Using Statistical Learning to Forecast Monthly Crime Data in Chicago

---

A Project Presented
to the Faculty of the Graduate School
at the University of Missouri-Columbia

In Partial Fulfillment
of the Requirement for the Degree
Master of Arts

---

by
Andrew Hillard
Project Supervisor
Dr. Christopher K. Wikle

Table of Contents

I. ABSTRACT

The purpose of this study is to use statistical learning to predict whether serious crimes per capita in a Chicago census tract is above or below the four year median. The analyzed data comes from the City of Chicago Crime Records, Weather Underground, the 2010 U.S. Census, and the U.S. Bureau of Labor. The best model is selected from the following methods: Logistic Regression with L1 Penalization, Logistic Regression with L2 Penalization, Random Forest, Gradient Boosted Classification Trees, Support Vector Machine, and Hierarchical Logistic Regression Models with Spatio-Temporal dependence.

II. INTRODUCTION

The year 2016 has been a record year for violent crime in Chicago. The Brennan Center for Justice projects a 47.1% increase in murder compared to 2015 (Mock, 2016). Though the national rate of homicide is expected to increase 13% this year, half of the increase comes from Chicago alone (Mock, 2016). As of November 2016, the homicide count in Chicago was 616, exceeding New York and Los Angeles combined, two U.S. cities with larger populations (Gorner, 2016).

Predictive analytics is an increasingly popular technique to fight rising crime rates. Large consultant companies such as IBM now offer technology solutions that can predict the who, what, when, and where of crime. When implemented for the Manchester Police Department, the IBM crime solution reduced robberies by 12 percent, burglaries by 21 percent, and theft from motor vehicles by 32 percent (Williams, 2016). More recently, the Chicago Police Department (CPD) implemented a machine-learning algorithm that creates a list of individuals who are "most likely to be shot soon or to shoot someone" (Davey, 2016). The high-risk individuals are predicted using data such as past arrests, shootings, affiliations with gangs, and other variables. The CPD uses this list to start conversations with at risk individuals and offer social services to help prevent the crime before it occurs. Of the people shot in Chicago in 2016, 70% have been on the predicted list, demonstrating its efficacy.

The models proposed in this paper provide a different approach than the CPD algorithm. Instead of predicting high-risk individuals, the models predict high-risk locations, specifically which Chicago census tracts will have serious crimes per capita that are above the four year median. More importantly, however, the proposed models are accessible to citizens,

not just law enforcement (the CPD algorithm is only accessible to law enforcement). Public data is used in the analysis and the conclusions can be replicated using the open source software, R. Furthermore, the predictions of the models are tailored to citizen need. The response variable is based on a per capita calculation and can indicate a citizen's risk in a location. The crime category studied in this paper is also broader than the CPD algorithm. Predictions focus on serious crimes; the CPD algorithm only deals with homicides. Serious crime is a category that includes the crimes that concern most individuals: homicide, but also sexual & aggravated assault, robbery, aggravated battery, burglary, larceny, arson, and motor vehicle theft. As a citizen centric model, the techniques proposed in this paper can be publicly replicated to help individuals make safer day to day decisions.

The models can also be used by the CPD. Predictions of higher risk census tracts give a low-resolution indicator of where the occurrence of serious crime is most dense. This information can be used to identify locations where CPD resources are needed most.

Though useful, crime prediction has a problematic limitation: the number of recorded crimes can be a indicator of the amount of policing rather than the actual crime level. The reason for this limitation is because more police in a location results in a higher likelihood that a criminal will be caught, and vice versa. Two locations with identical crime levels can have different levels of reported crime if one location has higher policing levels than the other. This problem is complicated by a positive feedback loop between policing and recorded crime. If an area has a high count of reported crime, the number of police are increased. With more police, the count of reported crime increases. With higher reported crime, more police are sent, ad infinitum. Unfortunately, over-policing, which causes inflated levels of reported crime, tend to occur in areas of high poverty and minority populations (Lum, 2016). As such, the danger of a predictive model is that it would reinforce the practice of over policing in communities that are already disadvantaged. Therefore, the models presented in this paper are to be used by law enforcement with caution, keeping in mind the limitations of reported crime.

III. DATA SOURCES

The following datasets are used in the analysis.

(1) Chicago Crimes – 2001 to Present

Chicago crime data are downloaded from the City of Chicago data portal. The dataset contains "reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to the present" (Crimes - 2001 to present, 2016). For each reported crime, the dataset contains the date and time, block identification, type of crime, crime description, latitude and longitude, and FBI code. The link to the data is given below.

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

(2) Weather Underground Data

Daily weather data from 2010 to the present are downloaded from Weather Underground for Chicago's Midway Airport. Though slightly outside the boundaries of Cook County (downtown Chicago), the airport is chosen because its weather records are detailed and there are no missing observations. The data include 25 daily variables including the maximum, minimum, and average values for temperature, dew point, humidity, pressure, sea level, visibility, wind speed, and precipitation. The data also provide a measure of cloud cover and the weather event for the day (i.e. rain, thunderstorm, snow, or fog). The link to the data is given below.

https://www.wunderground.com/history/airport/KMDW/

(3) U.S. 2010 Census Data

The data for the U.S. 2010 Census are downloaded for Cook County, Chicago, using the R packages "UScensus2010" and "UScensus2010tracts." The data include 457 observations on population and housing demographics for each census tract. The variables in the dataset are summarized in the 2010 Census Summary File (2010 Census Summary File). The U.S. 2010 Census data are used rather than American Community Survey data because of its increased accuracy and decreased variance. The link to the data is given below.

http://lakshmi.calit2.uci.edu/census2000/R/src/contrib/UScensus2010tract_1.00.tar.gz

(4) Census Population Estimates

Data for the estimated annual population of Chicago are downloaded from the U.S. Census Population Estimates Program. The population estimates are for Cook County, Chicago and given for the years 2010 through 2016. The link to the data is given below.

http://www.census.gov/popest/data/historical/index.html

(5) Bureau of Labor Statistics Data

Economic data are downloaded from the Bureau of Labor Statistics (BLS) for the greater Chicago area (Chicago-Joliet-Naperville). The BLS data include monthly values for unemployment, hours & earnings, and consumer price index from January 2006 to September 2016. The link to the data is given below.

http://www.bls.gov/eag/eag.il_chicago_md.htmci

(6) Chicago Census Tract Boundaries

The boundaries of the 2010 census tracts are downloaded as shape files from the City of Chicago data portal. The link to the data is given below.

https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Census-Tracts-2010/5jrd-6zik

(7) Chicago Crime Categories

The FBI crime classification table is obtained from the Chicago police CLEARMAP, Chicago Police Department's Geographic Information System. The table classifies FBI crime codes into the following crime categories: crimes against Persons, Property, or Society, and crimes that are More or Less Serious. The link to the data is given below.

http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html

IV. DATA MANIPULATION

The Chicago Crimes dataset is imported, filtered for complete cases, and manipulated so that each recorded crime gives the following observations: date, latitude, longitude, and FBI Code.

A. Creation of Predictor Variables

The following predictors are created.

