University of California

Los Angeles

A Study of Crimes Near Marijuana Dispensaries Through a Slew of Modern Statistical Approaches

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Michael Tzen

2012

Abstract of the Thesis

A Study of Crimes Near Marijuana Dispensaries Through a Slew of Modern Statistical Approaches

by

Michael Tzen

Master of Science in Statistics

University of California, Los Angeles, 2012

Professor Robert Gould, Co-chair

Professor Frederic P. Schoenberg, Co-chair

In the past decade, marijuana laws have come under scrutiny with Los Angeles serving as an epicenter and testing ground for regulations concerning the use and distribution of marijuana. The topic is polarizing, acting as a divide amongst the inhabitants of Los Angeles. In this paper, an analysis of crimes near legal marijuana dispensaries is pursued through a few modern statistical tools. This paper is motivated and built upon the 2011 research conducted by RAND authors Chang and Jacobson which sparked a commotion amongst supporters of anti and pro marijuana laws. An alternative view of Chang and Jacobson's result is pursued through the quantification of temporal sources of variation in the sampling distribution of a key regression parameter as well as classical spatial statistical tools. Further, we will make use of Voronoi tessellations in two ways. First, we will provide an alternative spatial regularization when interpreting crimes clustered around the marijuana dispensaries. Lastly, an application of the Voronoi Estimator, which estimates the intensity of an inhomogeneous point processes, will be implemented.

The thesis of Michael Tzen is approved.

Hongquan Xu

Mark S. Handcock

Frederic P. Schoenberg, Committee Co-chair

Robert Gould, Committee Co-chair

University of California, Los Angeles

2012

My sincerest thanks to:

. . . the professors of UCLA Statistics for the invaluable time and guidance,

. . . my friends and past mentors for shaping my views and skills,

. . . my family for the never-ending support along with the reminder to strive for the best.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In the Summer of 2011, RAND, the Santa Monica based think tank, published a report, later retracted, which riled up both parties that form the great divide amongst the medicinal marijuana community. The authors Chang, Jacobson, et.al, presented a technical report entitled "Regulating Medical Marijuana Dispensaries" [JCa11] in which the derived result cut against conventional wisdom. Their analysis, which focused on the effects of medical marijuana dispensary closures on crime counts, was a novel look at a hot topic amongst Californians. In this paper, we will study the effects and limitations under the assumptions made by the RAND authors.

First, a brief overview of RAND's analysis is presented. Using crime data obtained through crimereports.com, Chang and Jacobson looked at crimes occurring within a 21 day window centered around Dday = June 7^{th}, 2011.

$$[Dday - 10, Dday + 10]$$

Further, the attention was focused specifically on crimes occurring within the union of circular regions (Fig. 1.1) each centered at the 600 dispensaries scattered throughout LA city with varying radii of 0.3, 0.6, 1.5, and 3 miles.

$$\text{Domain} = \bigcup_{j=1}^{600} B_r(Dispensary_j) \text{ for j = 1,...,600 and r} \in \{0.3, 0.6, 1.5, 3\}$$

We will look specifically at the context when radius r = 0.3 miles, since this is the setting that is showcased in RAND's technical report (Note: a sensitivity

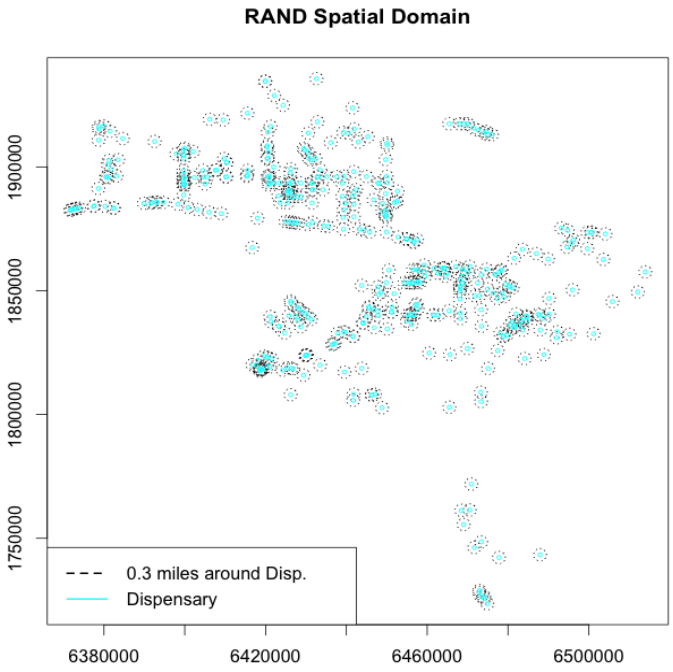

Figure 1.1: Spatial Domain: Union of Balls

analysis using the various radii was also conducted by RAND). With 21 days of crime and 600 dispensaries (170 open, 430 'closed'), RAND approached the analysis by fitting a linear model of the following form

$$Crime_{dt} = \alpha_d + \beta * 1_{(date>june7)} * 1_{closed} + \delta_t + \epsilon_{dt} \tag{1.1}$$

α_d is the fixed effect for dispensaries

β the interaction effect of the two indicator covariates

δ_t is the fixed effect for day

d is the dispensary index {1,2,. . . ,600}

t is the time index corresponding to the day {1,2,. . . ,21}

The conclusion that crimes increased when dispensaries were "shut down" was derived by interpreting the $\hat{\beta}$ coefficient (an estimated value of 0.013 with standard

error 0.006) which corresponds to the interaction $1_{(date>june7)} * 1_{closed}$ [JCa11].

Some skeptics, e.g. Dennis Romero, have pointed out that a handful of marijuana dispensaries listed as "closed" were in actuality open [Rom] . Other sources of criticism were directed at the 21 day time window Chang and Jacobson used, suggesting that the window length was not large enough to capture the effect of interest. The RAND paper was later withdrawn, with acknowledgements to the above critiques.

With the above context and possible problems in mind, this paper explores alternative analytical methods of the proposed model. We will look at a possible "null" distribution of the $\hat{\beta}$'s found in RAND's proposed model. The null distribution corresponds to the hypothesis that the $\hat{\beta}$ value RAND obtained for June 7^{th}, 2011 is no different then the $\hat{\beta}$ value for any other day of the year 2009. We proceed by using various 21 day time frames throughout the year of 2009, (in a sampled manner) and later extending the analysis to the full year. We still consider the same model RAND used. Essentially, the analysis will "relabel" the days of the year; with each relabeled day acting as a possible designated closure day.

Further, a look into the intrinsic spatial characteristics of the marijuana dispensaries and crimes are also pursued. Tools employed in the analysis of point patterns are leveraged in an exploratory manner. The purpose is to provide an alternative exploration of clustering (or inhibition) of the marijuana dispensaries touted by Chang and Jacobson. Also, an alternative regularization of the spatial domain is implemented with the use of Voronoi Tesselations.

CHAPTER 2

Data Used

The data used in this paper's analysis are gathered from two sources. All of the data processing and analysis are done in [R], the statistical programming language [R D12]. The crime data is supplied by the LAPD. Each row entry corresponds to a recorded crime incident. While the marijuana data is pulled from the API of the LA Times website.

```
> dim(lapd09.data.raw)
[1] 339688     57
```

	1	2	3	4	5
1	DR	PREMISE	WEAPONCODES	VICAGE	BEGWEEK
2	CRIMETYPES	PREMISECODE	WEAPONCODE	VICVEHMAKE	ENDYEAR
3	CRIMETYPE	SUSPECTS	MOCODES	VICVEHMODEL	ENDMONTH
4	CASESTATUS	POENTRY	NARRATIVE	VICVEHSTYLE	ENDDAY
5	CRIMECLASSCODES	POENTRYCODE	SUSPVEHMAKE	VICVEHYR	ENDWEEK
6	RD	POEXIT	USPVEHMODEL	VICVEHTOP	DISTRICT
7	BASICCAR	POEXITCODE	SUSPVEHSTYLE	VICVEHBOT	x
8	BEGTIME	PROPERTY	SUSPVEHYR	TYPEWATCH	y
9	ENDTIME	PROPCODE1	SUSPVEHTOP	TYPE_WATCH	dateconv
10	WATCH	PROPCODE2	SUSPVEHBOT	BEGYEAR	
11	DOW	PROPCODE3	VICSEX	BEGMONTH	
12	ZIP	PRIMARYWEAPON	VICDESCENT	BEGDAY	

```
> dim(disp.data)
[1] 431  10
```

	1	2	3
1	Collective	Notified	x
2	Address	lon	y
3	Applied.for.lottery	lat	
4	Ordered.to.close	index	

The raw LAPD dataset contains 7 observations that have missing entries and 183 observations where the geo-location pairs of (x, y) take on values $(0, 0)$. We make the strong assumption that these 190 observations are missing completely at random. As the missing data mechanism is not the focus of our study, we treat the 190 observations as ignorable and move on.

We will make use of crime occurrences that were recorded by the LAPD in 2009 with focus on the number of crimes, the date, and the location. From the LA Times, we will use the locations of the marijuana dispensaries and the binary variable depicting if the dispensary was ordered to close. The reader should be aware that there are obvious differences between the data used for our analysis and the data used by RAND. The following highlights the key differences between the datasets.

Differences in Datasets:

1. Mine: 431 Dispensaries
   Rand: 600 dispensaries

2. Mine: March 7^{th} 2011 Ordered to Close (Letters Sent Out)
   Rand: June 7^{th} 2011 Ordered to Close

3. Mine: LAPD for Crime Data (Recorded Crimes)
   Rand: crimereports.com for Crime Data (Reported Crimes)

4. Mine: 2009 Crime Data
   Rand: 2011 Crime Data

With the above differences, RAND's results cannot be compared to ours in a direct cross referential fashion. However, by paying careful attention in constructing a similar context, analysis of our results provide additional insight into RAND's conclusion.

For the time being, we work with crime data from 2009 recorded by the LAPD consisting of 339,688 recorded crime incidents throughout LA (city). Emperically, crime rates have seem to be historically declining from 2005 to 2009 (Fig. 2.1-2.3). It would be of interest to work with data from the past, specifically 2009, and see how it stacks up with a result obtained from 2011 data.

