UCLA Computational and Applied Mathematics REU

Final Report for
LA Homicide Narratives

Trained to Kill: Analyzing Homicide Data in Los Angeles County

February 23, 2022

Team Members

Elijah Gross-Sable (University of California, Los Angeles)
Jacky Lee (Harvey Mudd College)
Xia Li (University of California, Los Angeles)
Tyler Sam (Harvey Mudd College)
Nate Sands (City College of New York)

Principal Investigators

Andrea Bertozzi
Jeffrey Brantingham

Mentors

Michael Lindstrom
HanQin Cai

Abstract

Provided with homicide data from public resources and classified documents from the Los Angeles Police Department, we explore applications of machine learning and other modeling methods to discover the nuances between homicides. While homicides are generally unpredictable by nature, we believe that there may be some element of underlying structure to homicide execution (on the suspect end) and solvability (on the policing end). We employ a novel method of optical character recognition to read the data out of a typewritten book and proceed with various techniques, including latent Dirichlet allocation and variations of it, dynamic modeling, and geographically weighted regression. We discover interesting trends in the data with respect to solvability, dynamics, and other patterns.

Acknowledgments

This research project was supported the following grant: ATD: Sparsity Models for Forecasting Spatio-Temporal Human Dynamics NSF DMS-1737770.

We would like to acknowledge Professor Andrea Bertozzi and Professor Jeffrey Brantingham for sponsoring our project for the REU. We would also like to thank Dr. Michael Lindstrom and Dr. Hanqin Cai for mentoring us and providing guidance and insight throughout the course of our project. We would additionally like to thank the Los Angeles Police Department for providing homicide data to us for analysis. Furthermore, we want to thank the UCLA Math Department and IT staff for providing support for us throughout the project.

Contents

Abstract i

Acknowledgments iii

1 Introduction 1

2 Data Collection and Preprocessing 3
- 2.1 LA Times Homicide Report . . . 3
  - 2.1.1 Web Scraping . . . 3
  - 2.1.2 Latitude and Longitude Extraction . . . 4
- 2.2 LAPD Records . . . 4
  - 2.2.1 Optical Character Recognition . . . 4
  - 2.2.2 Text Extraction . . . 5
- 2.3 LA Open Data . . . 7
- 2.4 Census Tract Data . . . 7
- 2.5 Zillow Data . . . 8

3 Data Visualization 9
- 3.1 Data Trends . . . 9
- 3.2 Rate Parameter Estimation . . . 9
- 3.3 Spatial Correlation . . . 10
- 3.4 Factor Analysis . . . 11

4 Methods 17
- 4.1 Topic Modeling . . . 17
  - 4.1.1 Non-negative Matrix Factorization . . . 17
  - 4.1.2 Word2Vec . . . 18
  - 4.1.3 Semantic Non-negative Matrix Factorization . . . 18
  - 4.1.4 Latent Dirichlet Allocation . . . 19
  - 4.1.5 SCHOLAR . . . 20
- 4.2 Recurrent Neural Networks . . . 21
  - 4.2.1 Long Short-Term Memory . . . 21
  - 4.2.2 Gated Recurrent Unit . . . 22
- 4.3 Dynamic Models . . . 22
  - 4.3.1 Uniform Criminal Population . . . 23
  - 4.3.2 Multiple Gang Populations . . . 24
- 4.4 Geographically Weighted Regression . . . 25
  - 4.4.1 Bandwidth Selection . . . 26
  - 4.4.2 Data Preparation . . . 26

5 Results 27
- 5.1 LA Times Dataset . . . 27

5.1.1 Geographically Weighted Regression 27
5.1.2 Topic Modeling 28
5.1.3 Homicide Forecasting 29
5.2 Topic Modeling on the LAPD Records 30
5.3 Dynamic Models 31
5.4 LA Open Data 32

6 Conclusion **37**
6.1 LA Times 37
6.2 LAPD Records 38
6.3 Dynamic Models 38
6.4 LA Open Data 38

7 Future Work **41**
7.1 Web Scraping 41
7.2 Optical Character Recognition 41
7.3 Spell Checking 42
7.4 Improving the Binary Classifier 42
7.5 Dynamic Models 42
7.6 LA Open Data 42
7.7 Outlier Detection 42

OCR Process on Sample Image **43**

Derivation of the Poisson MLE **59**

Section 1

Introduction

Homicide is a crime that has long been difficult to characterize. Defined in California as "the unlawful killing of a human being or fetus with malice aforethought," homicides can be difficult to distinguish from one another because of this very broad definition (36). However, behind these killings, there are motivations, witnesses, weapons, and a plethora of smaller details which might be useful for exploring and preventing them in the future.

Los Angeles has long harbored serious gang violence, which has made up a significant portion of homicides throughout the city's history. But as the second most populous city in the United States, its homicides are not isolated to the gangs. While a general trend can be observed of higher homicide rates in these gang-dominated areas, there are frequently altercations throughout the city resulting in homicides separate from gang violence (33). We were interested in exploring any underlying structure that might be found within the data.

Social disorganization theory originally sought to correlate the occurrence of crime with conditions in a certain location, or neighborhood. According to the theory, originally developed out of the University of Chicago in the 1920s, high levels of residential instability and poverty lead to communities that are socially disorganized, which in turns produces high levels of juvenile delinquency. In 2011, Regoeczi et al. attempted to extend that theory beyond the occurrence of crime, observing instead how neighborhood context affects the aftermath and responses to these crimes (37). We were interested in these qualitative features, which might allow us to extract latent information about the homicides themselves.

More recently, crime topic modeling has entered the spotlight as the key to finding this latent information. In 2017, Bertozzi et al. applied a hierarchical rank-2 non-negative matrix factorization (NMF) algorithm with frequency-inverse matrix to detect crime topics and considered the crime types as mixtures of different crime topics (31). There are certain issues when it comes to these types of analyses. For one, the text records entail a massive loss of information due to the topical upper limit. Second, the crime type labels themselves may harbor both intentional and unintentional errors. In our analysis, we sought to expand on this method and examine techniques that might avoid similar roadblocks.

However, there have also been outcries regarding these predictive methods. Lum et al. demonstrated how biased police data sets lead to discriminatory reporting, suggesting that a predictive policing software would create a feedback loop to reinforce stereotypes and racial bias (32). Thus, it is imperative that in the future we decide how to move forward with the data that we have, with respect to homicides or anything else.

For the purposes of this analysis, we had access to two primary data sets: a blog-style report of homicides in LA compiled by the LA Times, and a classified book of homicides from the year 1980, which we received directly from the Los Angeles Police Department (LAPD). While these two sources each contain a mix of individualized information over different time frames, there are certain aspects of each that can be considered comparable for the purposes of our analyses.

Both the LA Times and the LAPD data sets contain a text description of the homicide in question. We perform a series of topic modeling techniques on these descriptions, including non-negative matrix factorization (NMF), latent Dirichlet allocation (LDA), and variants of each. However, it is important to note the nuances of how each of these text descriptions was formulated. In the case of the LA Times data, the descriptions were written by journalists, with varied levels of engagement and detail. For the LAPD data set, the descriptions were all written by a detective within 48 hours of the homicide. For this reason, the perspectives of the text descriptions are different and their topic modeling results need be interpreted as such.

Separate from the text descriptions, each data set includes metadata relevant to each homicide. These include weapon type, victim and suspect characteristics, and location. We use this metadata from the LA Times to perform preliminary spatial and temporal analyses since the year 2000. With the LAPD data, we utilize the case status field in order to develop a predictive model for whether a case will be open or closed within 48 hours using its latent topics.

In Section 2, we introduce the various data sources and describe how each was processed in order to perform necessary analyses. It is in this section that we also describe a novel method of Optical Character Recognition (OCR) for reading poorly scanned, typewritten pages of the LAPD book. In Section 3, we describe some preliminary data visualization results in order to give the reader an idea of the Los Angeles homicide landscape over the relevant time frame. We proceed in Section 4 with descriptions of the mathematical and computational techniques employed in order to obtain our results, which are set forth in Section 5. We summarize our findings in Section 6 and describe opportunities for future work in Section 7.

Section 2

Data Collection and Preprocessing

Before we were able to perform any analyses, we had to first collect the data and perform some preprocessing. Each dataset demanded a different collection and preprocessing method.

2.1 LA Times Homicide Report

The Los Angeles Times (6) website has a homicide report (5) that keeps track of every homicide that has occurred in Los Angeles County since 2000. Each page of the homicide report contains information such as name, address, age, cause of death, race/ethnicity, and description. The description of the homicide is of interest to us since we would like to analyze the descriptions using topic modeling.

2.1.1 Web Scraping

We collected the data from the homicide report by writing a web scraper that would go through each page of the website, scrape relevant information, and collect it in a data frame. We collected approximately 17,000 records from the website. The information we were able to collect included:

1. Name of victim
2. Article post date
3. Date of homicide
4. Neighborhood of Homicide Location
5. Address of Homicide Location
6. Age of victim
7. Gender of victim
8. Cause of death
9. Race/ethnicity of victim
10. Police agency of record

11. Existence of police involvement

12. Description of homicide

13. Comments by site visitors

One notable field is the comments made by site visitors in the comments section. As the homicide report is online, people are able to put comments in any of the pages of the homicide report, generally expressing condolences or other sorts of emotions. While we did not focus any of our analyses on this section, the emotional and unfiltered nature of this text could prove valuable in future work.

2.1.2 Latitude and Longitude Extraction

Although we were able to extract the neighborhood and address of the homicide location, we were not able to directly use this since the computer does not know how to interpret this information. To make this more interpretable for the computer, we used the Google Maps API to query the addresses and receive the latitude and longitude of the location of the homicide.

2.2 LAPD Records

We received from the Los Angeles Police Department (LAPD) a book (9) containing confidential records of homicides that occurred in 1980 in Los Angeles County. Each page of the book contained information for a single homicide and the book in total contained 984 records.

To extract this information for analysis, we first scanned each page of the book using a Samsung Galaxy S9+ phone. We did not use a prepackaged scanner because it proved to be less effective in getting high quality images given the age and density of the book. Given that this book has been classified, we were also unable to unbind the pages.

Since these images were not professionally scanned, there were some artifacts that made processing challenging. One of the biggest issues was the bleedthrough from the other pages. Since the pages of the book were thin, text from the page underneath remained exposed. Another issue was the curvature of the page. Since the book was quite thick, some of the pages were curved during scanning. Another issue that came up was that the background outside of the page was noisy. This made it hard for the computer to determine whether something was junk or text.

Our processing of the book images can be broken down into optical character recognition (OCR) and spell checking.

2.2.1 Optical Character Recognition

We used image processing techniques and OCR to localize and extract text. The main idea was to determine regions of text that were locally adjacent so we could both read the text and have information regarding where the letters and words were relative to each other.

Image Binarization

First, we binarized the image to make it easier for the computer to understand. We set each pixel value to either black (0) or white (255). This was a two step process.

First, we used Otsu thresholding to binarize the image. Otsu thresholding determines, based on the distribution of pixel values over the whole image, an optimal value to set as the threshold. Pixel values less than the threshold are set to 0 while pixel values greater than the threshold are set to 255. For the most part, Otsu thresholding did a good job in finding a threshold that eliminated most of the bleedthrough text.

Once Otsu thresholding was applied, we applied adaptive Gaussian thresholding, another binarization technique. As opposed to Otsu thresholding, adaptive Gaussian thresholding uses local information to set the value of a pixel. The purpose of this was to get rid of artifacts which may have been incorrectly classified by Otsu thresholding.

Text Localization

Once binarization was complete, we wanted to group characters based on proximity and put bounding boxes around groups of characters that were close to each other. To do this, we first analyzed connected components in the image to better filter out artifacts and noise. A connected component is a group of white pixels. Based on the size and density of the connected component, we either kept it in the image or removed it.

We then applied an erosion kernel in preparation for the next step, convolving our image with a minumum kernel. Essentially, the means that we iterate using a sliding window and at each iteration we take the minimum value in the sliding window. We did this to get rid of any noise that might not have been taken care of by our noise removal step described above. The erosion risks making the text unreadable due to most of the pixels being eroded; however, we could ignore this since we were only trying to localize the text at this point in processing.

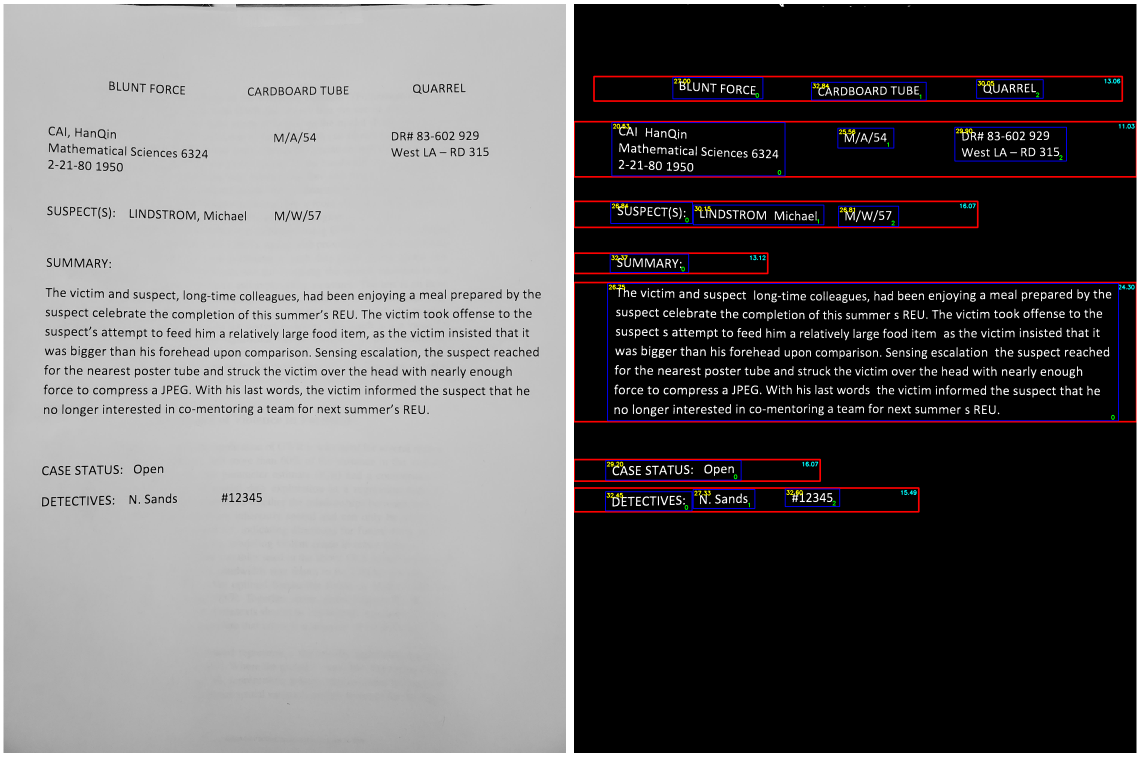
Once the erosion kernel was applied, two dilation kernels were applied. Here, dilation kernel refers to convolving our image with a maximum kernel where maximum kernel refers to an idea similar to a minimum kernel except we take the maximum instead of the minimum. The first convolution localizes text and the second convolution organizes the groups of text into a hierarchy by rows. The goal of the second convolution was to place text horizontally next to each other in the same hierarchy so we can simulate reading text from top-to-bottom and left-to-right.

We then found contours and their corresponding bounding boxes in the two dilated images using a simple chain approximation algorithm. We then needed to sort the bounding boxes, since some images could have bounding boxes not in the same horizontal row be placed in the same hierarchy due to the curvature of the page. The bounding boxes in each row get sorted by a modified greedy scheduling algorithm.

The following shows our image processing pipeline being applied on a sample image. A real image was not used to due the confidentiality of the data. For a longer sequence of images showing the image processing steps in more detail, we refer to the appendix.

2.2.2 Text Extraction

Once we were able to simulate reading top-to-bottom and left-to-right, we could employ an OCR engine to read the text on the page.



BLUNT FORCE CARDBOARD TUBE QUARREL

CAI, HanQin
Mathematical Sciences 6324
2-21-80 1950

M/A/54

DR# 83-602 929
West LA – RD 315

SUSPECT(S): LINDSTROM, Michael M/W/57

SUMMARY:

The victim and suspect, long-time colleagues, had been enjoying a meal prepared by the suspect celebrate the completion of this summer's REU. The victim took offense to the suspect's attempt to feed him a relatively large food item, as the victim insisted that it was bigger than his forehead upon comparison. Sensing escalation, the suspect reached for the nearest poster tube and struck the victim over the head with nearly enough force to compress a JPEG. With his last words, the victim informed the suspect that he no longer interested in co-mentoring a team for next summer's REU.

CASE STATUS: Open

DETECTIVES: N. Sands #12345

Figure 2.1: Before and after image processing.

Tesseract OCR Engine

Tesseract (29) is an OCR engine capable of extracting text from documents. The default Tesseract 4 uses an LSTM-based engine to recognize characters. Since our data did not include certain characters such as 'æ', we set a whitelist of characters for the engine to recognize. Since whitelists are not supported yet in Tesseract 4, we fall back on a legacy version of Tesseract.

Spell Checking

Once the text has been extracted using Tesseract, we performed some spell checking in order to get cleaner data and make field extraction easier. To do this, we use the SymSpell algorithm (15).

Field Extraction

Depending on the field we are extracting, we looked for different things. The four easiest fields to extract were suspects, summary, case status, and detectives since those are in general explicitly stated on the page. Some of the other information could be extracted using heuristics such as being on the top of the page or coming before some long strings of text.

Currently, the fields we have extracted are

1. Cause of death
2. Weapon used

3. Reason for homicide
4. List of suspects
5. Summary of homicide
6. Case status
7. List of detectives

The suspects, summary, case status, and detectives fields were extracted with high accuracy since most of the pages explicitly declared where each of those fields began in the document. The cause of death, weapon used, and reason for homicide were extracted with lower accuracy due to inconsistencies found in the images.

2.3 LA Open Data

The City of Los Angeles provides a portal to many datasets that are relevant to the investigation of crime patterns (8). This includes a database of crime reports that were recorded by the LAPD from 2010 to the present (2). This dataset was used to generate features matrices which were run through recurrent neural networks to generate predictive models of assaults.

The crime reports database contains over 2 million entries, recording everything from incidents of petty theft to homicide. It includes fields giving date, time, street address, latitude and longitude, crime type,*modus operandi*, and victim demographics. It can be downloaded in CSV format for processing and analysis.

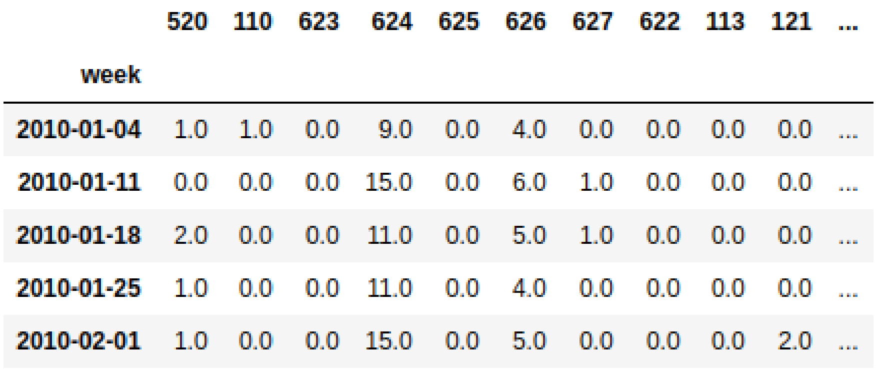
For each crime, the police area of record is listed (e.g. Rampart, Hollenbeck). We required an additional column listing the neighborhood. Shape files of LA neighborhoods are available via an API maintained by the L.A. Times (10). We ran a Python script using the shape files that returned the neighborhood for each crime. Since certain neighborhoods intersect more than one police area, we created a separate column that consisted of a concatenation of the police area and neighborhood (e.g. "Hollenbeck_El Sereno"). This allowed us to partition every police area into a set of pairwise-disjoint regions. It should be noted that street addresses are rounded to the nearest one-hundred block to maintain privacy, which is also reflected in the latitude and longitude data. This adds an element of uncertainty to the location of crime occurrences.

For each police area, and for each neighborhood, we created features matrices, each row representing a week of data, and containing counts of crimes of each type that occurred during that week. In addition we created target vectors consisting of the total number of assaults with a deadly weapon that were recorded in an area or a neighborhood during a given week. The whole process of generating the matrices was automated using Python scripting.

A portion of a features matrix is shown in Figure 2.2. The numbers across the top are LAPD numerical codes for different crime types.

2.4 Census Tract Data

The US Census Bureau (17) provides a variety of statistical data in different topics and geographies. The census tract data can be downloaded as a CSV file through the American Fact Finder advanced search (1). We were particularly interested in data that are relative to homicides from a demographic standpoint, including the poverty rate, housing value, household income, gross rent, employment status, and education status in each census tract. We processed the data into a pandas data frame with index being the `geoid` of census tracts in the LA County. We then computed the median household income, median gross rent,



week	520	110	623	624	625	626	627	622	113	121	...
2010-01-04	1.0	1.0	0.0	9.0	0.0	4.0	0.0	0.0	0.0	0.0	...
2010-01-11	0.0	0.0	0.0	15.0	0.0	6.0	1.0	0.0	0.0	0.0	...
2010-01-18	2.0	0.0	0.0	11.0	0.0	5.0	1.0	0.0	0.0	0.0	...
2010-01-25	1.0	0.0	0.0	11.0	0.0	4.0	0.0	0.0	0.0	0.0	...
2010-02-01	1.0	0.0	0.0	15.0	0.0	5.0	0.0	0.0	0.0	2.0	...

Figure 2.2: Features matrix with crime type counts. Each column corresponds to a crime code. For instance, code 520 refers to extortion of money by way of force, and during the week of January 4, 2010, there was only a single such case.

median housing value, unemployment rate and education level for each census tract.

We matched the geolocations of the homicides in the LA Times data set with the census tracts which they corresponded using a geocoder from the US Census Bureau (18). Later, we incorporated the processed census data with the LA Times data set to implement a factor analysis.

