

Crime Density S/S-T Analysis

Washington, DC

Peter Trubey

Chieko Seto

San Diego State University

STAT 596 Spring 2014

Professor Alexander Kolovos

Abstract

Estimating Crime is essential, not only for ordinary citizens, but also policy makers and businesses. In the United States, the annual cost of crime is approximately \$3.2 trillion per year, which is roughly the combined cost of the wars in Afghanistan and Iraq from 2001 to 2011. This includes victim costs, criminal justice costs, crime career costs, and some intangible costs. Crime rate tends to affect housing prices, as people may be willing to pay more to be in a safer neighborhood. So there exists a clear need for accurate prediction of crime. The objective of this analysis is to predict the violent crime density in the area of Washington, District of Columbia, by using the annual violent crime incident data for the region. The data set was obtained from the District of Columbia Metropolitan Police Department. These data were collected as individual observations from April 25, 2013, to April 25, 2014. In total, we have 2136 observations of crime in the district in that period of time. However, this analysis will focus specifically on violent crime, including arson, homicide, sexual abuse, robbery, and assault with a dangerous weapon. In this analysis, correlation between crimes committed with respect to distance between crimes as well as time period was analyzed within the framework of Spatial/Spatio-temporal Analysis; specifically Kriging. From this analysis, we generate a crime density map for the district for the year. Results showed that the neighborhoods on the Northwest side of the district were vastly safer than those on the Southeast. The safest areas recorded included the neighborhoods of Tenleytown, Woodley Park, Spring Valley, and Georgetown. The more dangerous neighborhoods recorded included Anacostia, Congress Heights, and Lincoln Heights. Of particular surprise is the concentration of crime towards the center of the map, in the areas of Shaw, U-Street, and Columbia Heights.

Introduction

Although crime has been dropping steadily for the last 30 years, the cost of crimes committed is still a large burden on society. In the United States, the number of victimizations in 2010 dropped to approximately 20 million, which is almost half of the number of victimizations in 1995.[3] However, the same study showed the aggregate cost of crime has increased since the 1900's. The total cost of crime annually today is 3216 Billion, which comes to 5284 per person in the United States in 2012.[1] This includes direct cost, as well as indirect cost such as the opportunity cost of pursuing a life of crime, time lost both by criminals and by victims, as well as the burden of fear by crime. We also see other costs borne by persons unrelated, in housing prices in region of crime tend to go down. One study observed that a 10 % increase in violent crime in an area tended to drop property values by 6 %.[2] Because of this loss of housing value, along with other economic opportunity losses, we can see that higher crime is antithetical to tax base, which in turn may cause underfunding of police, which may lead to more crime. We can see that there exists clear and compelling reasons for estimating the crime rate by areas in the city, if we are a consumer looking to purchase property, a Law Enforcement policy setter looking to maximize the effectiveness of a police force, or a policy planner trying to make my city a better place. The purpose of this analysis is to analyze the correlation of violent crime by area in the boundaries of Washington, DC.

Data Description/ Preparation

Data are collected as incident reports from the Washington DC Metropolitan Police Department's website.[4] From there, we take a history of one year of crime data. Raw data is collected in point pattern style: It describes individual incidents, and all

incidents that occurred in the area are in the dataset. Because the Kriging method we intend to use is not applicable to this form of data, we are required to aggregate individual observations by area to reach a rate of crime for a given area, for which the Kriging method will be applicable.

