Mining Crime Data Using Spatiotemporal Correlation Analysis

Madhumita Duvvuri[1], Prithvi Yarlagadda[1], Natasha Kapoor[1], Gowtham Srungarapu[1], Hakan Gogtas[2], Edward Moskal[1], and Sylvain Jaume[1]

[1] Data Science Program, Saint Peter's University,
2641 John F. Kennedy Blvd, Jersey City NJ 07036, USA,
http://www.saintpeters.edu/data-science
[2] Model Risk Governance, American Express

Abstract. The availability of detailed crime records over long period of time offers new applications for data mining and the emergence of crime analytics. Are crimes happening in one neighborhood related to crimes happening later in another neighborhood? By applying time series and correlation analyses on crime records collected from 2001 to 2015, we investigate the existence of this relation across the City of Chicago and at different time scales. Our results indicate that the most significant relation exist at the quarterly scale and bi-monthly scale.

Keywords: crime analytics, correlation, time series, geospatial analysis, big data, smart cities, data visualization

1 Introduction

Mining crime data is a new field that is of high value for law-enforcement and intelligence-gathering organizations. In an effort to accurately and efficiently analyze the growing volumes of crime data, Chen et al. [1] review crime data mining techniques and present four case studies done with Phoenix police departments. Nath [2] focus on k-means clustering with some enhancements to identify crime patterns. Estivill-Castro and Lee [3] incorporate clustering and association-rule mining into an exploratory tool for the discovery of spatio-temporal patterns. They have also developed an application to visualize cluster boundaries and mining association rules. Keyvanpour et al. propose [4] to extract important entities from police narrative reports written in plain text and entered into a database, in law enforcement agencies. They apply self-organizing maps clustering and plan to use the clustering results to perform crime matching process. Yu et al.[5] build a model that takes advantage of implicit and explicit spatial and temporal data to make reliable crime predictions.

Crime analysis helps the police departments in finding patterns and relationships between the crimes. They provide us with who, what, when and where kind of information about crimes, with this information we can build better strategies to prevent or reduce the crime from happening. Crime analytics main agenda is to find valuable patterns and relationships from the previous committed crimes

and data sets we are provided with. This in turn helps law enforcement to establish crime prevention actions, which leads to the improvement of overall safety and quality of life in a state or country. As the advancement in the technology, research and development in this area are coming up with new software which can help in understanding, reporting and analyzing crime patterns. So many discoveries in relationships between locations and crimes are emerging due to these technologies by using spatial locations.

Crime data can be classified into two groups, point data and areal data based on the spatial locations. Point data will have the information of where the crime happened with locations to maximum accuracy, mainly latitudes and longitudes. Areal data will have the information of crime locations in a boundaries they defined. Point data can be used as an areal data in some situations, as we have location points we can easily identify them to which area it belongs (state, county or district). For visualizing different crime patters there are so many tools which can plot graphs like time-series plot, bar graph, choropleth map or scatter plot.

To visualize correlation patterns of crimes in Chicago using spatio-temporal with our data set, we used three techniques heat map based on the number of crimes per district, time series of number of crimes per day throughout the year and correlation heat map where the crime correlation between districts are plotted as discussed in the section 3. This helps in better understanding and step-wise approach towards the patterns of crimes. We also need to recognize the importance of different factors like type of crimes, general location of crime (street, shops etc.) and the neighborhood for better assessment of the relationships, that may underlie in the process of understanding the patterns of crime activities in the neighborhood and analyze it. However, there are less number of resources which can accomplish the multi-dimensional visualization from the given data and analyze it. Multi-dimensional approach may include the factors like type of crimes and location. Different types of crimes and different locations will tell us different story and different kind of patterns with respect to time and space, so we need to categorize them and perform analytics. In our paper, we took the crime theft and applied the three step technique discussed before. When we thought of using single factor for the analysis, we calculated different crimes that are registered and number of crimes committed in each type. Among all those types of crimes theft has the highest records, so considered to use theft as our base crime for the analysis and started from there. Discovering the factors which are too small to be ignored and using them to understand, build and analyzing the patterns of the crime activities will provide us knowledge of how space and time are influencing the crimes. Using this we can build the effective model, use them in law enforcement and develop theories and methodologies for crime patterns so that we can build better crime prevention programs. From the beginning we stick to the idea of correlation in crimes between the districts. Which states data to consider, what data to use, what we are going to accomplish.

In this paper, we make experiments to verify and measure the relation of crimes between different neighborhoods at different times in the city of Chicago. Section 2 describes the data sets that we used for this study. Section 3 develops

the correlation analysis method that we have developed. The section 4 our results are reported. Section 5 presents a qualitative comparison of our results with earlier work. Section 6 concludes our paper.

