Chicago Crimes: A Study of temporal factors

2012 - 2016 using Chicago Police Department CLEAR data

By Nathaniel Schwamm, Andrew Jones, Christopher Broll, and Toufik Bouras

Introduction

Our initial interest into crime datasets was agreed upon after exploring possible topics as lately it has been a major subject within contemporary politics and news. So after agreeing on crime as a broad topic we were surprised and excited to find a dataset that described specifically Chicago in detail. While there are other metropolitan areas that have higher incidences of crime, the reasons Chicago is so often focused on is because of its size. As the third largest city in America, small upticks in the crime rate of Chicago can have drastic influence over the crime rates of the United States at large. For this reason we were excited to focus on this dataset and attempt to describe and explain some of the trends and temporal factors around it.

Dataset Background

In a comprehensive way this dataset covers reported crime within Chicago over the period of 2001 - 2017. It includes a wide array of crime types and definitions within. Possible types of crime include but are not limited to: Battery, Theft, Narcotics etc. Furthermore it classifies crimes as “index” vs. “non index” crimes. Index crimes are the eight crimes the FBI combines to produce its annual crime index which was necessary for comparison purposes between states. In layman's terms this this loosely corresponds between “serious” crimes vs. non-serious crimes. The dataset also includes descriptive fields for each crime incident such as type of location, block, district as well as longitude and latitude. This data set contains 1,456,714 crime records and could qualify as a “medium” sized dataset or one that at least cannot be effectively analyzed in just MS Excel. The dataset is continually updated as more crimes are committed, or information the redefines previous crimes is discovered. One item to note is that we focused on the subset of this data for the years 2012 - 2016. The reason behind this decision are both practical, in that it reduces load times and data cleanings, as well as potentially insightful as this will make the numbers fall within one Presidential administration.

For the technical details of how this data set was compiled we need to look at the sources. The original source of this data is the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. This system was developed by Oracle Corporation and launched in 2002. After receiving a crime report the officer inputs the information of the crime report in the system. The system then organizes these records in a logical and accessible way in the form of tables, charts and maps for publication and use.

While the dataset is very comprehensive there are some key limitations that must be considered in the analysis. For example the largest obvious flaw in this dataset (which is a function of practicality) is that the crimes in this dataset are strictly *reported* crimes. While obviously police follow up on many reported crimes and they are confirmed as legitimate crimes, we do not know the percentage of reported crimes that are investigated or “confirmed” crimes. Another not crucial but still considerable flaw is that with any system that requires manual data entry it poses the risk of data entry error. Second, this system was developed around the turn of the century and undoubtedly has morphed over that time until now. It is reasonable to assume that a small percentage of records may be erroneous due to simple typos, misunderstandings, and database schema problems over the years. Third, since these crimes are reported by the general public it is anticipated that even the people experiencing the reported crime may not be communicating it correctly. Perhaps the person reporting the crime was a victim rattled by the crime itself or simply confused about their location. Perhaps the reporting person had poor cell reception or simply incorrectly described the event. Ultimately there could be many confounding reasons for why the crime as it actually occurred was not communicated to the receiving officer accurately. Fourth, while a minor point for this type of analysis, the data is only reported at the block level and this is for privacy reasons. The posting of the dataset actually strictly prohibits attempts to reverse engineer reported crimes to exact addresses. Additionally, once we dug into the types of crime, the coding of crimes is often not perfect. For example, the dataset uses IUCR (Illinois Uniform Crime Reporting) codes, but around 10,000, which is still relatively small compared to the size of the data, ICUR codes that mostly represent financial crimes, unique crimes, and non-criminal police activity, were not specifically classified in the publically available list of ICUR codes. Thus, we could not give a complete analysis of index crimes (to be discussed later) across the entire dataset. Finally, there was a handful of missing latitude and longitude points, which could impede future geospatial information. From this understanding of the dataset, it becomes important to consider the context of this analysis will be completed; an understanding of crime (both in Chicago and generally) provides a useful starting point.