The dataset contains 2136 reports of criminal activities, in 4 columns. Those columns include the *XBLOCK* and *YBLOCK*, which can be thought of as Latitude and Longitude coordinates in this dataset. In fact, they correspond to a standard measurement basis used in Maryland, but for this analysis it is sufficient to know that a change in *XBLOCK* of 1 is equivalent to 1 meter distance. The total change in *XBLOCK* and *YBLOCK* are approximately 17500 times 20100 corresponding to at 17.5 times 20 kilometer area. For this analysis, we decided to aggregate these observations by square kilometer, which led to a grid of 18 times 20. The other columns included time of incident report, which we aggregated into 30 day blocks, as well as the offense, by which the data was classified into 1 of 8 categories. In order to get a worthwhile rate of crime, we needed to weight the types of crime by severity. Using homicide as a baseline, we weight the other types of crime relative to homicide, then sum the total crime in an area by month in order to get a rate of crime per area. We call this rate the number of Homicides (equivalent) per square kilometer per month. Weighting is as follows: homicide is weighted as 1, while sexual abuse, assault with a dangerous weapon, and arson are weighted as 0.7. Robbery is weighted as 0.5. The R code used to transform the data into this form is included in the appendix. The color plot indicating count of homicides (equivalent) by grid point is given in figure 1. We then added an arbitrarily small value to those observations that were 0, so that we could take a natural log transform of the data.

Methodology

After the S/ST data was developed, the spatiotemporal distance between the grid observations was calculated. We set a spatiotemporal parameter $v=1$ to reflect that we expect spatial and temporal proximity to other crime to be equally contributing. Then, exploratory analysis was conducted. The surface trend analysis is conducted using the kernel exponential smoothing method, which is a non-parametric method of fitting a moving window to the data and smoothing or fitting splines within that window. The color plot of fitted trend can be seen in figure 2. We call this fitted smooth the trend, and take residuals as the difference between observed values and this fitted trend. The residuals color plot can be seen in figure 3. The correlations between these residuals are then analyzed, and with that analysis we fit a prediction model using the Kriging method.

Results

The empirical covariance plots can be seen in figure 4. To fit this, the spatiotemporal covariance model we selected included two nested models, which we will hereafter term the short and long term models. The long term model included an exponential spatial model with a range of 1, and an exponential temporal model with a range of 100. We estimated this model to comprise approximately 40 % of the sill. The short range model, which we estimated to comprise 60 % of the sill, also had an exponential spatial model component with a range of 1, and a spherical temporal model component with a range of 2.6. These combine to give the fast drop-off and extremely long duration in the temporal axis, as well as the fast drop to 0 in the spatial axis. This fitted model is visible in figure 5. Using this covariance function, we were able to run the Kriging models, and map out the more dangerous portions of the district.

Conclusion

The predicted crime levels at time periods 7 and 12 are visible in figure 6. We saw little to no violent crime in the Northwest portion of the district, in the areas surrounding Woodley Park, Tenleytown, Spring Valley, and Georgetown. However, we saw significant violent crime in the southwest portions of the district, covering the areas of Anacostia, Congress Heights, and Lincoln Heights. Of particular note, and of some surprise to us, was the amount of violent crime in the center of the district, in the area encompassing Columbia Heights, Shaw, and Adams Morgan. This region encompasses many popular late-night hangout spots, 'bar districts', and the like. These may serve as attractors to crime. However, Georgetown also hosts several bars and nightclubs, and doesn't seem to see this level of rise in crime. It is interesting to see the difference between these two areas.

Reference

[1] Anderson, David Anderson. 2012 September 18. *Cost of Crime*. Retrieved from

http://archive.centre.edu/cost_of_crime.pdf

[2] Ihlanfeldt, Keith. Mayock, Tom. 2009 February. *Crime and Housing Prices*

<http://coss.fsu.edu/dmc/sites/coss.fsu.edu.dmc/files/CrimeHousingPricesFEB25.pdf>

[3] National Crime Victimization Survey (1995-2010)

