Case Study 2 (CS2): Evaluating Crime Prediction Performance

"On my honor, I pledge that I have neither given nor received help on this assignment."

Name: Prateek Agrawal
Computing ID: pa7sb

SITUATION

The Chicago Police Department (CPD) was pleased with the kernel density estimate (KDE) visualizations you produced. However, despite their nice visual properties, it is difficult for crime analysts to quantify how good KDEs are. Thus, CPD has requested additional work. Specifically, they want to know the following:

1.) How should the goodness of a KDE be evaluated? Be specific, including equations for the required calculations?

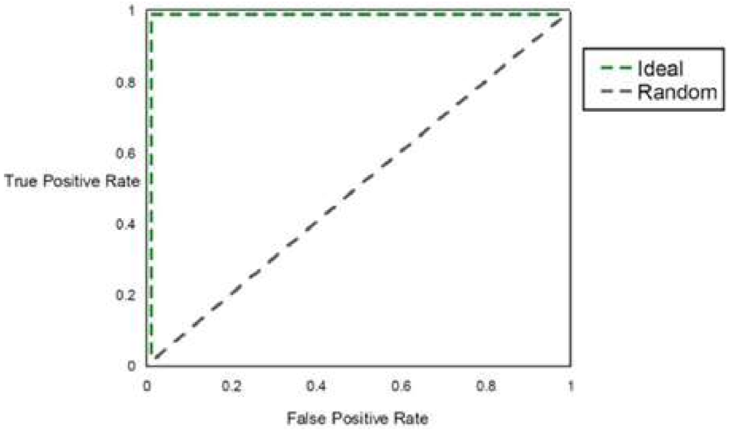
Solution:

In our Kernel density estimate we set the grid size to 200mtr. This divides the KDE in squares of 200mtr. The top 10% cells above the two standard deviation s from the mean were considered “hot spots” for the KDE techniques.

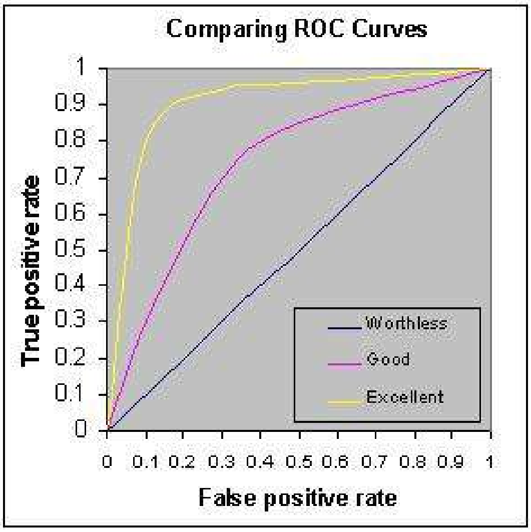
For evaluation of our Kernel density estimates we used a surveillance plots which uses ROC curves that plots the true positive rate to false positive rate. Theoretically it can be written as:

FPR = False Positive/Ground Negative
TPR= True Positive/Ground Positive;

An ideal ROC can be shown below:



Therefore, as the random curve reaches the ideal curve better the curve or we can say better the model as shown below,



In our KDE, the squares are sorted in descending order of higher probability of crimes occurring in the future. We do this because it helps us differentiate between the true positive theft cases and true negative theft cases (i.e. area where no theft occurred.) In our model's prediction, a surveillance plots measure the true positive crime that occurred within the x% most affected crime area.

For the purpose of this question, we summed the number of True thefts occurring in the most affected squared area(the size of the square is created by us) of the town square on each prediction hour, day or month and then dividing it by the total number of true Thefts occurring across all prediction hour, day or month.

%crime = True Positive in top affected area/ Total True Positive
%Area= area in the square under surveillance/ total area of the region

The area under the surveillance plots (AUC) is used to compare different KDE's. Better performing Surveillance plots are those which approach the upper-left corner, therefore, larger AUC.