Criminology Background

According to the *Atlantic Monthly*’s article ‘*What’s Causing Chicago’s Homicide Spike*’, in 2016, there were 762 homicides in the city of Chicago, which is a “...58 percent jump in homicides...”, with respect to the number of homicides in 2015, making Chicago one the most dangerous, larger cities in the United States (Ford). On January 24, 2017, President Trump tweeted ‘If Chicago doesn’t fix the horrible ‘carnage’, 228 shootings in 2017 with 42 killings (up 24% from 2016), I will send the Feds!” As one of the primary focuses of the president at the beginning of the year, researchers at the *University of Chicago Crime Lab (UCCL)* published a report, attempting to provide insight into the causes of such violence in Chicago. The report included data on fields such as “social programs, mental-health funding, policing strategies, criminal-investigation clearance rates, gun ownership and more” (Ford). The researchers found no evidence to suggest that these ‘typical’ contributing factors to crime, in particular, murder, have changed significantly over the past few years. For example, educational funding on a 'per pupil' has nearly remained the same, and graduation rates increased by 10 percent--as well as Chicago increasing their funding to 'Department of Public Health, the Commission on Human Relations, and the Department of Family Services' (Ford). Since there does not seem to be a connection between socioeconomic factors and the rise in the murder rate, researchers turned to the police enforcement as a possible explanation.

Considering police enforcement methods in Chicago, the UCCL focused on the number the of 'street-stops' , which is stopping people in the streets who are under suspicion of criminal activity, conducted in the past few years (Ford). The number of 'street-stops' reached 80,000, 60,000 and 10,000, in 2014, and at the beginning and end of 2015, respectively, which could in principle explain the increase in criminal activity and specifically murder (Ford). However, comparing this to a similar program in New York City, the 'stop and frisk', a program which had stopped in 2011, arrests and crimes committed declined annually post 2011 (Ford). Accordingly, it is hard to argue that a drop in stopping/arresting people led to an increase in crime rates in Chicago. Without sociological, economical, or police enforcement serving as explanatory variables to the increase in the number of murders in Chicago, we shift our focus to the broader discipline of criminology.

In Don Weatherburn's article '*What Causes Crime?*', he breaks crime into three categories: *crime-prone individuals*, *crime-prone places* and *crime-prone times*. Under the category *crime-prone individuals*, he lists factors such as biological, familial, intelligence, substance abuse, peers, etc.; under *crime-prone places*, he lists economic inequality, poverty/unemployment, weak informal social controls, gangs/organized crime, lax/insufficient law enforcement, etc.; and finally, under *crime-prone times*, he lists short-term influence and long-term influences.Some of these 'factors' require further explanation, as the names give little definition to the term. For example, *biological factors* point to the notion that criminal behaviors are inherited, while *family factors* point to 'poor/inadequate parenting'(Weatherburn). Furthermore, *weak informal social controls* means a community's inability or unwillingness to police itself, while *lax/insufficient law enforcement* is simply policemen not doing their jobs (Weatherburn). Finally, and most importantly, we have *short-term influence* and *long-term influence*. The former refers to investigating criminal trends in a span of days, months and at most one year, while the latter analyzes the same, however, trends spanning several years. According to Weatherburn, *short-term influence* is still not 'well understood', due to the volatile nature of that time frame. Together with *long-term influence*, we will study both, mainly focusing on *short-term influence*, referring to this as temporal analysis.

Beyond individual, socioeconomic, and policy factors, criminology research has also focused on these temporal factors. In 2007, the *Home Office of England* published preliminary findings, linking seasonality to criminal trends. In summary, Celia Hird and Chandni Ruparel studied 29 crime 'types', where 25 out of 29 of them showed evidence of occurring seasonally. They were able to group them into three categories: crimes that peak in the summer and plummet in the winter, crimes that peak in the winter and plummet in the summer, and crimes that have 'peaks and troughs' each year but do not follow the seasons (Hird). Crimes that belong to the first group were violent and sexual assaults as well as bicycle theft and arson (Hird). Crimes that belong to the second group were property crimes such as auto/tampering theft and domestic/business burglary (Hird). The third group included various crimes stemming from thefts to criminal damages (Hird). Because of these preliminary findings, perhaps seasonal patterns would be of interest to persons analyzing crime in Chicago.

As for crimes that produce trends on an hour-to-hour basis, in a demonstration provided in Felson's '*Simple Indicators of Crime by the Hour of Day*', we find how robbery occurs over the course of a day beginning at 5am and ending at 4:59pm, split into quartiles and having a 'median minute of crime'; that is to say, 50 percent of all robberies occur before the median time, tracing back to 5am (Felson). Collecting data from the police departments of "12 middle-sized American cities" from 1991 to 2001, Felson and Poulsen showed that roughly 33.5 percent of robberies occur during the day (5am-4:59pm) out of a 24 hour period (Felson). Therefore, considering this general trend of robbery, like seasonality, this would also be something to consider when analyzing particular crimes on an hourly basis in Chicago.

