Crime in Chicago 2001-2017

Ryan Rouleau
University of Colorado
Department of Computer Science
Boulder, CO

Kylee Budai
University of Colorado
Department of Applied Mathematics
Boulder, CO

ABSTRACT

This paper aims to predict the average severity of crimes at different locations in Chicago. The purpose of doing this is to show what areas, if any, are prone to have more severe crimes. It is common knowledge that criminal activity varies with the seasons – there is generally more crime in the summer and less in the winter. By the end of this paper, the reader will have a clear understanding of the behavior of crime in Chicago over the past 15 years and how we used that behavior to generate severity heat maps that accurately capture the general spatial trend of the data.

PROBLEM STATEMENT AND MOTIVATION

Chicago is a hot spot for unlawful acts and we aim to look at the changes in the number of crimes and the average severity of crimes on different scales. We will begin by looking at crime as a whole in Chicago and then we will separate Chicago into many different regions and consider crimes in those regions. We hope to be able to predict both the number of crimes in Chicago (and around different points in Chicago) and the average severity of crimes throughout the entirety of Chicago. In order to make these predictions, we must answer the question "How have the number of crimes and their severities been changing over time throughout Chicago?". We plan on answering this question by ranking crimes based on severity, binning those severities into three bins, and looking at data local to different centers at different scales. We will develop temporal fits for these data and will perform prediction by extrapolating these fits. Once we achieve decent temporal prediction, we will be able to extend our model to more spatial locations and develop predictive crime severity heat maps by using ordinary kriging.

PREVIOUS WORK

Crime data in Chicago has been tracked by the Chicago Police Department since 2001. From this data, the FBI and other governmental agencies have identified where and when crime takes place to better allocate their resources. In addition to government studies, third-parties have also used this data to discover other trends.

Last year, the government conducted a study focusing on which types of demographics they could expect to either unlawfully discharge a weapon into someone, or be shot themselves. They figured that although there are a large number of shootings in Chicago, the majority of them are caused by a small group of people. The idea behind this study was to predict who is most likely going to shoot someone or be shot themselves, which would help the authorities know which areas/types of people needed more protection. They were able to make predictions that were about 70% accurate using a model that incorporated where the person lived, if they had been to jail before, and if they had been shot before. This study was conducted using an extended version of the publicly available data set which included ten additional attributes. This study was covered by the New York Times in an article titled *Chicago Police Try to Predict Who May Shoot or Be Shot* [1].

In 2015, a study was done that considered the effects of weather on crime. It was interested in predicting whether crime would be higher or lower than a yearly average given different weather statistics. This study, titled *Predicting Crimes in Chicago by Weather* [2], was performed by a student at Northwestern, Alan Fu, who was curious to see if the weather had a significant enough effect on people's moods to affect their criminal behaviors. He used a weather database which included parameters such as temperature, humidity, wind speed, etc. Using decision trees, he discovered that the weather attribute which contributed most significantly to crime was temperature which is not entirely surprising.

3 years ago a study was conducted to determine if Chicago's sports team's games affected crime in the city [3]. The study found that generally for games involving the city's teams, crime during the matches dropped by roughly 15%. This drop increased to 25% for more important contests such as the Super Bowl. The study deduced these results by comparing the amount of crime during NFL, NBA, and MLB games to corresponding times with the same day of week, and same month as the original game where there were no matches being played. The study also binned the crimes into general areas to identify if certain crimes were more or less likely to occur during these games. The study found that there was generally an identical drop in all types. The two researchers obtained their data from sports-reference.com (for the dates times of the games) and the identical data set we are using for the crime data.

Most of the past work we've seen looks at the type of crime and correlates it to location. What makes our approach unique is that it looks at the severity of crime (e.g. murder being worse than gambling) by area and tries to predict the change of severity in the future. This could allow law enforcement to plan for the future and better address community issues before they occur.

DATA SET

This project looks at the publicly available Chicago crime data set that contains ∼6.2 million rows and has 22 attributes [4]. It includes all of the crime data in Chicago from 2001 to the present and is updated each week, giving us the ability to test our temporal predictions against real data. This data includes fields such as when the crime occurred, what type of crime it was, the location of the crime, a description of the location, the IUCR code of the crime (which maps to a more specific type of crime), and whether there was an arrest or not.

Since it is more interesting to consider severity of crime as opposed to type of crime, we must map crimes to severity levels. In order to do so, we will use the IUCR (Illinois Uniform Crime Reporting) code to identify crimes. There is an easily accessible data set that explains each of the codes [5]. It helped us to narrow down severity into more categories than using the the attribute "Primary Type" would have. For a clear distinction between primary type and IUCR code, an example of primary type is homicide whereas the IUCR codes break that primary type of homicide into first, second, third degree homicide and manslaughter.

For this project, we will consider 5 of the data's attributes: Location (Latitude/Longitude), IUCR (which specifies type of crime), Date(Month and Year), and Arrest. An example of a few lines of our condensed database is given below.

Month	Year	IUCR	Lat	Long	Arrest
08	2008	1330	41.896	-87.630	1
08	2008	1320	41.699	-87.618	0
08	2008	0486	41.763	-87.615	0

Merging this data file with both a binning and a severities file added two more attributes to our database. The severity provides a numerical value representing the severity of a particular crime while the binning provides a separation between petty, medium, and severe crimes which we labeled as bin 1, 2, and 3 respectively.

TOOLS

The tools used for this project are

- Python
- R
- D3.js
- git
- bash
- AWK
- sed

We planned to use a MySQL database to hold and manipulate our data, but we found working with plain CSV's and writing scripts worked better for us. We also used git for version control, Python for the majority of the scripting, R for data visualization and analysis, and bash for running our aforementioned Python scripts automatically. We used d3 for less trivial graphs, visualizations, and animations.

PREPROCESSING AND SEVERITY MAPPING

In order to perform a sentiment or severity analysis on the data, we needed to consider how to be the least subjective with our data. In order to reduce subjectivity, we removed all crimes with the primary type "OTHER OFFENSE".

We used a severity ranking for crimes that was developed in Canada in 2015 [6] as a loose structure. Using this structure, we were able to identify a mapping for a large amount of the IUCR codes. With a little more work, we added in the rest of the crimes, fitting them as best as we could, using data about maximum and minimum prison sentences and fines in order to not be too subjective [7].

After a preliminary plotting of the data, we found that the heat map did not show any spatial significance which was surprising. Because of this, we rethought the idea of using a severity map by adding in a binning. We put petty crimes in bin 1, non-aggressive crimes in bin 2, and aggressive crimes in bin 3. Petty crimes ranged from gambling to criminal abortion, non-aggressive crimes from concealed carry license violation to lower level sexual assault, and aggressive crimes from high level assault to homicide. We performed all of the analysis on the binned data.

Additionally, we decided that there being an arrest would indicate that a crime was more severe than the same crime without an arrest. For this reason, we decided to separate the data into two parts: crimes with arrests and crimes without arrests. The table below shows the counts of arrest versus no arrest data throughout each of the bins.

	Bin 1	Bin 2	Bin 3
Arrest	816441	243560	327850
No Arrest	1524584	1381745	1168094

This table yields a χ^2 test statistic value of 213500 with two degrees of freedom and a p-value less than 2.2×10^{-16}. This says that there is significance in separating the data between arrest and no arrest and gave us more of a reason to separate our data.

DATA ANALYSIS

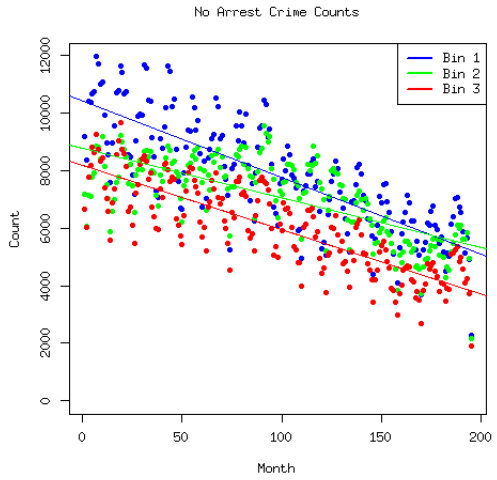
Here is a brief breakdown of the material in each of the subsections.