- *Census Tract ID* – Using the shape file that defines the boundaries of Chicago's census tracts, each observed crime is matched with the census tract ID in which the crime occurred. The created predictor is a factor with 801 levels corresponding to the 801 census tracts in Cook County.
- *Population and Housing Demographics* – The 457 observations from the 2010 U.S. Census are merged with the observed crimes using the census tract ID. The created predictors are a mixture of factors and numeric variables, as defined by the 2010 Census Summary File.
- *Year and Month* - The date column is manipulated to create two columns that provide the numeric value of the year and month for each reported crime.
- *Daily Weather* – The twenty five Weather Underground observations are aggregated to give average values by month. The aggregated weather data are merged with the Chicago crime data using the year and month. The weather predictors are numeric predictors.
- *Annual Population of Chicago Area* – The annual estimated population of the Chicago area is merged with the Chicago Crime data using the year variable. The annual estimated population is a numeric predictor.
- *Monthly Bureau of Labor Statistics* – The BLS monthly data for unemployment, wages & earnings, and consumer price index, are merged with the Chicago Crime data using month and year. The BLS predictors are numeric.

Table 1. Summary of Predictors.

Category	Number of Variables	Classification	Resolution
Census Tract ID	1	Spatial	Census Tract
2010 Census Data	457	Spatial	Census Tract
Weather	25	Temporal	Average By Month
Date	2	Temporal	Year and Month
Population	1	Temporal	By Year
Unemployment	1	Temporal	By Month
Wages & Earnings	1	Temporal	By Month
Consumer Price Index	1	Temporal	By Month
Total	**489**		

B. Creation of Response Variable

The following steps are taken to create the response variable.

(1) Each recorded crime is classified as "more serious" or "less serious" using the crime type classification table.
(2) The monthly number of more serious crimes is calculated for each census tract.
(3) The monthly number of more serious crimes per capita is calculated for each census tract by dividing the monthly number (calculated in step 2) by the tract's population. Census tract population is given in the 2010 U.S. Census Data as variable "P0010001."
(4) The median value of more serious crimes per capita across all census tracts and years is calculated for 2010 to 2013, the years used in the training data.
(5) For each month, a census tract is classified as having an above median or below median level of more serious crimes per capita.

C. Creation of Training and Test Dataset

The data is split into a training and test data set. The training data are the observed values for the response and predictor variables from 2010 to 2013. The test data are the

observations from 2014 to 2015. The year 2016 is not included in the analysis because the year is incomplete and would lack any predictors that are annual.

V. EXPLORATORY DATA ANALYSIS

The data are explored to see if the response variable substantially varies by time. The count of census tracts that exceed the median value of serious crimes per capita is aggregated and plotted for different values of year and month (Figure 1 and 2). Visually, the response variable changes with the year and month, indicating the possibility of a temporal effect.

Figure 1 shows a decline in the count of the response variable between the years 2010 through 2015. The decline in Chicago crime during these years has been observed in previous research (Crime in Chicago, 2016; Papachristos, 2013). The downward trend is also supported by a logistic regression that uses the years 2010 through 2015 as predictors. Each year except 2011 is a significant predictor at the .01 level, and the coefficients become more negative as the year increases.

Figure 2 shows a higher count of above median serious crimes during the summer months. The increase in crime between June and August is a documented crime phenomenon (Lauritson, 2014). The effect of the month on the response is also supported by a logistic regression that uses the months as predictors. Nine out of the twelve months are significant and the coefficients reflect the trends in Figure 2, a higher probability of crime during the summer months.

The interaction between year and month is also explored by plotting the response variable as a time series over the months between 2010 and 2015 (Figure 3). The time series reflects the trends already identified in Figure 1 and Figure 2: decreasing crime rates over the years and higher crime during the summer months. The time series trend indicates a possible temporal dependence.

Figure 1: Total Number of Census Tracts Above the Monthly Median of Serious Crimes Per Capita in Given Year

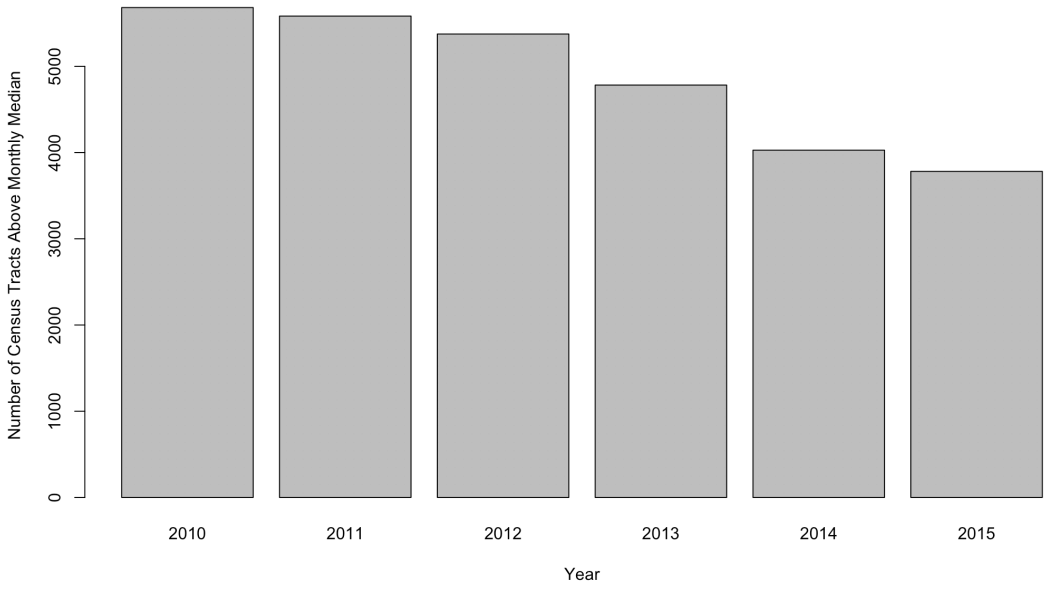


Figure 2: Total Number of Census Tracts Above the Monthly Median of Serious Crimes Per Capita in Given Month

