Case Study 4 (CS4): Non-Linear Crime Models

"On my honor, I pledge that I have neither given nor received help on this assignment."
Name: Sharath Chand PV
ID: vp4pa
Date: 10-23-2015

Introduction:

In the previous study we developed linear models for predicting crime using KDE estimate and spatial factors as predictors. We evaluated its effectiveness in identifying hotspots using surveillance plots.
The hypothesis was theft occurrence is correlated with proximity to certain locations and it varies linearly with the distance. But in reality the linearity assumption might not hold good. In this report we tried to explore and test non-linear relationships between various spatial factors and risk of theft around them. We focused our study only on spatial factors in this study, even though many other factors could play a role in it and have non-linear correlation with the risk.

1. Spatial Factors with Non-linear effect on Theft

We discussed and tested possible affects of various spatial factors on theft in our previous study. We assumed that the risk vary linearly with the distance from the hypothesized points. Since linear assumption is unlikely to hold good in reality, we will assume non-linear relationship and test whether the risk prediction would improve over the models based on linear assumptions.

Our hypothesis is that following spatial factors have non-linear effect on theft risk.

Major Streets:

The locations closer to the major streets are more likely to be prone to thefts. Since many small businesses develop around major streets, the locations would get crowded during business hours, which is conducive for thefts.

The theft risk might not be uniform as the locations get farther from the major streets. We assume that the risk might be quite higher near the streets and decreases steeply when we slightly move away from the street surroundings implies non-linear association.
Following are hypothesized reasons for high theft risk around streets

a. Major activities (shops, small business establishments, restaurants etc.) happen just next to the streets and as we move a bit farther the activity drops steeply in case of most of the streets
b. Easy for offenders to disguise themselves on streets
c. A lot of exit options after the crime is committed
d. Transportation and coordination gets easy for offending groups in locations around the streets

CTA rail stations:
CTA has 1,356 rail cars operate across the city and cover 224.1 miles of track. CTA trains make about 2,250 trips each day and serve146 stations. Given the number of people use train services we hypothesize that the CTA rail stations, being major crowded places in Chicago, would have higher risk of theft around them and the risk drops non-linearly as moved away from those stations.

References

1. http://www.transitchicago.com/about/facts.aspx

Non-Linear Models for Crime Prediction:

We have hypothesized that distance decay effect will not have uniform association with theft. We will test our hypothesis by building and evaluating non-linear crime models

Approach:
Our goal is to test the non-linear correlation of a particular spatial factor (Ex: Proximity to major street lines) with the probability of occurrence of theft.

We have considered following spatial factors

1. Proximity to Major street lines
2. Proximity to CTA railway stations (CTA: Chicago Transit Authority)
3. Proximity to Police Stations
4. Proximity to parks

To understand whether the above correlation is non-linear or not we have build Support Vector Machines model and Logistic regression model on the same data. We compared both the models using surveillance plots to test whether our hypothesis has improved the prediction accuracy.

Step-by-step Methodology:

Predictors (Base data) – March theft data
Response (Train data) – April theft data
Validation set (Test data) – May theft data

Following is the step-by-step approach taken to build models

1. Train data preparation:
    a. In our train data (April theft) we have got only those locations (coordinates) where the theft has occurred, called as positive observations. To train our classification model we have added some locations where the theft has not occurred, called as negative observations, to our train data.

    b. Add a response variable, 1 for positive observation and 0 for negative observation
2. Adding Spatial factors to train data:
    a. Extracted point coordinates of Major Streets, CTA rail stations, police stations, parks from shape files downloaded from Chicago data portal
    b. Calculated minimum distance from each point of *training data* to identified factors (proximity to Major street, CTA rail station, police station, park)
    c. Add those minimum distances to training data as spatial factors
3. Add KDE estimate to the training data. Get crime density for each training point based on base data (March theft) records
4. Build Linear and Non-linear models, Response as response variable and spatial factors and KDE estimate as predictors.
    a. Probabilities should be obtained instead hard classification to crime or non-crime
    b. Probabilities are required to evaluate the effectiveness using surveillance plots
5. Predict response on Validation data (May), using the fitted model and predictors from April
    a. Prepare test data – Predictors from April thefts
    b. Predict crime for prediction points using the fitted SVM (non-linear)
    c. Predict using Logistic regression model (linear model)
6. Compare both the models by validating them on ground truths (May thefts)