- *Total No Arrest Crime Counts*: This section considers different models for fitting the counts per bin of the no arrest data It includes model fits in addition to the predictions of counts of crimes in the following months.
- *Total Arrest Crime Counts*: This section fits the same models as used for the no arrest crimes and uses those fits to generate predictions.
- *Localized Crime Counts*: This section separates Chicago into different regions by randomly selecting points across the Chicago area. It considers crime counts in neighborhoods to those points and predicts future crimes surrounding those points.
- *Localized Severities*: This section considers the average severity of crimes at the same points as the previous section and fits models to perform prediction.
- *Heat Maps*: This section extends the idea from the previous section and uses ordinary kriging as a method to develop severity heat maps across Chicago.

Since the goal is to develop a generalized model that is easy to extend, the same analysis will be done on all bins, arrest and no arrest.

TOTAL NO ARREST CRIME COUNTS

We began by developing linear fits for the number of crimes in Chicago over time, ignoring location. We did this for each of the three bins and separated them based on whether there was an arrest or not. This section will consider only no arrest data. Below is a plot of the counts of crimes per bin without an arrest, fit with a simple linear regression.



Clearly, the counts on average have been decreasing per month per bin. Using a linear regression and minimizing the least squared errors in each of the bins yields a model

$$Y = \alpha + \beta X$$

for each bin where Y is the count and X is the month. Recall that the first month, January 2001, was mapped to zero while the last month, December 2016, was mapped to 192. Both coefficients in addition to the standard error of the coefficients are given in the table below for each of the three bins.

Bin	α	$\hat{\sigma}_\alpha$	β	$\hat{\sigma}_\beta$
1	10418.4	169.6	-26.53	1.50
2	8783.9	140.7	-17.32	1.24
3	8167.9	127.7	-22.22	1.13

With corresponding R^2 values

Bin	R^2
1	0.6182
2	0.501
3	0.6672

which says that the models for bin 1, 2, and 3 explain 61.82%, 50.1%, and 66.72% of the variability in the data respectively.

Based solely on the linear regression, predictions for the next few months are given in the table below

Bin	January 2017	February 2017	March 2017
1	5362	5335	5309
2	5509	5493	5476
3	3916	3894	3872

with corresponding confidence intervals given by

Bin	January 2017	February 2017	March 2017
1	(5028,5696)	(4999,5672)	(4971,5649)
2	(5233,5783)	(5217,5768)	(5199,5754)
3	(3664,4169)	(3640,4149)	(3616,4129)

Since it is common knowledge that there is more crime in the warmer months, we decided to bin the residuals based on month and fit a smoothing spline to those data to see if that would improve prediction. Let the residuals per month be denoted by ε and the months be denoted by M (note that $1 \leq M \leq 12$). A smoothing spline is a model given by

$$\varepsilon = \mathrm{f}(M) + \text{error}$$

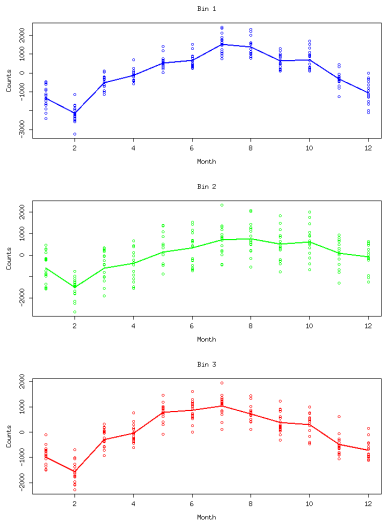
where $\mathrm{f}(M)$ is the function that minimizes

$$\sum_{i=1}^{n} (\varepsilon_i - \mathrm{f}(M_i))^2 + \lambda \int_1^{12} \frac{d^2\mathrm{f}}{dM^2} dM$$

where λ is a set penalty parameter that decides how much weight to put on curvature. A smoothing spline favors models with less total curvature and it can be shown that the optimal smoothing spline is a natural cubic spline with knots at each of the data locations.

The no arrest crime counts in addition to the smoothing spline fits are shown in the plot below, separated for each bin.

No Arrest Fitted Residuals



which results in predictions of residuals given by the table below

Month	$\hat{\varepsilon}_1$	$\hat{\varepsilon}_2$	$\hat{\varepsilon}_3$
January	-1344.24	-607.02	-908.31
February	-2128.38	-1504.21	-1563.45
March	-531.86	-603.46	-299.48
April	-120.55	-373.28	-49.42
May	542.83	146.04	781.68
June	677.07	337.59	867.08
July	1535.10	715.29	1044.49
August	1386.14	766.43	712.15
September	629.88	506.89	389.01
October	712.35	607.54	301.15
November	-311.81	80.48	-486.39
December	-1046.52	-72.30	716.53

where $\hat{\varepsilon}_i$ is the predicted residual for bin i in the given month.

Adjusting the previous predictions by the estimated residuals results in predictions given by

Bin	January 2017	February 2017	March 2017
1	4018	4728	4401
2	3381	3989	3913
3	3384	3291	3573

In order to determine whether considering the residuals was beneficial to prediction of total crime counts, we generated the following plot that shows the predicted counts based solely on linear regression, the counts adjusted by residuals, and the actual values. The vertical lines in these plots represent the 95% confidence intervals from the regression.

Looking at these plots, it is clear that prediction using linear models without factoring in estimated residuals, is not terrible. In fact, prediction is made much worse when factoring in residuals. Despite the fact that we could have used a linear model, we wanted to be able to incorporate the monthly changes. For this reason, we decided to rethink the model.

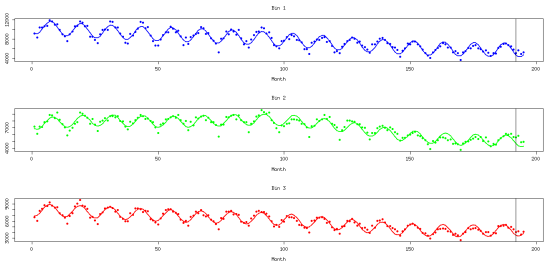
Instead of using a basic linear model, we decided to fit the data with a natural cubic spline and add in variability that corresponds to months. A natural cubic spline is a cubic polynomial between knots that is linear outside of user-defined knots. We choose to use this model so that we could extrapolate the fit in order to make time predictions with confidence intervals. The model used to fit the data is given by

$$Y = \beta_0 + \beta_1 \cos\left(\frac{2\pi X}{12}\right) + \beta_2 \sin\left(\frac{2\pi X}{12}\right) + f(X)$$

where Y is the count of crime, X is the month, $f(X)$ is a natural cubic spline with knots at 25, 55, 100, and 150 and the added sines and cosines account for seasonal variation.

The data fitted with these models is shown below.

Non-Parametric Fits

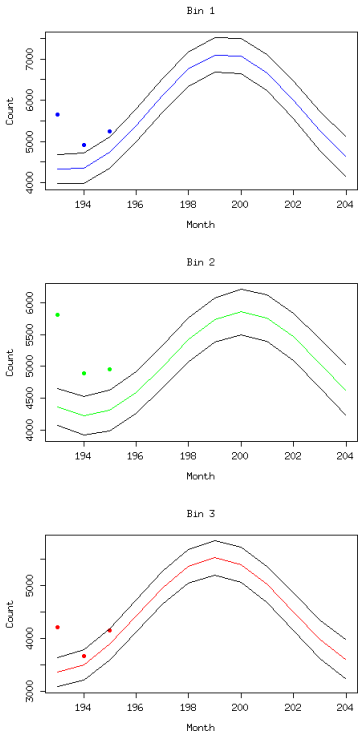


where the vertical black lines show where the model stops and the prediction begins. The R^2 values for these models are significantly improved from the linear models. Each of these models explains about 90% of the variability in the data whereas the linear models only explained about 60%. The R^2 values are given in the table below.