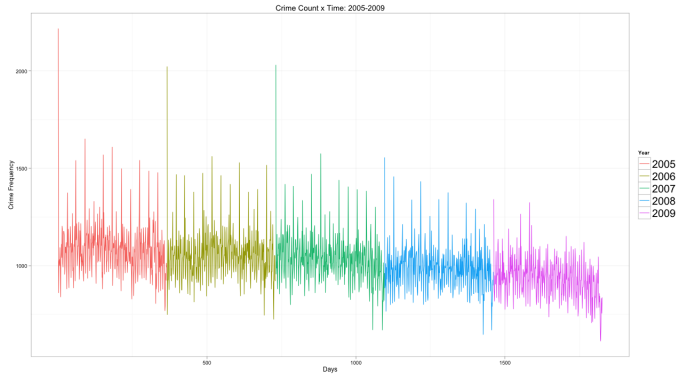

Figure 2.1: Crime Counts (all) across time from 2005-2009

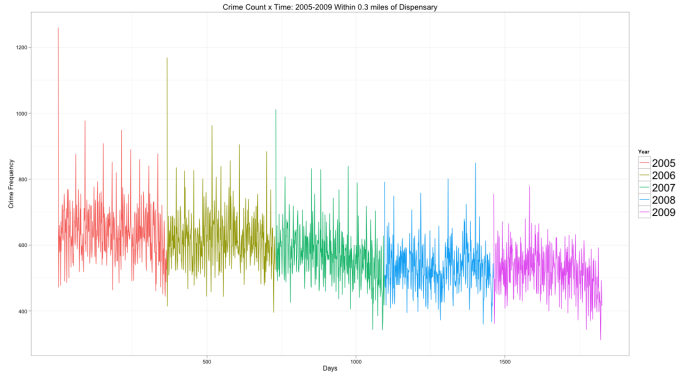

Figure 2.2: Crime Counts (0.3 mi around disp.) across time from 2005-2009

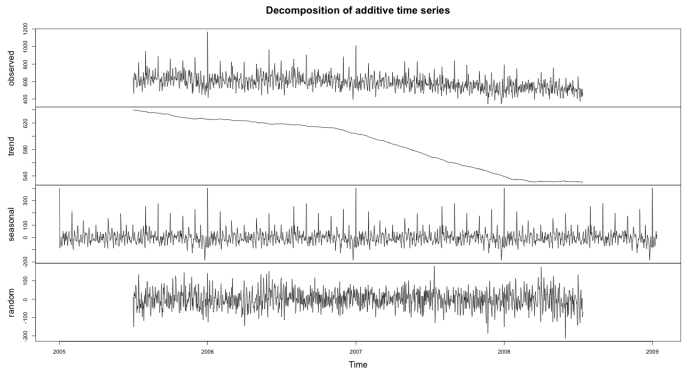

Figure 2.3: Decomposed time series of crimes 0.3 mi around disp. from 2005-2009

CHAPTER 3

Quantifying Sources of Temporal Variation in $\hat{\beta}$

3.1 Intro

Chang and Jacobson's novel use of an OLS regression model, although nice, has its limitations. One major obstacle is the narrow 21 day time frame employed in the regression model. By using such a model, the question of window size comes to question. Of concern would be the reported standard error of the estimated $\hat{\beta}$ when used to interpret $\hat{\beta}$'s significance. By taking Chang and Jacobson's result, as is, no further context is provided when trying to discern if their estimated $\hat{\beta}$ is atypical. Interpretation relies on the assumed asymptotic qualities commonplace to a classical simple regression analysis. In this part of the paper, we provide a possible null distribution of β's with the null hypothesis of "no label difference" when looking at relabeled closure dates. The method of analysis draws conceptual similarities to Fisher's permutation test and Efron's bootstrap. Essentially, we embed the 21-day RAND regression model into a larger dataset containing 365 days, and look at possible contiguous 21-day regression models. The sampling distribution of $\hat{\beta}$ would thus serve as the null distribution previously described.

3.2 Appx. Null Distribution using 100 sampled days

We first proceed by drawing randomized k=100 days sampled in 2009 without replacement. We work under the assumption that each of the 100 days acts as the designated closure date denoted as "$rDday$." This assumption implies that any of

the sampled $rDday$'s can act analogously to June 7^{th} 2011, the day RAND drew their gaze upon.

One detail worthwhile of mention is how the values of the two indicator covariates in "$\beta * 1_{(date>rDday)} * 1_{closed}$" are coded (equation 3.1). The term 1_{closed} should really be interpreted as $1_{\textit{ordered to close on March } 7^{th}}$. (We are thus making the underlying assumption that the March 7^{th} closure from LATimes, corresponds to RAND's closure date of June 7^{th} 2011). For ease of notation, we consider the two indicator variables discussed above as functions of the indices $d, t,$ and k, where d is the dispensary index, t is the time index, and k is our rDday index.

$$1_{date} = f(t,k) = \begin{cases} 1 & t > k \\ 0 & \text{otherwise} \end{cases}$$

$$1_{closed} = g(d,k) = \begin{cases} 1 & Disp_d \text{ ordered to close on day } k \\ 0 & \text{otherwise} \end{cases}$$

This is to illustrate that we will fix the values of $1_{closed} = g(d,k)$ for $k = 1, \ldots, 100$. This reduces the indicator of 1_{closed} to a function depending only on the dispensaries. We will fix the values of 1_{closed} using the observed data from "gmarijuana.txt"[Tim]. The justification of doing this is warranted. If we let 1_{closed} represent if the dispensary 'truly' was ordered to close on day_k (i.e. a varying function of the sampled rDdays), almost all the interaction terms will drop out, with the only remaining interaction term corresponding to the regression where the sampled rDday = March 7^{th}.

Fixing 1_{closed} (i.e. a fixed function of the sampled rDdays where the values are fixed across all rDdays) keeps the interaction terms (corresponding to RAND's regression). Hence, the choice of fixing the 1_{closed} indicator ensures correspondence to the dispensaries RAND focused on when looking at June 7^{th}, 2011. Hence, throughout all possible 21-day windows of 2009, we are looking at similar models as

the one proposed by Chang and Jacobson in 2011. With the discussed precautions, extra care is needed when wrangling the data. If the reader is interested, the documentation on the data wrangling procedure can be viewed in the *Code* section.

Proceeding with the construction of our approximate null distribution, we draw, without replacement, 100 rDdays $\in \{10, 11, \ldots, , 354, 355\}$ and focus on $[rDday - 10, rDday + 10]$. For each $rDday_k$ where $k = 1, \ldots, 100$, we run an OLS regression of the form

$$Crime_{dt} = \alpha_d + \beta * 1_{(date > rDday)} * 1_{closed} + \delta_t + \epsilon_{dt} \tag{3.1}$$

Note that the linear equation in (3.1) is analogous to the RAND equation shown in (1.1). We essentially ran $k = 100$ OLS regressions corresponding to the sampled central cut-off day rDday. We then analyze the distribution of $\{\hat{\beta}_k : k = 1, \ldots, 100\}$.

We now "retroactively" compare our results (2009 data) to RAND's (2011 data) by looking at the the following summary statistics in Table 3.1 and the distribution in Figure 3.1. One thing to note, RAND's numbers, in the first row, will be point estimates from a single OLS model, while the values in the second row (using 100 subsets) will be averages of the 100 point estimates output from the OLS models of Equation 3.1.

Table 3.1: Radius = 0.3 Miles and Total Crime Counts as a response

	$\hat{\beta}$	SE($\hat{\beta}$)
RAND (2011)	0.013	0.006
	mean($\hat{\beta}$)	sd($\hat{\beta}$)
100 Subsets (2009)	0.010	0.062

We see that RAND's 2011 $\hat{\beta}$ coefficient estimate falls near the mean of our Null distribution derived from 2009 data. The direct interpretation of our findings tell

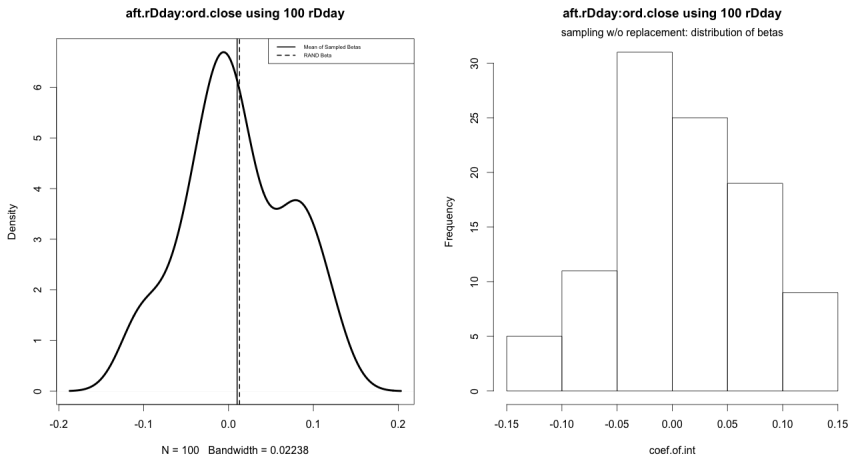

Figure 3.1: Distribution of $\hat{\beta}^*$ through 100 samples w/o replacement

us that RAND's results would not be atypical if they were to conduct the same study in 2009. Specifically, under the null hypothesis of 'no label difference' that is, the $\hat{\beta}$ estimate using Dday = June 7^{th} 2011 is not more unusual than the $\hat{\beta}$ estimate when using any random day in 2009 should not be rejected.

The reader might see how the use and interpretation of our preceding analysis has similarities to an approximate Fisher's permutation test, where the distinction is we are looking at possible combinations in the "ambient space" the 21-day model is embedded in as opposed to possible combinations within the 21-day window itself. Our null distribution is deemed approximate since we only used k=100 sampled 21-day windows. Thus, our obtained null distribution is a Monte Carlo approximation of the full null distribution using all the days of 2009. In the next section, we explore the full null distribution of $\hat{\beta}$.

3.3 Full Null Distribution Using 345 Days

The developments in the preceding section laid the foundational groundwork for our method of analysis. The approximate null distribution obtained (using 100

sampled center days) served as a proof of concept which also provided some insight on the behavior of RAND's $\hat{\beta}$. With a slight modification to the schematic and code, we can look at the full null distribution of $\hat{\beta}$.

Utilizing all of the days, as opposed to just 100 random days, in 2009, we run the same scheme of looking at OLS regression models using 21-day windows (Equation 3.1). Incorporating all 365 days in 2009 serves two fold. We get to check for the robustness of our previous approximate null distribution by comparing it to the full null distribution. Specifically, we would like to see if the properties of the null $\hat{\beta}$ distribution we saw previously contained artifacts due to the sampling scheme. Further, the assessment of the temporal variation in the coefficient of interest $\hat{\beta}$ is preferred when using the full null distribution.

Instead of sampling the designated closure dates (or equivalently the center day in the 21 day interval), we now look at all designated closure dates by "lagging" the designated closure date across the whole year of 2009. The null distribution of $\hat{\beta}$ is thus the distribution of the set of parameter estimates when using all possible 21 day subsets in 2009.

Table 3.2: Radius = 0.3 Miles and Total Crime Counts as a response

	$\hat{\beta}$	SE($\hat{\beta}$)
RAND (2011)	0.013	0.006
	mean($\hat{\beta}$)	sd($\hat{\beta}$)
345 Subsets (2009)	-0.0008	0.057

We see that the distribution has a mean that is still extremely close to 0 as well as being uni-modal and more symmetric. A difference between our previous "sampled" distribution is that the "full" distribution has the mean on the negative side. Most importantly, we still see that the difference between RAND's beta estimate and the mean of our null distribution is small. Most of the properties

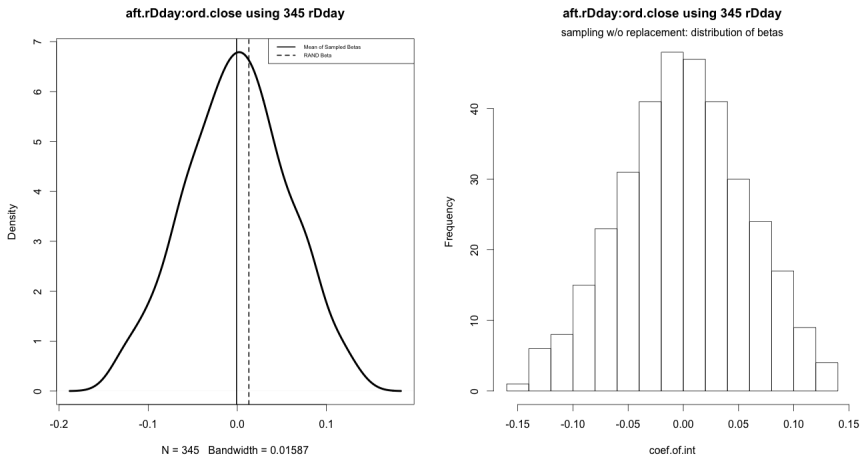

Figure 3.2: Distribution of $\hat{\beta}^*$ using 345 samples w/o replacement

of the approximate null distribution is represented in the full null distribution. A conclusion based on the full null distribution tells us that once again the $\hat{\beta}$ Chang and Jacobson obtained is not atypical of a $\hat{\beta}$ obtained by using any day of 2009. Specifically, Chang and Jacobson's $\hat{\beta}$ estimate of 0.013 is enveloped by temporal variations in $\hat{\beta}$.

