Predicting Crime Rate Based on Monthly Event Data

Thomas Collas, Trevor Garner, Sam Lefar, Morgan McGuinn, Ziad Moukadem, Alex Wang

Introduction, Problem, and Motivation

Most cities of any appreciable size host public events, where an event can be anything from a large-scale concert to a convention to even a small-scale sale at a popular pub. In general, anything that can be described as "something that happens in public and draws citizenry to a singular area to participate" can be considered an event, for the purposes of city planning. Most public events tend to be fundamentally good for the city; but despite this, we were worried about potential drawbacks and costs of holding events—namely the potential for more crime. There is a large body of literature across many countries suggesting that when there are many events in a city, crime rates go up significantly [2, 6, 7, 12]. In particular, locations designed to attract tourists have a very large impact on a city's crime rate compared to other events, as a result of encouraged rowdiness and liveliness [9, 10].

There have been numerous attempts to use these data to predict future crime, i.e., given historical crime data for neighborhoods of a city, where is it most likely for future crime to occur? These past attempts use many different models, such as Naïve Bayes, Decision Trees, Bayesian structures time series, regression, LSTM and heat spot tracking; and they produce results that have accuracy ranging between 50-75% [1, 11, 14, 17, 18]. These models can be incredibly useful for standard day-to-day operations, but they do not tackle the idea of events and how they may change crime rates. Some approaches even explicitly talk about how their results are meant to be used for one specific problem, like finding optimal police patrol routes or determining how pandemic outbreak affects the long-term crime prevention [4, 13].

All of the above rely on the idea that crime rate is generally static, that unless there's some great upheaval in population, demographics, or economic state of an area, then crime rate will stay roughly the same—this idea is so deeply ingrained in the current literature that at least one paper has been written that actively encourages use of long-term trends *over top of* short-term trends [16]. The above models don't consider that there is, for example, a Hawks playoff game today, which may cause crime rate in downtown Atlanta to skyrocket. In particular, this is problematic for visitors to the city, i.e., tourists.

There are many events stretching across a wide variety of interests and scales (playoffs for any of Atlanta's four major sports, Music Midtown and other concerts, DragonCon and other conventions, political rallies, and many more) that draw visitors to the city, visitors for whom the regular crime rate predicted by the above models would not matter nearly as much as event-driven spikes. And this is to say nothing of how those in charge of city planning may be led to make poor decisions in terms of which events are *worth it* for the city to host. Destruction of property, city image, emotional costs to inhabitants, and other side-effects of crime, all have reasonably estimable costs that should be considered when determining how much benefit there is to gain from hosting an event [5]. It's still possible to make intuition-driven decisions regarding what events to host or not—e.g., hosting the Olympics is never worth it by any objective measure, but cultural prestige is too great for many to pass up—but if any decision is to be *data driven*, it's irresponsible to ignore this cost-benefit analysis.

Somewhat separately, but still a problem, is the idea that crime rate appears to follow different distributions depending on where in the world you are. Some studies have been done, for example, that explicitly confirm that the topography of an area is important to determining crime rate; others explicitly deny that topography is important at all [6, 8]. Because of this, any model that is made in the abstract to predict crime rates is not likely to work properly; it must be trained on data specific to a particular area (in

this case, Atlanta). No matter how broad the body of literature across the entire world, the results are suspect unless they are contextualized to a specific area; Atlanta needs its own models to describe it.

In response to this lack of research into factors that cause short term crime spikes, especially in the city of Atlanta, our group has decided to create a model that will consider event data and figure out how it affects neighborhoods and a graphical user interface on a website that will allow users to figure out on a monthly basis, which areas of Atlanta are safest. Our innovative product and ideas lend themselves to a superior model and easily analyzable results for the city of Atlanta.