	Bin 1	Bin 2	Bin 3
R^2	0.9157	0.8882	0.9241

Predictions for the next entire year using these models along with 95% confidence intervals are shown in the plot below. Although the predictions are still not great, they are significantly better than the previous predictions (with the added estimated residuals).

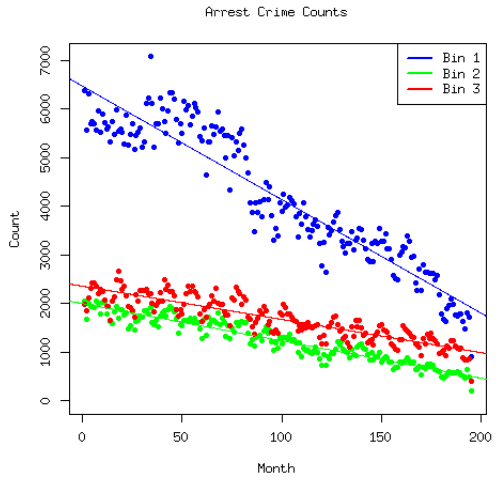
2017 No Arrest Crime Count Predictions



In these plots, the solid colored lines represent the extrapolated fits (or predictions), the black lines are the 95% confidence intervals, and the colored points are the actual values for January, February, and March. Although the actual values did not often fall within the 95% predictive intervals, notice how the fit correctly captures the trend of the data. It predicts a rise in crime where there is a rise in crime and we expect that it will predict the trend of the data this year rather accurately, although it may not be able to correctly predict the values.

TOTAL ARREST CRIME COUNTS

To verify that the second model is the better of the two, we will perform the same analysis and prediction for the arrest crimes. Below is a plot of binned crimes where there was an arrest in addition to least squared linear regression fits.



The corresponding coefficients are given by

Bin	α	$\hat{\sigma}_\alpha$	β	$\hat{\sigma}_\beta$
1	6473.6	69.00	-23.3	0.61
2	1995.7	19.46	-7.62	0.17
3	23.52.2	29.7	-6.84	0.26

and R^2 values given by

Bin	R^2
1	0.8833
2	0.9102
3	0.7791

Looking at the diagnostics plots in addition to the R^2 values, these three models explain much more of the variation in the data than the models in the no arrest case (which may be the case just because there is clearly less variation in these data).

Based on the linear regression, predictions for the next couple months are given by

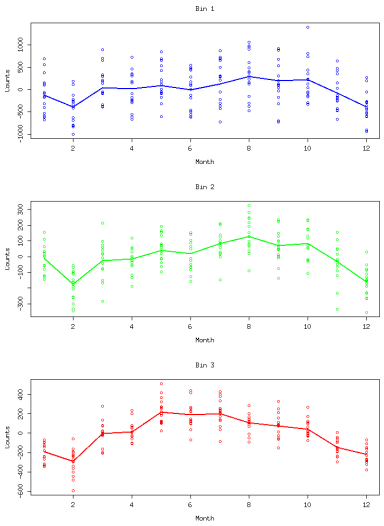
Bin	January 2017	February 2017	March 2017
1	2000	1977	1954
2	530	522	514
3	1051	1045	1038

with 95% confidence intervals given by

Bin	January 2017	February 2017	March 2017
1	(1864,2136)	(1839,2114)	(1815,2092)
2	(492,568)	(483,561)	(475,554)
3	(994,1109)	(987,1103)	(980,1097)

In order to hopefully improve prediction, we generated smoothing splines from the residuals as was done with the no arrest data. Doing this yields the following residual plots

Arrest Fitted Residuals



in addition to the predicted residuals

Month	$\hat{\varepsilon}_1$	$\hat{\varepsilon}_2$	$\hat{\varepsilon}_3$
January	-127.81	-11.47	-188.72
February	-393.01	-174.50	-290.98
March	36.29	-25.65	-3.64
April	16.16	-16.04	13.09
May	87.30	39.83	219.31
June	-2.93	18.83	190.51
July	130.29	82.51	201.60
August	296.89	128.61	106.88
September	201.40	69.66	74.43
October	220.33	84.06	41.77
November	-74.96	-32.79	-148.42
December	-389.95	-163.04	-215.82

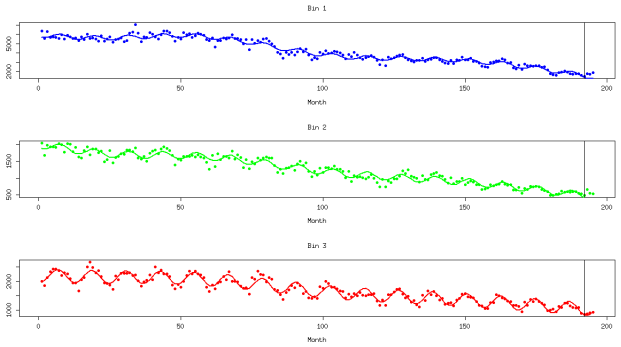
where ε_i corresponds to bin i, just as it did before. Adjusting the previous predictions by the estimated residuals yields the updated predictions given in the table below.

Bin	January 2017	February 2017	March 2017
1	1873	1584	1990
2	519	348	489
3	862	754	1035

Just as was the case for the no arrest case, this method of prediction is okay but not great. However, the prediction using a linear model is better for the arrest data, mainly because of the lower counts and smaller variability in the data. Although this was expected, we decided that providing the plots would be useful to the reader.

Hoping to improve prediction, we used the same natural cubic spline model on the arrest data. Here is the data with non-parametric fits that capture seasonal variability.

Non-Parametric Fits

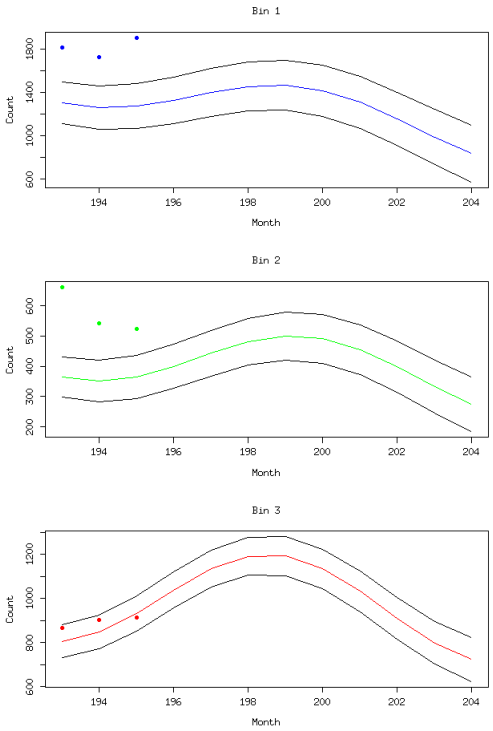


These models have corresponding R^2 values given in the table below

	Bin 1	Bin 2	Bin 3
R^2	0.9539	0.9482	0.9305

which suggest that these models are significantly better than the linear models. Below is a plot of predictions using these models for the entire next year. The black lines are 95% confidence intervals, the colored lines are the fits, and the points are the actual values.

