temperature variables and directional kernel regression (Xu et al. (2011)) was used to smooth wind direction.

The baseline model that the authors begin with depends only on the season associated with time $t$, and the background rate $m(x, y)$ of wildfires for the location in question. This simple reference model is

$$\lambda(t, x, y) = \gamma m(x, y) + \alpha S(t), \tag{3.1}$$

where $\gamma$ and $\alpha$ are parameters to be estimated in model fitting. The background rate $m(x, y)$ was estimated using the 1950-1975 wildfires instead of the 1976-2000 wildfires to avoid overfitting. Similarly, the seasonal rate $S(t)$ is estimated using the times of wildfires during the previous years. Standard kernel density estimators are used to estimate both the trends.

$$m(x, y) = \frac{1}{n_0 \beta_m} \sum_{j=1}^{n_0} K\left(\frac{(||(x, y) - (x_j, y_j)||}{\beta_m}\right) \tag{3.2}$$

$$S(t) = \frac{1}{n_0 \beta_t} \sum_{j=1}^{n_0} K\left(\frac{T^*(t) - T^*(t_j)}{\beta_t}\right) \tag{3.3}$$

$K$ is the kernel function, $\beta_m$ and $\beta_t$ are the bandwidths estimated during model fitting, $(x_j, y_j)$ and $t_j$ indicate the $j$th wildfire coordinate and time between 1950-1975, and $n_0$ is the number of observed wildfires. $||(x, y) - (x_j, y_j)||$ is the Euclidian distance and $T^*(t)$ represents the number of days since the beginning of the year to time $t$. For $S(t)$, $K$ is

a wrapped kernel function allowing January 1 and December 31 to be treated as one day apart.

The subsequent models presented in Xu and Schoenberg (2011) are extensions to the baseline model presented in equation (3.1) above. Model 1(shown in equation 3.4) adds the interpolated Burning Index (available only at RAWS locations) as a term to the intensity function. Model 2(shown in equation 3.5) adds spatial interpolation of temperature, relative humidity, wind speed, and precipitation at time $t$ to the intensity function. The smoothing kernels in Model 2 are also weighted by the wildfire burn area to allow for nonlinear dependence of burn area on weather variables. Model 3 extends Model 2 by adding a directional kernel term for wind direction. Finally, Model 4 incorporates fuel age as a final term. Fuel age is the time since the location's last recorded burn, has an nonlinear relationship with burn area.

$$\lambda_1(t,x,y) = \gamma m(x,y) + \alpha S(t) + \mu_{BI} B(t,x,y) \tag{3.4}$$

$$\lambda_2(t,x,y) = \gamma m(x,y) + \alpha S(t) + F(t,x,y) \tag{3.5}$$

where $B(t,x,y)$ is a kernel smoothed Burning Index and $F(t,x,y)$ represents the distance-weighted kernel smoothing of temperature, ralative humidity, directed wind speed and precipitation.

The authors conclude that while the model with the Burning Index fits better than the baseline model, all other models using weather variables directly fit better than the Burning Index model. While the major goal of the paper was to assess the Burning Index and its short term, forecasting effectiveness of wildfire hazard, authors acknowledge that the models could be improved further by incorporating factors such as vegetation, land use and public policy. Moreover, long-term temporal trends could be incorporated by allowing yearly seasonal components to change.

The final paper summarized here is one by Syphard and Keeley (2015), and explores wildfire *ignition* causes. So far, the two papers by Schoenberg et al. (2007) and Xu and Schoenberg (2011) only analyze the *centroid* locations of wildfires' burn areas and no mention of the wildires' *origin* or *ignition* locations is made. However, from Balch et al. (2017) and

California Wildland Fire Coordinating Group(CWCG), we know that approximately 95% of all wildfires in California are caused by humans. Syphard and Keeley (2015) is a paper that explores the issues of wildfire ignition causes, and how they affect the spatial and temporal distribution of wildfires.

The paper begins with the goal of identifying whether different ignition sources cause distinct spatial or intra-annual temporal patterns of wildfires. The idea is to aid adoption of specific ignition prevention programmes optimized to target high frequency and potent ignition causes. Syphard and Keeley (2015) focus on wildfires in two sub-regions including the Santa Monica Mountains in Ventura and Los Angeles county, and the western portion of San Diego county that falls within the South Coast Ecoregion in southern California. These regions are not only topographically and biologically diverse, but 95% of all ignitions here are caused by humans.

The analysis is on wildfire ignitions from 2006 to 2010, and variables included are elevation, slope gradient, Calveg vegetation data and fuel age. Moreover, as human-caused ignitions occur frequently near transportation corridors, the authors included interpolated maps of distance to roads and residential structures to capture human and housing density. The housing density is an important factor because with the expansion of Wildland-Urban Interface(WUI), the number and areal extent of human ignitions has escalated over the years. Wildland-urban Interface is defined as the area where structures and other human development meet and intermingle with undeveloped wildland (Radeloff et al. (2005)).

For the statistical analysis, Syphard and Keeley (2015) use a Maximum Entropy model to estimate probability of ignition given the wildfire ignition locations and various meteorological and human variables. The authors fit a separate model for each ignition type and conclude that while equipment fires were the most frequent, arson and power line ignitions caused the most area burned in the Santa Monica Mountains region. It is also stated that distance to roads and structures predominantly affect ignition, especially ignition causes such as arson, equipment use, playing with fire, smoking and vehicular fire. As the modeling was ignition cause specific, 30% of the ignitions labeled unknown or miscellaneous were deleted.

Maximum Entropy is a presence-only, statistical algorithm that iteratively minimizes the relative entropy between probability densities of explanatory variables at ignition locations vs. locations randomly distributed across the sub-region. To estimate the probability of wildfire ignition, given the environmental factors, the MaxEnt framework uses the Bayes' rule and the ratio of presence and background rate:

$$Pr(y = 1|\mathbf{z}) = \frac{f_1(\mathbf{z})Pr(y = 1)}{f(\mathbf{z})}, \tag{3.6}$$

where $Pr(y = 1)$ is the marginal probability of wildfire ignition, $f_1(\mathbf{z})$ is the marginal distribution of presence only (ignition location) environmental factors and $f(\mathbf{z})$ is the distribution of environmental factors over the entire region. Syphard and Keeley (2015) do not adopt a specific value for $Pr(y = 1)$ and only model the log of the density ratio as:

$$\begin{aligned} \log \frac{f_1(\mathbf{z})}{f(\mathbf{z})} &= \eta(\mathbf{z}) \\ f_1(\mathbf{z}) &= f(\mathbf{z})e^{\eta(\mathbf{z})}, \end{aligned} \tag{3.7}$$

where $\eta(\mathbf{z}) = \alpha + \beta h(\mathbf{z})$. Here, $f(\mathbf{z})$, the background variation of all the environmental factors, is estimated through a random sample of 10,000 points, and separate MaxEnt models are fit for each ignition type. The paper presents variable importance, as a function of information gain, as the final results.

The Maximum Entropy model framework is contentious among Statisticians due to its assumption of identifiability of prevalence. It is also a bit *adhoc* and does not provide a framework to directly model wildfire ignition probabilities, unlike logistic regression models. In addition, the authors make an erroneous assumption that non-ignition points are *not* true absence points, treating ignition locations as presence-only data. This results in a model that only provides ordinal values of variable importance and does not explain ignition likelihood directly. The model assumptions and results are further discussed in relation to the separable model results presented in Section 3.5.

## 3.3 Data Overview

### 3.3.1 Road network

To aid comparison with Syphard and Keeley (2015), the Santa Monica Mountains region is chosen to model the wildfire origins with respect to the existing road and trail network. The road and trail network was extracted using OpenStreetMap (`https://www.openstreetmap.org`), an open source website providing downloadable map data from all over the world. The raws vector map object of the Santa Monica Mountains region, within a specified bounding box, is downloaded from OpenStreetMap and preprocessed to a R compatible spatial object through the R package '*osmar*'. The bounding box of the extracted network is $[-119.079888, -118.44189]$ for longitude and $[33.99637, 34.17661]$ for latitude. The linear network that will be used from here on to analyze wildfire origin point processes is shown below:

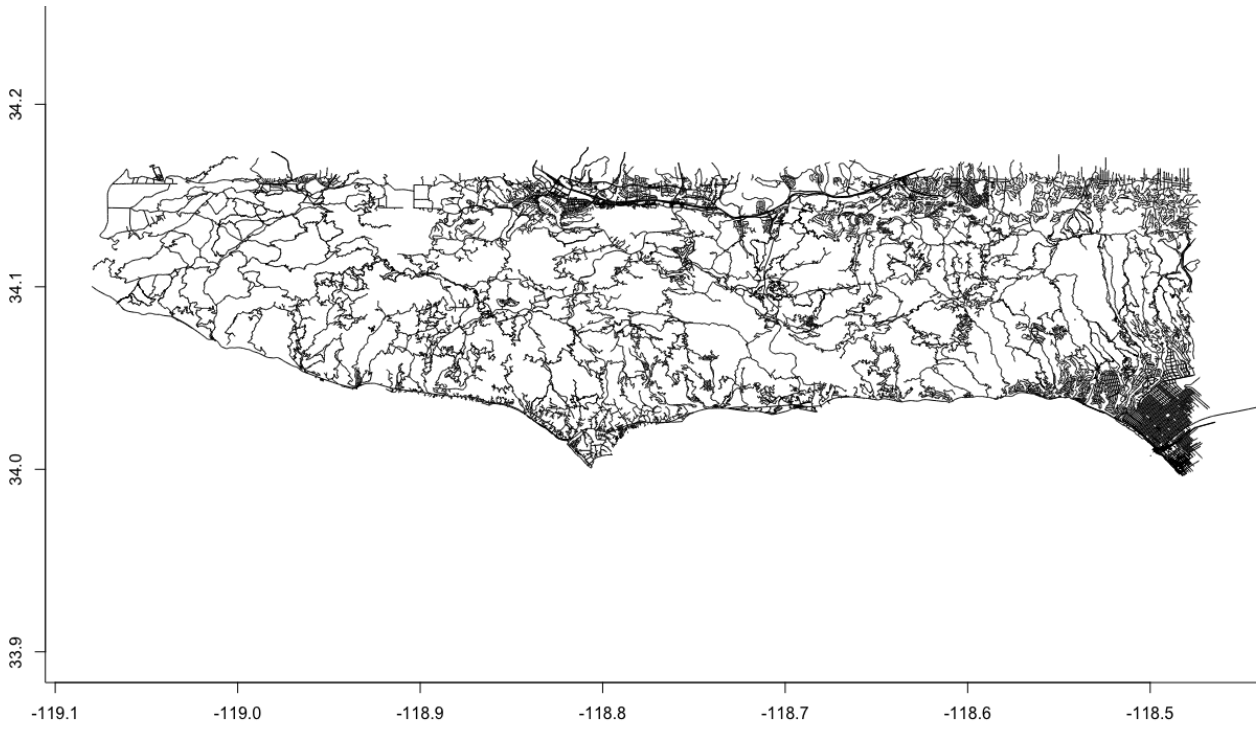


Figure 3.1: Linear Network in the Santa Monica Mountains Region

The OpenStreetMap(osm) object also provides a road tag for each of the ways, and

some of the 22 roadtags include footway, bridleway, motorway, service road etc. As the road categorizations are too specific, they have been consolidated into 5 broad categories depending on their traffic and foot activity, and whether they are commercial or residential. The 5 road categories are in Table 3.1 and the map is in Figure 3.2. And finally, the percentage of the linear network comprised by each road type is shown in Table 3.2, with paths and then residential road types making up most of the linear network at 45% and 30% respectively.

| Road Category | OSM Road tag |
|---|---|
| Residential | living street + residential |
| Path | cycleway + footway + path + bridleway + steps + track + pedestrian |
| Road1 | motorway + motorway link + trunk + trunk link + primary + primary link |
| Road2 | secondary + secondary link + tertiary + tertiary link + service |
| Unknown | road + unclassified |

Table 3.1: Road categorization

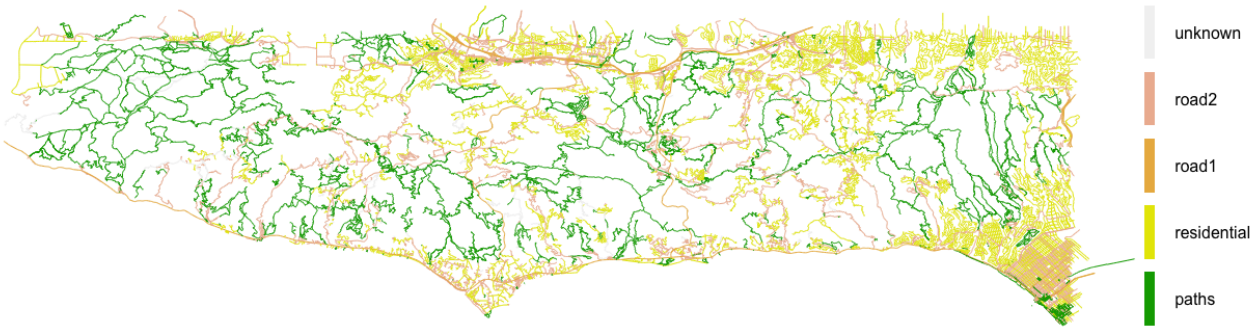


Figure 3.2: Road type in the Santa Monica Mountains Region

### 3.3.2 Wildfire ignition data

The wildfire data analyzed in this dissertation is from the database, 'Spatial wildfire occurrence data for the United States, 1992-2013' (Short (2015)), compiled and cleaned by Karen C. Short of the United States Forest Service. The data publication contains data on

| Road Category | Percentage of Network |
|---|---|
| Residential | 29.5 |
| Path | 45.9 |
| Road1 | 1.9 |
| Road2 | 19.7 |
| Unknown | 2.8 |

Table 3.2: Road type breakdown in the Santa Monica Mountains Region

all wildfires that occurred in the United States from 1992 to 2013. Newer editions of the database extend it to year 2015. The wildfire records were acquired from reporting systems of federal, state, and local fire organizations. The main variables include discovery data, final fire size, ignition cause and a point location at least as precise as the Public Land Survey System (PLSS) section (1-square mile grid). The resulting database includes 1.73 million geo-referenced wildfire records, representing a total of 126 million acres burned during the 22-year period (Short (2014)).

The wildfire dataset was subsetted to just the Santa Monica Mountains region, resulting in 229 wildfires between 1992 and 2013. Due to meteorological data limitations, all following analyses will be on wildfires between year 2000 and 2013. This resulted in 175 wildfires in the Santa Monica Mountains region. The ignition cause breakdown for these 175 fires is shown in Table 3.3. While 69 out of the 175 fires were labelled unidentified, human-related causes like equipment use, and campfire contribute their fair share in terms of acres burned. Despite the availability, the ignition cause variable is *not* used in model fitting as the main goal of the linear networks methodology is to account for human activity, and the model's performance is best judged without the cause variable.

The 175 wildfires are projected on the linear road network introduced in Section 3.3.1 and the spatial map is shown in Figure 3.4. As the *formation* model needs the point process to fall *on* the linear network, the wildfire origin locations are projected to the nearest linear network path and the projected wildfires are shown in Figure 3.5. Aside from two wildfires with origin locations in the water, off the coast, the rest of the wildfire locations are not

obviously odd. The projected distance of the 175 wildfire locations to the nearest path on the linear network range anywhere from 1173 meters to 0.05 meters. The two aberrant wildfires in the water have the farthest projections at 1173 and 660 meters each and these two wildfires have been removed from the dataset for the rest of the analysis. The mean projected distance for the 175 wildfires is 66 meters and 150 (85%) of the 175 wildfires are within 100 meters to the nearest path on the linear network. Once again, the original wildfire origin locations are at least as precise as 1-square mile grid (1609.34 meter-square grid). All projection lengths for the 175 wildfires are under a mile in length and a histogram of the distances can be seen in Figure 3.3.

| Ignition Cause | # of Fires between 2000 and 2013 | Acres burned |
|---|---|---|
| Arson | 10 | 26.51 |
| Campfire | 4 | 4916.2 |
| Children | 4 | 42.21 |
| Debris Burning | 4 | 13.3 |
| Equipment Use | 36 | 109.66 |
| Fireworks | 3 | 3.1 |
| Lightning | 1 | 0 |
| Miscellaneous | 23 | 134.91 |
| Missing/Undefined | 69 | 4732.33 |
| Powerline | 12 | 904.9 |
| Smoking | 8 | 1.81 |
| Structure | 1 | 2 |

Table 3.3: Wildfire Ignition Cause breakdown in Santa Monica Mountains Region

### 3.3.3 RAWS meteorological data

This subsection outlines the elevation and meteorological data used in the analyses of Santa Monica Mountains' wildfires on the road network. As elevation is an important factor in determining ignition probability, and spread of wildfires, it is added as one of the spatial

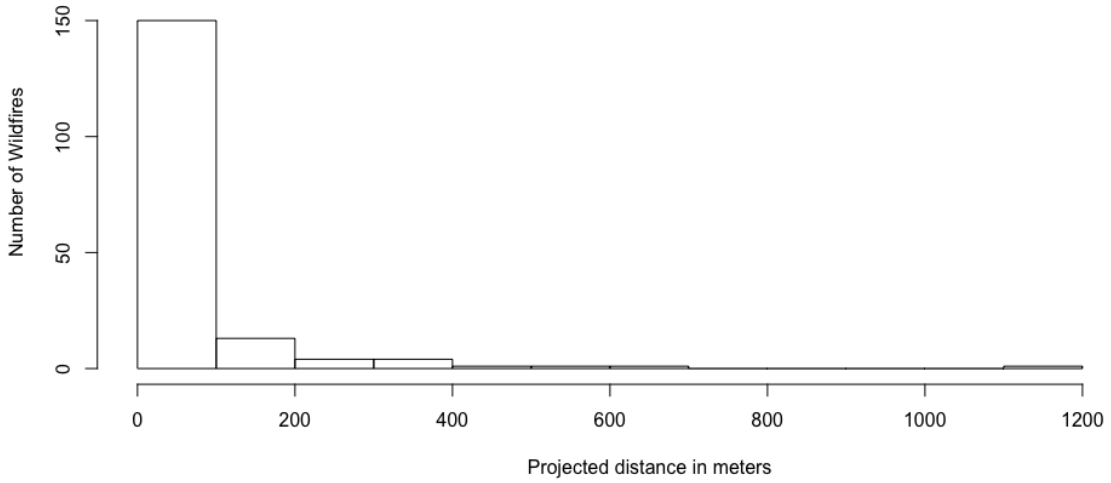


Figure 3.3: Projection Distance in meters of the 2000-2013 wildfire origins to the nearest road

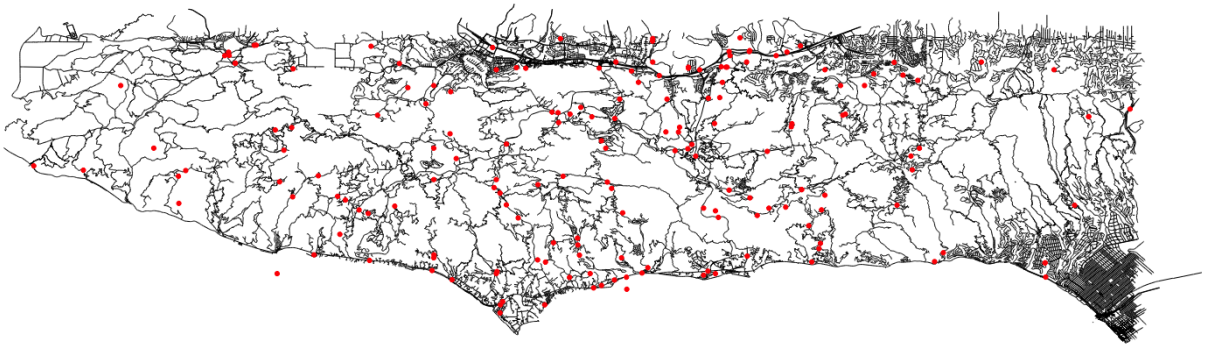

Figure 3.4: Original Wildfire origins in the Santa Monica Mountains Region between 2000 and 2013

covariates in estimating the wildfire ignition *formation* intensity. The elevation data for the Santa Monica Mountains region was sourced from the National Map website (`https://nationalmap.gov`) of the United States Geological Survey. The elevation data sourced is at 1/3 arc-second resolution, which is the highest resolution of elevation data avaiable for the

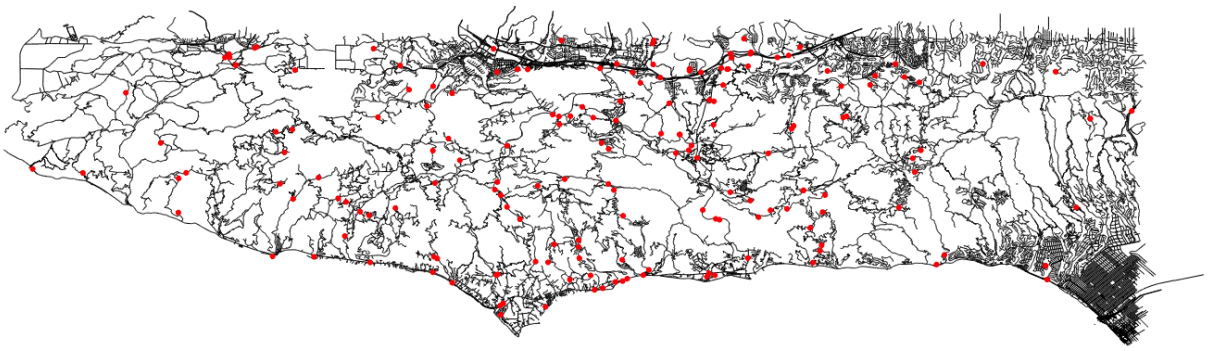

Figure 3.5: Projected Wildfire origins in the Santa Monica Mountains Region between 2000 and 2013

U.S. The resolution or ground spacing is approximately 10 meters north/south, but variable east/west due to convergence of longitudes with latitude. The elevation map over the Santa Monica Mountains region is shown in Figure 3.6.

