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: Sharath Chand PV
ID: vp4pa

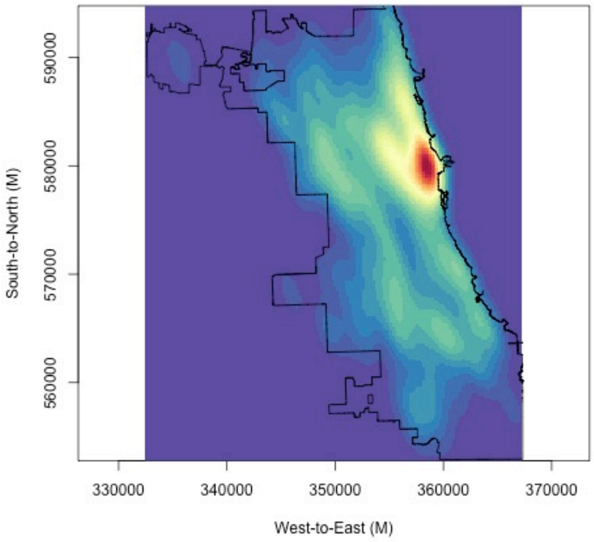
Introduction:

In the previous study we have presented how various crimes in Chicago can be predicted using KDE techniques. We have also studied how crime distribution changes based on hours of day, days of week and months. In this report we evaluated the performance of KDE techniques in crime prediction and tried to provide insights on how the police department can best use KDE findings through Surveillance Plots.

In the latter part of report, we further analyzed whether KDE techniques performs better for some specific hours, days or months. We also tried to find out if the performance is dependent on duration of past data used for estimating KDE. For example, if the crime is predicted better from past 3 months data than past 6 months data etc.

KDE Evaluation:

Below is the KDE hotspot plot, which essentially shows locations with high probability of theft. Before incorporating these findings in policing strategies, it is imperative to understand how good is the technique in predicting future crimes.



We have carried out our studies with a premise that crime is of cyclical nature and its distribution can be predicted from the distribution of past occurrences. We tested this hypothesis by validating predictions for a certain period against its actual occurrences. We used surveillance plots for the validation.

Surveillance Plots:

We used surveillance plots to evaluate the goodness of a KDE. In surveillance plots, we plotted aggregated probability of crime occurrence against %area covered. The x-axis scale has areas sorted from higher probability / higher risk to lower.

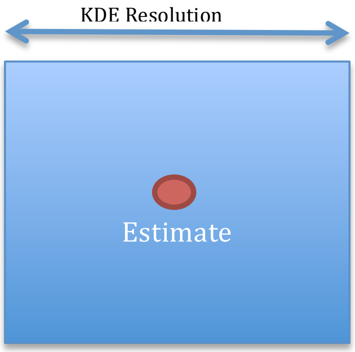
Surveillance plots check how accurate is the KDE in identifying the hotspots by calculating AUC (Area under curve) metric. The mechanism used to validate KDE prediction is discussed below

Following is the surveillance plot for theft cases in May in the city of Chicago.

Below is the approach used to evaluate the goodness of KDE:

1. Using KDE, estimate the theft distribution from the data from Jan to April month. Create a square around each estimate and sort the squares based on their threat levels. The square with the highest threat level would be at the top.

   Square around each estimated point



2. Count how many locations where the theft has occurred in May are captured in each square. Calculate the % of crime captured in each square.

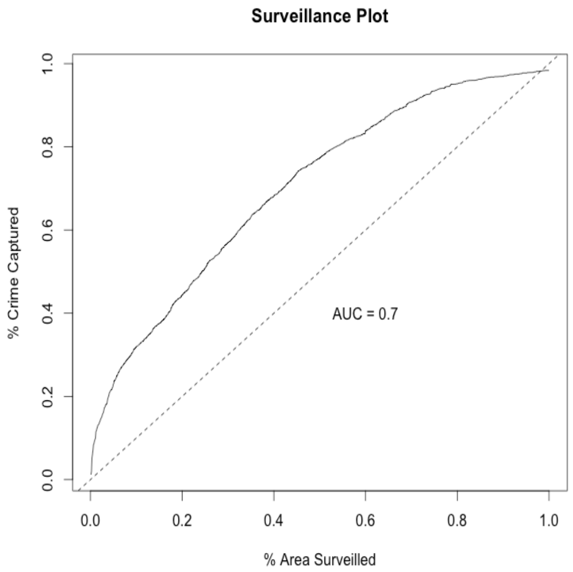
   %Crime captured in a square = No. of crimes captured / Total crimes in May

   % Area of a square = 1/total squares

   Aggregate the % crime captured corresponding to the % of covered area (aggregated area of squares)

3. If the KDE is effective, the squares with high threat level should have more number of captured crimes or more % of captured crimes.

4. **Plot % of aggregated captured crime against the % covered area.**
5. The plot helps us validating the effectiveness of a square / location that is predicted as part of a hotspot in KDE
6. Validate KDE performance by finding out how well we identified the hotspots / locations with high probability of crime occurrence.
7. The AUC (Area under curve) gives us the effectiveness of identifying a particular location as hotspot using KDE techniques. The more the AUC the better the identification of hotspots / KDE.
    - Area under curve is calculated by aggregating areas of each point on x-axis and its average corresponding y –values
    - AUC = Aggregate ((x2-x1) * Average (y1, y2))
8. In the following plot, we can observe that 10% of the city has 38% of theft occurrences. Which tells us that KDE technique has performed well in identifying hotspots of theft for the month of May.



Surveillance Plot: Thefts prediction for May using data from Jan-Apr

In the above case the past four months data is used for identifying the hotspots for theft in the month of May. Surveillance plots as discussed above can validate the goodness of KDE.