2.5 Zillow Data

Zillow collects monthly data on home and rental value in each section of LA County. The data can be sectioned off by zip code and downloaded to CSV, so for the purposes of this study, minimal preprocessing was required. We merged this data with shape files of each zip code region and treated distance between their respective centroids as a measure of distance between regions. While we did not utilize this data for any specific studies within this report, it would be particularly interesting to incorporate the time series of home and rental indices into a further implementation of geographically weighted regression (GWR), in which we might seek to predict local homicides in a nearby time step.

Section 3

Data Visualization

Before we discuss the methods used in our data analysis, we first discuss and visualize some trends that can be seen from our data.

3.1 Data Trends

Preliminary analysis of the LA Times shows some interesting trends. For example, an overwhelming majority of the homicide victims are male and most homicides were caused by gunshot.

Gender	Count
Male	12461
Female	1936
None	14
	3

Cause	Count
Gunshot	11432
Stabbing	1331
Blunt force	868
Other	537
Strangled	132
Unknown	67
Pending	34
Undetermined	13

Race	Count
Latino	7548
Black	4646
White	1593
Asian	509
None	91
Other	23
	4

Table 3.1: Counts broken down by various categories. The blank items under “Gender” and “Race” refer to empty strings in the data.

We can better visualize this data by plotting it. Note that in Figure 3.1 the distribution of ages for males has a much more noticeable mode at the age of 20. Similarly in Figure 3.2, we see that the distribution for Latinos also has a noticeable peak around age 25.

3.2 Rate Parameter Estimation

We assessed various patterns in time with respect to the homicide weapons used. Binning homicides by week for each weapon, we generated histograms for each two-year period and fitted these for the Poisson distribution using a maximum likelihood estimation (MLE).

Applying MLE (see appendix) gave us the value of the Poisson parameter λ which makes the observed data “most probable”.

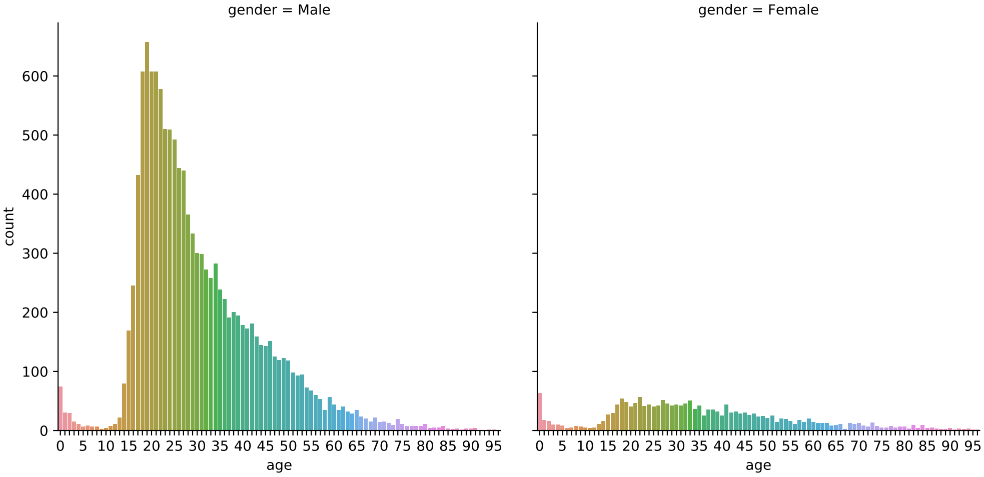


Figure 3.1: Distribution of ages broken down by gender.

We then plotted each of these rate parameters to observe their changes over time, as can be seen in Figure 3.3. For each type of homicide, we can observe at least a minimum decrease in λ since 2000. Notably, gunshot homicides have steadily declined, while still dominating the landscape of homicides in general. While fits for other weapons were less convincing, the plethora of data available for gunshot homicides suggests this to be a fair interpretation of the rate parameter's evolution over time.

3.3 Spatial Correlation

We were interested in the correlations between certain types of homicides over a range of distance. For example, we are interested in the correlation between those victims killed by gunshot and those who are of a certain race as well as how the correlation may change as the distance increases. Intuitively, as the distance between homicides increases, the correlation between them decreases.

To verify this thought, we first created uniform $N \times N$ mesh grids on the map of the LA area and denoted the set of the grids as M. Then we counted the number of homicides belonging to two types, say type x and type y, in each grid. For each grid, we searched for a grid p that was within the distance $(r, r + \Delta r)$ and denoted them as a pair (m, p). Here the metric $d(-, -)$ was used for the distance between grids—the physical distance between the centroids of the grids—and Δr is a parameter for the annulus. Then the set $S(r)$

$$S(r) = \{(m, p) : m \in M, d(m, p) \in (r, r + \Delta r)\} \tag{3.1}$$

denotes the set of the every possible pair that are within the distance $(r, r + \Delta r)$ of each other.

The formula for correlation coefficient in terms of r is

$$K_{xy}(r) = \frac{\sum_{(m,p)\in S(r)}(x_m - \bar{x}(r))(y_p - \bar{y}(r))}{\sigma_x(r)\sigma_y(r)} \tag{3.2}$$

where $\bar{x}(r)$ and $\bar{y}(r)$ are the mean of the counts of homicides of type x and y respectively. Similarly $\sigma_x^2(r)$ and $\sigma_y^2(r)$ are the variances. x_m is the number of the homicides that belong to type x in grid point m and

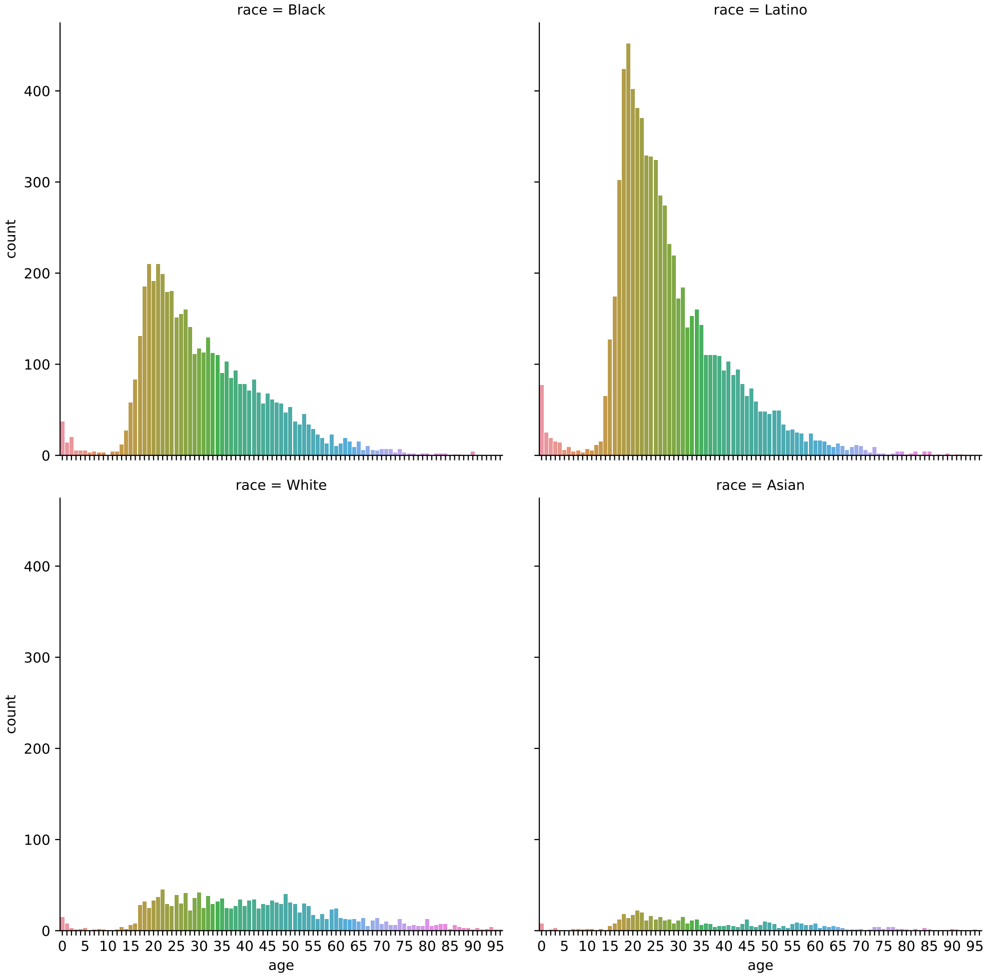


Figure 3.2: Distribution of ages broken down by race/ethnicity.

y_p is the number of homicides that belong to type y in grid point p.

Figure 3.4 shows the spatial correlation between gunshot homicides and victims of different races. By comparing different races, Latino victims maintain a relatively higher correlation coefficient with gunshot homicides and the correlation between all three depicted races. Besides, the correlation drops drastically when the distance is greater than 10km.

Figure 3.5 reveals that the correlation between female and Latino victims drops by 50 percent at a distance of approximately 10 km; between female and Black victims at 8 km; and between female and white victims at 18 km. We speculate that the last result is due to whites being the least segregated of the races in the city.

3.4 Factor Analysis

Factor analysis (28) (34) is a traditional method of finding the most important features of the data by using dimensionality reduction. Here we implement a variation of principle component analysis on our data.

Principle component analysis (PCA) uses the orthogonal transformation such that most of the variance in

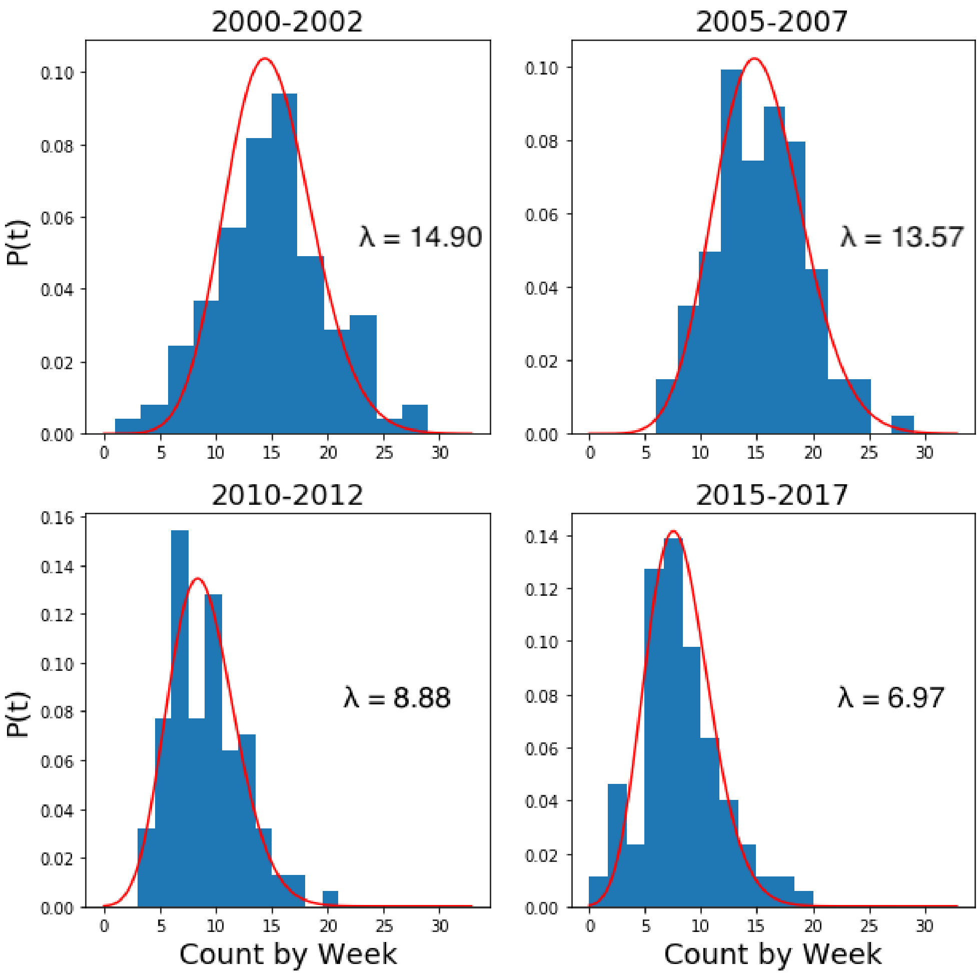


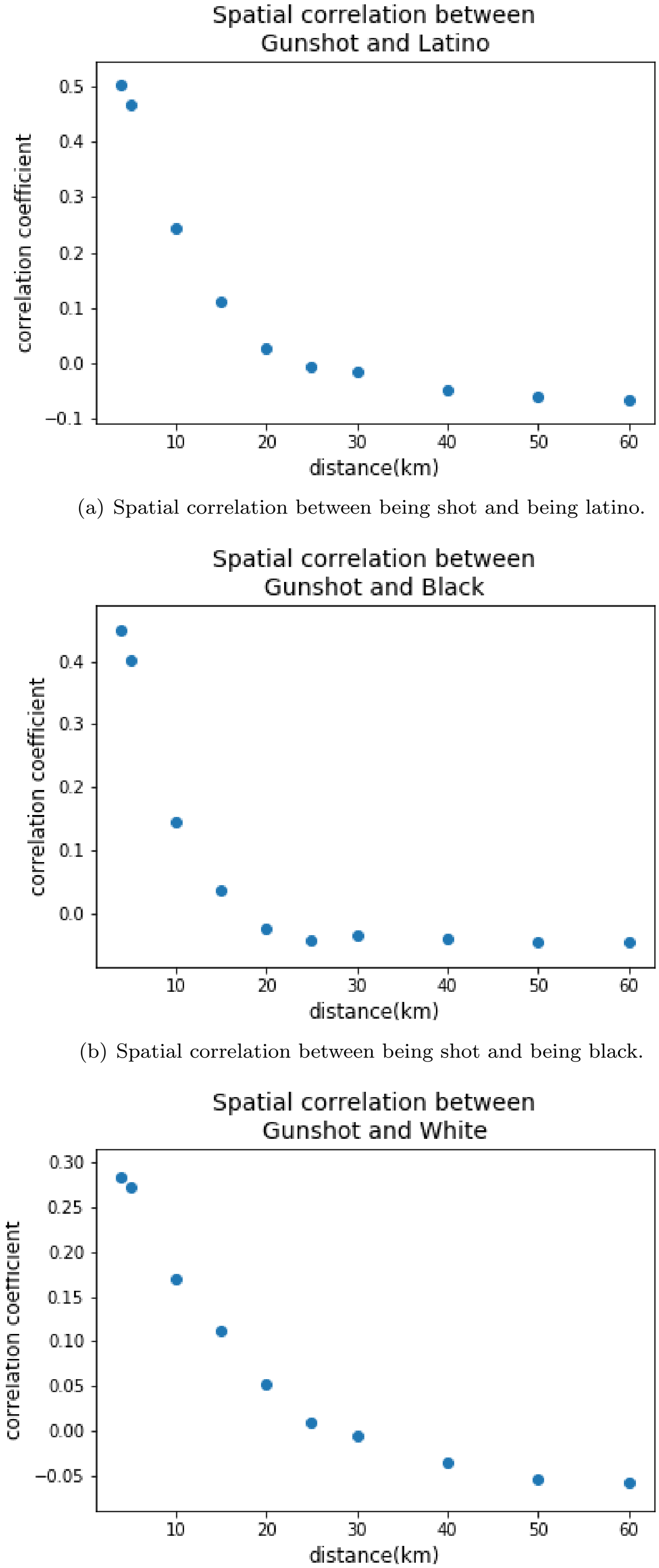
Figure 3.3: Poisson distribution fits for gunshot homicides.

the data is captured by the first component of the new coordinate system, the second largest variance is captured by the second component, and so on.

Since our data had both categorical variables such as gender, race, and cause, as well as numerical variables such as poverty rate and median housing value, we sought a way to compare them simultaneously.

For categorical data, the one hot encoder is used to generate a binary representation of the data. Concatenating matrices of two types of data, we applied PCA on the global matrix(14). With the variance of the first three components being 64%, 13% and 0.06%, it is sufficient for us to project the data into the subspace of the first two principle components. It is not surprising that the factors that have higher correlation coefficients with the first two components are factors associated with the neighborhoods where the homicides occurred. In fact, for the first component, the correlation coefficients of median housing value and median household income are −0.99 and −0.59. For the second component, the highest correlation coefficients are median housing value 0.89, median gross rent 0.8 and median household income 0.75.

Figure 3.7 (a) is the 2D projection of our data labeled by gender. There is a general trend along component 0: the male cases tend to have a higher value. We interpreted this as suggesting that the males are more likely to be murdered in less affluent areas whereas females are likely to be murdered in more affluent areas. Figure 3.7 (b) is labeled by death cause. Similarly, we can interpret this as suggesting that gunshot is more likely to happen in less affluent areas.

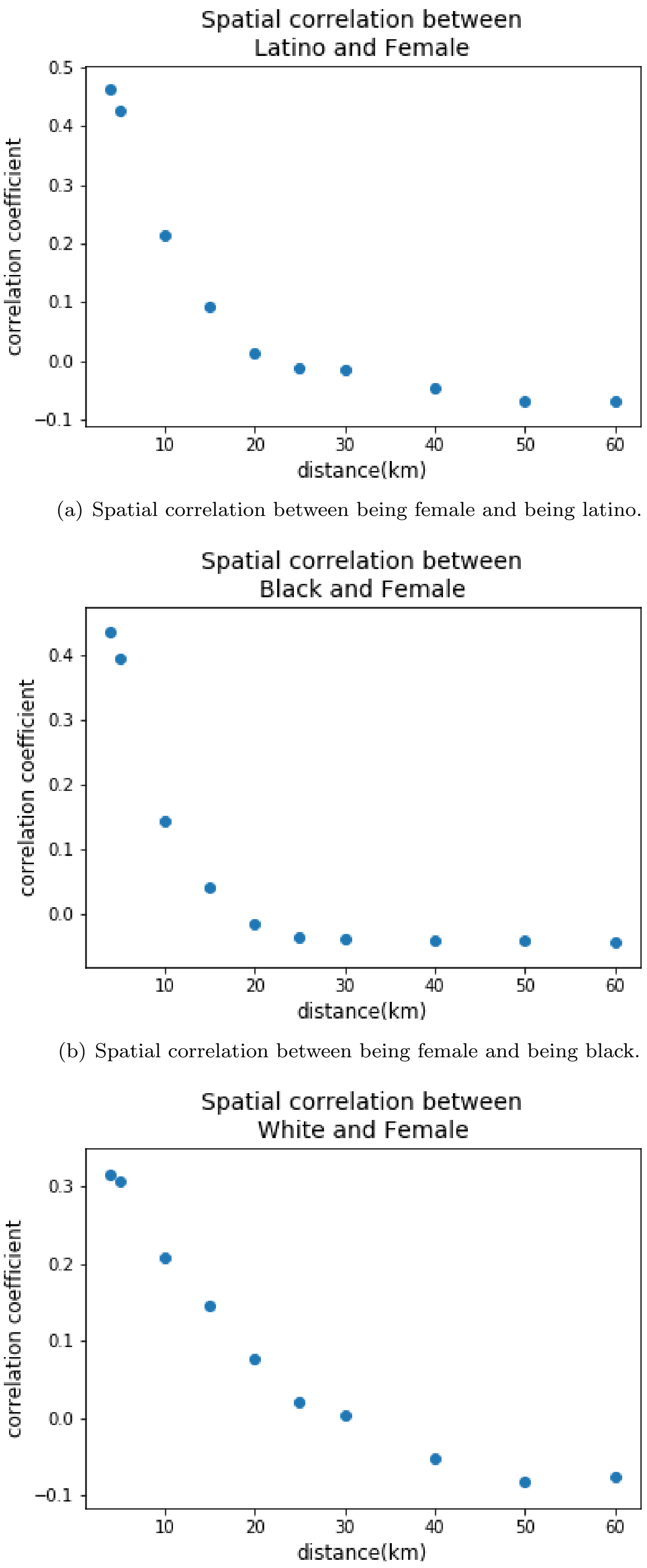


(a) Spatial correlation between being shot and being latino.

(b) Spatial correlation between being shot and being black.

(c) Spatial correlation between being shot and being white.

Figure 3.4: Spatial correlation with annulus $\Delta r = 5$km.

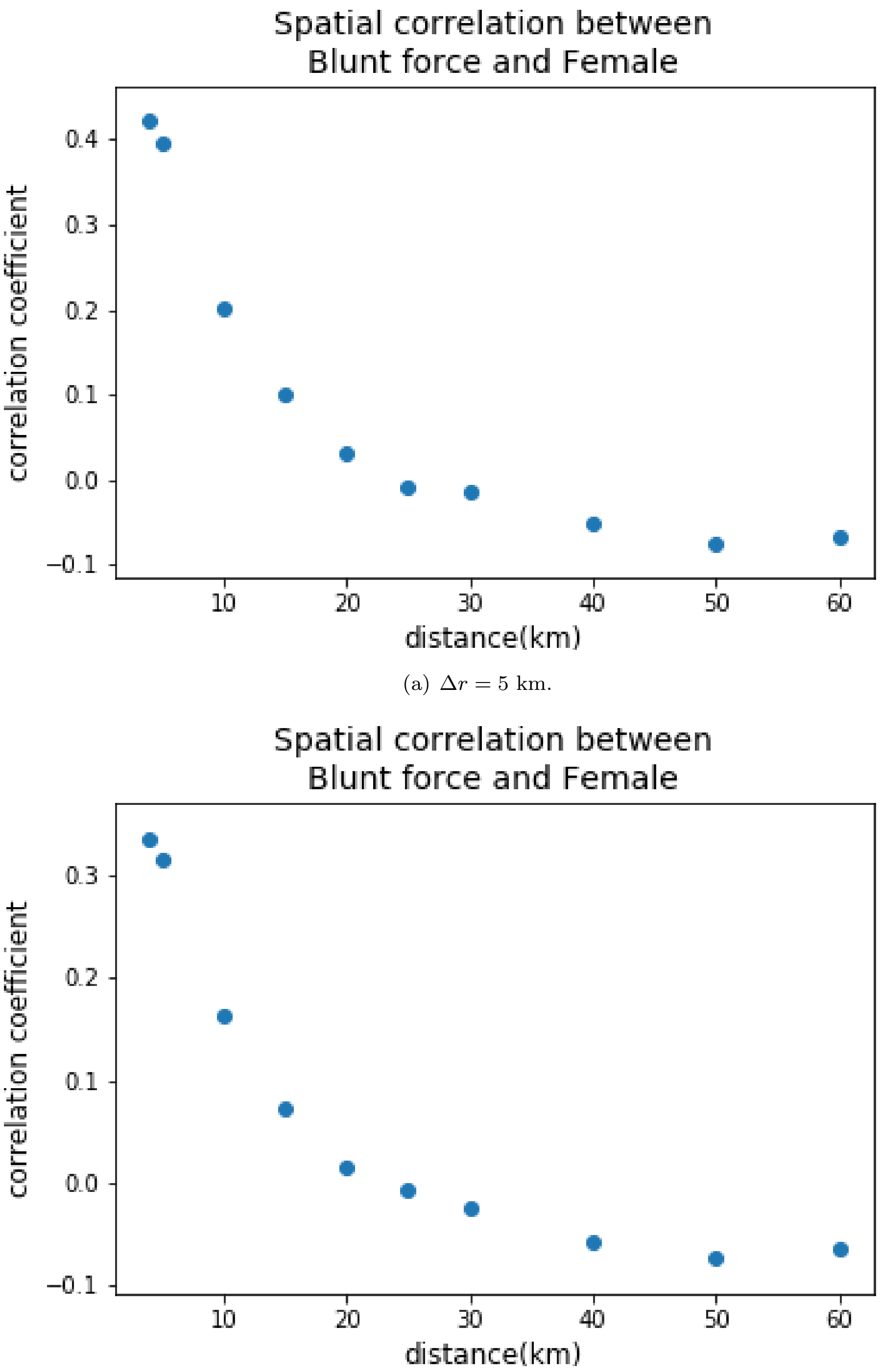


(a) Spatial correlation between being female and being latino.