Methodology

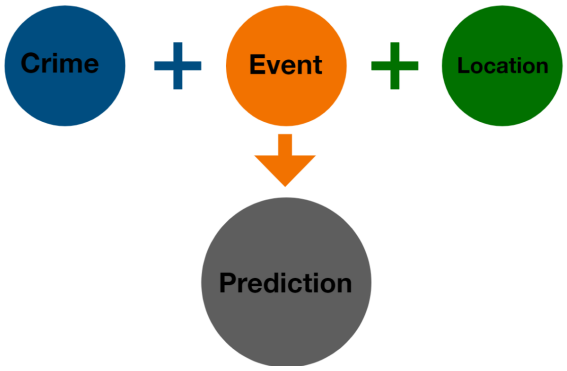

Figure 1: Methodology of prediction

Our final product can be decomposed into two basic pieces: the model and the visualization. The model outputs the predicted number of crimes committed for a month and illustrates how local events affect the amount of crime within a specified neighborhood monthly. A user can interact with and analyze the model through a website that presents the monthly predictions in the form of a geographic map and allows the user to view metrics describing the current and predicted crime rate of any 'beat.' For reference, a beat is a police-drawn boundary typically consisting a of a few blocks, used to specify a location for stationing on-duty officers.

Our final dataset consists of information given by the following three datasets: A violent crime dataset which is available in already-processed form on the APD public website; an Atlanta demographic dataset from the City of Atlanta Neighborhood Statistical Areas [19]; and an event dataset put together by scraping from Google maps, Google events, and the PredictionHQ events API. The final dataset provides the number of violent crimes committed within each neighborhood for the month of March in 2021. The dataset additionally includes the beats, population, percent of the population that is Caucasian, median household income and number of major events for each neighborhood. This particular combination of features sets this model apart from previous implementations. The incorporation of local events as a feature should be especially noted. In terms of features within event data, the date, name, location, and category/type of each event were tracked.