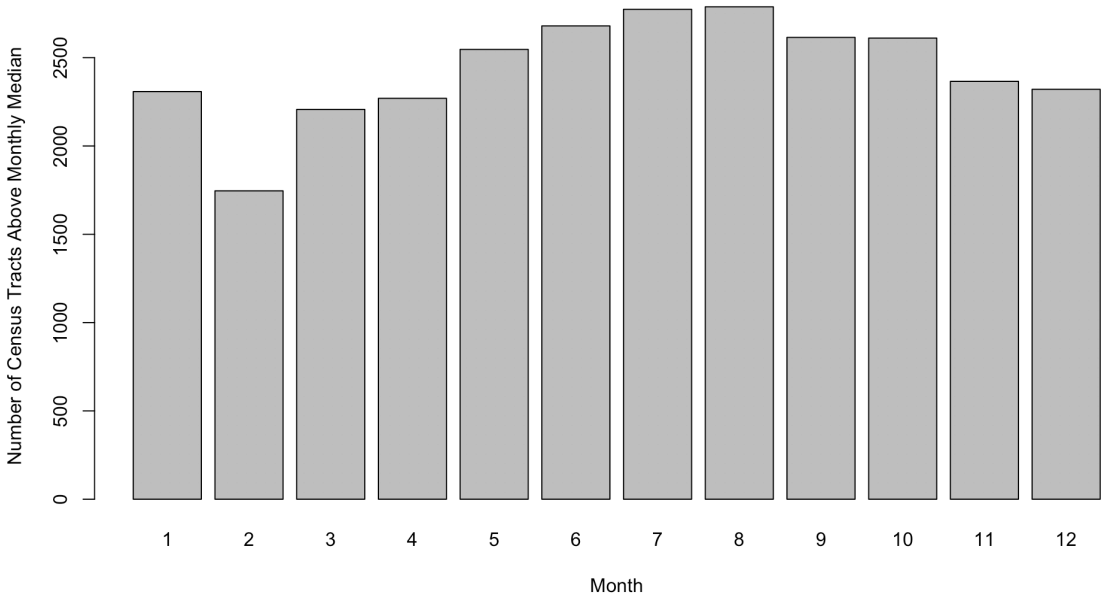
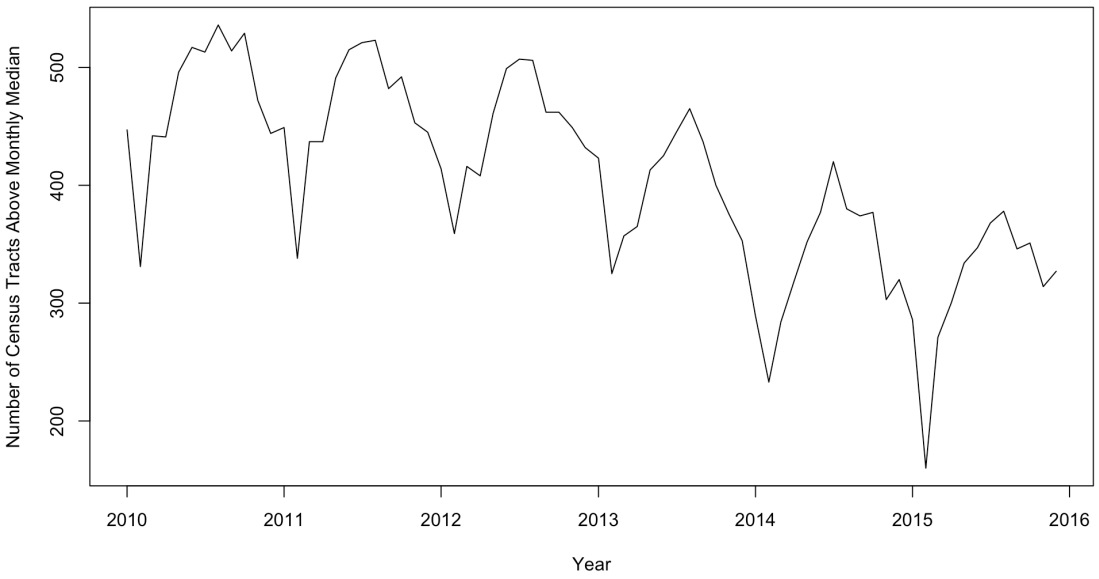


Figure 3: Total Number of Census Tracts Above the Monthly Median of Serious Crimes Over Time - By Month



The data are then explored for spatio-temporal dependence by plotting the map of census tracts and overlaying the response variable for January, April, July, and October of 2010 (Figure 4). These months were used as a proxy to see if the spatial arrangement of the response variable changed with month.

Figure 4 exhibits two notable trends: the census tracts are clustered into high and low crime areas, and clusters vary slightly as the month changes. The areas in which the response variables are clustered correspond to areas that are commonly assumed to have higher levels of crime, poverty rates, and racial minorities, West and South Chicago (James, 2016). The patterning of clusters indicates potential usefulness of the 2010 Census data as well as neighborhood dependence in hierarchical models. The variations in clusters as the month changes also suggests that there is a month and space interaction that can possibly be incorporated into a spatio-temporal model.

The spatial map is also analyzed over four years (month held constant) to explore whether there may be a spatio-temporal relationship between year and space (Figure 5). Similar to Figure 4, the map displays clear clustering in expected areas. The clusters also change over time, which is especially noticeable when one compares 2010 to 2013. This suggests the possibility of a year and space interaction that can be incorporated into a spatio-temporal model.

Figure 4: Census Tracts Above the Monthly Median of Serious Crimes for Different Months in 2010 - Blue: Below Median and Red: Above Median

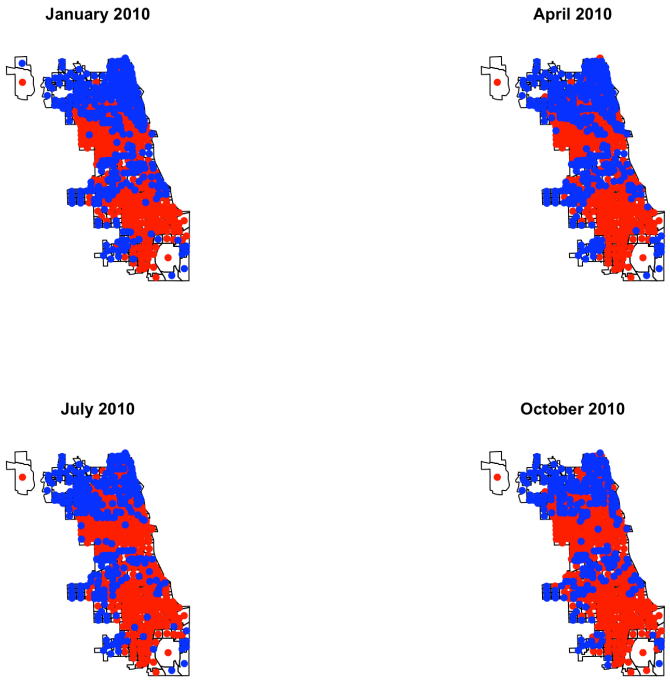
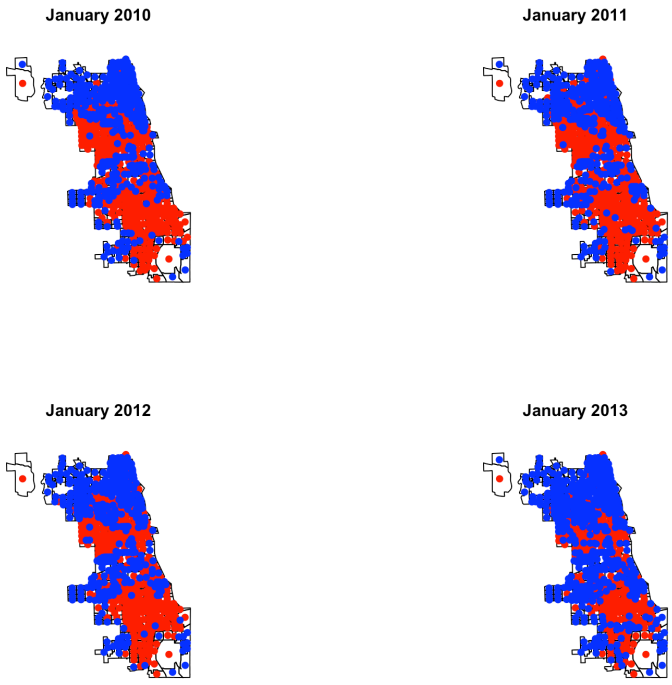
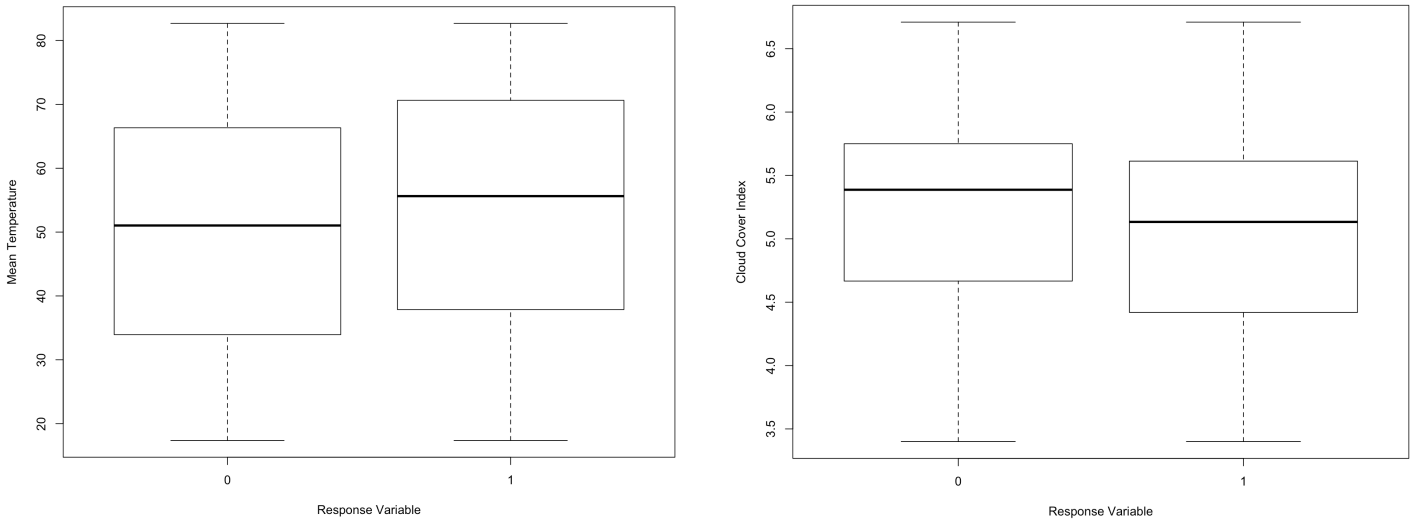