(b) Spatial correlation between being female and being black.

(c) Spatial correlation between being female and being white.

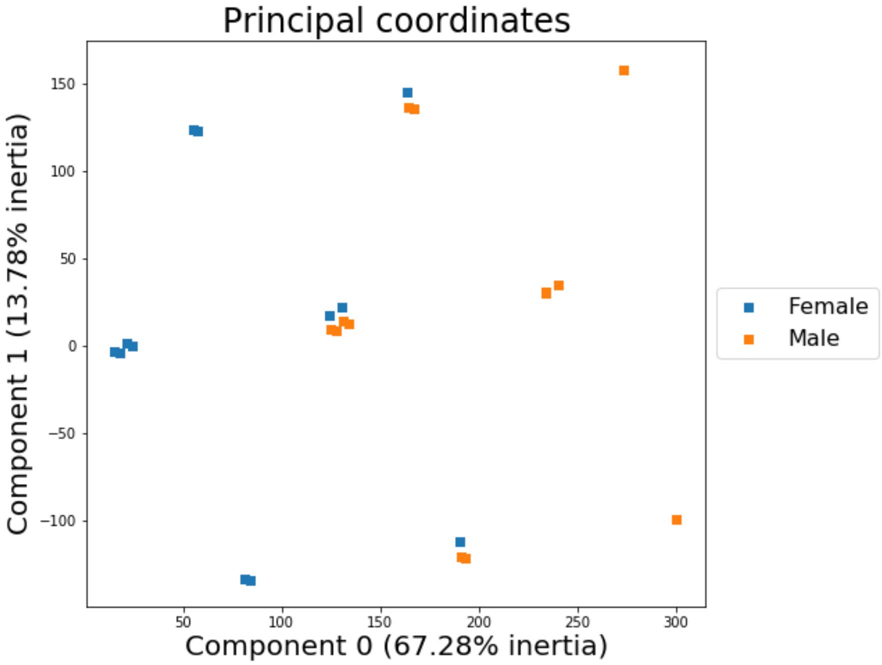
Figure 3.5: Spatial correlation with annulus $\Delta r = 5$km.



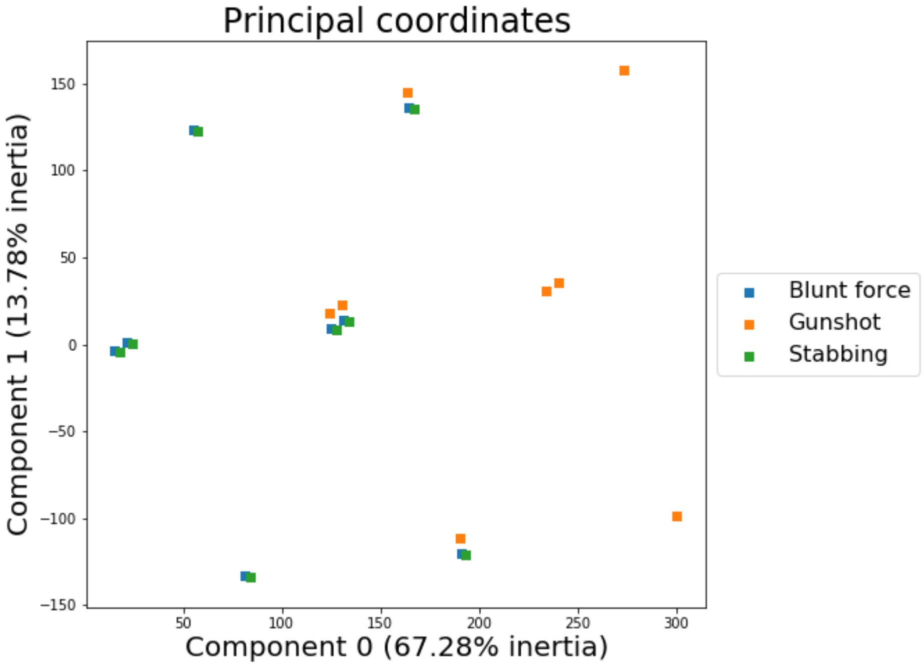
(a) $\Delta r = 5$ km.

(b) $\Delta r = 8$ km.

Figure 3.6: Spatial correlation between being female and killed by blunt force with different annulus.



(a) Factor Analysis: colored in gender.



(b) Factor Analysis: colored in cause.

Figure 3.7: Results from factor analysis.

Section 4

Methods

We present the methods used to analyze our datasets. We used techniques from machine learning, topic modeling, and dynamic modeling.

4.1 Topic Modeling

Topic modeling is useful for discovering hidden topics or structure within text in our corpus. The main methods we used are non-negative matrix factorization, semantic non-negative matrix factorization, latent Dirichlet allocation, and sparse contextual hidden and observed language autoencoder.

4.1.1 Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is a matrix factorization method that decomposes a given matrix $X \in \mathbb{R}^{n \times d}$ into two low-rank, non-negative matrices $W \in \mathbb{R}^{n \times r}$ and $H \in \mathbb{R}^{r \times d}$ for some r to be specified. To find the W and H, we solve for the minimization problem:

$$\min_{W,H} \|X - WH\|_F^2 \tag{4.1}$$

NMF can be very useful in topic modeling. To format the narratives into inputs for our topic modeling methods, we used a bag-of-words model, which takes in the corpus and creates a vocabulary out of each unique word in the corpus. It then models each document as a vector with length equal to the number of words in the vocabulary where entry i in the vector corresponds to how many times word i occurred in the document. Thus, the bag-of-words model gives us a matrix X with dimension $n \times d$ where n is the number of words in the vocabulary, and d is the number of documents. The parameter r represents the number of the topic which need to determined. For matrix W,the words per topic matrix, each column of the W represents a word distribution of a topic. For matrix H, the topic per document matrix, ith column represents the topic distribution of ith document.

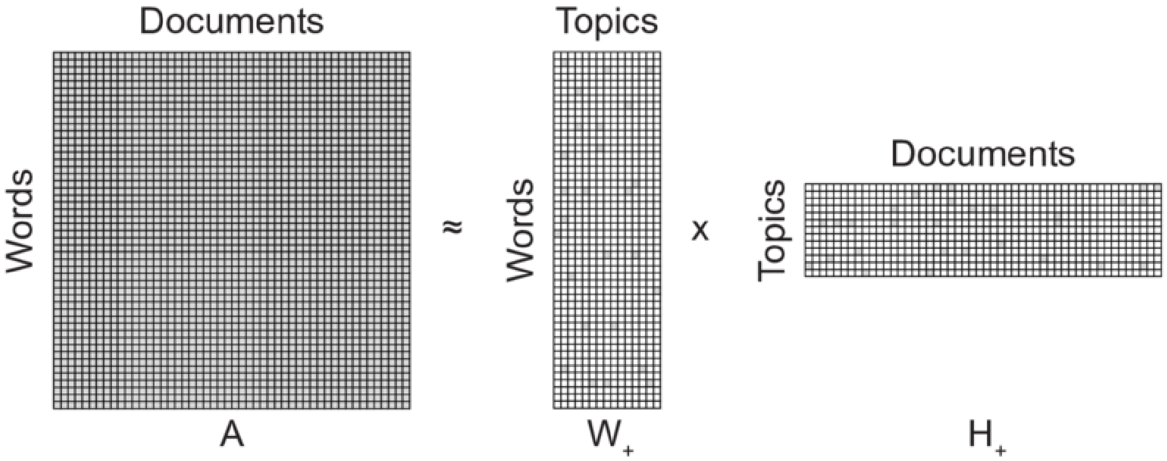


Figure 4.1: An illustration of NMF.

4.1.2 Word2Vec

One issue with NMF is that it does not account for the meanings of each word. For example, the model views the difference between cat and dog the same as the difference between cat and gun. Thus, we use Word2Vec (35) to project our data into a lower dimensional space, or semantic space, that accounts for the word interpretations.

Mathematically we treat every word in the vocabulary as a vector in a vector space and introduce a inner product of two words that measures how close they are. In this way, we can reduce the number of dimensions by learning a linear map from our vocabulary to the semantic space by projecting words with similar meanings to similar locations.

4.1.3 Semantic Non-negative Matrix Factorization

Incorporating Word2Vec with NMF gives us Semantic Non-negative Matrix Factorization (SNMF) (25), which is similar to NMF but accounts for the meanings of the words.

SNMF differs from NMF by calculating the reconstruction error in the semantic space. The problem boils down to minimizing the distance between X and WH in the semantic space by multiplying V, the word embedding matrix obtained from Word2Vec, to the NMF minimizing object, our minimization problem becomes:

$$\min_{W,H} \|VX - VWH\|_F^2 \tag{4.2}$$

Then we can apply hierarchical alternating least square (HALS) algorithm to solve the minimization problem.

4.1.4 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (19) is a commonly used topic modeling technique. It assumes a structure on the corpus, i.e. every document is composed of a mixture of topics, and each topic has a word distribution over the corpus's vocabulary. The generative process for LDA with k topics is:

1. For $i \in [1, \ldots, k]$, choose word distributions $\phi_i \sim \text{Dir}(\beta)$
2. For each document $d \in \mathcal{D}$:
   (a) Choose a topic proportion for document d, $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each word position j in document d:
      i. Choose a topic $z_j \sim \text{Multi}(\theta_d)$
      ii. Choose a word $w_j \sim \text{Multi}(\phi_{z_j})$

Figure 4.2 visualizes the document part of LDA's generative process.

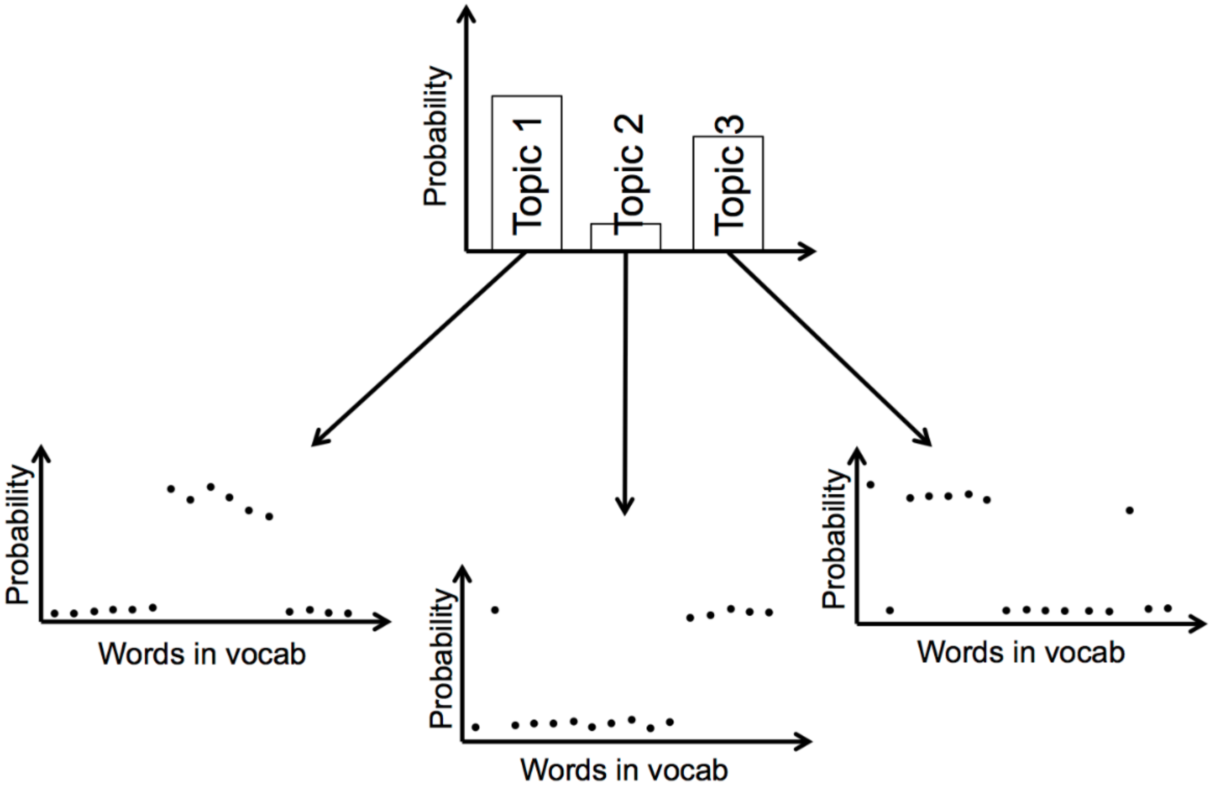


Figure 4.2: An illustration of part of LDA's generative process. Each document is represented by a topic proportion, and each topic has a word distribution (11)

The goal of inference is to find α and β, which is done by computing the MLE of the posterior distribution of the hidden variables θ and $\mathbf{z}$. In other words, compute $p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta)$. This is intractable so we turn to variational inference (or potentially other approximate inference algorithms).

We find a variational distribution with parameters γ and ϕ such that the Kullback-Leibler (KL) divergence between the variational distribution and posterior distribution is minimized. In other words, find

$$(\gamma^*, \phi^*) = \underset{(\gamma,\phi)}{\operatorname{argmin}}\ D(q(\theta, \mathbf{z}|\gamma, \phi) \mid\mid p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)) \tag{4.3}$$

This can be done using the iterative algorithm described by the paper.

We then want to find optimal values for α and β, the parameter of our Dirichlet prior and the word distribution per topic, that maximize the marginal log likelihood of the data:

$$l(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d|\alpha, \beta) \tag{4.4}$$

Since the likelihood is intractable, this is done using variational expectation-maximization where in the E-step we find γ^* and ϕ^* as described above and in the M-step we find α and β such that the likelihood is maximized. The β update can be done analytically while the α update can be implemented using an efficient linear-time Newton-Raphson method.

Upon convergence, we have α and β as desired. To figure out our latent topics, we find the expectation of the posterior $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$.

4.1.5 SCHOLAR

While LDA is very popular topic modeling technique, Sparse Contextual Hidden and Observed Language Autoencoder (SCHOLAR) (21) has many advantages. Specifically, SCHOLAR gives the option of using a sparsity-inducing prior to produce more interpretable topics, a variational auto-encoder to add in word-embeddings, and categorical metadata.

Metadata can be used as either covariates, labels, or both. Giving the model categorical metadata as covariates forces the model to treat documents differently, regarding topic proportions, based on the value of that variable. Feeding the model categorical metadata as labels guides the model to produce topics relevant to the labels' values.

SCHOLAR differs from LDA by using a generative network, f_g, to alter the word distributions per topic in each document based on that document's covariate value. Also, it uses a multilayer perceptron (MLP), or a simple feed forward neural network with at least 3 layers, to predict the label of a document given its topic proportion and covariate value. Ultimately, a goal of this method is to determine the parameters for both the generative network and MLP that best fit the corpus with metadata. These parameters are found via stochastic gradient descent.

By finding those variables, we have the topic representations of each document, and all the word distributions per topic. Furthermore, we have the MLP that predicts a label given the topic representation of a document. We can find the topic representations of unseen documents given the word distributions per topic and then use it to predict the unseen document's label.

To evaluate multiple SCHOLAR models, we used perplexity, shown in Equation 4.5,

$$\text{perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{4.5}$$

where $\mathbf{w_d}$ is the word count vector for document d, $p(\mathbf{w_d})$ is the probability of seeing that word count vector, and $\mathbf{N_d}$ is the number of words in document d (19).

Perplexity is a commonly-used measure to assess the coherence of topics generated by the model. This measure is useful as the value decreases with an increase of the log likelihood of observing the words in a test corpus. So, a small perplexity indicates a good topic model.

4.2 Recurrent Neural Networks

A recurrent neural network (RNN) (39) is a special type of neural network that has temporal connections in the hidden layers. This enables an RNN to exhibit temporal behavior and process sequences of inputs better than a vanilla neural network. This is achieved by having a layer that keeps an internal hidden state.

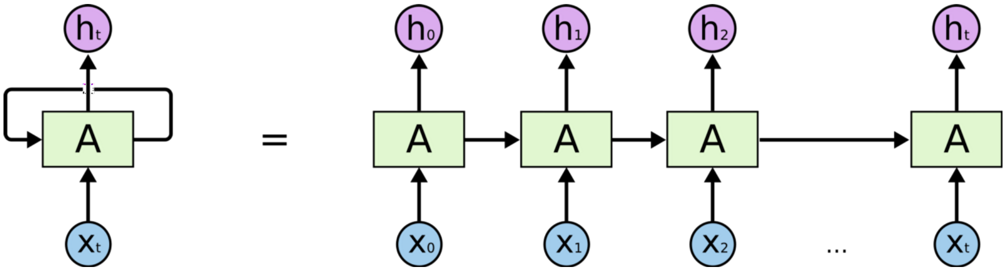


Figure 4.3: An illustration of an RNN (3).

Here, the x_i are the sequence of inputs, A is a cell, and h_i is the hidden layer at time i. For every x_i that is passed into the RNN, the cell processes it and updates the hidden state h_i.

Because of the RNN's ability to process sequential data, we wanted to use it to predict homicide data. Specifically, given three consecutive weeks of data containing the number of homicides that occurred in each region of LA, we wanted to predict the number of homicides that were going to occur in each region of LA in the next week. To make our task easier, we normalized the counts into a crime density map.

To evaluate our model and train our neural network, we used the KL divergence shown in Equation 4.6, a common measure used in statistics to determine how different one probability distribution from another distribution.

$$D_{KL}(P||Q) = -\sum_{x \in \mathcal{X}} P(x) \log \left(\frac{Q(x)}{P(x)} \right) \tag{4.6}$$

4.2.1 Long Short-Term Memory

Vanilla RNNs often running into the vanishing gradient and exploding gradient problems. This is an issue where during gradient descent, the updates can be effectively zero or unbounded since RNNs build up long chains of derivatives during backpropagation. To alleviate the situation, we used a long short-term memory (LSTM) network (39). An LSTM essentially keeps some memory cells that are more resistant to change. The main components of an LSTM are the cell, input gate, output gate, and forget gate. LSTMs have proved to be superior to vanilla RNNs due to their ability to avoid the vanishing and exploding gradient problems.

The main idea behind LSTMs can be found in the horizontal black line near the top of the cell. This serves as the memory of the LSTM and is the cell state. Since it has minimal interactions with the inputs, it can retain its state for a long period of time.

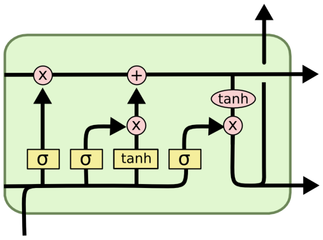


Figure 4.4: An illustration of an LSTM cell (16).

4.2.2 Gated Recurrent Unit

One downside of LSTMs is that they are more complex than vanilla RNNs. Since our dataset is relatively small, we potentially run into issues of overfitting. To remedy this situation, we use a newer and simpler modification of RNNs, the gated recurrent unit (GRU) (23). This architecture has fewer parameters to fit but still achieves similar results compared to LSTMs.

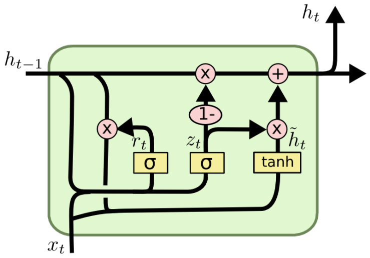


Figure 4.5: An illustration of a GRU cell (13).

A GRU achieves a simpler architecture by reducing the number of gates. It uses an update and reset gate to control its memory cells.

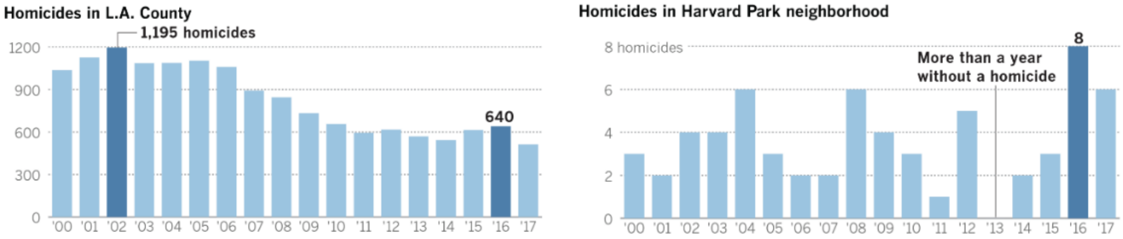
4.3 Dynamic Models

Since 2010 Los Angeles county as a whole has experienced a decline in the number of homicides that have occurred each year. While this is certainly a desirable trend, it is not always one that is observed when a more microscopic view of the data is taken. Figure 4.6 shows how one South L.A. neighborhood, Harvard Park, has had a much different experience of crime over the past two decades. There, the trend has not been one of steady decline, but of a cyclic return of violence (12).

