THE APPLICATION OF LONGITUDINAL MIXED NON-GAUSSIAN ZERO-INFLATED MODELS FOR ESTIMATING RATES AND TRENDS OF HATE CRIMES IN THE U.S. FROM 2005-2019

By

LAUREL ELESA BEATY

B.S., Seattle University 2017

A thesis submitted

to the Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Biostatistics and Informatics

2021

This thesis for the Master of Science degree by

Laurel Elesa Beaty

has been approved for the

Department of Biostatistics and Informatics

by

Matthew J. Strand, Chair

Kathryn Colborn

Nichole Carlson

Date:

Beaty, Laurel Elesa (M.S., Biostatistics)

The Application of Longitudinal Mixed Non-Gaussian Zero-Inflated Models for Estimating Rates and Trends of Hate Crimes in the U.S. from 2005-2019

Thesis Directed by Associate Professor Matthew J. Strand

**ABSTRACT**

Understanding how the prevalence and patterns of hate crimes vary across time and across the country is imperative to implementing changes in policy and protecting communities from further crimes. Hate crimes in the US operate in distinctive ways in comparison to other types of crimes and the data collected for these crimes offer unique analytical challenges.

This study aims to investigate the spatial and temporal clustering and trends of hate crimes in the U.S. from 2005 to 2019. The data was collected from the FBI Uniform Crime Reporting Database, and therefore this study investigates hate crimes reported from individual police departments rather than the actual occurrence of crimes. These data have an overabundance of zeros compared to traditionally spatially varying count data and may have over-dispersion. We investigated the use of generalized linear mixed models using distribution mixtures to estimate the prevalence of hate crimes in the United States. Several different distributions and link functions were compared along with various parameterizations of both fixed and random effects. The utility of zero-inflated mixture models was compared to non-zero-inflated mixed models. Random effects estimates and fitted values were used to identify various trends across time on the city level, and these trends were further explored to isolate significant differences in locations based on the overall fitted trend.

We found that the overall population trend was a slight linear decrease by about 3.8% each year in the rate of reported hate crimes (p = 0.02). However, this trend only captured approximately 3.7% of the variability of rates of hate crimes. The random effects captured an additional 26.2% of the variability and 9 distinct temporal trends were identified. Temporal trends across cities varied, and this variability was captured by the polynomial random effects. Based on AIC and model interpretation, the zero-inflated Negative Binomial model fit the data the best in comparison to all other modeling techniques. We were able to estimate the dispersion and excess zeros present in the data. The results suggest that further investigation into the true rates of hate crimes is necessary, and the national data may indicate many irregularities and under-reporting.

The form and content of this abstract are approved. I recommend its publication.

