America’s Warzone: Modeling Armed Robberies in Chicago

Reuben K. McCreanor[*]
reuben.mccreanor@duke.edu

Anna K. Yanchenko[*]
anna.yanchenko@duke.edu

Lei Qian[*]
lei.qian@duke.edu

Megan S. Robertson[*]
megan.robertson@duke.edu

1 Introduction

The City of Chicago is frequently listed as one of the most dangerous and crime-ridden cities in the US. President Donald Trump frequently discusses the high-rate of crime in Chicago. According to the Chicago Tribune, there were 4,367 shooting victims in Chicago in 2016. In the same year there were also 785 homicides.[2] However, other reports conclude that Chicago should not be called the crime capital?? of America, as Chicago’s violence rate is lower than cities like St. Louis and Detroit. [1] The goal of this project was to examine crime in Chicago, specifically armed robberies, from 2012-2016.

2 Data

The crime data used for this project came from the City of Chicago’s website. [1] The data contained every reported crime in Chicago from 2001 to the present (Table 1). In addition to the type of crime reported (battery, assault, etc.), there was information on the location and time of the crime. The data set was reduced to only consider armed robberies.

Case Number	Date	Block	Description	Beat	District	Ward	Community Area	Location
HN180091	02/16/2007 03:20:00 PM	012XX W 103RD ST	OTHER WEAPONS VIOLATION	2232	22	21	73	(41.706819022, -87.654048084)
HN184333	02/15/2007 07:00:00 PM	093XX S WOODLAWN AVE	TELEPHONE THREAT	413	4	8	47	(41.725252492, -87.594860893)
HN182527	02/17/2007 09:00:00 PM	042XX S COTTAGE GROVE AVE	OTHER VIOLATION	213	2	4	38	(41.81741558, -87.606719823)
HN183814	02/18/2007 10:06:37 PM	063XX N SHERIDAN RD	TO PROPERTY	2433	24	49	77	(41.996866019, -87.655592844)
HN182579	02/18/2007 12:20:00 AM	033XX W HURON ST	DOMESTIC BATTERY SIMPLE	1121	11	27	23	(41.893682761, -87.710701702)
HN182986	02/18/2007 10:20:00 AM	042XX N CENTRAL AVE	OVER $500	1624	16	38	15	(41.957385814, -87.767141739)
HN183716	02/18/2007 08:40:00 PM	037XX S MICHIGAN AVE	DOMESTIC BATTERY SIMPLE	211	2	3	35	(41.827167256, -87.623160687)
HN184010	02/19/2007 02:40:00 AM	010XX N LAWNDALE AVE	DOMESTIC BATTERY SIMPLE	1112	11	27	23	(41.89998654, -87.71890157)
HN181071	02/17/2007 02:21:00 AM	051XX S CALUMET AVE	POSS FIREARM/AMMO:NO FOID CARD	232	2	3	40	(41.801609105, -87.617736187)
HN182079	02/17/2007 04:00:00 PM	034XX W FLOURNOY ST	DOMESTIC BATTERY SIMPLE	1133	11	24	27	(41.872709648, -87.711929688)
HN184350	02/19/2007 12:00:00 AM	013XX N WOLCOTT AVE	FROM BUILDING	1424	14	1	24	(41.905889056, -87.674299261)
HN184306	02/15/2007 12:01:00 AM	001XX N PARKSIDE AVE	ILLEGAL USE CASH CARD	1512	15	29	25	(41.88301997, -87.766605275)
HN183370	02/09/2007 10:00:00 AM	024XX W DEVON AVE	SIMPLE	2413	24	50	2	(41.99771689, -87.690448237)
HN183500	02/18/2007 05:10:00 PM	070XX S THROOP ST	DOMESTIC BATTERY SIMPLE	734	7	17	67	(41.7658997, -87.65656296)
HN184055	02/19/2007 05:30:00 AM	005XX W ROSCOE ST	TO RESIDENCE	2331	19	44	6	(41.943338611, -87.64332075)
HN184352	02/19/2007 10:45:00 AM	003XX N MICHIGAN AVE	FROM BUILDING	122	1	42	32	(41.887845852, -87.624560336)
HN181018	02/17/2007 01:31:00 AM	056XX S WABASH AVE	TO LAND	233	2	20	40	(41.792323044, -87.624025834)
HN182857	02/18/2007 07:39:41 AM	014XX W LUNT AVE	FORCIBLE ENTRY	2431	24	49	1	(42.009107852, -87.666843608)
HN177373	02/15/2007 08:44:00 AM	054XX S CORNELL AVE	TO VEHICLE	2132	2	5	41	(41.796263314, -87.585435453)
HN183280	01/14/2007 12:00:00 PM	021XX W BIRCHWOOD AVE	DOMESTIC BATTERY SIMPLE	2424	24	49	1	(42.017948603, -87.683418074)

Table 1: The City of Chicago website provides a data set containing information on crimes committed in the city from 2001 to present day.

[*] Department of Statistical Science, Duke University

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

3 Time Series Analysis

The City of Chicago is divided into regions known as sides (Figure 1), where each side is comprised of several neighborhoods. There is a lot of variation in the population (Figure 2) and the number of armed robberies per capita (Figure 3) for these sides. Additionally, some sides are more residential, while others are more commercial.

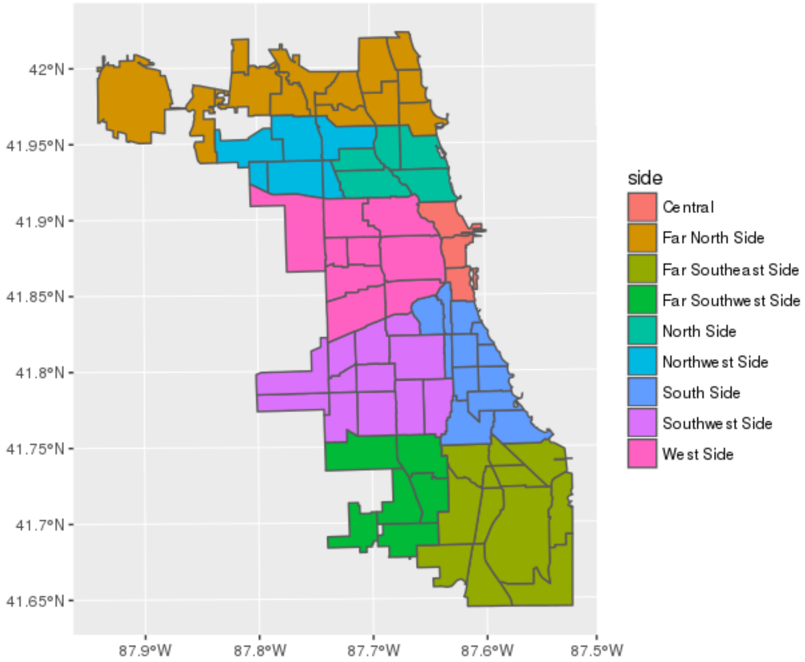


Figure 1: The "sides" of Chicago. The borders correspond to the boundaries of the community areas colored by the side.