Non-Parametric Fits

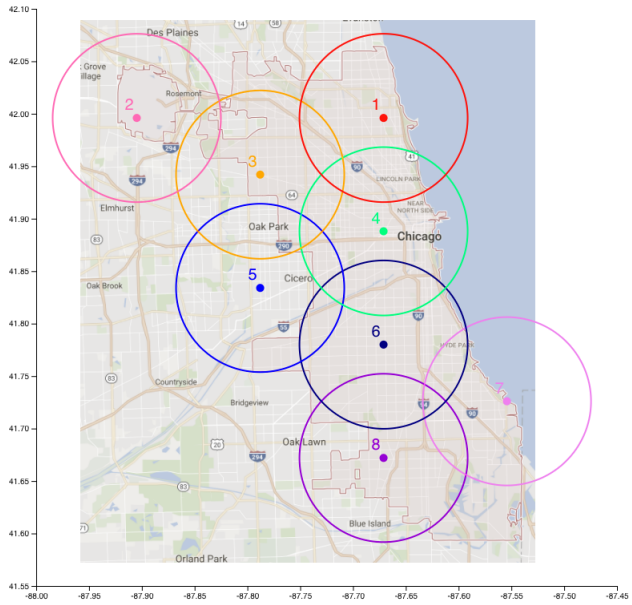


Looking at these plots, it is clear that, although these models are improvements, they do not extrapolate accurately. However, for bin 3, the predictions are very good and the actual values fall within the confidence intervals. For bin 1, as with the no arrest predictions, the model predicts the general trend of crimes and is off on the values. As for bin 2, the model is not effective in predicting at all. However, fitting these data with any other models did not yield any true improvement in prediction.

LOCALIZED CRIME COUNTS

This section continues to look at crime counts per bin but instead of looking at the entirety of Chicago, we chose 8 points throughout Chicago and considered crimes local to those centers. The centers are shown in the plot below and their latitude and longitude coordinates are given in a table for reproducibility's sake. Note that the colors in this plot are consistent with the coloring in the rest of this section and the colors align with the label on the centers (labels 1-8).

Center Label	Longitude	Latitude
1	-87.67107	41.996
2	-87.90479	41.996
3	-87.78793	41.942
4	-87.67107	41.888
5	-87.78793	41.834
6	-87.67107	41.780
7	-87.55421	41.726
8	-87.67107	41.672



To determine the count of crimes local to each of the chosen centers, we decided that a crime would be considered local to that center if the Euclidean distance between that crime and the center was less than 0.07956. This radius was chosen such that all the crimes were covered while minimizing the overlap.

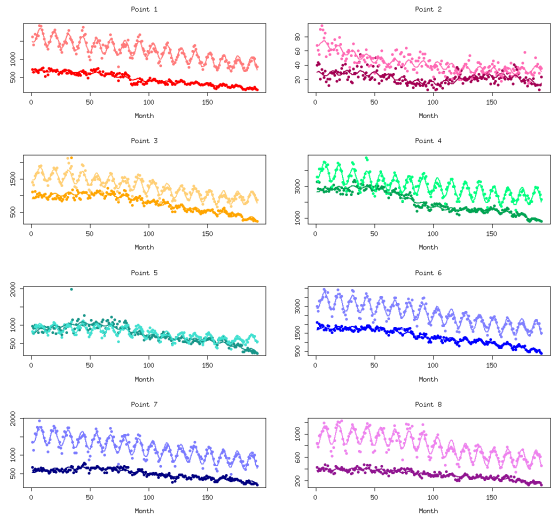
Unlike in the first section, this section will be broken up by bin as opposed to arrest/no arrest because there are interesting trends within each bin. This section will first consider petty crimes, then non-aggressive crimes, and finally aggressive crimes, maintaining the separation of arrest and no arrest throughout. Since the predicted counts in the last section were improved greatly using natural cubic spline models, this section will fit the data using a similar model. The models will be of the form

$$Y = \beta_0 + \beta_1 \cos\left(\frac{2\pi X}{12}\right) + \beta_2 \sin\left(\frac{2\pi X}{12}\right) + f(X)$$

where Y is the crime count, X is the month, and $f(X)$ is a natural cubic spline with knots at month 55, 100, and 150. These models were sufficient with fewer knots because there was less data to fit and less variation in those data.

Below are plots of crime counts in the first bin around each center. The arrest data is in color and corresponds to the center's location while the no arrest data is added to each plot in a lighter tone of that same color. Additionally, each plot includes a regression fit of the form above.

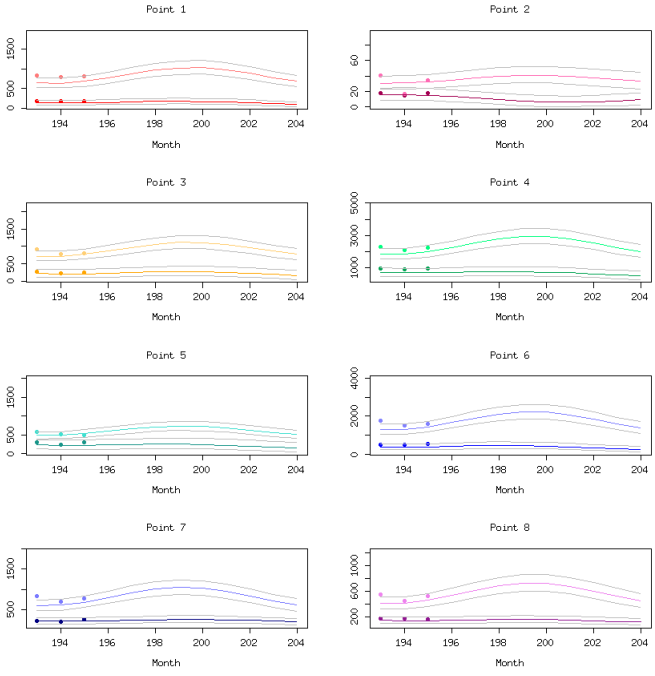
Bin 1 Crime Counts



Looking at the plots, it is obvious that these fits are significantly better than simple linear fits (or natural cubic splines) would be. This can also be verified by comparing the Akaike information criterion for the two models. For the sake of space, this table is not included.

We used these models to predict the next full year and included a plot with the fits, confidence intervals, and actual values for January, February, and March. In the plots, the colored lines represent the fits, the black lines the 95% confidence intervals, and the data points are the actual values.

Predictive Counts and Confidence Intervals

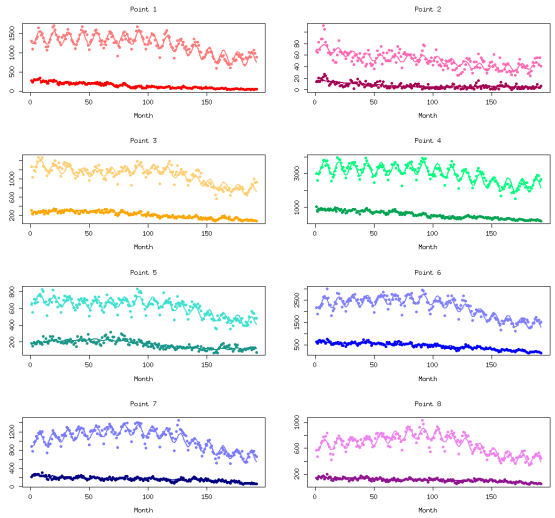


The three data points in these plots represent January, February, and March of 2017. It is clear that our models

predict the general trends of the data relatively well, despite the fact that the predicted values are not perfect. For bin 1, the least severe crimes, the actual values fall in the 95% confidence intervals 100% of the time for arrest crimes but only 58.3% of the time for no arrest crimes. A reason for this vast difference lies in the variability of the data. The no arrest crimes are more difficult to predict because they differ more month to month whereas the arrest crimes remain very close to constant, as is supported by the previous plot. However, we decide that changing the model would not significantly improve prediction for no arrest crimes.

Below is the fitted data for bin 2 crimes, arrest and no arrest. As in the previous figures, the arrest crime is in the darker color whereas the no arrest crime is the lighter shade.