2 Data

We first gone through every states available data sources and searched for crime data which is open source. Going through each state we found that Chicagos crime data is more structured and is available in many different formats like excel, CSV, JSON, Xml etc. They are also providing APIs to plug them into any applications. As this data provides lots of possibilities in reaching what we intended to, we thought of starting with Chicago crime data and move forward from here. This data-set have lot of columns which helps in describing the crime like time, county, place, even latitudes and longitudes.

For calculating correlation of crimes between the districts of Chicago we need total number of crimes which we dont have in the data set. So we had to calculate the total number of crimes committed per district. This gave us refined data set which we can use for the calculation of correlations between the districts. There is also a time lags for some districts which needs to be eliminated. If there are no crimes committed on some particular day or if there is no crime registered on that day there wont be any records for that day. Other districts may have crimes registered on that day, so there will be a time lag between two districts. Eliminated this time lags by adding identifying the missing dates in the data-set and adding 0s to those dates. After eliminating the time lag we can calculate the correlations very effectively and plot time series without any errors or loosing accuracy. In calculating correlation of crimes between different districts on daily basis, eliminating time lag helped in achieving accurate correlations. Categorized the type of crimes committed and found that theft are the most common crime. So, the correlation are calculated on theft. For this we need to filter the records where we get only theft based crimes. After doing all these steps, data set is finally refined to be used for calculating correlation between districts based of crime, theft.

The data used in this paper was extracted from City of Chicago data portal. The data that is publicly available in the Chicago data portal, in turn extracts the data from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. Most of the crime data is kept confidential and is not for public use[6]. This site provides the temporal, spatial and crime data required for a data analyst to identify the crime hot spots in Chicago for various crime types. This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present. Each recorded crime incident contains information on case ID, case number, date/time, type, location, description, addresses(shown at block level only), District and geo-coordinates. It is observed that over 5 million crimes were reported from 2001 through 2015 in 22 different districts of Chicago.

In the process of crime analysis the crime data can be categorized into three types:

- Spatiotemporal data (crime locations and type of occurrence)
- Crime natural specifications (crime scene description, offenders behavior)
- Offender profile (offenders specification (age, sex, race))

Our paper uses the spatiotemporal data in the analysis to investigate whether crimes happening in one neighborhood are related to crimes happening in other neighborhoods.

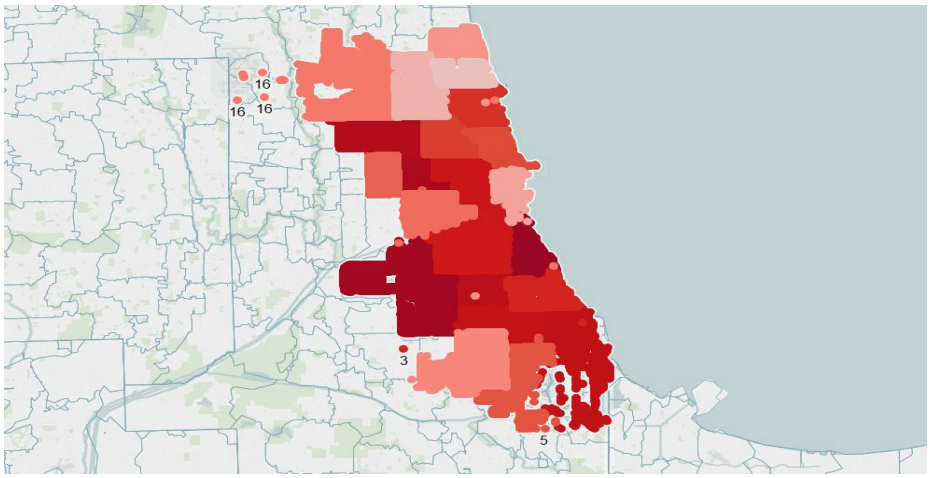

Fig. 1. All the crime events in the city of Chicago from 2001 to 2015. Geo-coded events reveal features of the city such as street, alleys, and grocery food stores.

Figure 1 indicates a higher crime rate in the southern part of Chicago. Plotting geographical overlay of the city makes it possible to analyze the crime data at district or neighborhood level.

3 Method

Initially we looked into how crimes are distributed in Chicago, which districts are with most crimes and are there any relationship in it. Plotted the heat map of crimes on the Chicago map based on number of crimes in that district, but we didnt find any relationship we can relate it to the correlation. Next, needed to know how crimes are distributed over the year long. This showed us some patterns like, on Christmas Eve the rate of crimes every year are low when compared to all other days in every district. This seems reasonable, as

there might be less number of cops working that day or everyone is busy in celebrating their own. Second thing we observed is the rate of crimes in winter are low compared to spring or summer. There is no big difference between these seasons but it is not so small that we can ignore this analysis. This is also observed in all districts of Chicago.