Another view of the temporal variations in $\hat{\beta}$ is seen by looking at how $\hat{\beta}$ changes over time. This analysis draws similarities to using time plots of the outputted samples from an MCMC algorithm. The time plots are used to diagnose if adequate "mixing" of samples in an MCMC algorithm is prevalent. In our case, we are interested in any striking patterns prevalent in the plot of $(\hat{\beta}_k, k)$. A plot of the $\hat{\beta}$ over ordered rDdays is presented in Figure 3.3. We see that there is fluctuations of $\hat{\beta}$ reminiscent of random noise. Upon inspection, there does not seem to be any grotesque structure. If the values of $\hat{\beta}$ were consistently low for days before a specific day and consistently high for days after, the null distribution in the form of the density provided previously would not be able to pick up this striking time dependent pattern.

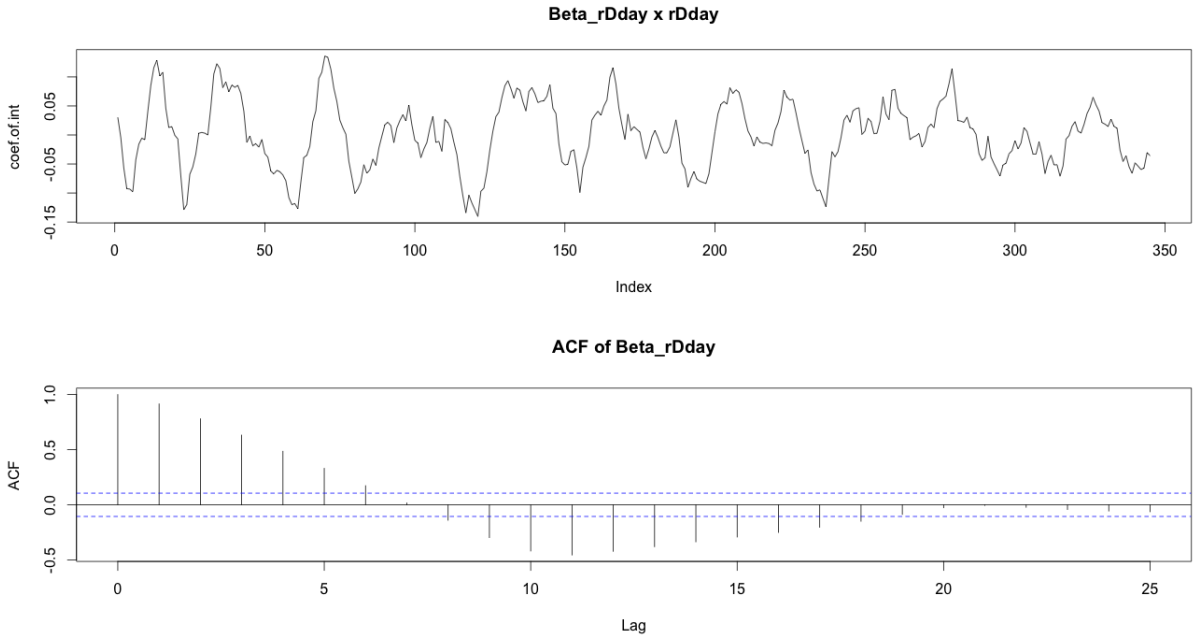

Figure 3.3: $\hat{\beta}^*$ against ordered rDdays

3.4 Simulated Null Distribution using 10,000 samples w/ replacement

Our previous null distributions drew similarities with Fisher's permutation test, in an "approximate" followed by an "exact" manner. By allowing the rDdays to be drawn with replacement, we would have a null distribution similar to Efron's bootstrap. Once again, the difference is we are drawing samples from the "ambient" temporal space in which the 21-day regression model is embedded in.

Using data from 2009, we have conducted various sampling schemes on the parameter of interest found in RAND's regression model. The same methodology (and code) can be applied to crime data from 2011. Using crime data from 2011 would give us a more direct comparison to RAND's regression result. At the time of this study, 2011 crime data was not available. However, combining our result (using 2009 data) and the empirical evidence of declining crime, it would not be

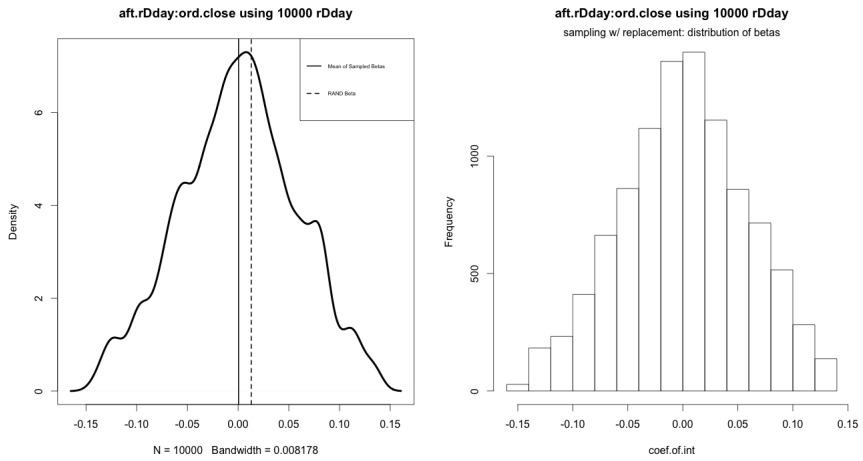

Figure 3.4: Distribution of $\hat{\beta}^*$ using 10,000 samples w/ replacement

surprising to see a similar result using 2011 crime data. The value of $\hat{\beta} = 0.013$ is quite common for any contiguous 21-day subset of 2009.

3.5 Translating the Analysis to a 31-Day Window

In the preceding sections' analysis, results suggested that the allocation of the observed $\hat{\beta}$'s variability was associated with the randomly chosen designated closure date. In this section, we translate the preceding analysis to a 31-Day window from the, previously studied, 21-Day window.

The regression equation found in Equation 3.1 is still the one of study. However, the modification is looking at crimes in the window $[rDday-15, rDday+15]$ instead of $[rDday - 10, rDday + 10]$. The motivation in pursuing this analysis comes from addressing the "small window" size critiques of Chang and Jacobson's 2011 paper. The posed problem suggested the 21-Day window was potentially too small to correctly capture the correlation of crime counts and the dispensary closures. For this paper's study, the sensitivity of $\hat{\beta}$'s null distribution to the window size is presented.

We want to compare the null distribution of the sampled $\hat{\beta}$'s when using a 31-day window to a 21-day window. Figure 3.5 is the null distribution of (the same) 100 sampled days in 2009. When compared with Figure 3.1, we immediately see striking similarities. Moving on, we compare and contrast the "full" null distribution using all days of the year by looking at Figure 3.6 and Figure 3.2. Both "full" null distributions are, uni modal, symmetric, and supported on similar ranges of $\hat{\beta}$. The similarities between the $\hat{\beta}$'s when using the 31 and 21 day windows continue to be seen in the time plots seen in Figure 3.7 and Figure 3.3. As a last comparison, we turn to Figure 3.8 and Figure 3.4, which are the two distributions when 10,000 closure dates are sampled with replacement. The 31-day distribution once again mimics the 21-day distribution in range, symmetricity, and even in the pattern of local modes.

Our exploration of our null distribution's sensitivity to an increase in window size has presented two conclusions. The 31-day and 21-day windowed null distributions are very similar in their moments. Since the qualities of the null distributions are mirrored, the analysis in the preceding sections (using the 21-day window) can be carried over when we talk about a 31-day window. Whether we use a 21-day window or a 31-day window, the variability in the observed $\hat{\beta}$ is associated with the randomly chosen designated closure day. This suggests that the estimate Chang and Jacobson obtained is not atypical in both window size contexts.

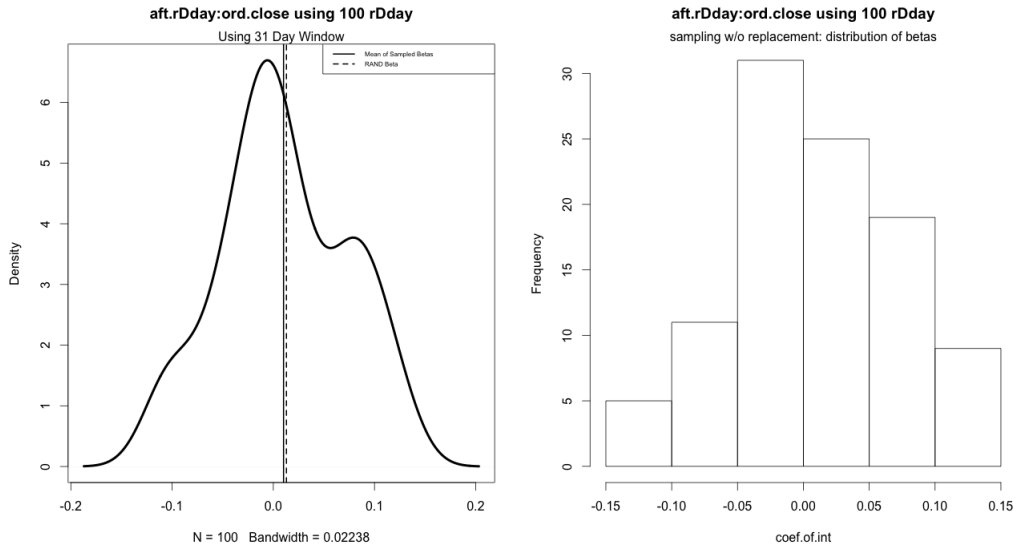

Figure 3.5: 31 Day: Distribution of $\hat{\beta}^*$ using 100 samples w/o replacement

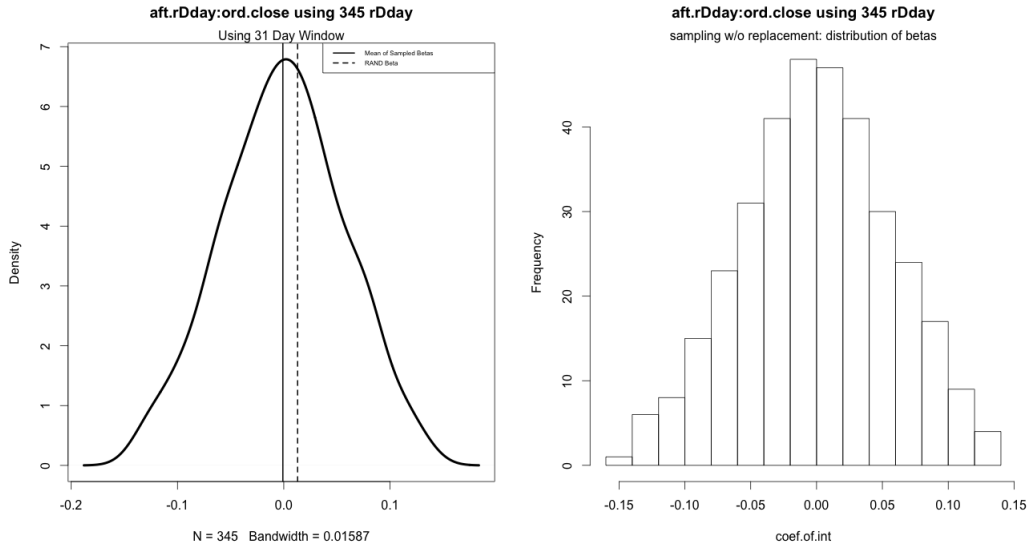

Figure 3.6: 31 Day: Distribution of $\hat{\beta}^*$ using 345 samples w/o replacement