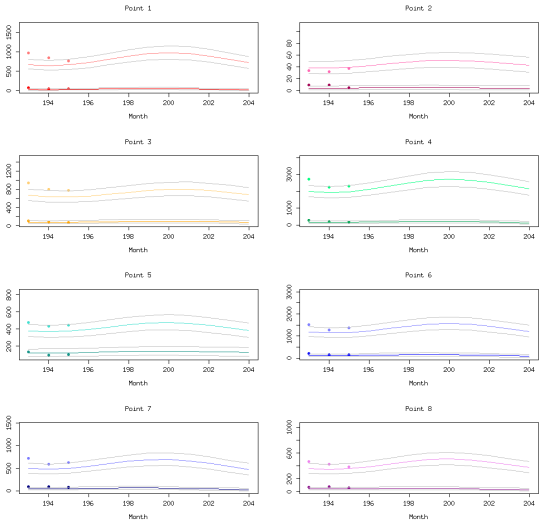
Bin 2 Crime Counts



Notice the increase in crimes at all points around month 100 which corresponds to 2008. We believe that this is related to the great recession that officially begin in December 2007 and lasted until June 2009 [8]. There is a historical correlation between recessions and rises in crime and as such, it's expected there would be in crime during this time as well [9]. The general trend in all these data is about the same for all points except for point 2. Looking at the map, point 2 is the point furthest away from all other data. It includes the crimes in Northern Chicago and there is not much significant variability, hence the worse fit.

The following plot shows the predictions for the next full year in addition to confidence intervals and the data that we have thus far (for January, February, and March this year). The coloring is the same as the previous plot.

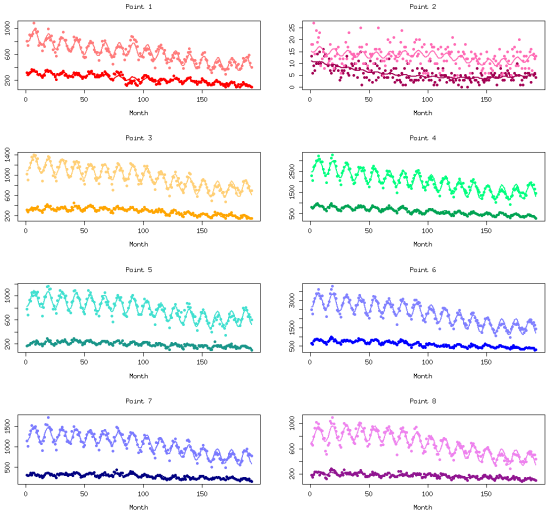
Predictive Counts and Confidence Intervals



These fits describe the data marginally worse than in bin 1. Our guess as to why this is, is because the data at each point is less predictable. It falls suddenly around 2008 and appears as though it may be steadily increasing by the end of 2016. Our fits, however, become sinusoidal around a linear function as opposed to a polynomial function (as they did where we fit the data). This is a characteristic of natural cubic splines and we had to use them so that we could extrapolate our data. Our predictions for arrest crime fall within the 95% confidence intervals 58.3% of the time while the predictions for no arrest crimes fall within the intervals about 50% of the time. Despite the fact that these predictions are much worse than the predictions for bin 1 crimes were, it makes sense from looking at the data.

Below are the fits for the bin 3 crimes - the most severe. Unlike the crimes in bin 2, there more of a steady, constant decrease in crimes over time so we expected that our predictions will be much better with these data.

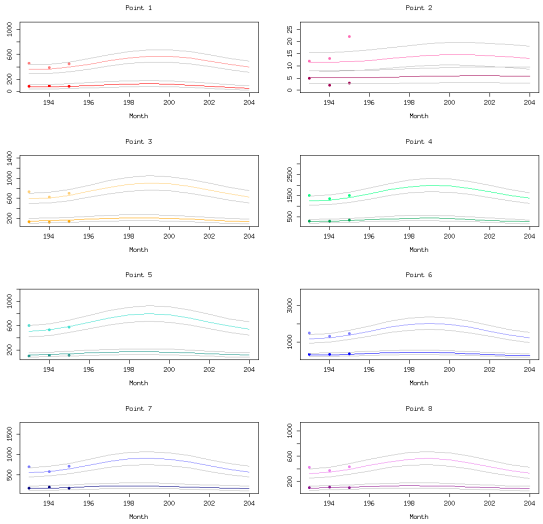
Bin 3 Crime Counts



We were correct in our expectation that predictions for

bin 3 would be much better. The actual values fell within our confidence intervals 95.8% of the time for arrest crimes and 66.7% of the time for no arrest crimes, an improvement from both of the other bins. Below is the same plot as was shown for bin 1 and bin 2 of predictions for the next full year.

Predictive Counts and Confidence Intervals



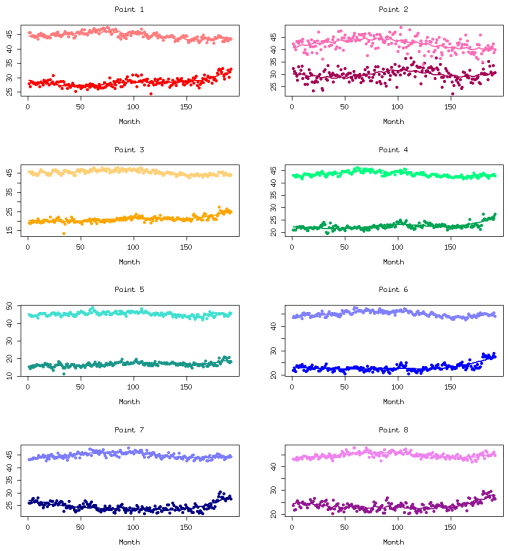
LOCALIZED SEVERITIES

This section will consider the same 8 centers as the previous section but instead of considering counts of crimes local to those points, it will consider the average severity of localized crime. The goal of this section is to see if crimes are more or less severe on average surrounding different points. It will lead into the next section which attempts to generate predictive severity maps.

The models in this section are of the same form as the models in the previous section: natural cubic splines with variation that is dependent on month. Average severity at each point is much more constant than crime counts. This means that although the number of crimes per location may differ, the severity of the crimes is about the same. Although severity across all bins is approximately constant, it is interesting to consider the potential historical influence on severity of crime. This paper will not delve into the history but will leave that analysis to the reader.

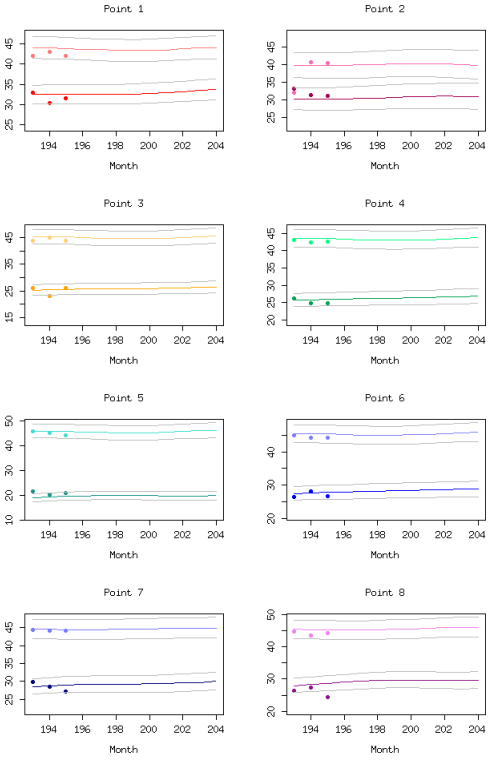
Below is the average severity at each point for bin 1. As for with the counts, there is much more variability at point 2 than at all of the other points. The effect of this is a much less accurate fit. Notice that the of severity of the arrest crimes is lower than the severity of the no arrest crimes at every point. Assuming that this is not a problem with the severity mapping (which it very well may be), it provides more support for splitting the data into arrest and no arrest crimes. It does not make logical sense that people would get arrested for crimes that were less severe than the crimes that people did not get arrested for.