ARIMA models were fit to predict the counts of monthly armed robberies in each side of the city between 2003 and 2016. In order to determine the type of model, the ACF and PACF plots were examined for the data for each of the sides. For example, if there was structure in the PACF plot beyond one lag, moving average terms were added. The model residuals were also examined to ensure that there was no remaining structure in the residuals. The PACF and ACF plots for the data from the South Side are displayed below. The ACF plot showed evidence of seasonality at lag 12 (i.e. yearly trends). After lag 1, there was no large values for the PACF, so no moving average terms were included in the model.

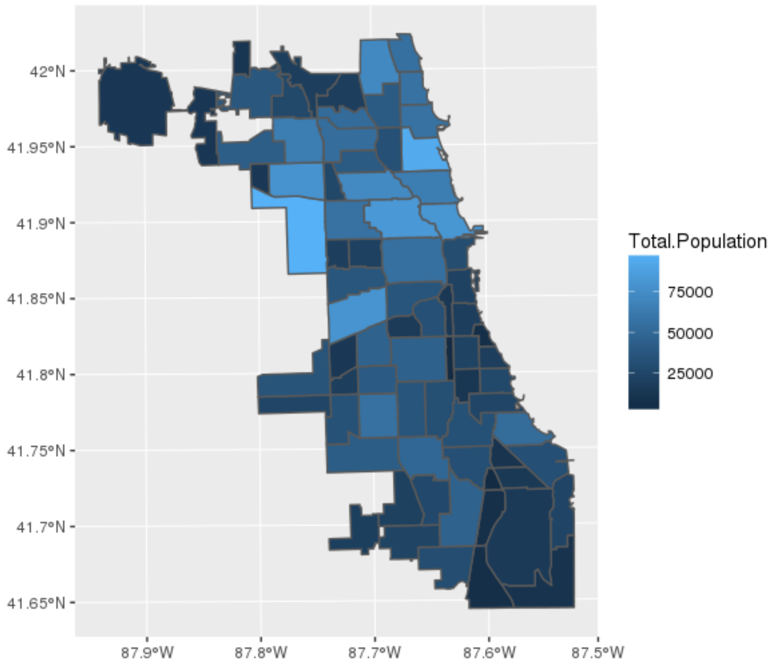


Figure 2: The population distribution of the "sides" of Chicago. The borders correspond to the boundaries of the community areas colored by the side.

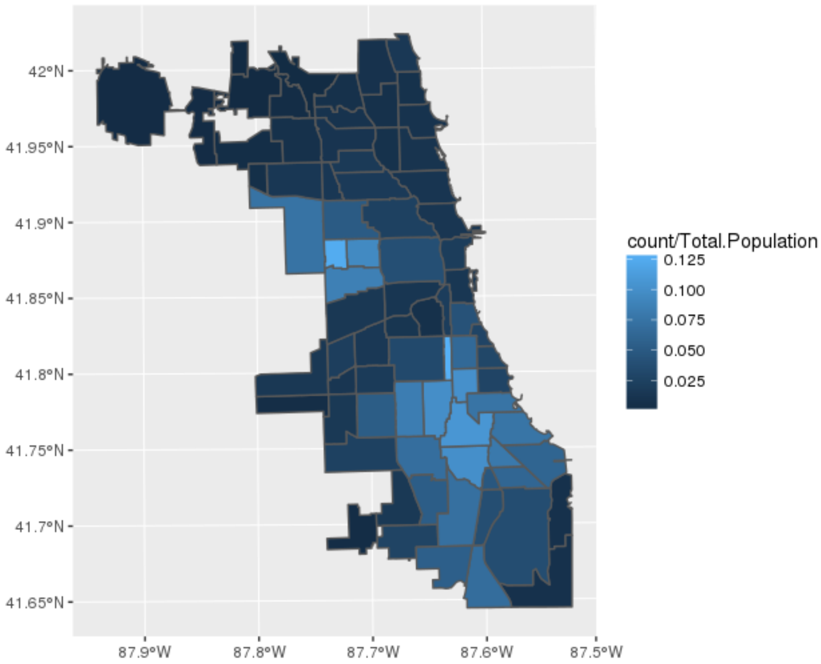
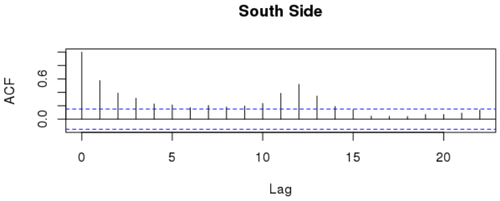


Figure 3: The number of armed robberies per capita for the "sides" of Chicago between 2003 and 2016. The borders correspond to the boundaries of the community areas colored by the side.



	ar1	ar2	ar3	ar4	sar1
Coefficient	0.3850	0.1328	0.0825	0.0167	0.4784
Standard Error	0.0793	0.0832	0.0824	0.0771	0.0708

Table 2: Summary of model fit for the AR(4) with period 12 seasonal component fit to the monthly count of armed robberies for the South Side.

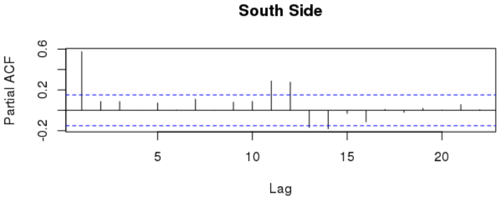


Figure 5: PACF plot for the number of monthly armed robberies in the South Side of Chicago between 2003 and 2016.

Based on the ACF Figure 6 and PACF Figure 3 plots, an AR(4) model was fit with a seasonal component with period twelve. The residuals plot for this model did not display any remaining structure in the data and the coefficients are in **??**.

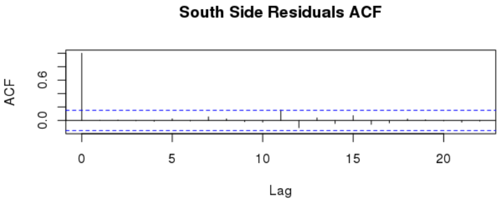


Figure 6: ACF plot of the residuals of an AR(4) model with a period twelve seasonal component fit to the monthly count of armed robberies for the South Side.