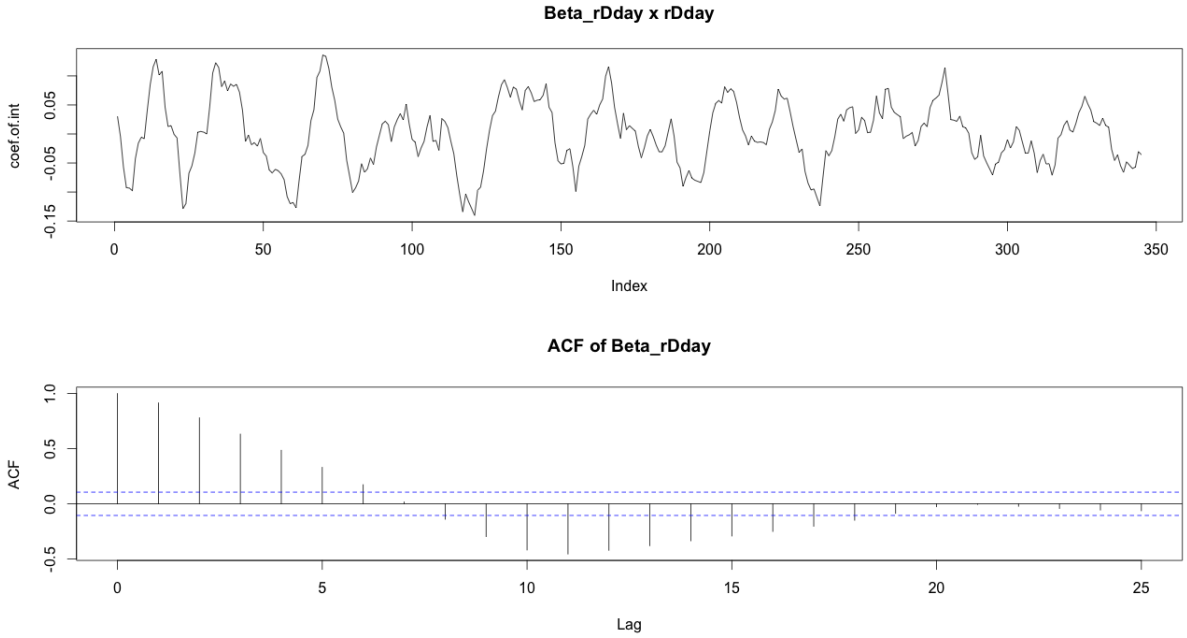

Figure 3.7: 31 Day: $\hat{\beta}^*$ against ordered rDdays

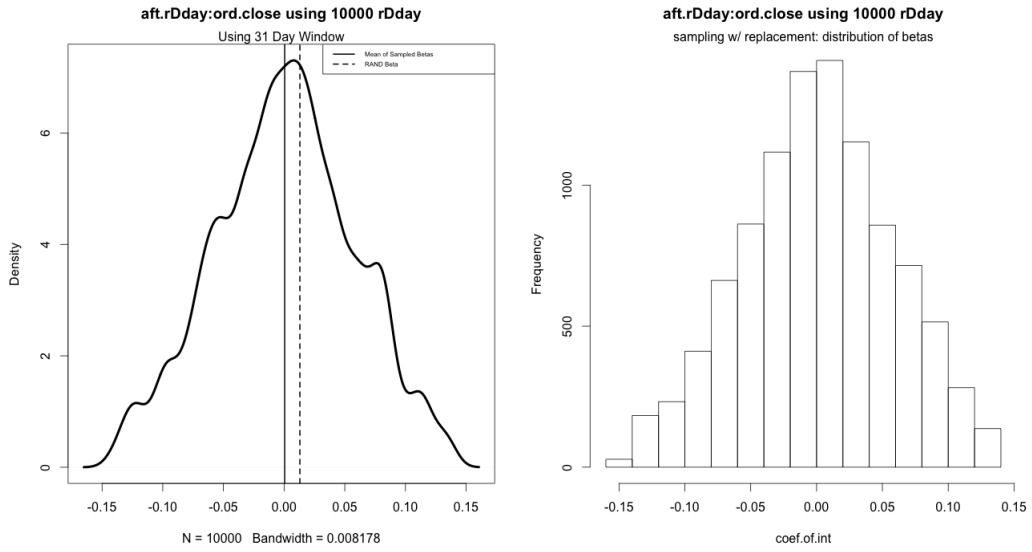

Figure 3.8: 31 Day: Distribution of $\hat{\beta}^*$ using 10,000 samples w/ replacement

CHAPTER 4

Spatial Clustering and the Voronoi Tessellation

4.1 Intro

In the preceding discussion, we have looked at Chang and Jacobson's regression model and came to the conclusion that their interpretation of the proposed model was not an "atypical" result (if it were 2009). We now shift our analysis on the assumptions Chang and Jacobson proposed in regards to defining the spatial domain of interest. Noting the irregular shape of Los Angeles' (city) boundary (Fig. 4.1-4.2), as well as the slight nuances of zoning restrictions, the question that comes to mind is whether the use of the spatial domain depicted in Fig. 1.1 is a reasonable one. We present two applications of Voronoi Tesselations when dealing with the point pattern of dispensary locations. Our first Voronoi application proposes an alternative spatial regularization when analyzing the regression model of Chapter 3. The unique properties of the Voronoi cells are also leveraged when used as an estimator, the Voronoi Estimator.

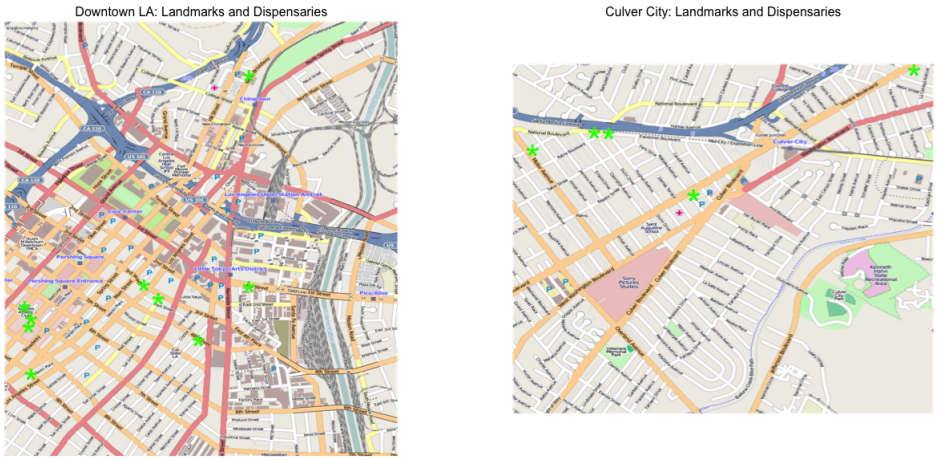

Figure 4.1: Dispensaries in DTLA and Culver City with Neighboring Landmark References

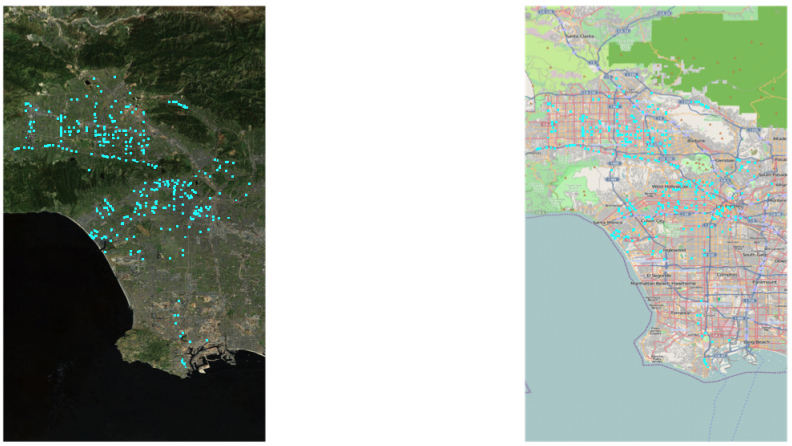

Figure 4.2: Dispensaries in LA with Natural and Street References

4.2 Clustering of Dispensaries as Point Patterns

Recall that the previous analysis' (in both this paper d=431 and RAND's d=600) focused exclusively on crimes that fall in the union of circular regions given by

$$\text{Domain} = \bigcup_{j=1}^{600} Ball_r(Dispensary_j) \text{ for j = 1,,d and } r = 0.3.$$

Although this approach of binning crimes into such a domain seems intuitive, we will explore if it's a reasonable one. The RAND author's specification of this domain was motivated by previous research conducted by the same authors and also referenced in [JCa11]. The paper was a preceding and separate analysis conducted by Chang and Jacobson where they investigated the spatial characteristics of crime and marijuana dispensary locations.

Specifically, the problem of interest was whether dispensaries (open and closed) showed signs of clustering. The clustering of dispensaries is an important characteristic to diagnose, since the estimates of the effects in the OLS linear regression model were at the dispensary-level. Through regression analysis, the authors concluded that "...clustering is independent of closure status, meaning that the likelihood that a closed dispensary is near another closed dispensary is the same as the likelihood that an open dispensary is near a closed dispensary."

$$Pr(Disp_j = Closed | Disp_i = Closed) \approx Pr(Disp_k = Open | Disp_l = Closed) \tag{4.1}$$

where {i,j} are the indices of dispensaries that are nearest neighbors

We will investigate this result through a point process framework with the aim of assessing the plausibility of the defined spatial domain aforementioned.

By only modeling the crimes within the union of epsilon balls, the analysis masks and restricts data which might be relevant. This potential problem can be

seen in the point plots (Fig 4.3) of crime and dispensaries.

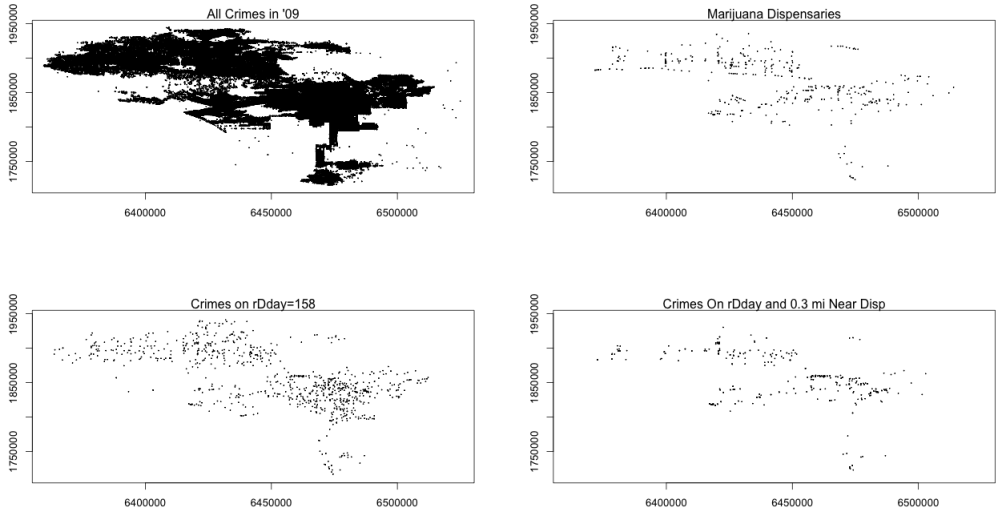

Figure 4.3: Various Point Plots of Crime

We proceed by utilizing the J, K, and L functions that are commonly used to investigate the clustering (or inhibition) of point patterns. We first look at the dispensary locations provided by the LA Times API [Tim].