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)

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

Shape Files:
To extract spatial factors, following shape files are downloaded from https://data.cityofchicago.org Chicago data portal

1. Major_20Streets (Lines shape file)
    a. All the points used to define the major street lines are extracted using slot() function. Total 16000 points extracted
    b. 2000 points sampled from which minimum distance to prediction points calculated
2. Parks_Aug2012 (Polygon shape file)

3. PoliceStationsDec2012 (Points shape file)
4. CTA_Stations (Points shape file)

Analysis:

We have built SVM (3 kernel functions) and logistic regression models on March theft data as predictors and April data as response.

To check non-linear association, we have used 3 different kernels for SVM

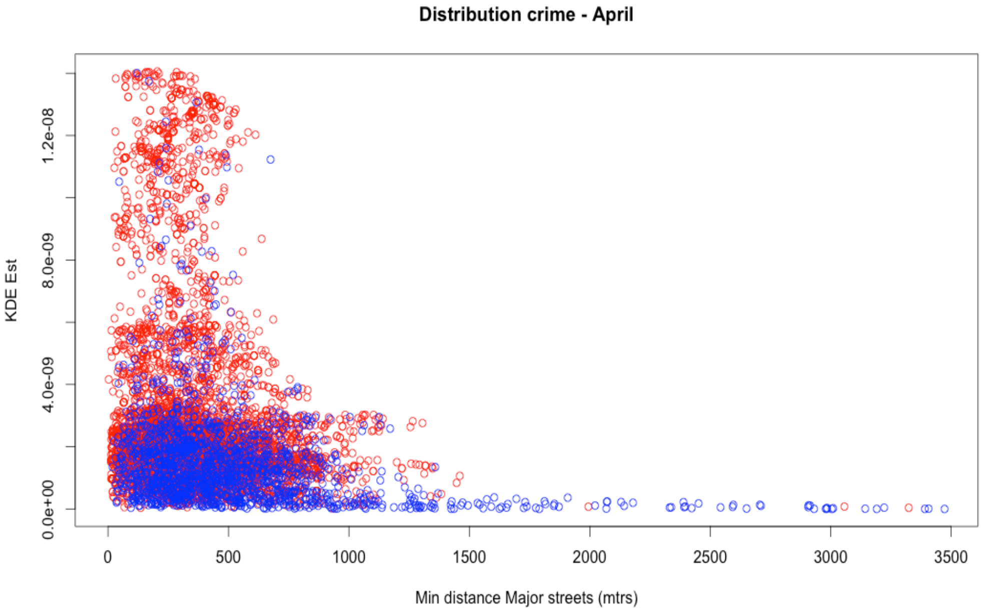
- Linear kernel
- Polynomial kernel
- Radial basis

Initial Plots:

To understand the distribution of crime and non-crime points in the feature space we plotted some basic plots
The following plot is between KDE estimate and proximity to Major street lines for April theft data. We can observe that probability of theft is expected to be high in the locations close to major streets

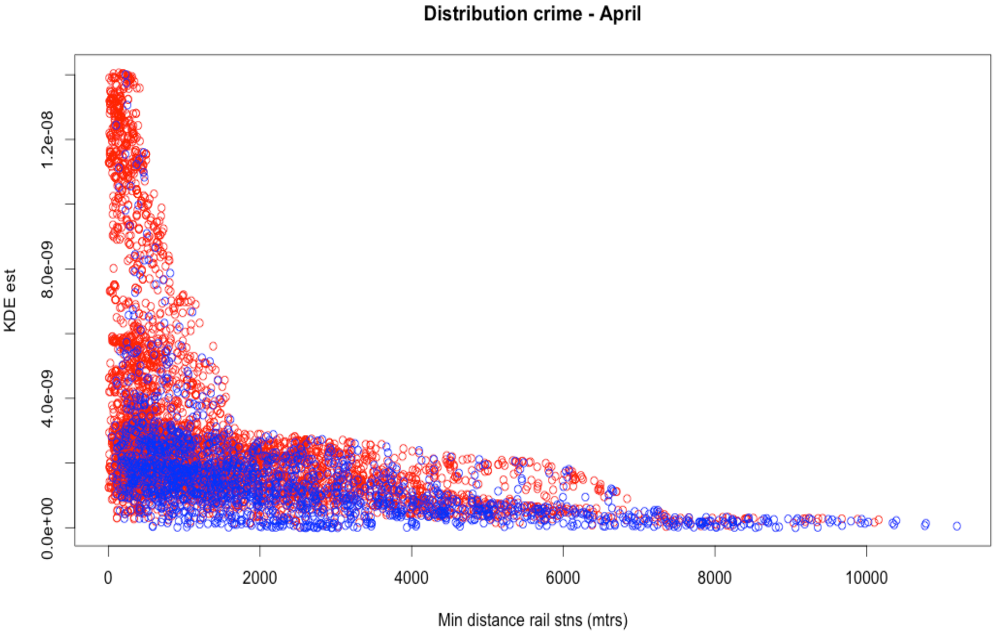
We can also see that theft risk is higher with in 1 km distance from major street lines and drops steeply beyond that