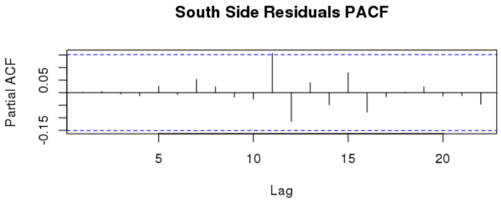


Figure 7: PACF plot of the residuals of an AR(4) model with a period twelve seasonal component fit to the monthly count of armed robberies for the South Side.

The coefficient estimates for all of the different sides were very similar. While some sides displayed evidence of higher order autoregressive structure or the addition of moving average terms, as compared to the South Side, all sides had a clear period 12 seasonal component, indicating strong yearly trends for all sides of the city. The coefficient estimates were positive for the autoregressive terms, indicating that there was a positive correlation between the amount of monthly armed robberies over time. Plots of the various model fits can be found in subsection 5.1. It is interesting that although the sides of Chicago are quite diverse in terms of population and demographics, as well as the number of monthly armed robberies, the temporal trends for all of the sides are very similar. Although the count of the monthly armed robberies differs by side of the city, the overall temporal trend is the same across Chicago and has a strong yearly, autoregressive trend.

4 Spatial Models

4.1 Introduction

The City of Chicago is comprised of 77 distinct community areas. We predict for counts of armed robbery in each area using a Bayesian Spatial Latent Gaussian Process Poisson Regression Model (BSLGPPR) and a 100 explanatory variables provided by Chicago, **??** gathered from a range of years within that of the armed robbery data. We do make the assumption that these variables have not changed much over the next/previous few years and remain applicable. Since we have numerous explanatory variables that range from demographic information to counts of graffiti art, we want to narrow down the number of variables to improve the model's prediction accuracy and interpretability. A popular frequentist method is the penalized LASSO regression; however, it does not take into account spatial information. Our model, the BSLGPPR, will simulate a LASSO regression while also modeling spatial random effects.

As the response variable of our data are counts of armed robbery in community areas, we model our observations $\{y_i\}_1^N$, $N = 77$, with a Poisson distribution.

$$y_i \sim \text{Poisson}(\lambda_i) \tag{1}$$

We know that λ_i must be positive, which is why we let it equal to the exponential of $\mathbf{X}\boldsymbol{\beta}$. $\boldsymbol{\beta}$ are the coefficients for the explanatory variables and the intercept, which comprises the design matrix$\mathbf{X}$.

$$\log(\boldsymbol{\lambda}) = \mathbf{X}\boldsymbol{\beta} \tag{2}$$

Equation 2 is for the LASSO regression, but we will add another variable ω_i to model random spatial effect in the BSLGPPR model.

$$\log(\boldsymbol{\lambda}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\omega} \tag{3}$$

4.2 The LASSO Model

Our first model comprises of a simple penalized LASSO regression. Given the positive value of Moran's I, 0.5118399, this model is highly unlikely to outperform one that takes into account the spatial nature of crime. However, we implement this model as a baseline in order to illustrate why the BSLGPPR model performs better.

In this model, without the random spatial effect, we do make the assumption that the observations are independent of one another.

From Equation 2, it is simple to realize that

$$\boldsymbol{\lambda} = e^{\boldsymbol{\beta}'\mathbf{X}}$$

. Given a set of parameters $\boldsymbol{\beta}$, and explanatory variables $\mathbf{X}$, we observe the counts of armed robbery $\mathbf{Y}$ with probability

$$p(y_1, ..., y_N|\mathbf{x}_1, ..., \mathbf{x}_N, \boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{e^{y_i\boldsymbol{\beta}'\mathbf{x}_i} e^{-e^{\boldsymbol{\beta}'\mathbf{x}_i}}}{y_i!} \tag{4}$$

Equation 4 can be obtained by plugging Equation 2 into the Poisson probability distribution.

We then want to use the maximum likelihood method to find a set of $\boldsymbol{\beta}$ that will maximize the likelihood, Equation 4 which is the same as maximizing the log-likelihood:

$$l(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{N}(y_i(\boldsymbol{\beta}'\mathbf{x}_i) - e^{\boldsymbol{\beta}'x_i}) \tag{5}$$

To implement the penalized LASSO regression, we instead optimize the penalized log-likelihood:

$$\min_{\boldsymbol{\beta}} -\frac{1}{N} l(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) + \lambda\frac{1}{2}\left((1-\alpha)\sum_{i=1}^{N}\beta_i^2) + \alpha\sum_{i=1}^{N}|\beta_i|\right) \tag{6}$$

and set $\alpha = 1$.

This method will allow us to obtain more accurate predictions than regular OLS and perform variable selection to prevent overfitting and for interpretability purposes.

4.3 The Bayesian Spatial Latent Gaussian Process Poisson Regression Model

Our second model uses a double exponential or Laplace prior to emulate the LASSO penalized regression model as the distribution sharply peaks at zero; concentrating the probability mass at zero. While this prior will not cause our coefficients to go to zero as in the case of a LASSO–instead behaving more like Ridge regression–we get around this by constructing 95 percent credible intervals around the coefficients and finding the ones that contain zero.

In our Bayesian model, we use the same data as in subsection 4.2 and supplement in spatial data in terms of spatial polygons for the 77 areas.

For our model, we take Equation 1 and Equation 3 and further specify by setting:

$$\begin{aligned}
\beta_j &\overset{iid}{\sim} \text{Laplace}(0, \eta)\\
\boldsymbol{\omega} &\sim \text{MVN}(\mathbf{0}, \tau(D - \phi W))\\
\tau &\sim \text{Gamma}(2, 2)\\
\phi &\sim \text{Unif}(0, 0.99)\\
\eta &\sim \text{Unif}(0.001, 10)\\
\text{where } \{D : d_{\{}jj\} &= \text{ total number of neighboring community areas for community area j}|j \in 1...77\}\\
\text{where } \{W : w_{\{}jk\} &= \text{whether community area j shares boundaries with community area k}|j, k \in 1...77\}
\end{aligned}$$

Like before, we say that the armed robbery data can be modeled using a Poisson distribution with λ_i. We know that λ_i must be positive, which is why we let it equal to the exponential of $\mathbf{X}\boldsymbol{\beta} + \omega_i$. $\boldsymbol{\beta}$ are the coefficients for each explanatory variable $\mathbf{X}$ and the intercept. We model the random spatial effect by putting a multivariate gaussian prior on ω_i and setting the mean to zero,we expect the average effect to be 0, and a correlation matrix $\tau(D - \phi W))$, we believe community areas to be affected by neighboring community areas and the total number of neighboring community areas, which is then scaled by τ and ϕ. We set τ to have a Gamma(2,2) prior because we believe τ is positive and also heavily concentrated between 0 and 4. ϕ has a Unif(0, 0.99) prior to allow for equal probability for any value within the specified range; as we are not biased towards any weight for W. The $\boldsymbol{\beta}$s are iid Laplace to emulate a penalized LASSO regression. We set η to have a Unif(0.001, 10) distribution because we believe that is how concentrated the $\boldsymbol{\beta}$'s will be around the mean, 0; the smaller the η the more heavily concentrated at the mean.