Working under the point process framework treats the dispensary locations as a stochastic process. Theoretically, the marijuana dispensaries can set up shop anywhere in LA. The real dispensary locations (Figure 4.3) would be an instance or realization of this underlying stochastic point process. We first address if the dispensaries are clustered or inhibited (as a function of distance s) through the use of Ripley's K-Function (and related L-Function) [DV03].

$$\widehat{K}(s) = \lambda^{-1} n^{-1} \sum_{i \neq j} I(d_{ij} < s) \tag{4.2}$$

$$\widehat{L}(s) = \left[\frac{\widehat{K}(s)}{\pi} \right]^{\frac{1}{2}} \tag{4.3}$$

By default, the K and L functions are utilized by defining some a priori boundary window with an applied "boundary correction." The values of the K-Function

is sensitive to the boundary window of choice [ESR], and thus the interpretation of clustering is reliant on a suitable apriori window choice. We choose a non-convex polygon (the boundary of LA City) as our reasonable a priori window choice (Fig. 4.4).

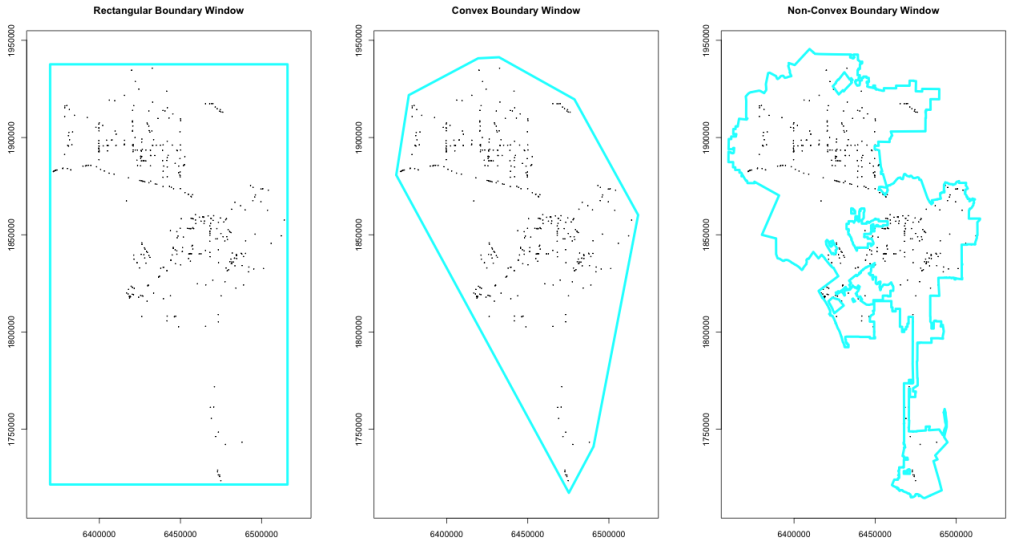

Figure 4.4: Boundaries: Rectangular / Convex / Non-Convex

As an aside, we have also looked at both rectangular and convex boundaries and its effect on the K-Function (Fig. 4.5). The use of the non-convex polygon as our a priori boundary is suitable and the most reasonable.

Interpreting the L-Function$_{non-convex}$, we can say that the locations of dispensaries (up to 100,000 feet) are a smidgeon more clustered than what a pure (or homogeneous) poisson process would be. That is, the intensity (or rate) parameter in a theoretical model for dispensary counts is not the same as the rate parameter of a homogeneous poisson process. This tells us that marijuana dispensaries (without the consideration of closure status) are clustered.

Moving forward, we now look at the clustering of different types of dispensaries by incorporating an indicator variable representing if the dispensary was ordered

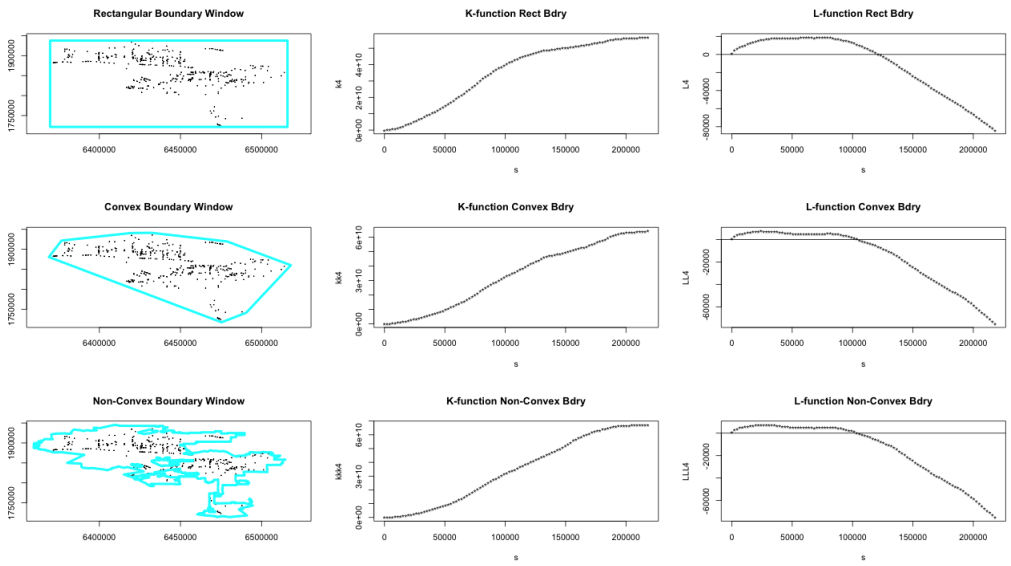

Figure 4.5: K-L Functions on 3 boundaries

to close. This point process is now "marked" (Eq. 4.4) which allows us to distinguish between the dispensaries that were ordered to close relative to dispensaries that were not.

$$z(x,y) = \begin{cases} 1 & Disp_d \text{ ordered to close} \\ 2 & \text{otherwise.} \end{cases} \tag{4.4}$$

The generalization of the previous single-type K and L functions (Eq 4.2 and 4.3) to their multi-type counterparts are used to look at the marked point process. Like the earlier version of the K and L functions (single-type), the values of the multi-type K and L functions tell us the amount of clustering as a function of distance. However, with two "subprocesses" (open dispensaries and closed dispensaries) we are interested in comparing the interdependencies between one subprocess relative to the other across all distances.

The observed multi-type K and L functions (Fig. 4.6) suggests that relative clustering is prevalent between the two types of dispensaries.

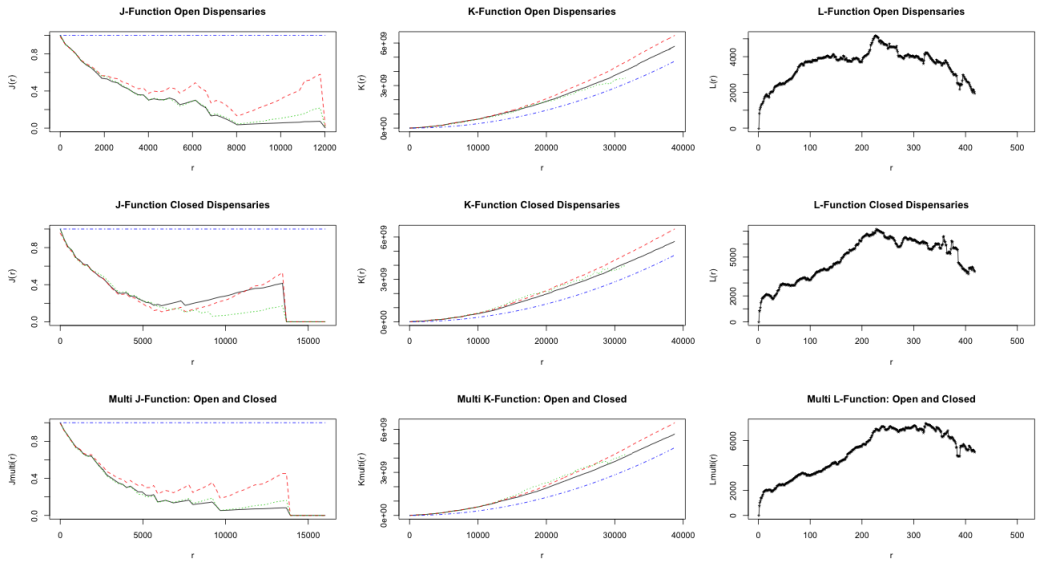

Figure 4.6: Multi-J/K/L functions of Open and Closed Disp.

Specifically, the two types of dispensaries exhibit similar K-L function values. Dispensaries that were ordered to close are found near dispensaries that were not ordered to close. Through alternative tools, we arrive at similar conclusions in regards to Chang and Jacobson's (2010 unpublished) analysis of dispensary to dispensary clustering. With the assessment of dispensary-status independence, The preceding spatial exploration confers the ability to talk about dispensary-level effects in the regression models of equation 1.1 and 3.1. However, we can also look at an alternative spatial regularization instead of the one used in Fig 1.1.

4.3 Spatial Regularization of a Domain: Voronoi Tessellation

We introduce a spatial domain that serves as an alternative to the one used by RAND authors Chang and Jacobson. We employ the Voronoi tessellation (also

called Dirichlet tessellation) on our marijuana dispensaries. There have been many successful applications of the Voronoi tessellation in various fields [OBS00]. The classical example can be found in the use of the tessellations conducted by the Cholera inquiry Committee (1855). The intent was to study potentially problematic water pumps with concerns to cholera outbreak.

The motivation for entertaining the use of the Voronoi tessellation are both intuitive and statistical. Medicinal marijuana dispensaries are brick and mortar operations whose spatial location should acknowledge competing dispensaries in regards to territorial concerns over customers (and therefore the potential for crime). Further, the Voronoi tessellation provides a disjoint spatial regularization, a quality not present in RAND's union of epsilon balls. Another attractive feature of the tessellation are the non-parametric qualities present in its derivation.

For a point set $P = \{p_1, \ldots, p_n\}$ in R^2, the planar ordinary Voronoi diagram generated by P is defined as

$$V(p_i) = \{x : ||x - x_i|| \leq ||x - x_j|| \text{ for } j \neq i, j \in I_n\} \tag{4.5}$$

We let each cell of the Voronoi tessellation, whose generating point is the unique dispensary, represent the "territory" attributed to its corresponding dispensary.

Figure 4.7 shows the $431 - 7 = 424$ cells of the Voronoi Tesselation corresponding to the set of 424 unique marijuana dispensaries (Note: There are $3 + 3 + 2 + 2 + 2 = 12$ marijuana dispensaries that are found in 5 of the Voronoi cells. Therefore, $12 - 5 = 7$ dispensaries did not generate their own unique Voronoi cell. The red tiles in Figure 4.7 depict the 5 cells that contain an additional marijuana dispensary). We proceed in analyzing the 424 dispensaries that correspond to the Voronoi Tesselation composed of 424 unique cells.