<http://www.bjs.gov/index.cfm?ty=dcdetail&iid=245>

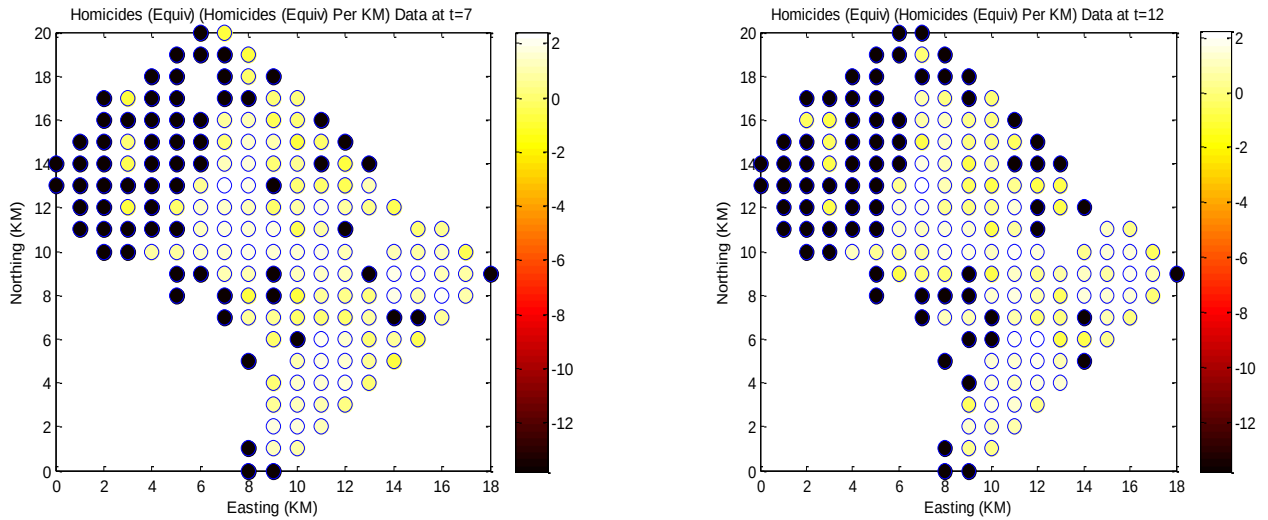
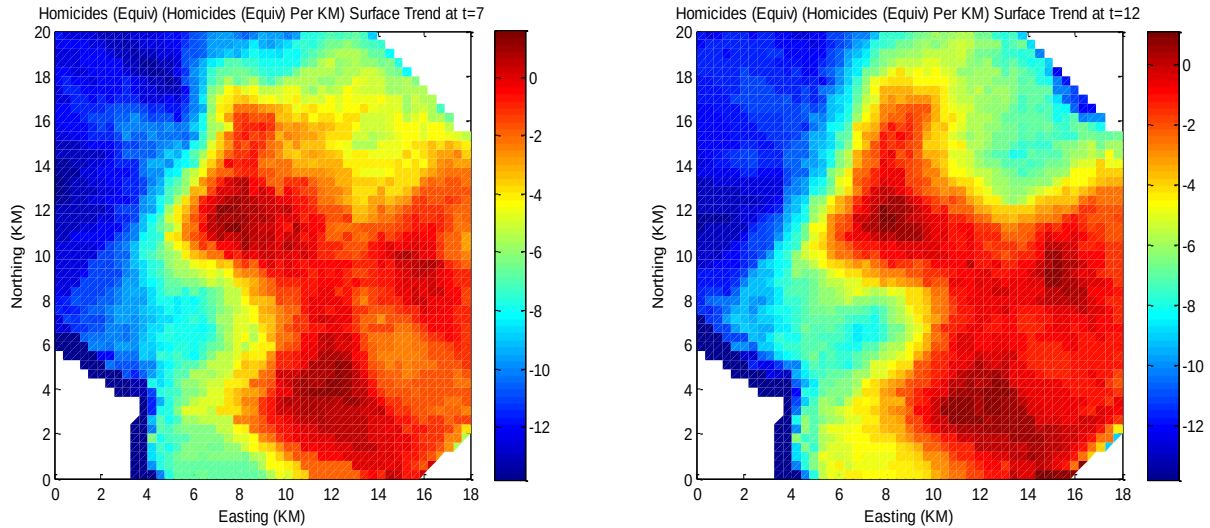
[4] District of Columbia Metropolitan Police Department

<http://crimemap.dc.gov/CrimeMapSearch.aspx> (Accessed April 25, 2014)