Is KDE's performance uniform across all periods?

We have observed that the Crime is of cyclical nature, which can be predicted from past occurrences. But the cyclical nature might be more observed in some specific periods when compared to other periods?
It is similar to saying that KDE can better estimate some specific durations of day, week or month from its corresponding historical values?
Let us check KDE performance in various scenarios using the surveillance plots.

DATA:
2014_THEFT.csv – All the theft incidents occurred in Chicago from Jan 1st 2014 to Dec 4th 2014 (Date, time, location coordinates, location description, incident description, address, arrest made etc. are the important features)

For each dataset, following data **preprocessing activities** are done:

- Step1: Coordinate Ref system of city boundary is in meters. So the location coordinates of our data to be converted to the same CRS (epsg:26971)
- Step2: parse the date variable (ex: 12/04/2014 11:30:00 PM) to extract Hour of day, day of week and month

Step3: Addition of Extracted features to the dataset for further analysis

KDE performance depends on Hour of the day?

In this analysis we tried to understand occurrence of crime on a particular time of the day is better estimated by the corresponding times in the previous days for some hours when compared to others.
Ex: Does KDE estimate of thefts from 4PM to 5PM better than 6AM to 7AM from their corresponding past data?
We also tried to understand if the past 4 months data is better than 8 months data or vice versa

Approach Taken:
For each hour of the day, we performed 3 KDE tests for changing the training and validation data sets.

- Step1: Divide the data in to 3 sets
  - Set1: Jan – April
  - Set2: May – August
  - Set3: September – December

- Step2: Subset Set1, Set2 and Set3 for a particular hour Ex: Hour == 6 gives us the thefts occurred from 6AM to 7AM only.

- Step3: Run KDE for following tests and plot surveillance plot. Since crime data is temporal in nature, we can only run validations sequentially and randomization cannot be done

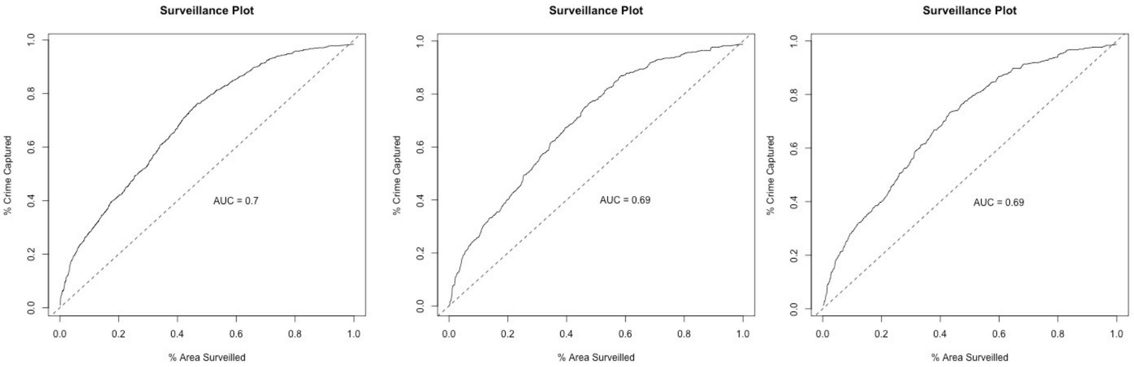
- Test1: KDE on Set1 and Evaluation on Set2
- Test2: KDE on Set2 and Evaluation on Set3
- Test1: KDE on Set1+Set2 and Evaluation on Set3

- Step4: Understand KDE performance for various hours based on AUC values

Analysis:
From the above methodology, 72 surveillance plots are created. 3 plots for every hour with different estimation and evaluation sets

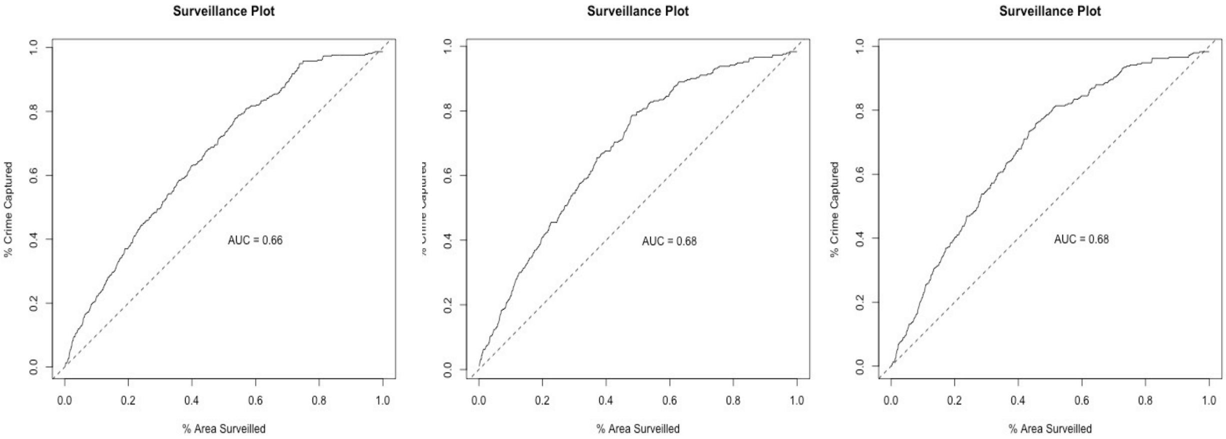
Following are the surveillance plots for 00.00 to 1.00 hour of the day.
For all the three tests, the AUC values are in the same range. For Test2 and Test3, the AUC values are same. Which means that our hotspot identification is same for theft cases at 00.00 to 1.00 hrs when past 4 moths data or past 8 months data considered for KDE estimation.