Proximity to MAJOR Streets Vs KDE



Red – Crime
Blue – Non-Crime

Proximity to CTA Train Station Vs KDE



Red – Crime
Blue – Non-Crime

The above plot suggests the theft risk drops steeply as we move beyond 1500 meters from CTA train stations

Model Building and Evaluation:

For SVM classification, we have used e1701 library in R

Response:
Theft occurred - 1
Not occurred - 0

Predictors:
KDE = Probability of theft occurrence at the location
ST = Minimum distance from major streets in meters
PR = Minimum distance from Park in meters
PS = Minimum distance from Police station in meters
RS = Minimum distance from CTA rail station in meters

Initially, to understand correlation clearly we have developed models with 2 different sets of predictors

- Spatial predictors alone
- Spatial predictors + KDE

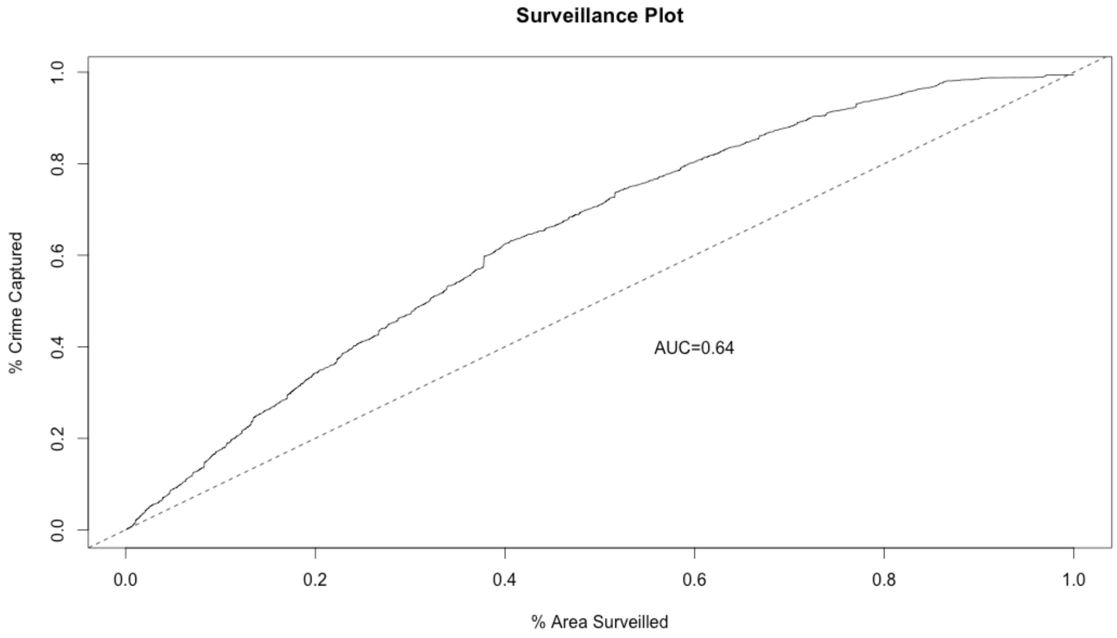
Models with Spatial Predictors alone:

Predictors used:
ST = Minimum distance from major streets in meters
PR = Minimum distance from Park in meters
PS = Minimum distance from Police station in meters
RS = Minimum distance from CTA rail station in meters