The temporal and spatial characteristics available in the crime data makes it easy to show different patterns in crimes. With time and location details in each crime report known the crime rate on daily, weekly, or monthly scale can be measured for a district within city of Chicago. A time series for the number of crimes in the city for 2001 displays seasonal trends on a monthly scale. Figure 2 shows seasonal trends that indicate peaks in crime at the beginning of summer and around the midpoint of fall. It is important to note that there is a decrease around the holiday season time.

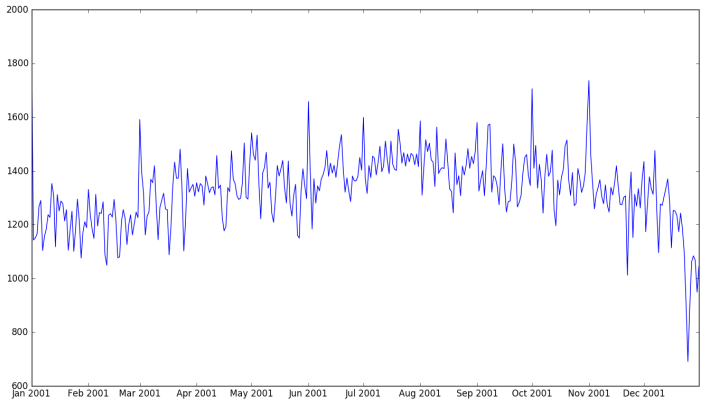

Fig. 2. The figure represents a monthly times series plot against number of crimes in the city of Chicago in 2001.

With aggregation on the types of crime we test the relationship of the major crime committed in the city for a long period of time with the various neighborhoods. Figure 3 shows that theft has the highest count and is the most commonly committed crime in the city through the years 2001 to 2015. To investigate the existence of a relation between the crimes happening in different districts, we use the time series analysis and the correlation matrix analysis to get the most significantly correlated districts.

We perform a correlation matrix analysis to investigate the dependence of multiple variables at the same time. Following the notation of Toole et al.[**?**] we build a K T correlation matrix Y where K is the number of time series taken from the location and T is the length of each time series. Since each

time series corresponds to a location in the matrix, we are able to associate locations in Chicago with correlations. The result for this analysis is a table which has the correlation coefficients of each variable with the other variables. The higher the value of the correlation coefficient, the stronger the significance of the relationship between the districts will be. Using the time series and the crime rate, we can find if the correlation exists between the different neighborhoods on theft-based crimes on multiple time scales.

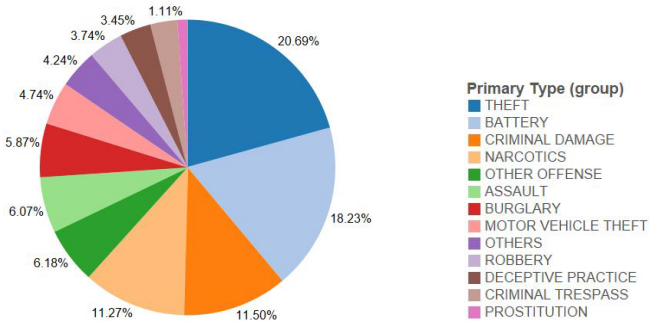

Fig. 3. Percentage Breakdown of Crime Types. The highest percentage of crimes in Chicago is Theft followed by Battery.

Table 1. Categorical Groupings of Different Crime Types from 2001-2015

Category	Offenses Included	Crimes (%)
All	all reported offenses	5973040 (100%)
Theft	burglary, robbery, auto	2092785 (35.04%)
Violent	assault, homicide, gun	1481233 (24.8%)
Vandalism	deceptive practice, arson, criminal damage	1077077 (18.03%)
Sex Offenses	public indecency, prostitution	683602 (11.4%)
Drug	narcotics, possession, sale	672899 (11.26%)
Other Offenses	obscenity, intimidation, involving children	425967 (7.13%)

4 Results

Now that the required fields and data to find the significance between the districts is established, we seek to analyze the data. For this, we calculated the correlation based on time-series with each district of Chicago. We created a M*N matrix where M is time series of number of crimes and N is Districts. We found a small amount of time lag in each district, as if there is no crime committed in any particular day in a district. So we need to eliminate them and make

it unbiased based on time-series. To achieve this, we retrieved the missing dates of each year grouped by district and addressed the time lag issue in the dataset.