Surveillance Plots for 12.00 AM theft cases



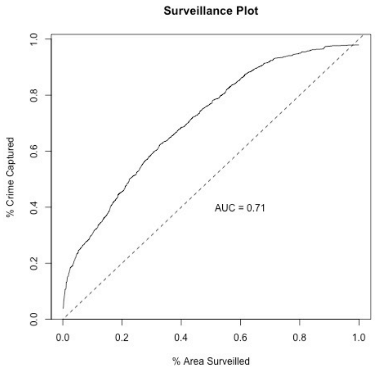
Test1: AUC 0.70 Test2: AUC 0.69 Test3: AUC 0.69

Surveillance Plots for 6.00 AM theft cases

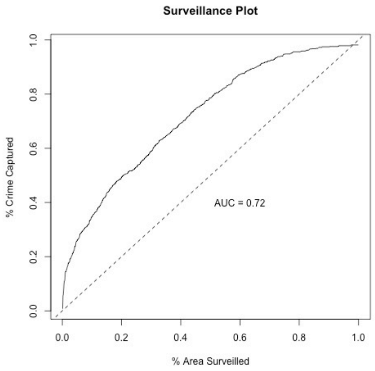


Test1: AUC 0.66 Test1: AUC 0.68 Test1: AUC 0.68

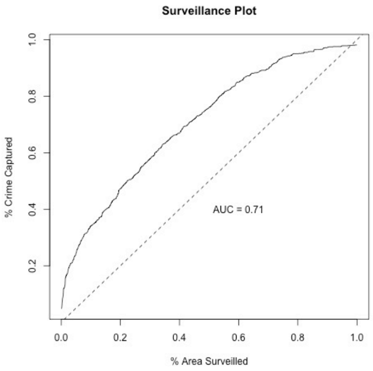
Surveillance Plots for 12.00 PM theft cases



Test1: AUC 0.71

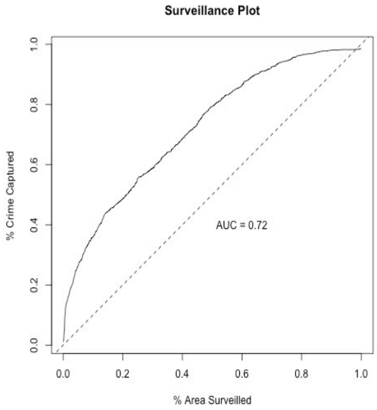


Test2: AUC 0.72

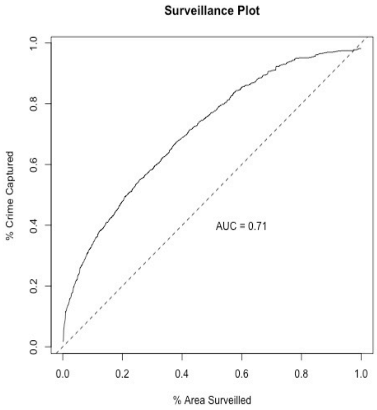


Test3: 0.71

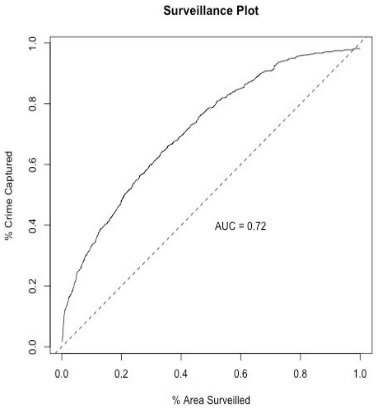
Surveillance Plots for 6.00 PM theft cases



Test1: AUC 0.72



Test2: AUC 0.71



Test3: 0.72

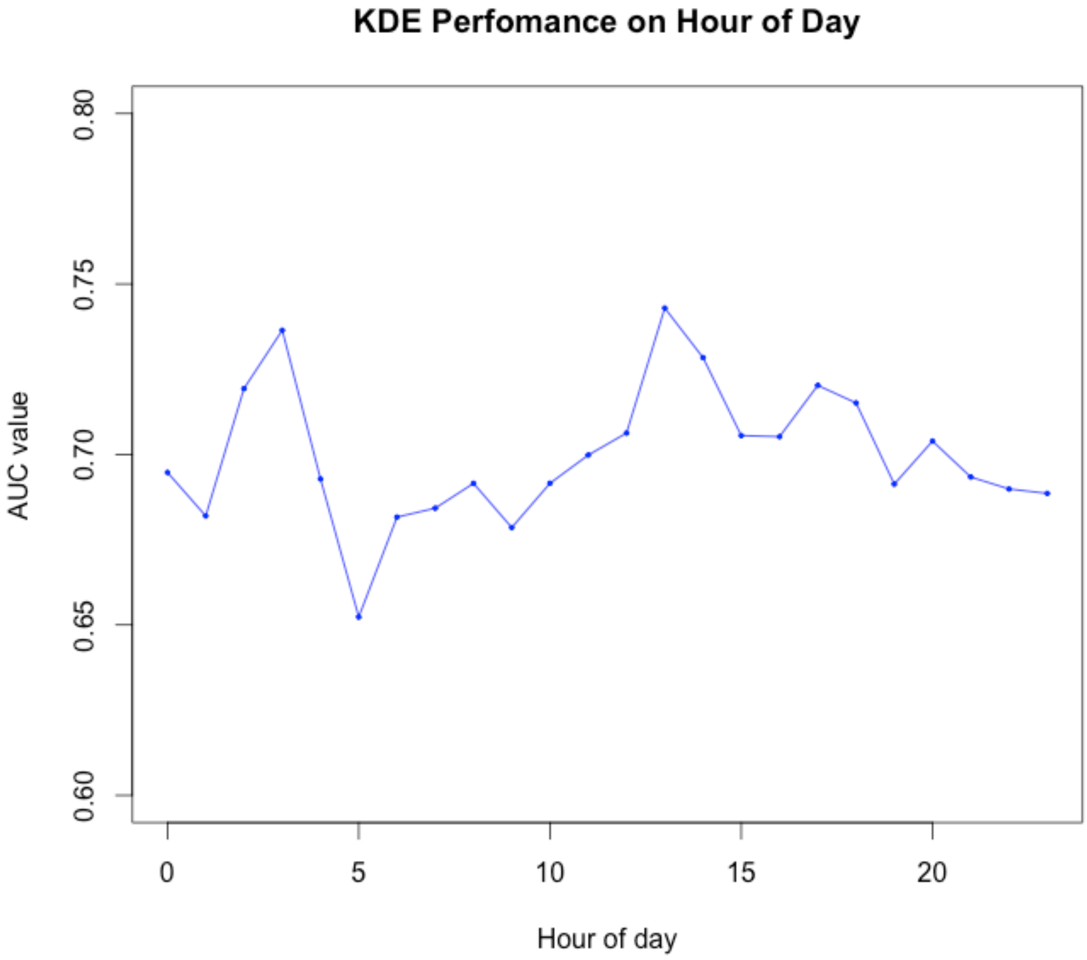
Observations:

From the above plots it is observed that for 12 PM and 6 PM the KDE estimation technique worked better when compared to 12 AM and 6 AM.

There is no considerable change in estimation effectiveness with an increase in training data from past 4 months to 8 months

Hr	Test1	Test2	Test3
12:00 AM	0.70	0.69	0.69
6:00 AM	0.66	0.68	0.68
12:00 PM	0.71	0.72	0.71
6:00 PM	0.72	0.71	0.72

Plotting Test3-AUC values (KDE's performances) on various hours of the day:



X-Axis Scale: 0 is 12 AM and 23 is 11PM

From the above map, it is quite evident that KDE is better for some hours of the day 2 AM, 3 AM, 12PM to 6PM when compared to other hours of the day. KDE performance is the least at 5AM to 6 AM window.

Interpretation & Recommendations:

- For mid day hours 11 AM to 6 PM the prediction is very effective. That means the distribution of crime is better correlated with distribution in the same hours of previous days
- There is not much difference in prediction effectiveness for 4 month past data and 8 months past data.
- From the surveillance plots, it can be observed that by covering 1.5% of the total city almost 40% of the thefts can be handled.

- From the hour-hour analysis we understood that during day hours and 2AM to 4AM, the probability of crime reoccurrence in the same places is high. The police can take these periods into account and focus on hotspots to get better results

Does KDE perform better on some days of week? **(Ex: Does the KDE estimation is better for Saturdays predicted from previous Saturdays when compared to Sundays predicted from previous Sundays?)**

Approach Taken:
For each day of the week (Sunday to Saturday), we performed 3 KDE tests by changing the training and validation data sets.

- Step1: Divide the data in to 3 sets
  - Set1: Jan – April
  - Set2: May – August
  - Set3: September – December

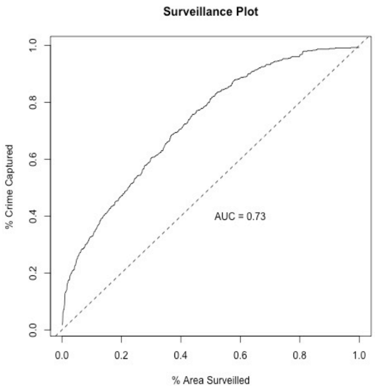
- Step2: Subset Set1, Set2 and Set3 for a particular hour Ex: Day == 1 (Sunday) gives us the thefts occurred Sundays only.

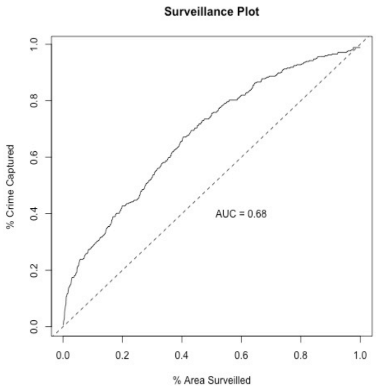
- Step3: Run KDE for following tests and plot surveillance plot. Since crime data is temporal in nature, we can only run validations sequentially and randomization cannot be done
  - Test1: KDE on Set1 and Evaluation on Set2
  - Test2: KDE on Set2 and Evaluation on Set3
  - Test1: KDE on Set1+Set2 and Evaluation on Set3

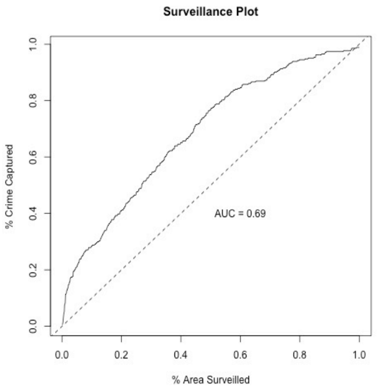
Step4: Understand KDE performance for various days based on AUC values

Analysis:
21 plots are generated 3tests for 7 days. 3 plots for every day with different estimation and evaluation sets. Some of the surveillance plots are shown below.