A variety of regression models were tested using the features in the dataset, including both linear and negative binomial models, and a Poisson regression model was settled upon as the superior model due to the comparably lower standard error and root mean squared error (rMSE). Because we are using crime as count, assuming the data follows a Poisson distribution is reasonable.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.372e+00  8.669e-02  27.357  < 2e-16 ***
white        1.323e-02  2.242e-03   5.903 3.58e-09 ***
income      -1.164e-05  2.435e-06  -4.781 1.75e-06 ***
events_count 1.939e-01  9.743e-02   1.990   0.0466 *
pop          7.542e-05  1.006e-05   7.499 6.41e-14 ***
```

Figure 2: crime_rate ~ white + income + events_count + pop

Though the model indicates the number of events in the month as a statistically insignificant variable, the rMSE for the model with the events feature did decrease compared to the model without the events feature. A plethora of types of events are incorporated in the dataset, including but not limited to concerts and flight delays. However, only sporting events, festivals, and events with large public gatherings were included. The additional events incorporated noise and did not have the desired effect on the model. The rMSE dropped slightly on filtering events, and the p-value for event features dropped significantly.

The final visualization was created using HTML, CSS, JavaScript and d3.js. And designed so a user can quickly and easily make informed safety decisions for their time in Atlanta. The visualization encompasses the following topics: 'Predicted Total Crime,' 'Previous Total Crime,' 'Total Crime Per Neighborhood,' and 'Total Events Per Neighborhood.' While alternative data could have been displayed, these topics were chosen because they encapsulate the most critical information to the user.
The visualization can be broken down into three pieces: the map, the written data and the graphs. The map is segmented by beat and color coded based on the predicted number of crimes committed. When a user clicks on a beat the written data appears on the upper right of the interface. This includes information specific to the selected neighborhood where the beat is located.

The information shown includes the beat number, neighborhood name, key landmarks in the beat (e.g. Georgia Institute of Technology will be listed for beat 504) and the number of crimes committed in March and predicted in April. The crime values are both color coded based on whether the crime value is above or below the average and accompanied by additional metrics - the mean, minimum and maximum. The graph section displays two bar graphs which show the ten neighborhoods with the highest number of crimes committed and total events for the month. When a beat is clicked the neighborhood where the beat is located is highlighted on the bar graphs. If the neighborhood associated with the beat is not in the top ten, the lowest bar will be replaced with the values for the specified neighborhood. This feature was incorporated so the user is better able to understand both the correlation certain events with a higher amount of crime. Additionally, a user can pinpoint a location on the map by entering an address or popular landmark for the user to contextualize his exact surrounds.

Experiments/Evaluation

In this section, we talk about the three fundamental goals our model wishes to accomplish, some hurdles we overcame in the implementation, and the potential sources of error that must be acknowledged and considered before we can declare the model a success. These fundamental goals are divided into three considerations, and will be discussed below in descending order of importance.

The primary consideration our model is designed to answer is, "In the city of Atlanta, given historical crime rates, how do those crime rates in individual beats of the city change on a monthly basis as a result of the events that are occurring in the city across all beats?" e.g., if there is a medium-sized festival near Atlantic Station, how does that affect crime around Five Points? Depending on the exact neighborhood in which that event takes place, our model may give a very good answer—but for lower-crime neighborhoods, it may not. Because event data tends to be sparse, some beats ended up with 0 events for long stretches of time, and as a natural consequence, there isn't much data to describe how they work. We can only aggregate units of time so much before the results become meaningless (nobody is interested in how events affect crime *yearly*), and so we have to accept that our model is bad at predicting crime in low-crime areas. While not very clean mathematically, practically this is an acceptable source of error; in the sense that safe areas, even if more dangerous during events, still tend to be safe. Northside at *triple* its usual crime rate is still less dangerous than Downtown. Making a mistake in Northside's crime rate is not nearly as problematic, practically speaking, because it is less likely to make a difference in how residents, planners, or visitors perceive the area; nor is it likely to accidentally spur more crime in the area as a result of lax precautions.