In order to begin to understand how periodic fluctuations in crime levels like the ones seen in Harvard Park might arise, we developed two dynamic models that explore the relationship between crime, violence, and police presence across multiple quasi-geographic sites.

The underlying assumption behind these models is that criminals will shift the focus of their activities to areas with lower police presence, and that police will adjust their numbers depending on the local level of crime.

Classes for both models were implemented in Python to perform simulations.



(a) Homicides in LA County since 2010 (LA Times).

(b) Homicides in Harvard Park since 2010 (LA Times).

Figure 4.6: Homicides in two regions since 2010.

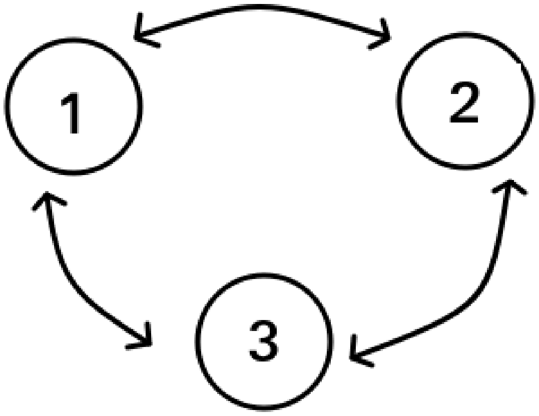


Figure 4.7: A three-site model.

4.3.1 Uniform Criminal Population

Discrete

We pick an arbitrary number of sites n. Although numbered in sequence, there is no strict geographic relationship between them, (i.e., site 2 is not necessarily 'between' sites 1 and 3), and no fixed scale. They are merely abstract spaces which serve as collection points for "bad actors" who can pass freely between them as seen in Figure 4.7.

Each site i is assigned a baseline police presence p_i, and a source term S_i, $i = 1, \ldots, n$, representing an underlying level of criminality that is added to the site at every step. At time step $t+1$, we calculate the number of bad actors B_i and police P_i at site i using the formulae:

$$B_i(t+1) = B_i(t) - \overbrace{\sum_{j \neq i} \Pi_{ij} B_i(t)}^{\text{CRIME OUT}} + \overbrace{\sum_{j \neq i} \Pi_{ji} B_j(t)}^{\text{CRIME IN}} - \overbrace{\beta P_i(t) B_i(t)}^{\text{ARRESTS}} + \overbrace{S_i}^{\text{Source}} \tag{4.7}$$

$$P_i(t+1) = p_i + \theta B_i(t) \tag{4.8}$$

where $\beta > 0$ and $\theta > 0$ are constants. Π_{ij} is a transition function taken to be the probability of a group of criminals in site i moving to site j. Its value depends on the difference in police populations between sites i and j.

$$\Pi_{ij} = \frac{\ell(P_i - P_j)}{\sum_j \ell(P_i - P_j)} \tag{4.9}$$

$$\ell(x) = \log(1 + e^x) \tag{4.10}$$

Continuous

For purposes of analysis, we restated the above equations as a system of first order non-linear differential equations. For a two-site model, we have

$$\dot{B_1} = \zeta(\Pi_{11} - 1)B_1 + \zeta\Pi_{21}B_2 - \beta P_1 B_1 + S_1 \tag{4.11}$$

$$\dot{B_2} = \zeta(\Pi_{22} - 1)B_2 + \zeta\Pi_{12}B_1 - \beta P_2 B_2 + S_2 \tag{4.12}$$

$$\dot{P_1} = \theta\dot{B_1} \tag{4.13}$$

$$\dot{P_2} = \theta\dot{B_2} \tag{4.14}$$

with ζ being a constant.

4.3.2 Multiple Gang Populations

A slightly more complicated model can be introduced to calculate levels of crime and inter-gang violence at different sites. Here we have z rival gangs $G^1, G^2, \ldots, G^z$, each with its own level of criminal activity determined by a constant α^k, $k = 1, \ldots, z$. The contribution of the kth gang to the crime at each site i is then ${C_i}^k = \alpha^k {G_i}^k$. Our equations for the gang and police populations become

$${G_i}^k(t+1) = {G_i}^k(t) - \sum_{j \neq i} \Pi^k_{ij} {G_i}^k(t) + \sum_{j \neq i} \Pi^k_{ji} {G_j}^k(t) - \beta P_i(t) {G_i}^k(t) + {S_i}^k \tag{4.15}$$

$$P_i(t+1) = p_i + \gamma C_i(t), \qquad C_i(t) = \sum\nolimits_k {C_i}^k(t) \tag{4.16}$$

Transitions of gang members between sites depend not only on the police population, but also the relative populations of rival gang members. We assume that the probability of gang members in site i moving to site j is higher if there are more members of the same gang present at site j, along with fewer police and fewer rivals. We modify our transition function accordingly, and add a weighting constant $\lambda, 0 \leq \lambda \leq 1$, that determines which portion of avoidance is based on the presence of police, and which is based on the the presence of rivals. Hence the probability that gang k at site i will move to site j is

$${\Pi_{ij}}^k = \frac{\ell(\lambda[P_i - P_j] + (1-\lambda)[({G_j}^k - {G_i}^k) + \sum_{l \neq k}({G_i}^l - {G_j}^l)])}{\sum_j \ell(\lambda[P_i - P_j] + (1-\lambda)[({G_j}^k - {G_i}^k) + \sum_{l \neq k}({G_i}^l - {G_j}^l)])} \tag{4.17}$$

where $\ell(x)$ is defined as in Equation 4.10 .

Violence between rival gangs at site i may be calculated as the sum of the products of rival gang populations times a constant δ.

$$V_i = \delta \sum_{\ell} \sum_{k < \ell} {G_i}^k {G_i}^\ell \tag{4.18}$$

4.4 Geographically Weighted Regression

Geographically weighted regression (GWR) is a technique born out of ordinary least-squares (OLS) regression, in which regression coefficients are estimated locally based on nearby features. At its most simple, GWR can be thought of as an enhanced local regression, as in Figure 4.8. In this diagram, we have a kernel centered at the origin weighting the value of nearby points higher in a regression local to the y-axis. It is unlikely that a linear model would fit well to the curve as a whole; however, when we localize our regression, we can break the fit down into smaller linear components and develop a well-fitted linear model with local regression coefficients.

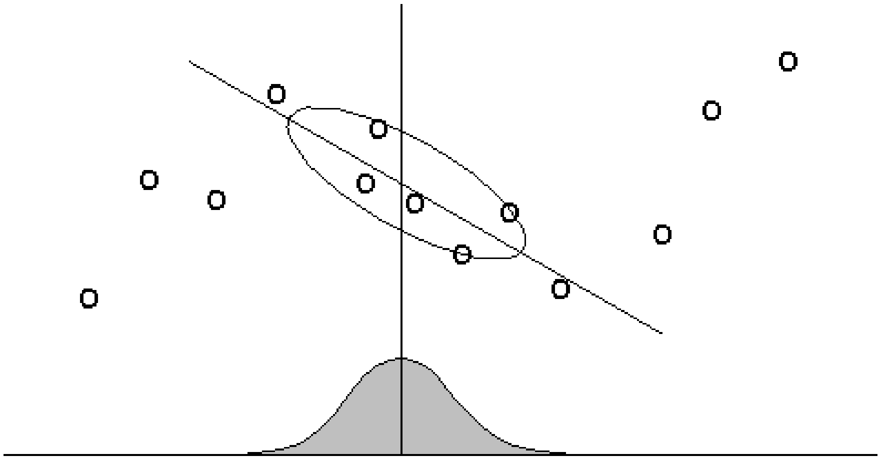


Figure 4.8: Local regression in 2-dimensions. Here we have a linear model being fitted to a subset of data which otherwise is nonlinear. Nearby points are weighted more heavily in order to calibrate the model. (7)

The goal of GWR is to capture spatial heterogeneity by calibrating an ensemble of linear models at any number of locations through the "borrowing" of nearby data. For our purposes, it made sense to apply GWR in order to analyze spatially varying relationships in homicides. With the dynamic model described in Section 4.3, we noticed that a jump in crime in one area might cause instability in surrounding areas. This suggests that these local relationships are worth exploring. The result of GWR is a surface of location-specific regression coefficients for each relationship in the model. Additionally, we require a single bandwidth parameter to specify geographic scale. We discuss selection of this parameter in the following section.

For the model itself, we populate a features matrix θ and calibrate the model using the following formula:

$$f(\beta, x, y) = \sum_{i \text{ locations}} w(x, y, x_i, y_i)(\sum_j \theta_{ij}\beta_j - h_i)^2 \tag{4.19}$$

where β is a vector of estimated regression coefficients, w is a weighting function, θ is a matrix of independent variables (features), h_i is the dependent variable that we are interested in predicting, and $\hat{\beta} = \hat{\beta}(x, y)$. We find $\hat{\beta}$ by minimizing over β, solving

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} f(\beta, x, y) \tag{4.20}$$

4.4.1 Bandwidth Selection

Within our GWR model, we have the freedom to choose a bandwidth parameter that controls the size of our weighting kernel. In general, there are two types of kernels: fixed and adaptive. With a fixed kernel, we specify a bandwidth distance over which to perform the regression. Alternatively, an adaptive kernel uses a bandwidth which represents the number of nearest neighbors to include in the regression and the kernel adapts accordingly.

Consider a Gaussian weighting function

$$w_{ij} = \exp\left\{-\frac{1}{2}(\frac{d_{ij}}{b})^2\right\} \tag{4.21}$$

where d_{ij} is the distance between two locations i and j and b represents some bandwidth. One could imagine choosing b by minimizing the quantity

$$z = \sum_{i=1}^{n}[y_i - \hat{y}_i(b)]^2 \tag{4.22}$$

However, suppose this is the case and b is made very small. In this case, the weighting of all points except i itself risk becoming negligible, and the fitted values will tend to actual values so that the value of z becomes zero (26). Instead, we can refer to a cross-validation (CV) approach to local regression (24)

$$CV = \sum_{i=1}^{n}[y_i - \hat{y}_{\neq i}(b)]^2 \tag{4.23}$$

in which $\hat{y}_{\neq i}(b)$ is the fitted value of y_i with observations for point i left out of the calibration process. With this process, we avoid the risk of disrupting the model when b becomes very small.

4.4.2 Data Preparation

In preparing the data for running the GWR model, we required that features be separated by region. Initially, we set out to build a data structure in which the user could specify the size of a region and generate arbitrarily-sized features matrices for these types of models. However, while this may be feasible in the future, it proved difficult given the time constraint of the project. Instead, we looked to data that was already sectioned off by region (zip code) and paired with its features.

In order to run the model, we needed a dependent variable (homicide count), which also had to be included in the features matrix along with its respective zip code in each case. Processing data for a GWR model and be time-consuming, especially if the user is interested in incorporating a time series for each location and over each feature.

Section 5

Results

We describe the results of our methods in detail. We organize our results by the dataset being analyzed.

5.1 LA Times Dataset

Since each homicide report contained the longitude, latitude, and narratives, we used geographically weighted regression and topic modeling to analyze the LA Times dataset.

5.1.1 Geographically Weighted Regression

We performed a geographically weighted regression using regions bounded by zip code and very general demographic data associated with those. The demographic data includes total population, total males, total females, median age, total households, and average household size. We populated a features matrix θ and calibrated the model using Equation 4.19 with h_i being the number of homicides that actually occurred at location i.

In Figure 5.1 we have plotted the local R^2 statistic, which provides a measure of goodness of fit for the regression. An $R^2 \approx 1$ suggests that the model is very well fit to the data. We observe a fairly high value throughout LA, suggesting that the model fits fairly well to the data. While this may be promising initially, it is worth noting certain issues that might arise within this method which should be improved in future work. For instance, in GWR it is not uncommon to deal with issues of multicollinearity, in which we are attempting to fit a linear model to parameters which may already correlate fairly well linearly (27). This might occur between total males/females and total population, for example, since the male-to-female ratio does not fluctuate much between zip codes. In the future, we would hope to justify with more reasonable intuition which features to include in the analysis.

A second issue arises regarding the predictability of the model itself. Even though GWR allows us to analyze *spatially* varying relationships, it neglects to incorporate a temporal aspect, which, for the purposes of prediction, would allow for more meaningful insights. In addition to running the algorithm with more pointed features, we hope to incorporate this aspect in future work to improve the predictability of the model.

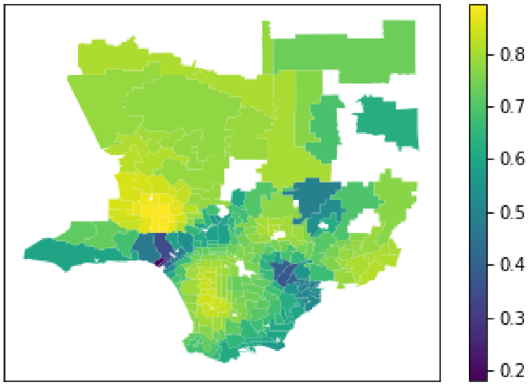


Figure 5.1: Local R^2 statistic plotted with an adaptive bandwidth of 57.0.

5.1.2 Topic Modeling

As the summaries of the LA county homicides may contain latent features that are useful for crime prediction, we analyzed the LA Times corpus with the following topic models, NMF, SNMF, Latent Dirichlet Allocation, and SCHOLAR. Before using those models, we first cleaned the homicide reports. Since the first sentence of almost every report only contained information given by the metadata, we removed that sentence as it didn't contribute to the topics found in the narrative portion of the reports. Furthermore, proper nouns, numbers, punctuation, stop words, and any variation of sentences asking for more information about the homicide were eliminated from the narritives with regular expressions. We then were left with roughly 6000 nonempty narratives. We used the `CountVectorizer` method from `sklearn` (20) to create our word frequency matrix. `Gensim`'s Word2Vec model (38) was used to create our word embedding with 50 dimensions for SNMF.

Even though SNMF with 7 topics gave the most interpretable results, NMF, SNMF, and LDA all yielded poor results, regarding topic coherence, on the case summaries. We also experimented with using term frequency inverse document frequency (tf-idf) instead of bag-of-words but got similar results. The topics shown in Table 5.1 were produced with SNMF.

Topic 1:	shooting	gang	shot	information	case	homicide
Topic 2:	officers	officer	shooting	police	deputies	release
Topic 3:	murder	year	charged	office	records	attorney
Topic 4:	**family**	**mother**	would	**home**	year	shooting
Topic 5:	car	shooting	vehicle	dead	pronounced	shot
Topic 6:	dead	pronounced	found	gunshot	records	scene
Topic 7:	police	year	shot	found	home	authorities

Table 5.1: Above are the results of running SNMF with 7 topics and the top 6 words associated with each topic. Topic 4 seems to be the best as those words suggest a domestic topic.

While topic 4 shows some signs of coherence, the other topics include randomly chosen words that are commonly found in homicide reports. Thus, we used SCHOLAR for topic modeling. One useful feature of SCHOLAR is option of incorporating metadata into the model as we believe that the number of homicides should affect the topic proportions of the report. For example, we would expect a homicide report from a death in Beverly Hills to be quite different from a report from a death in Harvard Park; there is significantly more gang activity near Harvard Park compared to gang activity around Beverly Hills. As SCHOLAR takes in categorical variables, we created a new variable for each homicide, the relative homicide amount, by binning each region into three bins, low, medium, or high based on the total number of homicides in that region. Running SCHOLAR with the relative homicide amount as a covariate produces the following topics found in Table 5.2. Similarly to the previous results, the number of topics was chosen based on the interpretability of the associated words of each topic.

Topic 1:	gunshots	shot	area	found	ambulance
	scene	dead	chief	lying	heard
Topic 2:	sedan	colored	light	dark	north
	vehicle	drove	west	standing	person
Topic 3:	responded	release	news	call	officers
	gunshot	suffering	shots	wounds	arrived
Topic 4:	trauma	serial	wife	body	lieu
	slayings	deaths	matches	children	death
Topic 5:	always	drop	wanted	officer	never
	even	officers	holding	cameras	kids
Topic 6:	handling	details	soon	update	contacted
	arrests	obtain	spokesman	contact	additional
Topic 7:	count	attorney	charged	due	guilty
	pleaded	district	allegation	murder	arraignment
Topic 8:	youths	offered	young	youth	graduate
	spot	cleared	crowd	brawl	dozens

Table 5.2: The topics and associated words above are significantly better than the results from NMF, SNMF, or LDA. Topics 2, 4, 7, and 8 are good as they roughly represent driving, serial killings, legal issues, and homicides of young people, respectively.

5.1.3 Homicide Forecasting

Since the topic representations of the homicides contain latent features excluded in each entry's metadata, we used them to predict future homicide counts. We first averaged the topic representation of each homicide by week over our entire dataset to obtain a time series. Using the L.A. Times Mapping L.A. dataset of L.A. County region boundaries, we computed the number of homicides in each region and then divided by the total number of homicides per week to obtain a time series of homicide densities (4). In order to predict next week's homicide density from the topic representations, we used a RNN as it captures the sequential dependency in our data. Specifically, we used a GRU by passing the topic representations from weeks $t-4, t-3, t-2, t-1$, and t as inputs to predict the homicide density of week $t+1$. We used a GRU instead of a LSTM as it requires fewer parameters to fit because of our lack of data.

We implemented the GRU with Keras (22). The GRU's architecture contains GRU layer, a flatten layer, and a dense layer with a softmax activation function as our output is a probability density. We used the ADAM optimizer (30) and the KL Divergence as the loss function. We randomly selected 10% of the data as the test set and 10% of the remaining data set as the validation set. We then trained the model with a batch size of 16 for 25 epochs as the average KL Divergence on the validation set started increasing, a sign of overfitting. The KL Divergence on the test set was 10.99. However, this result means little without other average KL Divergence scores to compare it to. Thus, to add context to our results, we compared it with two other benchmark distributions, a uniform distribution and an average homicide density. The average homicide density was computed by totaling the number of homicides in each LA County region and then dividing by the total number of homicides. Table 5.3 shows the average KL Divergence scores of our model and benchmark densities on the test set.

	GRU	Average Homicide Density	Uniform Density
KL Div.	10.99	10.88	15.60

Table 5.3: Above are the average KL Divergence scores of the model and benchmark densities on the test set.

Since KL Divergence is a measure of the difference between two probability distributions, a larger score indicates that the distribution deviates more from the true distribution compared to another distribution with a lower score. Thus, our model performs marginally worse than the average homicide density but better than the uniform density. However, this result is not surprising as a heatmap of the homicide density by week shown in Figure 5.2 reveals that there is no discernible pattern.

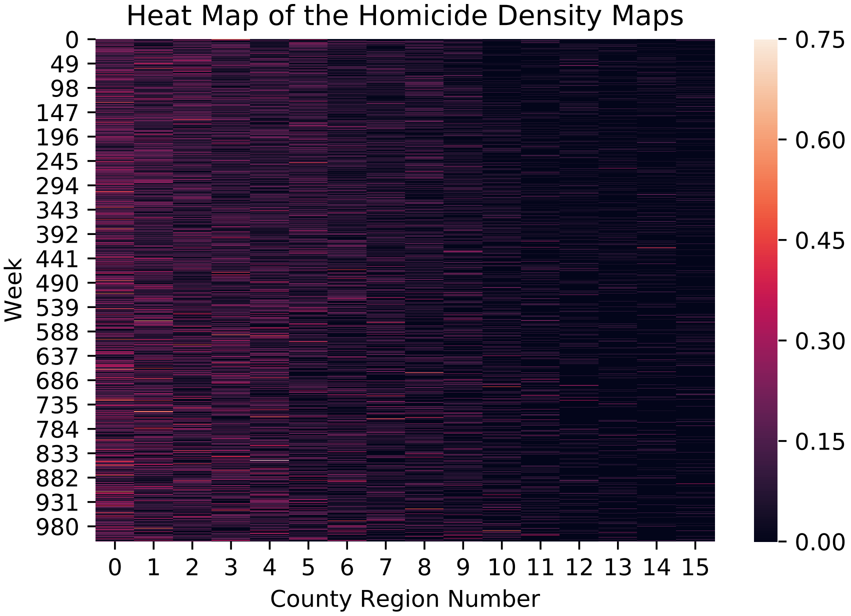


Figure 5.2: The heatmap of the homicide densities by week.

5.2 Topic Modeling on the LAPD Records

Similarly to the LA Times dataset, we used topic modeling to extract latent features in the LAPD book records. We preprocessed the data by removing numbers, punctuation, and stop words in snowball. We also used a bag-of-words model with a vocabulary of only the 2000 most frequent words to represent the records in order to apply SCHOLAR. In contrast to the LA Times dataset, we have access to the case status, open or closed, of each homicide. Since predicting this given the report is of much interest to the LAPD, we ran scholar with the case status as the label. To illustrate how labels are generated with words from the topic distribution, we ran SCHOLAR with an 80-20 split of training and test data with two topics. Table 5.4 contains the top 10 words associated with the topics while Table 5.5 shows the probabilities of a document that is only composed of topic 1 or 2 being labeled open or closed. The words in each of these topics seem to align with their label, open or closed. Topic 1's words, discovered, bound, found, information, and investigation, suggest that this topic represent prolonged investigation. On the other hand, topic 2's words, verbal, altercation, knife, and bar, demonstrate that this topic represents a brawl, which would likely alert witnesses and police to the fight.

Topic 1:	discovered	bound	found	multiple	information
	lying	motive	investigation	remain	believed
Topic 2:	verbal	altercation	knife	started	involved
	rejected	pulled	bar	self	went

Table 5.4: Above are the associated words of the two topics.

	Closed Topic Probability	Open Topic Probability
Topic 1	0.2740	0.7260
Topic 2	0.9448	0.0552

Table 5.5: The probabilities shown above represent the likelihood of a document composed of only topic 1 or 2 being assigned open or closed.

While these words seem to form logical topics, the trade-off is that the decrease in test accuracy of the binary classifier. To determine the optimal number of topics to use, we compared the accuracy and perplexity averaged over 5 different seeds of SCHOLAR models with 5, 10, 15, 25, 50, and 100 for 2000 epochs each. 20 percent of the data was randomly chosen as the test set while 20 percent of the remaining data was randomly selected for the validation set. Table 5.6 shows the results of the 6 models evaluated on the validation dataset.