Surveillance Plots for Thefts on Sundays



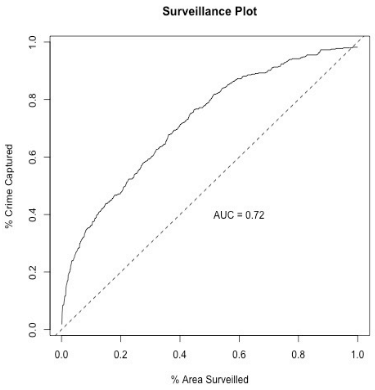


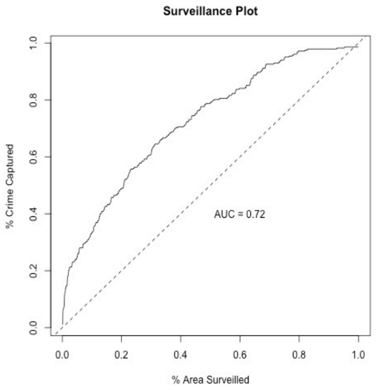


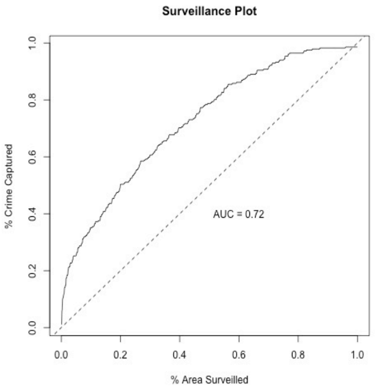
Test1: AUC 0.73 TEST2: AUC 0.68 Test3: AUC 0.69

From Sunday surveillance plots we can observe that Jan-Apr data effectively predicted the theft distribution of May-August. But prediction for Sept – Nov from the past data hasn't been as effective.

Surveillance Plots for thefts on Mondays

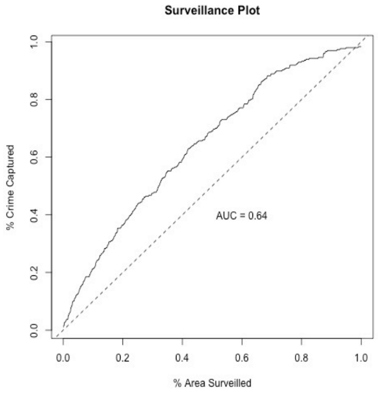




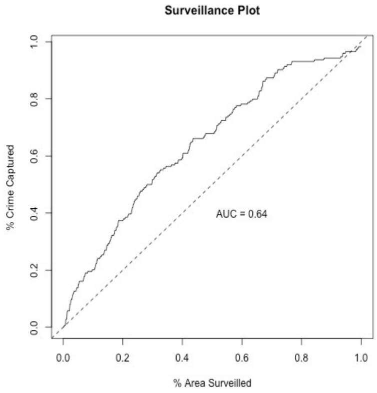


Test1: AUC 0.72 TEST2: AUC 0.72 Test3: AUC 0.72

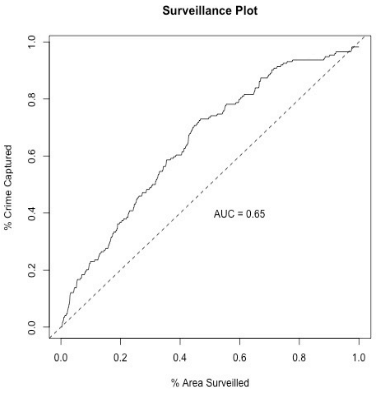
Surveillance Plots for thefts on Thursdays



Test1: AUC 0.64

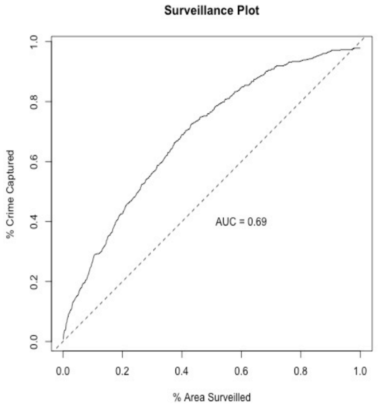


TEST2: AUC 0.64

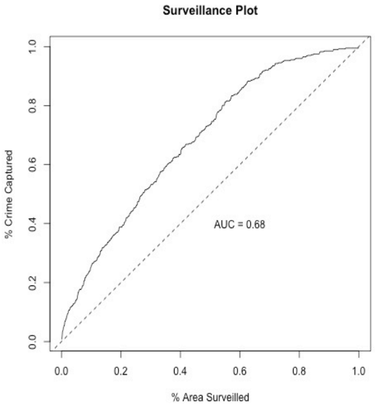


Test3: AUC 0.65

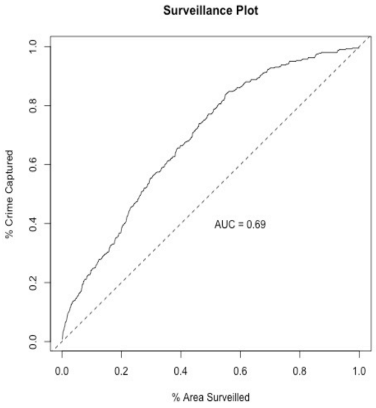
Surveillance Plots for thefts on Saturdays