4.4 Results

4 Spatial Models Intro To Data Spatially - Different Community Areas - Added Variables (100) - What We Want to Predict - Why Spatial: Moran?s I - How Many Data Points - Why Armed Robbery LASSO Model Bayesian Spatial Model Results + Conclusions - Talk about Austin - Cancer - Residuals Difference - Demographics - Top Ten Coeffs - Add Betas in Appendix - Add Reference for Data Sources - Add Reference for Data Sources

References

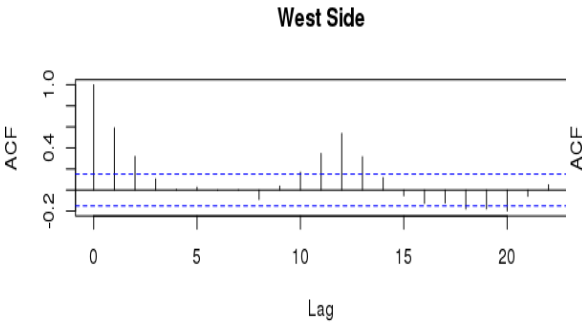
[1] Papachristos, Andrew V., ”48 Years of Crime in Chicago: An Analysis of of Serious Crime Trends from 1965-2013,http://isps.yale.edu/sites/default/files/publication/2013/12/48yearsofcrime_final_ispsworkingpaper023.pdf, December 2013.

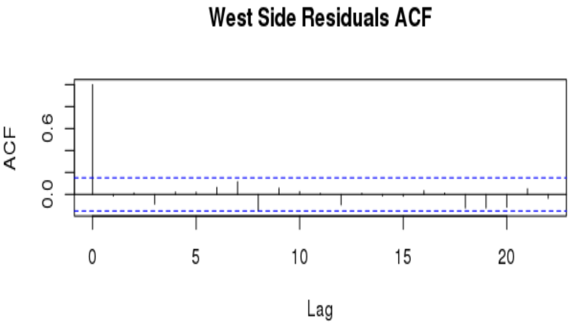
[2] Pearson, Rick, ”Trump Again Assails Chicago gun violence in speech to Congress”, *Chicago Tribune*, http://www.chicagotribune.com/news/local/politics/ct-donald-trump-congress-speech-chicago-met-20170228-story.html, March 2017.