Figure 5: Census Tracts Above the Monthly Median of Serious Crimes for Different Years (Month Held Constant) - Blue: Below Median and Red: Above Median



The relationship between weather data and the response variable is explored by plotting a box plot of each weather variable against the response. The box plots for mean temperature and mean cloud cover are shown below as examples of weather predictors that display a visual trend (Figure 6). The box for the mean temperature of an above median response, coded as 1, is higher than the below median response, coded as 0. The box for the cloud cover of an above median response, coded as 1, is lower than the below median response. The Mann-Whitney U Test is applied to determine whether the visual trend is statistically significant. The Mann-Whitney U Test is purely exploratory and no attempt is made to consider possible spatial or temporal dependencies in the data. The populations of 0's and 1's are found to have statistically significant differences in mean temperature and cloud cover at a 0.01 level, indicating that weather may play a role in predicting the response variable.

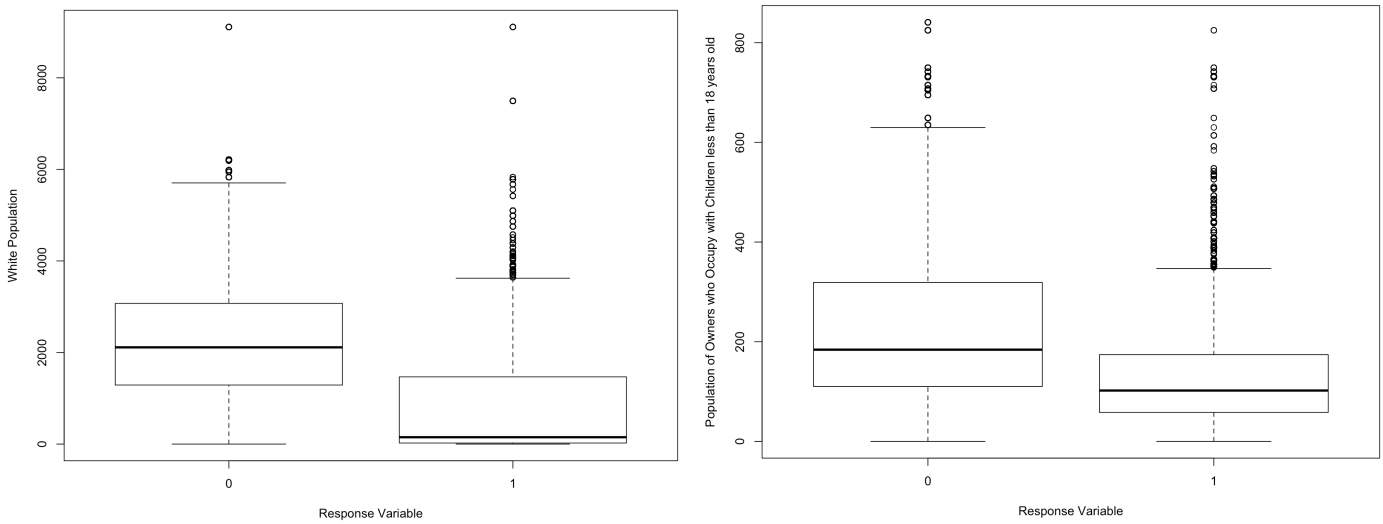
Figure 6: Box Plots of Mean Temperature vs. Response Variable (Left) and Cloud Cover vs. Response Variable (Right)



The relationship between the census data and response variable is explored by plotting box plots of select census variables against the response. Two of the more striking box plots are shown below (Figure 7). The left box plot is a census tract's population of white individuals plotted against the response variable. The right box plot is the census tract's population of housing owners who occupy their housing unit with children under the age of

18 plotted against the response variable. In both boxplots, lower populations of either white race or owners who occupy their housing unit with children under 18 are visually associated with below median serious crimes per capita. The Mann-Whitney U Test is applied to determine whether the visual trend is statistically significant. The Mann-Whitney U Test is purely exploratory and no attempt is made to consider possible spatial or temporal dependencies in the data. The populations of 0's and 1's are found to have statistically significant differences in white population and population of owners who occupy with children less than 18 years old, indicating that census variables may play a role in predicting the response variable.

Figure 7: Box Plots of White Population vs. Response Variable (Left) and Population of Housing Owners who Occupy their Housing Unit with Children under the Age of 18 vs. Response Variable (Right)