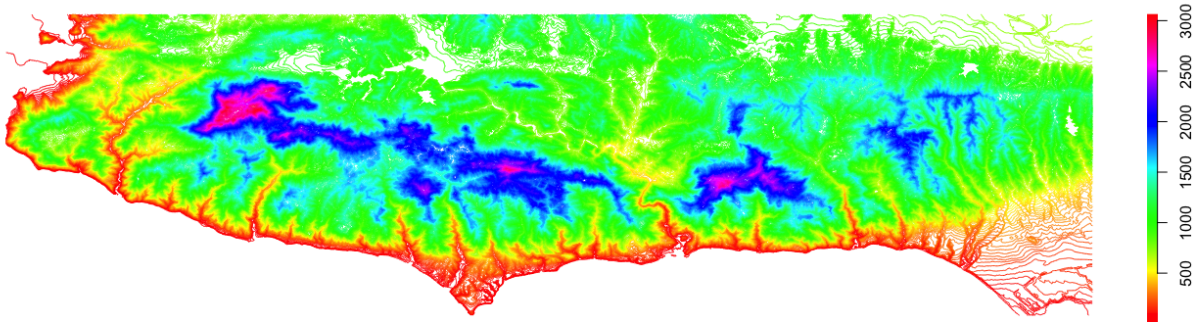


Figure 3.6: Elevation in feet of the Santa Monica Mountains Region

Finally, the meteorological data on daily average temperature, precipitation, humidity, wind speed, and wind direction are sourced from the RAWS (Remote Automated Weather Station) archival website hosted by Desert Research Institute at `https://raws.dri.edu`. There are only 4 RAWS that fall within the Santa Monica Mountains region and their

locations can be seen in Figure 3.7.

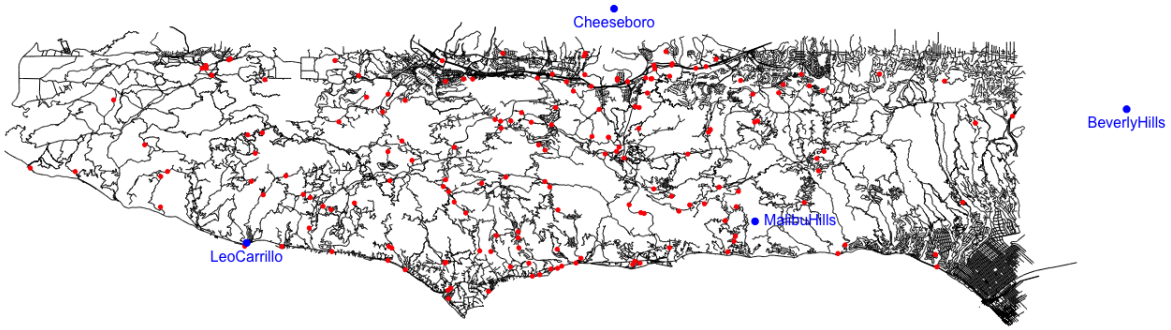


Figure 3.7: Location of RAWS nearest to the Santa Monica Mountains Region

### 3.3.4 Smoothing of RAWS data

As the meteorological data is only available from the four RAWS locations shown in Figure 3.7, spatial and temporal smoothing was performed to be able to include the data as covariates into the linear point process model. The two methods explored to smooth the meteorological data were thin plate spline regression models and generalized additive models. Each of them are summarized below.

#### 3.3.4.1 Thin Plate Spline Regression Models

Thin plate spline regression models (Nychka (2000)) are fitted on each of the daily meteorological variables with longitude, latitude and year of the day variables as predictors. Multiple fits of the thin plate spline models (with quadratic and cubic spatial trends) were explored.

The R package '*fields*' was used to fit the thin plate spline regression models on each of the four meteorological variables (total precipitation, average temperature, average speed, average humidity). A separate spline model was fitted for *each* of the 13 years. The thin plate spline model minimizes the residual sum of squared errors subject to a constraint that the function has a certain level of smoothness (or roughness penalty). The assumed model

is additive and the smoothing parameter is chosen by cross-validation.

$$Y = \beta_0 + f(X) + \lambda, \tag{3.8}$$

where $\beta_0$ is a constant, $f(X)$ is a 3-dimensional surface comprised of sum of flexible functions for longitude, latitude and year of the day and $\lambda$ is the error term that also provides a built-in smoothing function based on a penalized least squares method. This is equivalent to minimizing the penalized sum of squares

$$E_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n} W_i(y_i - f(\mathbf{x}_i))^2 + \lambda J_m(f) \tag{3.9}$$

where $J_m(f)$ is the roughness penalty function and $\lambda$ represents the smoothness constraints (Nychka (2000)). When $\lambda > 0$ and two dimensions, the roughness penalty function is

$$J_2(f) = \iint_{\mathbf{R}^2} (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2)dxdy \tag{3.10}$$

$\lambda = 0$ represents zero smoothness constraints and $\lambda = \infty$ corresponds to fitting the polynomial base model by ordinary least squares.

However, due to the limited number of RAW stations, resulting in just four unique longitude and latitude pairs, multiple fitting issues (collinearity) were faced; especially for quadratic and cubic splines models. The linear thin plate spline models, while too simplistic, also heavily affected the linear model predictions, as explored in section 3.5.3. As a result, Generalized Additive models were pursued as an alternative.

#### 3.3.4.2 Generalized Additive Models

A generalized additive model (Hastie and Tibshirani (1990)) is a nonparametric case of a generalized linear model where the predictors are related to the dependent variable via individual functional forms and additivity is assumed between these predictor smooth functions. A simple example is:

$$g(E(y_i)) = \alpha + f_1(x_{1i}) + f_2(x_{2i}) \tag{3.11}$$

where $y_i$ is the response variable, $f_1$ and $f_2$ are smooth functions of covariates $x_1$ and $x_2$, and $g$ is the link function. The smooth functions need to adhere to constraints for the model

to be identifiable. In the case of the meteorological variables, two smooth functions were used; one was a function of longitude and latitude while the other was a cyclical function on the day of the year. The R package, '*mgcv*' (Wood (2011)) was used to fit the generalized additive models on the meteorological variables. Once again, a separate GAM was fit on each of the 13 formation years and the GAM estimates were compared to those from the thin plate spline models. The GAM estimates were comparatively smoother, and after close inspection, GAM yearly estimates were used as the yearly meteorological covariates.

The daily wind direction variable was smoothed using directional kernel regression through the R package '*directional*'. Once again, longitude, latitude and the day of the year were used as predictor variables to smooth out the daily wind direction, and a separate directional model is estimated for each of the 13 years. The estimated angle value is then categorized into eight directions: North, North-East, East, South-East, South, South-West, West, and North-West.

Due to extensive missing data before year 2000, the RAWS data used here are only from years 2000 to 2013; as a result, the wildfires analyzed are also from years 2000 to 2013. Year 2002 has the largest amount of missing RAWS data at 20%; the rest of the years have 12% or less missing values.

## 3.4 *Formation* Model on Wildfire origins along Road Networks

Given the model specification outlined in Section 2.2, the current section presents the *formation* model that estimates the conditional intensity of wildfire ignition occurrence in the Santa Monica Mountains region. To recap, the separable temporal model presented in Chapter 2 outlines the ideas of *formation* and *dissolution* processes that are conditionally independent, but come together to capture the process evolution. The conditionally independent processes individually explain what factors affect the *formation* and *persistence* of a point event on a linear network over time. In applying this framework to wildfire origin locations, a *formation* process model capturing ignition occurrence intensity over time is estimated; a *dissolution* model is skipped as extinguished wildfires are not point events and involve concerted human

effort.

To fit the *formation* model, the yearly wildfire origin locations are first consolidated into the *formation point processes* ($\mathbf{x}^t$), where $\mathbf{x}^+ = \mathbf{x}^t \cup \mathbf{x}^{t+1}$. Each wildfire *formation point process* is a union of the wildfires that occurred in 2 consecutive years. For example, the *formation point process* between the years 2000 and 2001 is:

$$\mathbf{x}^+_{2000-01} = \mathbf{x}_{2000} \cup \mathbf{x}_{2001} \tag{3.12}$$

The 14 years of wildfire origin locations between 2000 and 2013 result in 13 *formation point processes* that form the basis of the *formation* model. Each *formation point process* is assigned with dummy points that are placed equidistantly on the linear network and weighted accordingly. The data points (wildfire origin locations) are also assigned weights based on the length of the network segment they lie on. Through the Berman-Turner device (Berman and Turner (1992)) introduced in Sections 1.2.3 and 2.3, the weighted data and dummy points along with their corresponding spatial covariate values (meteorological and network) produce the following estimate to the wildfire ignition *formation* model (conditional intensity of ignition occurrence), seen in Table 3.4.

An important note to mention about the results in Table 3.4 is that the maximum pseudolikehood coefficient estimates are presented with bootstrap standard errors. While the Berman-Turner method and the GLM software serve as convenient computational tools to estimate the Papangelou conditional intensity (through a log-linear model), the GLM software incorrectly assumes identical and independently distributed Poisson observations, which is not the case here. As a result, non parametric methods are adopted for model inference; the bootstrap method for dependent data used here is outlined in Section 3.5.2 and model assessment is covered in Section 3.6.

### 3.4.1 Model Interpretation

Looking at the estimate values in Table 3.4, all of the roadtypes–in comparison to residential roads–contribute positively to the conditional intensity of wildfire ignition. Paths, Road1, and Road2 categories are all likely to increase wildfire ignition intensity by at least a factor

of 1.5 when compared to residential roads. Road1, representing the highways and primary roads, increases ignition intensity by a factor of 7.7. This is possibly due to the high speed vehicular traffic, equipment use, and power lines along Road1 category roads. The Unknown road type, despite its small representation of the full network at just 2.8%, also has a relatively high effect on ignition intensity. These roadtype coefficients clearly indicate that human and vehicular activity are leading factors in determining ignition frequency of wildfires. The road network acts as a perfect proxy for human and vehicular activity and find that it plays an important role in understanding the spatial distribution of wildfire ignitions over the years.

The effect of months varies depending on the month, with August(month 8) as the reference factor level; August is a relatively dry month that also receives the dry, high speed Santa Ana winds from the East, making it a high fire hazard month. Compared to August, almost every other month has a lower intensity of ignition at varying degrees (see Figure 3.8). February has the lowest rate of ignition intensity while July has the highest rate of ignition intensity compared to August. June and September display only slightly lower intensities than August; overall the Summer and early Fall months (time of Santa Ana winds) indicate higher ignition intensities than the Winter and Spring months. The monthly seasonal effects on ignition intensity can be seen in Figure 3.8. Despite the monthly effect, there has been a sustained expansion in the length of fire seasons due to increase in human-related causes (Balch et al. (2017)). While this expansion in the fire season is not captured due to the time inhomogeneity of the model, the inclusion of the road network as a human proxy does account for the human involvement in wildfire ignition intensity over the 14 year time period.

Two meteorological variables, total daily precipitation and average daily humidity, slightly decrease ignition intensity for every unit increase in value. The daily average temperature and daily average windspeed variables slightly increase ignition intensity for every unit increase in value. Compared to these meteorological variables, all the wind direction categories, compared to winds blowing from the West, positively effect ignition intensity. Specifically, winds blowing from the North-East and the South-East have some of the highest positive effects on ignition intensities in comparison to winds blowing from the West. Note that the

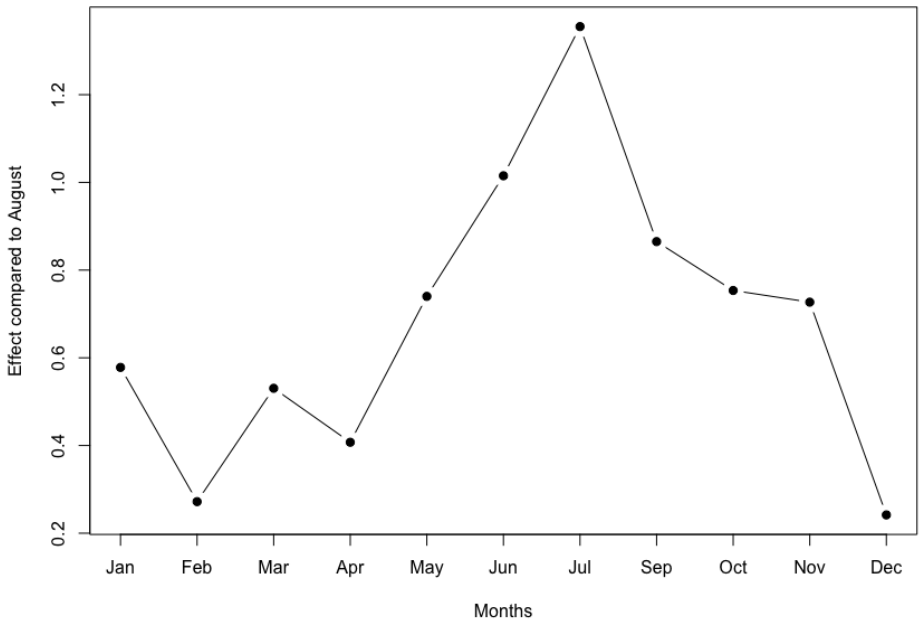


Figure 3.8: Monthly Estimated Conditional Intensity of Wildfire Ignition.

dry, high speed Santa Ana winds blow towards the coastline from the East and North-East directions. While wind itself does not ignite fires, strong winds carry hot embers and debris causing ignitions of dry vegetation elsewhere. Moreover, burning fires are fanned and spread rapidly in the presence of strong winds. The directional effect of wind categories on wildfire ignition intensity is illustrated in Figure 3.9. Once again, due to adopting smoothed meteorological data, year specific, extreme wind events are not accounted for in this model, making the current wind estimates an approximate of the long-term wind effects.

Average daily temperature, Elevation and Slope (change in elevation) also have small positive effects on ignition intensities. It is known that wildfires spread faster uphill but not much is known about elevation's affect on ignition rates. The range of elevation in the Santa Monica Mountains region is from 0 to 2700 feet; while this is not high enough to cause snowtops, there are multiple trails leading to hill tops across the region, as seen in Figure 3.5. While human foot traffic might get sparse with increasing elevation, the road network acts as a proxy for building structures and vehicular traffic that cause structural and vehicular


Figure 3.9: Effect of Wind Direction on Wildfire Ignition.

wildfires.

While the results in Table 3.4 provide indications on how certain covariates affect ignition occurrence intensity, the spatial distribution of ignition intensity is not apparent. The spatially plotted, predicted conditional intensities for the 13 *formation* years are shown in Figures 3.10, 3.11, and 3.12 with the legend in Figure 3.13. The predicted intensities are categorized into 5 quantile intervals with the lowest quantile represented in dark blue and the top 20th quantile represented in yellow.

| Parameter | Coefficient Estimate | MBB SE | MBB 95% CI |
|---|---|---|---|
| (Intercept) | -1.577 | 2.855 | (-6.753, 1.644) |
| Elevation | 0.004 | 0.001 | (0.0005, 0.006) |
| roadtype-Paths | 0.419 | 0.273 | (0.271, 0.970) |
| roadtype-Road1 | 2.044 | 0.097 | (2.028, 2.299) |
| roadtype-Road2 | 0.898 | 0.265 | (0.702, 1.468) |
| roadtype-Unknown | 1.407 | 0.253 | (1.000, 1.866) |
| January | -0.548 | 0.852 | (-1.863, 1.055) |
| February | -1.302 | 0.739 | (-2.841, -0.107) |
| March | -0.634 | 0.787 | (-2.319, 0.082) |
| April | -0.898 | 0.638 | (-1.996, 0.149) |
| May | -0.301 | 0.527 | (-1.308, 0.509) |
| June | 0.015 | 0.467 | (-0.961, 0.509) |
| July | 0.304 | 0.124 | (-0.087, 0.371) |
| September | -0.145 | 0.287 | (-0.903, 0.179) |
| October | -0.283 | 0.383 | (-1.099, 0.093) |
| November | -0.319 | 0.679 | (-1.295, 1.008) |
| December | -1.420 | 0.668 | (-1.974, 0.475) |
| TotPrecipitation | -0.209 | 0.106 | (-0.475, -0.125) |
| AveTemperature | 0.022 | 0.091 | (-0.100, 0.183) |
| AveHumidity | -0.006 | 0.007 | (-0.007, 0.021) |
| AveWindSpeed | 0.009 | 0.209 | (-0.344, 0.371) |
| WindDir-S | 0.780 | 0.397 | (0.173, 1.517) |
| WindDir-E | 0.886 | 0.942 | (-0.341, 2.632) |
| WindDir-N | 0.617 | 0.443 | (-0.324, 1.115) |
| WindDir-NE | 1.307 | 0.933 | (-0.274, 2.766) |
| WindDir-NW | 0.050 | 0.540 | (-1.426, 0.311) |
| WindDir-SE | 1.686 | 0.799 | (-0.510, 2.050) |
| WindDir-SW | 0.745 | 0.432 | (0.185, 1.629) |
| Slope | 0.008 | 0.004 | (-0.003, 0.010) |

Table 3.4: *Formation* model of Wildfire ignitions in the Santa Monica Mountains Region between 2000 and 2013. This Table presents the time-invariant, log linear model of the Papangelou conditional intensity of wildfire ignition occurrence. A unit change in the coefficients represents a log effect on the expected conditional intensity. The factor reference levels for the roadtype, month, and wind direction coefficients are Residential, Month 8(August) and West respectively. The standard errors and confidence intervals are produced using the Moving Block Bootstrap method.

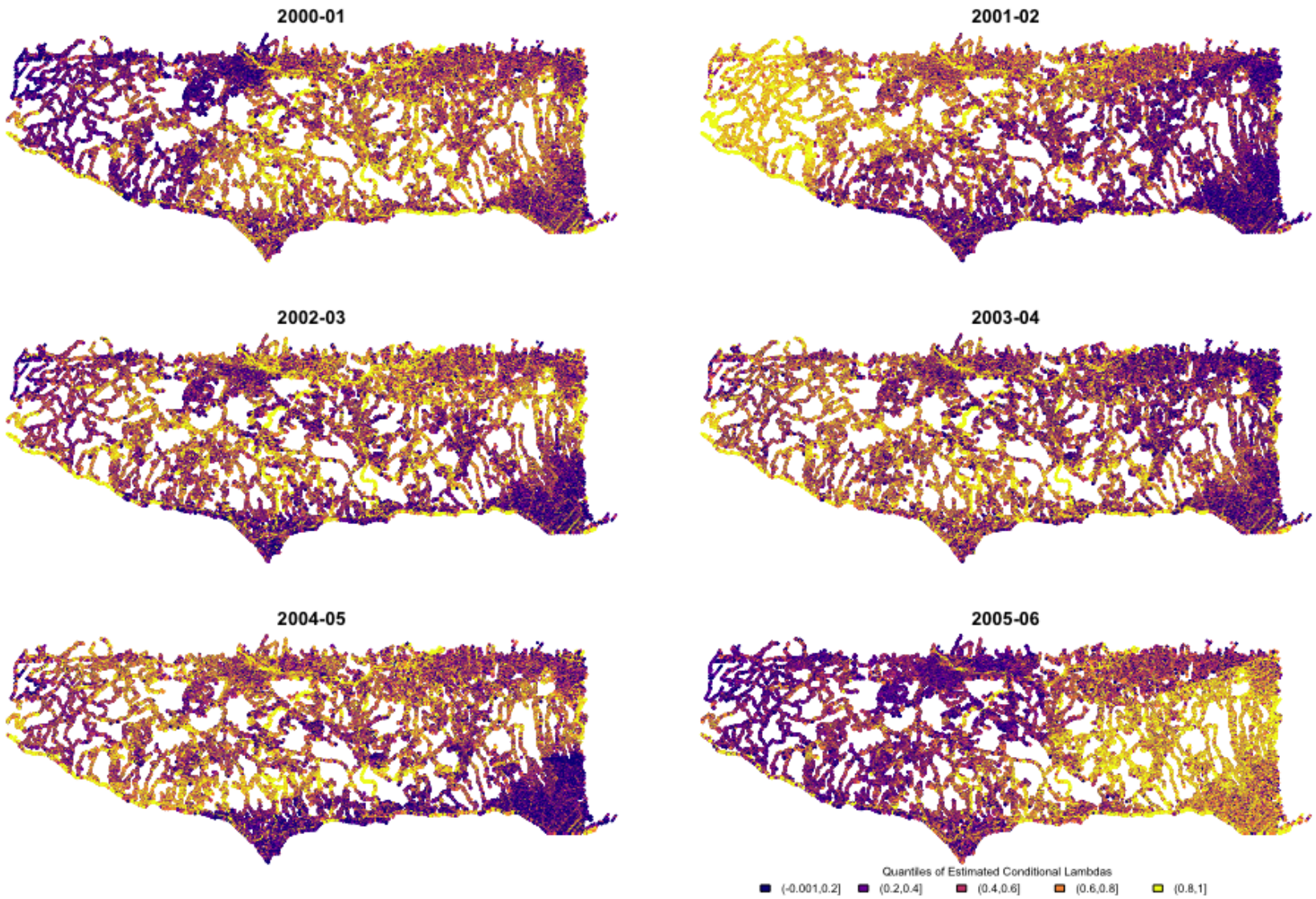


Figure 3.10: Estimated Conditional Intensity of Wildfire Ignition Occurrence from 2000 to 2006. See figure 3.13 for the legend.