A visual comparison of the union of 0.3 mile radius balls and the Voronoi

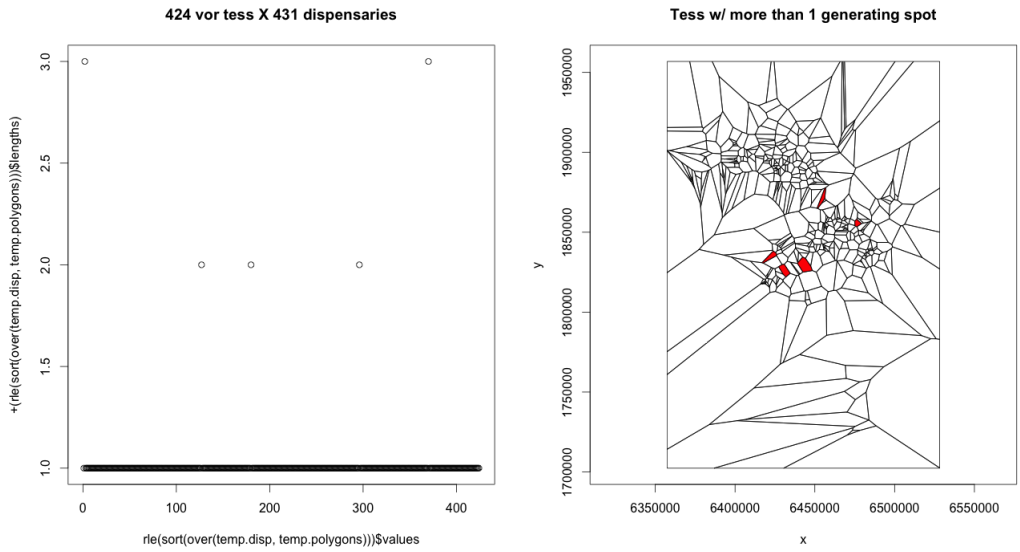

Figure 4.7: Removed 5 Dispensaries to Generate 424 Vor Tess

tesselation is presented in Figure 4.8. The trade off for gaining non-overlapping bins, by using the Voronoi tessellation, is seen in the exchange for larger area of the cells generated by fringe dispensaries. One can imagine that the Voronoi cells with large area, generated by the fringing dispensaries, would capture more crime occurrences than its radius ball counterpart.

For the whole year of 2009, we compare the two spatial domains in regards to the number of crime occurrences falling within a dispensary's region. Looking at Figure 4.9, we see how assigning larger area to a dispensary's territory (in our trade-off) attributes more crime counts to fringing dispensaries.

The difference in scale is highlighted when looking at the overlaid plots in Figure 4.10 as well as the comparison of the distributions in Figure 4.11. We see that by using the Voronoi tesselation, a handful of dispensaries allocate more crime counts than the ball bin does. The re-allocation of crime counts is seen in how the Voronoi bin's density plot contains more mass in the right tail than the density plot of the ball bin.

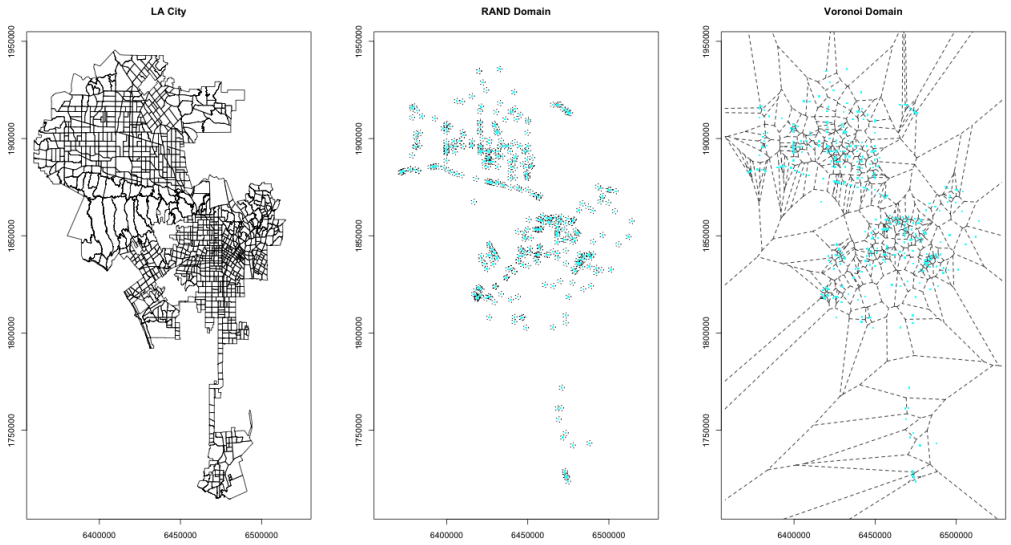

Figure 4.8: LA City, Union Ball Domain, Voronoi Tess. Domain

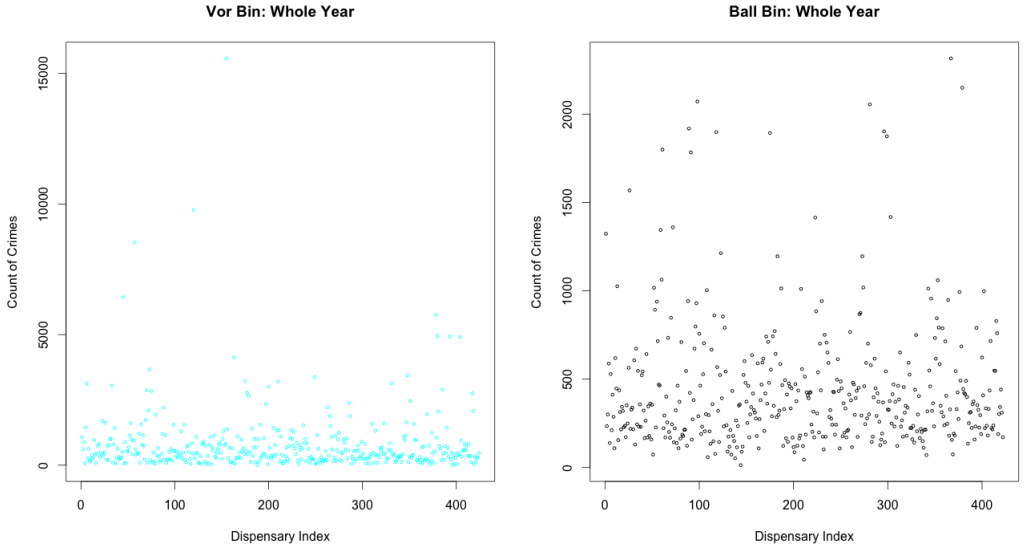

Figure 4.9: Voronoi and Ball Bin: Crime Counts

For a last visual exploration of the Voronoi Tesselation's effect on the allocation of crime counts, we turn to Figure 4.12. The 424 Voronoi tiles are plotted along with each tile's crime count (all of 2009) in the form of a gray-scale gradient fill.

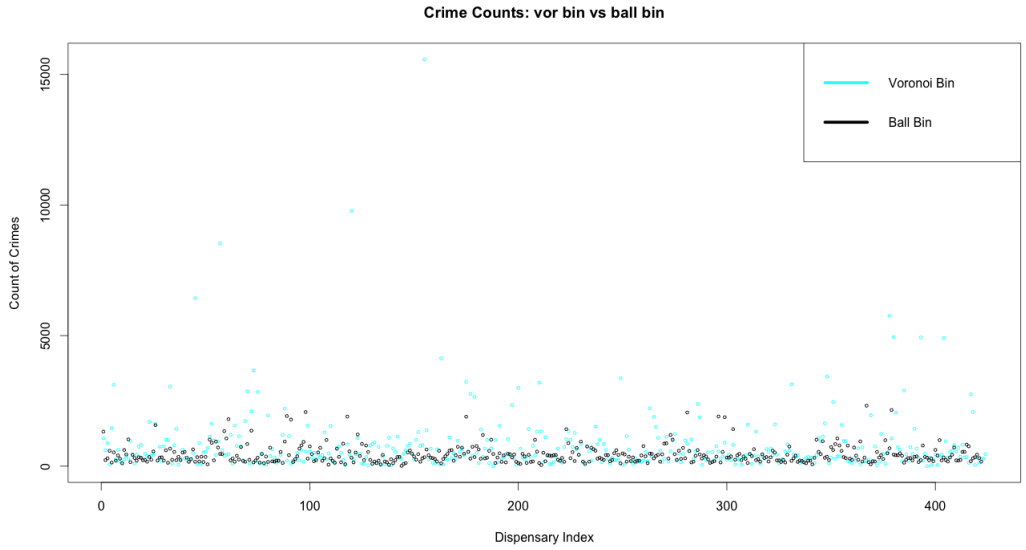

Figure 4.10: Voronoi and Ball Bin: Crime Counts Overlaid

We see that for the annual raw crime counts, a relatively few number of tiles take on high values. A log scaled version is presented alongside to get a sense of how $log(crime)$ differs from tile to tile. In the raw crime count scale, the standout black tile corresponds to areas near Compton.

The preceding exploration has suggested that on an annual level, a few Voronoi cells attribute more crime counts to its corresponding dispensary than their ball bin counterpart. However, when we focus on a smaller time frame, say 21-day windows, the argument can be made that gaining the non-parametric and non-overlapping qualities of the Voronoi regularization outweighs the annual-level "inflation" seen in a few fringe dispensaries. The next section pursues the implementation of the Voronoi tesselation's spatial regularization in such a context.

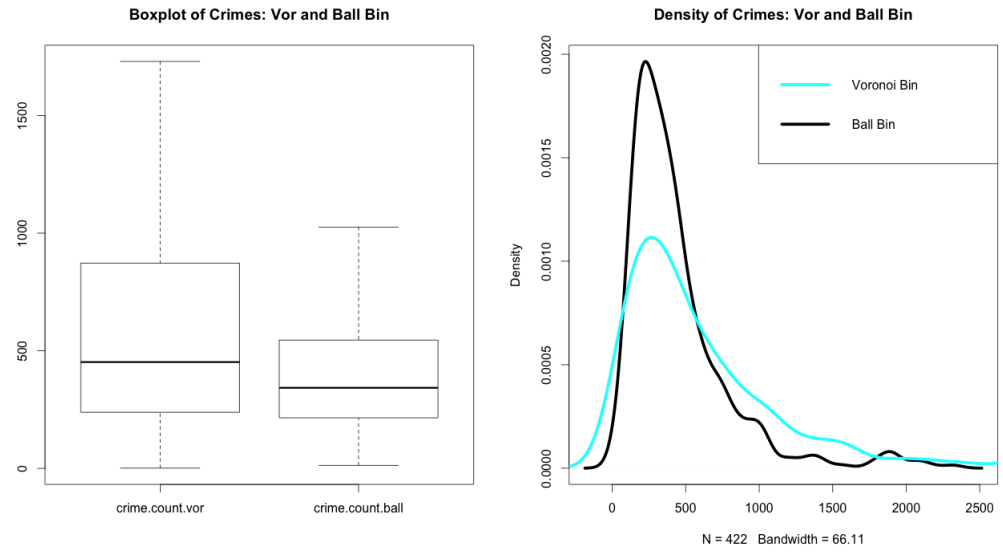

Figure 4.11: Voronoi and Ball Bin: Distribution of Crime Counts

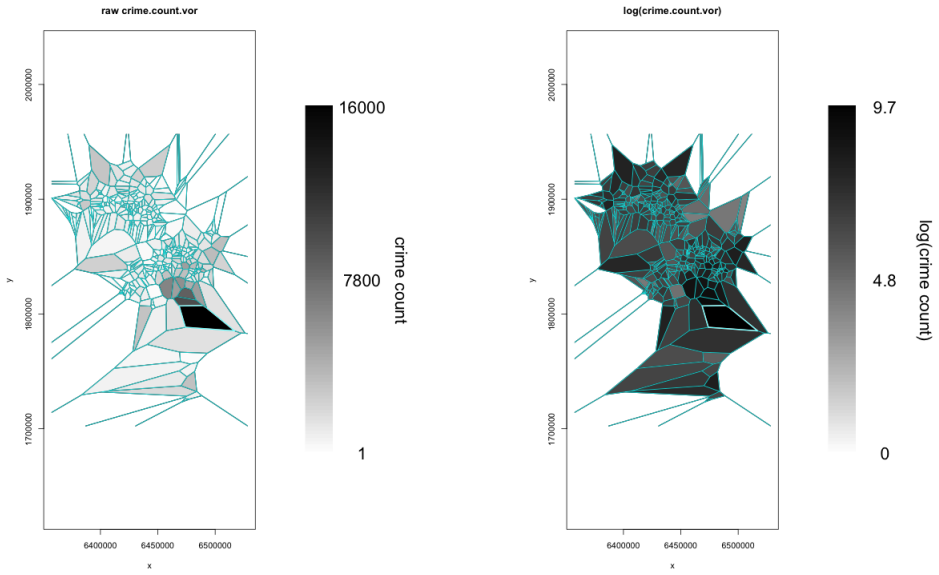

Figure 4.12: Voronoi Bin: Crime Counts and log(Crime Counts)