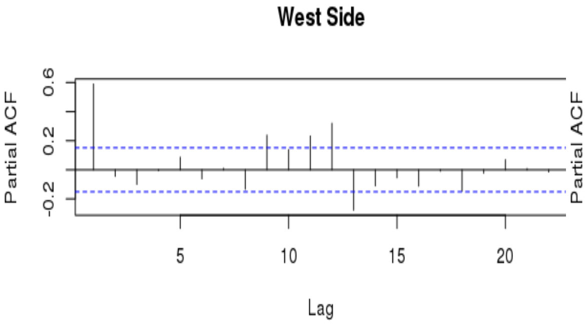
5 Appendix

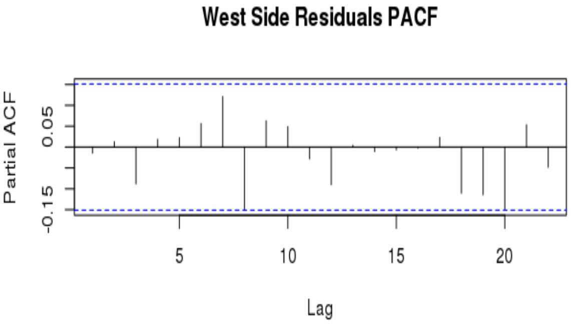
5.1 Time Series Modeling Plots

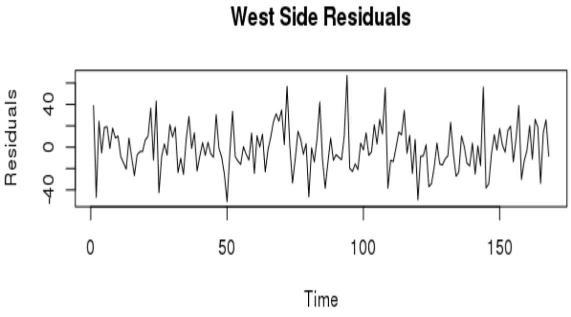
5.1.1 West Side



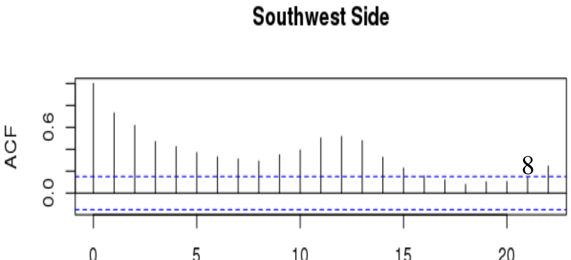


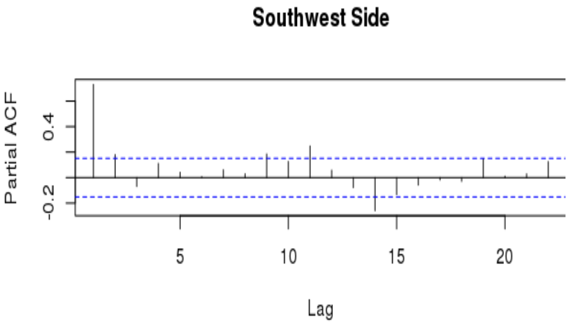


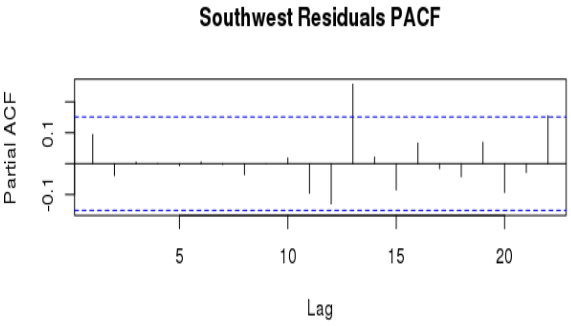


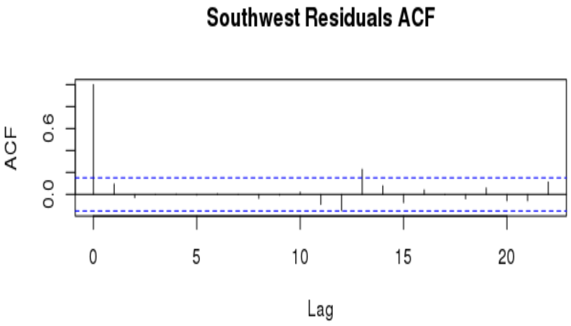


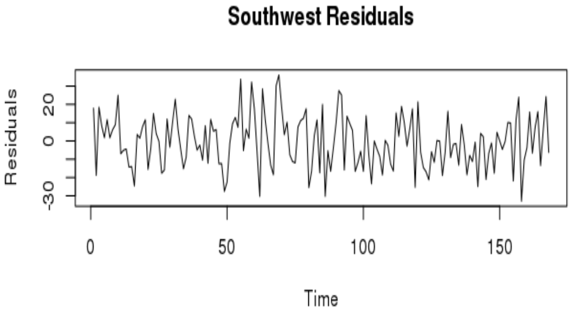
5.1.2 Southwest Side



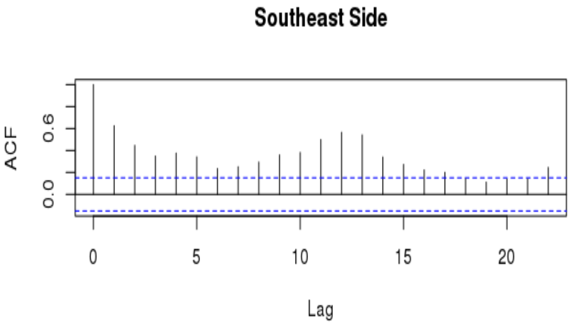


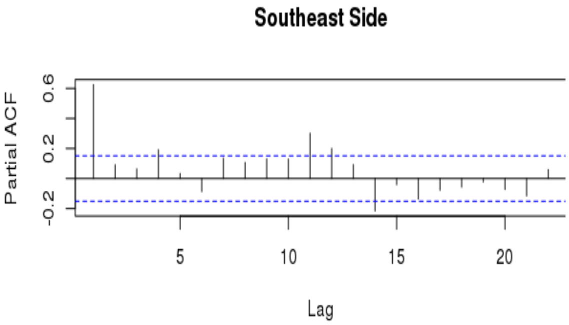


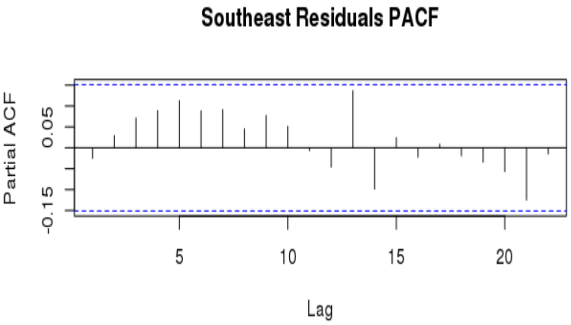


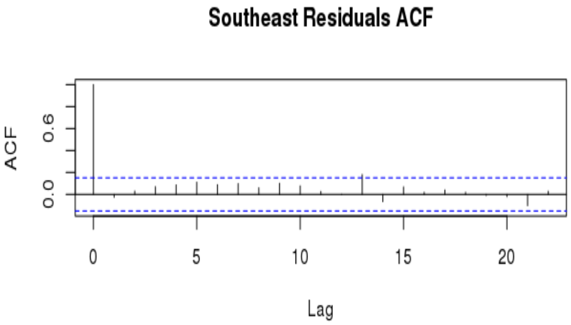


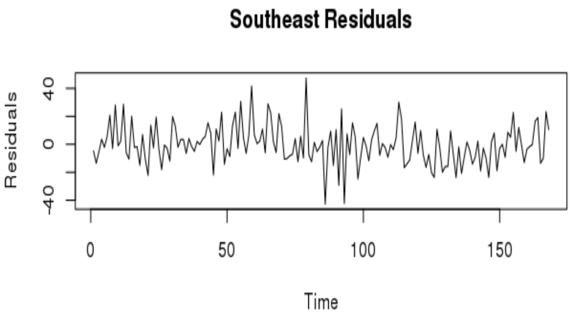
5.1.3 Southeast Side



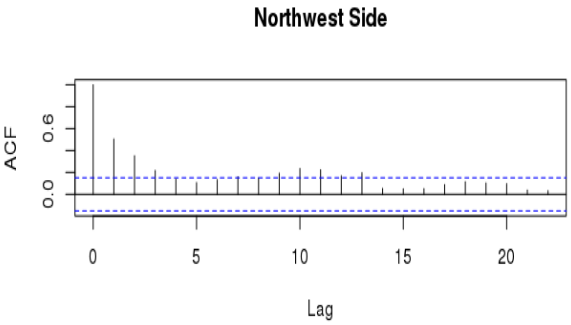

Southeast Side

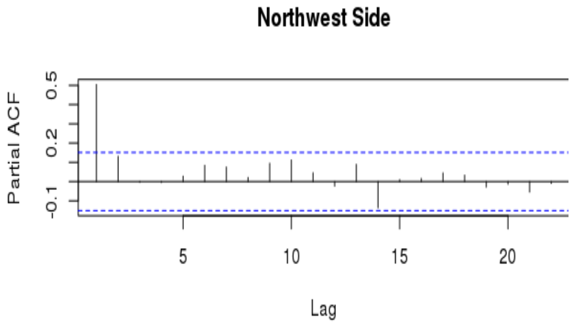

Southeast Residuals PACF

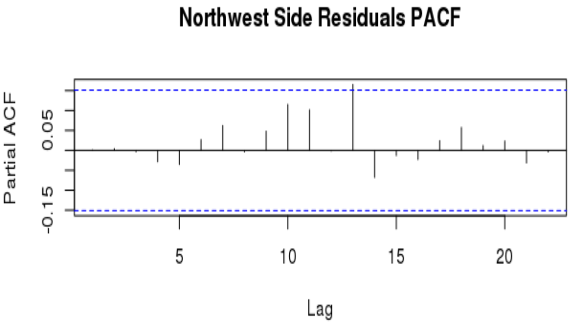

Southeast Residuals ACF

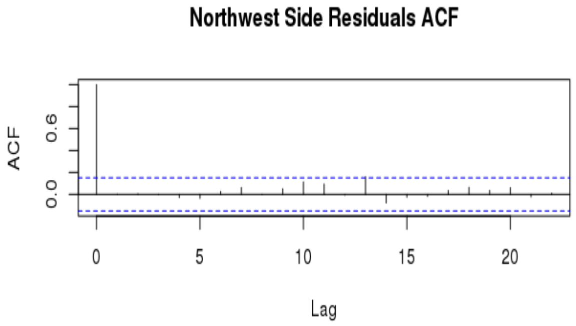

Southeast Residuals

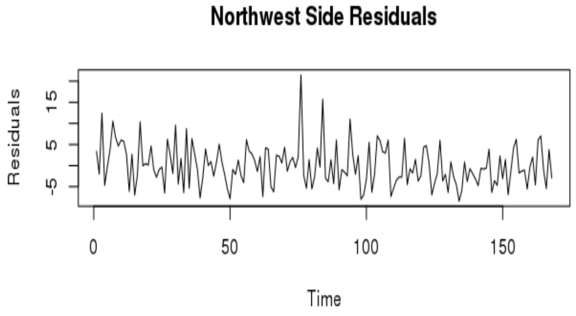
5.1.4 Northwest Side


Northwest Side

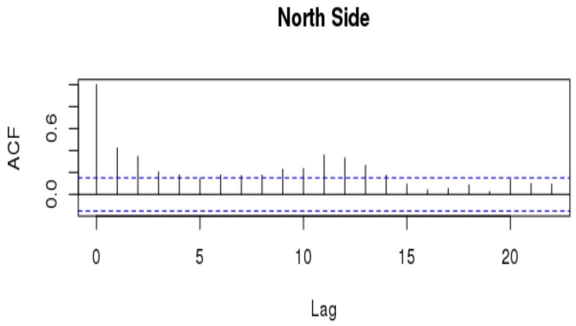


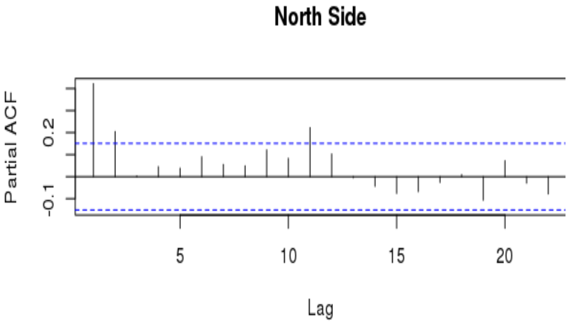


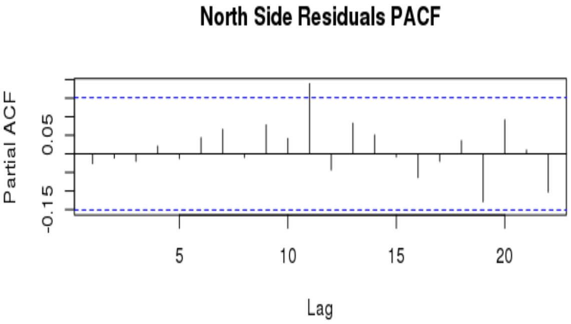


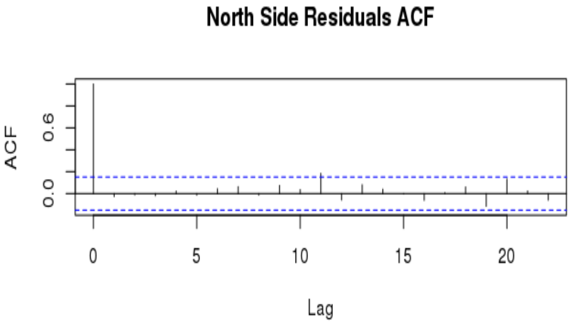


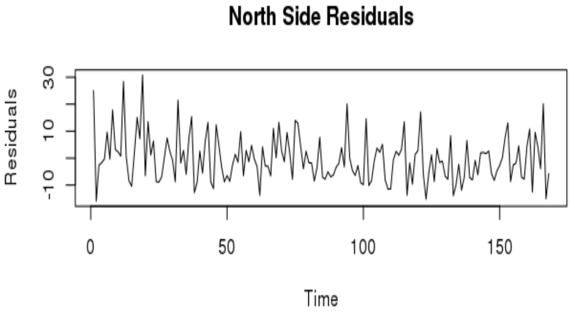
5.1.5 North Side



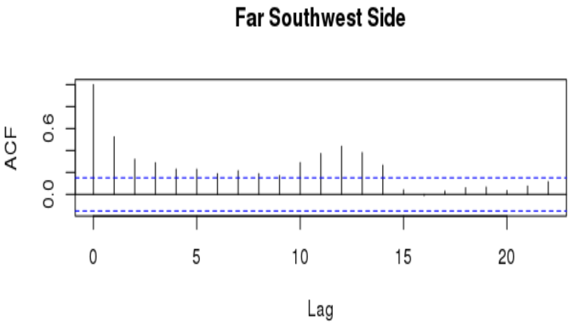


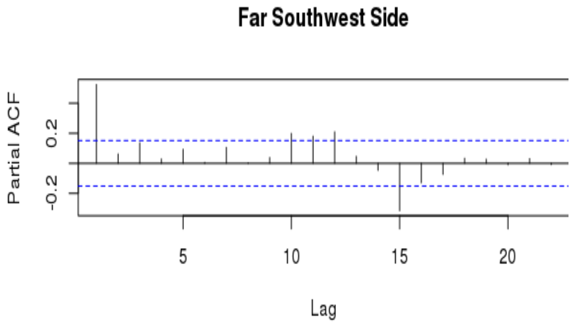