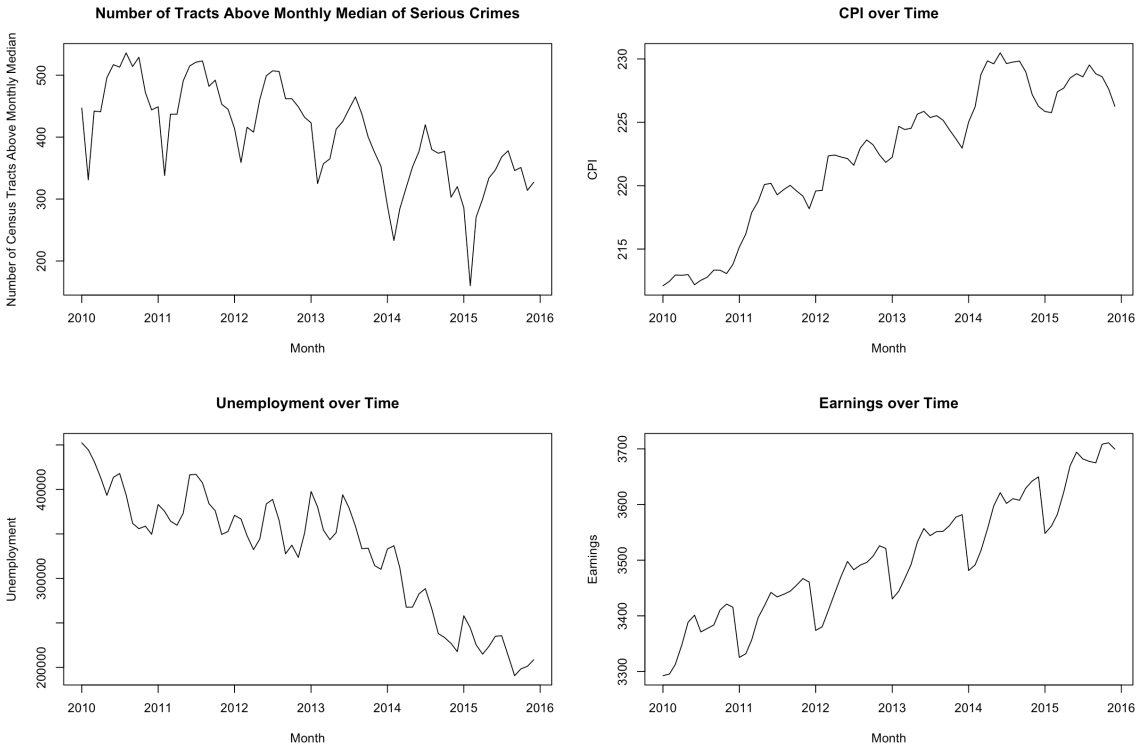


The relationship between the Bureau of Labor Statistics and the response variable is explored by comparing time series graphs between the count of tracts above the median of serious crimes (the response variable), CPI, Unemployment, and Earnings (Figure 8).

Visually, the time series for the response variable (upper left in Figure 8) and CPI (upper right in Figure 8) display a possible inverse relationship. The time series for the response variable and unemployment (lower left in Figure 8) show a possible direct relationship as both trend downward. The most striking relationship is between the response variable and the earnings time series, (lower right in Figure 8), which both exhibit annual seasonality with depressions occurring at the start of each year. The annual seasonality of

CPI, unemployment, and earnings are documented trends in econometric practice (Labor Force Statistics from the Current Population Survey, 2016).

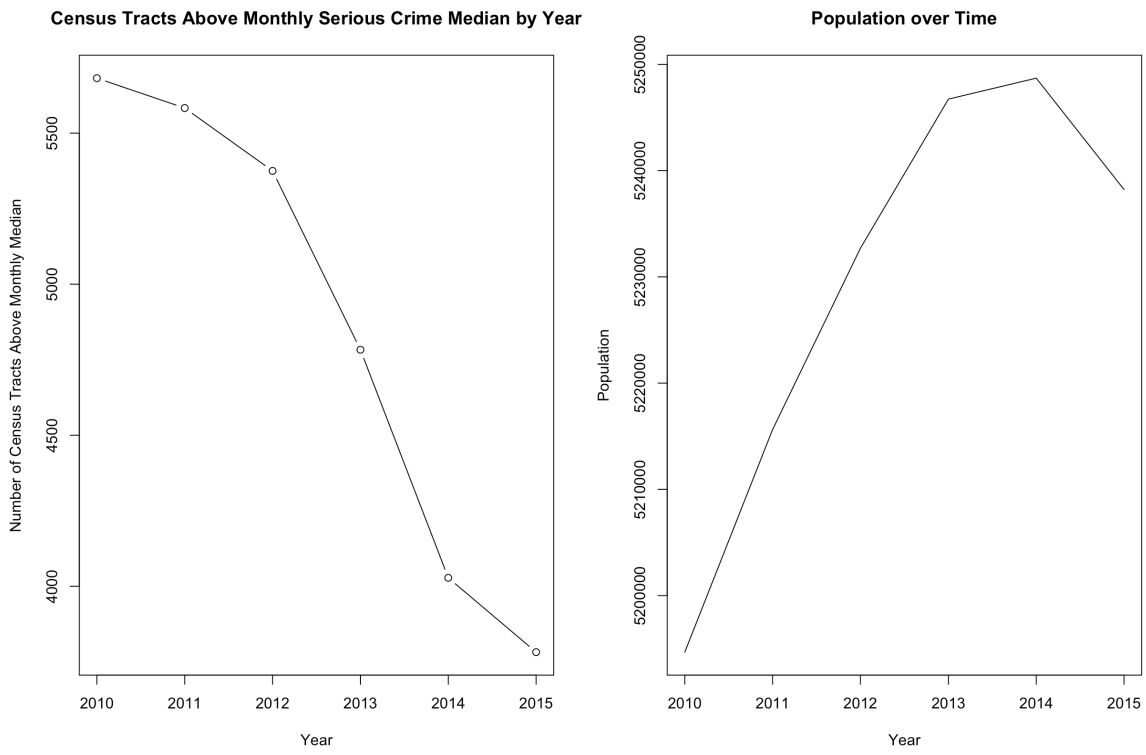
Figure 8: Time Series of Response Variable, CPI, Unemployment, and Earnings



The time series of the population of Chicago is plotted next to the time series for the response variable by year (Figure 9). The two graphs show opposite trends over time, indicating a possible inverse relationship.

Overall, the exploratory analysis seems to suggest that there may be associations between many of the predictors and the response variable.

Figure 9: Time Series of Response Variable and Population



VI. MODELING METHODOLOGY

The data is analyzed using the following statistical methods: Logistic Regression with L1 Penalization, Logistic Regression with L2 Penalization, Random Forest, Gradient Boosted Classification Trees, Support Vector Machine, and Logistic Regression Hierarchical Models with Spatio-Temporal dependence. A brief description of each method and its application is given below.

A. Logistic Regression with L1 Penalization

Logistic Regression with L1 penalization is a technique that uses a L1 (lasso) penalty to shrink regression coefficients to zero. Unlike L2 (ridge) penalization, the L1 penalty can force coefficients to be exactly zero, performing variable selection. Though increasing bias, the L1 penalization can decrease variance and overall test error relative to non-regularized logistic regression. The amount of shrinkage is determined by the value of the parameter

lambda. A large value of lambda (a large penalty) creates more shrinkage and forces more coefficients to zero (James, 2015).

The Logistic Regression with L1 penalty is implemented in R using the "glmnet" package. The "glmnet" package also performs k-fold cross validation to select the optimal value of lambda. Once the lambda parameter is optimized, predictions are generated on the test data and the test error is calculated.

B. Logistic Regression with L2 Penalization

The Logistic Regression with L2 penalization is a technique that uses a L2 (ridge) penalty to shrink regression coefficients to zero. Unlike L1 penalization, the L2 penalty will not force coefficients to be exactly zero. The L2 penalty shrinks low variance components the most, similar to principal component analysis. Though increasing bias, the L2 penalization can decrease variance and overall test error relative to non-regularized logistic regression. The amount of shrinkage is determined by the value of the parameter lambda. A large value of lambda (a large penalty) creates more shrinkage and forces more coefficients to be closer to zero (James, 2015).

The Logistic Regression with L2 penalty is implemented in R using the "glmnet" package. The "glmnet" package also performs k-fold cross validation to select the optimal value of lambda. Once the lambda parameter is optimized, predictions are generated on the test data and the test error is calculated.

C. Random Forest

The Random Forest Classifier is a tree based method that builds a committee of decision trees via bootstrapping. The committee of trees is de-correlated by using a random sample of input variables for each tree split. Because its predictions are based on a committee of de-correlated decision trees, the Random Forest Classifier typically has low bias and variance when compared to individual and bagging tree methods (James, 2015).

The Random Forest Classifier is implemented using the R package "ranger," which is an implementation of Random Forests particularly suited to high dimensions. The "ranger" package is much more efficient than the typical Random Forest forest package, "randomForest." The only parameter to be tuned is the number of input variables randomly selected at each split. This parameter is tuned using the "caret" package, which compares the

out-of-bag error rates for different parameter values. Once tuned, the Random Forest Classifier is built, predictions are generated on the test data, and the test error is calculated.