Number of Topics	5	10	15	25	50	100
Perplexity	414 ± 3.28	412 ± 2.93	419 ± 2.59	438 ± 4.62	500 ± 6.39	618 ± 23.0
Accuracy	0.86 ± 0.049	0.89 ± 0.026	0.86 ± 0.017	0.88 ± 0.025	0.85 ± 0.030	0.86 ± 0.025

Table 5.6: Above are the mean and standard deviation of the accuracy and perplexity of the model with the respective number of topics.

As 10 topics yield the best results in both accuracy and perplexity, we ran the model with 10 topics for 2000 epochs because the validation accuracy converged at that time step. The training and test accuracy are given in Table 5.7. The model performed well on the test set given that the proportion of closed cases is 63%. Furthermore, some of the summaries were not correctly captured by the OCR, which may have decreased the accuracy of the model.

5.3 Dynamic Models

Because of the number of different variables and constants present in our models, they allow configurations which give widely different outcomes when run in simulation. Because we have not yet had the opportunity to fit the models to any empirical data, or even decide on any physically meaningful scale for the quantities involved, it is difficult to know which sets of results, if any, shed light on the patterns of crime that result from the interplay of police and criminals as their respective populations change over time.

That being said, the models suggest at least one interesting avenue for further exploration, namely the question of how police forces are to be deployed when an area experiences a unusual increase in crime.

Train Accuracy	Test Accuracy
0.922	0.875

Table 5.7: Above are the train and test accuracies for SCHOLAR with 10 topics run for 3000 epochs.

A natural policing strategy in response to spikes in crime is to shift more resources directly to the affected area. The effect of such a strategy can be simulated using our first model.

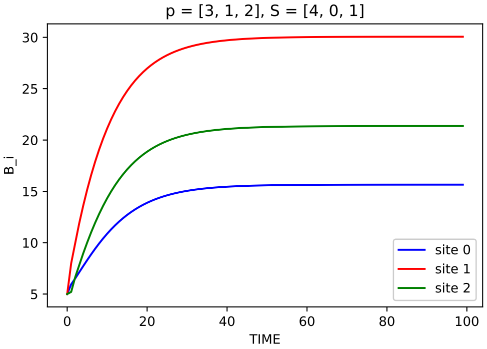


Figure 5.3: Equilibrium is reached.

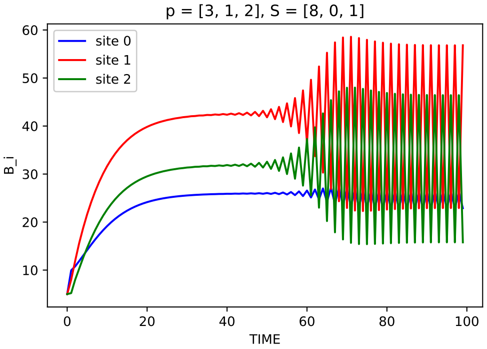


Figure 5.4: Crime spike at site 0 leads to instability.

Figure 5.3 shows a three site model at equilibrium. When crime in site 0 spikes, the neighboring low-crime areas are destabilized Figure 5.4. Increasing the police presence at site 0 merely reduces crime there, but the instability in neighboring sites goes unchecked Figure 5.5. An alternate strategy of adding the same number of additional police forces to the low-crime sites, on the other hand, brings the whole system back to stability Figure 5.6.

5.4 LA Open Data

For each LAPD patrol area, and for each Los Angeles neighborhood, we created matrices containing week-by-week counts for crimes of all types using data provided by the LA Open Data website (2).

We trained a GRU recurrent neural network on each matrix (Keras package/Theano Backend), using consecutive pairs of weeks as inputs to take advantage of the RNN's sensitivity to sequenced data. 200 epochs were used during training. The target value for each input was the number of assaults with a deadly weapon that

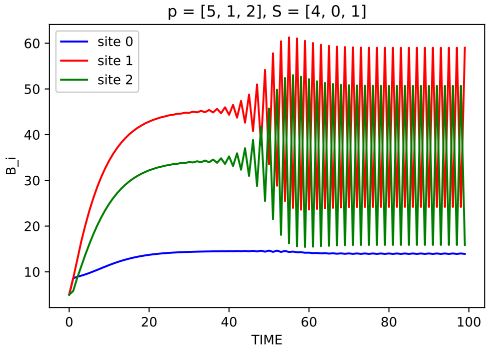


Figure 5.5: Increasing police at site 0 fails to stabilize

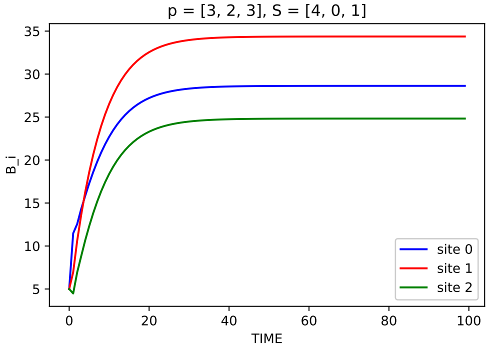


Figure 5.6: Reinforcing low-crime sites restores stability.

occurred in the week following the last week contained in each input. We set aside the final 100 week-pairs of each matrix to use as testing data.

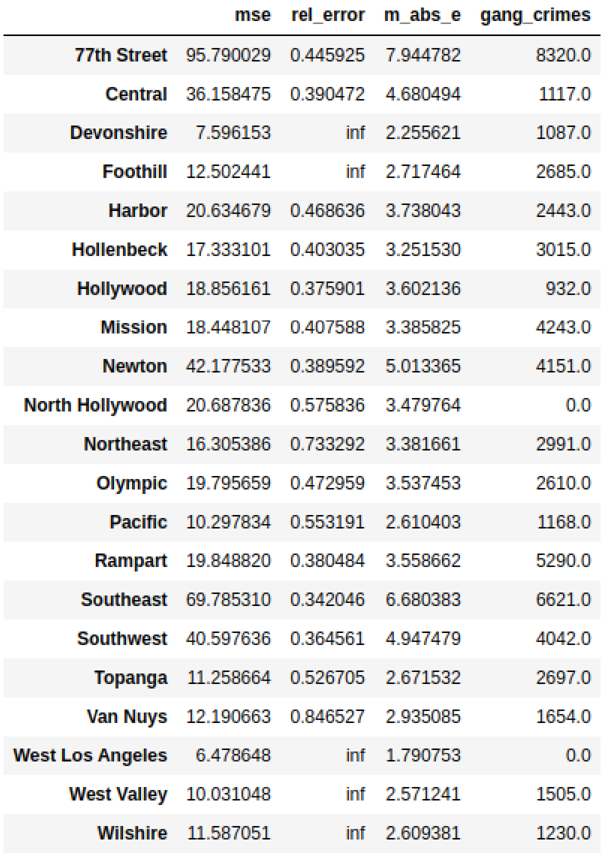
The accuracy of our RNN model in predicting the number of assaults to occur in a given LAPD patrol area in a given week is summarized in Figure 5.7. For each area, the mean squared error, mean absolute error, and relative error (where it could be calculated) are given.

The final column in Figure 5.7 relates the accuracy of the model to the total number of gang-related crimes reported in each area over the entire time span of the dataset. The figures in this column were tabulated from records containing an M.O. code of '0906' to indicate that they were gang-related.

We note that the relative error ($\frac{target\_assaults - predicted\_assaults}{target\_assaults}$) varies from approximately 37 percent (Hollywood) to 85 percent (Van Nuys). A graph of the number of gang crimes per area versus the relative error is shown in Figure 5.8. The dashed line represents a linear function fitted to the data using least squares ($R^2 = 0.14$), and indicates a possible inverse relationship between the volume of gang activity in a region and the ability of an RNN to predict violent crime. (But R^2 is small.)

Next we explored the relative accuracy of the RNN models trained on area-level data vs. neighborhood-level data. We expected that our models would become more accurate if training and prediction were conducted on smaller geographic regions. This was based on the assumption that if criminal events have any influence whatsoever on other criminal events in space and time, their influence diminishes with distance.

To compare the results from our models that were done on patrol areas to those done on neighborhoods, we calculated the sum of the number of assaults predicted for each week for all neighborhoods contained within a particular area, and subtracted the actual number of assaults in the area to get the error. We then



	mse	rel_error	m_abs_e	gang_crimes
77th Street	95.790029	0.445925	7.944782	8320.0
Central	36.158475	0.390472	4.680494	1117.0
Devonshire	7.596153	inf	2.255621	1087.0
Foothill	12.502441	inf	2.717464	2685.0
Harbor	20.634679	0.468636	3.738043	2443.0
Hollenbeck	17.333101	0.403035	3.251530	3015.0
Hollywood	18.856161	0.375901	3.602136	932.0
Mission	18.448107	0.407588	3.385825	4243.0
Newton	42.177533	0.389592	5.013365	4151.0
North Hollywood	20.687836	0.575836	3.479764	0.0
Northeast	16.305386	0.733292	3.381661	2991.0
Olympic	19.795659	0.472959	3.537453	2610.0
Pacific	10.297834	0.553191	2.610403	1168.0
Rampart	19.848820	0.380484	3.558662	5290.0
Southeast	69.785310	0.342046	6.680383	6621.0
Southwest	40.597636	0.364561	4.947479	4042.0
Topanga	11.258664	0.526705	2.671532	2697.0
Van Nuys	12.190663	0.846527	2.935085	1654.0
West Los Angeles	6.478648	inf	1.790753	0.0
West Valley	10.031048	inf	2.571241	1505.0
Wilshire	11.587051	inf	2.609381	1230.0

Figure 5.7: Accuracy of predicted number of weekly assaults with a deadly weapon for LAPD patrol areas using recurrent neural networks.

compared the mean absolute error generated from neighborhood figures to the mean absolute error generated by area-level predictions. Figure 5.9 shows the results.

The difference between the neighborhood- and area-level mean absolute errors ($MAE_{area} - MAE_{neighborhood}$) is plotted against the number of gang crimes per area in Figure 5.10. A least squares linear fit ($R^2 = 0.36$) is also shown. The linear function suggests that as the number of gang crimes increase, the difference in error goes from negative to positive, that is, the neighborhood-level model becomes more accurate than the area-level model.

The relation between model accuracy and gang crime will be discussed in the conclusion.

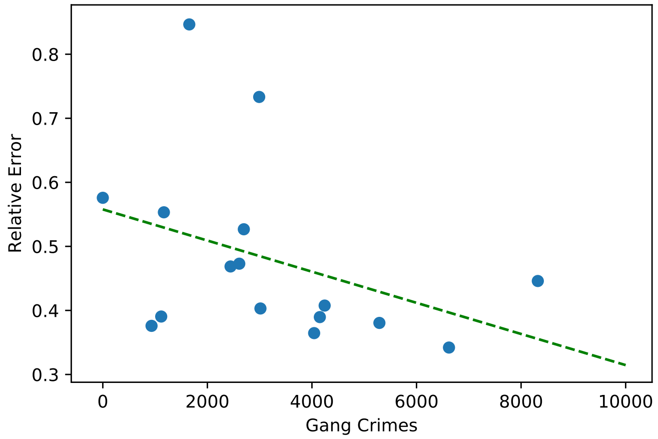
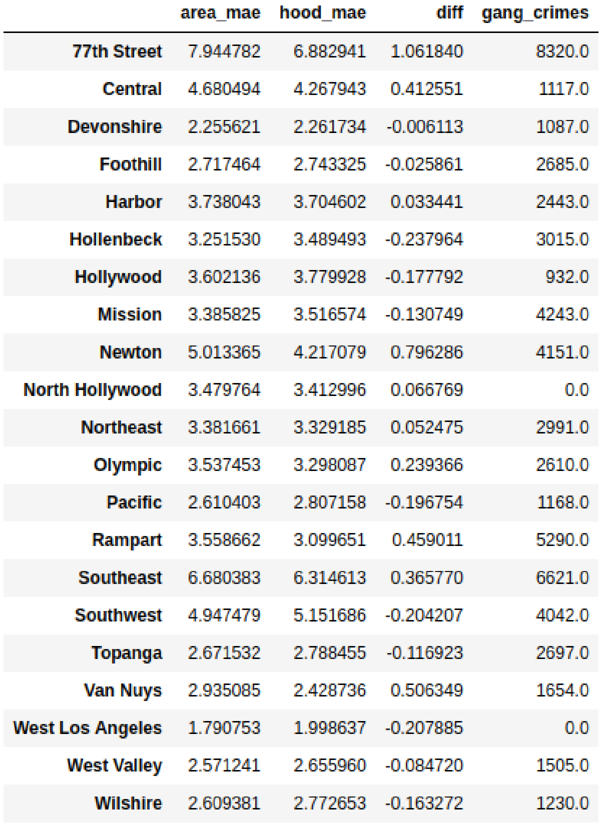


Figure 5.8: Relative error of predictive RNN model for each patrol area versus number of gang crimes in area, with least-squares fit ($R^2 = 0.14$).



	area_mae	hood_mae	diff	gang_crimes
77th Street	7.944782	6.882941	1.061840	8320.0
Central	4.680494	4.267943	0.412551	1117.0
Devonshire	2.255621	2.261734	-0.006113	1087.0
Foothill	2.717464	2.743325	-0.025861	2685.0
Harbor	3.738043	3.704602	0.033441	2443.0
Hollenbeck	3.251530	3.489493	-0.237964	3015.0
Hollywood	3.602136	3.779928	-0.177792	932.0
Mission	3.385825	3.516574	-0.130749	4243.0
Newton	5.013365	4.217079	0.796286	4151.0
North Hollywood	3.479764	3.412996	0.066769	0.0
Northeast	3.381661	3.329185	0.052475	2991.0
Olympic	3.537453	3.298087	0.239366	2610.0
Pacific	2.610403	2.807158	-0.196754	1168.0
Rampart	3.558662	3.099651	0.459011	5290.0
Southeast	6.680383	6.314613	0.365770	6621.0
Southwest	4.947479	5.151686	-0.204207	4042.0
Topanga	2.671532	2.788455	-0.116923	2697.0
Van Nuys	2.935085	2.428736	0.506349	1654.0
West Los Angeles	1.790753	1.998637	-0.207885	0.0
West Valley	2.571241	2.655960	-0.084720	1505.0
Wilshire	2.609381	2.772653	-0.163272	1230.0

Figure 5.9: Comparison of error between predictive RNN models trained on patrol area-level versus neighborhood-level data.

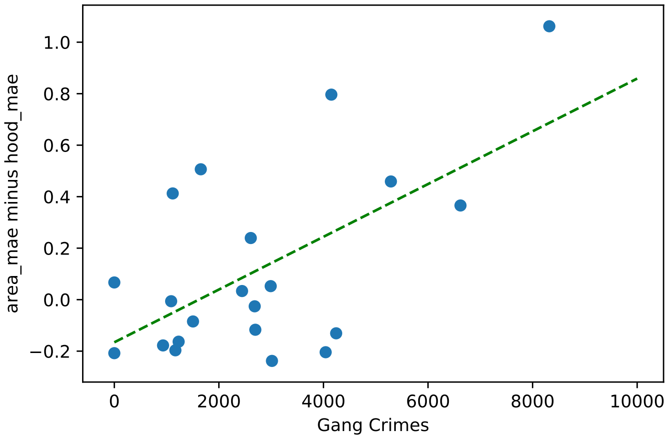


Figure 5.10: Difference in error between area-level and neighborhood-level RNN models plotted against quantity of gang crimes in area, with least-squares fit ($R^2 = 0.36$).

Section 6

Conclusion

We present our conclusions based on our results below.

6.1 LA Times

We first applied a spatial correlation analysis on the LA Times data set. Intuitively, as the distance increases, two variables should be less related. This is verified by the data. The correlation between two variables of homicides decreases to 0 as the distance increases. Also for different pair of variables, the distance at which the correlation coefficient decays to 0 changes drastically.

Incorporating the census tract data where each homicide belongs, we implemented a factor analysis to find the most significant features of homicides. Homicides is highly correlated with the surrounding areas. Among all the features, the median housing value, median household income, and median gross rent are the most significant. Qualitative conclusions are drawn from the 2D projections of our data. Generally, the males are likely to be murdered in less affluent areas compared to the females. Homicides caused by gunshot are likely to happen in less affluent areas than homicides caused by blunt force.

We also analyzed the LA Times descriptions with SCHOLAR using the relative crime level of the location of each homicide as the covariate. The generated topics seem to represent various aspects of homicides, such as driving, legal issues, and youth. Detectives can analyze these topics along with location of homicides to identify patterns in homicides in certain areas. Furthermore, we created a model, GRU, to predict the a homicide density over LA County. However, the performance of this model was poor as most of the homicides occurred in one or two regions.

- The correlation between variables in homicide tends to zero as the distance increases.
- Variables in homicide data are highly correlated with median housing, median household income, and median gross rent.
- Males are more likely to be murdered in less affluent areas than females.
- Gunshot homicides are less likely to occur in affluent areas than homicides by blunt force.
- There are patterns inherent in the narratives.

6.2 LAPD Records

Homicide report descriptions can contain latent features, in other words topics, that highlight the key aspects of the deaths. We used SCHOLAR to predict case status with the topic representations with high accuracy. Furthermore, the case status can be used to guide topic words, e.g. with two topics, one topic's words resemble a lack of information about the dead body and homicide while the other topic's words suggest a fight between the victim and suspect. Using the predictive model and topics, detectives can gain intuition about features of homicides that are open or closed after 48 hours by analyzing the topic words.

- Case status can be used to find trends hidden in the homicide records.
- Case status can be accurately predicted using topic modeling.

6.3 Dynamic Models

Citywide crime statistics can be reported in such a way as to obscure the dynamic processes that are occurring at the level of neighborhoods and at even smaller geographic subdivisions. We proposed a way of modeling the interactions between police and criminal populations that could be scaled appropriately to simulate the unstable crime patterns seen in neighborhoods like Harvard Park. Although the model needs to be validated against real-world data, and subjected to more rigorous analysis, our simulations raise interesting questions about how best to deploy police forces when areas are experiencing a spike in crime. We found in our model that under certain circumstances a local spike in crime can destabilize neighboring lower-crime areas. Simply increasing the number of police in the area experiencing the spike did not return the neighboring sites to equilibrium, whereas making an equivalent increase across the lower-crime sites did. We conclude that reinforcing neighboring sites with added police presence must be a consideration when addressing crime spikes.

Police forces are by nature finite, and we hope that future work with dynamic models can help make decisions about the optimal distribution of resources for bringing crime numbers down across all areas of the city.

- Crime spikes in one area can destabilize neighboring areas.
- Increasing police presence in the area of a spike may not cure instability in neighboring areas.
- Neighboring areas must be reinforced during a crime spike.

6.4 LA Open Data

We attempted to build a predictive model for forecasting assaults with a deadly weapon using recurrent neural nets and publicly available information. We trained RNN's on weekly crime report data from 2010 to the present, splitting the data into subsets based on LAPD patrol areas first, then on neighborhoods. We found on average that the models were mediocre to poor predictors. However we noted a possible relation between the accuracy of the models in a given area and the volume of gang crime in that area. Our models tended to have less average relative error in areas that recorded a higher number of gang-related crimes. Also, models trained on neighborhood-level data became more accurate than area-level models as the area gang crime count increased.

The results were by no means conclusive, and it is unclear if gang crime count is merely a proxy for another more meaningful statistic. However, one possible interpretation is that violence in areas with high gang activity has more antecedent causes in the proximate "ambient" crime than violence in other areas.

RNN's are particularly suited to sequenced data, which made them a first choice for our time-sequenced data. Even if their ability to forecast crime in the present case is limited, they may still be able to tell us something about the variable nature of violent crime across the city.

- Our RNN models were not reliable predictors of assault violence.
- There was some indication that accuracy increased in areas of high gang activity.
- Violent crime in these areas may be more of a function of "ambient" crime.
- RNN's could be used to detect qualitative differences in crimes of the same type from region to region.

Section 7

Future Work

Although we have promising results, there are still some things we can do to improve our analyses.

7.1 Web Scraping

Currently, the web scraper is still unable to parse all the information correctly. For example, sometimes in the case descriptions, some of the HTML is extracted as is, meaning that during topic modeling, things like `</i>` might be in the data. Another issue with the webpages is that sometimes, the format is not consistent so some information does not get parsed correctly. Being able to have more flexible parsing would enable better extraction and data.

One thing we did not get to do was to perform topic modeling on the comments section. Some pages do not have comments while some have over 100 comments. This information could provide a lot of insight into the homicides.

7.2 Optical Character Recognition

The methods we used to clean up artifacts cause as a result of imperfect scans can be greatly improved. Currently, heuristics are used to determine whether a connected component is text or not. Training a neural network on the connected components to classify them is something that can greatly improve both automation and accuracy.

More fields could also be extracted. Because of the quality of the scans and the layout of the pages, it is not always possible to extract everything from each page. Further processing can be done to try to extract more fields and better scans can be made so our OCR can segment the image better.

The Tesseract OCR engine settings could also be better set to have better extraction. Currently we use the default page segmentation mode which performs full automatic page segmentation. There is a page segmentation mode that assumes a single block of text. This seems to be more suitable since we are passing bounding boxes into Tesseract and not the whole image.

7.3 Spell Checking

Currently in spell checking, we use a default dictionary as a reference. Since we know approximately what kind of words will appear in our text, we can use a custom dictionary to better spell check our OCR results.

7.4 Improving the Binary Classifier

While the binary classifier performed well on the test set, we did not experiment with adding various categorical data to our SCHOLAR model as a covariate. The cause of death, age, race, or gender may significantly influence the homicide records. Thus, using any of those categorical variables as a covariate may lead to more coherent topics as well as higher accuracy on the case status. Furthermore, as more LAPD books become available, the increase in data should boost accuracy and topic interpretability.

7.5 Dynamic Models

The behaviour of our discrete models could be better understood with more fixed-point analysis of continuous versions. Some preliminary simplification of the models could help in this regard. Validation of the model with empirical data would be the next obvious step.

7.6 LA Open Data

We used an arbitrary time interval of a week to construct our features matrices. It would be interesting to see how much our results would change if a shorter time period, say, three day intervals, was used. Exploration of the use of RNN's to detect qualitative differences in the nature of crime of the same type from region to region is another possibly fruitful area of research.