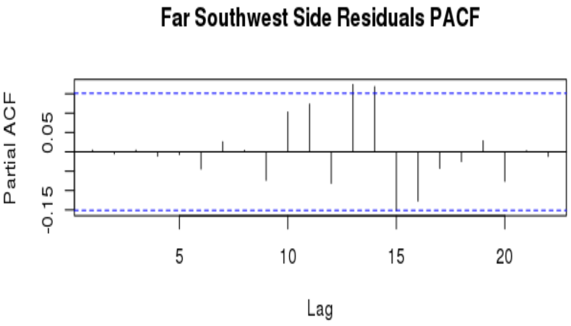


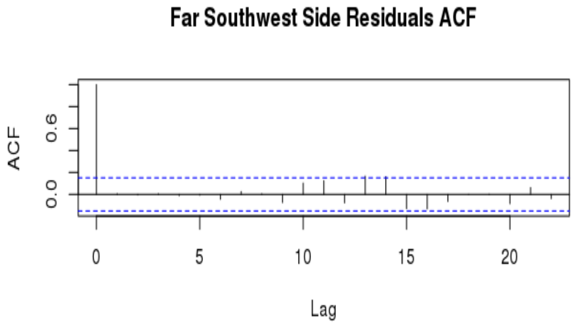


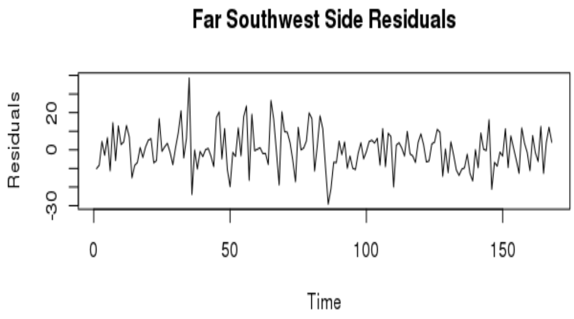
5.1.6 Far Southwest Side



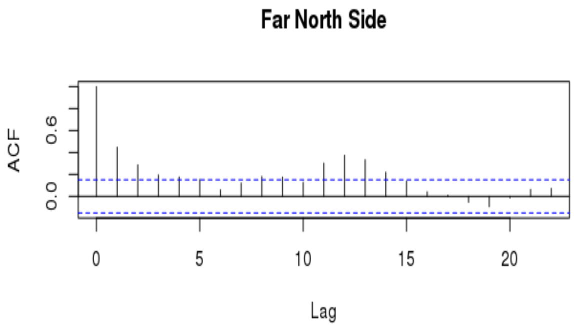


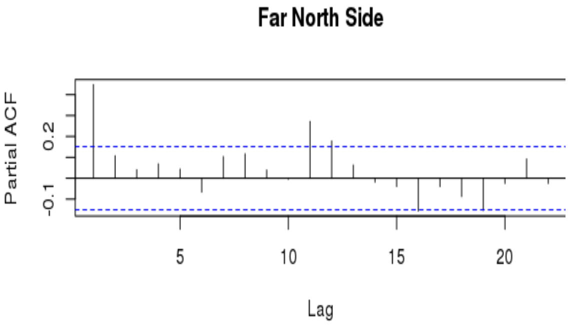


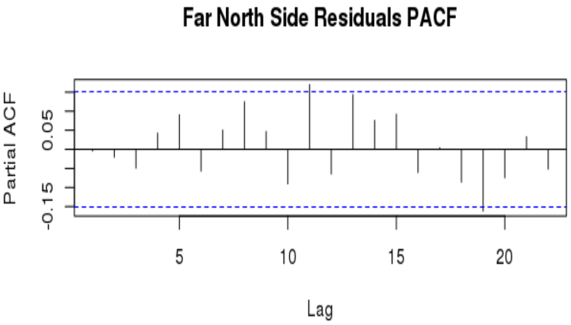


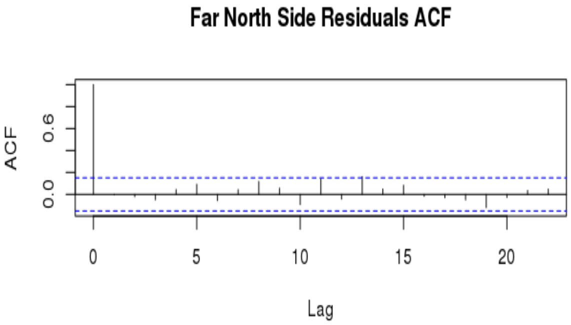


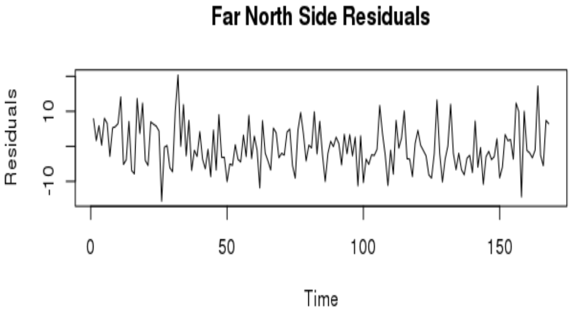
5.1.7 Far North Side



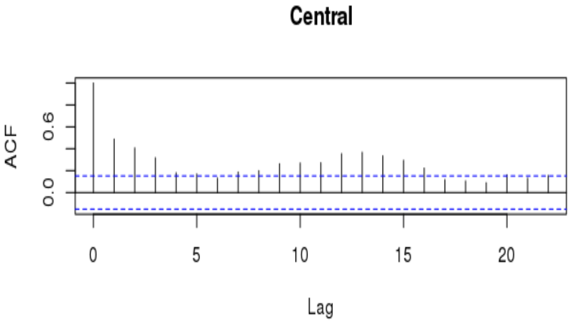

Far North Side
Partial ACF
0.2
-0.1
5
10
15
20
Lag

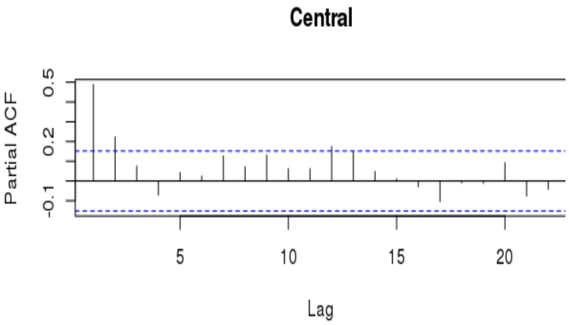

Far North Side Residuals PACF
Partial ACF
0.05
-0.15
5
10
15
20
Lag

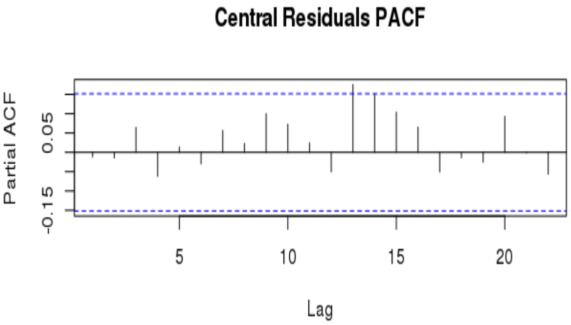

Far North Side Residuals ACF
ACF
0.6
0.0
0
5
10
15
20
Lag

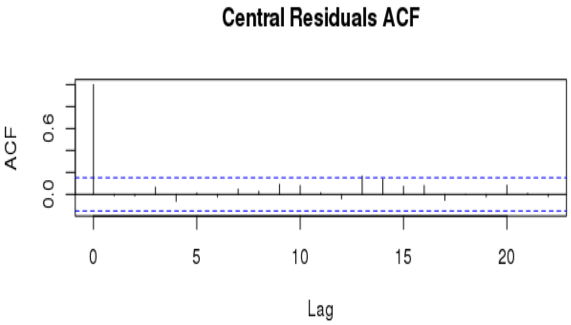

Far North Side Residuals
Residuals
10
-10
0
50
100
150
Time

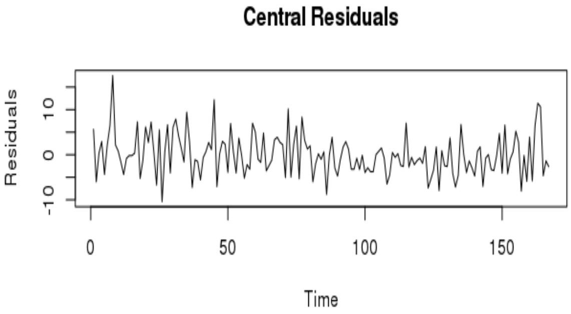
5.1.8 Central


Central
ACF
0.6
0.0
0
5
10
15
20
Lag


Central
Partial ACF
0.5
0.2
-0.1
5
10
15
20
Lag


Central Residuals PACF
Partial ACF
0.05
0.0
-0.15
5
10
15
20
Lag


Central Residuals ACF
ACF
0.6
0.0
0
5
10
15
20
Lag


Central Residuals
Residuals
10
0
-10
0
50
100
150
Time

5.2 Spatial Analysis

Variable	Lasso	2.5%	97.5%	Spatial	IncludeZero
Intercept	6.55	6.21	6.27	6.24	FALSE
KWH Total SQFT	0.00	-0.17	0.42	0.07	TRUE