For the second consideration, implicitly, because it needs to track spikes in crime rate, our model also needs to be able to identify resting crime rate (or else it wouldn't be able to determine changes in crime from 'the usual'). So, although not quite as novel as its first function, our model must also be able to answer questions regarding which neighborhoods of Atlanta are *usually* dangerous or safe overall. In the end, we decided the best way to draw a frame of reference for our model's predictions (i.e., the frame of reference for the prediction being considered "high" versus "low") was to suggest whether the crime rate was up or down from last month. We initially considered other metrics, such as using the average crime rate over all time as a benchmark, but because event data tends to be rather sparse, the most recent data should be considered as more important and given a stronger bias than older data. The most elegant way we found to account for that consideration was to simply make a comparison with the previous month's prediction. With a large enough amount of extra time, a more precise solution could possibly be found, weighting past months according to a sliding scale of recency; but with the given production time for this model, we believe that the decision is the best we could have made. This also has the additional side-benefit of not requiring the website to store, make calculations on, and automatically handle an ever-growing number of crime data files; it only needs the two most recent, and it will *always* be small.

As a consequence of question number two, our model needs to be able to figure out which factors specifically are important to driving Atlanta's crime rates. We don't want to misrepresent the importance of monthly events on crime rates, so we need to make sure our model still includes the usual demographic and long-term data as others as well. This makes sure our model can also answer questions regarding what causes Atlanta *specifically* to tick, potentially providing a cleaner insight into what sorts of crime prevention measures may work best. Just as with question 2, we first experimented with the idea of just including *everything*—every demographic feature we could get our hands on, as detailed and granular as possible. This led to a bit of a mess, where p-values were high and the coefficients of certain factors would change radically depending on which *other* factors were included. The most prominent example is was that including a separate factor for each ethnicity and racial group inhabiting an area caused them to *change signs* depending on how many there were. We believe that this irregular behaviour is a result of highly-interrelated factors, and simplified down to just tracking white population percentage. All other demographic factors, like population, were kept, and the model performed acceptably.

Throughout the process of attempting to answer the above questions, we experimented with training our model on different combinations of factors, attempting to find the one that had the lowest mean squared error and standard error. Over the course of our experiments, we found that not all event data are created equal. Including the entirety of all event types in our model *did* decrease the rMSE, but the p value was around 0.5, indicating that the variables may not all be statistically significant. Our lowest p-value of 0.046 was obtained when we only included events with the festival, sport, or terror classifications. The grand majority of events in the city of Atlanta are those with the "concert" classification, but including those in our data filter raises the p-value to 0.17. We also tried filtering events by whether or not they were *planned*, which removed most of the terror events as well; when we did, the rMSE dropped by a small margin, but this did not largely affect which neighborhoods are the safest or not.

One known weakness is that our model cannot consider future terror events when making predictions (since terrorist events are *never* known ahead of time)—but no model meant solely to predict crime rates can. A model that considers future terror events *must* be paired with an additional model that is meant to predict how many terror events will happen in a given area. Due to the sparseness of events as a whole, let alone terror events, and due to how such a model only intersects with our goals at a very specific point, building such a helper model is considered an exercise that extends beyond the intended scope and allotted time of the project.