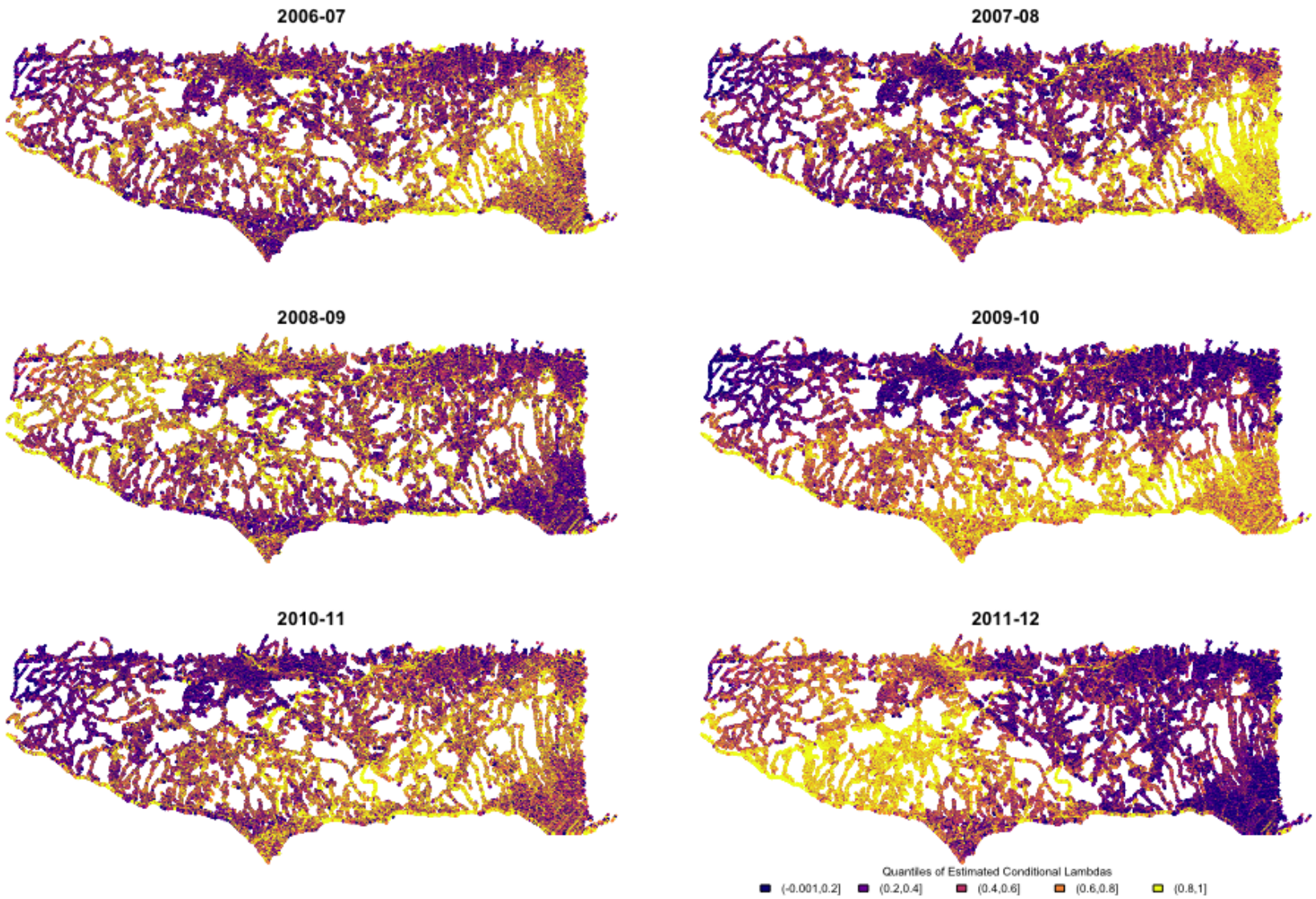


Figure 3.11: Estimated Conditional Intensity of Wildfire Ignition Occurrence from 2007 to 2012. See figure 3.13 for the legend.

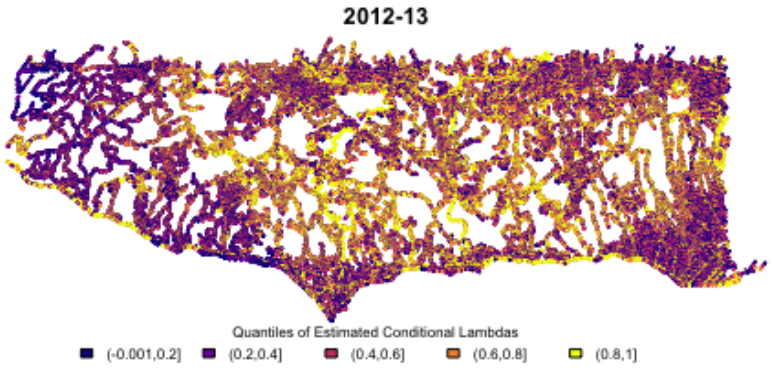


Figure 3.12: Estimated Conditional Intensity of Wildfire Ignition Occurrence from 2013. See figure 3.13 for the legend.




Figure 3.13: The legend for Figures 3.10, 3.11, and 3.12. Each interval represents the quantile of the predicted wildfire ignition formation intensity.

At first glance, Figures 3.10, 3.11, and 3.12 provide a general view of how the formation model's ignition intensities spatially change over the 13 formation years. Some years like 2000-01 and 2001-02 have clear spatial demarcations that indicate stark intensity changes, while other years are less so with a wide range of intensities over small regions. The model generally estimates lower ignition intensities in regions with high residential roads such as the South-East corner of Santa Monica city, the North-East region of Woodland Hills. However, this trend is reversed for a few years in between before reverting back to low intensity rates at the end of the 13 years. The large central track of the Santa Monica Mountains region experiences a range of low to high ignition intensities over the 13 years. The western track of the window, representing part of Point Mugu State Park and Malibu Springs, experiences generally low ignition intensities (except for 2 years); this region is generally remote, with not many areas categorized as Wildland-Urban Interface and Intermix regions in it. The regions with regular, yearly changes in ignition intensities over the 13 years are generally regions that are categorized as Wildland-Urban Interface and Intermix regions (See Figure 3.22 for WUI regions). These regions include the Tuna Canyon Park, Will Rogers State Park, Brentwood, Tarzana, Hidden Hills and Agoura Hills on the outer edges of the road network, and Topanga, Monte Nido and Cornell in the central regions of the road network.

For a closer look at how the intensity estimates are changing based on the road type, year 2012-13's estimates are plotted by road category in Figure 3.14. The formation year 2012-13 was chosen for this breakdown as it is one of the years displaying a wide range of intensity estimates over small regions, resulting in an overlay of multiple colors on the full network as shown in Figure 3.12. The residential roads have the highest degree of intensity range, with the highly dense regions of Brentwood and Pacific Palisades displaying a mosaic of colors. The same could be said of dense residential regions of north and north-east regions for the network. The residential roads on the western regions display the lowest ignition intensities; the western regions are also the most remote regions, with most of it belonging to the Point Mugu State Park.

The intensities on the Road1 category roads are relatively stable, with the Pacific Coast Highway (PCH) displaying higher ignition intensity than the 101 freeway. However, the

Malibu Canyon road and the Kanan Dume road show varying degress of intensities all along their length. While the 101 freeway is mostly surrounded by high building density areas, the Canyon roads connecting the PCH and the 101 freeway are heavily flanked by Wildland-Urban Intermix regions. This presence of high vegetation, lower building density and high traffic roads is a possible reason for the wide range of ignition intensities along these roads. The roads under Road2 category follow a slightly similar pattern where possible structural density subdues the ignition intensity range, and large regions of WUI regions indicate higher intensities. Finally, all roads categorized as Paths show relatively high rates of ignition intensities. The western region, in Point Mugu State Park, which is relatively remote display the lowest intensities of ignition. Overall, it is evident that the model is able to measure the regions of high Wildland-Urban Intermix and Interface, and roads flanking these regions have the largest change in ignition intensities spatially and temporally.

The results in Table 3.4 represent the *formation* conditional intensity, resulting from modeling the combined set of formation processes, over the 13 years. This requires close consideration while interpreting the model results. Table 3.4 represents the conditional intensity of a new ignition occurring at a specific point, given all the ignitions that occurred around it in the network region and the year before. This spatial and temporal conditional intensity is a measure of ignition likelihood at an infinitesimal point, given the state of its spatial surroundings and one year history. Given this setup, the model in general indicates that the regions with the most volatile ignition intensities are those with low structural density, but with considerable human and vehicular traffic. The model, with the aid of the road network, has essentially mapped out regions of high Wildland-Urban Intermix and Interface.

While the ignition causes (shown in Table 3.3) were not used as covariates in this model, Syphard and Keeley (2015) model each ignition cause separately to understand the affect of covariates. Their major conclusion is that distance to road, distance to structure, and structure density are the most influential factors in affecting ignitions with human-related causes. These human-related causes include arson, vehicular fire, equipment fire, campfire, powerline, and playing with fire. While their approach was in understanding how factors

affect each ignition cause, the separable temporal linear point process (STLPP) model aims to model the spatial and temporal change in ignition locations with respect to human activity, for which the linear road network is a perfect proxy. The medium term view of how the spatial and temporal distributions of ignition locations changed should provide insights into developing alternative fire management programs. Note that only ignition locations, and not the spread and extent of the wildfire, were studied to focus primarily on understanding the distribution of origin locations with respect to human movement and density.

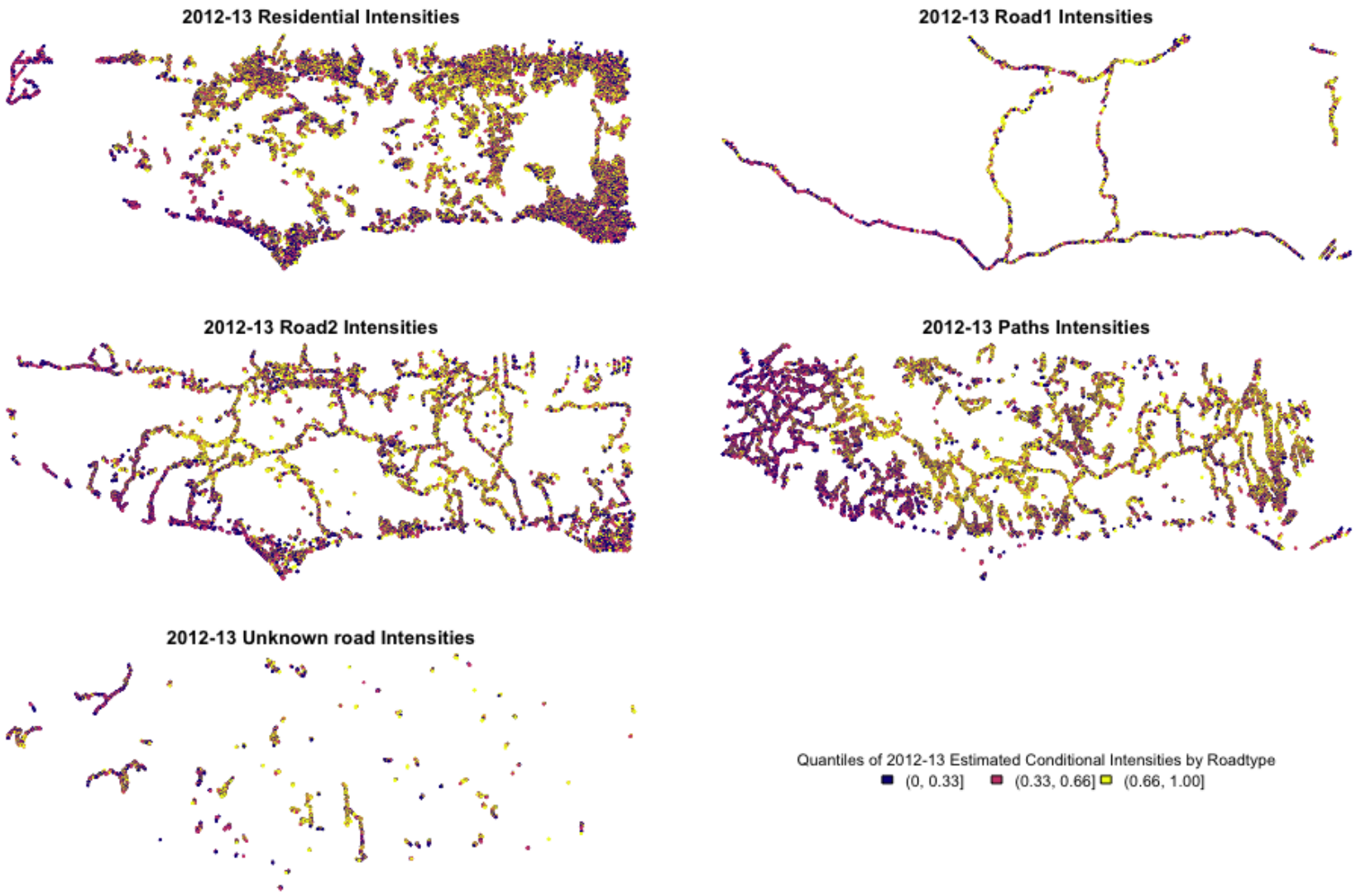


Figure 3.14: 2012-13 Estimated Conditional Intensity of Wildfire Ignition by Roadtype

### 3.4.2 Moving Block Bootstrap Standard Errors

Due to the implicit spatial dependency in the Papangelou conditional intensity, the standard errors of the log-linear model produced by the GLM software are incorrect. As a result, nonparametric bootstrap methods for dependent data are used to produce the standard errors shown in Table 3.4. The Moving Block Bootstrap, independently introduced by Kunsch (1989) and Liu and Singh (1992), is used to estimate the standard errors. Instead of randomly drawing a sample, the Moving Block Bootstrap resamples *blocks* of consecutive observations to retain the temporal or spatial dependency.

Assume that $X_1, X_2, \ldots$ is a sequence of stationary random variables, and let $\mathbf{X_n} = (X_1, \ldots, X_n)$ denote the observations. In the wildfire case, $\mathbf{X_n}$ represents the complete set of data and dummy points from the 13 *formation* years. Given the data, blocks of size $l$ are defined as $B_j = (X_j, \ldots, X_{j+l-1})$, $j = 1, \ldots, N$ where $N = n - l + 1$ denotes all the possible overlapping blocks in $\mathbf{X_n}$. The bootstrap blocks are obtained by selecting a random sample of $b$ blocks from the full set of overlapping blocks $\{B_1, \ldots, B_N\}$.

For the standard errors in Table 3.4, the moving block bootstrap method was implemented with block size of 1.3 million data points. The entire set of data and dummy points over the 13 years is 2.3 million data points. This resulted in a set of about 1 million overlapping blocks. A thousand samples were drawn and modeled through the GLM software to obtain the empirical distribution of each of the model coefficients. The sample standard deviation of this empirical distribution provides the moving block bootstrap standard errors of the model coefficients. Note that this method is appropriate for the separable temporal model as it assumes time inhomogeneity, and the data and dummy points are ordered to be spatially consecutive to retain spatial dependency.

A few other methods such as the Non-overlapping block bootstrap and Block-Jackknife methods for dependent data were considered. While non-overlapping block bootstrap should result in similar standard errors, the block-jackknife method is computationally intensive for the wildfire dataset. Hall (1985) introduced a moving tile block bootstrap method that is also appropriate for spatially dependent data. However, the moving tile method did not work for

the wildfire data as the moving spatial tile restricted year-long temporal sampling, resulting in zero estimates for some of the months. However, this is not an issue with the moving block bootstrap method, used above, since the data is ordered to be spatially contiguous, and the sampled blocks $B_j$ are akin to moving tiles which also retain the temporal coverage of a year span.

## 3.5 Model Assessment

This section outlines a few model assessment methods for spatio-temporal point process models on linear networks. While not all of them are implemented, they are evaluated for their appropriateness for linear network models.

### 3.5.1 Monte Carlo Tests on Model Deviance

For an overall assessment of how the model is performing, one suggested option is the Pseudolikelihood ratio test conducted through Monte Carlo tests. It is a reasonable substitute for the likelihood ratio test and is based on the the log *pseudolikehood* ratio. Consider a null hypothesis $H_0$ and an alternative hypothesis $H_1$, where $H_0$ is nested in $H_1$. The estimates of the canonical parameters (by maximum pseudolikehood) are $\hat{\theta}_0 = \hat{\theta}_0(\mathbf{x})$ under $H_0$ and $\hat{\theta}_1 = \hat{\theta}_1(\mathbf{x})$ under $H_1$. The test statistic will be twice the log pseudolikehood ratio

$$\Delta = \Delta(\mathbf{x}) = 2\bigg( \log PL(\hat{\theta}_1(\mathbf{x}); \mathbf{x}) - \log PL(\hat{\theta}_0(\mathbf{x}); \mathbf{x}) \bigg) \tag{3.13}$$

and is analogous to the deviance in likelihood theory (Baddeley and Turner (2006)).

Since there is no statistical theory available to support inferential interpretations of $\Delta$, the alternative is to use it as a test statistic in a Monte Carlo test. First, if $H_0$ is simple enough, $m$ independent realizations of point processes $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ are generated. Then the corresponding test statistics, $\Delta_i = \Delta(\mathbf{x}^{(i)})$ for $i = 1, \ldots, m$ are computed and ranked along with the $\Delta$ in equation 3.12. Under $H_0$, the rank R (of the $\Delta_i$s)is uniformly distributed on $\{1, 2, \ldots, m+1\}$ assuming no ties. The Monte Carlo test rejects $H_0$ when $R \leq k$ has size

$\alpha = k/(m+1)$ exactly. The associated p-value is

$$p = \frac{R}{m+1} \tag{3.14}$$

This method of testing the log pseudolikehood ratio was not implemented due to a couple of reasons. The null hypothesis $H_0$ necessary to estimate the test statistic is required to be nested or contained in the alternative hypothesis $H_1$. While it is easy to choose a smaller, nested $H_0$, it is not apparent as to what constitutes the best null hypothesis in the case of wildfire ignition location modeling with respect to road networks. One option is to assume a constant baseline trend on the entire road network, despite the different categories. Even then, the null hypothesis parameters $\hat{\theta}_0(\mathbf{x})$ are unknown and need to be estimated using the observed data $\mathbf{x}$. This major caveat makes the Monte Carlo test possibly approximate and strictly conservative. This violates the essential requirement of Monte Carlo testing, that the observed and stimulated point patterns should be statistically equivalent if the null hypothesis is true. Under the procedure described above, the simulated point patterns have been generated from the null model with parameter value $\hat{\theta}$, while (if the null hypothesis is true), the observed data came from a null model with unknown parameter value $\theta$. Since $\hat{\theta}$ is not exactly equal to $\theta$, either due to choice of $H_0$ or adopted estimators, the simulated and observed point pattern do not come from the same random process, making the Monte Carlo test only approximate. (Baddeley et al. (2014b)).

### 3.5.2 Residual Analysis for Spatial Point Patterns

An alternative method of assessing model fit is to study the residuals. For spatial point process models, the idea of residuals is not immediately intuitive as there is no natural ordering of spatial point events without a temporal variable. Baddeley et al. (2005) generalize the idea of residuals, common in regression and survival analysis setting, to the spatial setting by employing the Papangelou conditional intensity.

In the current case of wildfire ignition modeling, as the *formation* model is framed as a time inhomogeneous, Papangelou conditional intensity, the residuals introduced in Baddeley et al. (2005) are implemented and presented below. Consider a parametric model for a

spatial point process $\mathbf{X}$ with density $f_\theta$ (assuming it satisfies the positivity condition), the innovation process of the model is defined as

$$I_\theta(B) = n(\mathbf{X} \cap B) - \int_B \lambda_\theta(u, \mathbf{X}) du \tag{3.15}$$

for any set $B \subseteq W$, where $n(\mathbf{X} \cap B)$ denotes the number of randome points falling in B and $\lambda_\theta(u, \mathbf{X})$ is the Papangelou conditional intensity at point $u$. This definition is closely analogous to the residuals in time and space-time, except for the use of the Papangelou conditional intensity. The innovations $I_\theta$ constitute a random signed measure, with a mass $+1$ at each point $x_i$ of the spatial point process and a negative density $-\lambda(u, \mathbf{X})$ at all other spatial locations $u$. They satisfy

$$\mathbb{E}_\theta[I_\theta(B)] = 0 \tag{3.16}$$

when the weight function is set to $h(u, \mathbf{x}) = \mathbf{1}\{u \in B\}$ in the *Georgii-Nguyen-Zessin (GNZ) formula* (Baddeley et al. (2005)). Given the data $\mathbf{x}$ and the parameter estimate $\hat{\theta} = \theta(\hat{\mathbf{x}})$, the raw residuals are defined as

$$R_{\hat{\theta}}(B) = n(\mathbf{X} \cap B) - \int_B \hat{\lambda}(u, \mathbf{X}) du \tag{3.17}$$

Baddeley et al. (2005) introduce a few different weight functions $h(u, \mathbf{x})$ to scale residuals at every point $u$ a certain way. The different weight functions result in the Inverse-Lambda residuals, Pearson residuals and Pseudo-score residuals. The raw residuals are a result of the weight function $h(u, \mathbf{x}) = \mathbf{1}\{u \in B\}$ and their value at every quadrature point $u_j$ is

$$r_j = z_j - w_j \lambda_j \tag{3.18}$$

where $z_j$ is the indicator equal to 1 if $u_j$ is a data point and 0 if $u_j$ is a dummy point; $w_j$ is the quadrature weight attached to $u_j$ and $\lambda_j = \hat{\lambda}(u_j, \mathbf{x})$ is the conditional intensity of the fitted point at $u_j$. The raw residuals of the separable temporal model over the Santa Monica Mountains region are shown in Figures 3.15, 3.16, 3.17, and 3.18.

The idea behind this assessment is to check whether the raw residuals (estimating the Innovation measure) have an expectation that is approximately zero, indicating a good model fit.

$$\mathbb{E}_\theta[R(B, \hat{h}, \hat{\theta})] \approx 0 \tag{3.19}$$

The raw residuals presented in Figures 3.15, 3.16, and 3.17 range between -0.072 and 0.999. More than 90% of the residuals fall in the narrow range of (-0.000474, -0.00000657] that is very close to zero. This indicates that for most of the points on the linear network, the separable temporal model overestimated the conditional intensity of ignition, but only by a very narrow margin. The greatest overestimation by the model, the lower 5% of the residuals, are represented in navy blue in the plots. Over the years, this overestimation is consistently around the dense network of Pacific Palisades and the 101 Freeway. These areas have high structural density, making them relatively urbanized regions with low swaths of vegetation around them. One way to improve model performance on these regions would be to add a covariate identifying WUI, urban and wildland regions on the network. Visually, there are very few pockets of under estimation, with residuals greater than 0, as noticed by the yellow regions in the plots; these represent the few positions that possibly had wildfires but the model estimated less than expected ignition intensity during that year.

In general, as there are far more dummy points than data points, and the model estimates a non-zero ignition intensity at most points, most residuals are negative but very close to zero. Aside from the regions that display overestimation (which can possibly be remedied by including a model variable on structural density or WUI), the residual plots are not very indicative of directions for model improvement. As a result, prediction performance of the separable temporal model is considered and presented in the following subsection.