THERMS Total SQFT	0.00	-0.29	0.38	0.07	TRUE
N_Graffiti	0.13	0.10	0.34	0.19	FALSE
Birth Rate	0.00	0.06	0.40	0.21	FALSE
General Fertility Rate	0.00	-0.39	0.02	-0.14	TRUE
Low Birth Weight	0.09	0.01	0.25	0.10	FALSE
Prenatal Care Beginning in First Trimester	-0.16	-0.33	-0.09	-0.21	FALSE
Preterm Births	0.00	-0.17	0.05	-0.06	TRUE
Teen Birth Rate	0.00	-0.15	0.12	-0.02	TRUE
Assault (Homicide)	0.22	-0.12	0.31	0.09	TRUE
Breast cancer in females	0.14	0.07	0.22	0.14	FALSE
Cancer (All Sites)	0.00	0.16	0.50	0.29	FALSE
Colorectal Cancer	0.01	-0.03	0.19	0.08	TRUE
Diabetes-related	0.00	-0.14	0.07	-0.04	TRUE
Firearm-related	0.04	-0.09	0.11	0.03	TRUE
Infant Mortality Rate	0.20	0.13	0.33	0.23	FALSE
Lung Cancer	0.00	-0.10	0.14	0.00	TRUE
Prostate Cancer in Males	0.02	-0.08	0.14	0.02	TRUE
Stroke (Cerebrovascular Disease)	-0.04	-0.14	0.06	-0.04	TRUE
Tuberculosis	0.08	0.01	0.22	0.13	FALSE
Below Poverty Level	0.00	-0.06	0.27	0.08	TRUE
Crowded Housing	-0.05	-0.35	0.00	-0.19	TRUE
Dependency	0.00	-0.05	0.23	0.07	TRUE
No High School Diploma	0.00	-0.16	0.30	0.08	TRUE
Per Capita Income	0.00	-0.28	0.24	-0.01	TRUE
Unemployment	-0.21	-0.82	-0.39	-0.58	FALSE
N_Aff_Housing	0.00	-0.09	0.02	-0.03	TRUE
PERCENT OF HOUSING CROWDED	0.00	-0.18	0.09	-0.03	TRUE
PERCENT HOUSEHOLDS BELOW POVERTY	0.00	-0.20	0.24	-0.00	TRUE