Linear Kernel – Spatial Predictors alone

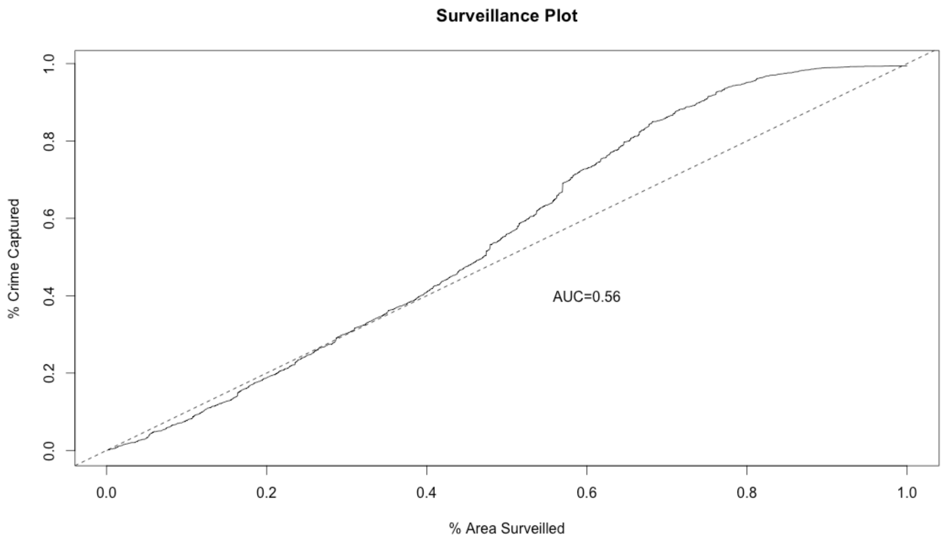
Type = C-classification
Kernel = linear

Following surveillance plot suggests there is strong correlation between location proximity to hypothesized spatial factors and theft risk