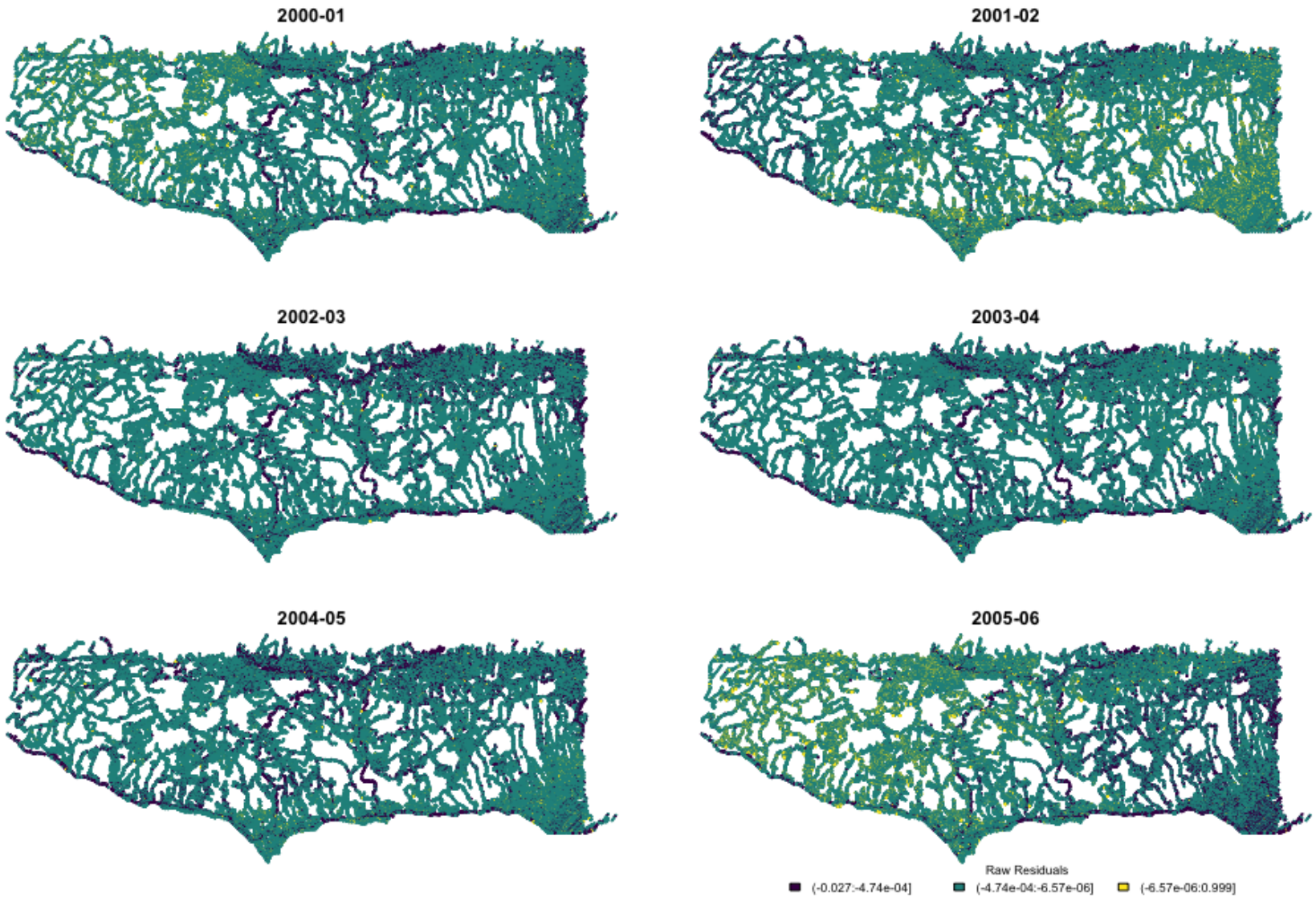


Figure 3.15: Raw Residuals of the Separable Temporal Linear Point Process model on Wildfire ignition data. See figure 3.18 for legend.

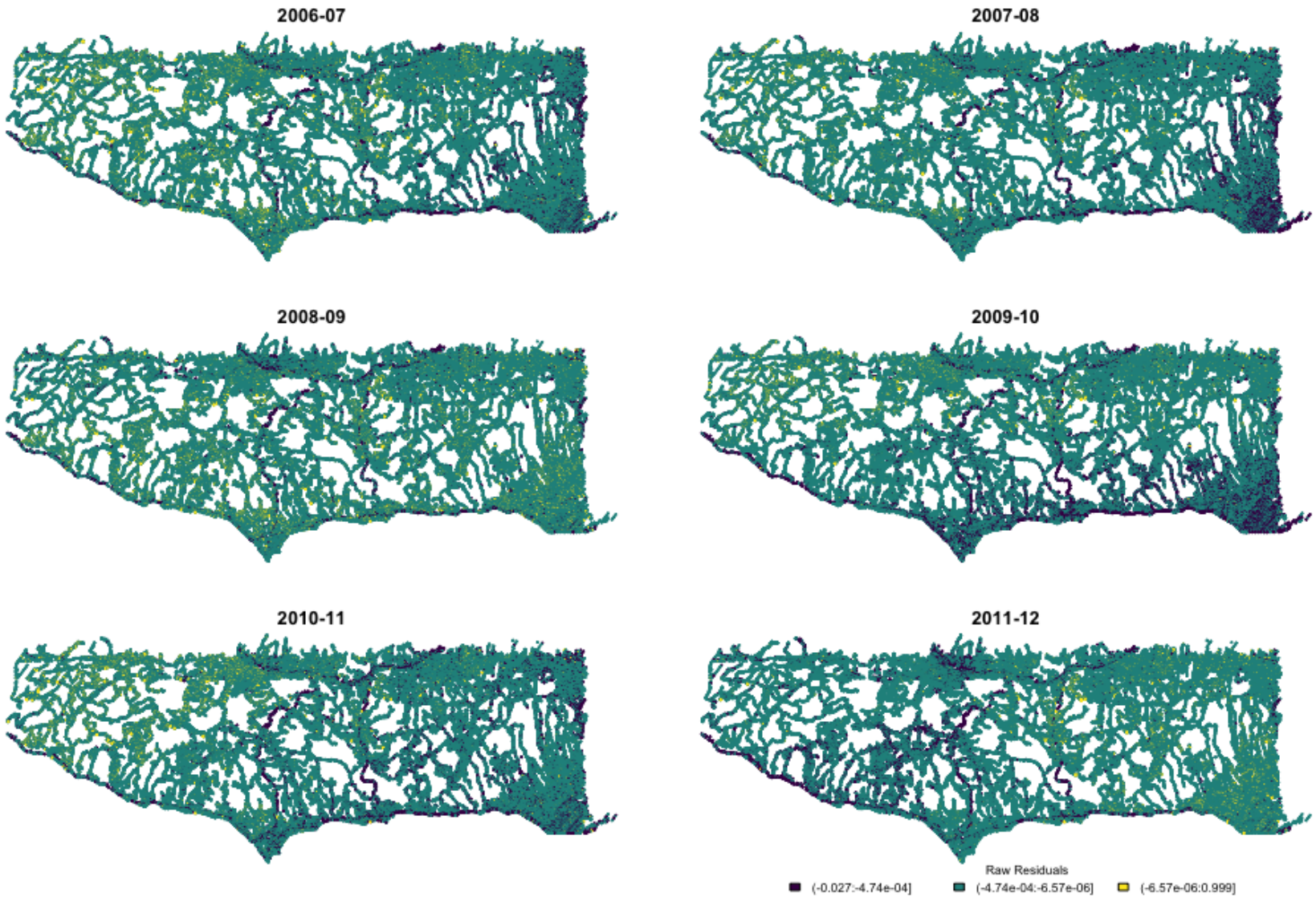


Figure 3.16: Raw Residuals of the Separable Temporal Linear Point Process model on Wildfire ignition data. See figure 3.18 for legend.

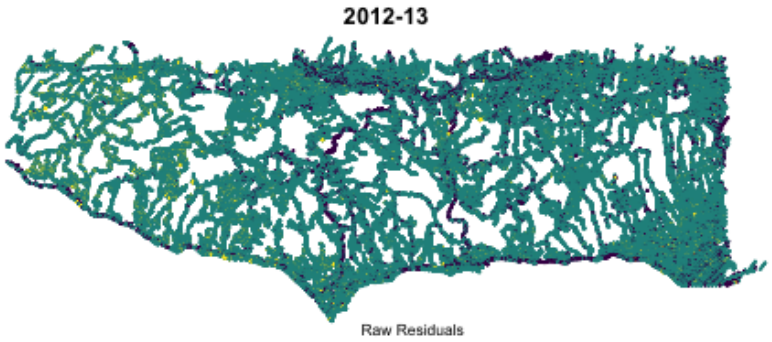


Figure 3.17: Raw Residuals of the Separable Temporal Linear Point Process model on Wildfire ignition data. See figure 3.18 for legend.



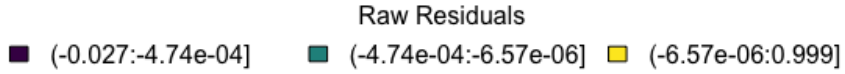


Figure 3.18: The legend for Figures 3.15, 3.16, and 3.17. Each interval indicates the actual residual values and it also represents the 5th quantile, 95th quantile and 100th quantile cutoffs respectively.

### 3.5.3 Residual Analysis on Predicted Conditional Intensities

For prediction, the separable temporal model was first estimated on a smaller set of *formation* point patterns between 2000 and 2008; then the model was used to predict conditional intensities on years 2009 to 2013. Note that while the model is estimated on *formation* point processes, the prediction is performed on just the yearly point patterns, using that year's smoothed RAWS covariates. The predicted conditional ignition intensities can be seen in Figure 3.19 and the raw residuals on the predicted intensities in Figure 3.20.

The predicted conditional intensities look somewhat similar to the estimated conditional intensities shown in Figures 3.11 and 3.12. The most promising feature is that the predicted intensities are consistently on the lower end for regions of high structural density (urban areas). This includes the Pacific Palisades, Malibu Riviera, Hidden hills and Tarzara regions. This was not the case under the estimated intensities where the model estimated high ignition intensities over the highly dense regions of Pacific Palisades and Brentwood over multiple years.

The other most noticeable feature of the predicted intensities is the stark spatial demarcation due to a linear shift in the predicted values. This is most likely an artifact of the smoothed meteorological variables using GAM models. The main reason for exploring various GAM and TPS models was to achieve smooth, non linear coverage of the meteorological covariates over the network window. However, due to collinearity issues, as mentioned in Section 3.3.4, this was not fully achieved and hence visible through the spatial distribution of the predicted intensities. However, the predicted intensity demarcations were much worse on TPS model smoothed data.

The raw residuals are, once again, not very indicative as most of the values are very close to zero. There is consistent overestimation on some of the larger roads over the 5 years; the residuals over the high structurally dense areas are very close to zero, indicating a relatively good predicted value. There are some pockets of underestimation on the North-central regions of road network.

Overall, the residuals on the separable temporal model are not very informative. Future

ways of improving model assessment could include thinning the Papangelou conditional intensity to a poisson process and test for model goodness-of-fit; however, this option requires close consideration in terms of thinning the process along the linear network or in an envelope of space along the linear network. Another option could include model fitting on perturbed ignition locations to test the distance to network measure.

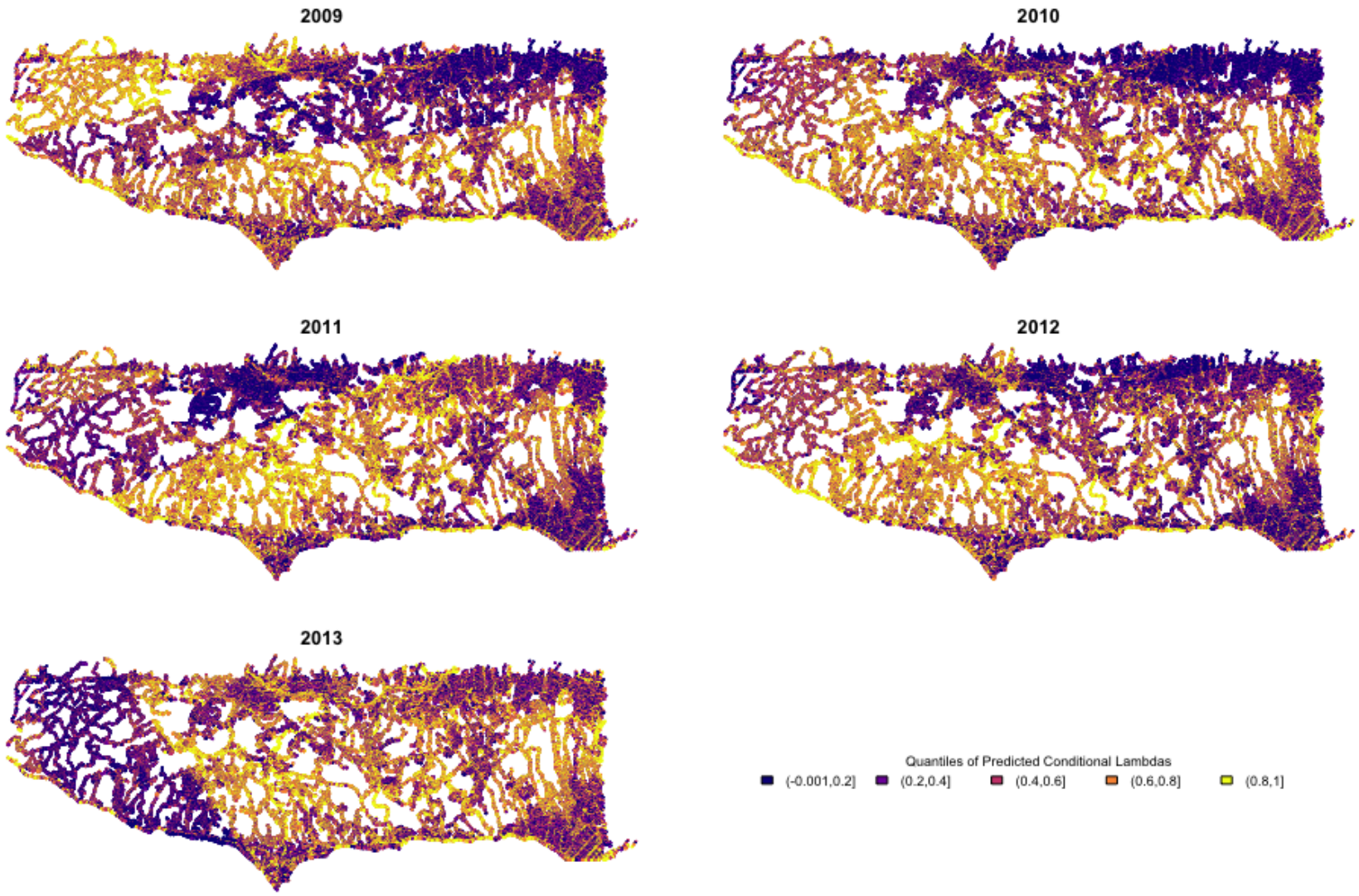




Figure 3.19: The predicted conditional ignition intensities from 2009 to 2013. Once again, the lightest color, yellow, indicates the top 20% of the predicted intensities while the darkest color, dark blue indicates the bottom 20% of the predicted intensities.

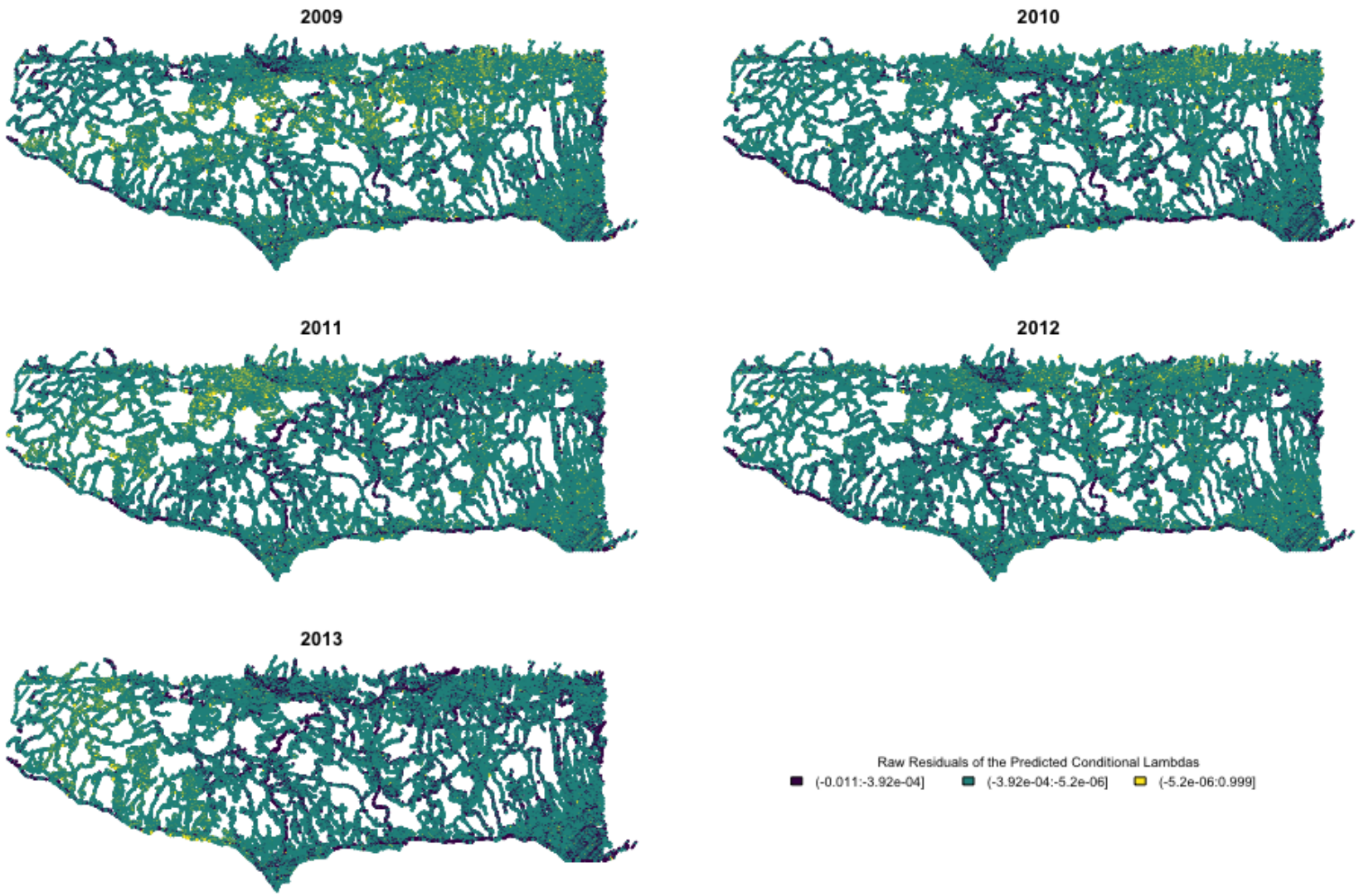


Figure 3.20: The predicted conditional ignition intensities from 2009 to 2013.

### 3.5.4 Super-Thinning

Another method that allows one to assess point process models is super-thinning, a combination of thinning and superposition introduced in Clements et al. (2012). The idea is to transform the point process $N$ (which might not be a poisson Process) into a residual poisson process with rate $k$, where $inf\{\hat{\lambda}(\mathbf{x}_i, t_i)\} \leq k \leq sup\{\hat{\lambda}(\mathbf{x}_i, t_i)\}$. The first step is to thin $N$, keeping each point $(\mathbf{x}_i, t_i)$ independently with probability $min\{k/\hat{\lambda}(\mathbf{x}_i, t_i), 1\}$ to obtain the thinned residual process $Z_1$. The second step is to simulate a Cox process $Z_2$ directed by $max\{k - \hat{\lambda}(\mathbf{x}_i, t_i, 0\}$. This is equivalent to simulating a homogeneous poisson process with rate $k$ and independently keeping each simulated point $(\tilde{\mathbf{x}}_j, \tilde{t}_j)$ with probability $max\{(k - \hat{\lambda}(\tilde{\mathbf{x}}_j, \tilde{t}_j))/k, 0\}$. The points of the residual point process $Z = Z_1 + Z_2$, obtained from superposing the thinned residual process and the simulated process, form the super-thinned residual points.

As $Z$ is a homogeneous poisson process with rate $k$ if and only if $\hat{\lambda} = \lambda$ almost everywhere, the super-thinned residuals can be examined for homogeneity to assess the goodness of fit of $\hat{\lambda}$. The choice of the tuning parameter $k$, that controls the rate of thinning and superposition, is still open to further research. Clements et al. (2012) suggests choosing $k$ that is either the mean or median of $\hat{\lambda}$ as they minimize the absolute and squared deviations of the estimated conditional intensity from $k$.

The mean of the estimated conditional intensity from the formation model is around 0.65. For the first attempt of super-thinning, the tuning parameter $k = 1$ was chosen. Figure 3.21 shows the original wildfires and the super-thinned residuals over the Santa Monica Mountains region.

In general, there are more super-thinned residuals than there are original wildfires. To assess the goodness-of-fit of the formation model, the super-thinned residual wildfires should be similar to a homogeneous poisson process with rate $k = 1$. It is important to note that super-thinning was limited to the linear network over the Santa Monica Mountains region and the uniformity of the super-thinned process should be assessed with respect to the network density. As a result, the super-thinned residual fires are much more common in high-density

network regions of Malibu, Santa Monica and along the 101 Freeway in the north. Visually, the super-thinned residuals seem to be fairly uniform given the linear network density around them; that is, high network density regions have more fires than low network density regions; this is in keeping with a homogeneous poisson process on a linear network. This indicates that in general, the estimated *formation* model has captured the wildfire origin intensity decently well.

A more thorough way of assessing homogeneity is to fit the network K function, estimated on the temporal spider data (Ang et al. (2012)), on the super-thinned residual fires. However due to computational limitations with regard to memory(Santa Monica Mountains linear network is 18MB), this was not possible.