D. Gradient Boosted Classification Trees

Gradient Boosting Classification Trees is a method that combines a series of weak classification trees (weak learners) to make predictions. Weak learners are built sequentially, and the newest weak learners are trained to focus on observations that previous weak learners struggled to classify. Final predictions are made by a majority vote where the weak learners are weighted for accuracy (Brownlee, 2016).

The "xgboost" package in R performs gradient boosting for decision trees. The xgboost model is tuned by adjusting the values for the following 6 parameters (Aarshay, 2016).

- nrounds: The number of trees built.
- max_depth: The max depth of each tree built.
- eta: The learning rate for gradient boosting.
- gamma: The minimum loss reduction required to make a split.
- colsample_bytree: The number of columns to sample and use in each tree split.
- subsample: Fraction of observations to be randomly sampled for each tree.

The parameters are iteratively tuned using a grid search in the "caret" package. Once tuned, predictions are generated on the test data and the test error is calculated.

E. Support Vector Machine

The Support Vector Machine is a method that builds a hyperplane that best separates the data into the target classes. When training a Support Vector Machine, the number of misclassified observations (those on the wrong side of the hyperplane) is controlled by the cost parameter. Lower cost results in more misclassification, higher bias, but lower variance. Higher costs results in less misclassification, lower bias, but higher variance. The hyperplane can be a variety of linear and non-linear shapes by transforming the inputs using kernels. Outside of cost, each kernel has its own unique parameters to be tuned (James, 2016).

The Support Vector Machine is implemented using the R package called "e1071." The package allows you to select from a variety of kernels and tune the parameters using a grid search. The kernel selected for testing in this analysis is the linear kernel. For the linear

kernel, only the cost parameter needs to be tuned. Once the parameters are optimized, predictions are generated on the test data and the test error is calculated.

F. Spatio-Temporal Hierarchical Models with INLA

The spatio-temporal models in this paper are Bayesian hierarchical models for logistic regression with spatial and temporal random effects. The main challenge for these models is computational efficiency, since the typical Markov Chain Monte Carlo methods are constrained by model complexity and database dimension. A solution to this challenge is provided by the R package "INLA," which uses the Laplace approximation instead of MCMC. The Laplace approximation is a good approximation and highly efficient (Blangiardo, 2012).

For each model, the i^{th} census tract is modeled as a binomial distribution and the link function for the logistic regression is assumed to be the logit. The specific distributions of the regression covariates for the linear predictor are given below for each spatio-temporal model analyzed.

Spatio-Temporal Model 1

Spatio-Temporal Model 1 assumes structured and unstructured spatial components as well as a linear interaction between time and space. The structured spatial component uses a Besag-York-Mollie specification for neighborhood dependence (Blangiardo, 2012). No covariates are used in Model 1.

$$\eta_{it} = \alpha + v_i + v_i + (\beta + \delta_i) \times t,$$

where,

η_{it} represents the linear predictor transformed on the logit scale,

α = intercept constant,

$v_i \mid v_{j \neq i} \sim Normal(m_i, s_i^2)$; spatial random effect,

$m_i = \frac{\sum_{j \in N(i)} v_j}{\#N(i)}$ and $s_i^2 = \frac{\sigma_v^2}{\#N(i)}$,

$\#N(i)$ is the number of areas which share boundaries with the i^{th} census tract,

$v_i \sim Normal(0, \sigma_v^2)$,

β = main linear trend or global time effect,

δ_i = differential trend or interaction between time and space,

t = time predictor or sequence of months from 2010 to 2015.

Spatio-Temporal Model 2

Spatio-Temporal Model 2 assumes the same structured and unstructured spatial components of Model 1, but replaces the linear time space interaction with a structured and unstructured time component. The structured time component assumes a random walk (Blangiardo, 2012). No covariates are used in Model 2.

$$\eta_{it} = \alpha + v_i + v_i + \gamma_t + \phi_t,$$

where,

$$\gamma_t | \gamma_{-t} \sim Normal(\gamma_{t+1}, \tau_\gamma), \qquad \text{for } t = 1,$$

$$\gamma_t | \gamma_{-t} \sim Normal(\tfrac{\gamma_{t-1}+\gamma_{t+1}}{2}, \tfrac{\tau_\gamma}{2}), \qquad \text{for } t = 2, \dots, T-1,$$

$$\gamma_t | \gamma_{-t} \sim Normal(\gamma_{t-1}, \tau_\gamma), \qquad \text{for } t = T,$$

$$\phi_t \sim Normal(0, \tau_\phi).$$

Spatio-Temporal Model 3

Spatio-Temporal Model 3 assumes the same structured and unstructured time and space components as Model 2, and adds a space time interaction component (Blangiardo, 2012). No covariates are used in Model 3.

$$\eta_{it} = \alpha + v_i + v_i + \gamma_t + \phi_t + \delta_{it},$$

where,

$$\delta_{it} \sim Normal(0, \tau_\delta).$$

Spatio-Temporal Model 4

Spatio-Temporal Model 4 assumes the same structured and unstructured components as Model 3, and adds covariates that are selected from the most important variables in Random Forest.

$$\eta_{it} = \alpha + v_i + v_i + \gamma_t + \phi_t + \beta_1 x_1 + \cdots + \beta_5 x_5,$$

where,

x_1 = H0060002 predictor,

β_1 = the constant coefficient for the H0060002 predictor,

x_2 = H0130005 predictor,

β_2 = the constant coefficient for the H0130005 predictor,

x_3 = CPI predictor,

β_3 = the constant coefficient for the CPI predictor,

x_4 = mean wind predictor,

β_4 = the constant coefficient for the mean wind predictor,

x_5 = earnings predictor,

β_5 = the constant coefficient for the earnings predictor.

VII. RESULTS AND ANALYSIS

A. Model Selection

The parameters for the non-Bayesian models are optimized using 5-fold cross validation. The values for the tuned parameters are given in Table 2 and the test error for all models is given in Table 3.

Table 2: Optimized Parameter Values

Model Type	Parameter Values
L1 (Lasso) Logistic Regression	lambda = 0.00033
L2 (Ridge) Logistic Regression	lambda = 0.02667
Random Forest	mtry = 576
Gradient Boosting	nrounds = 45, max_depth = 7, eta = 0.1, gamma = 0, colsample_bytree = 1, subsample = 1,
Support Vector Machine	cost = 71

Table 3: Test Error Results

Model Type	Test Error
L1 (Lasso) Logistic Regression	15.93%
L2 (Ridge) Logistic Regression	16.00%
Random Forest	16.26%
Gradient Boosting	17.03%
Support Vector Machine	23.05%
Spatio -Temporal 1	15.86%
Spatio -Temporal 2	15.23%
Spatio -Temporal 3	15.23%
Spatio -Temporal 4	16.46%

The models with the lowest test error, 15.23%, are Spatio-Temporal Model 2 and Spatio-Temporal Model 3. The deviance of these two models is very close, 24276.3 and 24275.4 respectively (Table 4). Therefore, applying the principle of parsimony, Spatio-Temporal Model 2 is selected as the best model for predictions.

Spatio-Temporal Model 4 is also considered since it has lower deviance than the other Spatio-Temporal models and includes covariates, making it potentially robust for future predictions. However, Spatio-Temporal Model 4 is not selected as the best model since it has the second highest test error.

