Case Study 3 (CS3): 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: 10-10-2015

Introduction:

In the previous studies we have studied crime prediction using KDE techniques. We also studied its effectiveness in identifying hotspots using surveillance plots. In this report we tried to understand why some locations are high-risk areas of crime, what factors or attributes might influence occurrence of crime in those areas. We focused our study on spatial factors in this study, even though multitude of other factors could play a role in it.

1. Spatial Factors:

Various spatial factors influence the crime occurrence, for example locations close to bars and liquor establishments are more prone to robberies (*Tilley et al., 2005; Wright and Decker, 1997*). Different spatial factors affect different types of crimes. Locations close to rail stations and busy areas could be more prone to thefts than assaults. Conversely, locations close to bars and construction sites could be more prone to assaults than thefts. We studied spatial factors influencing thefts in this report.

Intuitively, following are the spatial factors that could influence occurrence of theft in a city space.

Rail/Metro Stations:

Locations proximity to metro stations could have higher theft rates, which means more theft cases can be observed in locations close to these stations when compared to locations that are far. These stations are likely to be more crowded, it could be convenient for pick-pocketers and other offenders to steal. Following features of rail/metro stations could be conducive to offending

a. People frequent for a specific purpose
b. Heavily crowded due to small business developments around stations
c. Unmanned parking facility (auto thefts)
d. Large number of unmanned vehicles and availability of many escape routes could make these locations attractive for auto thieves

Malls / Shopping Centers:

Locations closer to malls and shopping centers might also have higher rate of theft cases. There can be some temporal factors such as weekends and holidays that could further influence theft cases. Following attributes of these locations could be drivers of theft

a. Busy places

b. People with distracted attention
c. Night time entertainment venues, bars/night clubs especially when clustered in a single place could be more prone to thefts

Police Stations:
Locations closer to police stations should have less theft activity when compared to other farther locations. In general, there could be increased activity of policing around those locations and offenders don't tend to get attracted to those locations. But in case of police stations, which are in the city centers, other spatial factors such as proximity to large shopping center could dominate the effect of police station

Parks:
Parks and locations close to parks could be hotspots for theft. The above discussed factors such as crowded places and distracted people could be found conducive for the offenders

Landmarks:
Landmarks if very popular attraction points could have tourists and visitors, which could be easy targets for offenders. Our hypothesis that places around landmarks could have higher probability of theft occurrence.

Convenience Stores:
Locations proximity to convenience stores could again be a factor in influencing theft. Since convenience stores are more likely to be closer to residential spaces and the most frequent places for people, they could attract potential thieves.

References
1. Mapping the Spatial Influence of Crime Correlates: A Comparison of Operationalization Schemes and Implications for Crime Analysis and Criminal Justice Practice. Joel M. Caplan, Rutgers University. *Cityscape: A Journal of Policy Development and Research • Volume 13, Number 3 • 2011*
2. Spatial Analyses of Crime - by Luc Anselin, Jacqueline Cohen, David Cook, Wilpen Gorr, and George Tita
3. Predicting Crime Using Twitter and Kernel Density Estimation. Matthew S. Gerbera

Magnitude of correlation follows a distance decay effect

Linear Models for Crime Prediction:

We have hypothesized various spatial factors that could help us identify a location as a potential hotspot for thefts. We will test our hypothesis on the theft cases in Chicago.

Approach:
Our goal is to understand the correlation of a particular spatial factor (Ex: Proximity to police station) with the probability of occurrence of theft. Which means we will study if the locations closer to police stations have higher or lower probability of theft occurrence. We have considered following spatial factors

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

To understand correlation of above factors with theft occurrence we have built logistic regression model with occurrence of theft as a response variable and kde estimate and above spatial factors as predictors.

Crime prediction using Logistic regression model:
Predictors – January theft data
Response – February theft data
Validation set – March theft data

Following is the step-by-step approach taken to built to logistic model

1. Train data preparation:
    a. In our data (February theft) we have got only those locations (coordinates) where the theft has occurred, called as positive observations. To train our logistic 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 CTA rail stations, police stations, parks and landmarks from shape files downloaded from Chicago data portal
    b. Calculated minimum distance from each point of *training data* to identified factors (nearby CTA rail station, police station, park and landmark)
    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 January theft records
4. Build logistic regression model, Response as response variable and spatial factors and KDE estimate as predictors.
5. Analyze the model to identify which factors are highly correlated with the theft occurrence (Response = 1)
6. Predict response on March data, using the fitted model and predictors from February
    a. Prepare test data – Predictors from February thefts
    b. Predict crime for prediction points using the fitted logistic regression model
7. Validate the model on ground truths (March thefts) for its prediction accuracy

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. Parks_Aug2012 (Polygon shape file)
2. LandmarksNationalRegister_nov2012 (Polygon shape file)
3. PoliceStationsDec2012 (Points shape file)
4. CTA_Stations (Points shape file)