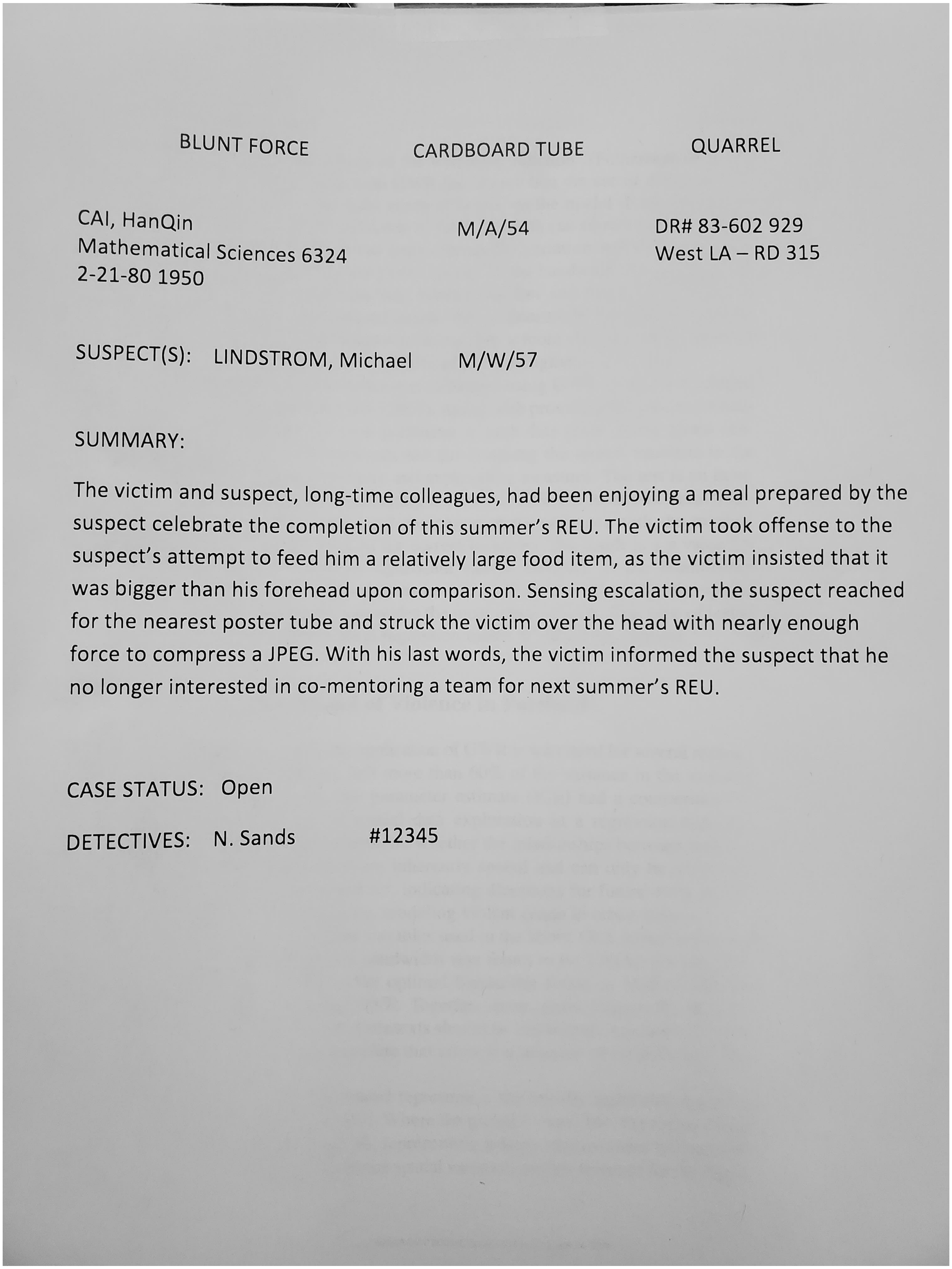
7.7 Outlier Detection

Apart from the factor analysis that we implemented, we are also working on a variation of robust PCA (RPCA) (40) that could be applied to the mixture of categorical data and numerical data. The advantage of doing RPCA is that we can find the structure of the data and at the same time detect the outliers. In the future, we may use those outliers to train a classifier that identify the anomaly of homicides.

OCR Process on Sample Image

We use a fake image similar to a typical page found in the LAPD records book to illustrate each step of the image processing used in OCR. Since this is a fake image, some of the steps might seem unnecessary but are needed to ensure our processing works for all scanned images.

The example image was a ‘nice’ image to process since it did not have much artifacts. To illustrate some difficulties we came across, we show three images from the book with the characters replaced by bounding boxes to respect confidentiality.



BLUNT FORCE CARDBOARD TUBE QUARREL

CAI, HanQin
Mathematical Sciences 6324
2-21-80 1950

M/A/54

DR# 83-602 929
West LA – RD 315

SUSPECT(S): LINDSTROM, Michael M/W/57

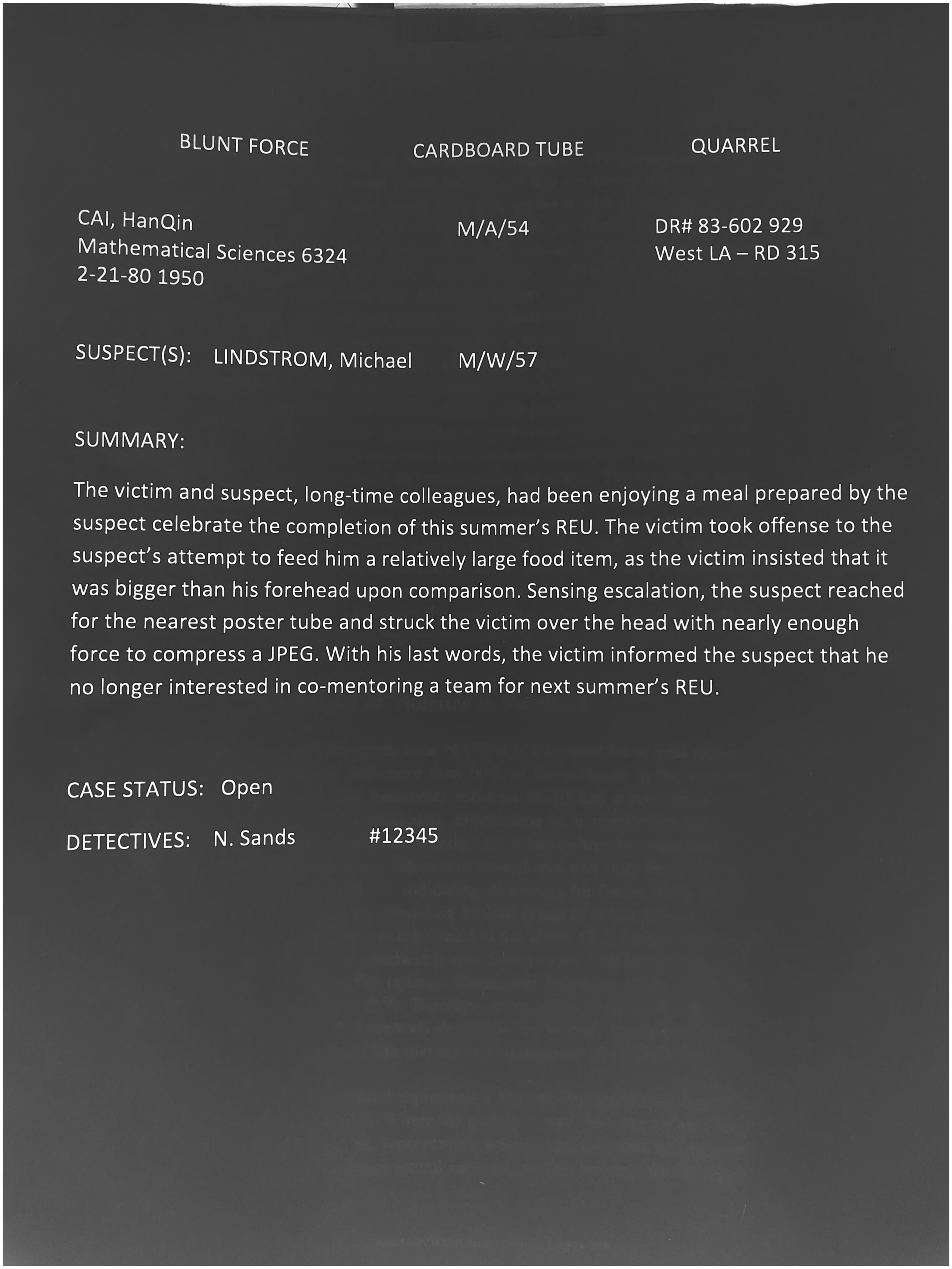
SUMMARY:

The victim and suspect, long-time colleagues, had been enjoying a meal prepared by the suspect celebrate the completion of this summer's REU. The victim took offense to the suspect's attempt to feed him a relatively large food item, as the victim insisted that it was bigger than his forehead upon comparison. Sensing escalation, the suspect reached for the nearest poster tube and struck the victim over the head with nearly enough force to compress a JPEG. With his last words, the victim informed the suspect that he no longer interested in co-mentoring a team for next summer's REU.

CASE STATUS: Open

DETECTIVES: N. Sands #12345

Figure 1: The original image.



BLUNT FORCE	CARDBOARD TUBE	QUARREL
CAI, HanQin Mathematical Sciences 6324 2-21-80 1950	M/A/54	DR# 83-602 929 West LA – RD 315

SUSPECT(S): LINDSTROM, Michael M/W/57

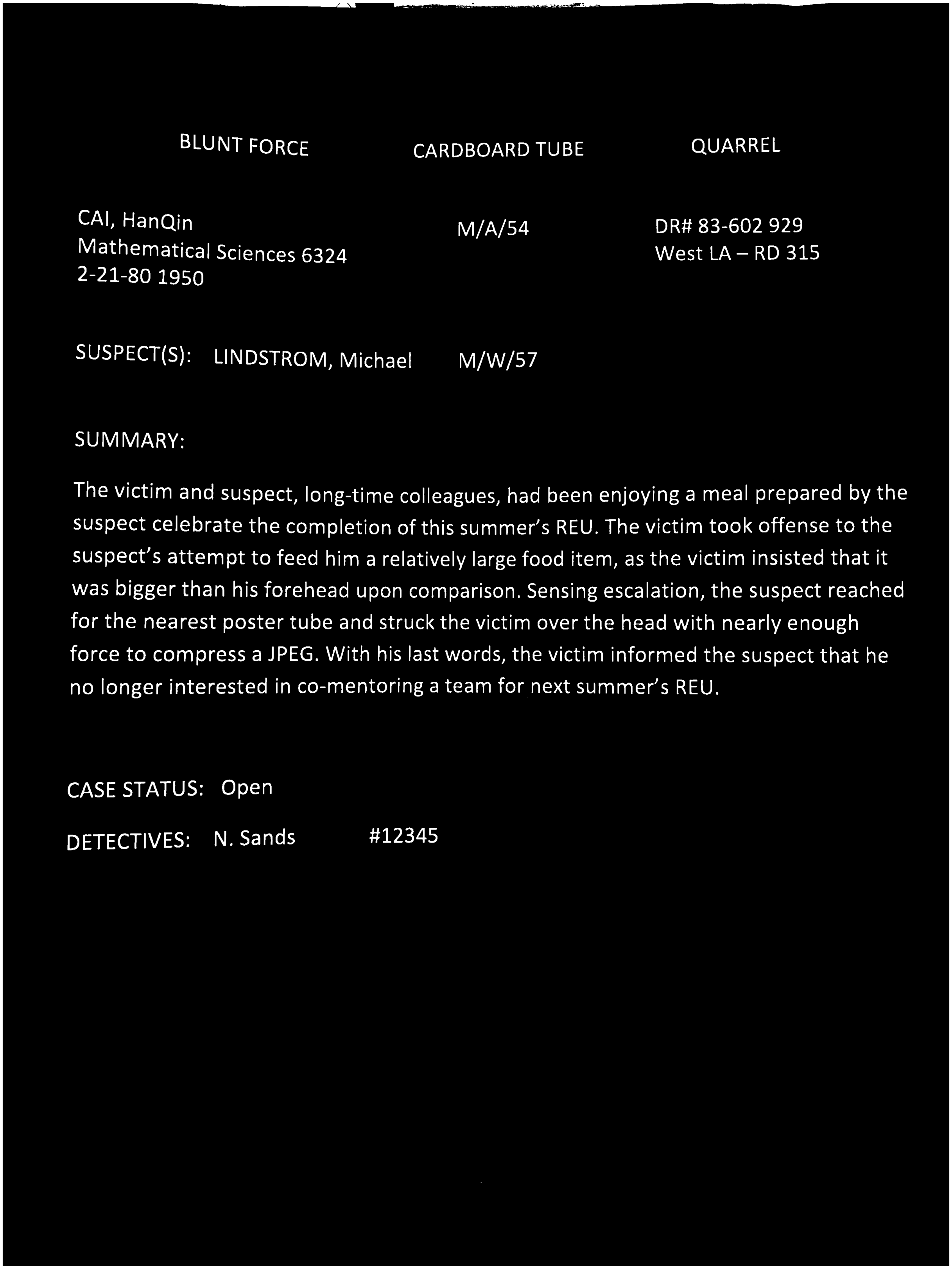
SUMMARY:

The victim and suspect, long-time colleagues, had been enjoying a meal prepared by the suspect celebrate the completion of this summer’s REU. The victim took offense to the suspect’s attempt to feed him a relatively large food item, as the victim insisted that it was bigger than his forehead upon comparison. Sensing escalation, the suspect reached for the nearest poster tube and struck the victim over the head with nearly enough force to compress a JPEG. With his last words, the victim informed the suspect that he no longer interested in co-mentoring a team for next summer’s REU.

CASE STATUS: Open

DETECTIVES: N. Sands #12345

Figure 2: The inverted image.



BLUNT FORCE CARDBOARD TUBE QUARREL

CAI, HanQin
Mathematical Sciences 6324
2-21-80 1950

M/A/54

DR# 83-602 929
West LA – RD 315

SUSPECT(S): LINDSTROM, Michael M/W/57

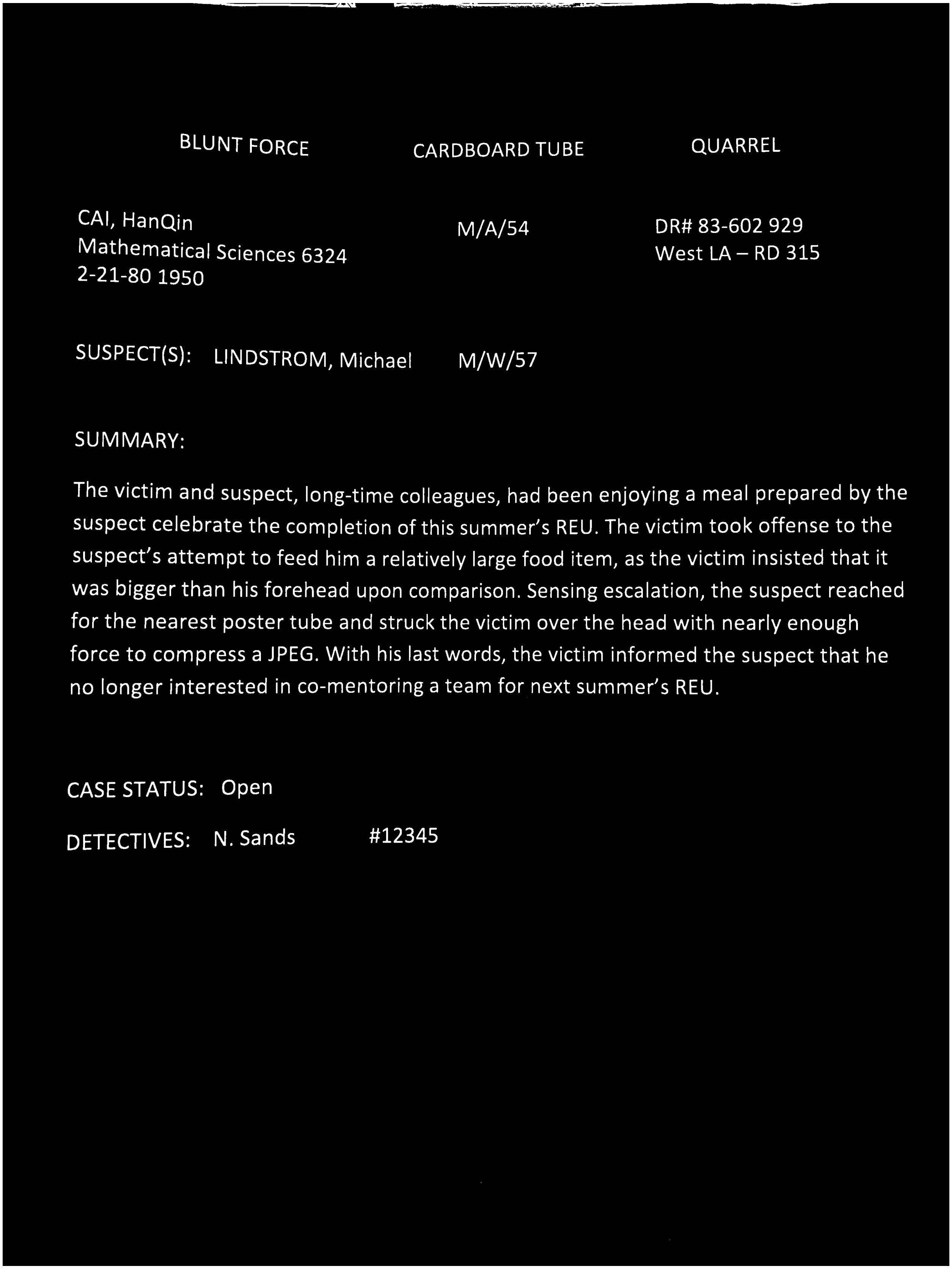
SUMMARY:

The victim and suspect, long-time colleagues, had been enjoying a meal prepared by the suspect celebrate the completion of this summer's REU. The victim took offense to the suspect's attempt to feed him a relatively large food item, as the victim insisted that it was bigger than his forehead upon comparison. Sensing escalation, the suspect reached for the nearest poster tube and struck the victim over the head with nearly enough force to compress a JPEG. With his last words, the victim informed the suspect that he no longer interested in co-mentoring a team for next summer's REU.

CASE STATUS: Open

DETECTIVES: N. Sands #12345

Figure 3: The image after Otsu thresholding is applied.



BLUNT FORCE CARDBOARD TUBE QUARREL

CAI, HanQin
Mathematical Sciences 6324
2-21-80 1950

M/A/54

DR# 83-602 929
West LA – RD 315

SUSPECT(S): LINDSTROM, Michael M/W/57

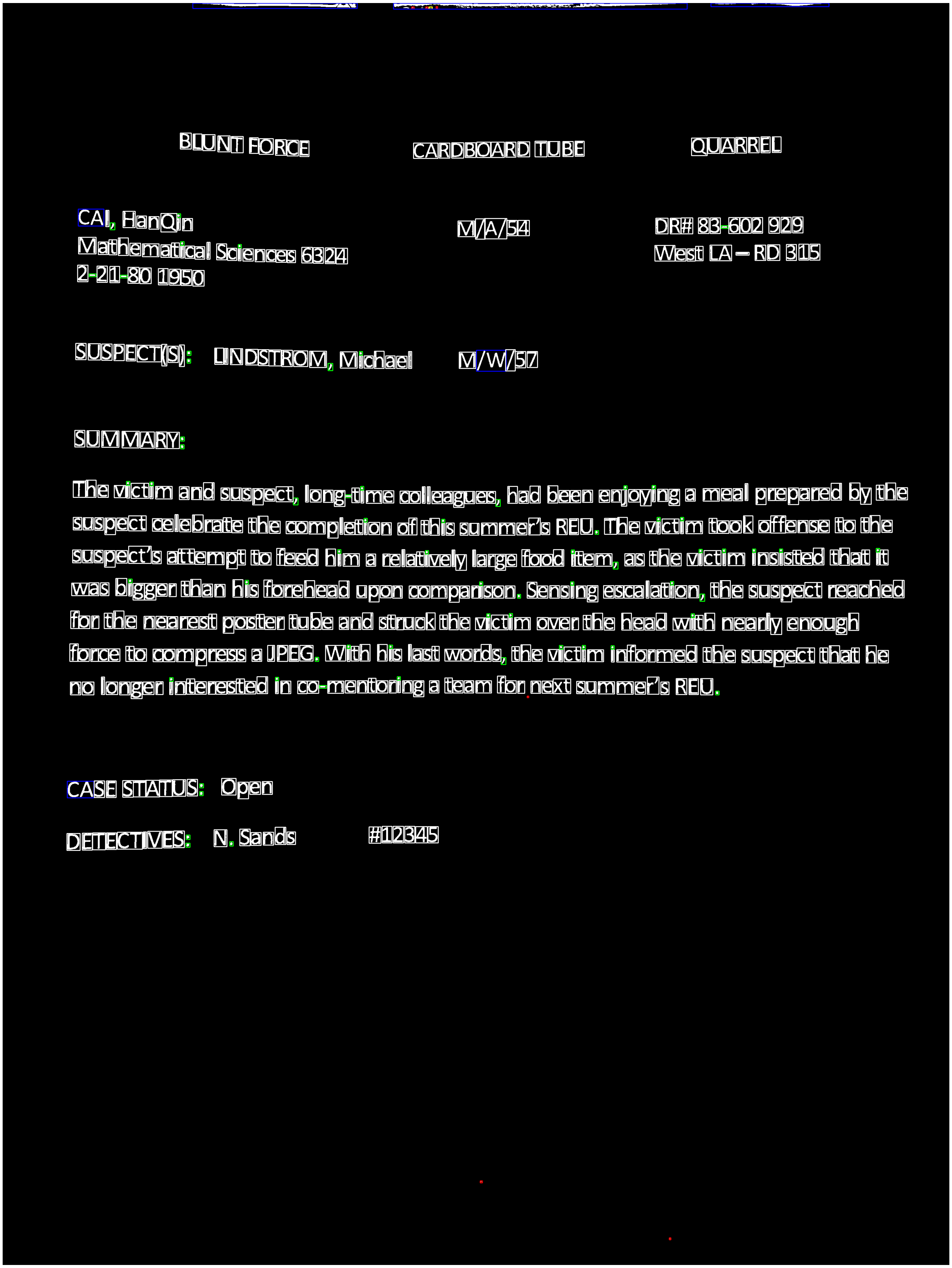
SUMMARY:

The victim and suspect, long-time colleagues, had been enjoying a meal prepared by the suspect celebrate the completion of this summer's REU. The victim took offense to the suspect's attempt to feed him a relatively large food item, as the victim insisted that it was bigger than his forehead upon comparison. Sensing escalation, the suspect reached for the nearest poster tube and struck the victim over the head with nearly enough force to compress a JPEG. With his last words, the victim informed the suspect that he no longer interested in co-mentoring a team for next summer's REU.