2.) Are KDEs better at some times of the day, days of the week, or months of the year? You may focus on a single crime type and one hour-hour, day-day, and month-month comparison.
Solution:

Approach Taken:

Different KDE (month-wise, day-wise and hour wise) plotted for theft cases have been created which are trained on the data based on the previous theft cases reported and evaluated it on different time of the year. For evaluation we plotted the Surveillance plots which gives the AUC values which is then used to compare the different KDE's.

Data:

I have used the 2014 theft crime data from the Chicago data portal. The data is then manipulated in different dataset based on the requirement of the analysis.

Month wise KDE

Analysis and Evaluation:

I have created the Kernel Density Estimate using the data from the month of January to April and then using this model to predict the theft cases in the month of May. Similarly, another model is created which is trained on the data from the month of July to October and then used to predict theft cases in the month of November.

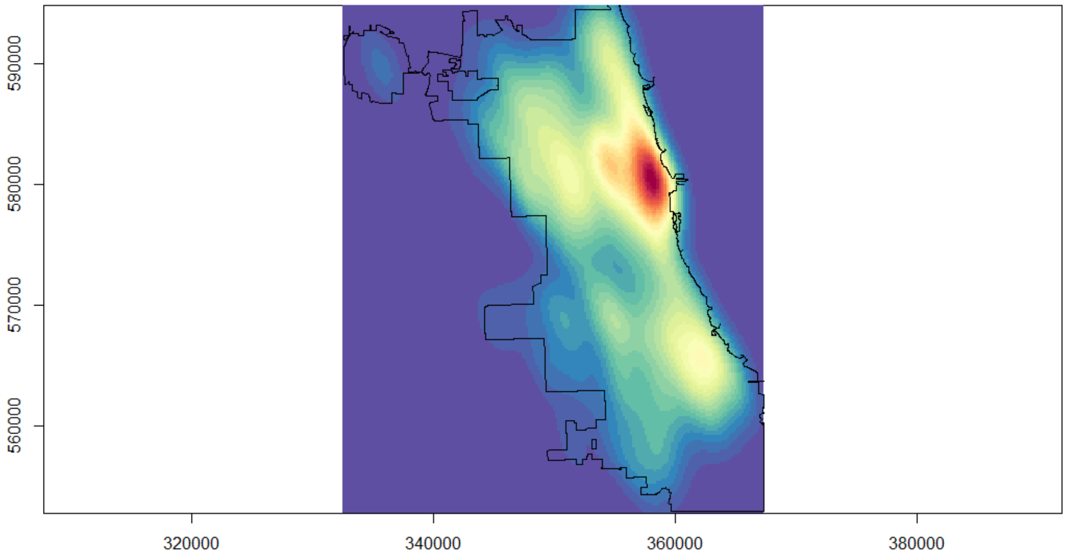


Figure1: KDE trained on the months of January to April

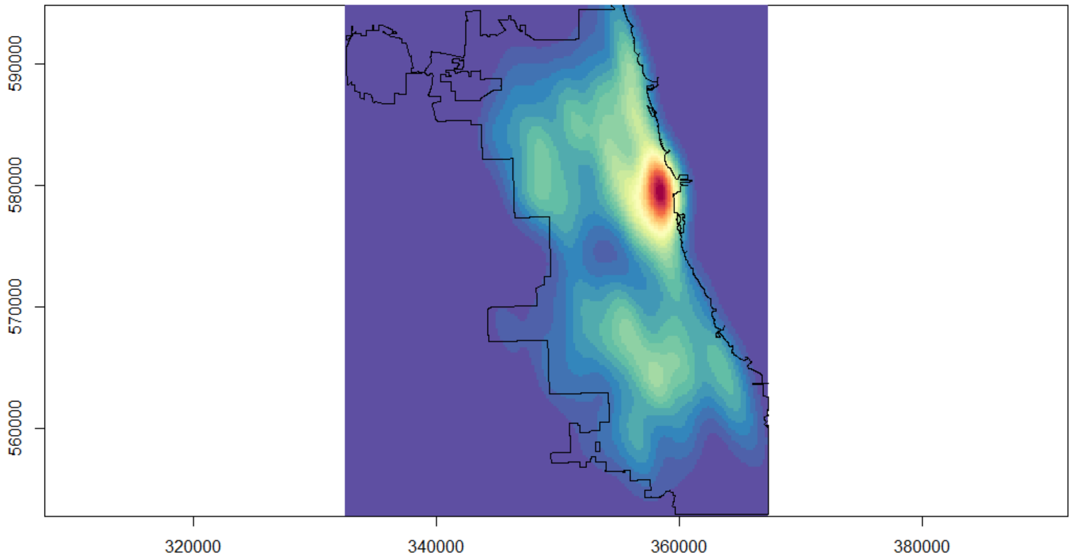


Figure2: KDE trained on the months of July to October

For Evaluation, Surveillance plots are used, which give the AUC values. We then compared the AUC values to evaluate the performance of our estimated monthly KDE plots. The higher the AUC value the better the KDE plot.