Table 4: Deviance Table for Spatio-Temporal Models

Model	Mean Deviance	P_d	DIC
Spatio - Temporal I	24877.4	887.2	25764.7
Spatio - Temporal II	23468.4	807.9	24276.3
Spatio - Temporal III	23466.7	808.7	24275.4
Spatio - Temporal IV	23449.1	776.0	24225.1

B. Variable Importance

Despite performing worse than the logistic regression methods, the Random Forest provides a sense of which variables may have a relationship to the response. The top twenty variables in the Random Forest are given in Table 5.

The census variables with the strongest effect on the Random Forest impurity measure are H0060002, H0080002, P0060002, P0080003, P0030002, H0130005, and H0130006. The first three of these variables correspond to the population of householders or citizens who are white or white with one or more other races. The connection between crime and race in Chicago is documented in previous studies; Chicago's highest rates of homicide are happening in the poor, "extremely segregated neighborhoods on the South and West sides" (Fessenden, 2016; Massey, 1994). The other census variables correspond to the number of 4 person households, or 5 person households.

The non-census variables with the strongest effect on the Random Forest include ten weather related variables, Consumer Price Index, Earnings, and Unemployment. The effect of weather on crime is documented in previous studies. Researchers theorize that during pleasant weather, more people are outside and more likely to become victims of crime: either interacting with criminals or leaving property unattended (Cohn, 1990). The research on CPI, Earnings, and Unemployment's effect on crime is less consistent with different studies reaching different conclusions (Fox, 1978; Box, 1987; Witt, 1998).

Table 5: Random Forest Variable Importance Plot

Variable	Gini Index Change
H0060002	1173.80278860882
H0080002	1158.05805846959
P0060002	1085.70338640557
P0080003	862.508212357395
P0030002	849.421962124525
H0130005	377.830644969064
H0130006	351.133558731341
CPI	331.599362252181
mean_wind	331.039443240164
Earnings	308.604305661514
cloud_cover	296.847553824876
winddirdegrees	292.862161450641
Unemployment	273.967905667117
min_hum	269.630997350149
precipitation	262.967197641126
min_dew	261.989060348286
mean_vis_miles	258.691592749687
mean_dew	256.437377327231
min_temp	255.311536704458
mean_hum	252.721558750243

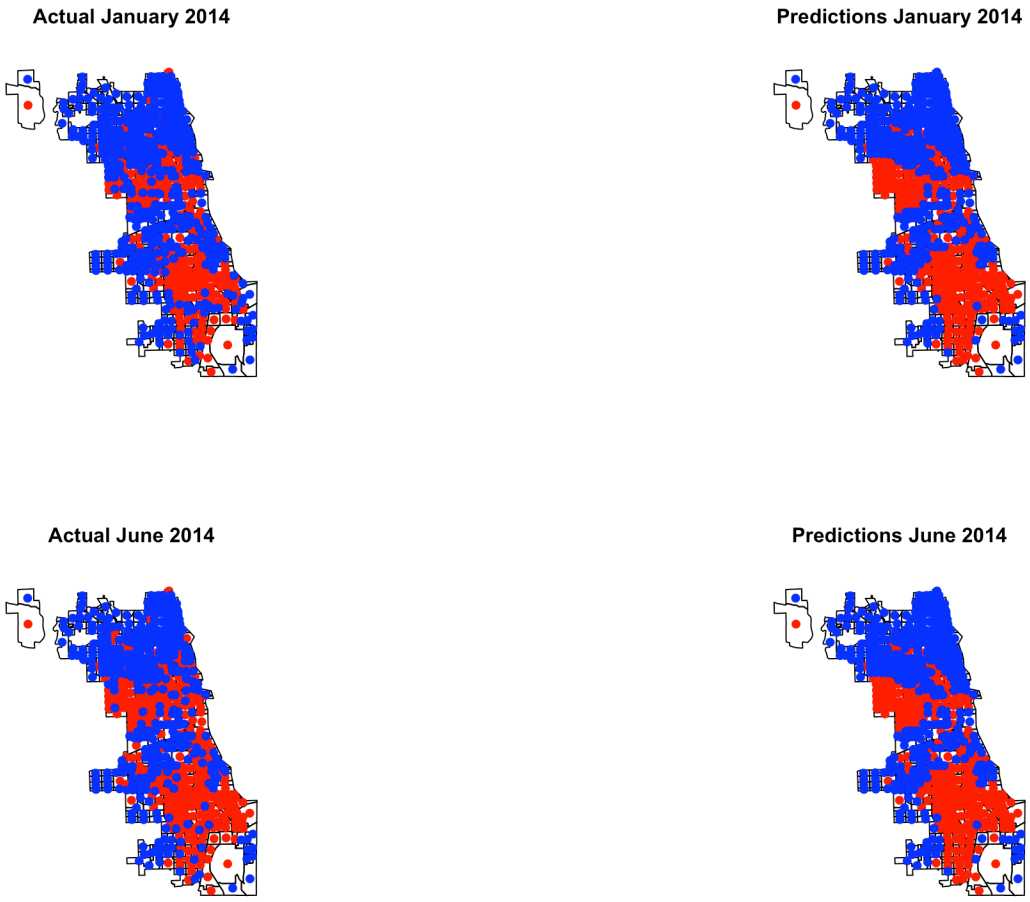
VIII. CONCLUSION

The best model for predicting whether a census tract is above or below the median level of serious crimes per capita for a given month is Spatio-Temporal Model II, a model that assumes structured and unstructured spatial and temporal components. Despite not

utilizing covariate data, the model provides predictions that yield a 15.22% error rate, which is the lowest error rate of the tested models. The next best error rates are 15.86%, Spatio-Temporal Model 1, and 15.93%, Logistic Regression with L1 Penalization.