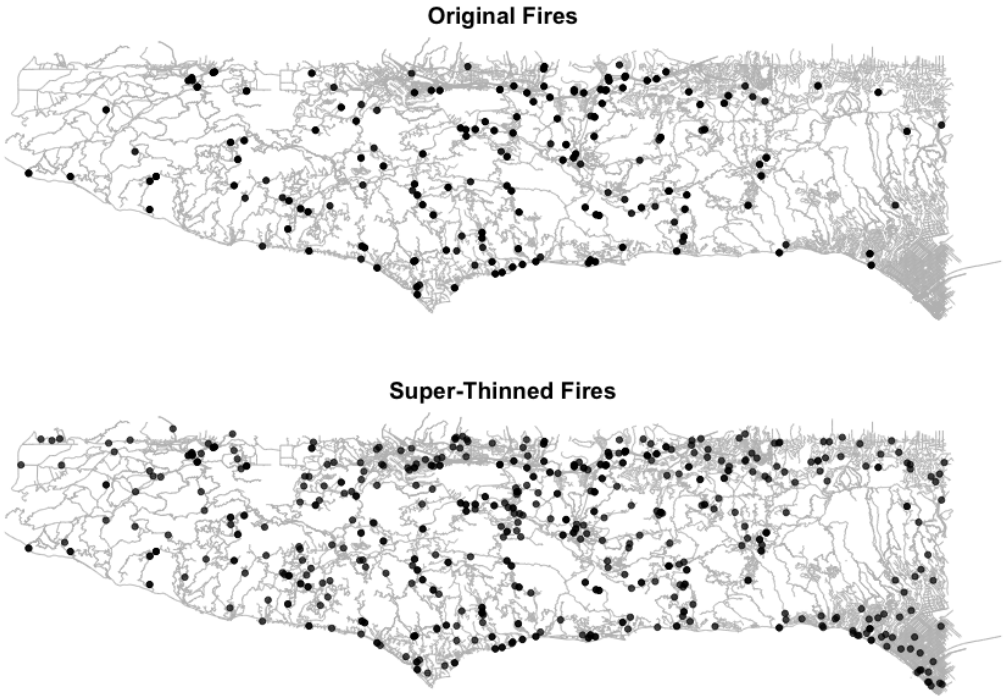


Figure 3.21: The original and super-thinned residual wildfires over the Santa Monica Mountains region with tuning parameter $k = 1$.

## 3.6 Conclusion

The first part of this dissertation presents a new approach to modeling point process data combined with linear networks. And one of the most critical and current applications to the approach has been modeling wildfire origin locations on road networks. As more than 95% of California wildfires have human-related origins, it was intuitive to model ignition locations on road networks, which form a perfect proxy for human traffic. Aside from the novelty of using linear network data, most of the current wildfire research models wildfire burn area *centroids* rather than *origin* locations. The presented seperable temporal model not only models *origin* locations but also models their relationship to road networks. This provides a more nuanced look at how human movement and activity affect wildfire ignitions in Southern California. The temporal aspect of the model also provides a medium to long term spatial view of where most of the ignitions have occurred, allowing greater insight into handling and running fire management programs. The model, combined with burning and fire hazard indices, can aid and direct park rangers and fire fighters to better implement programs such as park closures, fuel management etc.

The model's strongest suit is in providing a spatial distribution of conditional intensities dependent on the density and type of roads. This feature has also brought to light the importance of Wildland-Urban Interface & Intermix (WUI) areas and their high proclivity to fire ignitions (due to high fuel abundance and volatile human activity). There is research indicating the increasing scales of human encroachment into wildland areas, resulting in more ignition prone regions(see Mann et al. (2014)). As mentioned before, the presented model does not include WUI as a spatial covariate and would benefit heavily from doing so. This added spatial covariate might also remedy the overestimation of conditional intensities in the relatively urban areas as seen in the residual plots. The WUI areas over the Santa Monica Mountains region can be seen in Figure 3.22.

In terms of methodological improvements, the meteorological variables smoothed using TPS and GAM models can use a closer look. It is known that weather variables have a non-linear affect on wildfire spread. Moreover, the availability of burn fuel also affects ignition

rates. As a result, including fuel age variable and exploring better kernel methods to capture the true affect of weather could improve model prediction. Another option is to consider regions with more RAW stations that can provide a better spatial cover of weather variables. Another important option is to consider other forms of conditional intensities aside from the log-linear model. While the log-linear model allows for easy computation, other forms of intensity functions could provide better prediction. The separable model also assumes time inhomogeneity and this might not be the best option, especially with changing landscapes of burn fuel and human encroachment. The model also does not capture one-off, extreme weather events over the years. It is also important to test how the (ignition) distance to road variable changed over the long term.

Despite the inclusion of all relevant meteorological variables in the model, burn fuel (vegetation) cover changes depending on previous years' wildfires and seasonal precipitation; this naturally affects the spatial distribution of wildfire ignitions over the years. The current *formation* model for wildfire ignition does not allow for year to year inhomogeneity and as a result might not fully capture the yearly spatial distribution of ignition locations. For future research, fuel age (time from last wildfire), the covariate used in one of the models in Xu and Schoenberg (2011), could help capture the yearly inhomogeneity on wildfire ignition.

Finally, all the coefficient estimates fall in the MBB 95% confidence interval, with a few of them on the lower or upper edge of the interval. Aside from getting a sense for coefficient variability, since only one model was employed, there is no reference model to compare this model's performance with. This can easily be remedied by comparing this model to simpler null models without weather and elevation covariates.

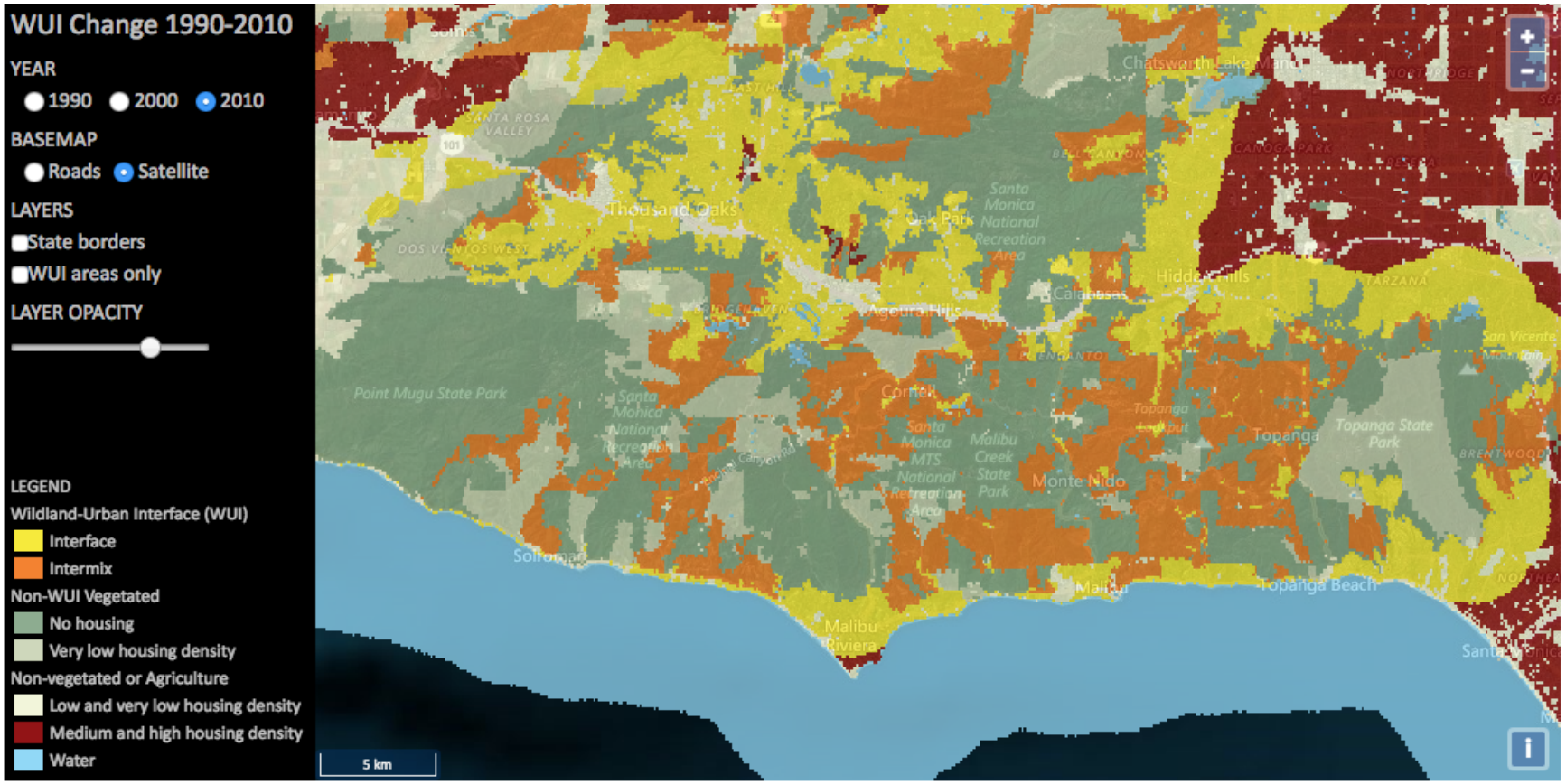


Figure 3.22: The Wildland-Urban Interface areas over the Santa Monica Mountains region. The yellow and orange areas represent the WU Interface and Intermix areas respectively. Source: `http://silvis.forest.wisc.edu/data/wui_change`

### 3.6.1 *Formation* Model without Wind direction

As a quick remedy to the visibly stark gradients produced by the estimated intensity plots in Figures 3.10, 3.11 and 3.12, the *formation* model was re-fitted by excluding the weather variables one by one. It is clear that the directionally smoothed wind direction category variable is causing the stark gradients, and excluding it has resulted in a model with similar coefficient values as Table 3.4 and smoother spatial distributions as can be seen in Figures 3.23, 3.24 and 3.25.

The immediate take aways from the new spatial distribution are that the year to year differences have predominantly subsided and the model has consistently estimated lower ignition intensities in high network density areas such as the Pacific Palisades, parts of Malibu and the northern regions flanking the 101 Freeway. However, a closer look reveals that the estimates *are* different from year to year at the granular level, but the overall sections of networks with high ignition intensities have remained the same. These high ignition intensity roads are represented in yellow, and seem to mostly align with the Wildland-Urban Interface and Intermix regions. This version of the model has clearly revealed certain sections of the road network to have consistently higher ignition intensities over the 13 years, and possibly denote the regions that require higher policing.

Once again, adding burn fuel and WUI variables might provide a clearer year to year evolution in the spatial distribution of high ignition intensity regions.

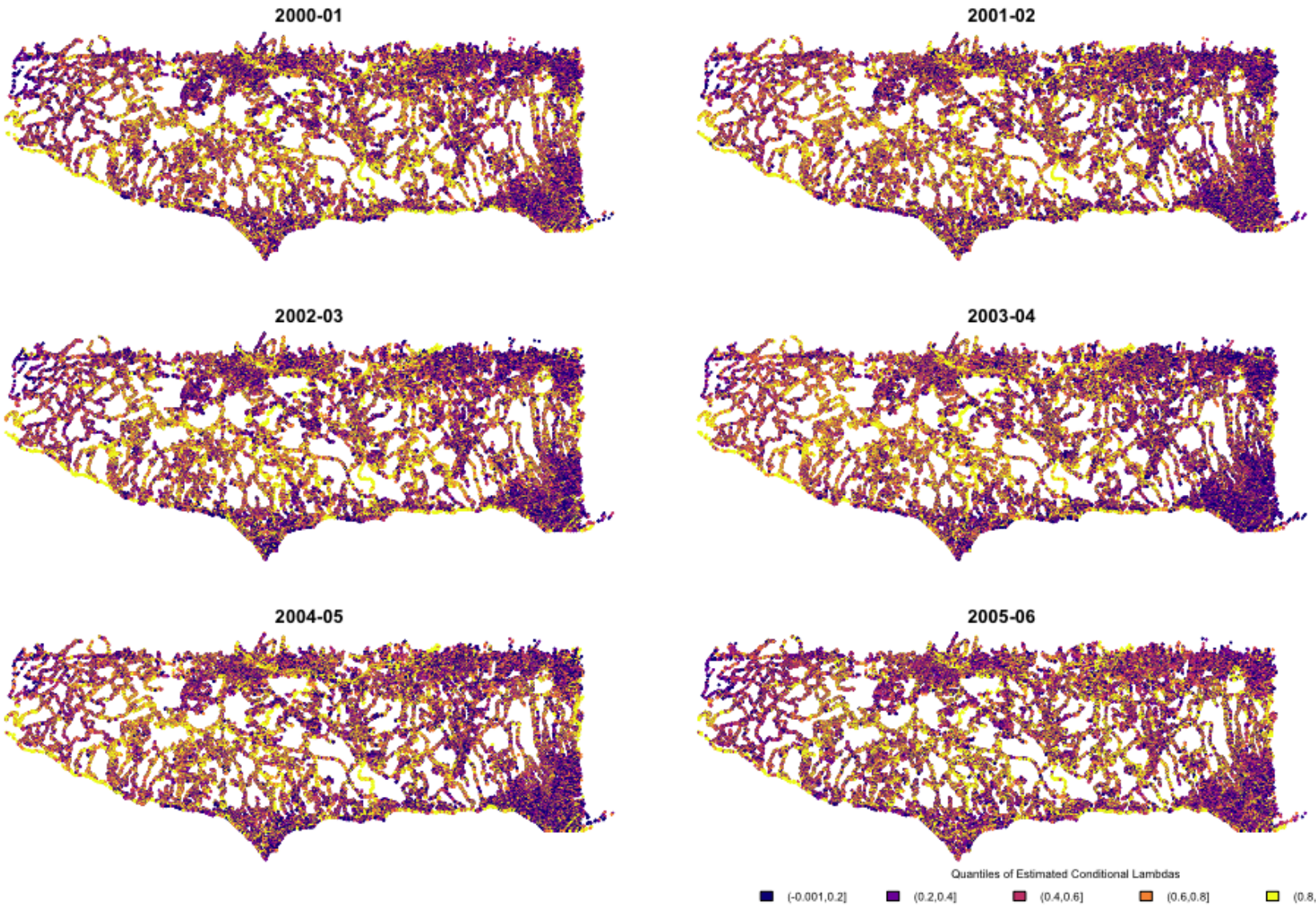


Figure 3.23: Estimated Conditional Intensity of Wildfire Ignition Occurrence from 2000 to 2006, without the Wind Category. See figure 3.26 for the legend.

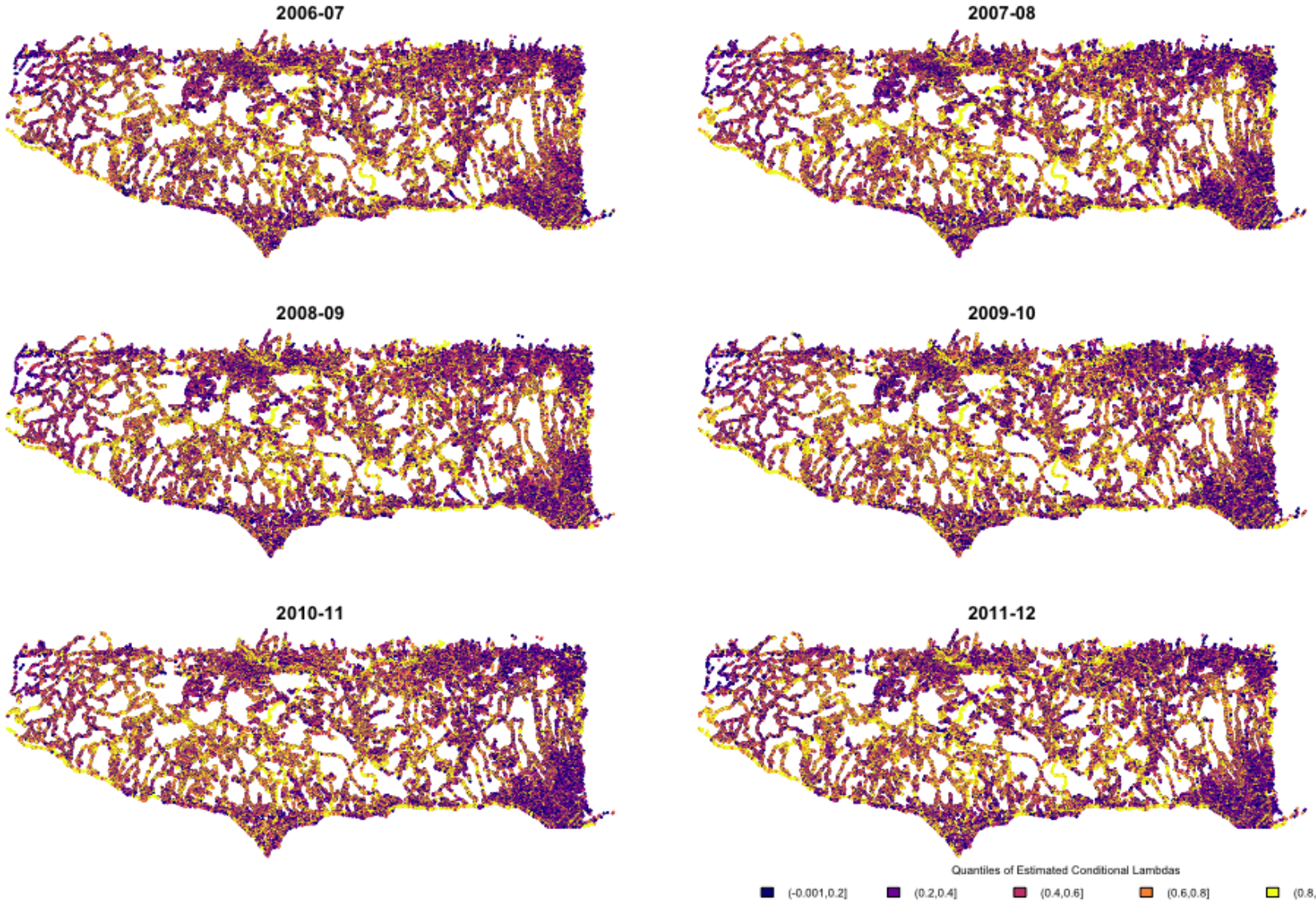


Figure 3.24: Estimated Conditional Intensity of Wildfire Ignition Occurrence from 2007 to 2012. See figure 3.26 for the legend.

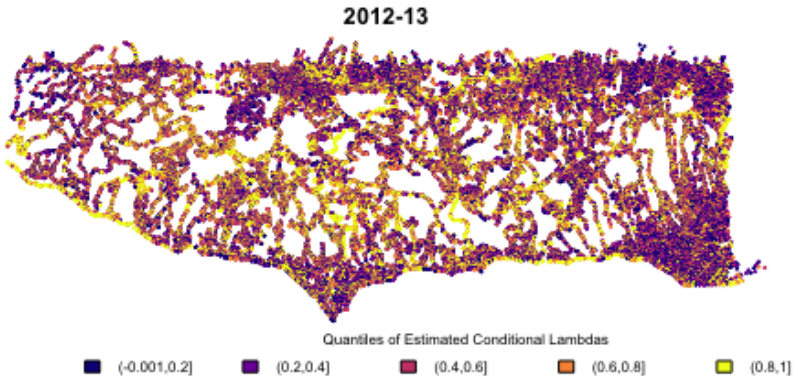


Figure 3.25: Estimated Conditional Intensity of Wildfire Ignition Occurrence from 2013. See figure 3.26 for the legend.




Figure 3.26: The legend for Figures 3.23, 3.24, and 3.25. Each interval represents the quantile of the predicted wildfire ignition formation intensity.

**Part II**

# Data Sufficiency and Privacy

# CHAPTER 4

# Introduction: Energy Atlas Database

The Energy Atlas project, set up and managed by the UCLA California Center for Sustainable Communities (CCSC), provides an exploratory framework access to the entire energy consumption of LA county (with the exception of a few cities) over a span of five years[1] This interactive framework is based on a core database that comprises of around 500 million records with individual service addresses and monthly energy consumption at parcel level. Along with the address-level energy consumption data ranging from 2006 to 2010, the database is augmented by the following variables:

- Levels of geocoding
    1. parcel
    2. street centerline
    3. ZIP code
    4. Utility provided locations
- Building type
- Building age
- Income characteristics using American Communities Survey (ACS) 2006-10 at block group level
- Demographic characteristics using ACS 2006-10 at block group level

[1]The database has now been expanded to cover 10 years, and also includes water usage data.

The public interface (Energy Atlas website: `http://energyatlas.ucla.edu`) of this project allows data downloads disaggregated by area levels, building types and consumption types, over the entire LA county. This complex undertaking provides the public access to the largest, disaggregated energy data in the nation. Moreover, the process of assembling this core database was the result of a multi-year negotiation between the CCSC, the LADWP and the California Public Utilities Commission (CPUC). It is crucial to note that, so far, most of the data collected by LADWP and other utility companies has not been shared with researchers, let alone the public. As a result, the Energy Atlas project sets a precedent in not only providing public access to high resolution databases but also expanding laws concerning data transparency. As statisticians, we understand the need for data to make the right decisions and researchers having access to such energy databases is paramount to making the right public policy decisions on energy sustainability.

## 4.1 Background and Motivations

The Energy Atlas database allows us the unique opportunity to not only study energy consumption patterns on a large scale, but also investigate the efficiency of such large databases with regard to scientific research. The core database comprises of multiple, sensitive covariates such as geographic location, square footage area and building usetype. As a result, the pressing issue the Energy Atlas project continually grapples with is how to make such a database accessible to be public, while balancing proper levels of data resolution and subject confidentiality. This chapter and the next extend this question a step further and ask if, and what an efficient level of resolution for large energy databases would be. This brings us to the issues of *data sufficiency* and *privacy* in the context of data transparency.

We can broadly define *data sufficiency* in this context as providing sufficient data resolution for scientific research while protecting individual confidentiality of the data subjects. In simpler terms, how much data is enough to gather for thorough scientific study while also maintaining proper subject anonymization? Another important question is regarding the sufficient level of transparency (data resolution) large database holders like energy utility

companies need to adhere to, which raises the issue of public information property. The US Census Bureau, one of the largest holders of public data, adheres to the "72 year rule" that permits them to release personally identifiable information only 72 years after it was collected for the decennial census. While this is to safeguard the privacy of the census subjects, such delayed releases of high resolution energy databases will not just be inefficient but futile in making immediate public policy decisions on energy use. This underscores the need to balance both data sufficiency and transparency with data privacy. Currently, the US Census is beginning to implement differential privacy methods on its websites (Dajani et al. (2017)) and this will be discussed further in Chapter 5.

Computer scientist at University of Warwick, Graham Cormode (Cormode (2015)) claims that "Guaranteeing privacy and being able to share useful data stand in fundamental opposition". However, it is important to explore and understand if complete data resolution is absolutely necessary for all scientific study or if certain omnibus statistics could fulfill *data minimization* requirements for specific scientific questions while maintaining proper privacy. Another important issue this brings forth is the need for data *confidentiality* mechanisms allowing scientific access while maintaining subject *privacy*; as maintained by the US Federal Statistical Data Research Centers for the US census data. These issues, along with ideas of differential privacy are explored further in Chapter 5.