PERCENT AGED 16+ UNEMPLOYED	0.08	0.17	0.52	0.34	FALSE
PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	0.00	-0.21	0.43	0.15	TRUE
PERCENT AGED UNDER 18 OR OVER 64	0.00	-0.22	0.22	-0.05	TRUE
PER CAPITA INCOME	0.00	-0.16	0.19	0.03	TRUE
HARDSHIP INDEX	0.00	-0.13	0.63	0.18	TRUE
vacantLots	-0.01	-0.02	0.09	0.05	TRUE
Total Population	0.00	-0.07	0.35	0.11	TRUE
Not Hispanic or Latino, White alone	0.00	-0.33	0.20	-0.07	TRUE
Not Hispanic or Latino, Black or African American alone	0.28	-0.12	0.18	0.06	TRUE
Not Hispanic or Latino, American Indian and Alaska Native alone	0.00	-0.15	0.22	0.08	TRUE
Not Hispanic or Latino, Asian alone	-0.00	-0.25	-0.02	-0.14	FALSE
Not Hispanic or Latino, Native Hawaiian and Other Pacific Islander alone	0.00	-0.05	0.09	0.01	TRUE
Not Hispanic or Latino, Some Other Race alone	0.00	-0.11	0.33	0.13	TRUE
Not Hispanic or Latino, Two or More Races	0.00	-0.32	0.33	0.02	TRUE
Hispanic or Latino	0.11	-0.21	0.18	0.02	TRUE
Male: Under 5 years old	0.00	-0.19	0.15	-0.02	TRUE
Male: 5 to 9 years	0.00	-0.26	0.07	-0.09	TRUE
Male: 10 to 14 years	0.00	0.08	0.55	0.29	FALSE
Male: 15 to 17 years	0.00	-0.31	0.22	0.02	TRUE
Male: 18 and 19 years	0.00	-0.03	0.78	0.34	TRUE
Male: 20 years	0.00	-0.00	0.36	0.15	TRUE
Male: 21 years	0.00	-0.37	0.13	-0.08	TRUE
Male: 22 to 24 years	0.00	-0.15	0.37	0.07	TRUE