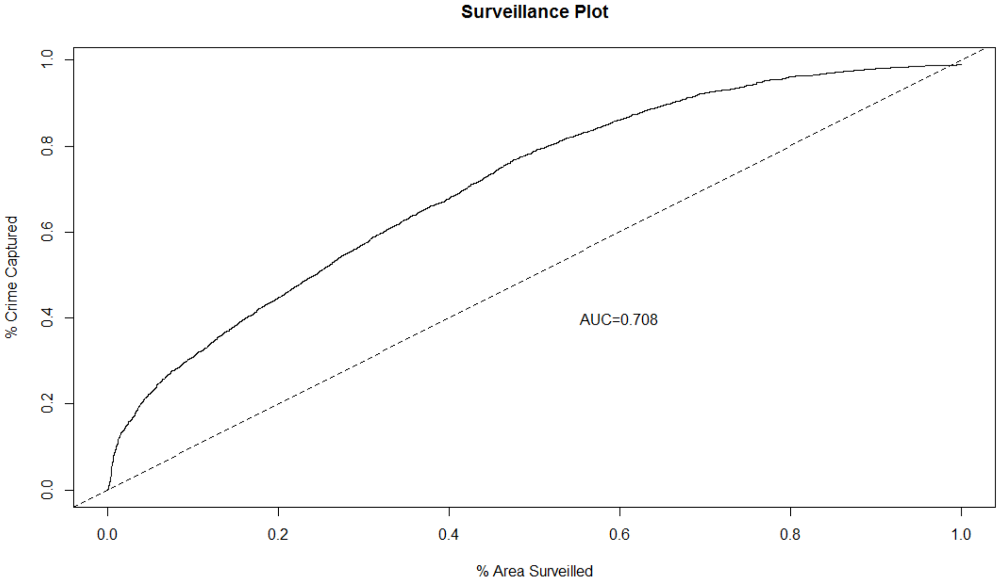


Figure3: Surveillance plot for the month of May

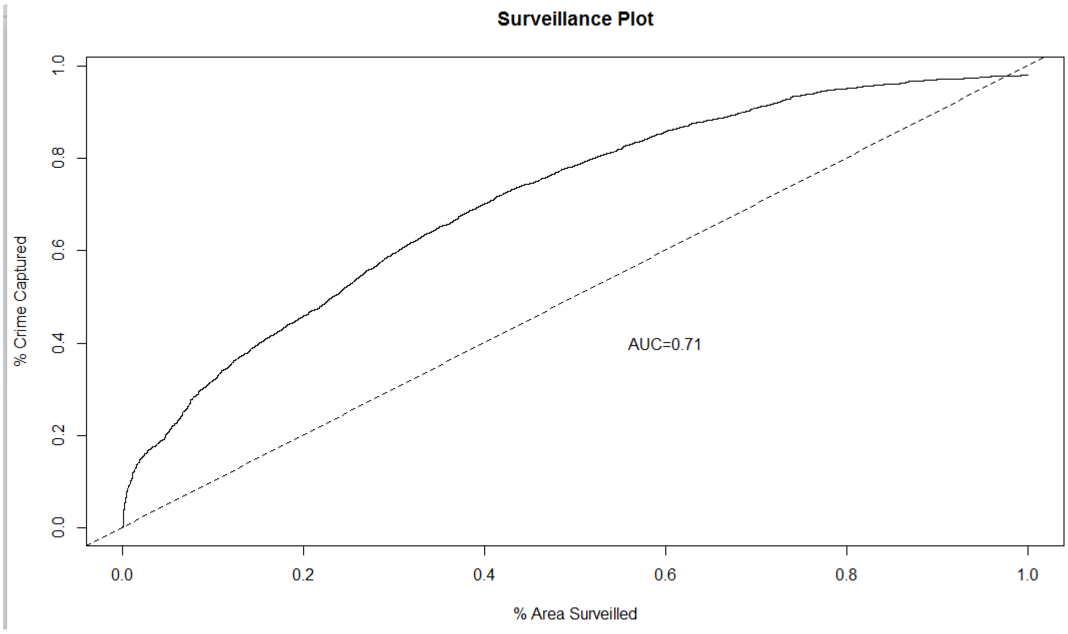
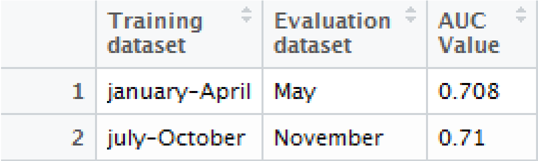


Figure4: Surveillance plot for the month of November

Therefore, the AUC values are:



	Training dataset	Evaluation dataset	AUC Value
1	january-April	May	0.708
2	july-October	November	0.71

It can be observed that the AUC value for the month of November is a little better than the AUC value for the month of May; therefore, we conclude that our estimate for the month of November is better than the month of May by a little margin.

Recommendations:

1. As there is a difference in the AUC values in different months of the year the crime analyst should focus on our KDE estimates for months which have better AUC.

Day wise KDE

Analysis and Evaluation:

Using the theft data for each Wednesday in the months of January to April, we have created our KDE. We then use this model to predict the theft cases occurring on each Wednesday in the month of May. Similarly, another model is created which is trained on the theft data from each Saturday in the months of January to April and then used to predict theft cases occurring on Saturday's in the month of May.