Approved: Matthew J. Strand

**TABLE OF CONTENTS**

CHAPTER

1. INTRODUCTION………………………………………………...…….…………….1

Overall Research Aims…………………………………………...…….……………..1

Statistical Background…………………………………………...…….……………...1

Counts and Overdispersion; Poisson and Negative Binomial…………………….1

Mixture Models…………………………………………...…….…………………3

Zero-Inflated Negative Binomial (ZINB) Model………………………..………..7

General Model Specification……………………………………………...7

Estimation and Optimization……………………………………………...9

Introduction to Subject Area; Hate Crimes………………………………….……..…9

1. METHODS…………………………………………………….…………………….12

Data Compilation and Cleaning……………………………….……………………..12

Descriptive Statistics……………………………….………………………………...14

Statistical Modeling……………………………….…………….…………………...14

Final Model Specifications……………………….…………….…….……………...17

Sensitivity Analysis……………………….…………….…………………………...18

Extensions of Final Model……………….…………….………………………….....18

Grouping of Cities by Random Effects and Predicted Values……………………….18

Comparing Expected and Observed Counts and Rates………………………………21

1. RESULTS……..………………………………………….………………………….23

Descriptive Statistics…………………………………….…………………………...23

Statistical Modeling…………………………………….……………………………26

Interpretation of Final Model Parameters…………………………………………....30

Sensitivity Analysis………………………………………………………………….31

Groupings of Cities by Trajectories………………………………………………….32

Expected vs. Observed…………………………………………………………….…37

1. DISCUSSION………………………………………………………………………..40

Model-Based Conclusions…………………………………………………….……..40

Barriers to Reporting………………………………………………………………....43

Type of Bias and Crime…………………………………………………………...…46

Limitations………………………………………………………………………...…46

Future Directions……………………………..…………………………………...…47

REFERENCES

APPENDIX

**LIST OF TABLES**

TABLE

1. Table 1: Locations in the U.S. based on Census delineations……...……………………13
2. Table 2: Intercept only generalized linear mixed model result………………………….26
3. Table 3: Parameterization of time results………………………………………………..28
4. Table 4: Random effects exploration results………………………………………….…29
5. Table 5: Final model comparisons between distributions……………………………….29
6. Table 6: ZINB final model results……………………………………………………….30
7. Table 7: Sample size per trajectory grouping……………………………………………32
8. Table 8: Tests of association for continuous variables by trajectory group……………..34
9. Table 9: Tests of Association for region by trajectory grouping along with descriptive statistics………………………………………………………………………………….35

**LIST OF FIGURES**

FIGURE

1. Figure 1: Simulated Gaussian Data from a Gaussian Mixture Model…………………….4
2. Figure 2: Simulated Poisson Mixture Model histogram…………………………………...5
3. Figure 3: Simulated zero-inflated Negative Binomial data………….……………………..6
4. Figure 4: Loess plots of average counts across time……………………………………..23
5. Figure 5: Loess plots of average rates across time………………………………………23
6. Figure 6: Boxplots of population adjusted rate stratified by region……………………..24
7. Figure 7: Fitted trajectories grouped into the 9 trajectories…………………………...…33
8. Figure 8: Map with latitude and longitude points color coded based on linear temporal trajectory…………………………………………………………………………………36
9. Figure 9: Map with latitude and longitude points color coded based on quadratic temporal trajectory………………………………………………………………………………....37
10. Figure 10: Difference in expected and observed Counts by year and region…………....38
11. Figure 11: Observed (green) and expected (blue) rates for each city across time and stratified by region……………………………………………………………………….39

**CHAPTER I**

**INTRODUCTION**

**Overall Research Aims**

This project aims to explore the application of generalized linear mixed models with mixture distributions to reported hate crime data. These data present unique modeling challenges, and the models developed to analyze the data incorporate many complex elements. The two primary questions that are explored are 1) What is the overall trend across time in the rates of reported hate crimes and 2) How do city-specific trends differ from national trends? Both of these questions are of great importance and will incorporate both a broad look at overall trends and a location-specific investigation into the potential heterogeneity of the rate of reported hate crimes. The utility of the models developed includes the ability to identify unusual reporting and the probabilities of zero-inflation.

**Statistical Background**

**Counts and Overdispersion: Poisson and Negative Binomial**

Count data are common in many applications including crime data and biomedical data. One of the primary tools developed to deal with counts was the Poisson model (Cameron & Trivedi, 1998). However, the Poisson distribution assumes that the mean and variance are equal (i.e. ) and this is a strict parameterization that fails to appropriately model data that does not meet this assumption. There are two types of failures of the assumption of equidispersion: underdispersion and overdispersion. Underdispersion is what occurs when the variance is smaller than the mean and overdispersion is what occurs when the variance is larger than the mean. When we assume equidispersion, the parameter estimates are likely to be biased and the standard error may be underestimated in the case of overdispersion and overestimated in the case of underdispersion (Fang, 2008).

True count data are more often overdispersed (Hoef & Boveng, 2007), and a common way to model overdispersed count data is to use Negative Binomial models instead. The Negative Binomial offers greater flexibility by way of introducing latent heterogeneity in the conditional mean of the Poisson (Greene 2008). Greene (2008) outlines this extension in Equation 1. Note that all equations presented for models are conditioned on the data ***x****i*.

[1]

Where is assumed to follow a Gamma distribution with one parameter . Therefore, the marginal Negative Binomial distribution is specified by Greene (2008) in Equation 2:

[2]

Where . Based on this parameterization, estimation is straightforward with Maximum likelihood (Greene 2008). Therefore, the Negative Binomial offers flexibility by including a parameter for overdispersion and allowing the variance to have a quadratic (or linear in some cases) relationship with the mean. In addition, we can extend a Negative Binomial model into the generalized linear mixed model framework through the use of link functions and random effects. The inclusion of random effects allows us to include geographic and temporal clustering, and to allow cities to have their own trajectories over time.