Analysis:
We have built a logistic regression model on January 2014 theft data as predictors and February data as response.
From the model, we can interpret the importance of factors that are influencing the theft occurrence.

```
Call:
glm(formula = response ~ . - x - y, family = binomial, data = train_full)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.4721  -1.0929   0.3286   1.0903   2.3401

Coefficients:
                    Estimate    Std. Error   z value      Pr(>|z|)
(Intercept)        -4.901e-02   6.996e-02    -0.701      0.483615
crime.density       2.967e+08   1.717e+07     17.276     < 2e-16 ***
Park.min.distance  -2.030e-04   5.523e-05    -3.676      0.000237 ***
PS.min.distance    -1.004e-04   2.457e-05    -4.085      4.41e-05 ***
Rail.min.distance  -1.174e-04   1.596e-05    -7.357      1.88e-13 ***
Land.min.distance   3.628e-05   2.413e-05    1.504       0.132651
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Null deviance: 10682.0 on 7713 degrees of freedom
Residual deviance: 9477.8 on 7708 degrees of freedom
AIC: 9489.8

From the above summary, we can interpret that minimum distances from parks, police stations, CTA railway stations are negatively correlated with probability of occurrence of theft. But minimum distance from landmarks is not significantly correlated with the theft. The high AIC value suggests that regression was significant.

Interpretation of factors:

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

= 2.967e+08* KDE + -2.030e-04 * PR + -1.004e-04 * PS + -1.174e-04* RS

KDE = Probability of theft occurrence at the location
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
KDE Estimate:
The theft probability is highly correlated with KDE probability. A 0.001% increase in KDE estimate would cause an increase of 2967 times in log of odds of theft occurrence if all other factors were held constant. It is plainly interpreted as KDE estimate is positively correlated and the risk of theft occurrence would increase if the KDE estimate increases

Minimum distance from Park in meters (P):
The coefficient is negative, which means that the increase in distance would decrease the probability of theft occurrence. It can be interpreted as the location closer to park, the higher the probability of theft occurrence when the other factors are kept constant.
The log of probability of odds decrease by 0.203 if the minimum distance from park is increased by 1km when other factors are kept constant. As hypothesized the locations closer to parks has higher risk of theft.

Minimum distance from Police Station in meters (P):
The coefficient is negative, which means that the location closer to police station has higher probability of theft occurrence. It is contrary to our hypothesis that location closer to police stations has low risk of theft.
The log of probability of odds decrease by 0.1 if the minimum distance from police station is increased by 1km when other factors are kept constant.

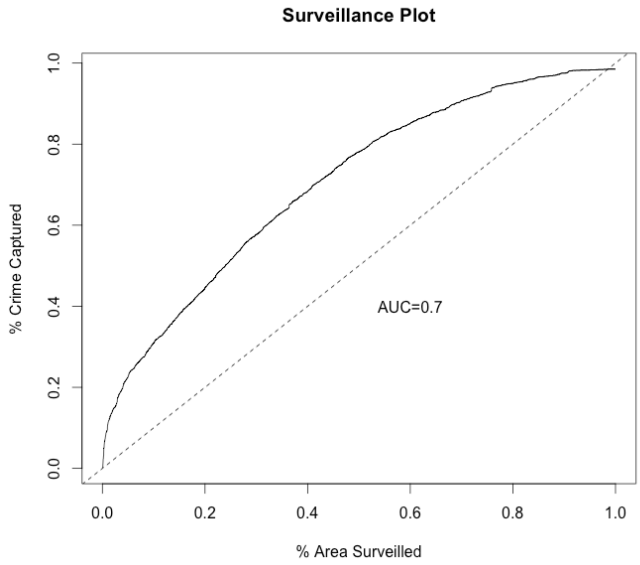
Minimum distance from CTA Railways Station in meters (P):
The low p-value (1.88e-13) shows that it is the most significant factor out of all the spatial factors considered for the regression. Similar to other factors, the proximity of CTA railway station is positively correlated, which means that, the closer the location to railways station the higher the probability of theft occurrence.
It is according to our hypothesis that location closer to crowded places such as railways stations has high risk of theft.
The log of probability of odds decrease by 0.117 if the minimum distance from railways station is increased by 1km when other factors are kept constant