Bin 1 Average Crime Severity



Across all bins, there is a slight increase in severity level starting at about month 150 (2012). Since our model is linear outside of the data (besides the sine and cosine variability), it will predict the increase in average severity as being linear. Below is the plot of predictions with confidence intervals and actual values for bin 1.

Predictive Severities and Confidence Intervals

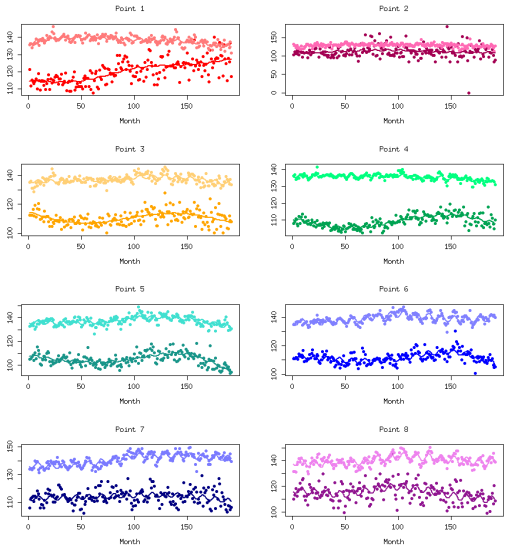


Based on the reduction of deviation from counts to average severities, it is not surprising that the actual values fall in

the predictive confidence intervals much more. In particular, 87.5% of the arrest data falls within the intervals and 95.8% of the no arrest data falls within the intervals.

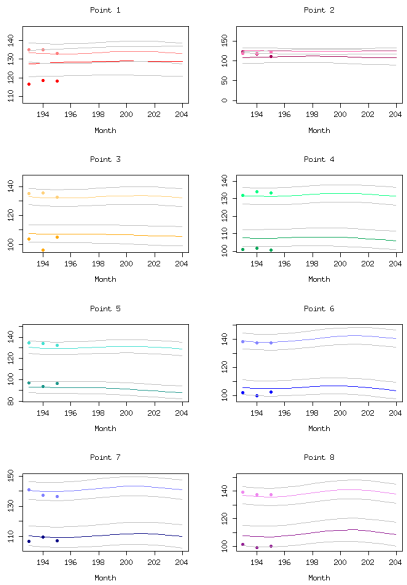
Bin 1 was relatively easy to model because of the lack of variation in the data. Bin 2, however, includes much more variation (especially in the arrest data) which yields potentially worse fits and worse predictions. The average severity data for bin 2 is shown below.

Bin 2 Average Crime Severity



Looking at this data, there is an interesting rise (between month 75 and 140) and fall (between month 140 and 190) in average severity. This corresponds to the period between 2006 and 2016 and may be due to the 2008 recession and later economic recovery. The predictions using these models are shown below.

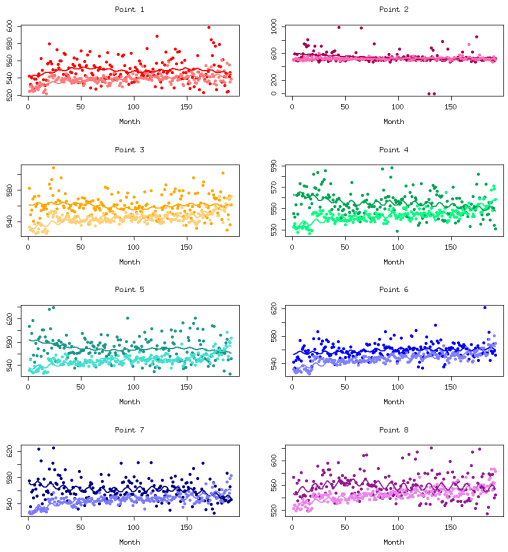
Predictive Severities and Confidence Intervals



Notice that these predictions are all relatively accurate besides at point 1 where the arrest predictions are nowhere near the true values. This indicates potential problems in our model that we will not address. Despite these problems, the true values fall within the confidence intervals 58.3% of the time for the arrest data and 95.8% of the time for the no arrest data. This drastic decline in predictive ability was foreseen because the arrest data has much more variability than the no arrest data.

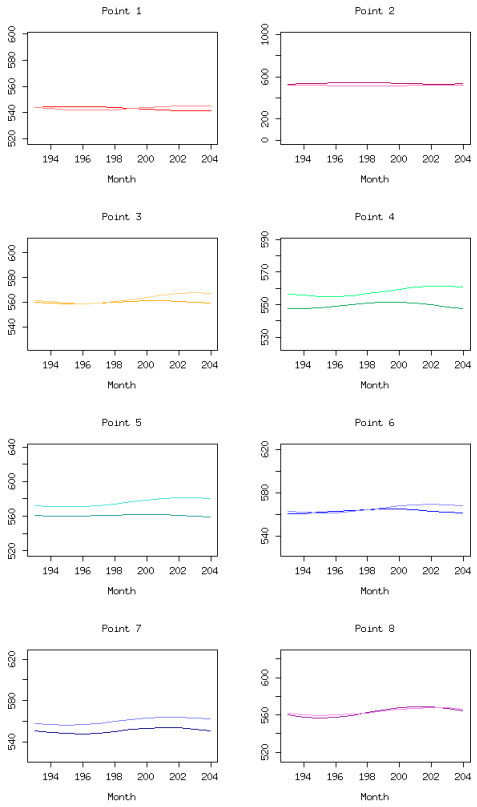
Looking at the plots of the bin 3 data, it seems as though there is much more variability. Since the variability in the bin 2 arrest data significantly impacted prediction it would not be surprising if prediction was bad for both no arrest and arrest crimes in bin 3. Below is the plot of the data in addition to the fits. Notice that, unlike the data in bin 1 and bin 2, there is not a clear and simple separation between the arrest and no arrest data.

Bin 3 Average Crime Severity



Because of the lack of separability of the no arrest and arrest data, it is not insightful to look at the predictive plots with the confidence intervals to assess the fits. For that reason, the plots below include only the prediction lines for 2017.

Predictive Severities and Confidence Intervals



66.6% of the actual arrest data values fall within the 95% confidence intervals whereas 87.5% of the no arrest data fall within the confidence intervals. The reason behind this difference is the same as the reason behind the difference in predictive ability for the models in bin 2. Although there is more variability in both data, there is more in the arrest data.

SUMMARY OF PREDICTIONS FOR LOCALIZED DATA

In the last two sections, we looked at the data local to eight points. At each point, we generated a model fit and attempted to predict the values for January, February, and March. The table below provides the percentage of actual values that fell within the confidence intervals for those values. This table is included to compare both sections.

Table to Assess Accuracy of Predictions

Arrest	Bin	Count Prediction	Severity Prediction
Yes	1	100%	87.5%
	2	58.3%	58.3%
	3	95.8%	66.6%
No	1	58.3%	95.8%
	2	50%	95.8%
	3	66.7%	87.5%

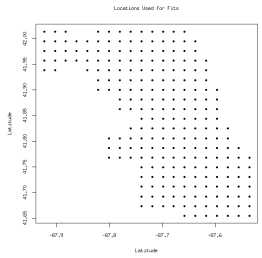
It is interesting to see that the count predictions were more accurate for the arrest data whereas the average severity predictions were more accurate for the no arrest data. Looking back at the plots however, it is obvious as to why that is. There is a lot of variability in the arrest data with respect to average severity but very little with respect to counts. Likewise, there is a lot of variability in the no arrest data when looking at counts and much less when considering severity.

Despite the fact that prediction is better for arrest counts and no arrest severities, the following section will only consider generating heat maps for severities. The reason behind this is that it does not make sense to generate a continuous map for a count because count is a discrete value whereas average severity is not.