Since many reports/analysis of crime in Chicago have been more concerned with socioeconomic factors and police enforcement, our analysis focuses more broadly on the strategies and ideas used in criminology, namely temporal analysis. Since Chicago experiences all four seasons, it would be reasonable to conduct temporal analysis at the month-to-month level as weather patterns change. We are also interested in analyzing hours of the day, as it is likely that certain crimes happen during the night and others during the day, as patterns found in Felson's article suggests. Fundamentally, our analysis will seek to apply the criminologist's perspective of temporal factors to Chicago by asking: "Do temporal factors such as time, date and month significantly (at the 5% level) affect the prevalence and type of reported crimes within Chicago between 2012 and 2016?"

Limitations

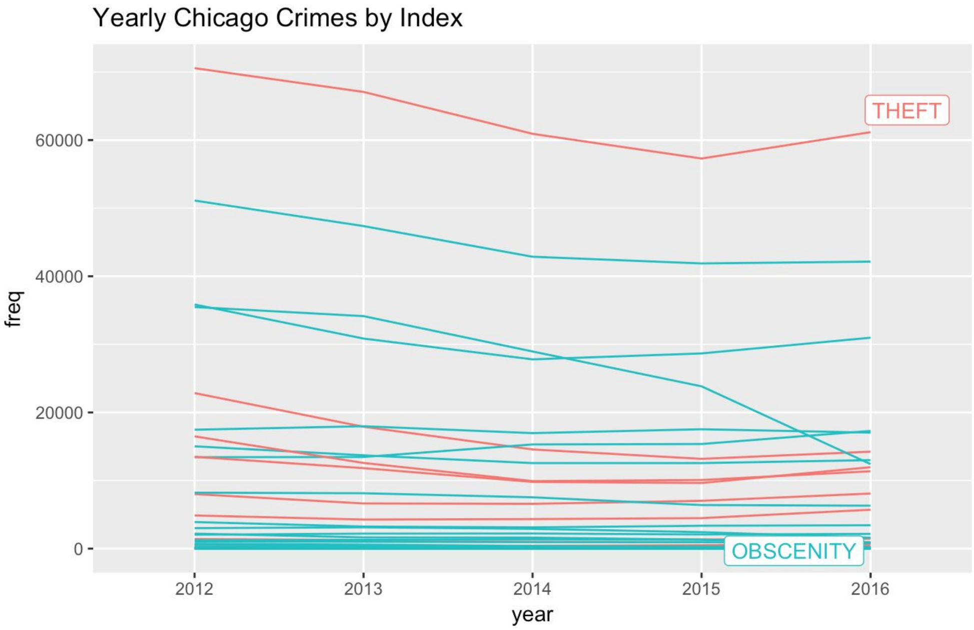
After examining our data closely, it would have been beneficial to have fields concerning financial value of crimes (e.g., theft), crime severity and court outcomes--additionally, collaborative evidence on causal links i.e., the hotter temperatures in the summer causing a spike in crimes and outdoor traffic to explain the dip in criminal activity. Also, we would have benefited from an audit on the data to confirm its accuracy and the ability to analyze the full dataset, information extending beyond the year 2012 (because this required more computing power). Even though we lacked this additional information, it ultimately led us to augmenting our SMART question to a point in which answers were achievable and the results were sound.

Exploratory Data Analysis

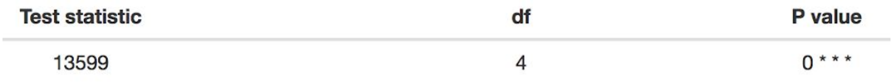
The exploratory data analysis helped us narrow our question and look at the interactions between the variables. When we initially went through the topic literature, most of the studies on either Chicago crime, or crime more generally, seemed to focus on the individual effects of specific factors. However, criminal behavior is often influenced by multiple factors, so it seemed helpful to look at the interaction between our temporal factors. Thus, an important part of analysis became looking at the dependence amongst both monthly and seasonal factors and the hourly factors discussed in the literature. Beyond that exploring the data in more detail also helped us narrow the time frame to 2012-2016. First, the data from 2007-2011 seemed to be coded much differently than the data in the more recent time frame. We determined this by first loading in all of the data and then examining the data from 2007-2011, noticing the large amount of incomplete data (especially on the time stamps and geolocations), and deciding to move in a more realistic direction. Furthermore, although the dataset is updated semi-frequently, analyzing the partial data from 2017 gave us an easy way to see that it was not nearly complete enough of data. Furthermore, for all the tests that we ended up running (with the exception of year to year comparisons), we also tested whether the inclusion of 2017 data would give different results. Given that the results were very similar, we decided to eliminate the 2017 data from our final analysis to ensure we looked at the same dataset for both our annual analysis and our hourly/monthly analyses.