**Mixture Models**

Mixture models are a very flexible class of models that are capable of modelling data with more than one underlying distribution. The phenomenon of multiple underlying distributions can be present in many different ways such as assumed heterogeneity in clusters, mixtures of continuous distribution, mixtures of discrete distributions, and mixtures of continuous and discrete distributions. The general form of a mixture model is defined in Equation 3 based Ghojogh et al. (2019).

[3]

Where K is the number of component distributions and are the weights constrained by . Each of the components has a corresponding set of parameters and can either be probability density functions or probability mass functions. The components can be from same parametric family (also called parametric mixture models) or can be of different distributions. Parametric mixture models can be used to model data that have assumed heterogeneity in the random effects (Verbeke, 2011) and can be used to model data that arises from one type of distribution but with several different means and standard deviations.

Bouguila and Fan (2020) illustrate a common form of parametric mixture models called Gaussian Mixture Models. They are an example of latent variable models (which can also be referred to as hidden models) and Bouguila and Fan (2020) derive the log likelihood of a mixture of normal and illustrate how the Expectation Maximization (EM) Algorithm can be used to obtain estimates. The Gaussian mixture models can be defined as follows in Equation 4 by Bouguila and Fan (2020):

[4]

Where the *N* is the normal probability density function, are the respective means and are the respective covariances.

Chart, histogram

Description automatically generatedThe applications of these model are illustrated in Figure 1. Data was randomly simulated from 3 different normal distributions and pooled into one outcome. The three curves show the individual distributions that are incorporated into the overall Gaussian Mixture Model. The overlap in the distributions illustrates the need for the which will weigh the probability of a value arising from one of the distributions (i.e. a 2 can come from any of the three distributions, and the weights determine this probability).

Figure 1: Simulated Gaussian Data from a Gaussian Mixture Model

Another example of mixture models are continuous-discrete mixture models. These models are useful when an outcome may have both a continuous outcome and a discrete outcome (binary or counts). Lord et al. (2005) illustrate how a Poisson-Gamma can be successfully used to model car crashes.

Oftentimes, mixture models can be used to model zero-inflated data. These data are characterized by an excess of zeros in relation to the modelling distribution. An example of this is rainfall data; data collected in the form of inches of rainfall can be modelled through a continuous distribution (i.e. Gamma, Lognormal, etc.) whereas the days with no rainfall can be incorporated through a discrete distribution. Rider (1961) and Cohen (1963) were some of the first to introduce the concept of zero-inflated models through their exploration of mixed discrete distributions such as the Poisson or Negative Binomial. A mixed Poisson distribution is characterized by data that can arise from two different Poisson distributions with probabilities and . A mixed Poisson distribution is fairly straight forward and is defined in Equation 5 following Lord et al. (2005):

[5]

[5]

[5]

Chart, histogram

Description automatically generatedWhere *n* is the observed count and and are the respective means of the two Poisson distributions. A histogram with simulated data that follows a mixture distribution of two Poisson distributions is visualized in Figure 2. A helpful feature of this parameterization is that we can derive the probability of a count arising from one of the two different underlying distributions that make up the mixture distribution.

Figure 2: Simulated Poisson mixture model histogram

Another way of conceptualizing a zero-inflated model is through the combination of two different processes; one process that creates only zeros and one process that creates zeros and positive integers. This type of mixture model is outlined in the following section. Figure 3 shows pooled data from both a randomly simulated Negative Binomial distribution and excess zeros. The dark blue line represents the smoothing of the model that only incorporates the Negative Binomial values, whereas the light blue line indicates the smoothing that includes the excess zeros as well. Clearly, there is a discrepancy that needs to be accounted for in a model.

One type of mixture model that was considered but was not utilized here was the hurdle model. A hurdle model assumes that all zeros come from a separate process (also called a Chart, line chart, histogram

Description automatically generatedstructural source) and all positive integers come from a separate process (or a sampling source) (Hu, Pavlicova & Nunes, 2011). This differs from the zero-inflated models used in this paper as a zero-inflated model assumes that zeros can arise from both processes. The reason a zero-inflated model was chosen instead of a hurdle model is that when considering counts, it is possible to have a true zero count that arises from the Negative Binomial process. Therefore the zeros can indicate a separate process (i.e. a lack of reporting) or the count process (a zero count of crimes).