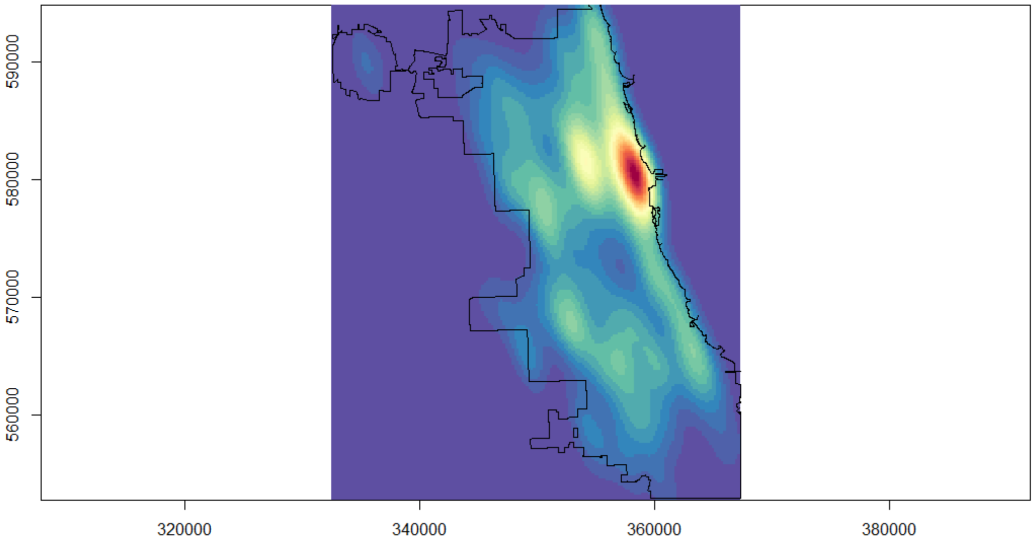


Figure5: KDE trained on Wednesday's in the months of January to April

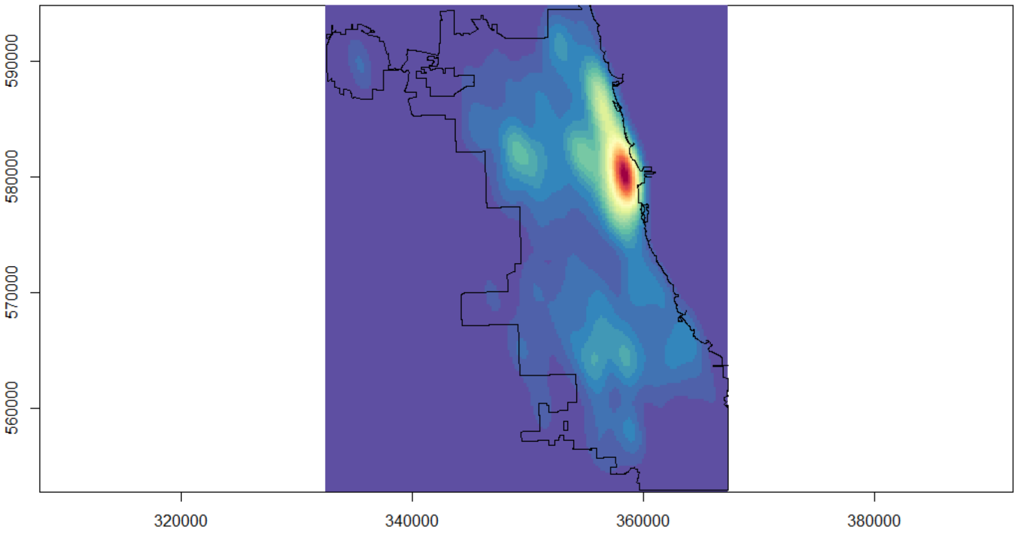


Figure6: KDE trained on every Saturday for the months of January to April

For evaluation of the KDE's, we use Surveillance plots which give the AUC values. We then compared the AUC values to evaluate the performance of our estimated monthly KDE plots. The higher the AUC value the better the KDE plot.