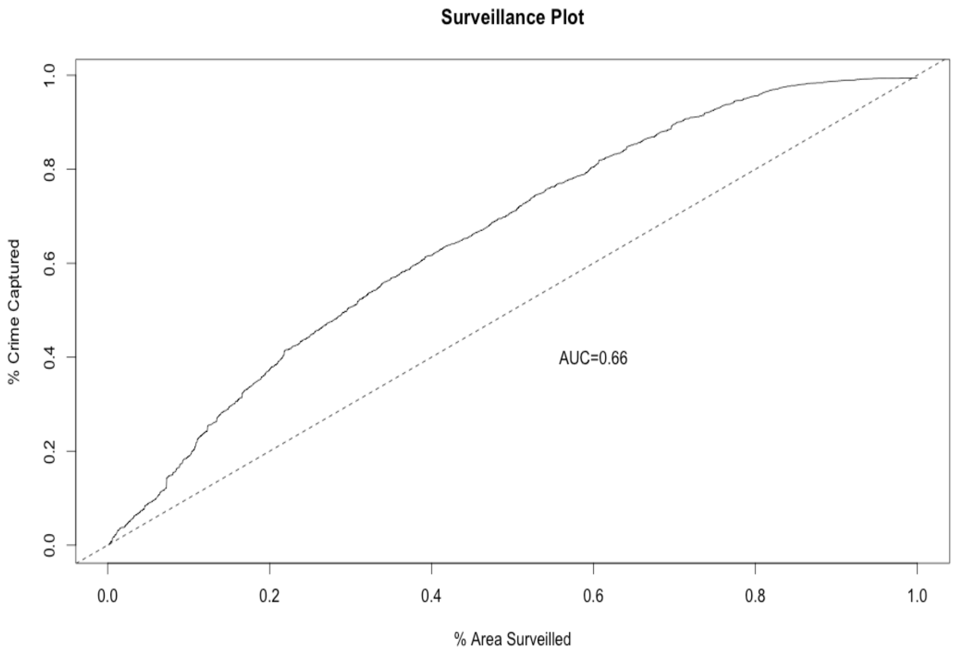


Polynomial Kernels – Spatial Predictors alone

Type = C-classification
Kernel = polynomial
Degree = 2
Gamma = 0.25

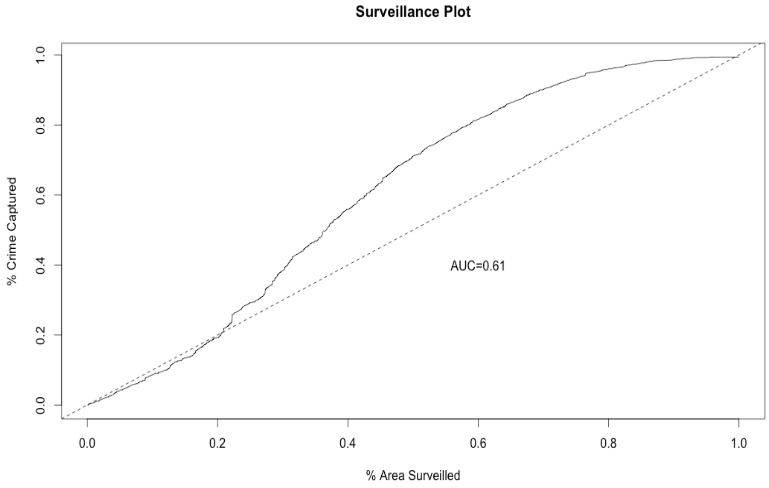


Kernel = polynomial
Degree = 3
Gamma = 0.25



Radial basis Kernel – Spatial factors alone

Type = C-classification
Kernel = radial basis
Gamma = 0.25



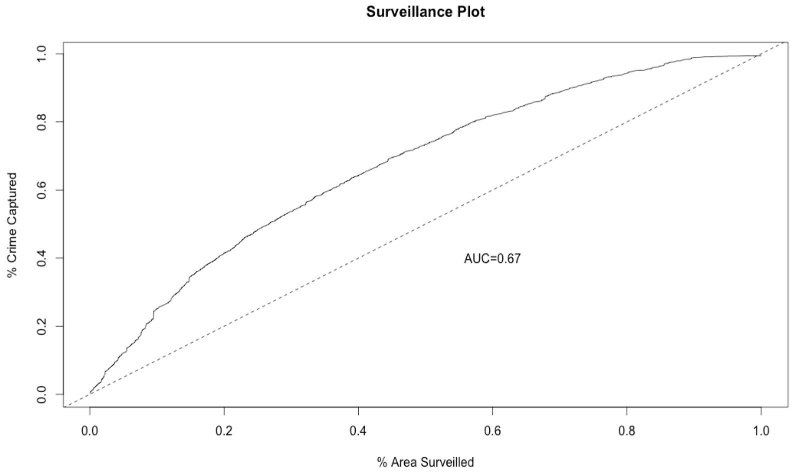
Logistic Regression with Spatial predictors alone

All the spatial factors hypothesized seem to be significant

Log of (probability of a location has theft / probability of location doesn't have theft)

= 2.09 - 9.019e-04* ST - 1.474e-04 * PR - 6.633e-05* PS - 2.142e-04 * RS

ST = Minimum distance from major streets in meters
PR = Minimum distance from Park in meters
PS = Minimum distance from Police station in meters
RS = Minimum distance from CTA rail station in meters

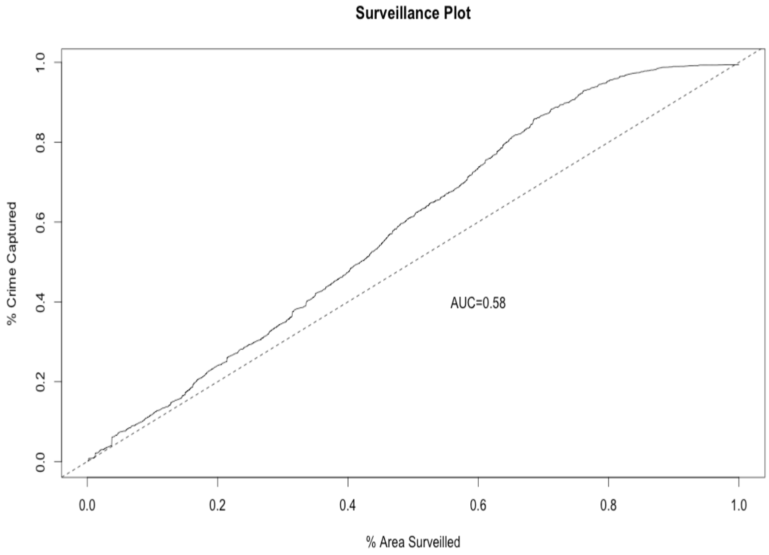


Predictors = Spatial factors + KDE estimate

Now, let us add KDE estimate to our predictors list and repeat the same process of building non-linear and linear models for the comparison