4.4 Null Distributions of $\hat{\beta}$ using Voronoi Bins

With our Voronoi tesselation generated by our set of dispensary locations, we take another look at the now familiar regression equations discussed in Chapter 3. In

this section, we repeat the analysis of Sections 3.1-3.2, where we analyze the null distribution of $\hat{\beta}$'s under various sampling schemes.

We begin our study by using the Voronoi Tesselation of 424 cells (Figure 4.7) obtained from the previous section. Specifically, instead of using the ball spatial domain (Figure 1.1), we will only look at crimes that fall into the 424 tiles $\{C(disp_i) : i = 1, \ldots, 424\}$ generated by

$$C(disp_i) = \{x : ||x - x_i|| \leq ||x - x_j|| \text{ for } j \neq i, j \in I_n\} \tag{4.6}$$

Note that the use of crimes in the Voronoi tiles, above, is the key difference from the analysis done in Chapter 3. We then use our Voronoi binned crime counts to estimate the regression model found in Equation 3.1.

We first look at 100 sampled $rDday$'s (designated closure dates without replacement), while using a 21-day window. The resulting null distribution of $\hat{\beta}$ is given in Figure 4.13. Comparing the same sampled 100 days, the null distribution when using a ball bin (Figure 3.1) differs from our null distribution when using the Voronoi bin. The difference in inflection points of the two density plots is the major difference. We can attribute this change in densities to the use of Voronoi binning. Of concern is how close the mean of our null distribution is from 0.013, the point estimate of RAND. When looking at our two null distributions, we see that the mean of our sampled $\hat{\beta}$'s are both close to 0.013. One difference is the Voronoi-binned mean is slightly larger than 0.013, while the ball-binned mean is slightly smaller.

We now proceed to look at all contiguous 21-day windows of 2009. Instead of using 100 sampled $rDday$'s, we look at 345. The resulting "full" null distribution of $\hat{\beta}$ is given in Figure 4.14. Compared to our approximate distribution (using 100 samples), our full distribution has similar spread yet contains more observations near its center (higher kurtosis). Similarly, when compared to the ball-binned

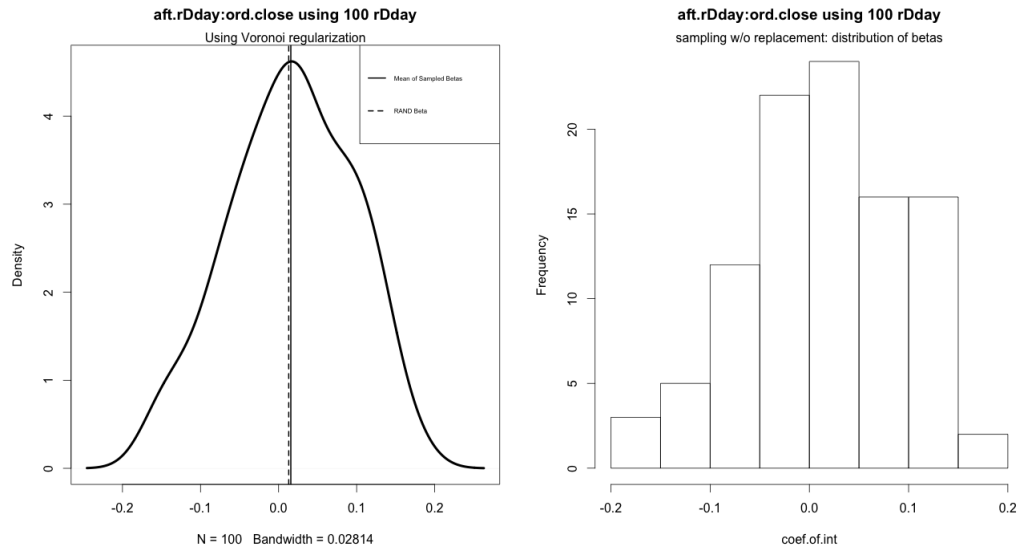

Figure 4.13: Voronoi Bin: Distribution of $\hat{\beta}^*$ using 100 samples w/ replacement

"full" null distribution (Figure 3.2) our "full" voronoi-binned distribution shows higher kurtosis as well as less symmetry. We still see how RAND's estimate of 0.13 is higher than the mean of our sampled $\hat{\beta}$, yet the difference is pretty small. Although slight, the mean being less than 0.13 was not seen in our voronoi-binned distribution when using 100 samples. This suggests that when using 100 samples, the mean of $\hat{\beta}$ having a value less than 0.13 is an artifact of of the sampling scheme. However, this sampling "artifact" was not seen in our ball-binned null distribution.

We conclude our analysis of the voronoi-binned full null distribution by looking at the time plot of $(\hat{\beta}_k, rDday_k)$ in Figure 4.15. The time plot does not seem to contain grotesque time dependent structure. We do however see a difference of the time plots when comparing Figure 4.15 (voronoi bin) to Figure 3.3 (ball bin). Although the actual time plots take on different values, the acf plots are very similar.

We have looked at how using a Voronoi tesselation, to regularize crime count

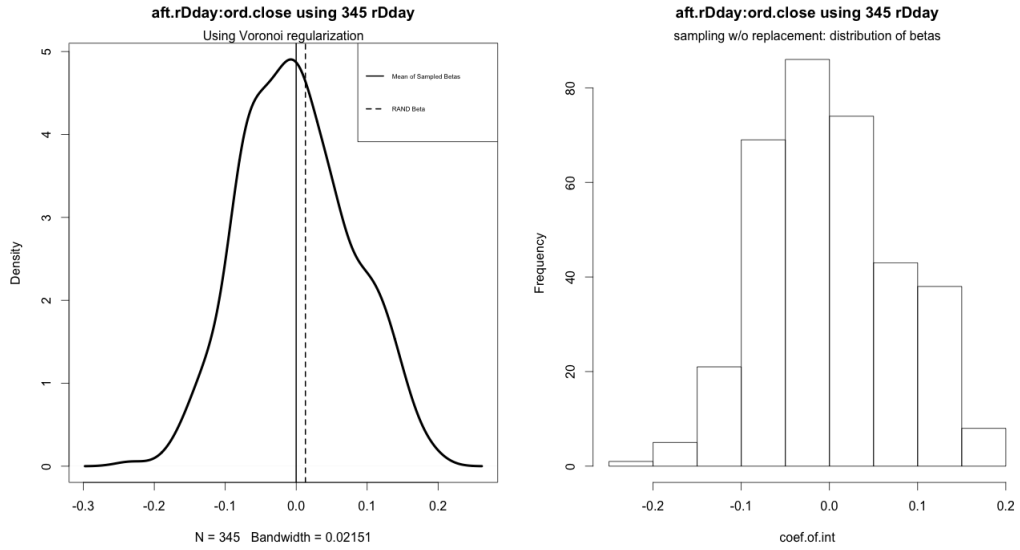

Figure 4.14: Voronoi Bin: Distribution of $\hat{\beta}^*$ using 345 samples w/ replacement

locations, affects the null distribution of $\hat{\beta}$, the parameter of interest. The Voronoi tessellation has attractive qualities (non-parametric and disjoint) when constructing the spatial domain to be analyzed via a regression model. The slight drawback

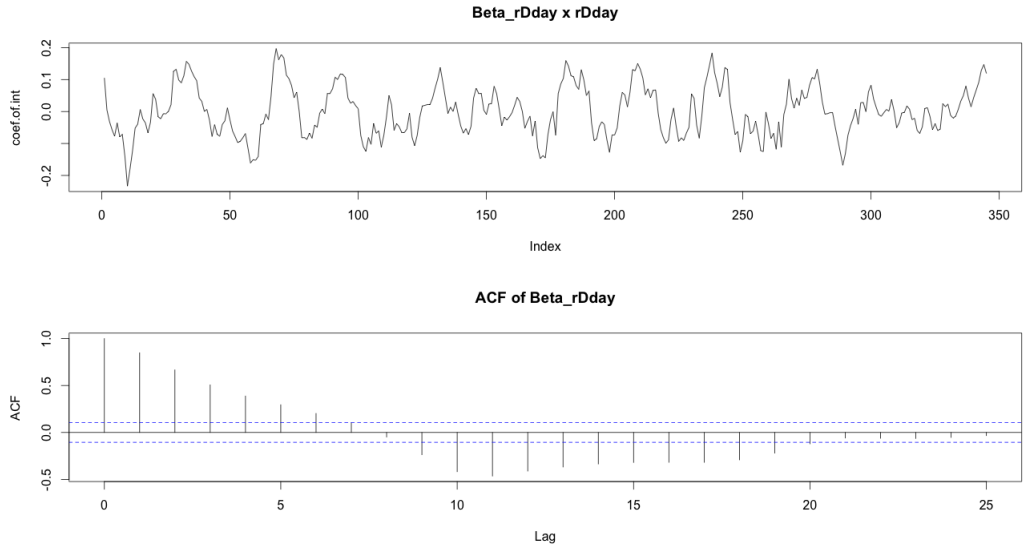

Figure 4.15: Voronoi Bin: $\hat{\beta}^*$ against ordered rDdays

is for some fringe generating spots, in our case marijuana dispensaries, the area associated to the dispensary, in the form of a tessellation, can be quite large. When aggregating the crime counts to an annual level, the crime counts falling within a large tessellation changes the distribution of crime occurrences quite drastically. However, for the purpose of studying a smaller time frame, 21-days, the inflation of crime counts due to a large Voronoi tile, diminishes. Interpreting the voronoi-binned null distribution of $\hat{\beta}$, we again see that the value of 0.013 is quite common for 2009.

4.5 The Voronoi Estimator for Dispensary Intensity

In the preceding, we have used the tessellation of one point pattern to regularize another point pattern. There are however, quite diverse applications of the tesselations. Another use of the Voronoi tiles would be to apply its properties as an estimator. Barr and Schoenberg, in a 2010 paper [BS10], have explored the statistical properties of a non-parametric estimator based on the Voronoi tesselation. The authors have shown that the Voronoi Estimator takes on low bias when estimating the intensity of an inhomogeneous poisson process. For our paper's purposes, we will implement the Voronoi Estimator on the point pattern of marijuana dispensaries. The goal would be to estimate the intensity of the expected number of dispensaries.