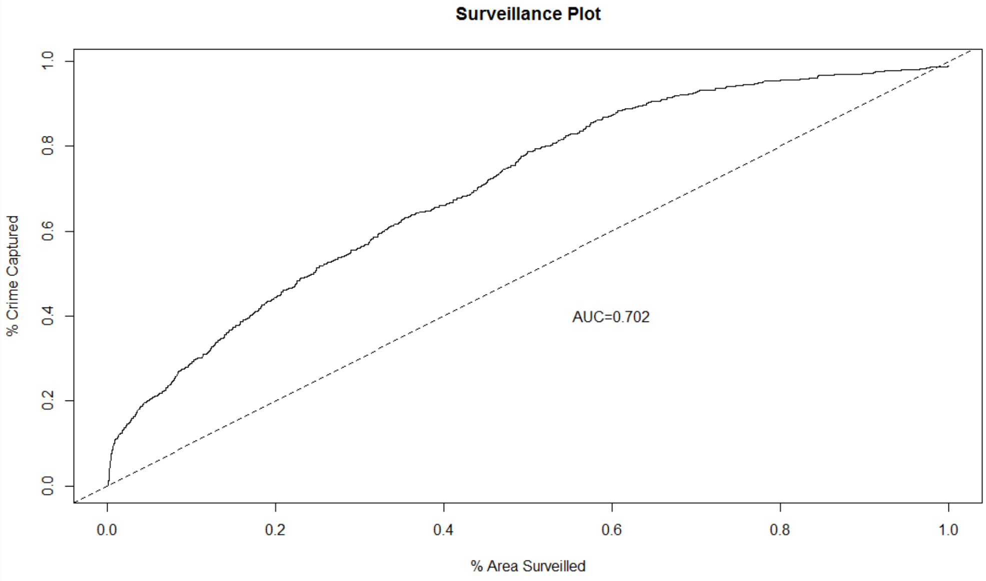


Figure7: Surveillance plot for Wednesday's in the month of May

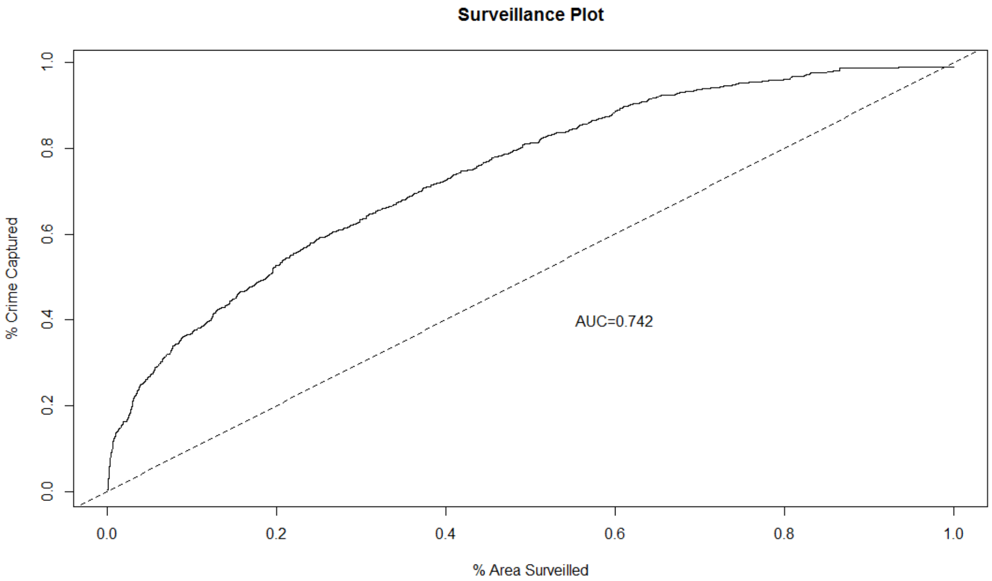
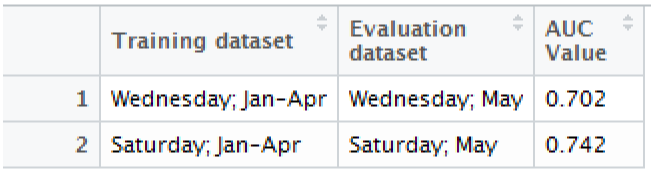


Figure8: Surveillance plot for every Saturday in the month of May

Therefore, the AUC values obtained are:



	Training dataset	Evaluation dataset	AUC Value
1	Wednesday; Jan-Apr	Wednesday; May	0.702
2	Saturday; Jan-Apr	Saturday; May	0.742

We can observe that the AUC value for Saturday's in the month of May has a better than the AUC value for Wednesday's in the month of May; therefore, we conclude that our estimate for the Saturday's in the month of May is better than Wednesday's in the month of May.

Recommendations:

1.) As there is a difference in the AUC values for different days, the crime analyst should focus on our KDE estimates of day's which have better AUC.

Hour wise KDE

Analysis and Evaluation:

I have created the Kernel Density Estimate based on the data of theft cases reported for 3am to 4am on every day of the month of January. This model is then used to predict the theft occurrences for each day of February from 3am to 4am. Similarly, another model is created which is trained on the data obtained from theft cases between the 3pm to 4pm on each day of January, which is then used to predict theft occurrences from 3pm to 4pm on each day in the month of February.