Initial Findings

After going through all of our analysis, we can start to answer our questions. First, at the annual level of analysis, it appears as though the crime rate has fluctuated over time. Specifically, there is a general reduction in crime from 2012 to 2016. In 2012, there were 335,670 recorded crimes, while there were only 265,462 recorded crimes in 2016. Furthermore, it appears that in general, for most types of crime, that decrease held throughout the time frame. However, a closer analysis reveals a slightly more complicated story. When looking at the difference in crime rates while holding the severity of a crime constant, it looks like the trend may not be as consistent. In fact, it looks like a lot of FBI index crimes showed an uptick in 2016.

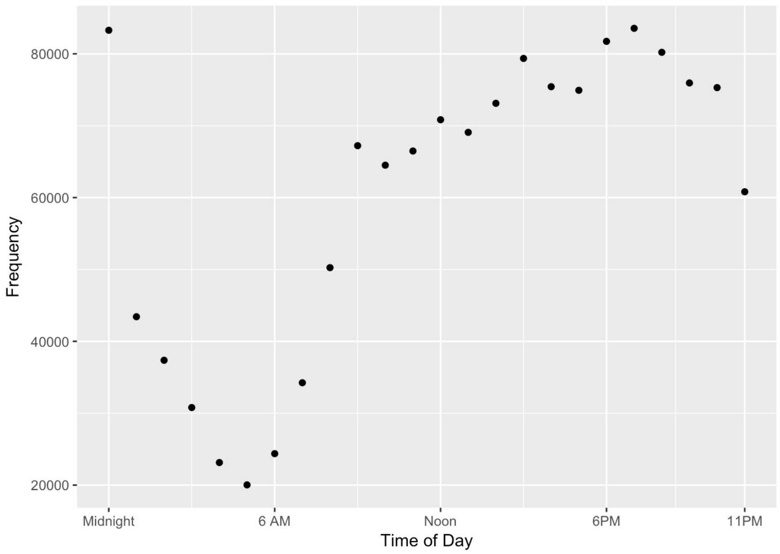


We can test this relationship more specifically using Chi-squared test of independence. The null hypothesis will be that the relationship between year and crime frequency is independent, and the alternative hypothesis is that the relationship between the two is not independent. The results of the test displayed below indicate that the relationship is highly significant and we can reject the null hypothesis.



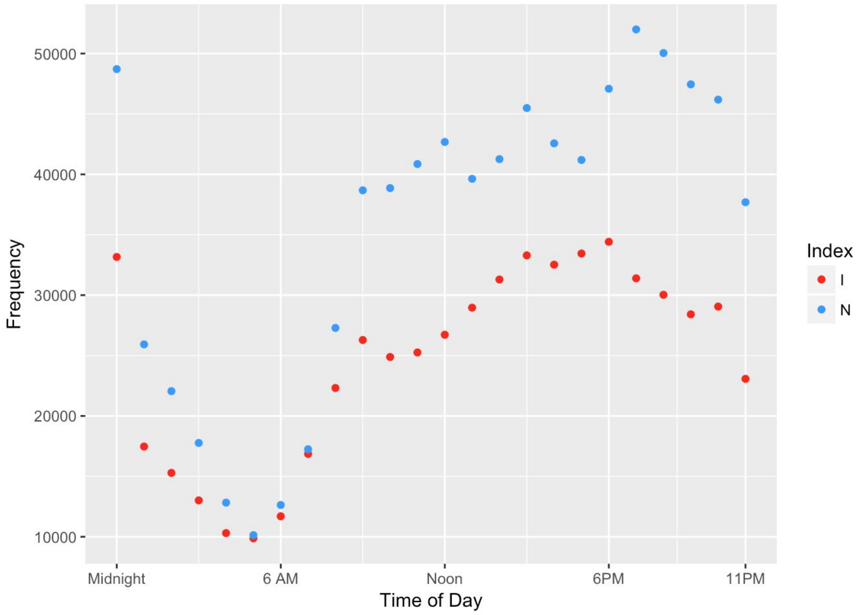
Test statistic	df	P value
13599	4	0 * * *

We can continue this analysis for hourly factors. It makes sense that the likelihood of crimes will fluctuate throughout the day because criminals will probably seek night cover which would make it appear like crimes spike at night. Additionally, factors like alcohol consumption probably spike at night, which could also explain crime spikes. Similarly, crimes will presumably start to decrease as more people go to sleep. The chart below shows that relationship, and a chi-squared goodness of fit test (testing the model that crime would occur at equal frequencies at every hour of the day) shows the strong statistical relationship.