### 4.1.1 Senate Bill 350: Contextual outliers

While the issues of data privacy, privacy preserving measures and data sufficiency are reviewed under Chapter 5, the rest of this Chapter aims to fulfill more immediate goals only possible through access to high resolution data. This is done by posing a series of narrower, specific questions that employ high resolution data; fulfilling these smaller goals could lend insight to the general *data sufficiency* problem.

A current policy goal that benefits from the Energy Atlas database is outlined in the California Senate Bill 350 titled Clean Energy and Pollution Reduction Act of 2015footnote-Was signed into law on October 7,2015. (de León (2015)). The two, overarching goals of

this bill are to increase Renewable Portfolio Standard of utility companies to 33% by 2020 and to 50% by 2030, and double statewide energy efficiency savings in electricity and natural gas end uses by 2030. The proposed steps to achieve the second goal include updating the Assembly Bill 758: Existing Building Energy Efficiency Action Plan, implementing the right energy efficiency retrofit programs etc.

One way to improve the proposed energy efficiency retrofit targeting is to fit a longitudinal quantile regression model, to the Energy Atlas database, which will provide contextual distribution of energy consumptions for a specific combination of covariates. In other words, a longitudinal quantile regression model can reveal how different covariates affect energy consumption at various quantiles. This can improve retrofit targeting towards buildings and units that disproportionately contribute to higher quantile energy consumption. The following sections summarize quantile regression methods, the Energy Atlas data on the Westwood neighborhood and the results of longitudinal quantile regression on Westwood data.

## 4.2 Quantile Regression

Simple quantile regression is outlined in Section 4.2.1 and longitudinal quantile regression through a Fixed-Effects model is outlined in Section 4.2.2

### 4.2.1 Review of Simple Quantile Regression

The quantile function of a scalar random variable Y is the inverse distribution function. The conditional quantile function of Y given X is the inverse of the corresponding conditional distribution function, i.e.,

$$Q_Y(\tau|X) = F_Y^{-1}(\tau|X) = \inf\{y : F_Y(y|X) \geq \tau\} \tag{4.1}$$

where $F_Y(y|X) = P(Y \leq y|X)$. The conditional quantile function of Y given X captures the relationship between Y and X. Quantile regression provides one way to study this relationship between X and Y. Given the following classical linear model

$$Y_t = \theta' X_t + u_t, t = 1, \ldots, n, \tag{4.2}$$

where $X_t$ are vectors of regressors including a constant, and $u_t$ are i.i.d. mean zero errors and are idependent of $X_t$, a regression of the above model can be conducted based on the following optimization problem:

$$\hat{\theta} = \min_{\theta} \sum_{t=1}^{n} \rho(Y_t - \theta' X_t), \tag{4.3}$$

where $\rho(\cdot)$ is a criterion (loss) function. If the routine quadratic loss function $\rho(u) = u^2$ is employed, the ordinary Least Squares estimator $\hat{\theta}_{OLS}$ is obtained from solving equation (4.3). However, for a conditional quantile estimate, Koenkar and Basset (1978) introduced an asymmetric loss function,

$$\begin{aligned} \rho(u) = \rho_t(u) &= u(\tau - \mathbf{1}(u < 0)) \\ &= (1-\tau)\mathbf{1}[u < 0]\,|u| + \tau \mathbf{1}[u > 0]\,|u| \end{aligned} \tag{4.4}$$

where $\tau \in (0, 1)$, and $\mathbf{1}(\cdot)$ is the indicator function. It is easy to see that when this loss function (4.4) is applied to equation (4.3), it minimizes the sum of asymmetrically weighted absolute errors. When $\tau = 0.5$, it is equivalent to minimizing (4.3) with $\rho(u) = |u|$, the Least Absolute Deviation criterion and this produces the Median of Y given X, or in other words, the conditional 50%th quantile regression solution.

Once $\hat{\theta}(\tau) = \min_{\theta} \sum_t \rho_\tau(Y_t - \theta' X_t)$ is solved using the asymmetrical loss function of (4.4), the $\tau$th conditional quantile function of $(Y_t|X)$ can be estimated by

$$\hat{Q}_{Y_t}(\tau|X_t) = X_t^{\top} \hat{\theta}(\tau) \tag{4.5}$$

Quantile regression is solved using the simplex method, a popular algorithm in linear programming.

### 4.2.2 Quantile Regression for Longitudinal data

Using the frameworks of classical linear fixed and random effects models, Koenker (2004) introduced quantile regression for longitudinal data. As the Energy Atlas database provides monthly energy usage at parcel level, the following model framework is presented in the context of monthly energy usage by individual megaparcels. Megaparcels are groups of

parcels that have the same GPS location, e.g. an apartment complex; more details about megaparcels are available under the Westwood data Section 4.3.

Within the context of monthly energy usage by a megaparcel, the conditional quantile function for the response $y_{ij}$ is,

$$Q_{y_{ij}}(\tau|x_{ij}) = \alpha_i + x_{ij}^\top \beta(\tau) + u_{ij}, \tag{4.6}$$

where $i = 1, \ldots, m$ megaparcels, $j = 1, \ldots, n_i$ months, $y_{ij}$ is the energy usage of the $i$th megaparcel in the $j$th month, $x_{ij}$ are the vectors of covariates on the megaparcels, and can be constants, $\alpha_i$ is the unobservable individual effects introduced by the $m$ megaparcels and $u_{ij}$ are the month specific errors. Estimating the above model (4.6), for several quantiles involves solving the following objective function:

$$\min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{j=1}^{n_i} \sum_{i=1}^{m} w_k \rho_{\tau_k}(y_{ij} - \alpha_i - x_{ij}^\top \beta(\tau)) \tag{4.7}$$

where $\rho_\tau(u)$ is the piecewise linear quantile loss function, as introduced in (4.4). The weights $w_k$ control the relative influence of the $q$ quantiles $\{\tau_1, \ldots, \tau_q\}$, on the estimation of the $\alpha_i$ parameters (Koenker (2004)).

In this fixed effects model, the unobservable, megaparcel effects $\alpha_i$'s represent a megaparcel-level heterogeneity that is not captured by the rest of the covariates $x_{ij}$, and are also *correlated* with them in an unspecified way. One example is to consider how certain megaparcel management and ownership styles could cause its consumption to correlate with their corresponding independent variables $x_{ij}$'s like building age, function and square footage.

In this formulation, the megaparcel specific effects provide a pure location shift on the conditional quantiles of energy usage. The effects of the covariates $x_{ij}$ are allowed to depend on the quantile of interest $\tau$, while the $\alpha_i$'s do not. It is possible to estimate quantile specific estimates of $\alpha_i(\tau)$'s, but care should be taken to make sure there are enough observations under each individual.

Koenker (2004) also introduces a *penalized* quantile regression with fixed effects, particularly when $m$ is much larger than the panel count of $n_i$. The main advantage of introducing

the $l_1$ penalty is to control the variability introduced by estimating a large number of $\alpha$ parameters. This penalized version of (4.7) is

$$\min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{j=1}^{n_i} \sum_{i=1}^{m} w_k \rho_{\tau_k}(y_{ij} - \alpha_i - x_{ij}^{\top}\beta(\tau)) + \lambda \sum_{i=1}^{m} |\alpha_i| \,. \tag{4.8}$$

When $\lambda \to 0$ we obtain the fixed effects estimator described under (4.7), and when $\lambda \to \infty$, the $\hat{\alpha}_i \to 0$ for all $i = 1, \ldots, m$ and we obtain an estimate of a model without any fixed effects. This penalized fixed-effects, quantile regression model is fitted on the Westwood energy consumption data using the Koenker co-authored R package '*rqpd*' available on R-Forge.

## 4.3 Westwood Energy Consumption Data

This sections provides an overview of the energy consumption data from the Westwood neighborhood from January 2006 to December 2010. Only the Westwood neighborhood is considered for now as it is a familiar area and allows for easier identification of data anomalies. The Energy Atlas team obtained monthly electric and gas consumption data at parcel level from all utilities that serve the entire Los Angeles county. A parcel, loosely defined, refers to a piece of land or lot owned or meant to be owned by an owner/individual. This can include individual apartment units, condominiums and floors in multi-floor buildings. This indicates that parcel level energy consumption data is the lowest resolution of energy consumption data that can be obtained from a utility company; it is the account level data that an individual or an owner might retain with an energy utility company.

The Westwood neighborhood, served by the Los Angeles Department of Water and Power (LADWP), consists of about 118,000 parcels. The Energy Atlas team augmented the parcel level data with information from the Los Angeles County Office of the Assessor (`https://assessor.lacounty.gov/`), and included variables on coordinate and street location, building usetype, age and area in square footage. Aside from single family houses, as most buildings contain more than one parcel, the Energy Atlas team introduced a new parcel unit called *megaparcel*. A megaparcel is comprised of multiple parcels with the same

coordinate and street location. As a result, the megaparcel's monthly energy consumption is the sum of monthly energy consumptions of all the parcels it contains. The megaparcel's area and usetype are cumulated similarly. The 118,000 parcels are contained in 4458 megaparcels with each megaparcel containing anywhere from 1 to 250 parcels.

Here onwards, energy consumption is only referred to and modeled at the megarparcel level. Out of the 4458 megaparcels in Westwood, 200 of them have been removed as they have more than 50% of missing data over the 60 months from January 2006 to December 2010. The rest of 4258 megaparcels have building usetype, building age and area as covariates. It is helpful to note that the Energy Atlas database is also augmented with demographic and income characteristics through the American Communities Survey (ACS) 2006-10 at the *block group* level, but the analysis done here is at a higher resolution of *megaparcel* level. While it is generally preferable to have more descriptive covariates at the unit level of the model, the model presented below does not make use of the *block group* level demographic features; this decision is discussed further under model results and discussion in Section 4.5.

The breakdown of building usetype of the Westwood megaparcels are shown in Table 4.1 below. The Table reflects the 4258 megaparcels left after removing missing data observations. The map in Figure 4.1 provides the spatial breakdown of building usetype in Westwood. Majority of the megaparcels are of the residential kind, including single family, condominium and multi family units. However, the area of the commercial, institutional and mixed use megaparcels are the largest. The breakdown of square footage of Westwood megaparcels can been seen in Figure 4.2. The UCLA campus and a few other megaparcels have missing data, but the Wilshire corridor of buildings have some of the largest square footage along with the Los Angeles Country Club at more than 50K sq. feet. The Wilshire corridor of buildings are a mix of commercial, mixed use and multi floor residential buildings. Finally, majority of the Westwood building structures were built before 1950 (See Figure 4.3). The Wilshire corridor buildings were build after 1978 and Westwood is peppered with a few buildings built after 1990. We are missing buiding age and square footage from UCLA owned megaparcels and a few other megaparcels of usetype "other"; these miscellaneous megaparcels include a few schools, a cemetary and a park.

| Usetype | Number of megaparcels |
|---|---|
| Mixed use | 8 |
| Institutional | 23 |
| Residential other | 39 |
| Other | 41 |
| Condo | 210 |
| Commercial | 243 |
| Multi family | 718 |
| Single family | 2976 |

Table 4.1: Breakdown of Westwood megaparcel usetypes

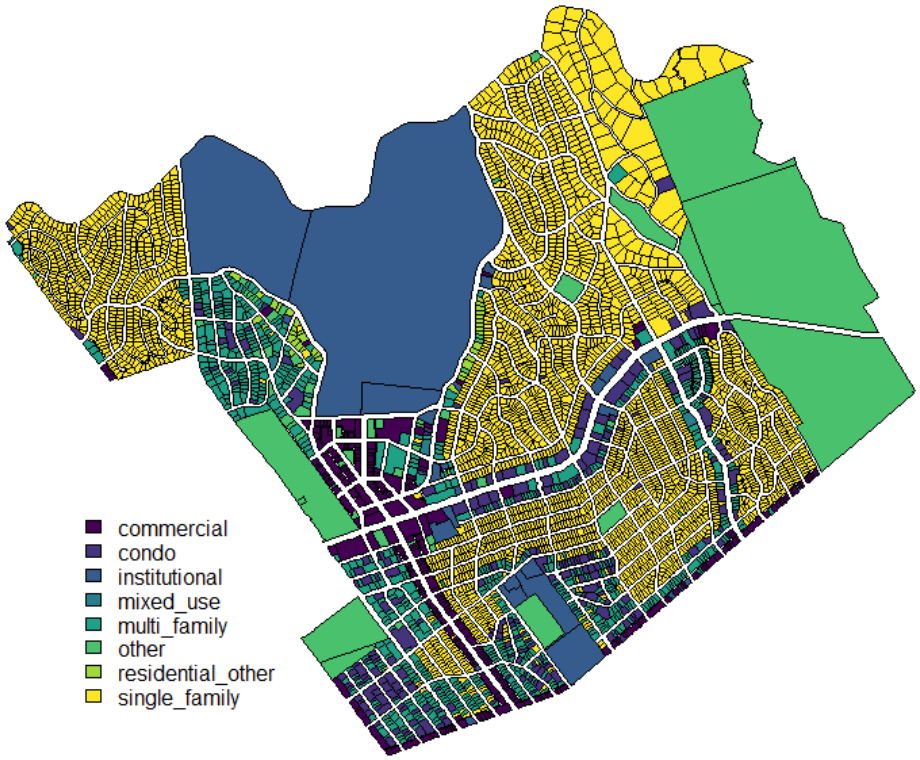


Figure 4.1: Usetype of Westwood Megaparcels

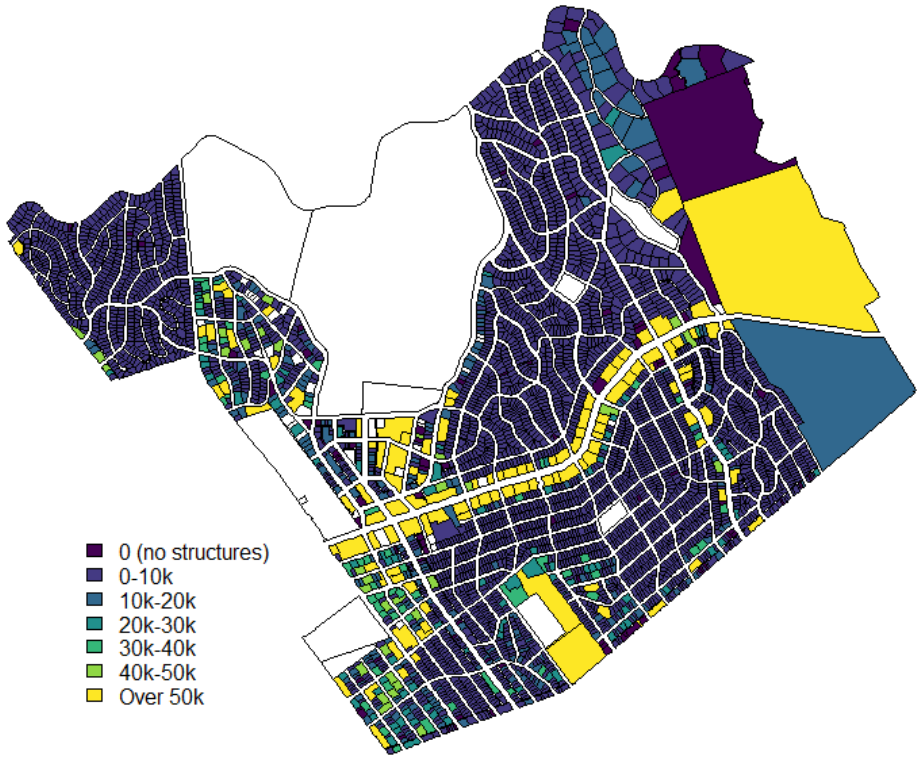


Figure 4.2: Square Footage of Westwood Megaparcels

## 4.4 Fixed Effects Quantile regression on Westwood data

A fixed effects, quantile regression model was fitted over the Westwood megaparcel data using the "*rqpd*" (Regression quantiles for panel data) package available on R-Forge. The model presented below is the penalized version of the fixed effects, quantile regression model outlined in Section 4.2.2. For each specified quantile $\tau$, the model minimizes

$$\min_{(\alpha,\beta)} \sum_{k=1}^{q} \sum_{j=1}^{n_i} \sum_{i=1}^{m} w_k \rho_{\tau_k}(y_{ij} - \alpha_i - x_{ij}^{\top}\beta(\tau)) + \lambda \sum_{i=1}^{m} |\alpha_i| \,. \tag{4.9}$$

Equal weights($w_k$) of 0.2 were chosen for the five specified quantiles($\tau$) of $(0.1, 0.25, 0.5, 0.75, 0.9)$. The L1(lasso) penalty term($\lambda$) that shrinks the fixed effects coefficients to zero is left at it's default value of 1. The results of the fixed effects quantile regression model are shown in Table 4.2.

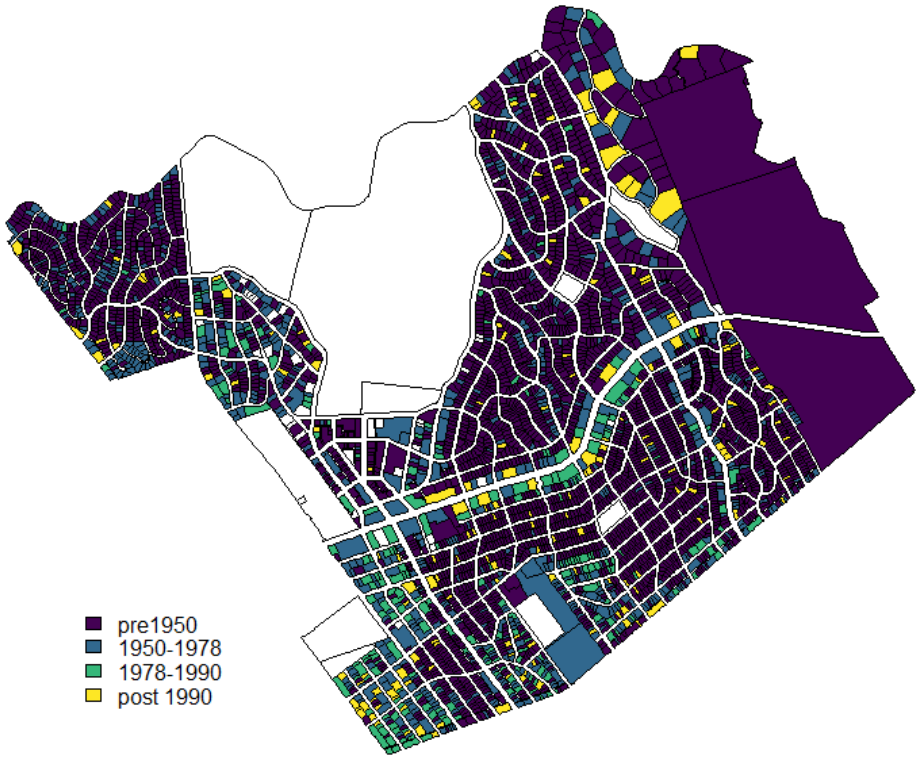


Figure 4.3: Building Age of Westwood Megaparcels

| Parameter | 10% | 25% | 50% | 75% | 90% |
|---|---|---|---|---|---|
| (Intercept) | 6681.07*** | 11390.62*** | 22230.21*** | 46837.68*** | 64200.30*** |
| | (1307.53) | (826.16) | (1243.14) | (2440.77) | (2258.43) |
| Sq. Footage | 0.21*** | 0.43*** | 0.62*** | 1.09*** | 1.50*** |
| | (0.01) | (0.008) | (0.11) | (0.03) | (0.02) |
| Year Built | -3.59*** | -5.37*** | -9.08*** | -18.61*** | -22.48*** |
| | (0.66) | (0.42) | (0.61) | (1.28) | (1.13) |
| Usetype-Condo | -1286.93*** | -4018.25*** | -8501.94*** | -14275.93*** | -23982.99*** |
| | (288.35) | (248.37) | (188.64) | (289.84) | (449.57) |
| Usetype-Institutional | -1484.30*** | -3899.66*** | -6897.35*** | -13855.94*** | -22254.61*** |
| | (292.90) | (268.40) | (221.17) | (316.103) | (646.33) |
| Usetype-Mixed Use | -236.06 | -2270.01*** | -7268.31*** | -12150.12*** | -22644.09*** |
| | (290.21) | (310.09) | (333.184) | (449.94) | (513.67) |
| Usetype-Multi Family | -849.77*** | -2571.53*** | -6675.86*** | -13508.86*** | -23389.48*** |
| | (166.34) | (112.07) | (175.39) | (284.56) | (437.71) |
| Usetype-Other | -2129.38* | 743.0425** | -1252.96** | -1668.64 | 15261.27** |
| | (1067.37) | (370.58) | (612.0) | (3128.28) | (6694.72) |
| Usetype-Residential Other | 229.77*** | -511.951** | -4135.86*** | -12003.74*** | -22921.48*** |
| | (535.12) | (229.88) | (241.37) | (345.13) | (470.13) |
| Usetype-Single Family | -405.71*** | -1763.714*** | -5498.87*** | -11852.50*** | -21638.21*** |
| | (136.22) | (74.35) | (143.68) | (285.97) | (450.53) |
| No. of parcels | 249.28*** | 178.149*** | 174.86*** | -239.19*** | -517.94*** |
| | (27.62) | (17.37) | (19.25) | (36.13) | (26.61) |

*: significance at 10%, **: 5%, ***: 1%

Table 4.2: Fixed Effects Quantile Regression model on Westwood Energy Consumption data with $\tau = (0.1, 0.25, 0.5, 0.75, 0.9)$. The bootstrap standard errors are produced using the method introduced in Bose and Chatterjee (2003) where the individual months rather than individual megaparcels are unit exponentially weighted.

### 4.4.1 Discussion

As seen in Table 4.2, only the megaparcel level covariates were used to fit the fixed-effects quantile regression model. The usetype category has 8 factor levels with Commercial megaparcels as the reference level. Majority of the megaparcels in the Westwood neighborhood are either Single family or Multi family.

At the 10% quantile, the area of the megaparcel (sq. footage) has a slight positive effect on the monthly energy consumption. The Year built has a negative effect on the monthly energy consumption, indicating that newer buildings are consuming less than older buildings at the 10% quantile. Compared to commercial megaparcels, almost all other usetypes express lower monthly consumption. The 'Institutional' and 'Other' megaparcels have the lowest levels of monthly consumption, while the 'Residential-Other' megaparcels have the highest level of monthly consumption among all usetypes. This indicates that at the 10% quantile, 'Residential-Other' megaparcels are the largest monthly consumers followed by 'Commercial' megaparcels. As can be seen in Figure 4.1, most of the 'Residential-Other' megaparcels are fraternity and sorority houses flanking the UCLA campus on Gayley and Hilgard Avenues.