Minimum distance from Landmarks in meters (P):
According to our model, the proximity to landmarks doesn't paly a significant role in estimating the probability of theft in a location. We hypothesized that landmarks would have a negative correlation which means locations closer to landmarks should have lower risk of thefts. But we did not find any correlation in our model.

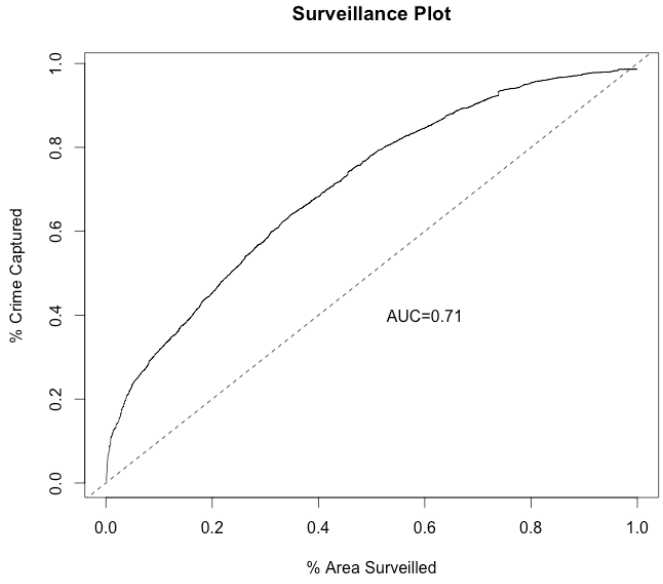
Evaluation:

Using the above model, which was built on January 2014 theft data, we can predict the risk of theft for any location in Chicago. In this scenario we used the model to predict the theft occurrence of February 2014. We evaluated the prediction accuracy of model by comparing with actual theft cases in February.

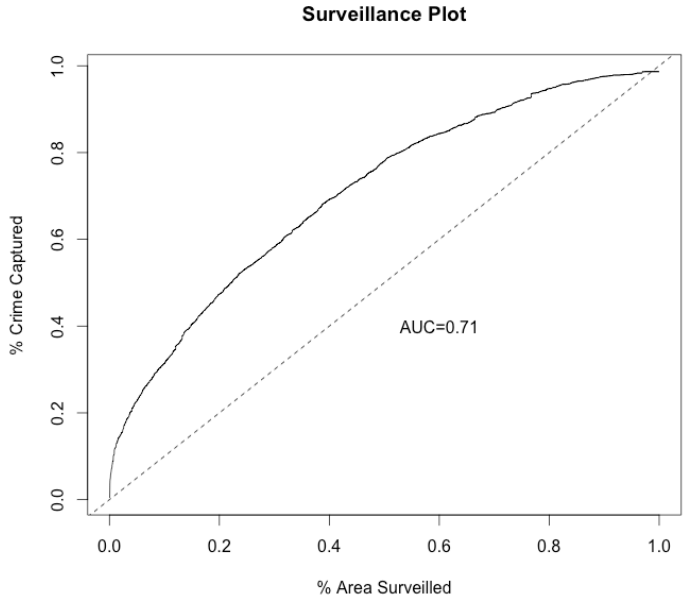
We used surveillance plots that were explained in our previous report to validate the goodness of model with hypothesized spatial factors.
Following plot has AUC value of 0.7, which means the model has done a good job of predicting March thefts from Responses from February and predictors from January theft data.



Logistic regression model on **responses from May** and **predictors from April**
Surveillance plot - evaluating the Model on ground-truth thefts from **June**



Logistic regression model on **responses from October** and **predictors from September**
Surveillance plot - evaluating the model on ground-truth thefts from **November**



3. Recommendations

Chicago police department can use our above findings in their patrolling strategies for the better prevention of theft occurrence. From spatial factor analysis we identified that locations close to crowded stations, parks and police stations have increased risk of theft. Following strategies could be employed in the regular patrolling and prevention activities

Prevention Strategies:

1. Warning signs and boards in nearby locations of high risk areas – railways stations, parks, crowded places etc.

2. Advertisements on screens inside railway stations would make customers attentive about thefts

3. Regular announcements about safety of belongings in stations and in trains

4. Visible cameras in parking lots near stations and shopping centers help in apprehending motivated offenders and could prevent auto theft

5. Parking a couple of empty police cars (Scarecrow cars) in parking lots could help in apprehending motivated offenders

6. Having fee parking lots, trained attendees with number plate identification systems would help prevent auto thefts from large parking spaces near stations, malls, sport events etc.

7. Directed patrolling based on operational hours of stations, parks. The visibility of cops, security guards, police dogs would discourage the offenders

8. Creating general community awareness about location risk factors using media or workshops could help in theft prevention

Intervention & Investigation Strategies:

1. Complaint desks in crowded places for quick tackling of cases when occurred

2. Installing hidden cameras at the areas highly prone to thefts as per the hot spot and spatial factor analysis and monitoring them during busy hours of the station could provide opportunities for quick intervention

3. Increased patrolling in about 1 km distance from the potential factors such as stations and parks. We have observed the closer the location to these factors the higher the risk of theft

4. Parking lots with video recording at ticket booths or exits could provide strong leads in investigation and quick intervention in to the reported incident