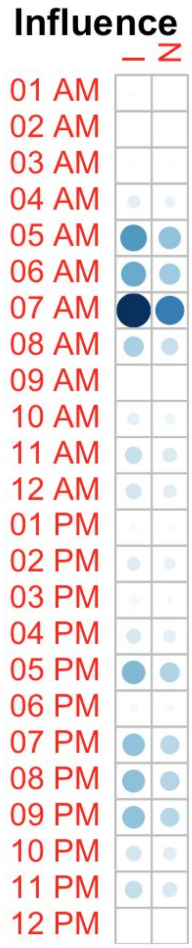
Test statistic	df	P value
172880	23	0 * * *

We can take this a step further, however, and test whether the severity of crime interplays with the hour of crime. It would seem that night cover may be more relevant for index crimes, whose perpetrators presumably have a larger incentive to hide. We can test this with a Chi-squared test of independence, with the null hypothesis being that the two factors are independent.

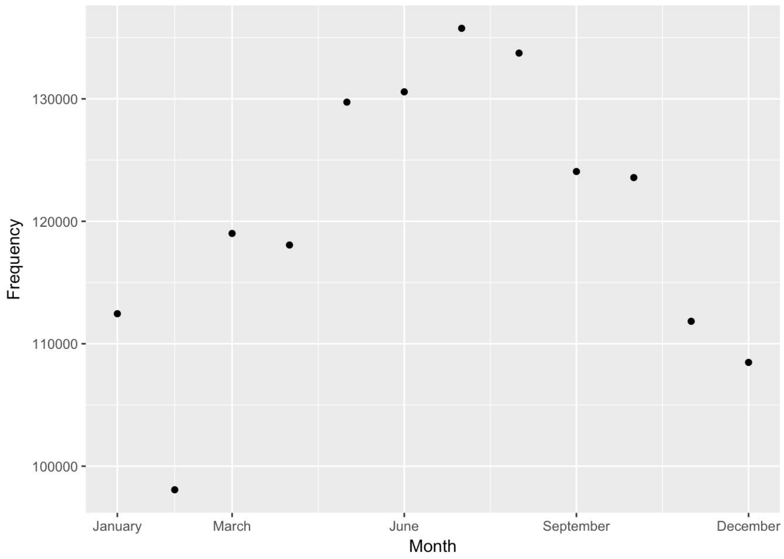


Test statistic	df	P value
5558	23	0 * * *

From these results, it is clear that we can reject the null hypothesis at most significant levels, and can assume that the severity of a crime interplays with the hour of the criminal behavior. The results make it seem as though this is propelled by a spike in non-index crimes throughout the day followed by a sharp decrease in the late night and early morning period, more so than a spike of index crimes at night. Looking at the influence of each hour on the total Chi-squared statistics, it shows that the drop off in crime in the earlier hours of the day, when index and non-index crimes are similar, contributes the strongest to the large Chi-squared value. This follows logically because if the severity and hour were truly independent, we would expect that non-index crimes would remain significantly higher than index crime, when in reality they both converge around 6 AM.