As expected, the number of parcels has a large positive effect on a megaparcel's monthly energy consumption. In Westwood, the number of parcels in *all* the 'Residential-Other' *and* 'Commercial' buildings is only 1. As a result, at the 10% quantile, controling for all other covariates, a 'Residential-Other' building's monthly energy consumption is about 230 units higher than a Commercial building's monthly consumption in the Westwood neighborhood.

At the 25% quantile, square footage of a megaparcel has a slight positive effect on the monthly energy consumption. The year built has an increased negative effect on consumption, indicating newer buildings are consuming less than the older buildings. Under megaparcel usetype, condos are the smallest energy consumers compared to commercial buildings, while usetype-Other are the largest consumers, at 740 units per month over commercial buildings. The major Westwood megaparcels categorized as usetype-Other include the Los Angeles Country Club, Holmby Park, an Elementary school, a Middle school, Westwood Park/Recreation Center and the UCLA owned housing/other buildings along Weyburn

Place. Due to tax categorization, most of the megaparcels categorized as 'Other' do not have square footage information on them, but most of them are considered single parcels with the Los Angeles Country Club having an area of over 50K square footage (Figure 4.2). This indicates that at the 25% quantile, the largest consumers are the megaparcels with relatively larger square footage and have miscellaneous and/or public functions to them. No. of parcels still maintains a positive effect on energy consumption but lower than it did at the 10% quantile.

At the 50% quantile, the area of the megaparcel has less than unit positive effect on energy consumption and the year built of the megaparcel has an increasingly negative effect on the energy consumption. Newer megaparcels have more energy savings than older megaparcels. With regard to usetype, at the 50% quantile, the commercial megaparcels are the largest consumers while the condominium megaparcels are the smallest consumers for a given parcel. The no. of parcels in the megaparcel has a positive effect on the energy consumption. As the no. of parcels in condominium megaparcels range from 1 to 250, the monthly energy consumption of a single parcel, commercial unit is the same as a 48 parcel, condominium building, for the same footage area at 50% quantile.

At both 75% and 90% quantiles, the area of the megaparcel has more than a unit positive effect on energy consumption for every square foot increase and the year built has a large negative effect on the energy consumption. As megaparcel consumption is growing, newer buildings are having increasingly more energy savings than the older buildings. At the 75% quantile, commercial megaparcels continue to be the largest monthly consumers for a single parcel, while the condominiums are the lowest consumers. At the 90% quantile, usetype-Other megaparcels are the largest monthly consumers for a single parcel, while condominiums are the lowest consumers. However, interestingly, the effect of no. of parcels on monthly consumption is *negative* for the 75% and 90% quantile consumers. At these higher percentiles, an increase in the no. of parcels is leading to a decrease in monthly consumption, possibly due to increased building efficiency. However, as all commercial megaparcels in Westwood only have a single parcel, they do not accrue these increasing benefits that multi-parcel usetypes accrue.

While each of the 5 quantile models present effects of covariates at that quantile level, the change of covariate effect across the quantiles is insightful in narrowing down specific users that could benefit from energy consumption management programs. The change in covariate values across the quantiles is presented in Figure 4.4 and 4.5 and the bands on the estimated covariate represent the bootstrap confidence interval. Each of the covariates and its trend over the quantiles is discussed in detail below.

**Square Footage**: As expected, area in square footage has an overall positive effect on monthly energy consumption. However, the effect is relatively small on the lower quantiles and as we reach 75% and 90% quantiles, the effect is more than unit increase. This indicates that the higher quantile consumers increase their energy consumption more than lower quantile consumers for every square footage.

**Year Built**: The Year built indicates the age of the megaparcel, and the variable ranges from 1843 to 2013. For every unit increase in the year built, the megaparcels experience decreasing energy consumption across the quantiles. This gain in energy savings is highest at the 90% quantile. This is most likely due to improved building efficiency rules laid on newer buildings. The model could be improved by using a categorical variable that represents ranges of years that implemented progressive building efficiency rules.

**Usetypes–Condominium, Institutional, Mixed Use, Multi Family and Residential Other**: These five usetypes have expressed a steep decreasing trend of energy consumption, compared to commercial parcels of the same size, over the range of quantiles. The increased energy savings at the 70% and 90% quantiles are most likely a result of the large area of the buildings under these usetypes. While the megaparcels under these usetypes have a range of parcels in them(institutional, multi family and Residential Other have only 1 parcel), the mean square footage of these buildings is greater than 20K. Aside from Institutional buildings, the rest of the usetypes are various kinds of residential buildings. The trend of lower energy consumption over the quantiles indicates that these usetype megaparcels are much more energy efficient than their commercial counterparts of the same size and age.

**Usetype-Other**: As mentioned above, the usetype-Other megaparcels are a combina-

tion of miscellaneous buildings that include parks, schools, churches, a country club and some institutional buildings. This is the only usetype that has shown an increase in consumption over the range of quantiles. Compared to commercial megaparcels, the usetype-Other buildings show a small spike at the 25% quantile and a larger spike at the 90% quantile. The large spike at the 90% quantile is likely due to the large area megaparcels under this category.

**Single Family**: The single family megaparcels have constantly consumed less energy than their commercial counterparts at all the quantile levels. However, their consumption takes a greater nose dive after the 50% quantile. Approximately 70% of the megaparcels in the Westwood neighborhood are single family units made up of a single parcel. Despite being comprised of a single parcel, the single family houses have a very wide range of area in square footage. The median area is 2700 sq. feet and about 16% of the single family homes have an area greater than 4000 sq. feet. In spite this, the single family units show large energy savings at the 75% and 90% quantile levels in comparison to their commercial counterparts with similar sq. footage.

**No. of Parcels**: This usetype, besides usetype-Other, starts off with a positive effect on energy consumption and takes a nose dive after 50% quantile, resulting in a large negative effect on energy consumption at 75% and 90% quantiles. It is important to note that due to taxation rules, most commercial, institutional, multi family and residential other usetype are considered single parcel units; single family are also single parcel units. As a result, only condominiums, mixed use and usetype-other megaparcels are made up of more than 1 parcel. This indicates that these multi-parcel buildings accrue energy savings extra savings for every parcel. These same buildings have higher energy consumption at the quantiles below 50%, when compared to their single-parcel commercial counterparts.

### 4.4.2 Actionable Results

The following outline some immediately useful results that can be used to target energy efficiency and other energy management programs for improved savings.

- At the 10% quantile, the largest consumers compared to commercial buildings are

the usetype Residential Other. Residential Other buildings are all single-parcel units mostly built in the 1950s and are spatially identified as the fraternity and sorority houses on Gayley and Hilgard Avenues flanking the UCLA campus. Controlling for area and year built, these fraternity and sorority houses are consuming more than any other usetype in Westwood at the 10% quantile. They would probably make good targets for energy efficient retrofits and other building standard updates.

- The commercial megaparcels (all single-parcel) have shown consistently higher energy consumptions at all quantile levels; a few exceptions include Residential Other and Other usetypes. At the higher quantiles, condominiums accrue savings due to having a large number of parcels but commercial buildings do not as they are all single-parcel units. Similarly, across the quantiles, even as sq. footage effect grows, large area institutional buildings are much more efficienct than their large area commercial counterparts. As a result, a closer look at all the commercial megaparcels is necessary to understand how building function and management differ from institutional and condominium buildings to cause the observed energy consumption disparity.

- Finally, the usetype-Other also requires a closer look as it exhibits a small spike at the 25% quantile and a larger spike at the 90% quantile, causing it to be a greater consumer than commercial units for a unit sq. foot. Most of usetype-Other units are single parcel but have missing area on most, possibly due to their tax categorization. From the map in Figures 4.1 and 4.2, we know that the largest building by area under usetype-Other is the Los Angeles Country Club and is possibly the largest consumer at the 90% quantile.

## 4.5 Conclusion

This chapter was begun by introducing the first of its kind, Energy Atlas database and the many questions of data privacy and sufficiency it brought with it. The intended approach was to fulfill smaller policy objectives through modeling high resolution data, while maintaining

an outlook on how data resolution and auxiliary databases affect modeling and its usefulness for policy goals. Fixed effects quantile regression was chosen to understand the factors leading to high quantile consumption. While there are a myriad of ways to analyze the Energy Atlas database, quantile regression was chosen specifically to underscore the need for high resolution data to understand usage patterns a small neighborhoods like Westwood.

While the model itself has revealed some interesting and useful results, as outlined in the previous section, this was not without extra aid from Google Maps and the LA County Office of the Assessor website that provided parcel area, age and usetype. Google Maps was specifically used to correct parcel mis-categorizations and identify the high consuming Residential-Other units as the Fraternity and Sorority houses in Westwood. This high degree of resolution is immediately useful in knowing where to utilize energy efficiency programs within Westwood, and also aid in future causal analysis of energy retrofit programs. Overall, the majority of the analyses performed in this Chapter is not possible without the high resolution, parcel level data on Westwood and the mentioned public data.

This brings us to the question of data privacy and sufficiency. While high resolution data is key to producing detailed results to fulfill policy objectives, the issues of Consumer Energy Use Data (CEUD) ownership and protection are still up for debate under most jurisdiction in the country. However, privacy preserving and disclosure limitation methods can be employed to protect consumer privacy while exploring sensitive data. These questions of data ownership, protection and privacy preserving methods in the context of Consumer Energy Use Data are explored in detail in the following chapter.

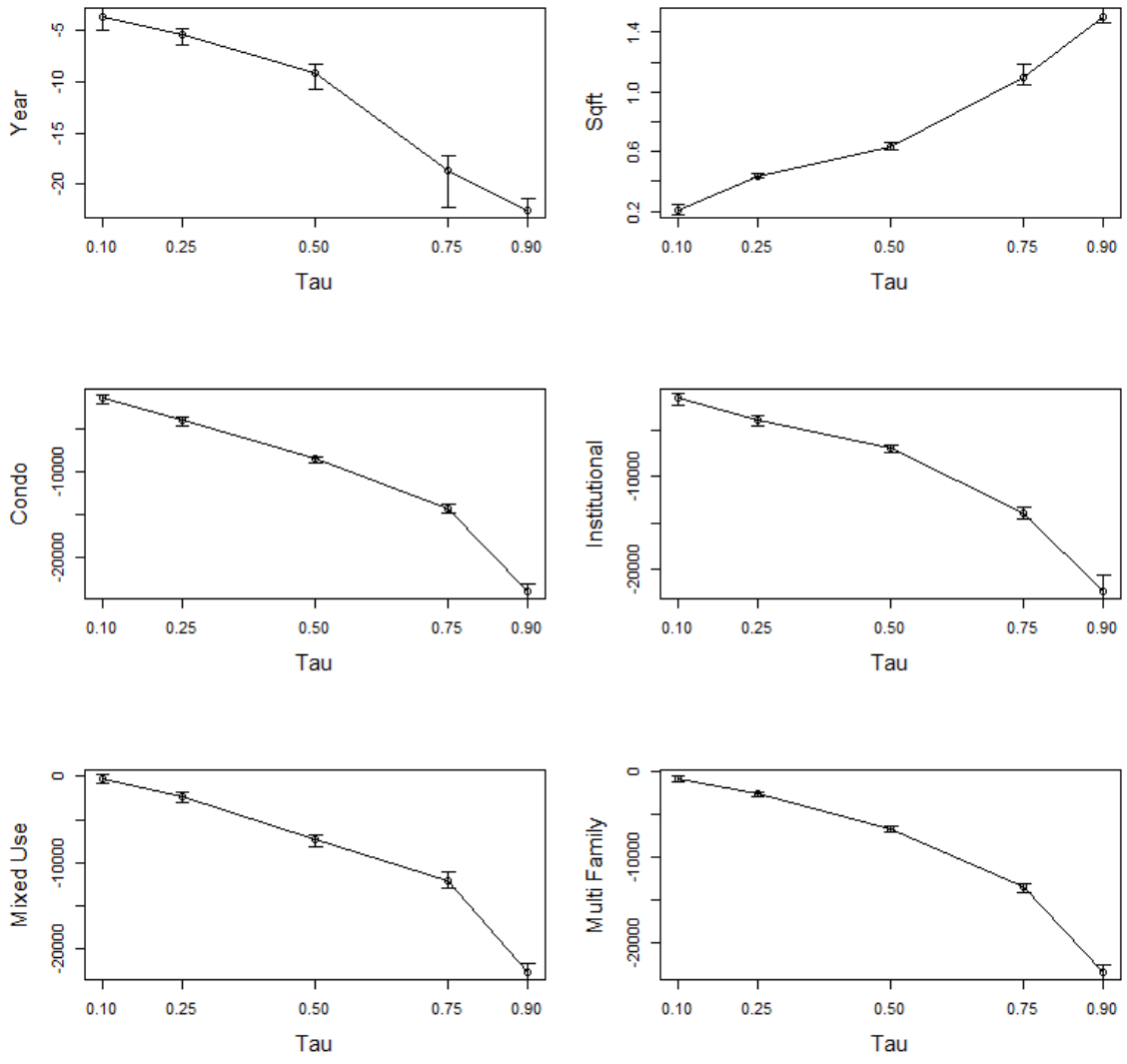


Figure 4.4: Quantile Regression coefficient trends

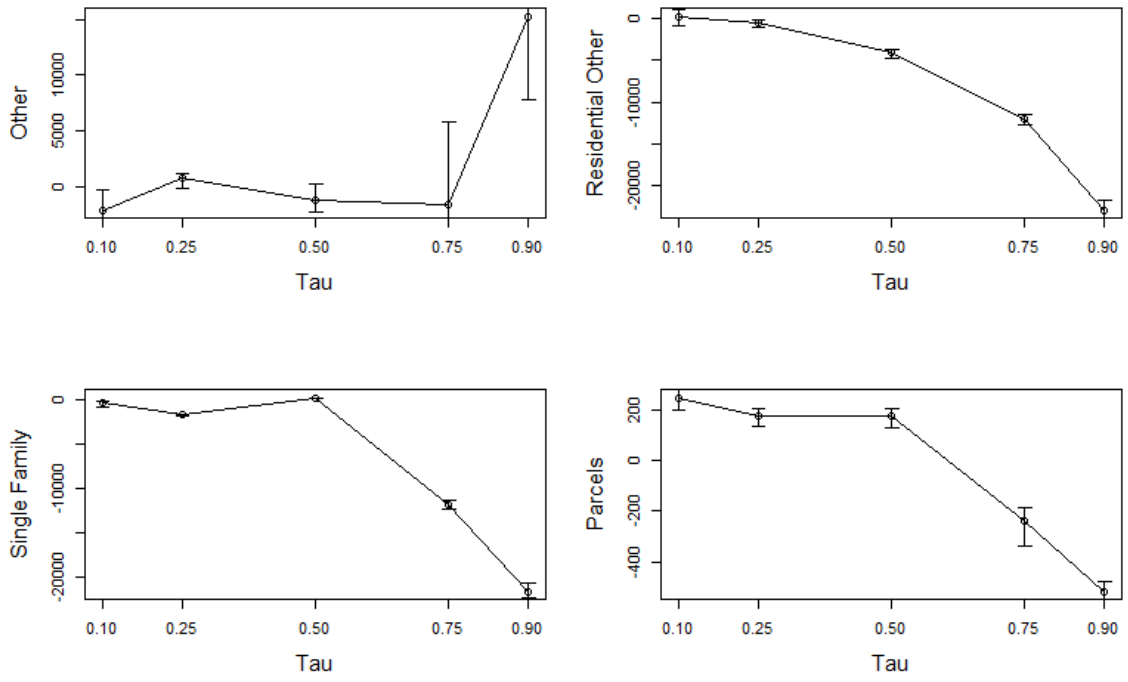


Figure 4.5: Quantile Regression coefficient trends

# CHAPTER 5

# Data Privacy and Confidentiality

## 5.1 Introduction

The motivation for this chapter has risen out of the compilation and existence of the Energy Atlas database and project. The Energy Atlas database, that supports the Energy Atlas website, is the largest set of disaggregated energy use data in the nation. The database consists of monthly, address-level (parcel-level), electricity and gas consumption from the entire Los Angeles County (and expanding) from 2006 to 2010. By order of the California Public Utilities Commission (CPUC), all utilities serving the Los Angeles County were required to share address-level energy use data with UCLA's California Center for Sustainable Communities, and the Energy Atlas project under a Nondisclosure Agreement.

The Energy Atlas Project is a unique undertaking as it not only compiles high resolution energy use data over a large area but also provides the public access to download aggregated data through the Energy Atlas website (`https::/energyatlas.ucla.edu`). This data is unique as it has never been analyzed either to understand long term consumption distributions or the effect of multiple energy efficiency programs implemented.

The opportunity for scientists and policy makers is immense as the high resolution data can be leveraged to understand, manage, and implement energy conservation and efficiency programs to fulfil policy objectives like those outlined in the Senate Bill 350. However, multiple unresolved questions of data ownership, consumer privacy and protection impose obstacles that make data access and usage complicated. Two of these main issue are outlined below.

### 5.1.1 Data Ownership

As utilities supply energy to consumers and collect usage data for billing, the utilities are considered to be the owners of the consumer energy use data. However, with the rise of Community Choice Aggregators (CCA), the issue of data ownership is murkier. Community choice energy is an alternative to traditional investor-owned utility supply, where local entities aggregate individual consumers' buying power to seek alternative, generally green energy supply contracts. Sometimes, the individual consumers that are part of a CCA also contribute to the supply grid through green sources like solar power. In such cases, the utility distributes energy and collects usage data for billing *on behalf* of the CCA. Here, the rightful data owner is undefined or at least legally unclear.

Data ownership has important implications on consumer privacy and protection, as the onus of securing consumer privacy is generally on the data owner. However, as mentioned, the boundaries of data ownership, in the energy sector, are still undefined, complicating the role of utilities, consumers and regulatory bodies with regard to safeguarding consumer privacy.

### 5.1.2 Data Resolution

One of the unique features of the Energy Atlas database is that it is parcel-level monthly energy consumption data. This is an unprecedented level of data resolution (spatially and temporally) in energy data ever made available for scientific study. Given the privacy concerns of handling address level data, it is only available for study in a secure location on the UCLA campus. However, privacy concerns of parcel-level data resolution seem trivial when compared to those arising from smart meter technologies. Smart metering technologies, increasingly employed by utilities around the country, are capable of collecting and transmitting minute by minute energy consumption data that also breaks down to the exact appliance used in the household. This raises serious concerns of data privacy as individual household and parcel patterns can be discerned at a daily scale. Mind that the utilities collecting this data can be Investor-owned (IOU) or Publicly-owned (POU) utilities resulting in

differences in business allegiances. With out regulation and oversight, utilities could employ exploitative pricing schemes or even sell this high-resolution data to third parties. On the other hand, access to high-resolution data is imperative to making guided policy decisions, and scientists and policy makers need access to such data.

While the uncertainty surrounding data ownership involves legal policy, the issues of data privacy, disclosure limitation and related methods are already being explored by statisticians and computer scientists. The U.S. Census Bureau has long been known to employ statistical disclosure limitation methods on its released microdata to safeguard individual privacy, and increasingly, newer theoretical definitions of privacy have paved way to broader privacy methodologies. The rest of this chapter provides a basic overview of these statistical disclosure limitation and privacy methods, and more importantly, contextualizes their role in the current arena of data privacy, applicability and usability.

## 5.2 Literature Review

### 5.2.1 Statistical Disclosure Control

Dalenius (1977) in "Towards a methodology for Statistical Disclosure control" begins to formalize the idea of disclosure in the context of data release. He emphasizes the importance of achieving a "reasoned balance between the right to privacy and the need to know", and by working through a framework, defines the concept of D-disclosure. Consider a specific object $O_K$ in a set of identifiable objects (total population), denoted by $\{O\}_T$. Given a characteristic $D$, which may be one of many survey statistics $X, Y, \ldots, Z$, $D_K$ is the characteristic value of object $O_K$. Here, D-disclosure has taken place if the release of the statistics $S$ makes it possible to determine the value of $D_K$ more accurately than is possible without access to $S$.

Dalenius (1977) continues to explore facets of the definition and extends its scope by considering external data, internal and external disclosure sources. Finally, he proposes the criteria of

- operational feasibility and

- maximizing benefit of statistics release while maintaining some accepted level of disclosure

for developing methodologies for Statistical Disclosure Control. Dalenius is honest about the difficulty of constructing balanced measures of disclosure levels, and the arena of data collection, release and ownership has changed by bounds since then. Dalenius' criteria are revisited and contextualized for the present times in the following sections, after a general overview of existing disclosure limitation and privacy methods.

### 5.2.2 Statistical Disclosure Limitation Methods

This section covers long-standing statistical methods used over the years to safeguard individual privacy while releasing microdata. The first step in anonymizing microdata is to remove obviously sensitive fields that are also knows as personal identifiable information (PII), which include name and date of birth. The subsequent methods used by many data releasing agencies (including the U.S. Census) include the following. See Matthews and Harel (2011) for more detail.

1. **Limitation of detail**: This technique includes reducing detail in a variable by binning or categorizing it into intervals. The U.S. Census does not release geographic identifiers that would leave a sub-population with less than 100,000 observations vulnerable.

2. **Top/Bottom coding**: This technique helps reduce disclosure risk of extreme values by categorizing them as 'greater than' or 'less than' a value rather than report the extreme value itself.

3. **Suppression**: In a contingency table, cells with too few observations are not released to the public as it is easy to infer the identity of the persons in that cell. Suppression of cells is used when some variable values are too unique that it allows easy de-identification.

4. **Rounding**: This is a technique that involves setting a rounding base and rounding the value of the cell to the base probabilistically. This technique safeguards sensitive

values by rounding them to the nearest base, but reduces the information retained in the microdata.

5. **Addition of noise**: This technique is extensively used and involves adding a random noise to the sensitive variable before being released. The perturbed data can be correctly modeled by accounting for the added noise. Noise can be added to discrete data by a technique called Post Randomization Method.

6. **Sampling**: Release of a sample rather than the whole population data has been used extensively for a long time. It is known to protect against linkage attacks where the attacker has auxiliary data about the population from external sources. The U.S. Census has long released Public Use-Microdata Sample (PUMS) data for quicker access.