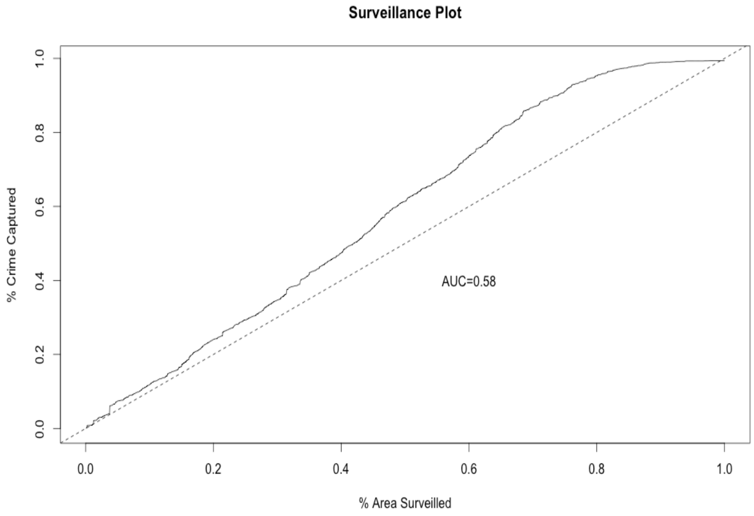
Linear Kernel – Spatial factors + KDE estimate as predictors