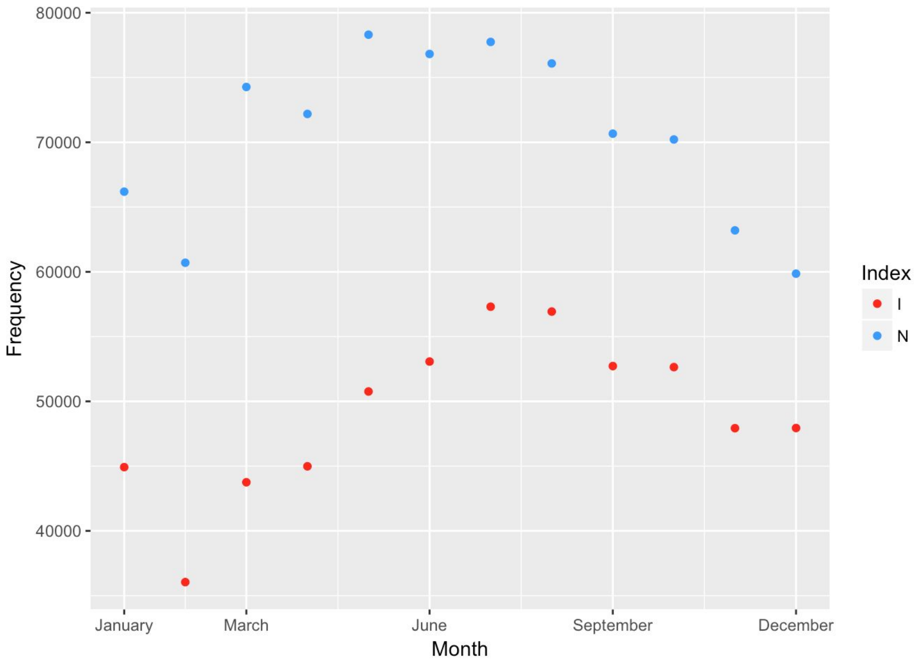


Finally, we can turn towards monthly factors. From the literature, we would assume that crimes will spike during the summer months. The chart below shows that relationship, and a Chi-squared goodness of fit test (testing the model that crime would occur at equal frequencies during every month) shows the strong statistical relationship.



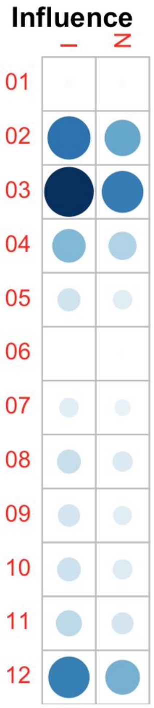
Test statistic	df	P value
11728	11	0 * * *

Like with the hourly analysis, we can test whether the severity of crime interplays also interplays with the seasonal effects of crime. We can test this with a Chi-squared test of independence, with the null hypothesis being that the two factors are independent.



Test statistic	df	P value
3170	11	0 * * *

From these results, the extremely low p-value will lead us to reject the null hypothesis, and we can assume that the factors are related. We can examine the influence plot to try to see what is causing this relationship.



From these results, it appears that the drop off in index crimes during the early months of the year is influencing the significance of the test the most strongly. This seems to be in line with the literature because Hird's research seemed to indicate that the seasonal effects of crime differ depending on the type of crime. It makes sense that there would be a sharper decline in index crimes during the winter because these types of crimes fell in Hird's first group, while lesser non-index crimes often would spike during the winter.

Conclusions

While Chicago crime is an interest of academics and politicians, an important component of the debate must be to look at the underlying causes of crime. However, any sort of advanced statistical test that will seek to provide an explanation of crime will inevitably need to consider other factors that underlay crime at its roots. Thus, this paper sought to provide an empirical test of general criminology hypothesis to data from Chicago. Through our exploratory data analysis, we were able to narrow down annual, monthly, and hourly factors as levels of interest, and test for the interaction of the severity of crime. Given the strong results from our chi-squared testing, it appears that these temporal factors influence both the severity and frequency of crime. Future research on criminology and future discussions about the Chicago crime rates should, therefore, consider how uncontrollable factors (like the year, hour, and month) may be influencing crime.

Bibliography

Felson, Marcus, and Erika Poulsen. *Simple indicators of crime by time of day*. International Journal of Forecasting, 2003, www.forecastingprinciples.com/files/pdf/Felson_and_Poulsen,_Simple.pdf.

Ford, Matt. “What's Causing Chicago's Homicide Spike?” *The Atlantic*, Atlantic Media Company, 24 Jan. 2017, www.theatlantic.com/politics/archive/2017/01/chicago-homicide-spike-2016/514331/.

Hird, Celia, and Chandni Ruparel. *Seasonality in recorded crime: preliminary findings*. Home Office, Feb. 2007, www.popcenter.org/problems/bicycle_theft/PDFs/Hird_Ruparel_2007.pdf.

Kaggle. “Crimes in Chicago.” 2016. https://www.kaggle.com/currie32/crimes-in-chicago.

Weatherburn, Don. “What Causes Crime? .” *Crime and Justice Bulletin*, NSW Bureau of Crime and Statistics Research , Feb. 2001, www.bocsar.nsw.gov.au/Documents/CJB/cjb54.pdf.