Figure 3: Simulated zero-inflated Negative Binomial data

**Zero-Inflated Negative Binomial (ZINB) Model**

*General Model Specification*

A ZINB model allows us to incorporate many elements into the model that include fixed effects, random effects, zero-inflation, and overdispersion. In general, a ZINB assumes that the data can arise from two different processes. The first process has probability and can only produce zero counts, and these are considered to be the structural zeros or the excess zeros. The second process is a Negative Binomial, and this process can produce both zeros and counts with probability (1 - ).

Based on an article by NCSS Statistical Software (accessed 2021), we can define the ZINB distribution as follows in Equation 6 where is the Negative Binomial distribution defined in Equation 7. In all following equations, the subscript *i* indicates subject and the subscript *j* indicates time.

[6]

[7]

In Equation 6, estimates the probability of the first process which essentially estimates the probability of zero-inflation. In Equation 7, are the random effects that are included in the mean () as defined in Equation 8, NSZ are the non-structural zeros (the zeros arising from the second process), is the overdispersion parameter which uses a log link, and is the underlying mean of the Negative Binomial which is defined in Equation 8.

*Underlying Mean of the Negative Binomial*

The following Equations (8 and 9) are based on Brooks et al (2017). The underlying mean of the Negative Binomial distribution is defined in Equation 8 and uses a log link.

[8]

We define as the linear combination of predictors, is the linear combination of the random effects, and we include a log offset for the population “at risk” for each subgroup. Both and are rows from either the or the matrix for a given city *i* in a given year *j*. This log offset can also be used to specify the amount of time at risk for each “subject”. An NSZ is a non-structural zero, which is a zero arising from the Negative Binomial process. We are assuming that where is the variance-covariance matrix for the random effects. Various variance-covariance structures are possible within ZINB models. We can see that the mean of the Negative Binomial is flexible enough to include fixed effects, random effects, and various offsets.

*Variance of the Negative Binomial*

There are typically two different parameterizations for the mean-variance relationship of the Negative Binomial; linear and quadratic. The quadratic parameterization was chosen here, as it allows for a quadratic relationship between the underlying mean and variance of the Negative Binomial which allows for more flexibility (Hardin, Hilbe, & Hilbe, 2007). In all cases, we are assuming that and this assumption can be checked by comparing a Poisson with a Negative Binomial and investigating the dispersion estimates. A Poisson model is restrictive in that and therefore does not allow for under or overdispersion. The Variance of the Negative Binomial is defined in Equation 9.

[9]

We define as the overdispersion parameter, and we can see that as approaches infinity, the distribution would approach the Poisson, for which the mean and variance are the same. The larger the estimate of , the smaller the estimate of overdispersion and the smaller the estimate of variance. The estimate of is obtained through a log link and with an intercept only overdispersion model.

*Estimation and Optimization*

Estimation can be performed in R with Maximum Likelihood Estimation and using the Laplace Approximation to integrate over the random effects parameters. In R, optimization can be performed using the “BFGS” method, which is a general-purpose optimization technique. It is based on 3 different algorithms including Nelder-Mead, quasi-Newton, and conjugate-gradient (R documentation for General-Purpose Optimization, accessed 2021). This technique is an iterative approach that is used to solve nonlinear optimization problems that is very flexible and robust (Kelley, 2003). In addition, these models can be fit in SAS using PROC NLMIXED.

**Introduction to Subject Area: Hate Crimes**

On the federal level, a hate crime is defined as “a crime motivated by bias against race, color, religion, national origin, sexual orientation, gender, gender identity, or disability” and they are differentiated from a bias or hate incident which are defined as “acts of prejudice that are not crimes and do not involve violence, threats, or property damage”. This study concentrates on the incidents that rise to the level of a true crime which “is often a violent crime, such as assault, murder, arson, vandalism, or threats to commit such crimes”. The United States Department of Justice (U.S. DOJ) notes that “Hate crimes have a broader effect than most other kinds of crime. Hate crime victims include not only the crime’s immediate target but also others like them. Hate crimes affect families, communities, and at times, the entire nation.” This ripple effect highlights the extreme power that hate crimes hold in society today and underscores the importance of investigating their prevalence and the nature of the biases that motivate them. (*Learn About Hate Crimes*, 2021).