Test1: AUC 0.69



TEST2: AUC 0.68



Test3: AUC 0.69

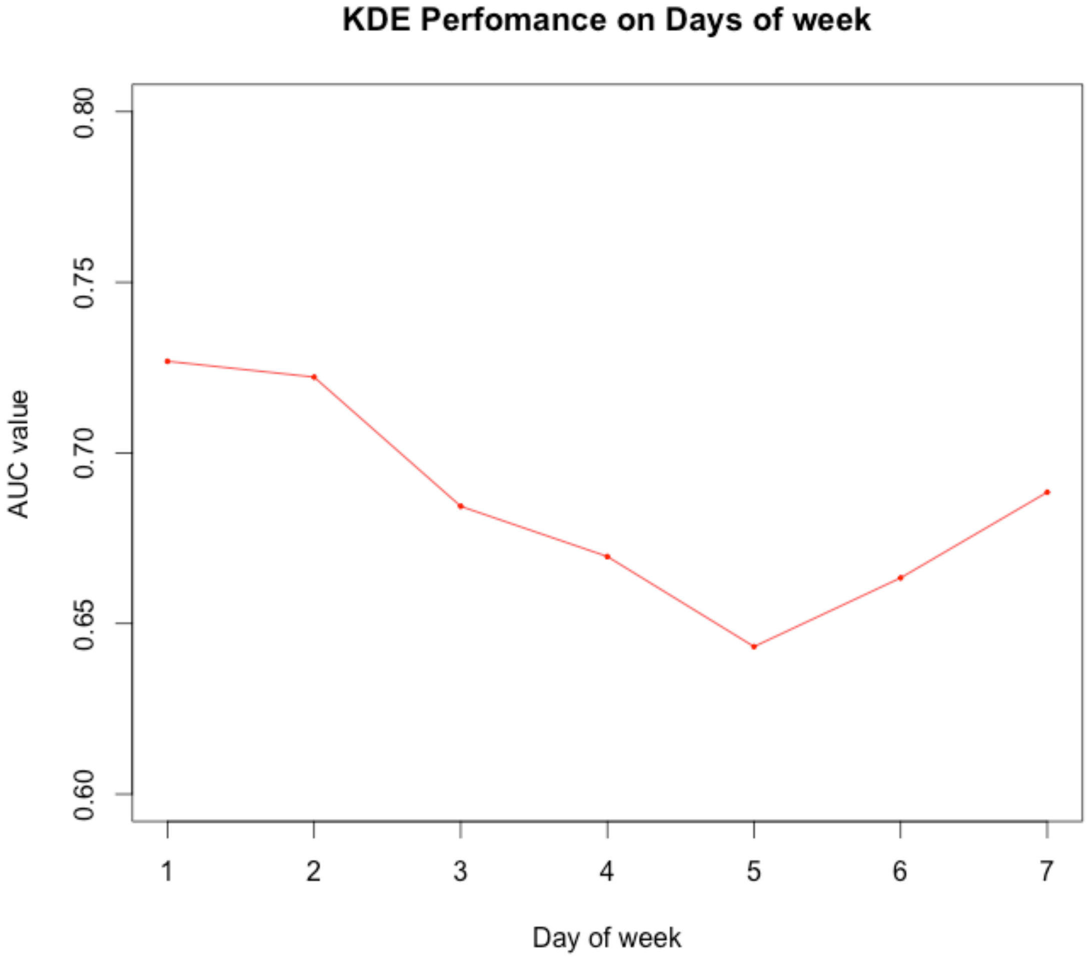
Below the table of AUC values for different days of week in 3 test scenarios.

Day	Test1	Test2	Test3
Sunday	0.73	0.68	0.69
Monday	0.72	0.72	0.72
Tuesday	0.68	0.73	0.73
Wednesday	0.67	0.67	0.69
Thursday	0.64	0.64	0.65
Friday	0.66	0.68	0.69
Saturday	0.69	0.68	0.69

- Test1: KDE on Set1 and Evaluation on Set2
- Test2: KDE on Set2 and Evaluation on Set3
- Test1: KDE on Set1+Set2 and Evaluation on Set3

AND

- Set1: Jan – April
- Set2: May – August
- Set3: September – December



X-axis: 1 – Sunday and 7 - Saturday
Comparing KDEs across the days of week (Using Test1 AUC values)

Interpretation & Recommendations:

- There is a clear pattern how KDE is performing way better on Sunday to Tuesday when compared to Wednesday to Friday, hitting lowest at Thursday and gradually improving there on.
- It is an indication that theft patterns on Saturdays, Sundays to Tuesdays are more cyclical in nature than the other days.

- 4 months data is as effective as 8 months data for predicting thefts on various days. But in some cases predictions on May to Aug thefts are better and other cases predictions for Sept-Dec are better. There is no particular pattern, we try to understand it while analyzing on monthly data.
- Same locations and probably similar type of thefts are being occurred every Sundays, Mondays. The further analysis on type of thefts at these locations would throw more insights on whether the same group of people are involved with this. In which case identifying those groups would effectively result in reduced number of thefts.

Does KDE perform better on some months? (Ex: Does the KDE estimation is better for May predicted from March, April data when compared to October predicted from August, September data?)

Approach Taken:

For each month (Feb to Dec), we performed 2 KDE tests by changing the training and validation data sets. We have considered prediction for Dec even though only 604 theft cases (for first 4 days) are available.