CASE STATUS: Open

DETECTIVES: N. Sands #12345

Figure 4: The image after adaptive Gaussian thresholding is applied.



BLUNT FORCE CARDBOARD TUBE QUARREL

CAI, HanQin
Mathematical Sciences 6324
2-21-80 1950

M/A/54

DR# 83-602 929
West LA – RD 315

SUSPECT(S): LINDSTROM, Michael M/W/57

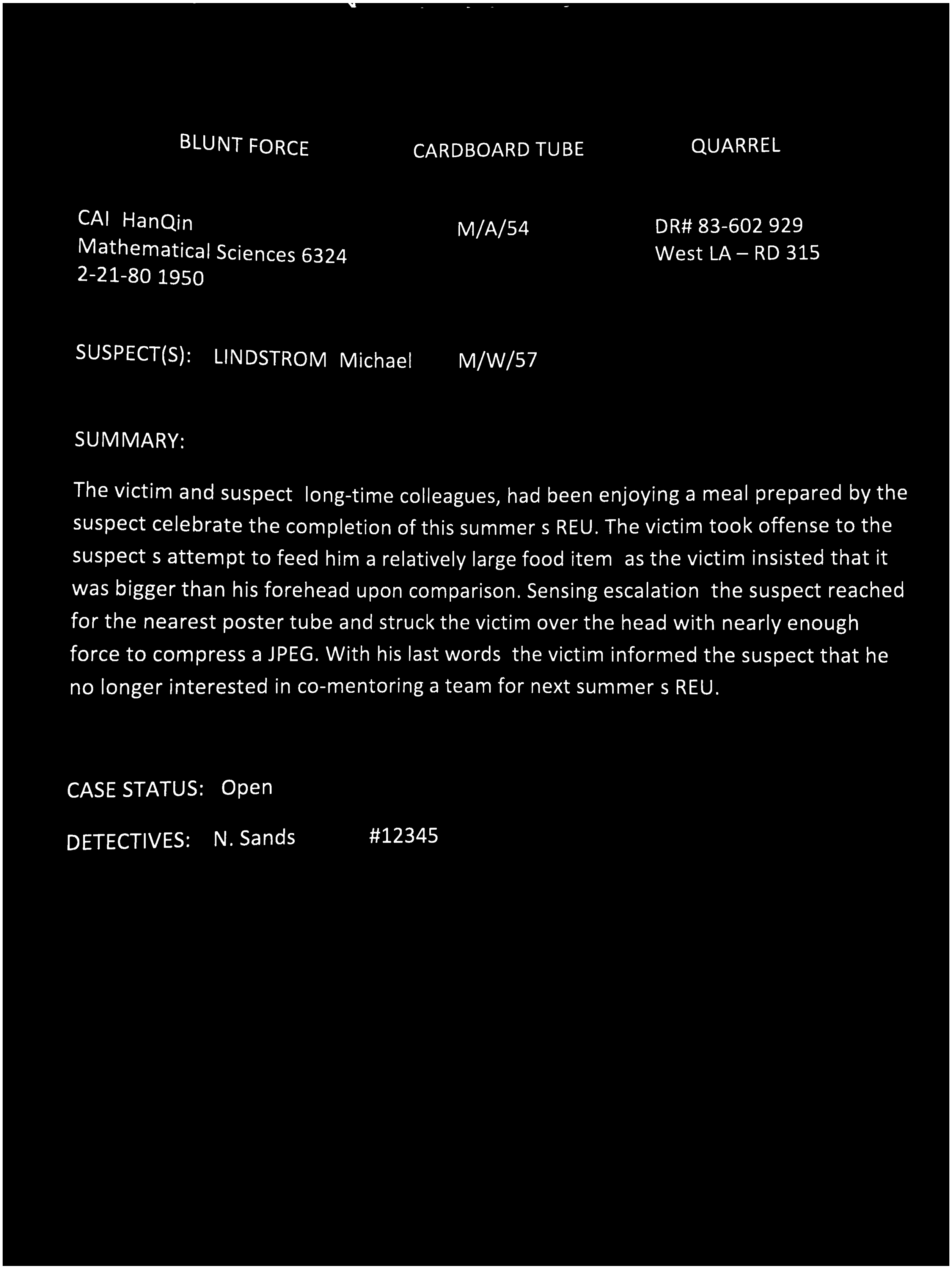
SUMMARY:

The victim and suspect, long-time colleagues, had been enjoying a meal prepared by the suspect celebrate the completion of this summer's REU. The victim took offense to the suspect's attempt to feed him a relatively large food item, as the victim insisted that it was bigger than his forehead upon comparison. Sensing escalation, the suspect reached for the nearest poster tube and struck the victim over the head with nearly enough force to compress a JPEG. With his last words, the victim informed the suspect that he no longer interested in co-mentoring a team for next summer's REU.

CASE STATUS: Open

DETECTIVES: N. Sands #12345

Figure 5: The image with bounding boxes around each connected component. Note that the differently colored bounded boxes refer to the different ways each connected component is processed.



BLUNT FORCE   CARDBOARD TUBE   QUARREL

CAI HanQin
Mathematical Sciences 6324
2-21-80 1950

M/A/54

DR# 83-602 929
West LA – RD 315

SUSPECT(S): LINDSTROM Michael M/W/57

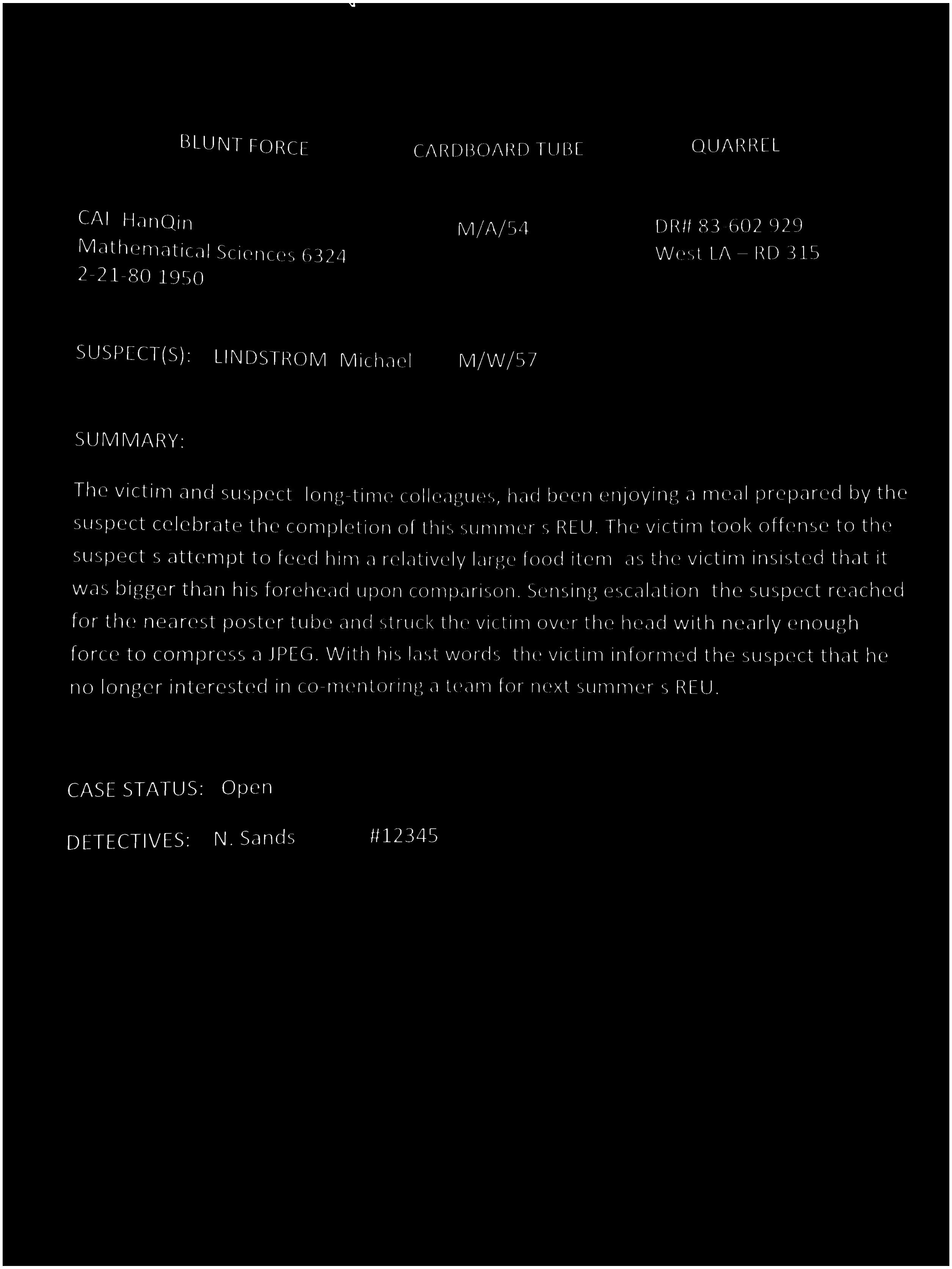
SUMMARY:

The victim and suspect long-time colleagues, had been enjoying a meal prepared by the suspect celebrate the completion of this summer s REU. The victim took offense to the suspect s attempt to feed him a relatively large food item as the victim insisted that it was bigger than his forehead upon comparison. Sensing escalation the suspect reached for the nearest poster tube and struck the victim over the head with nearly enough force to compress a JPEG. With his last words the victim informed the suspect that he no longer interested in co-mentoring a team for next summer s REU.

CASE STATUS: Open

DETECTIVES: N. Sands #12345

Figure 6: The image with noise removed.



BLUNT FORCE CARDBOARD TUBE QUARREL

CAI HanQin M/A/54 DR# 83-602 929
Mathematical Sciences 6324 West LA – RD 315
2-21-80 1950

SUSPECT(S): LINDSTROM Michael M/W/57

SUMMARY:

The victim and suspect long-time colleagues, had been enjoying a meal prepared by the suspect celebrate the completion of this summer s REU. The victim took offense to the suspect s attempt to feed him a relatively large food item as the victim insisted that it was bigger than his forehead upon comparison. Sensing escalation the suspect reached for the nearest poster tube and struck the victim over the head with nearly enough force to compress a JPEG. With his last words the victim informed the suspect that he no longer interested in co-mentoring a team for next summer s REU.

CASE STATUS: Open

DETECTIVES: N. Sands #12345

Figure 7: The image with an erosion kernel applied.

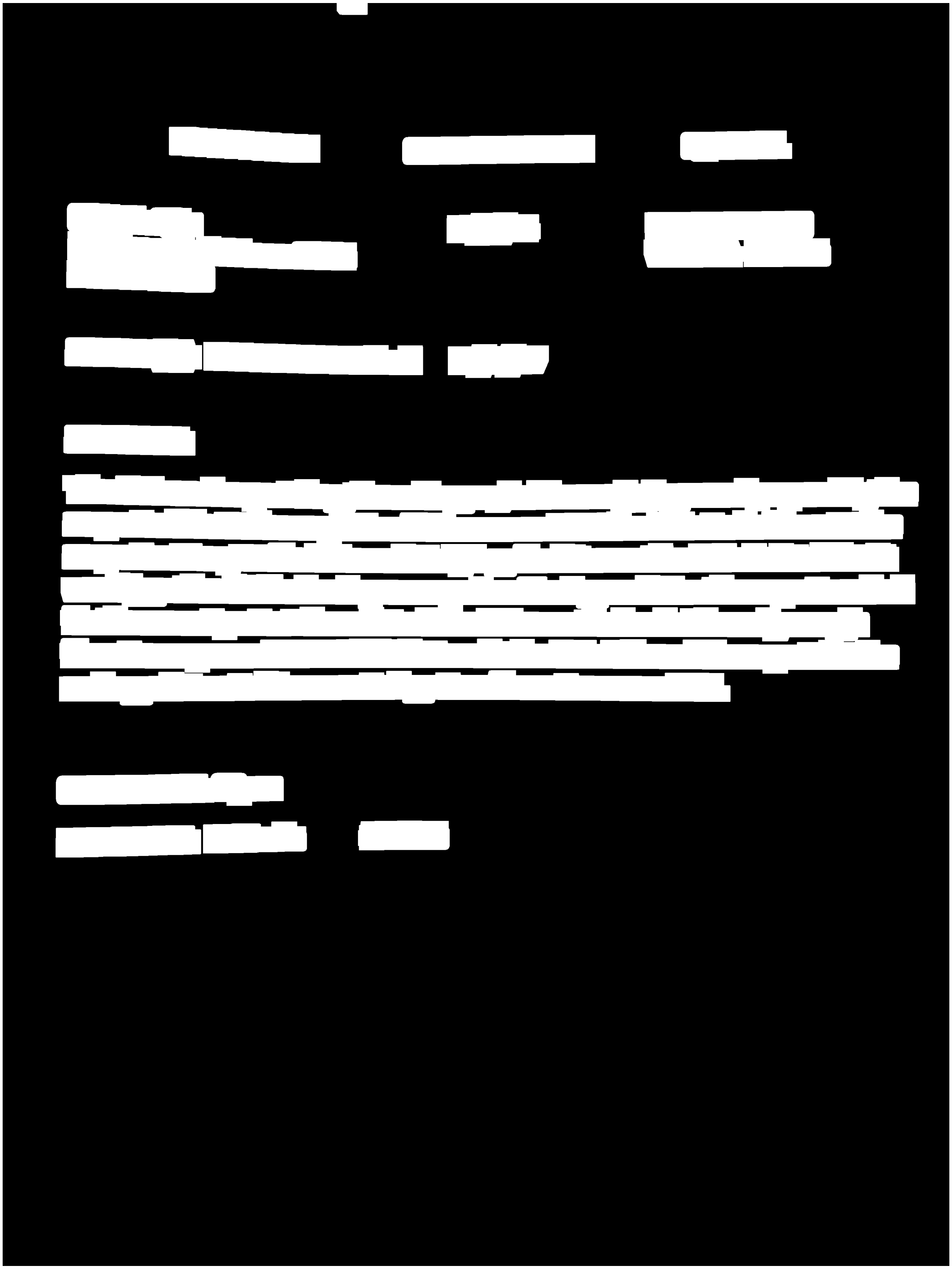


Figure 8: The image with a dilation kernel applied.

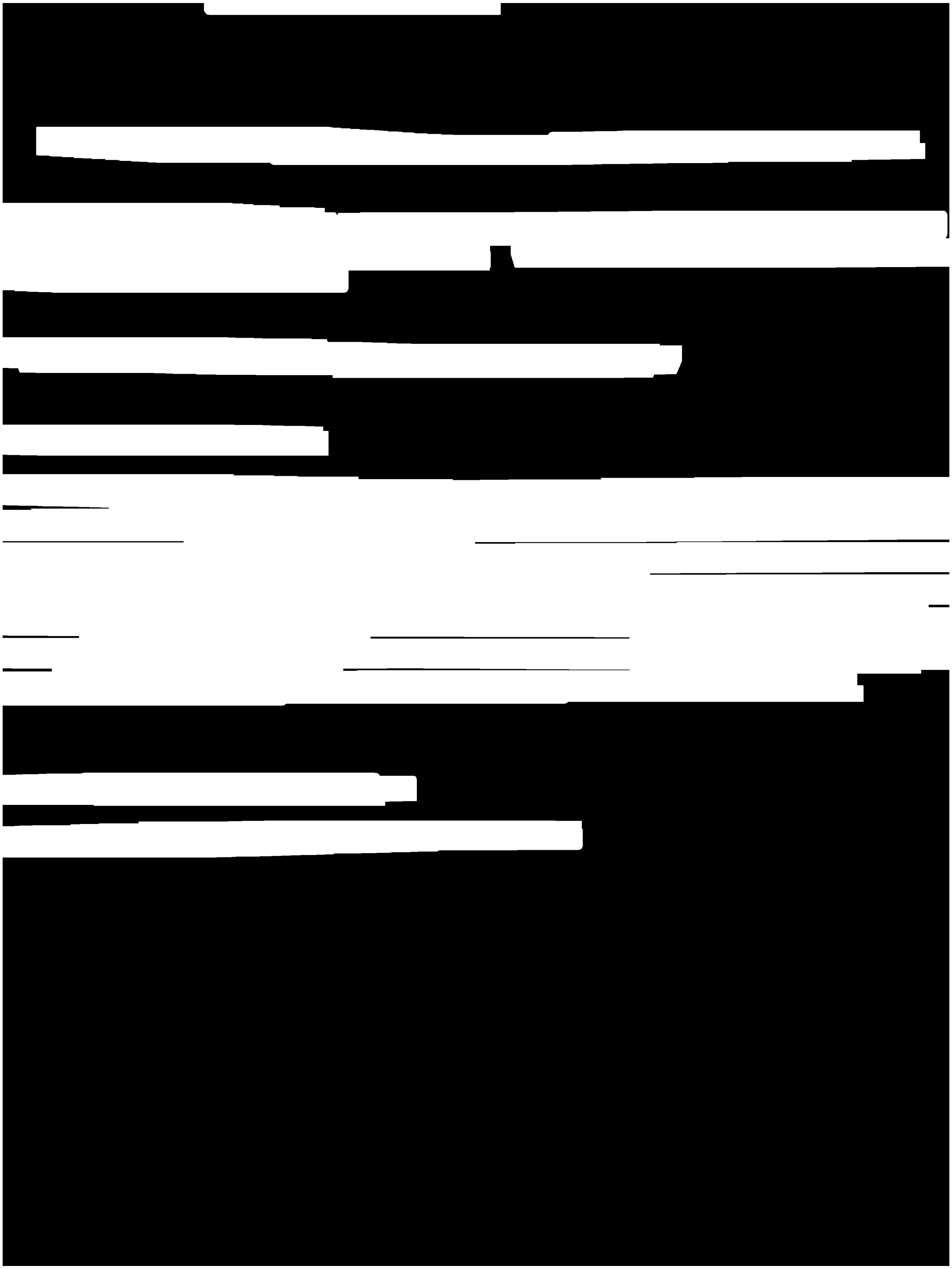


Figure 9: The image with another dilation kernel applied.