The Voronoi Estimator is defined by

$$\hat{\lambda}_y = 1/\mu(C_i). \tag{4.7}$$

For a point pattern $P = \{p_1, p_2, \dots p_n\}$, let the set of tiles $\{C_1, C_2, \dots C_n\}$ represent the Voronoi tessellation generated by the point pattern. Let $\mu(.)$ be the Lebesgue measure, which corresponds to area when considering S, a compact sub space of R^2. The subscript y in $\hat{\lambda}_y$ indicates that we are estimating the potentially

varying intensities

$$\lambda_y = \lim_{\delta \to 0} \frac{E[NB(y;\delta)]}{\pi\delta^2}. \tag{4.8}$$

Thus, we are considering an inhomogeneous poisson process. That is, we allow for different means (intensities) of any $y \in S \subset R^2$.

Entertaining the idea that our dispensaries are a realization of an inhomogeneous poisson process is motivated by our previous exploration. In section 4.2, we have looked at the presence of clustering and inhibition of our dispensaries through J/K/L functions. The set of marijuana dispensaries exhibited characteristics that are similar to an inhomogeneous poisson process. For our study, the set of marijuana dispensaries $P = \{d_1, d_2, \ldots d_{424}\}$ will be our point pattern; generating the Voronoi cells $\{C_1, C_2, \ldots C_{424}\}$ shown in Fig. 4.7. We can look at the areas of cells $\mu(C_i)$ (measured in ft^2 as recorded by the LAPD), through its distribution shown in the first row of Figure 4.16. Of interest would be the inverse of the areas; (second row of Fig. 4.16) as these are the Voronoi estimates for the inhomogeneous intensity.

The distribution of area is quite peaked near zero with a right tail since many of our cells C_i have small area (measured in ft^2) while certain fringe cells have rather large area. By looking at the log-scaled version, we see a more suitable visual. In Figure 4.17, we continue to look at the estimates of intensity through the Voronoi Estimator, by plotting the estimates overlaid on each cell C_i.

Looking at the scale of the intensity and log(intensity), returned from our Voronoi Estimator, we see that the intensity is small (and even negative on the log scale) since this is given in units per ft^2. This is saying for a given location $y \in S \subset R^2$ and looking at a neighboring area measuring a unit ft^2, there is a small expected number of dispensaries. A more reasonable interpretation would come after converting the intensity into a more practical scale such as $mile^2$ (this

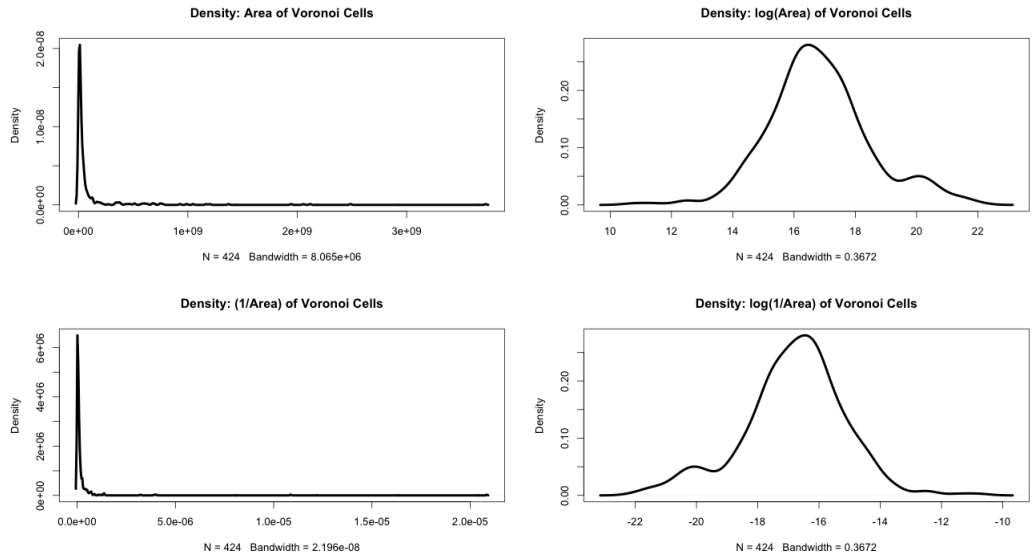

Figure 4.16: Distributions of Voronoi Area and Voronoi Estimator

is left to the interested reader).

As we're assuming our observed point pattern of dispensary locations are an instance of some stochastic process (inhomogeneous poisson), we infer that our Voronoi estimates of intensities are given by Equation 4.6 and shown in Figure 4.17. In conclusion, the Voronoi estimator is quite intuitive in general; and especially for our marijuana dispensaries. For a given point pattern where many of the points are found near each other, there would be many tiles with respectively small area. When the tile's area measure is inverted, we have a large intensity estimate. That is, when we observe many points near each other, the intensity estimate takes on an appropriately high value, as churned out by the Voronoi Estimator.

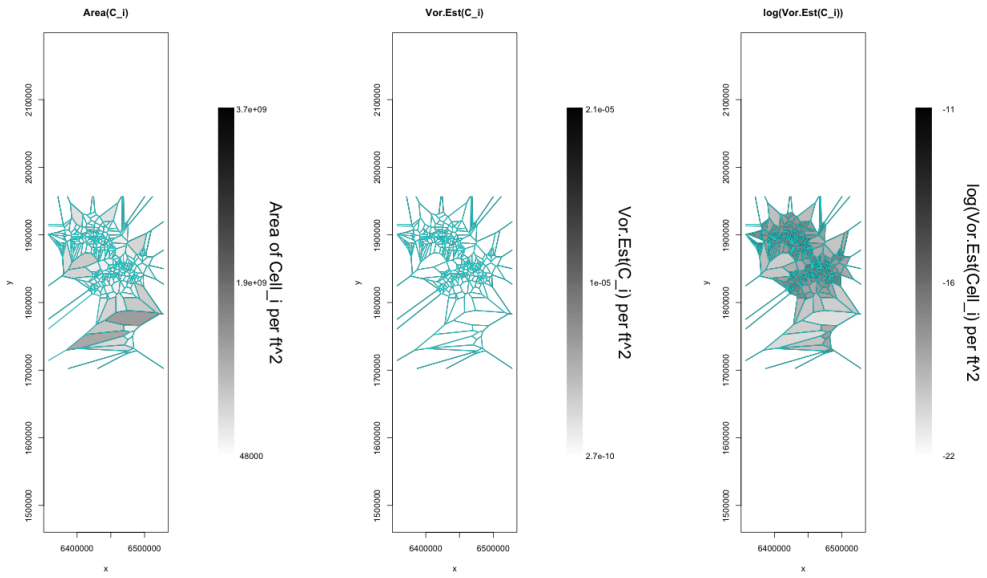

Figure 4.17: Values of Voronoi Estimators for Dispensary Intensity

CHAPTER 5

Concluding Remarks

Through various methods, we have assessed that RAND's estimated value of the 'Marijuana Dispensary Closure Effect' associated with crime counts around a local neighborhood of a dispensary is not an atypical value. By embedding a regression model into a larger dataset, all the while paying close attention in the preservation of an analogous structural form studied by Chang and Jacobson (2011 paper), we have been analyzing a context resembling the one set up by the RAND authors.

Chang and Jacobson's ambitious foray into modeling human behavior has garnered interest from the fervorous marijuana legislation community as well as statistical practitioners. However, the argument that the expected crime counts near marijuana dispensaries increase are associated with the closure of said dispensaries was hinged on a single "smoking gun" parameter estimate.

We have introduced and analyzed a few alternative methodological tools motivated by intuitive statistical underpinnings. The results presented in this paper suggest that the "smoking gun" $\hat{\beta}$ value, obtained by RAND, is not an atypical result. Specifically, when using any randomly chosen day in 2009 as a surrogate for the marijuana dispensary closure date of June 7^{th}, 2011, the observed $\hat{\beta}$ is quite common.

CHAPTER 6

Code

All of the analysis was done in [R] vers. 2.14.2 with use of packages: 'sp,'[PB05] 'fields,'[FNS12] 'spatstat,'[BT05], 'splancs,'[RDa12], 'maptools,'[LBE12] and 'deldir' [Tur11]. As well as a custom script 'geoCovertCoords.R' (by Dave Zes which converts latitude/longitude into the LAPD's coordinate system). The following scripts are suggested to be executed in order.

```
Data Scrubbing
'data_scrub.R'

EDA
'eda.R'
'vor_eda.R'

Ball Bin: Null Distributions
'gen_df_wrangle_ball_bin.R'
'general_samp_coef_lm_model_serv.R'

Point Pattern Analysis
'point_analysis.R'

Voronoi Tess: Null Distributions
'gen_df_wrangle_vor_bin.R'
'general_samp_coef_lm_model_vor_serv.R'

Analyzing Coefficient Output
'analysis_coef_output.R'

Voronoi Estimator
'vor_est_final.R'
```

References

[BS10] C. D. Barr and F. P. Schoenberg. "On the Voronoi estimator for the intensity of an inhomogeneous planar Poisson process." *Biometrika*, 2010.

[BT05] Adrian Baddeley and Rolf Turner. "Spatstat: an R package for analyzing spatial point patterns." *Journal of Statistical Software*, **12**(6):1–42, 2005. ISSN 1548-7660.

[DV03] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2003. Elementary theory and methods.

[ESR] ESRI Developer Network. *Multi-Distance Spatial Cluster Analysis (Ripley's k-function).*

[FNS12] Reinhard Furrer, Douglas Nychka, and Stephen Sain. *fields: Tools for spatial data*, 2012. R package version 6.6.3.

[JCa11] Mireille Jacobson, Tom Chang, and et al. "Regulating Medical Marijuana Dispensaries An Overview with Preliminary Evidence of Their Impact on Crime." Technical report, RAND, 2011.

[LBE12] Nicholas J. Lewin-Koh, Roger Bivand, contributions by Edzer J. Pebesma, Eric Archer, Adrian Baddeley, Hans-JÃrg Bibiko, Jonathan Callahan, StÃ©phane Dray, David Forrest, Michael Friendly, Patrick Giraudoux, Duncan Golicher, Virgilio GÃmez Rubio, Patrick Hausmann, Karl Ove Hufthammer, Thomas Jagger, Sebastian P. Luque, Don MacQueen, Andrew Niccolai, Tom Short, Greg Snow, Ben Stabler, and Rolf Turner. *maptools: Tools for reading and handling spatial objects*, 2012. R package version 0.8-14.

[OBS00] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams.* Series in Probability and Statistics. John Wiley and Sons, Inc., 2nd ed. edition, 2000.

[PB05] E.J. Pabesma and R.S. Bivand. *sp: Classes and methods for spatial data in R*, r news 5 edition, 2005.

[R D12] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[RDa12] Barry Rowlingson, Peter Diggle, adapted, packaged for R by Roger Bivand, pcp functions by Giovanni Petris, and goodness of fit by Stephen Eglen. *splancs: Spatial and Space-Time Point Pattern Analysis*, 2012. R package version 2.01-31.

[Rom] Dennis Romero. "Marijuana Dispensaries Aren't Magnets For Crime, According to RAND Study That Has Some Major Flaws.".

[Tim] LA Times. "Data on the Status of LA Marijuana Dispensaries." http://spreadsheets.latimes.com/status-of-la-marijuana-dispensaries/.

[Tur11] Rolf Turner. *deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation.*, 2011. R package version 0.0-16.