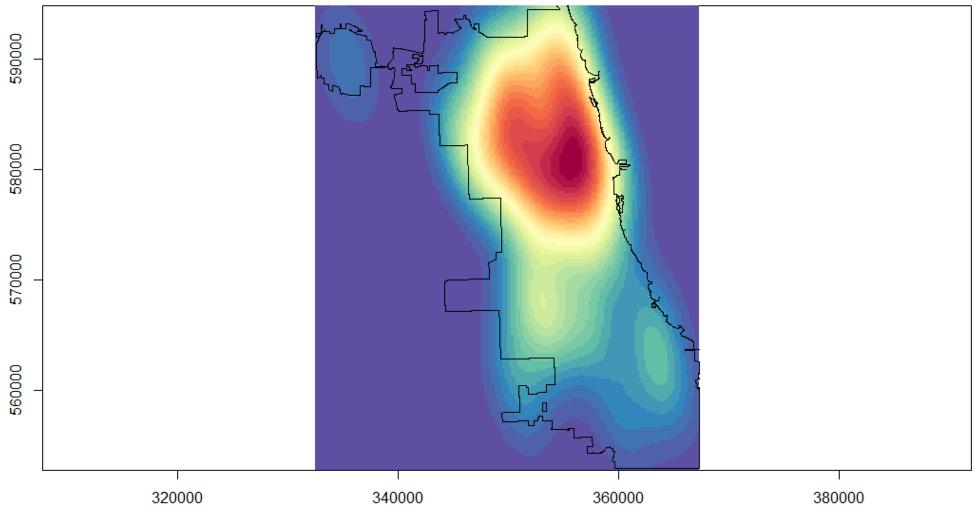


Figure9: KDE trained on theft cases from 3am to 4am on each day of the month of January.

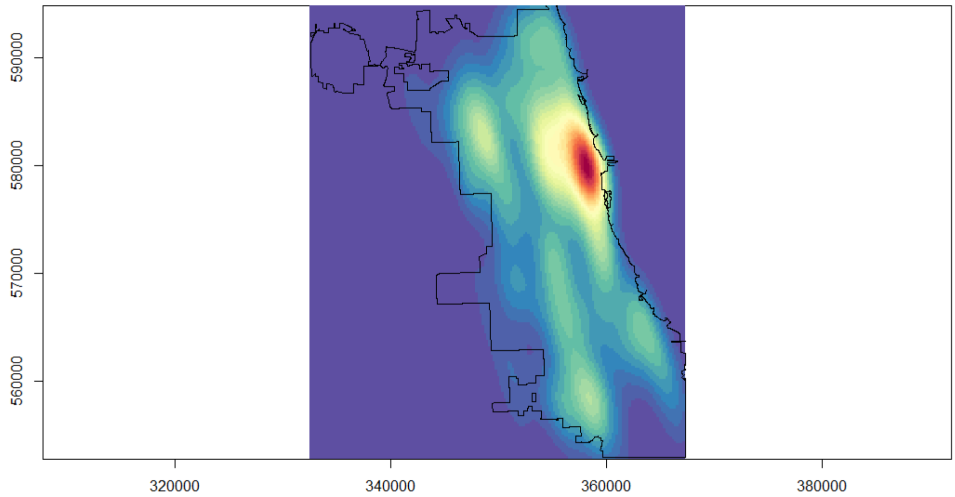


Figure10: KDE trained on theft cases from 3pm to 4pm on each day of the month of January.

For Evaluation, Surveillance plots are plotted which gives the AUC values; we then compared the AUC values to evaluate the performance of our estimated hourly KDE plots. The higher the AUC value the better the KDE.