Hate crimes in the US operate in very distinctive ways in comparison to other types of crimes, and it is of great importance to investigate the trends of hate crimes. In general, rates initially declined until 2014, then increased following this year (Uniform Crime Reporting Program (UCR) Publications, 2020). A report published by the Center for the Study of Hate and Extremism in 2018 highlighted that there was a 12.5% increase in hate crimes from 2016 to 2017, and that this was the fourth consecutive rise in hate crimes (Levin & Reitzel, 2018). Interestingly, this study noted that even as crime in general slightly decreased throughout this time, hate crimes increased. As more societal attention is being concentrated on the occurrence and prevalence of hate crimes, we are learning more about their broad reaching impact and the role that violence motivated by bias plays in the lives of so many.

The federal government has developed a definition of hate crime that has changed several times, and the reporting structure is relatively new. On April 23rd, 1990 Congress passed the Hate Crime Statistics Act which required data collection “about crimes that manifest evidence of prejudice based on race, religion, sexual orientation, or ethnicity”. In 1994 this act was updated to include bias crime based on disability. In 1996, hate crime statistics were made a permanent addition to the Uniform Crime Reporting Program (UCR), which is where the current data is obtained from. In successive years new types of biases were added including gender, gender identity, more subcategories for religion (including anti-Arab) and several others. This historical context of how the federal government defines and treats hate crimes is important to understanding how the data in this project were collected. (United States Department of Justice, 2021).

A key element of this study is the context of underreporting. According to a special report released by the Bureau of Justice Statistics in the U.S. Department of Justice estimated that 54% of hate crime victimizations were never reported to police during 2011-15 (Bureau of Justice Statistics (BJS), 2017). In fact, they estimate that an average of 250,000 hate crime victimizations took place each year from 2004 to 2015, which is substantially less than the Federal Bureau of Investigation (FBI) UCR reports. The FBI never reports over 10,000 hate crimes per year in this study period. The data collected in the special report are primarily from the Bureau of Justice Statistics’ National Crime Victimization Survey (NCVS) which is a nationally representative sample that consists of interviews with about 240,000 people in the US. The data are fundamentally different than the FBI UCR datasets used in this project as they are crimes reported by those who experience them to an interviewer, rather than crimes reported from police departments to the FBI. Using the FBI data, we will investigate if there are observable trends of under-reporting by police departments. Therefore, the context that we are only modeling reported crimes is very important.

Despite the many limitations of the UCR database, for the purposes of this study, the UCR database is utilized for several reasons. First, it is the only nationally available dataset certified by the FBI. Second, this dataset is nationally recognized and utilized to understand reported hate crime trends. Third, this dataset can potentially be used to understand a certain subset of reported hate crimes; specifically how hate crimes are viewed from a law enforcement perspective. And lastly, this dataset can potentially be used to identify trends of underreporting.

**CHAPTER II**

**METHODS**

**Data compilation and cleaning**

Data were collected and cleaned from two primary sources including the FBI UCR database for Hate Crimes and the U.S. Census Bureau. The FBI UCR database is updated once a year and includes the counts of hate crimes reported from different types of locations in the US and the type of bias that motivated the hate crime. The database includes counts at the state level, county level, institutional level (i.e. police departments, universities), and city level. To maintain direct comparability, crime counts on the city level were chosen and all other sources were excluded. There were 6 categories of bias for 2012-2019, including anti-race, anti-religion, anti-sexual orientation, anti-disability, anti-gender, and anti-gender identity. From 2005-2012, there were only 4 categories used, and they did not include the two gender-based categories. The primary outcome was total sum of counts, therefore all counts were maintained for a city regardless of type of bias that motivated the crime. Of great importance is the fact that these are the number of hate crimes that are reported to the FBI, and not the number of hate crimes that occurred.

The U.S. Census Bureau provides decennial counts along with intercensal estimates for the population of each city in the U.S. and U.S. territories. For the purposes of this project, we selected cities that have a population of 50,000 or larger in 2019, and there were 786 cities that met these criteria. All 15 fifteen years of census estimates were concatenated and data were checked for consistency by looking for outliers and large variation within a city. There were 7 cities that the census did not report estimates for the years of 2005-2009, and for these 7 cities we employed a first observation carried backward technique and imputed the estimate from 2010 for these 5 years. Consistency was checked for these years as well.

Table

Description automatically generated with medium confidenceThe U.S. Census defines 4 large regions in the U.S., (Northeast, Midwest, South, and West) with 9 smaller regions in total: refer to Table 1 for full description of each region. Both region classifications were then applied to the census data. For each location the data included latitude and longitude, city name, city state, city region (smaller), and city region (larger).