Type = C-classification
Kernel = linear

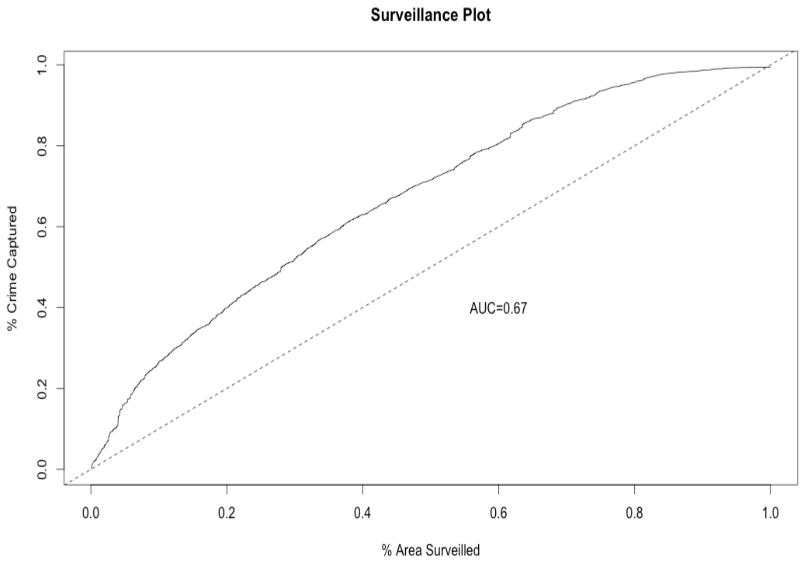


Polynomial Kernels – Spatial factors + KDE estimate as predictors

SVM-Type: C-classification
SVM-Kernel: polynomial
Cost: 50
Degree: 2
Gamma: 0.2

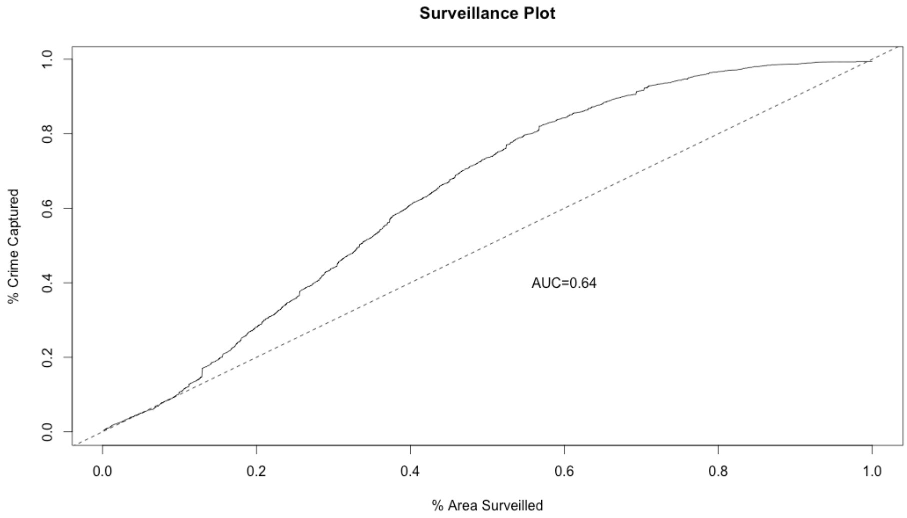


SVM-Type:	C-classification
SVM-Kernel:	polynomial
Cost:	50
Degree:	3
Gamma:	0.2



Radial Basis Kernel - Spatial factors + KDE estimate as predictors

Cost = 50 and Gamma = 0.2



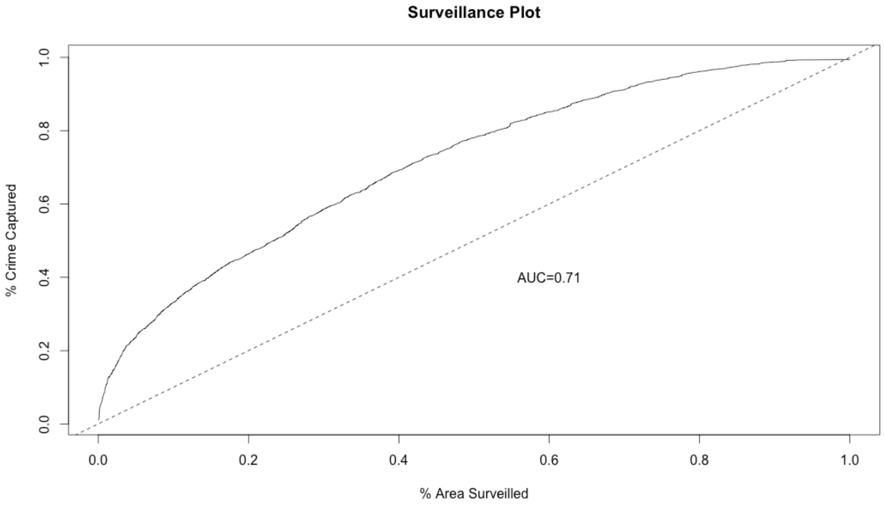
Logistic Regression - Spatial factors + KDE estimate as predictors

KDE and all the spatial factors hypothesized except minimum distance from parks seem to be significant.

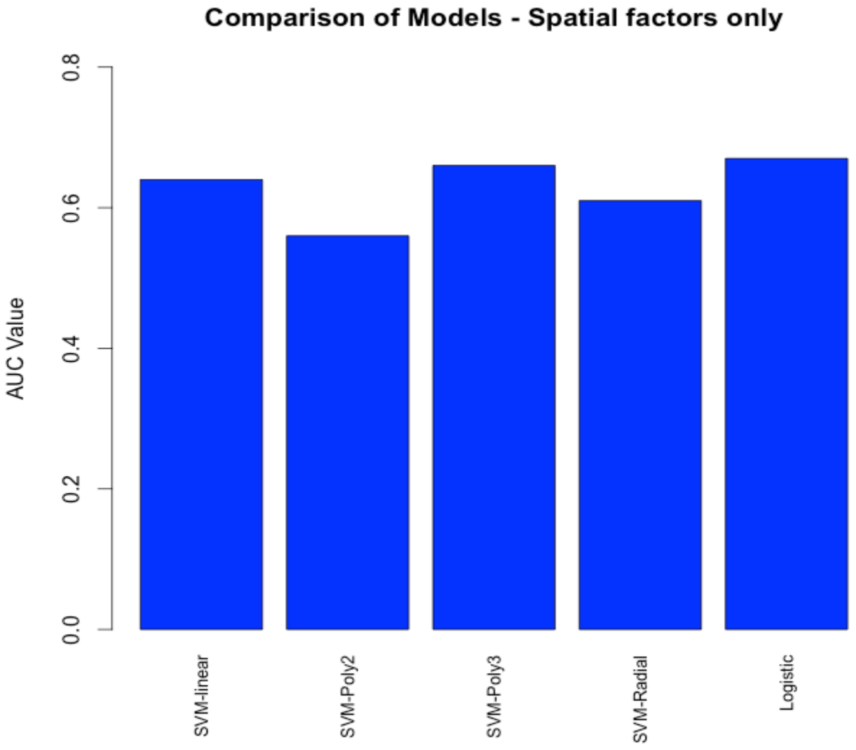
Log of (probability of a location has theft / probability of location doesn't have theft)