Trubey, Seto

STAT 596 Final Project

Appendix

Figure 1: *Homicide (Equivalent) Color Density Plot*Figure 2: *Trend Map from Kernel Smoothing at time 7, 12*

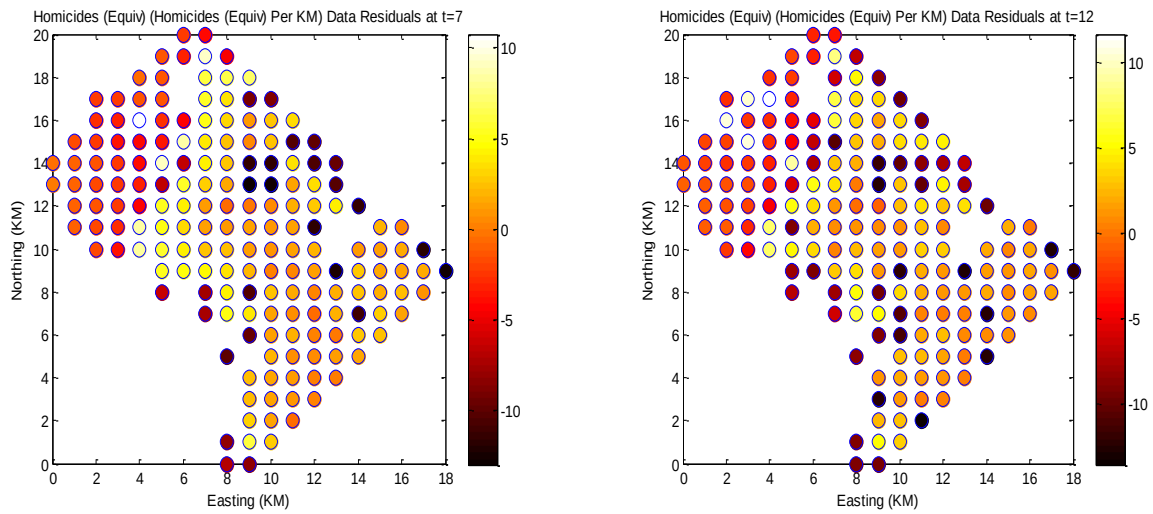