- Step1: Divide the data in to 3 sets, for every month 'm'
  - Set1: 1 Month past data (Ex: for May, April data is considered)
  - Set2: Total available past data (Ex: for May, Jan to April considered)
  - Set3: Month under evaluation (Ex: May)

- Step2: Run KDE for following tests and generate surveillance plots. Since crime data is temporal in nature, we can only run validations sequentially and randomization cannot be done
  - Test1: KDE on Set1 and Evaluation on Set3 (Set1 - Only 1 past month data)
  - Test2: KDE on Set2 and Evaluation on Set3 (Set2 - Available past data)

- Step3: Understand KDE performance for various months based on AUC values

Analysis:

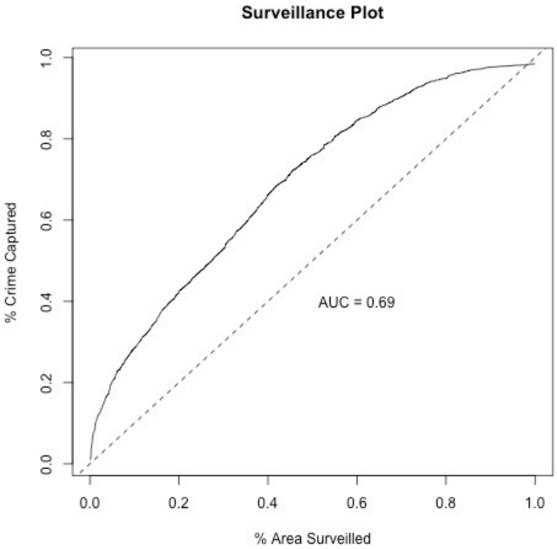
Months from February to December considered for evaluation. The KDE cannot be run for January, since there is no past data available for it. Two plots for each month from February to December generated.

We tried to understand two things

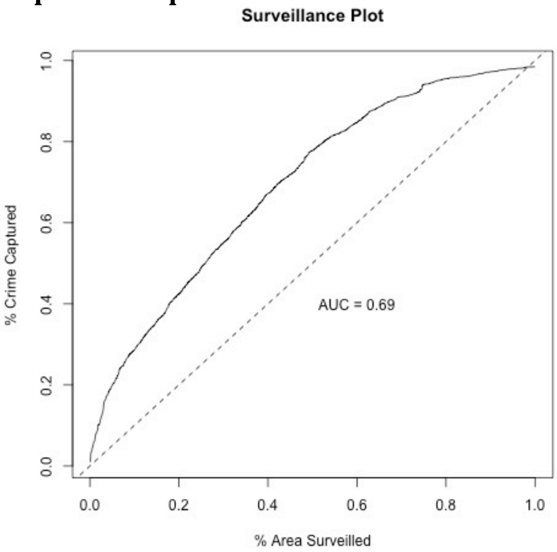
1. Can better prediction be obtained by considering total past available data for a particular month when compared to estimating using the immediate past month.
2. KDE performs better on some months when compared to others