Another unexpected result is that for the city of Atlanta, crime is only significantly affected by sporting events, terror events, and festivals, with a hanging question mark about whether concerts (by far the most common event type) are impactful. The naïve view is that this would lower the importance of our model—after all, if most events don't matter, who cares?—but even leaving aside the usefulness of *knowing* that most events aren't impactful and that festivals and sporting events *are* impactful, just on a broad level of analysis, we *still* can extract value from the product. Since the literature mentioned in the Introduction suggested that crime events are strongly correlated for many cities, it becomes interesting to think why Atlanta does not follow the trend—such an exploration could be the topic of an entirely new project all on its own. And, at the end of the day, we did indeed achieve our goal, which was to find an answer to the question of how an event affects crime on a monthly basis in Atlanta. The answer we found just happens to be "probably not much."

Conclusions/Discussions

By the end of the project, were able to create a model that tracks crime rates with a rMSE of 7.34, where our data has a mean of 14 and a standard deviation of 9.33. Despite the low performance that this may imply, we have already discussed and acknowledged several large sources of error. Because that error is somewhat predictable and appears in ways that are relatively harmless, the model is acceptable for the cases in which it needs to be used, and is therefore a good source of predictions for our website.

Now that we have the model, there are a limitless number of potential applications, only bounded by the creativity of the thinkers and the space we have to describe those applications. Specifically, one can ask about whether large and small events have effects directly proportional to one another, or about how the model could adapt to future changes in crime patterns, or even less technical questions like "when is a good time to visit Atlanta?"

The visualization website itself also succeeds at its stated goals, too. The intended use of the website is to let people know quickly and easily which areas of Atlanta are particularly dangerous this month, based on past data, and a user is able to quickly and easily do just that. We were even able to implement a feature of the website where a user searches for a large landmark of Atlanta and its location is displayed on the map, for easier locating of relevant beats. It's a lot harder to determine quality of a website than a statistical model, as user experience and web design is more subjective than a single error number, but given the color scheming and accompanying legend, the multiple graphs for general information, and the ability to search for landmarks, it can safely be said that at the very least, every visitor to the website will be able to understand what sort of metrics they are looking at, and have a good idea of how to find specific pieces of information that they want. In the abstract, as long as those goals are met, the website satisfactorily achieves what it has set out to do.

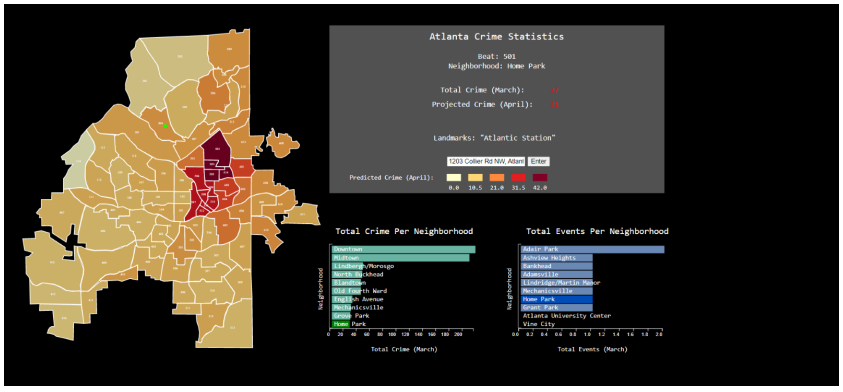

Figure 3: Website Overview

Work Repartition

In terms of effort contributed to the project, each member contributed similar amounts of effort. This report will end with a brief description of each group member's contributions:

Sam Lefar wrote the entirety of the proposal document, as well as the Introduction, Problem Statement, Experiments, and Conclusions sections of the Midterm Report paper. Finally, he is responsible for the Conclusions section of this Final Report, as well as general editing on the other sections. He is not responsible for most of the *contents* of the document, i.e., he did not decide what to include himself, but he took group discussions, etc., and translated them into readable format.

Ziad Moukadem was responsible, along with Morgan, for the entirety of the Project Proposal PPT and Video; and later did the entirety of the group poster on his own. In addition, while not an individual, concrete act, when at meetings he tended to take a pseudo- administrative role, asking good questions regarding scope and scale of what our project *ought* to accomplish, and generally keeping the conversation moving.

Morgan McGuinn was responsible, along with Ziad, for the entirety of the Project Proposal PPT and Video. In addition, she wrote the Methods section of and was chief editor for the midterm document; and finally, she rewrote the Methodology section of this paper to account for changes in strategy and updates to our model. During the initial planning phase for the project, she was most useful with brainstorming topics, including the one that we decided to flesh out into the project you are about reading now.

Alex Wang is responsible for nearly the entirety of the model itself. He gathered crime data, cleaned it and joined with the demographic data, built and tested possibilities off of Trevor's shortlist using a variety of different options each time, and selected the best results from among the candidates. Testing for him included not just different model types, but a large variety of feature arrangements as well, trying to figure out what sort of parameters were important to optimizing each type of model that he tried.