Figure 3: Color Plot of Residuals, after Detrending

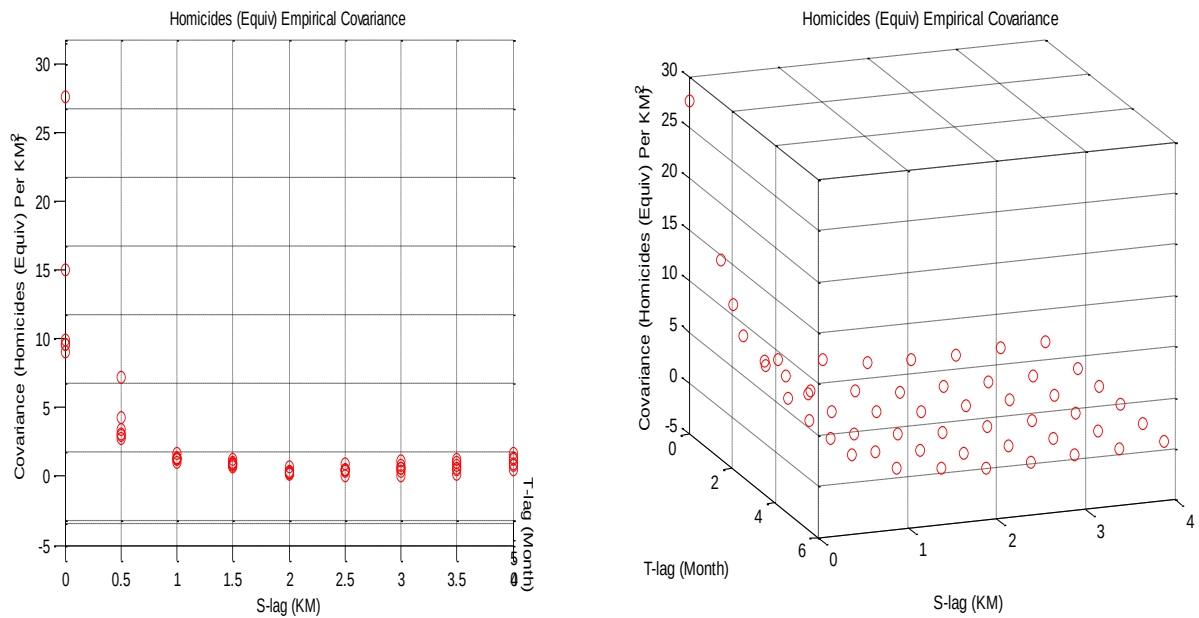
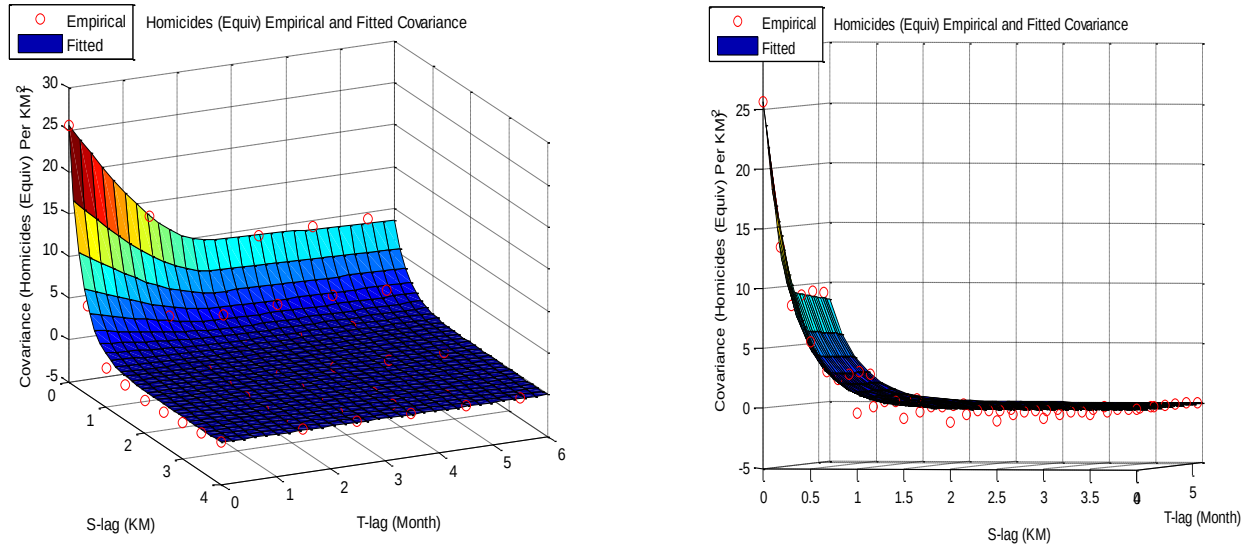
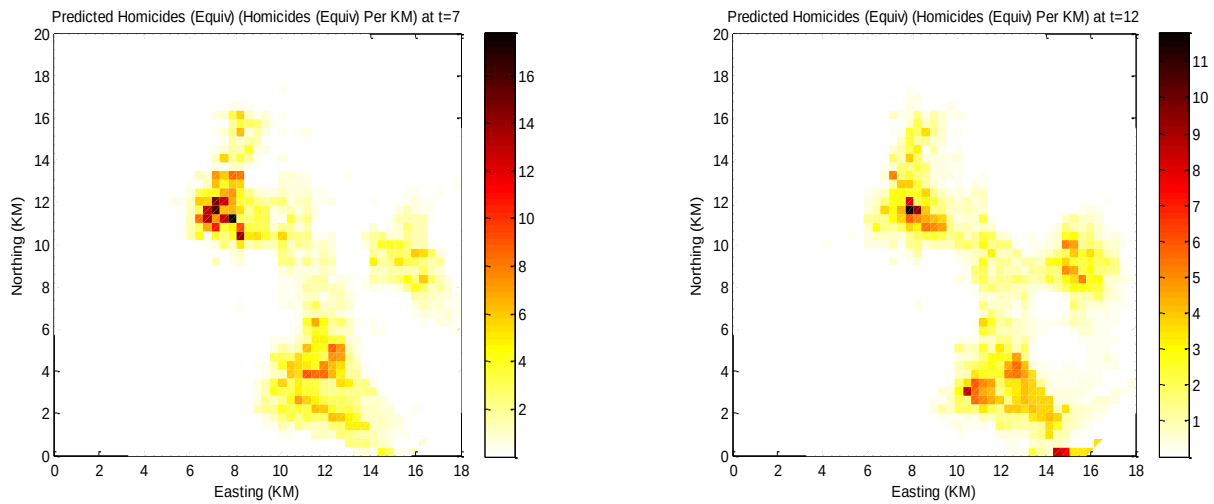