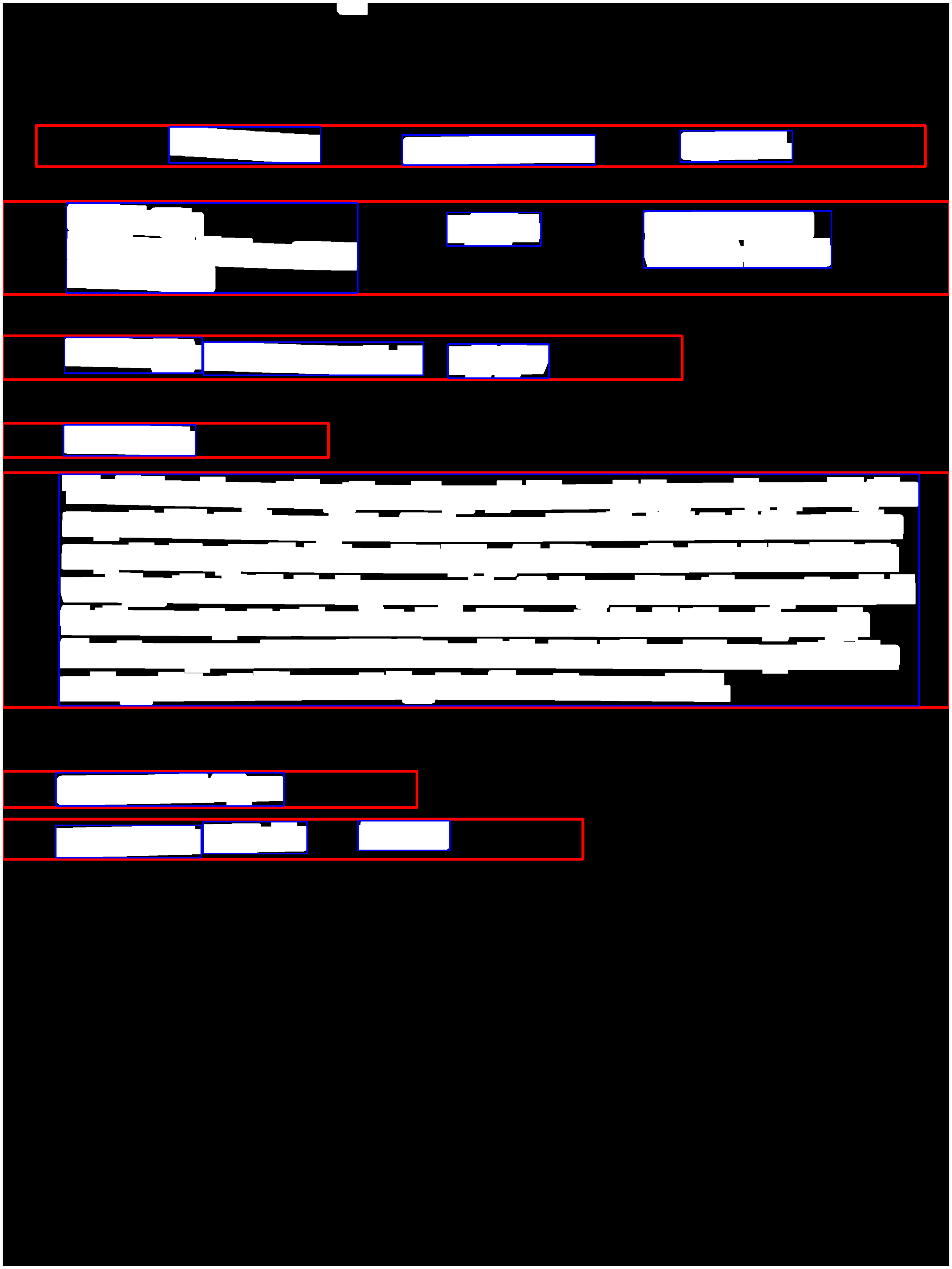
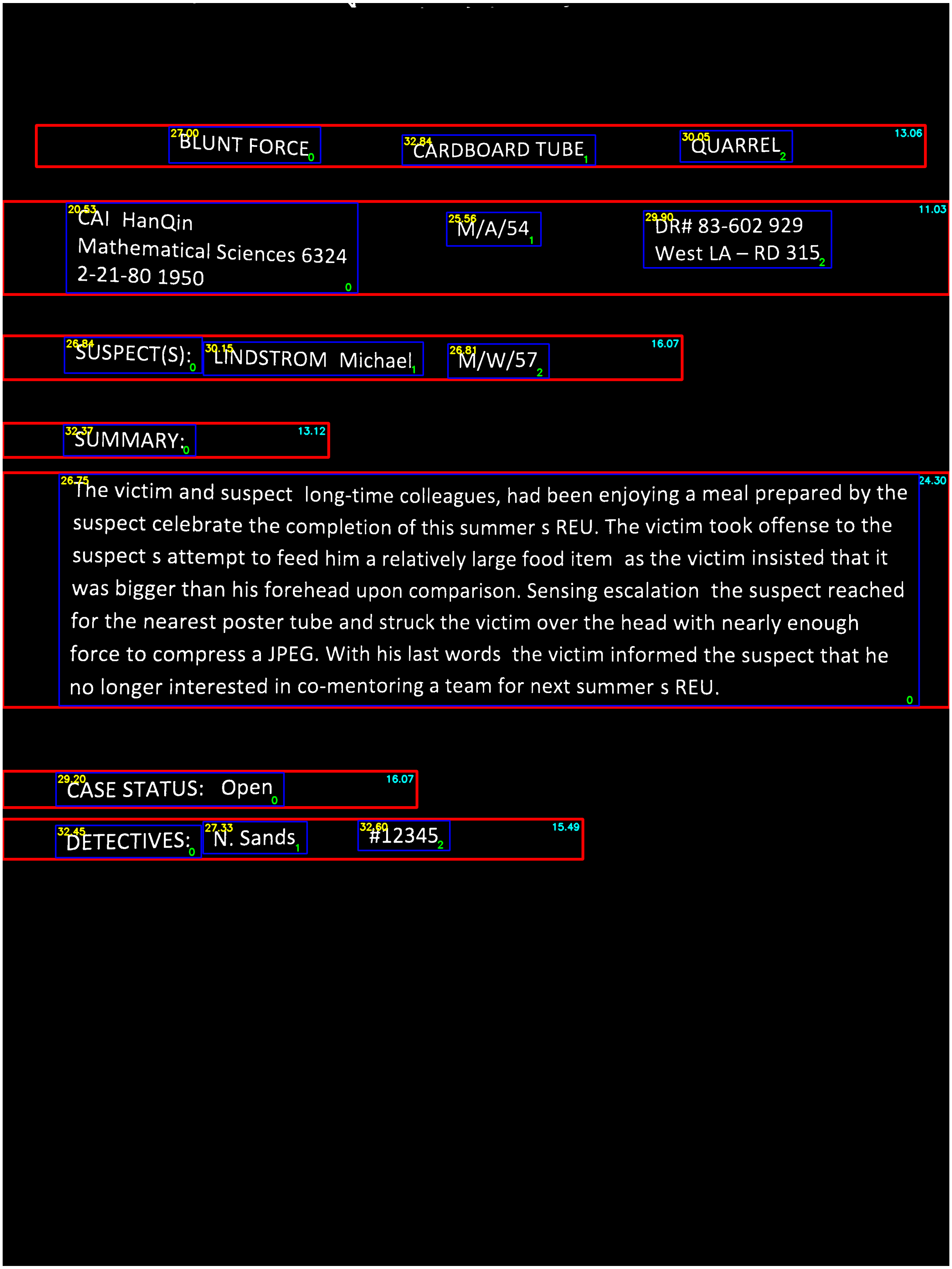


Figure 10: The dilated image with bounding boxes around localized text.



BLUNT FORCE CARDBOARD TUBE QUARREL

CAI HanQin
Mathematical Sciences 6324
2-21-80 1950

M/A/54

DR# 83-602 929
West LA – RD 315

SUSPECT(S): LINDSTROM Michael M/W/57

SUMMARY:

The victim and suspect long-time colleagues, had been enjoying a meal prepared by the suspect celebrate the completion of this summer s REU. The victim took offense to the suspect s attempt to feed him a relatively large food item as the victim insisted that it was bigger than his forehead upon comparison. Sensing escalation the suspect reached for the nearest poster tube and struck the victim over the head with nearly enough force to compress a JPEG. With his last words the victim informed the suspect that he no longer interested in co-mentoring a team for next summer s REU.

CASE STATUS: Open

DETECTIVES: N. Sands #12345

Figure 11: The image with bounding boxes and debug information.

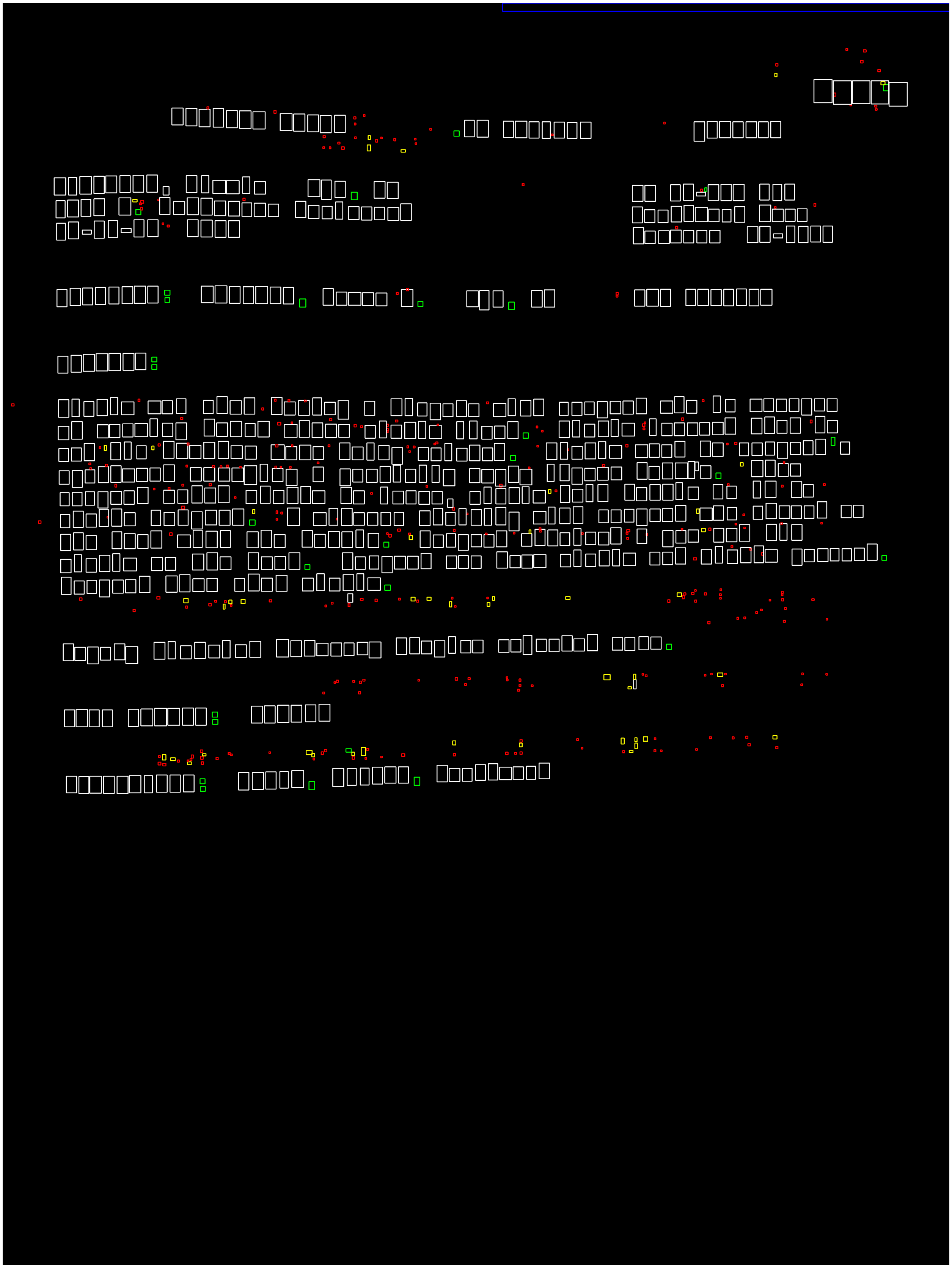


Figure 12: An average image from the book. Note that there are a lot of red contours resulting from bleedthrough from the previous page.

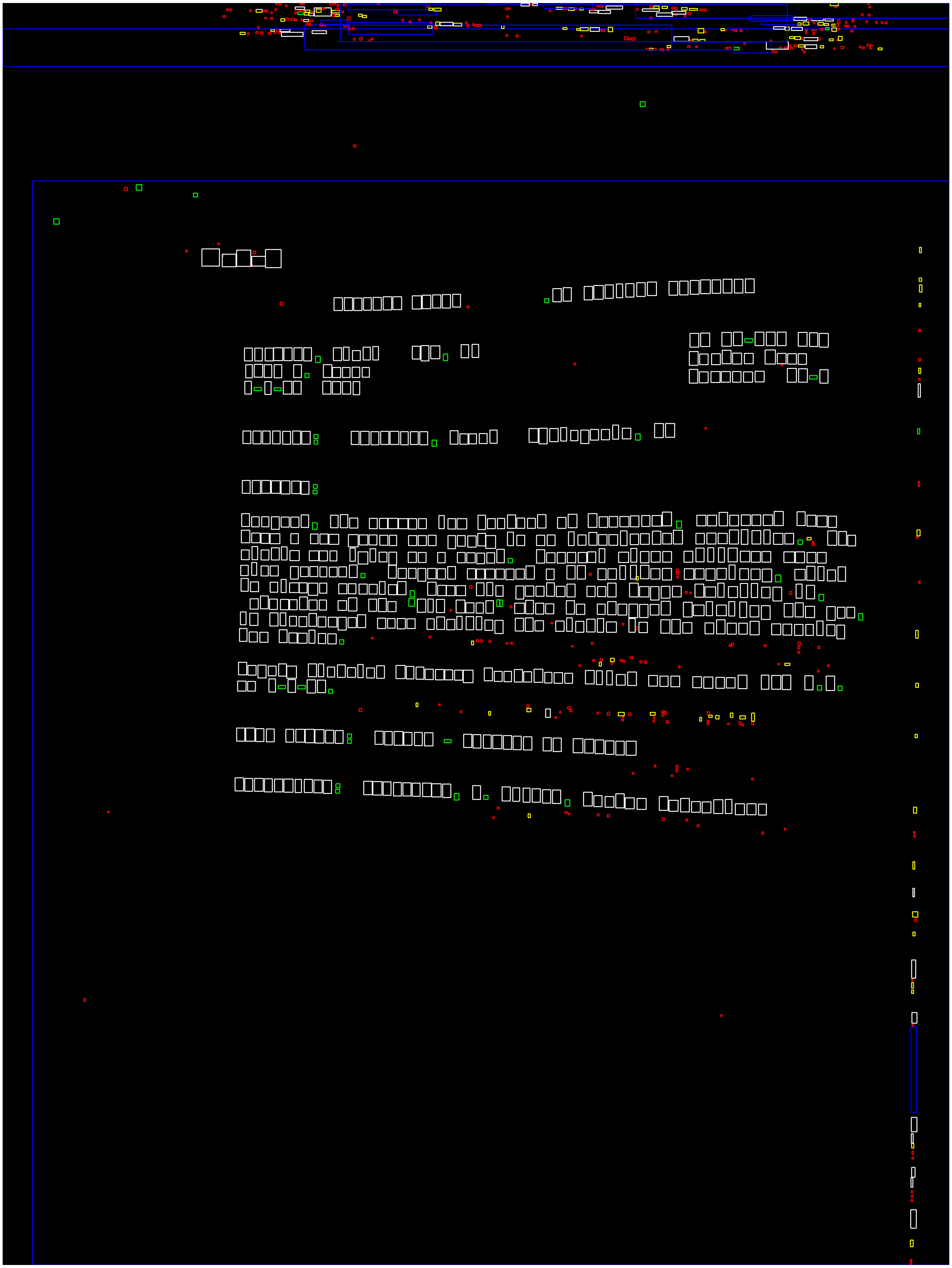


Figure 13: There are a lot of problematic contours in the top portion of the image. This is because the book was scanned on a wooden table and the texture of the wood grains is left as an artifact in the image.

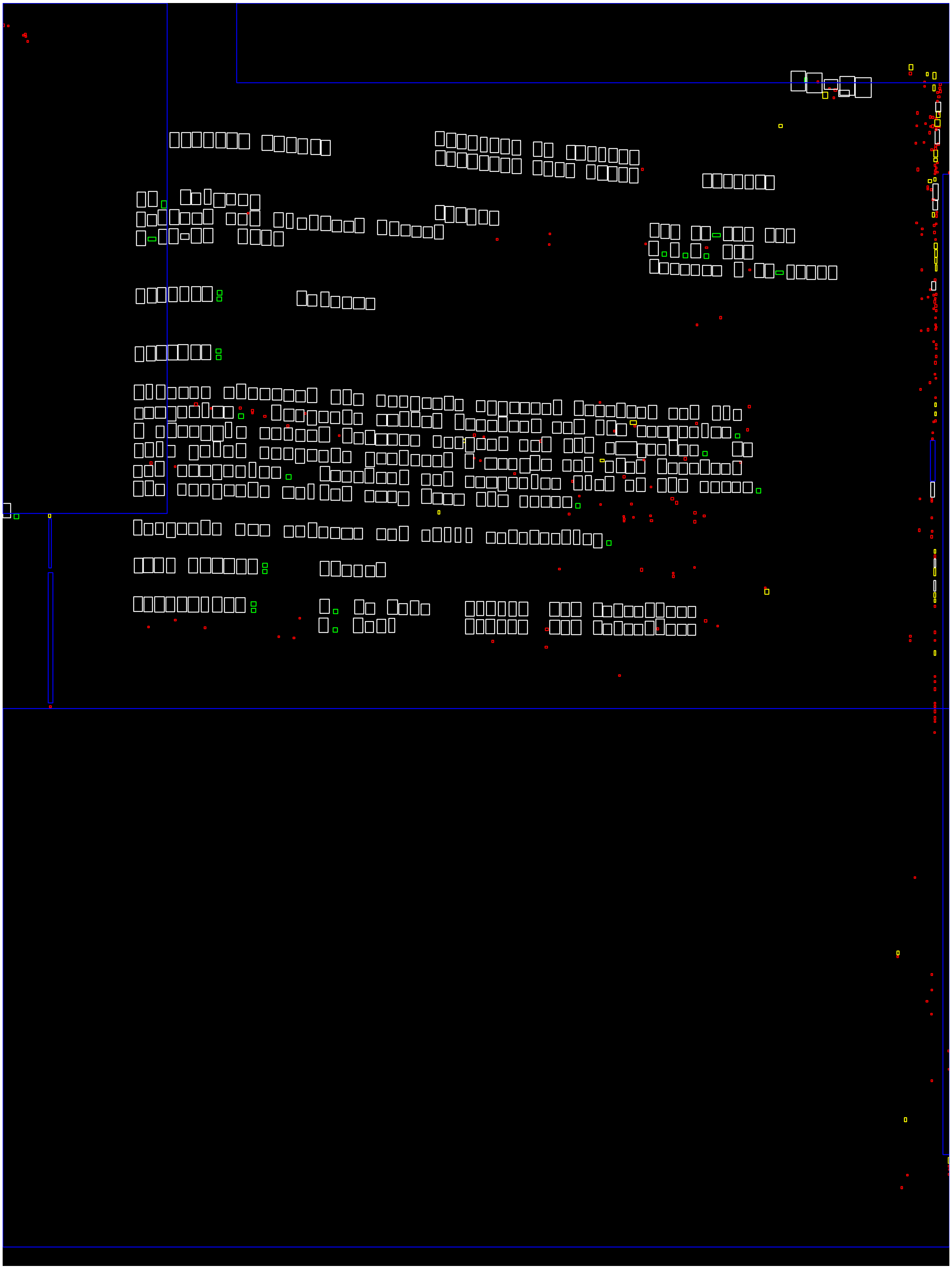


Figure 14: The image is bent, which can lead to incorrect text localization.

Derivation of the Poisson MLE

MLE is a common method to estimate the parameters of a certain distribution that the data may follow. Suppose we have n observations $X_1, \ldots, X_n$ which follow a distribution $f(\theta, X)$, where θ is the parameter to be determined, and f is the probability density function. Then we can estimate the parameter by maximizing

$$\max_{\theta} \prod_{i=1}^{n} f(\theta, X_i) \tag{1}$$

It is equivalent to minimize

$$\min_{\theta} - \sum_{i=1}^{n} \log(f(\theta, X_i)) \tag{2}$$

We find the minimizer by finding the critical point of the object function:

$$\sum_{i=1}^{n} \frac{\partial \log(f(\theta, X_i))}{\partial \theta} = 0 \tag{3}$$

In our case, the distribution is a Poisson distribution:

$$f(\theta, X) = \frac{\theta^X \exp(-\theta)}{X!} \tag{4}$$

The log likelihood function $\ell(\theta)$ will be:

$$\ell(\theta) = \log(\theta) \sum_{i=1}^{n} X_i - n\theta - \sum_{i=1}^{n} \log(X_i!) \tag{5}$$

By taking the derivative, the estimation for θ is:

$$\theta = \frac{\sum\limits_{i=1}^{n} X_i}{n} \tag{6}$$

Bibliography

[1] American factfinder - search. URL: https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml.

[2] Crime data from 2010 to present: Los angeles - open data portal. URL: https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/63jg-8b9z.

[3] Introduction to recurrent neural networks. URL: https://miro.medium.com/max/1838/1*NKhwsOYNUT5xU7Pyf6Znhg.png.

[4] L.a. county regions (v6). URL: http://boundaries.latimes.com/set/la-county-regions-v6/.

[5] La time homicide report. URL: https://homicide.latimes.com.

[6] La times. URL: https://www.latimes.com.

[7] Locally weighted regression. URL: http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume4/cohn96a-html/node7.html.

[8] Los angeles open data. URL: https://data.lacity.org.

[9] *Los Angeles Police Department Murder Investigations.*

[10] Mapping l.a. boundaries api. URL: http://boundaries.latimes.com/sets/.

[11] Nlp-guidance latent dirichlet allocation (lda). URL: https://moj-analytical-services.github.io/NLP-guidance/LDA.html.

[12] One corner. four killings. URL: https://www.latimes.com/projects/la-me-harvard-park-homicides/.

[13] Python deep learning tutorial: Create a gru (rnn) in tensorflow. URL: https://www.data-blogger.com/wp-content/uploads/2017/08/gru.

[14] Python factor analysis library (pca, ca, mca, mfa, famd). URL: https://github.com/MaxHalford/prince.

[15] Symspell: 1 million times faster through symmetric delete spelling correction algorithm. URL: https://github.com/wolfgarbe/SymSpell.

[16] Understanding lstm networks. URL: https://colah.github.io/images/post-covers/lstm.

[17] United states census bureau. URL: https://www.census.gov/academy.

[18] Us census bureau geocoder. URL: https://geocoding.geo.census.gov/geocoder.

[19] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. URL: http://dl.acm.org/citation.cfm?id=944919.944937.

[20] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[21] Dallas Card, Chenhao Tan, and Noah A. Smith. Neural models for documents with metadata. In *Proceedings of ACL*, 2018.

[22] François Chollet et al. Keras. `https://keras.io`, 2015.

[23] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014. URL: `http://arxiv.org/abs/1412.3555`, `arXiv:1412.3555`.

[24] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[25] Jeremy Du, Kelly Flood, Matin Ghavamizadeh, Bumsu Kim, Markus Plack, Samuel Tan, and Hanqing Yao. Dynamic topic modeling: Spatiotemporal analysis of los angeles twitter data.

[26] A Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2003.

[27] A Stewart Fotheringham and Taylor M Oshan. Geographically weighted regression and multicollinearity: dispelling the myth. *Journal of Geographical Systems*, 18(4):303–329, 2016.

[28] Francois "Husson and Julie" Josse. Multiple correspondence analysis. 01 2014.

[29] Anthony Kay. Tesseract: An open-source optical character recognition engine. *Linux J.*, 2007(159):2–, July 2007. URL: `http://dl.acm.org/citation.cfm?id=1288165.1288167`.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. URL: `http://arxiv.org/abs/1412.6980`.

[31] Da Kuang, P. Jeffrey Brantingham, and Andrea L. Bertozzi. Crime topic modeling. *Crime Science*, 6(1):12, Dec 2017. URL: `https://doi.org/10.1186/s40163-017-0074-0`, `doi:10.1186/s40163-017-0074-0`.

[32] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[33] Cheryl L Maxson, Margaret A Gordon, and Malcolm W Klein. Differences between gang and nongang homicides. *Criminology*, 23(2):209–222, 1985.

[34] Andrzej Makiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303 – 342, 1993. URL: `http://www.sciencedirect.com/science/article/pii/009830049390090R`, `doi:https://doi.org/10.1016/0098-3004(93)90090-R`.

[35] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013. URL: `http://arxiv.org/abs/1310.4546`, `arXiv:1310.4546`.

[36] James A Pike. What is second degree murder in california. *S. Cal. L. Rev.*, 9:112, 1935.

[37] Wendy C Regoeczi and John P Jarvis. Beyond the social production of homicide rates: Extending social disorganization theory to explain homicide case outcomes. *Justice Quarterly*, 30(6):983–1014, 2013.

[38] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[39] Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *CoRR*, abs/1808.03314, 2018. URL: `http://arxiv.org/abs/1808.03314`, `arXiv:1808.03314`.

[40] Namrata Vaswani, Thierry Bouwmans, Sajid Javed, and Praneeth Narayanamurthy. Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery. *IEEE*, Jun 2018. URL: `https://doi.org/10.1109/MSP.2018.2826566`, `doi:10.1109/MSP.2018.2826566`.