First we computed the correlation from one day to the next. However, we found no significant correlation between districts. Hence we performed the correlation analysis on different time scales: weekly, biweekly, monthly, bimonthly, quarterly and yearly. From this analysis, we observed that a significant correlation exists between districts for bimonthly and quarterly theft-based crimes as shown in Figure 4.

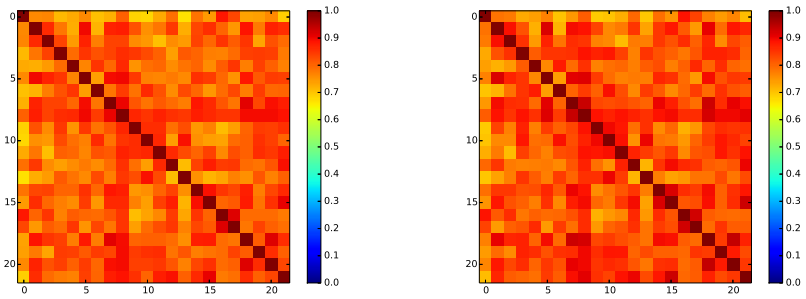

Fig. 4. This figure represents the results of our correlation analysis on theft records from year 2001 through 2015 on a bi-monthly scale (left) and on a quarterly scale (right).

5 Discussion

From the correlation analysis performed on the crime data the strongest relation between the neighborhoods was deduced. Using the correlation coefficients from the matrices we found the highly correlated districts (two highest correlations that is four districts) from bimonthly and quarterly correlations. Plotting the highly correlated districts geographically gives us a better understanding that the crimes happening in one neighborhood are related to crimes happening later in another neighborhood.

For the bimonthly correlation analysis, the top 4 significant correlated districts are on one hand the Near North district and the Town Hall district, and on the other hand the Foster district and the Rogers Park district (Figure 5 left). For quarterly top 4 significant correlated districts are on one hand the Chicago Lawn district and the Deering district and on the other hand the Foster district and the Rogers Park district (Figure 5 right). We observed that the crime rate (theft) is significantly high in the Foster district and the Rogers Park district in both bimonthly and quarterly correlation.

One of the usages from this calculating correlation is, if any two districts are correlated in crimes the police department can allocate the officers accordingly, as there can be approximately equal crimes in both districts officers can also be appointed approximately equal. With this we can expand our approach towards different types of crimes other than Theft alone. We can also develop correlations of crimes in other states and other types of crimes. With enough data we can even calculate the correlations between states too.

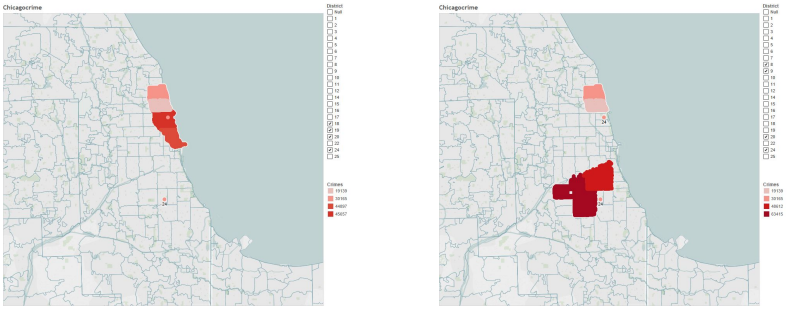

Fig. 5. Neighborhoods on Chicago with the highest correlation based on the bimonthly correlation analysis (left) and based on the quarterly correlation analysis (right).

6 Conclusion

We analyzed the correlation of crimes that occurred in different neighborhoods of Chicago at different times. We have observed that although there is no significant correlation was found for most of the crimes; However, a significant correlation could be observed between a small number of neighborhood. The highest correlation was found between Foster district and Rogers Park district.

References

1. H. Chen, W. Chung, Y. Qin, M. Chau, J. J. Xu, G. Wang, R. Zheng, and H. Atabakhsh, "Crime data mining: an overview and case studies," in *Proceedings of the 2003 annual national conference on Digital government research*. Digital Government Society of North America, 2003, pp. 1–5.
2. S. V. Nath, "Crime pattern detection using data mining," in *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*. IEEE, 2006, pp. 41–44.
3. V. Estivill-Castro and I. Lee, "Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data," in *Proc. of the 6th International Conference on Geocomputation*. Citeseer, 2001, pp. 24–26.

4. M. R. Keyvanpour, M. Javideh, and M. R. Ebrahimi, "Detecting and investigating crime by means of data mining: a general crime matching framework," *Procedia Computer Science*, vol. 3, pp. 872–880, 2011.
5. C.-H. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 779–786.
6. M. Kassen, "A promising phenomenon of open data: A case study of the chicago open data project," *Government Information Quarterly*, vol. 30, no. 4, pp. 508–513, 2013.