Trevor Garner provided backup support on each technical aspect of the project, aiding Alex and Thomas in their most data gathering/cleaning, but his most notable contribution is the grand majority of the D3 website. He chose how to build the interface, how the website ought to read outside files, etc. Perhaps most importantly, he was responsible for drawing the boundaries of each beat, which was a significantly larger challenge than we had expected up front.

Thomas Collas, aside from his general support for Alex and Trevor, was in charge of event data, which included finding sources, scraping Google Maps and Events, and really just managing to get whatever aspect of event data that we needed. His research and coding covered all sorts of local events and created an algorithm to generate a clean event dataset, and *also* figured out how not just to translate addresses into GPS locations, which was, like beat boundaries, a larger challenge than initially expected.

References

[1] Almanie, Tahani, et al. Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots. *International Journal of Data Mining & Knowledge Management Process*. 2015, 5, No. 4. https://arxiv.org/pdf/1508.02050.pdf

[2] Baumann, Robert, et al. Sports Franchises, Events, and City Livability: An Examination of Spectator Sports and Crime Rates. *The Economic and Labor Relations Review*. 2012, 23(2), 83-98. https://doi.org/10.1177/103530461202300207

[3] Belkhir, Mohamed El Amine Abdellah and Maamar Benaouda. Modélisation UML et mise en place des indicateurs d'analyse spatio-multidimentionnelle dans un SOLAP pour la gestion de sécurité dans la ville de Mostaganem. *Ministere de l'Enseignement Superieur et de la Recherche Scientifique*. 2015. http://e-biblio.univ-mosta.dz/bitstream/handle/123456789/9438/MINF103.pdf?sequence=1&isAllowed=y

[4] Campedelli, Gian Maria, et al. Exploring the Effects of COVID-19 Containment Policies on Crime: An Empirical Analysis of the Short-term Aftermath in Los Angeles. *American Journal of Criminal Justice.* 2020. https://doi.org/10.1007/s12103-020-09578-6

[5] Chalfin, Aaron. Economic Costs of Crime. In *The Encyclopedia of Crime and Punishment*, W.G. Jennings (Ed.) 2015. https://doi.org/10.1002/9781118519639.wbecpx193

[6] Chawla, Meenu, et al. A Clustering Base Hotspot Identification Appraoch for Crime Prediction. *Procedia Computer Science.* 2020, 167, 1462-1470. https://doi.org/10.1016/j.procs.2020.03.357.

[7] Demeau, Elodie and Geneviève Parent. Les facteurs de la distribution spatiale de la criminalité à Montréal : l'importance des bars. *Revue Internationale de Criminologie et de Police Technique et Scientifique*. 2017, 70. https://serval.unil.ch/resource/serval:BIB_6010837BB7F3.P001/REF

[8] Gleyze, Jean-François. Apport de l'information géographique dans l'analyse des risques. Application à l'étude des perturbations du réseau routier à la suite de catastrophes. 2001.

[9] Han, Sungil, et al. Crime Risks Increase in Areas Proximate to Theme Parks: A Case Study of Crime Concentration in Orlando. *Justice Quarterly*. 2019, 20. https://doi.org/10.1080/07418825.2019.1677935

[10] Homel, Ross and Steve Tomsen. Hot Spots for Violence: The Environment of Pubs and Clubs. In *Homicide: Patterns, Prevention, and Control*, Leather Strang and Sally-Anne Gerull (Ed.). 2009. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.840.4429&rep=rep1&type=pdf

[11] Khanwalkar, Sanket Subhash. Crime Intelligence 2.0: Reinforcing Crowdsourcing using Artificial Intelligence and Mobile Computing. *UC Irvine*. 2016. https://escholarship.org/content/qt6965r2v6/qt6965r2v6.pdf?t=oeipre

[12] Lee, Yong Jei, et al. How concentrated is crime at places? A systematic review from 1970 to 2015. *Crime Science Journal*. 2017, 6, 6. https://doi.org/10.1186/s40163-017-0069-x

[13] Lin, Ying-Lung et al. Grid-Based Crime Prediction Using Geographical Features. *ISPRS Int. J. Geo-Inf.* 2018, *7*, 298. https://doi.org/10.3390/ijgi7080298

[14] Perry, Walter L., et al. Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. *Rand Corporation*. 2013. https://doi.org/10.7249/RR233

[15] Ribaux, Olivier and Stéphane Birrer. Système de suivi et d'analyse des cambriolages appliqué dans des polices suisses. 2008. https://serval.unil.ch/resource/serval:BIB_6010837BB7F3.P001/REF

[16] Scheider, Stephen. Predicting Crime: A Review of the Research, Summary Report. *Canadian Department of Justice.* 2002. https://www.justice.gc.ca/eng/rp-pr/csj-sjc/jsp-sjp/rr02_7/rr02_7.pdf

[17] Wang, Hongjian, et al. Crime Rate Inference with Big Data. *Knowledge Discovery and Data Mining*. 2016. https://doi.org/10.1145/2939672.2939736

[18] Wang and Yuan. Spatiotemporal Analysis and Prediction of Crime Events in Atlanta Using Deep Learning. *IEEE 4th International Conference on Image, Vision, and Computing.* 2019. https://doi.org/10.1109/ICIVC47709.2019.8981090

[19] Atlanta Regional Commission Open Data and Mapping Group. "City of Atlanta Neighborhood Statistical Areas." Open Data, 15 July 2020, opendata.atlantaregional.com/datasets/d6298dee8938464294d3f49d473bcf15_196?geometry=-84.942%2C33.667%2C-83.899%2C33.867.