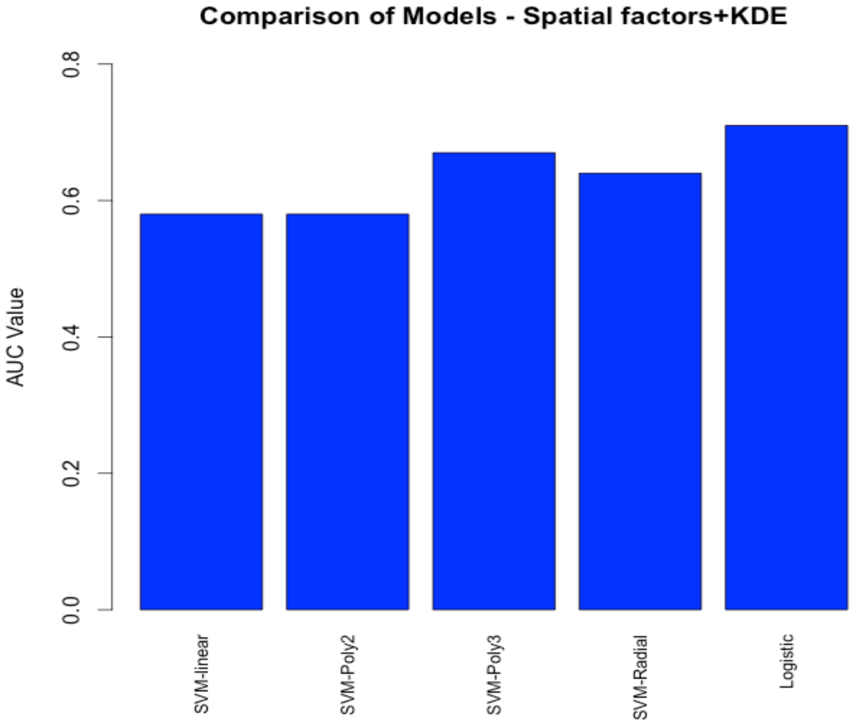
= 1.061 + 2.921e+08 *KDE - 7.155e-04* ST - 5.369e-05* PS - 8.808e-05 * RS

KDE = Probability of theft occurrence at the location
ST = Minimum distance from major streets in meters
PR = Minimum distance from Park in meters
PS = Minimum distance from Police station in meters
RS = Minimum distance from CTA rail station in meters
From the above relationship, we can interpret following



Comparison SVMs (Linear, Polynomial, Radial), Logistic:





From the above surveillance plots we can infer the following

- When KDE is added to the predictors, Logistic regression is the most effective of all the models. It suggests that KDE being the most significant predictor seem to have linear relationship with the crime occurrence.
- Logistic regression, SVM with 3rd Degree polynomial kernel have the best AUC values for models with spatial factors alone as predictors.
- The AUC value increased from 0.56 to 0.66 when degree of polynomial kernel changed from 2 to 3. The large difference in AUC values for polynomial kernels suggests non-linear hypothesis cannot be rejected and it should be further investigated with more number of factors.
- 3rd degree polynomial SVM has performed as effective as logistic regression shows that non-linear models can be used in those cases where crime occurrence doesn't depend much on earlier occurrences.
- SVMs are typically effective in higher dimensions and they perform best when the large number of predictors can almost determine the class. But in our case we are more concerned with likelihood of theft occurrence instead of hard classification of crime or not crime, for which logistic regression is more suitable.
- We have only considered 4 spatial factors. But the crime can depend on large number of factors (spatial, demographic, temporal, seasonal etc.). As discussed above non-linear models could be more useful if full modeling of crime needs to be done with all possible predictors of high dimensional data