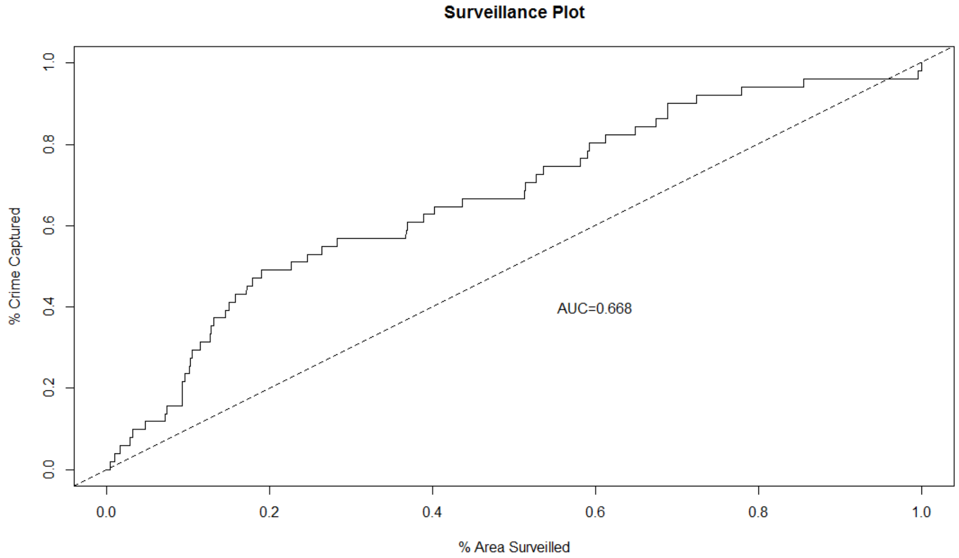


Figure11: Surveillance plot for theft cases occurring between 3am to 4am on each day of the month of February.

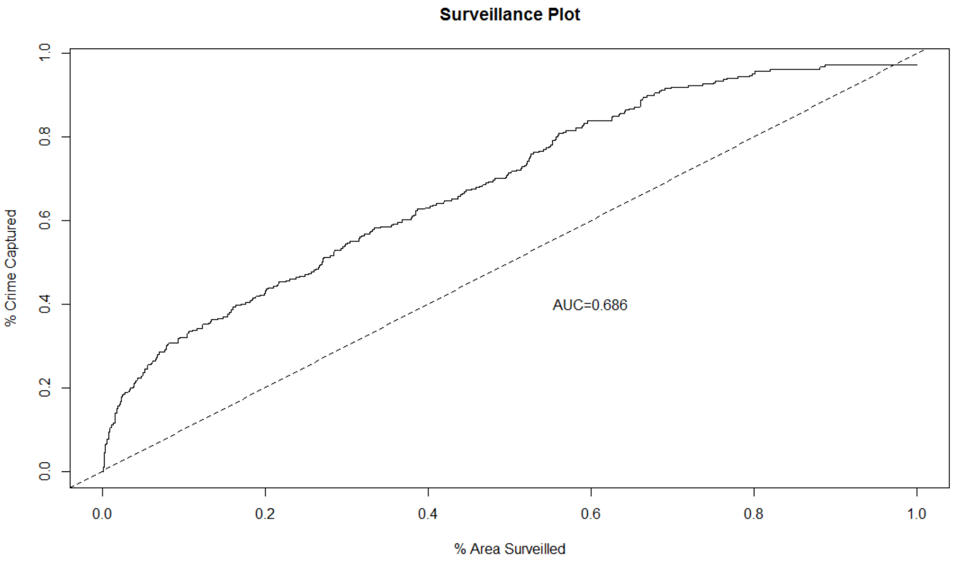
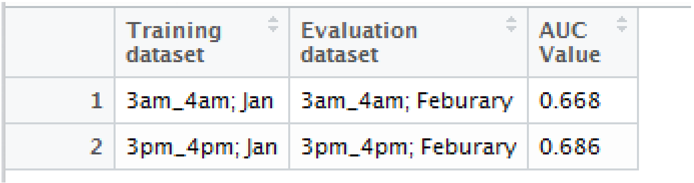


Figure12: Surveillance plot for theft cases occurring between 3pm to 4pm on each day of the month of February.

Therefore, the AUC values…



	Training dataset	Evaluation dataset	AUC Value
1	3am_4am; Jan	3am_4am; Feburary	0.668
2	3pm_4pm; Jan	3pm_4pm; Feburary	0.686

We can observe that the AUC value for our estimate of theft occurring between 3pm to 4pm on each day in the month of February is better than the AUC values of the estimate for 3pm to 4pm in the month of February by a little margin.

Recommendations:

1. As the AUC values are different for different time of the time, the crime analyst should focus on our KDE of the hour in the day which have better AUC values.

References:

1. Class Notes of Data Mining 6018
2. Matthew Gerber, Predicting Crime using Twitter and Kernel Density Estimation.
3. MetricComparisonPredictiveHotSpots&RTM_Drawve_2014.pdf