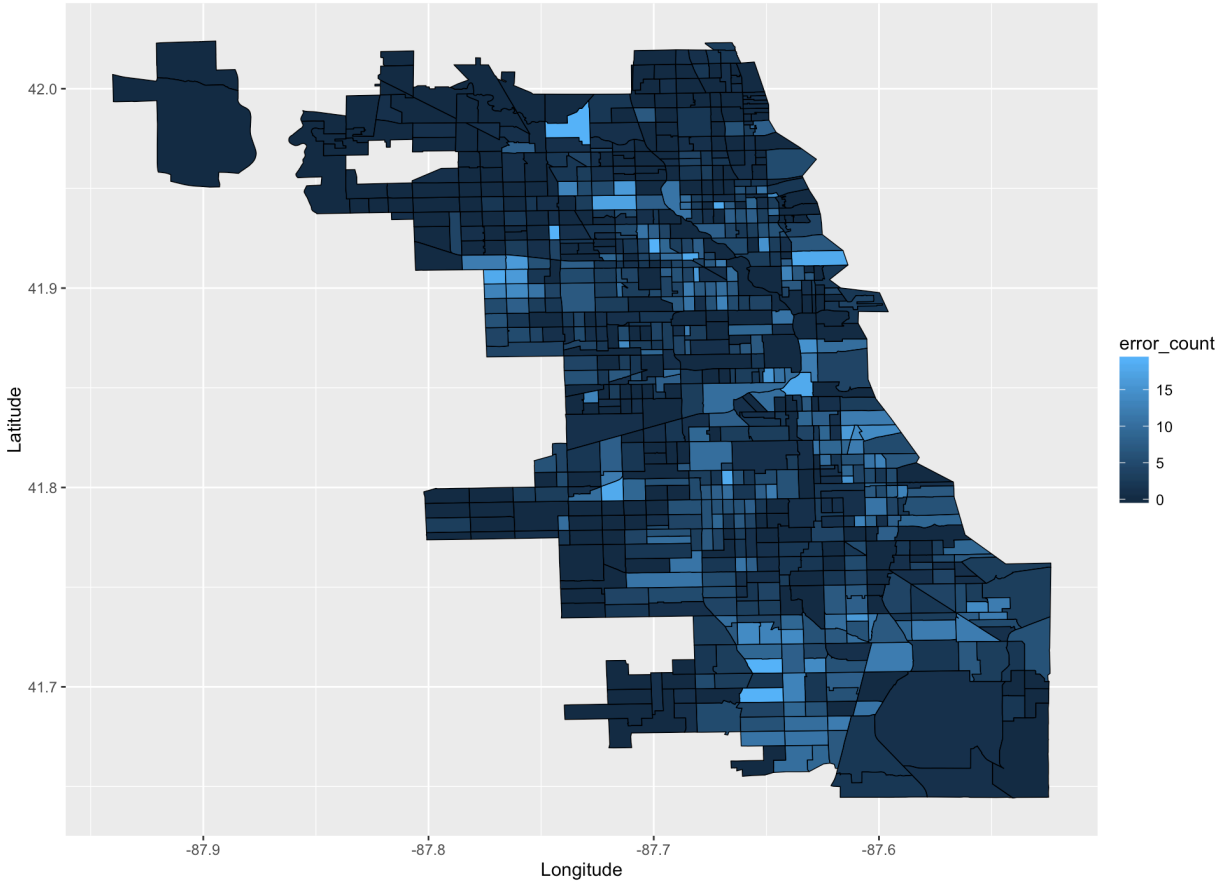
To get a sense of the predictions, the actual and predicted response variable for two months are mapped in Figure 10, January 2014 and June 2014. The predictions follow the clustering pattern of the actual values though fail to adjust to the month to month variability.

Figure 10: Actual vs. Predicted Census Tracts above the Median Level of Serious Crimes Per Capita - Blue: Below Median and Red: Above Median



The prediction errors for the 2014 and 2015 predictions are aggregated and mapped by census tract in Figure 11. The census tracts that are closest to the boundaries where areas of above median serious crimes per capita intersect with below median serious crimes per capita appear to have the highest error rates. This trend is in line with the assumption of neighborhood dependence assumed by Spatio-Temporal Model 2.

Figure 11: Prediction Error for Census Tracts -
Lighter Blue indicates Higher Number of Prediction Errors



Practically, citizens can use the predictions made in this paper to make month to month decisions such as where to drive or where to buy a home. For example, the ability to create driving routes based on safety is currently used by the driving application Waze, which warns drivers about high crime neighborhoods (Moffett, 2016). The predictions in this paper, however, are limited by the size of the studied location, census tract, and time period, month. The predictions cannot give information for a specific block or street. In traffic routing, this is a serious limitation as a safe highway could pass through a dangerous census tract. Likewise, predictions by month will not give information about the fluctuation of crime levels for smaller time periods such as a week, day, or hour. It is therefore recommended that citizens use the predictions as a high level indicator supplemented with more granular, local knowledge.

IX. FURTHER APPLICATIONS

The analysis in this paper focused on serious crimes per capita at a monthly level. The techniques can also be applied to crime categories other than serious, such as personal, property, and societal crimes. Likewise, the techniques can be applied to yearly, weekly, or daily data. In addition, the covariates used in this analysis only included census, weather, and BLS data. The same techniques can be applied to additional predictors such as Twitter, or Google places data.

Preliminary analysis of different crime categories and time granularity was done before focusing on monthly serious crimes. The data was initially aggregated by year and analyzed on personal, property, and societal crimes. The logistic regression model with and without L1 penalization was applied on this data using American Census Society data as predictors. The test errors depended on the type of crime being analyzed and ranged from 12.19% to 19.30%. Likewise Random Forest methods were run to analyze annual levels of personal, property, and society crimes and yielded error rates as low as 14.30%. The methods of this paper were also applied to aggregated crime at the daily level but were soon abandoned when simple random forests with one parameter took twelve plus hours to complete.

X. REFERENCES

Blangiardo, Marta, Michela Cameletti, Gianluca Baio, and Havard Rue. "Spatial and Spatio-Temporal models with R-INLA." Spatial and Spatio-Temporal Epidemiology, , 2012.

Box, Steven. Recession, Crime and Punishment. London, Palgrave, 1987.

Brownlee, Jason. "A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning." Machine Learning Mastery, 9 Sept. 2016.

Cohn, Ellen. "Weather and Crime." Brit. J. Criminol, Vol. 30, No. 1, Winter 1994.

"Crime in Chicago by month, 2001 to present." Chicago Tribune, 20 Nov. 2016, crime.chicagotribune.com/.

"Crimes - 2001 to present." City of Chicago Data Portal , City of Chicago, https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data.

Davey, Monica. "Chicago Police Try to Predict Who May Shoot or Be Shot." The New York Times, 23 May 2016.

Fessenden, Ford, and Haeyoun Park. "Chicago's Murder Problem." New York Times, 27 May 2016.

Fox, J. Forecasting Crime Data - An Econometric Analysis. New York, Lexington Books, 1978.

Gorner, Jeremy. "August most violent month in Chicago in nearly 20 years." Edited by Peter Nikeas and Elvia Malagon, Chicago Tribune, 29 Aug. 2016.

Jain, Aarshay. "Complete Guide to Parameter Tuning in XGBoost (with codes in Python)." Analytics Vidhya1 Mar. 2016.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning. 6th ed., New York, Springer, 2015.

James, Nick. "These are the 10 Worst Chicago Neighborhoods." RoadSnacks, 10 Feb. 2016.

"Labor Force Statistics from the Current Population Survey." Bureau of Labor Statistics, Bureau of Labor Statistics, 3 June 2016.

Lauritsen, Janet L. "Seasonal Patterns In Criminal Victimization Trends." Bureau of Justice Statistics, , 2014.

Lum, Kristian, and William Isaac. "To predict and serve?" Significance, 7 Oct. 2016.

Massey, Douglas S. "Getting Away with Murder: Segregation and Violent Crime in Urban America." University of Pennsylvania Law Review, 30 Oct. 1994.

Mock, Brentin. "Where Crime Is Rising in 2016." CityLab, The Atlantic, 20 Sept. 2016.

Moffett, Matt. "Waze can now warn you about high-crime neighborhoods in cities. But is that a good idea?." Quartz, 23 Aug. 2016.

Papachristos, Andrew. "48 Years of Crime in Chicago." Institution for Social and Policy Studies, , 2013.

Wikle, Christopher. Statistics 8330: Data Analysis III Lecture Notes. Columbia, MO, University of Missouri, 2016.

Williams, Paula. "6 simple ways to help fight crime with analytics." IBM Big Data and Analytics Hub, 5 May 2016.

Witt, Robert, Alan Clarke, and Nigel Fielding. "Crime, earnings inequality and unemployment in England and Wales." Applied Economic Letters, vol. 5, 1998.

"2010 Census Summary File." , United States Census Bureau, Sept. 2012, www.census.gov/prod/cen2010/doc/sf1.pdf.

XI. ACKNOWLEDGEMENTS

Special thanks to Dr. Wikle for his support and advice throughout the entire thesis writing process. And a big thank you to my friends and families who texted and asked how I was doing when I disappeared for days in the basement of Middlebush. And yes Maddy, I now have time to go see a movie with you.