Figure 4: Empirical Spatiotemporal Covariance Plot

Figure 5: *Fitted Spatiotemporal Covariance Model*Figure 6: *Predicted Homicides (Equivalent) per kilometer per month, at times 7, 12*

Trubey, Seto

STAT 596 Final Project

```

library(reshape2)
data = read.csv('SearchResults.csv', header = TRUE) # load source data
data$XBLOCK = as.numeric(as.character(data$XBLOCK))
data$YBLOCK = as.numeric(as.character(data$YBLOCK))
max(data$XBLOCK) - min(data$XBLOCK) # separate city into 18 x 20 grid marks
max(data$YBLOCK) - min(data$YBLOCK)

data$x = round((data$XBLOCK - min(data$XBLOCK))/1000) # by kilometer
data$y = round((data$YBLOCK - min(data$YBLOCK))/1000)
data$t = floor((strptime(data$REPORT_DAT, '%m/%d/%Y') - min(strptime(data$REPORT_DAT,
'%m/%d/%Y')))/(3600*24*30))+1 # 30 day increments
data = data[data$t <= 12,] #ensure only full 30 day increments are counted.

# crime per km per week
data$crime = NA
data$crime[data$OFFENSE == 'ARSON'] = .7
data$crime[data$OFFENSE == 'ASSAULT W/DANGEROUS WEAPON'] = .7
data$crime[data$OFFENSE == 'BURGLARY'] = 0
data$crime[data$OFFENSE == 'HOMICIDE'] = 1
data$crime[data$OFFENSE == 'MOTOR VEHICLE THEFT'] = 0
data$crime[data$OFFENSE == 'ROBBERY'] = .5
data$crime[data$OFFENSE == 'SEX ABUSE'] = .7
data$crime[data$OFFENSE == 'THEFT F/AUTO'] = 0
data$crime[data$OFFENSE == 'THEFT/OTHER'] = 0

coords = data[c('x','y')]
coords = unique(coords)
coords = coords[order(coords$x,coords$y),]

crime.wide = cbind(coords,matrix(NA,nrow=nrow(coords),ncol=max(data$t)))

for(i in 1:nrow(coords)){ #horribly inefficient method
  for(j in 1:max(data$t)){
    crime.wide[i,j+2] = sum(data$crime[data$x == coords[i,1] & data$y == coords[i,2] & data$t == j])
  }
}

crime.wide$x = factor(crime.wide$x)
crime.wide$y = factor(crime.wide$y)
crime.long = melt(crime.wide,id.vars=c('x','y'))
names(crime.long) = c('X','Y','t','Crime')

crime.long$X = as.numeric(as.character(crime.long$X))
crime.long$Y = as.numeric(as.character(crime.long$Y))
crime.long$t = as.numeric(as.character(crime.long$t))
crimedata = crime.long[order(crime.long$X,crime.long$Y,crime.long$t),]

write.csv(crimedata,file="crimedata.csv",row.names=FALSE)

```