HEAT MAPS

This section aims to expand the concepts from the previous sections, using the same models but increasing the number of points across Chicago and reducing the radius. The goal is to generate predictive severity heat maps for January, February, and March of 2017 and compare those to the actual maps.

The points chosen across Chicago are shown in the plot below. A crime was considered local to a point if the Euclidean distance between that point and the crime was less than 0.0189, about 7× smaller than the radius used in the previous sections.



These points were chosen by laying a 20×20 grid across the entire latitude-longitude range and picking the points that fell within Chicago.

Recall the model from the previous sections given by

$$Y = \beta_0 + \beta_1 \cos\left(\frac{2\pi X}{12}\right) + \beta_2 \sin\left(\frac{2\pi X}{12}\right) + f(X)$$

where Y is the response (either crime count or severity) and X is the month. To predict a severity map, a regression curve will be fit to each point and the curve will be extrapolated to the three consecutive months, just as was done in the previous section.

Extrapolating the fits to predict January, February, and March of 2017 yields the percentage of predictions that fell within the 95% confidence intervals that are given below.

Table to Assess Accuracy of Predictions

Arrest	Bin	Count Prediction	Severity Prediction
Yes	1	77.35%	70.51%
	2	74.79%	68.80%
	3	69.53%	71.24%
No	1	42.62%	62.45%
	2	31.93%	63.03
	3	39.41%	60.59%

These models were relatively good at predicting. Recall the small variability in arrest counts per location from the last section. This small variability allowed the predictions to be about 70% accurate for each of the three bins. However, because the variability was small in the no arrest severity data (from the last section), we expected that the predictions would be better than around 60%. Nevertheless, we will use these predictions to generate heat maps across the Chicago area. Prior to doing so, it is important to understand why heat maps for severity are the only ones being generated. The reason behind this is that we will be able to compare predicted severities to actual severities at each location but would not be able to do that for counts (since the predictions are made using larger radii and the actual values use smaller). Changing the radius should not change the average severity at each location too much.

In order to generate severity heat maps, we used a geostatistical method called Ordinary Kriging. This method assumes that the observed values, $Y(\mathbf{s})$ for $\mathbf{s} = \mathbf{s_1} \dots \mathbf{s_n}$, are of the form

$$Y(\mathbf{s}) = \mu + Z(\mathbf{s}) + \varepsilon$$

where μ is a constant, unknown mean, $Z(\mathbf{s})$ is a mean zero isotropic spatial process, and ε is the error. In this paper, $\mathbf{s}$ is a vector of latitude and longitude. For a new point, $\mathbf{s_0}$, the ordinary kriging predictor is given by

$$\hat{Y}(\mathbf{s_0}) = \mathbf{\Sigma_0}^T \Sigma^{-1} \mathbf{y} + \frac{1 - \mathbf{1}\Sigma^{-1}\mathbf{\Sigma_0}}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}} \Sigma^{-1}\mathbf{y}$$

where

$$\mathbf{y} = \Big(Y(\mathbf{s_1}), \dots, Y(\mathbf{s_n})\Big)^T \quad \text{is a vector of observed points}$$

$$\mathbf{1} = \Big(1, \dots, 1\Big) \quad \text{has dimension n}\times 1$$

$$\Sigma = \text{Cov}\Big(\mathbf{y}, \mathbf{y}\Big) \quad \text{is the covariance of the data with itself}$$

$$\mathbf{\Sigma_0} = \text{Cov}\Big(\mathbf{y}, \mathbf{s_0}\Big) \quad \text{is covariance between data and } \mathbf{s_0}$$

Prior to generating heat maps for our predictive data, we had to determine a covariance function for those data. Using empirical binned semivariograms, we decided that our heat maps would be good enough if we used an exponential covariance. Under this assumption, the covariance between $\mathbf{s_i}$ and $\mathbf{s_j}$ is given by

$$\text{Cov}\Big(\mathbf{s_i}, \mathbf{s_j}\Big) = \sigma^2 \text{e}^{-||\mathbf{s_i}-\mathbf{s_j}||/a} + \tau^2 \mathbb{1}_{\{\mathbf{s_i}=\mathbf{s_j}\}}$$

where σ^2 is the variance of the process, a is the range, $||\mathbf{s_i} - \mathbf{s_j}||$ is the Euclidean distance, and τ^2 is the variance of a white noise process that comes into account when $\mathbf{s_i} = \mathbf{s_j}$.

For the sake of space, we only included the spatial parameters and generated heat maps for January. The purpose of this is to give the reader a better idea as to what the effect of ordinary kriging is.

The spatial parameters for the arrest and no arrest data are given below. These parameters do not change much between months since the predictive values do not change a lot.

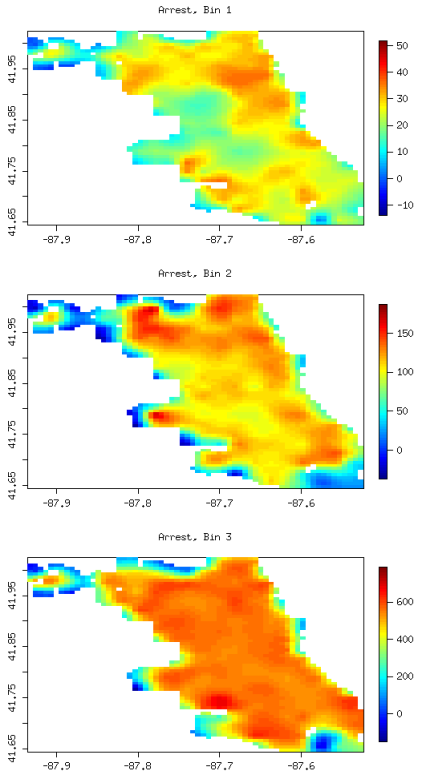
Covariance Parameters for January

Arrest	Bin	σ^2	a	τ^2
Yes	1	150.91	0.00637	50.30
	2	2848.19	0.00382	949.39
	3	55794.54	0.00476	18598.18
No	1	279.69	0.00447	55.94
	2	6117.69	0.00520	$1.886\times10{-}08$
	3	75434.37	0.00388	6.17×10^{-10}

Recall that σ^2 is the variance of the process. There is a consistent rise in variance as bin increases. Also, the magnitudes of the variances are very close between arrest and no arrest data which as expected based on the fact that there is more variability between points for bin 3 than for bin 2 and bin 1. The range is approximately the same for each covariance function and for that reason, is not very interesting. What is interesting is the noise. Under this model, there appears to be a lot of noise in the arrest data and much less in the no arrest data. This is interesting and could be due to the fact that there are more obvious clusters in the no arrest data.

Below are the heat maps generated using the covariance parameters for the arrest data. Recall that for each of the models, about 70% of the actual data points fell within the confidence intervals. That being said, these maps are not expected to be perfect since we used ordinary kriging on the already predicted values.

January Predictive Arrest Maps

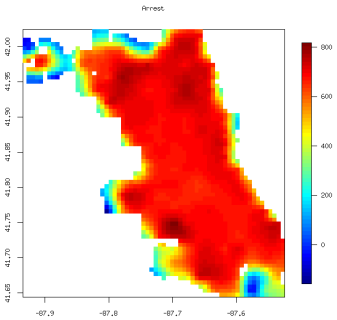


Looking at these plots, it is clear that there are regions in Chicago that have more severe crimes on average. These maps are interesting because they suggest that the more

severe areas are more condenseed for the less severe crimes. In particular, there is more obvious deviation in bin 1 than in bin 3.

In order to summarize the severity across Chicago, we added all of these heat maps and generated the following severity heat map for January 2017.