Surveillance plots of April, June and Nov are shown below

Surveillance Plots for April Theft prediction

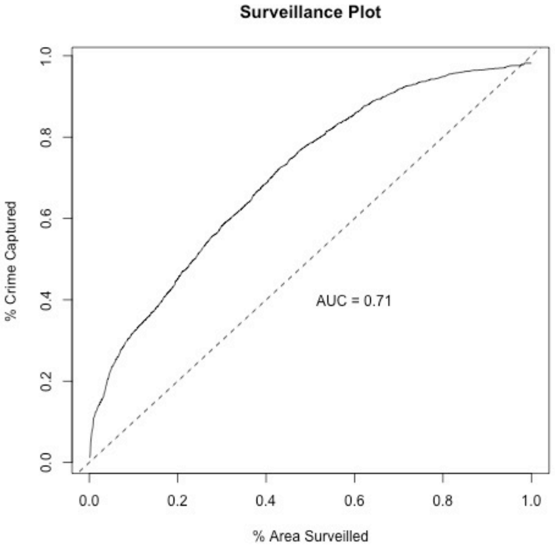


Test1: AUC 0.69

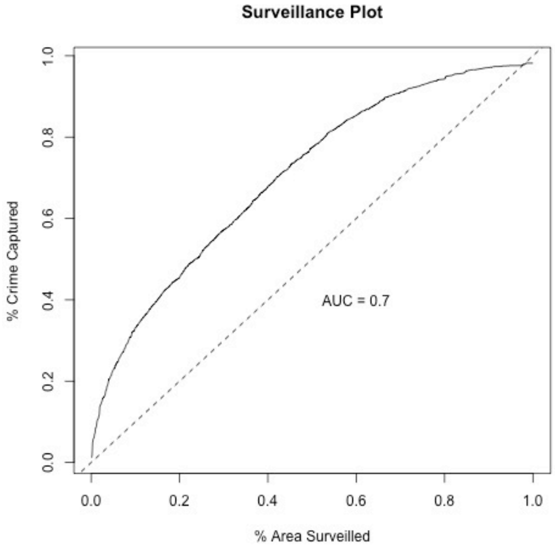


Test2: AUC 0.69

Surveillance Plots for June Theft prediction

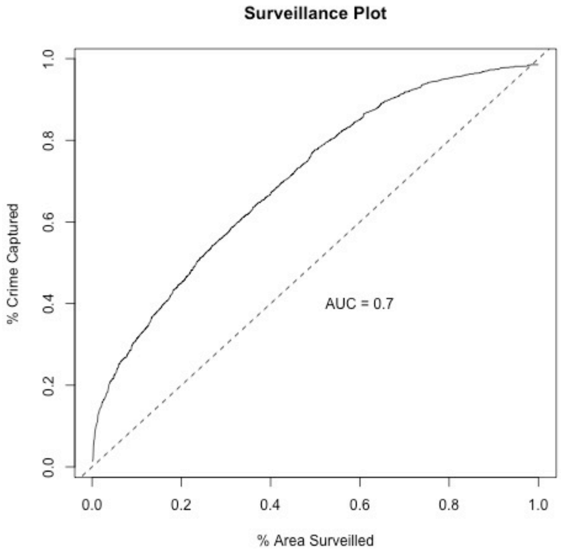


Test1: AUC 0.71

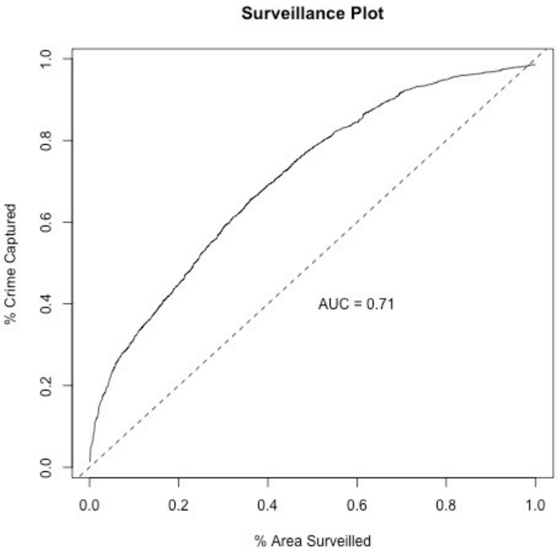


Test2: AUC 0.7

Surveillance Plots for Nov theft predictions



Test1: AUC 0.7



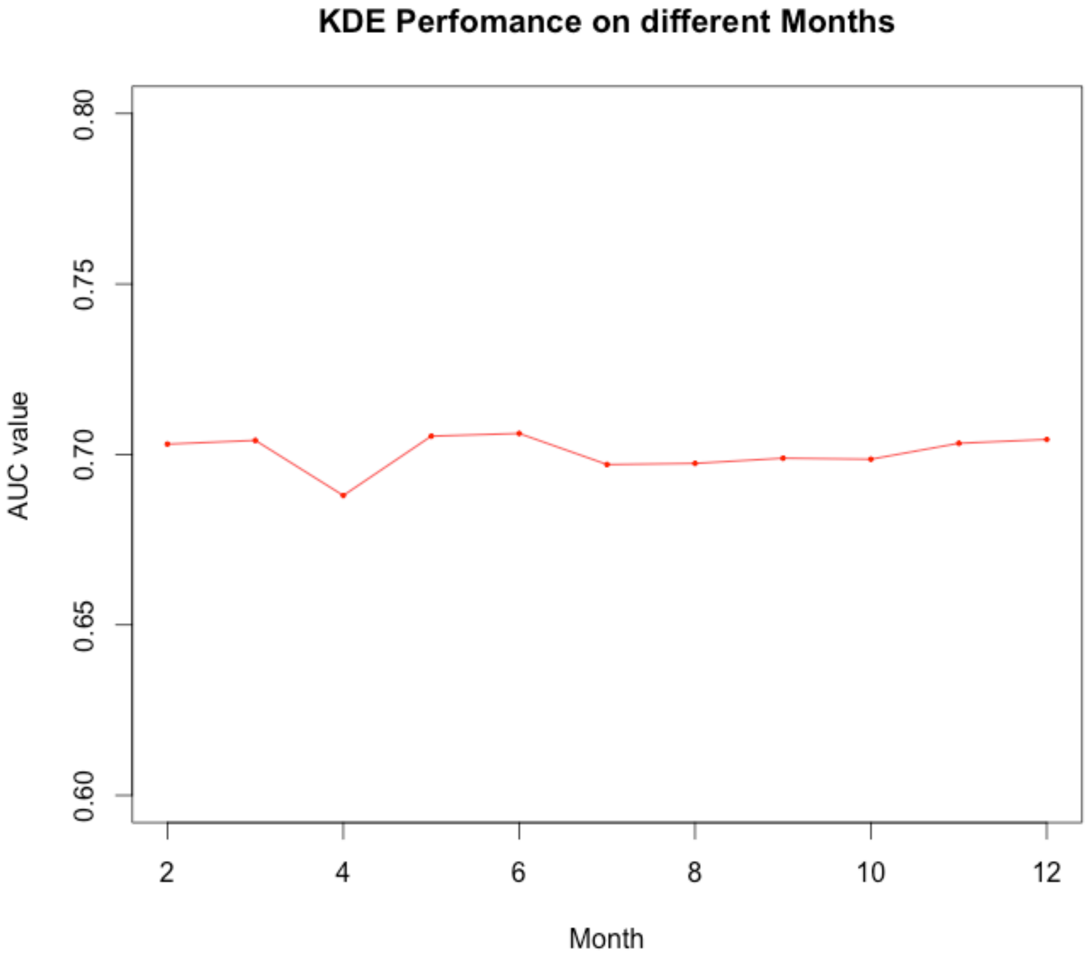
Test2: AUC 0.71

Following is the table with AUC values for Different Months

Month	Test1	Test2
Feb	0.70	0.70
Mar	0.70	0.70
Apr	0.69	0.69
May	0.71	0.71
Jun	0.71	0.70
Jul	0.70	0.70
Aug	0.70	0.70
Sep	0.70	0.70
Oct	0.70	0.70
Nov	0.70	0.71
Dec	0.70	0.70

It is evident from the above table that KDE performance is same if the training data is immediate past month data or multiple past months. It shows a strong dependence of current month's theft distribution on the previous month's theft distribution.

Plotting Test1 AUC values of different months



X-axis: 2 – February and 12 - December
Comparing KDEs across the Months (Using Test1 AUC values)

Interpretation & Recommendations:

- Unlike Days and Hours, KDE performance is consistent for prediction of theft for different months
- There appears strong dependence of current month theft distribution on previous month's distribution. It means that same locations are prone to thefts month after month. The AUC value ~ 0.7 shows KDE can work better for identifying the hotspots for thefts every month.
- The encouraging news is that, the locations are more or less same where the probability of thefts is higher. Strategizing policing activities especially on those locations could yield better results.
- As mentioned earlier, from the surveillance plots, 1.5% of the area has 40% risk of total theft cases