7. **Matrix Masking**: Proposed by Cox, matrix masking is a method of adding ambient noise to a matrix. Given an $n$ by $p$ data matrix $X$, one would release $Y = AXB + C$, where $A, B$, and $C$, are appropriate conformable matrices. Special cases of matrix masking include noice addition, sampling, cell suppression and adding simulated data. The drawback with matrix masking is that it makes it difficult to analyze the data, requiring the analyzer to know the masking procedure to reverse it.

8. **Randomized response**: This technique is a way to protect survey respondents' confidentiality when they respond to sensitive questions. The response to the sensitive question is a result of a probabilistic outcome, where the respondent answers truthfully with some probability $p$ or answers untruthfully with probability $1 - p$. While this protects respondent privacy, the data analyzer must have knowledge about the randomization mechanism to analyze the data correctly. This technique could also be applied to microdata post collection, before the data release.

9. **Data swapping and shuffling**: This technique involves swapping or permuting the values of a sensitive variable within a dataset. This removes the relationship between the record and the respondent. While this is very simple to apply, the major drawback is that the multivariate relationships in the dataset are no longer retained, possibly

making the dataset less useful for specific analyses. Another drawback of this method is that it is difficult to maintain variable relationships if some variables are weighted.

10. **Synthetic data**: This technique, as the title indicates, is a way to release simulated data produced from a model fit on the original data. The initial idea, proposed by Rubin(1993), treats the sensitive data as missing values and replaces them using multiple imputation techniques. The sensitive variable values are replaced by random draws from an appropriate posterior distribution. The advantage of this technique is that if the right model is used, the synthetic data can easily be analyzed producing the same results as the original data.

11. **Micro-aggregates**: This technique involves releasing new records that are averaged or aggregated from at least three original records.

12. **Spatial data techniques**: Some of the techniques used to secure spatial data include spatial cloaking, noice addition and rounding to the nearest grid point. Spatial cloaking involves masking all locations within a certain radius of the record's location. The center of the circle is chosen randomly near the origin location. Rounding to the nearest grid point allows to reduce location resolution and grid size can be based on the required level of privacy.

While the above listed methods provide a general overview of statistical techniques used to mask and secure microdata, it is by no means a comprehensive review of all methods that exist and are employed. The above summaries are mostly sourced from Matthews and Harel (2011) and for a thorough review of statistical confidentiality, disclosure limitation and disclosure risks, see Duncan et al. (2011).

Th methods outlined so far are relatively easy to implement, with some exceptions such as synthetic data and swapping methods; this fulfills Dalenius' first criterion of operational feasibility of disclosure control methods. It is also important to note that most of these methods were developed over the past five decades, where the main goal was to secure privacy of records in specific microdata. While there is research that probes the quality of

these methods, and their defense against external data linkage attacks and other disclosure risks, the current arena of data collection and release has immensely changed over the last few decades. This makes their fulfillment of Dalenius' second criterion of maximizing benefit of statistics release while maintaining an accepted level of disclosure, a little unclear. While well modeled synthetic data is as good as the original data with respect to analysis, some methods like randomized response, sampling and matrix masking specifically require knowledge of the randomization or masking method to correctly analyze the data. This is somewhat at odds with Dalenius' criterion of maximizing data release benefit, as the question of who accrues the benefit from data release is an important and somewhat undefined one. The benefitting parties could include researchers, policy makers, private companies, etc.

Given the growing ability of technologies to collect and save data, and the myriad of easily available online public data sources, the range of risks is increasing without a clear sense of data release benefits and the parties accruing the benefits. The line between 'right to privacy' and 'need to know' is getting murkier, not just as a result of technological advancements that can easily track and collect data (like smart meters) but also due to lack of exploration and understanding of data privacy boundaries. One way is to pursue Dalenius' second criterion as an overall lighthouse objective while slowly chipping away at more specific questions such as defining the confines of data release benefit, accepted levels of privacy, etc. Similar facets of privacy and data collection are acknowledged in varying degrees by the Institutional Review Board that oversees human-subject research.

One such goal already has burgeoning research from computer scientists and statisticians, and is aimed as specific definitions of privacy and developing methodologies fulfilling them. Three such privacy models are overviewed in the section below.

### 5.2.3 Privacy Methods

The approach to privacy and disclosure limitation has increasingly moved towards developing methodologies that fulfill certain formulations of privacy. While this is a necessary shift, given the expanding formats of available data, it is important to understand the contextual

positions of these new methodologies in the current field of privacy and disclosure limitation. Methodologies fulfill a wide range of goals and their usefulness needs to be understood in the context of applications, operational feasibility and privacy protection provided. This section briefly outlines 3 definitions of privacy and the following sections will discuss their feasibility and applicability to large databases like the Energy Atlas database. The privacy methodologies discussed below are sourced from Fung et al. (2010), which also offers a more comprehensive view of all the existing privacy methods. For a thorough overview of privacy methodology, see Chen et al. (2009).

***k*-Anonymity**: Aside from Personally Identifiable Information (PII), which explicitly identifies individuals, quasi-identifiers (QID) are information or attributes that could potentially identify record owners, especially given auxiliary data. $k$-Anonymity is a method that aims to prevent record linkage through quasi-identifiers: if one record in the dataset has some value of QID, at least *k-1* other records in the dataset should have the same value of QID. A dataset satisfying this requirement is called *k-anonymous*.

While a *k-anonymous* table not only safeguards the identity of a record owner by masking them in a group of similar records, it also possibly thwarts record linkage in other datasets as QID in the *k-anonymous* table are not unique. However, this method assumes knowledge of QID in the data table. While PII are obvious, QID attributes are not always straightforward to select, especially when the data publisher has to anticipate the various kinds of linkage attacks. Sometimes, multiple QIDs are used and this increases data security but results in information loss as more uniformity is introduced into the dataset.

$k$-anonymity has been extended to counteract some of the drawbacks giving rise to (X,Y)-anonymity, MultiR $k$-anonymity etc. (Fung et al. (2010)), but if a certain table is small and most of the records have the same QID, record linkage is inevitable. Returning to Delanius' criteria for disclosure limitation methodology, $k$-anonymity is operationally feasible if the data table is not too large, but there is a large degree of information loss depending on the number of QIDs anonymized in the table, affecting the usefulness of the table. For instance, in the Energy Atlas database, masking the street location (PII) and anonymizing on QIDs such as building type, square footage, number of parcels is possible but reduces the

usefulness of the data, especially for specific modeling goals like those pursued in Chapter 4. This once again raises the question of who accrues the benefit of released data. While privacy is meant to secure the identity of the individuals in the data, data release and transparency has multiple possible benefactors. This issue needs closer consideration depending on the data owner and type.

***l*-Diversity**: This method aims to reduce attribute linkage by diminishing the correlation between QID attributes and sensitive attributes. It requires that every QID group contains at least $l$ well represented sensitive values. While there are many interpretations of what well represented means, $l$-diversity automatically fulfills $k$-anonymity (if $k$=$l$), as each QID value has at least $l$ records under it. The drawback is that a probabilistic attack is still possible as sensitive attributes in datasets are not always uniformly distributed.

Continuing with the Energy Atlas database example, every value under QIDs building type, square footage and number of parcels should contain at least $l$ unique values of sensitive attributes like street or block address. While this is possible for housing and residential building types as they are a large sample, it is not simple to achieve for very specific building types such as institutional or mixed-use types that are smaller in number. Moreover, if the wrong QID like building age is incorporated, the size of specific subsets of buildings dwindle quickly, making the application of $l$-anonymity trickier. Depending on the chosen $l$, $l$-anonymity is reasonably feasible to apply but does face issues of skewed attribute distributions.

**$\epsilon$-Differential Privacy**: This privacy methodology provides theoretical guarantees about privacy by making sure that the risk to the record owner's privacy does not change whether the record is included in the database or not. The concept introduced by Dwork (2006) is also referred to as $\epsilon$-Differential privacy where the addition or removal of a single record does not change the outcome of any analysis. More formally, a randomized function $F$ ensures $\epsilon$-differential privacy if for all datasets $D_1$ and $D_2$ differing on at most one record,

$$|\ln \frac{P(F(D_1) = S)}{P(F(D_2) = S)}| \leq \epsilon \tag{5.1}$$

for all $S \in$ Range(F), where Range(F) is the set of possible outputs of the randomized

function $F$. This method does not protect against record and attribute linkage attacks but it does guarantee that the information known about an individual participant does not change much should they choose to be part of the database. Differential privacy allows the randomized output to be such that the addition or removal of a specific record does not change the output drastically.

The methodology is defined as an interactive query model where the dataset is accessible only through data queries submitted to the data publisher; it is supposed to be adaptive to multiple queries but it is unclear how detailed the data queries can be and if the number of queries reach a limit to secure privacy. The methodology has also been shown to fail for network data where tuples affect data attributes, requiring the data to be independent for the methodology to work (Kifer and Machanavajjhala (2011)).

Despite the challenges, one the of most practical applications of a probabilistic version of differential privacy was employed on the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD) in Machanavajjhala et al. (2008) and the online public interface that handles the queries is at `https://onthemap.ces.census.gov/`. The LEHD database is a commute pattern table representing every U.S. worker's employment and household location with complementary attributes on employment type and industry. While the employment or destination block is already public data, the household or origin block of the worker is treated as the sensitive attribute and masked through a probabilistic differential privacy algorithm. The authors are honest about having to tweak the current differential privacy method to the specific problem at hand to produce an appropriate privacy method for the commuter data.

## 5.3 Discussion

The above sections on Statistical Disclosure Limitation and Privacy methods are by no means comprehensive and there are numerous other methodologies that explore and provide varying degrees of privacy protection. However, since the arena of data collection and storage has changed immensely over the last two decades, the goals and operational feasibility of existing

methods in the present context are either undefined or unclear. This is mainly a result of changing methodologies over the years, without a deliberate assessment of applicability and feasibility of the methods in the field.

One of the biggest open issues is that the applicability of a disclosure limitation or privacy method needs to be addressed in the context of data release benefactors. The spectrum of data utility is wide and privacy methods provide various degrees of data utility. Researchers can work on specifying this spectrum of data utility and categorize existing methods based on utility they retain. While the legality of data ownership is out of the purview of statisticians, the spectrum of fulfilling privacy while retaining data utility is an interesting statistical problem. One way to explore this is to apply existing methods to databases like the Energy Atlas and specify the utility retained by each method. This also requires for a specific definition of statistical utility.

It is apparent that while there are no 'one size fits all' solution or method to secure privacy, researchers should actively seek to demonstrate where their methodology falls in the data utility spectrum while being honest about the privacy guarantees. As research expands, it is necessary to contextualize the methodology and understand its limitations of use. Once again, Delanius' criteria of operational feasibility and maximizing data utility against an accepted level of privacy are appropriate goals that all privacy methodologies need to address directly.

### 5.3.1 Future Research

In the case of the Energy Atlas database, future work could involve modeling and simulating synthetic data at parcel or block-group level to make available for public release. At the same time, the synthetic data could be analyzed at various resolutions to test for data utility, and specifically compare results with those in Chapter 4. This provides an interesting contrast to study the idea of securing privacy while retaining data utility and resolution that can fulfil useful policy goals.

Bibliography

Abrevaya, J. and Dahl, C. M. (2008), "The Effects of Birth Inputs on Birthweight," *Journal of Business & Economic Statistics*, 26, 379–397.

Altman, M., Wood, A., O'Brien, D. R., and Urs, G. (2018), "Practical approaches to big data privacy over time," *International Data Privacy Law*, 8, 29–51.

Ang, Q. W., Baddeley, A., and Nair, G. (2012), "Geometrically Corrected Second Order Analysis of Events on a Linear Network, with Applications to Ecology and Criminology," *Scandinavian Journal of Statistics*, 39, 591–617.

Arellano, M. and Bonhomme, S. (2016), "Nonlinear panel data estimation via quantile regressions," *Econometrics Journal*, 19, C61–C94.

Bache, S. H. M., Dahl, C. M., and Kristensen, J. T. (2013), "Headlights on tobacco road to low birthweight outcomes: Evidence from a battery of quantile regression estimators and a heterogeneous panel," *Empirical Economics*, 44, 1593–1633.

Baddeley, A., Gregori, P., Mateu, J., Stoica, R., and Stoyan, D. (eds.) (2006), *Case Studies in Spatial Point Process Modeling*, vol. 185 of *Lecture Notes in Statistics*, New York: Springer, New York, NY.

Baddeley, A., Jammalamadaka, A., and Nair, G. (2014a), "Multitype point process analysis of spines on the dendrite network of a neuron," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63, 673–694.

Baddeley, A., Rubak, E., and Turner, R. (2015), *Spatial Point Patterns: Methodology and Applications with R*, London: Chapman & Hall/CRC Press.

Baddeley, A. and Turner, R. (2000), "Practical maximum pseudolikelihood for spatial point patterns," *Australian & New Zealand Journal of ...*, 42, 283–322.

— (2006), "Modelling spatial point patterns in R," in *Case studies in spatial point process modeling*, eds. Baddeley, A., Gregori, P., Mateu, J., Stoica, R., and Stoyan, D., Springer, pp. 23–74.

Baddeley, A., Turner, R., Mateu, J., and Bevan, A. (2013), "Hybrids of Gibbs Point Process Models and Their Implementation," *Journal of Statistical Software*, 55, 1–43.

Baddeley, A., Turner, R., Moller, J., and Hazelton, M. (2005), "Residual analysis for spatial point processes," *Journal of the Royal Statistical Society. Series B*, 67, 617–666.

Baddeley, A. B., Diggle, P. J., Hardegen, A., Lawrence, T., and Milne, R. K. (2014b), "On tests of spatial pattern based on simulation envelopes," *Ecological Monographs*, 84, 477–489.

Balch, J. K., Bradley, B. A., Abatzoglou, J. T., Nagy, R. C., Fusco, E. J., and Mahood, A. L. (2017), "Human-started wildfires expand the fire niche across the United States," *Proceedings of the National Academy of Sciences*, 114, 2946–2951.

Berman, M. and Turner, R. (1992), "Approximating Point Process Likelihoods with GLIM," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41, 31–38.

Besag, J. (1977), "Some methods of statistical analysis for spatial data," *Bulletin of the International Statistical Association*, 47, 77–92.

Besag, J., Milne, R., and Zachary, S. (1982), "Point Process Limits of Lattice Processes," *Journal of Applied Probability*, 19, 210–216.

Bose, A. and Chatterjee, S. (2003), "Generalized bootstrap for estimators of minimizers of convex functions," *Journal of Statistical Planning and Inference*, 117, 225–239.

Brown, L. D. (1986), "Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory," *Lecture Notes-Monograph Series*, 9, 1–279.

CAL FIRE (2013), "2013 Wildfire Activity Statistics Annual Report," Tech. rep., California department of Forestry and Fire Protection.

Chen, B.-C., Kifer, D., LeFevre, K., and Machanavajjhala, A. (2009), "Privacy-Preserving Data Publishing," *Foundations and Trends® in Databases*, 2, 1–167.

Clements, R. A., Schoenberg, F. P., and Veen, A. (2012), "Evaluation of space-time point process models using super-thinning," *Environmetrics*, 23, 606–616.

Cobb, V. (1994), "Effects of Temperature on Escape Behavior in the Cribellate Spider, Oecobius annulipes," *The Southwestern Naturalist*, 39, 391–394.

Cormode, G. (2015), "The Confounding Problem of Private Data Release," in *18th International Conference on Database Theory (ICDT 2015)*, eds. Arenas, M. and Ugarte, M., Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, vol. 31, pp. 1–12.

Cormode, G., Procopiuc, C. M., Shen, E., Srivastava, D., and Yu, T. (2013), "Empirical privacy and empirical utility of anonymized data," *Proceedings - International Conference on Data Engineering*, 77–82.

Dajani, A. N., Lauger, A. D., Singer, P. E., Kifer, D., Reiter, J. P., Machanavajjhala, A., Garfinkel, S. L., Dahl, S. A., Graham, M., Karwa, V., Kim, H., Leclerc, P., Schmutte, I. M., Sexton, W. N., Vilhuber, L., and Abowd, J. M. (2017), "The modernization of statistical disclosure limitation at the U . S . Census Bureau," .

Dalenius, T. (1977), "Towards a methodology for statistical disclosure control," .

de León, K. (2015), "SB-350 Clean Energy and Pollution Reduction Act of 2015," .

Duncan, G. T., Elliot, M., and Salazar-González, J.-J. (2011), *Statistical Confidentiality: Principles and Practice*, New York: Springer, New York, NY.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015), "Preserving Statistical Validity in Adaptive Data Analysis," *Science*, 349, 636–638.

Eckardt, M. and Mateu, J. (2017), "Analysing highly complex and highly structured point patterns in space," *Spatial Statistics*, 22, 296–305.

Fritzén, N. R. (2013), "The synatropic Oecobius navus (Araneae: Oecobiidae) established indoors in southern Finland," *Memoranda Societatis pro Fauna et Flora Fennica*, 89, 32–34.

Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. (2010), "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys*, 42, 1–53.

Geyer, C. (1999), "Likelihood Inference for spatial point processes," in *Stochastic geometry : Likelihood and Computation*, eds. Barndorff-Nielsen, O., Kendall, W., and van Lieshout, M., Boca Raton, Fla. :: Chapman & Hall/CRC,, chap. 3, pp. 79–140.

Hall, J., Cattanach, B., and Hammer, B. (2015), "So, who owns your energy-use data," *Windpower Engineering & Development*.

Hall, P. (1985), "Resampling a coverage pattern," *Stochastic Processes and their Applications*, 20, 231–246.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall.

Hering, A. S. and Bair, S. (2014), "Characterizing spatial and chronological target selection of serial offenders," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63, 123–140.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008), "Goodness of Fit of Social Network Models," *Journal of the American Statistical Association*, 103, 248–258.

Jammalamadaka, A., Banerjee, S., Manjunath, B. S., and Kosik, K. S. (2013), "Statistical analysis of dendritic spine distributions in rat hippocampal cultures." *BMC bioinformatics*, 14, 287.

Kaza, N. (2010), "Understanding the spectrum of residential energy consumption: A quantile regression approach," *Energy Policy*, 38, 6574–6585.

Kifer, D. and Machanavajjhala, A. (2011), "No free lunch in data privacy," *Proceedings of the 2011 international conference on Management of data - SIGMOD '11*, 193.

Koenker, R. (2004), "Quantile regression for longitudinal data," *Journal of Multivariate Analysis*, 91, 74–89.

— (2017), "Quantile Regression: 40 Years On," *Annual Review of Economics*, 9, 155–176.

Koenker, R. and Hallock, K. F. (2001), "Quantile Regression," *Journal of Economic Perspectives*, 15, 143–156.

Koenker, R. and Xiao, Z. (2006), "Quantile autoregression," *Journal of the American Statistical Association*, 101, 980–990.

Krivitsky, P. N. and Handcock, M. S. (2014), "A Separable Model for Dynamic Networks." *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 76, 29–46.

Kunsch, H. R. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17, 1217–1241.

Lahiri, S. N. (2002), "On the Jackknife-after-Bootstrap Method for Dependent Data and Its Consistency Properties," *Econometric Theory*, 18, 79–98.

— (2003), *Resampling Methods for Dependent data*, Springer-Verlag New York, 1st ed.

Liu, R. Y. and Singh, K. (1992), "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence," in *Exploring the Limits of Bootstrap*, eds. Lepage, R. and Billard, L., Wiley, New York, pp. 225–248.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008), "Privacy: Theory meets practice on the map," *Proceedings - International Conference on Data Engineering*, 277–286.

Mann, M. L., Berck, P., Moritz, M. A., Batllori, E., Baldwin, J. G., Gately, C. K., and Cameron, D. R. (2014), "Modeling residential development in California from 2000 to 2050: Integrating wildfire risk, wildland and agricultural encroachment," *Land Use Policy*, 41, 438–452.

Marino, M. F. and Farcomeni, A. (2015), "Linear quantile regression models for longitudinal experiments: An overview," *Metron*, 73, 229–247.

Matthews, G. J. and Harel, O. (2011), "Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy," *Statistics Surveys*, 5, 1–29.

Mrkvička, T., Soubeyrand, S., Myllymäki, M., Grabarnik, P., and Hahn, U. (2016), "Monte Carlo testing in spatial statistics, with applications to spatial residuals," *Spatial Statistics*, 18, 40–53.

Nychka, D., Furrer, R., Paige, J., and Sain, S. (2017), "fields: Tools for spatial data," .

Nychka, D. W. (2000), "Spatial Process Estimates as Smoothers," in *Smoothing and Regression. Approaches, Computation and Applications*, ed. Schimek, M. G., Wiley, New York, pp. 393–424.

Okabe, A. and Sugihara, K. (2012), *Spatial Analysis Along Networks: Statistical and Computational Methods*, Chichester: John Wiley.

Radeloff, V. C., Hammer, R. B., Stewart, S. I., Fried, J. S., Holcomb, S. S., and McKeefry, J. F. (2005), "The wildland-urban interface in the United States," *Ecological Applications*, 15, 799–805.

Schoenberg, F. P., Chang, C.-H., Keeley, J. E., Pompa, J., Woods, J., and Xu, H. (2007), "A critical assessment of the Burning Index in Los Angeles County, California," *International Journal of Wildland Fire*, 16, 473.

Short, K. C. (2014), "A spatial database of wildfires in the United States, 1992-2013," *Earth System Science Data*.

— (2015), "Spatial wildfire occurrence data for the United States, 1992-2013 [FPA_FOD_20150323] (3rd Edition)." .

Syphard, A. D. and Keeley, J. E. (2015), "Location, timing and extent of wildfire vary by cause of ignition," *International Journal of Wildland Fire*, 24, 37–47.

Tsagris, M., Anthineou, G., Sajib, A., Amson, E., and Waldstein, M. J. (2018), "Directional: Directional Statistics," .

UCLA CCSC (2015), "Energy Atlas Methods," Tech. rep., UCLA California Center for Sustainable Communities.

Voss, S. C., Main, B. Y., and Dadour, I. R. (2007), "Habitat preferences of the urban wall spider Oecobius navus (Araneae, Oecobiidae)," *Australian Journal of Entomology*, 46, 261–268.

Wood, S. (2011), "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." *Journal of the Royal Statistical Society. Series B*, 73, 3–36.

Xu, H., Nichols, K., and Schoenberg, F. P. (2011), "Kernel Regression of Directional Data with Application to Wind and Wildfire Data in Los Angeles County, California," *Forest Science*, 57, 343–352.

Xu, H. and Schoenberg, F. P. (2011), "Point process modeling of wildfire hazard in Los Angeles County, California," *The Annals of Applied Statistics*, 5, 684–704.