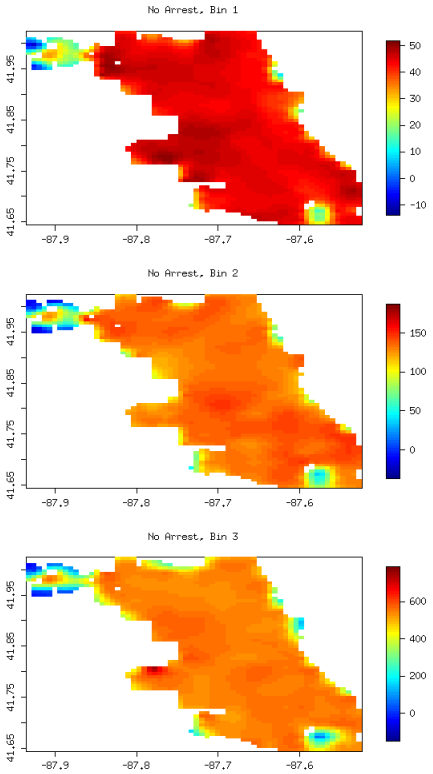
Arrest Data Severity Map



The most severe places in Chicago, with respect to arrest data, are up north and in the south. After adding the severity maps together, it appears as though the severity in central Chicago is approximately constant whereas crimes, on average, become more severe along the edges and drop off as you move outside of Chicago.

Performing the same method of ordinary kriging for January on the no arrest data yields the following heat plots.

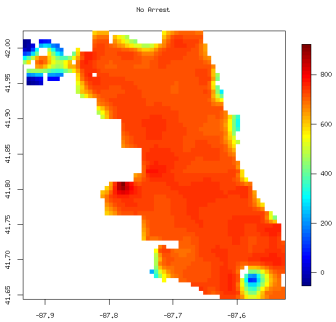
January Predictive No Arrest Maps



These maps suggest that on the eastern side of Chicago, there are more severe petty crimes than in the west. The other two maps suggest that there is more severe crime without an arrest in the southern side of Chicago.

Adding all of these maps together generates the map given below.

Predictive January Summary Heat Map

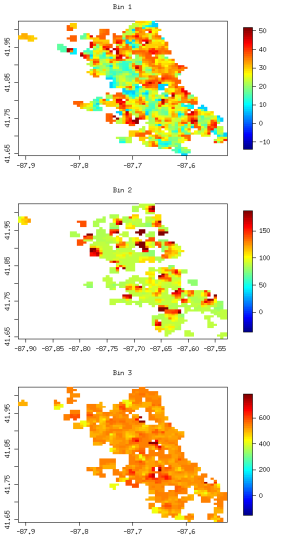


Notice that, although this map is relatively uninteresting, it does suggest that there are regions of more severe crimes.

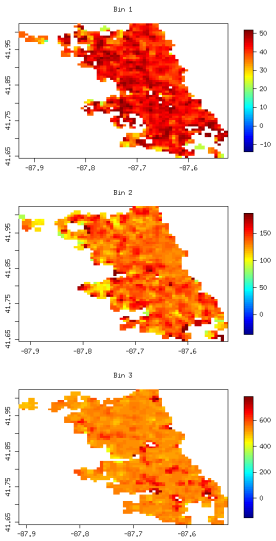
RESULTS

This section will discuss the results from the analysis section in addition to providing heat maps of the actual values and histograms of the difference between our final severity predictions and the actual data values. We determined the data values by collecting the counts of crime and severity of crime at 10000 points across Chicago, considering crimes to be local to a point if they fell within a Euclidean distance of 0.00756 from that point. The maps for January are given below and can be compared to the predictive heat maps given above.

Actual Arrest Severity Maps



Actual No Arrest Severity Maps



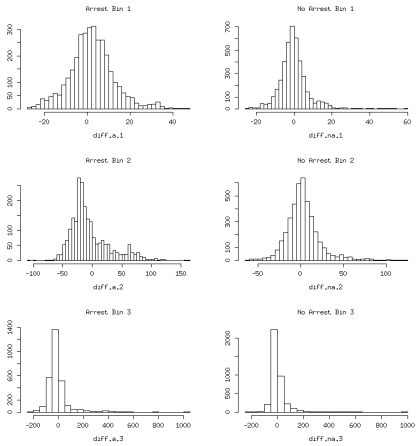
At a first glance, these maps might suggest that our predictions are bad for the arrest data and better for the no arrest data. This is merely due to the fact that the kriged predictions are smoother than the actual data values. However, looking at the general trends in the data, it is clear that the predictive maps are very good fits and capture a lot of the spatial variation of severity.

Because we predicted values and then generated heat maps according to those values, we are not able to accurately develop confidence intervals. To assess goodness of prediction, we will instead look at the histogram of the difference between actual and predicted values. In particular,

$$\text{Difference} = \text{Actual - Predicted}$$

Below are the histograms of the differences for the month of January 2017.

Histograms of January Differences



Notice that the histograms of the differences in bin are much more normally distributed than in bins 2 and 3. Bin 2 is much more weighted towards the negative side, as is bin 3. The means and standard deviations of the differences are shown in the table below.

Arrest	Bin	μ	σ
Yes	1	2.18	10.42
	2	-4.75	32.95
	3	-10.99	82.82
No	1	-0.31	7.40
	2	2.43	19.64
	3	2.37	63.02

For $\mu < 0$, the predicted values overestimate the actual values. Based on the means and standard deviations, the predictive fits are not that far off from the actual fits. Notice that the standard deviation for the no arrest crimes is much higher on average. The reason behind this being that the severities in bin 3 are much larger in magnitude than those in bins 1 and 2.

The heat plots and histograms for February and March look about the same as the heat plots and histograms for January and were not included for that reason. However, here are the corresponding means and standard deviations of the differences between the predictive fits and the actual values.

Month	Arrest	Bin	μ	σ
February	Yes	1	2.36	10.95
		2	-2.68	38.16
		3	6.81	186.41
	No	1	-0.40	6.60
		2	-0.40	6.60
		3	21.72	149.11
March	Yes	1	2.12	10.53
		2	-4.60	32.96
		3	-10.65	82.64
	No	1	-0.16	7.40
		2	3.48	19.74
		3	4.16	63.07

This table shows that even extrapolating our models yields relatively good heat maps. Prior to generating the heat maps, it may be worth noting that the actual values at the points we used to generate the maps fell within the confidence intervals about 65% of the time.

APPLICATIONS

The ability to predict the average severity of crimes across Chicago is useful for many different reasons. Firstly, it could help local law enforcement to better allocate their limited resources. It could also give potential residents a feel for the city without ever being there (like where to buy a house without knowing the area). It could also provide residents with a guideline of what and where to avoid. All in all, the ability to predict the severities of crimes across a city is very powerful and could potentially have many applications.

References

[1] *Chicago Police Try to Predict Who May Shoot or Be Shot*. New York Times. 2016.
https://www.nytimes.com/2016/05/24/us/armed-with-data-chicago-police-try-to-predict-who-may-shoot-or-be-shot.html

[2] *Predicting Crimes in Chicago from Weather*.
Alan Fu. 2015.
https://fuyuheng.github.io/EECS-349-Project/

[3] *Entertainment as Crime Prevention: Evidence from Chicago Sports Games*.
University of California, Berkeley. 2014.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2429551

[4] Chicago Crimes - 2001 to Present
https://catalog.data.gov/dataset/crimes-2001-to-present-398a4

[5] Illinois Uniform Crime Reporting Codes
https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e/data

[6] Canadian Crime Severity Index
http://www.statcan.gc.ca/pub/85-004-x/2009001/t001-eng.htmT001FN1

[7] Prison Sentence Data
http://famm.org/wp-content/uploads/2013/08/Chart-All-Fed-MMs-NW.pdf

[8] The Great Recession
http://www.investopedia.com/terms/g/great-recession.asp

[9] Correlation of Crime and Recessions
https://www.weforum.org/agenda/2015/03/do-recessions-increase-crime/