Table 1: Locations in the U.S. based on Census delineations

The FBI UCR count data were then paired with intercensal population estimates and location variables for each of the 786 cities. When a city did not report crime to the FBI for a certain year, zeros were imputed. There were 44 cities who never reported a crime to the FBI during the 15 study years and had a population of >50,000 in 2019, and zeros were imputed for each city-year combination as well. Therefore, we have 15 years of data for 786 cities, resulting in 11,790 city-year combinations.

The outcome was parameterized in three different ways; sum of all counts per city per year, population-adjusted rate based on these counts, and the log of the population-adjusted rate + 1. In the following sections, rate will be used to describe the population adjusted rate per 10,000 unless otherwise specified.

**Descriptive Statistics**

Descriptive statistics including means, standard deviations, ranges, and frequencies were calculated for both study year and region for each of the two of outcomes; count and rate. Smoothed plots of the average count and mean per year were constructed to assess the observable trend across time for each and to assess how these trajectories differed once accounting for population change. These smoothed plots were created using locally estimated scatterplot smoothing (loess) which is a non-parametric local polynomial regression technique for smoothing. For region, boxplots were created to illustrate the distribution of rates between regions. In addition, spaghetti plots of rates and counts were created per region to assess observable trends here as well. The spaghetti plots can also be used to visualize outliers and cities with large variation. Based on the rate, the eight cities with the largest standard deviation were identified and plotted to assess these trends visually and with means.

The distribution of the outcomes was first assessed graphically using histograms to plot the distribution of counts and rates. Histograms that include the zero counts and do not include the zero counts were constructed to compare the relative change in distribution with the potential zero-inflation.

**Statistical Modeling**

For the final model, several elements needed to be explored individually; the parameterization of time, additional fixed effects, the form and correlation structure of random effects, zero-inflation, overdispersion, and the distribution to apply.

First, we assessed which generalized linear mixed model distribution would be the best fit through intercept only models. As discussed previously, there are 4 types of generalized linear mixed models we considered: Poisson, Zero-Inflated Poisson (ZIP), Negative Binomial and Zero-Inflated Negative Binomial (ZINB). The intercept only models had no random effects, and we compared models with and without a population offset. Therefore, we compared 8 different models using Akaike information criterion (AIC). Based on the results of these fits, we were able to compare which distribution fit the unadjusted data the best.

The models fit with mixture distributions for the ZINB used the same parameterizations and equations outlined in 6 – 9. The utility of mixture models was compared to non-mixture models, and the integration of parameters for zero-inflation through mixture models were assessed.

Clustering is an integral element of the analysis and was investigated in many different ways. Spatial and temporal clustering were both explored. An intercept only spatial model was fit with Kriging with a normalizing transformation performed on the outcome. This model was fit by defining the spatial exponential structure on the error covariance structure. The estimated parameter allows correlation to be defined based on the Euclidian distance between the latitude and longitude. The spatial elements of these data were also addressed through the introduction of fixed effects, which are discussed below.

Clustering within a city was addressed in two ways. The first was with a random intercept for each city. This would allow each city to vary from the population mean at the beginning of the study. This random effect was included as there was a large amount of heterogeneity on the city level. In addition, the correlation due to repeated measures was assessed through random effects terms including a random intercept and random slopes for time by city. Time in these models was on the year level, and two different parameterizations of time were considered for the random effects; linear time and quadratic time. This would allow the trajectories of the cities to vary both linearly and quadratically from the overall population mean across time. Models with different levels of random effects were compared using AIC and Likelihood Ratio Tests. In addition, three different covariance structures for the random effects were fit and compared: unstructured, heterogenous compound symmetry and variance components. In this case, the heterogenous aspect allows the marginal variances of the compound symmetry model to have *n* additional variance parameters (where *n* is the number of dimensions) (Kristensun & McGillycuddy, 2021).

The primary covariate of interest in this study is time, and several parameterizations of time were explored. Categorical, linear quadratic, orthogonal polynomials (2nd, 3rd, and 4th degree), and B-splines (3rd degree and 4th degree) were considered as candidates for the best fit for time. Each of these different parameterizations of time were assessed first using a random intercept only model that allows the model to control for the within vs. between city variation in the rate of reported hate crimes. The decision for which parameterization to use was based on three criteria: AIC, parsimony of the final model, and interpretability. In addition, the fixed effect parameterization of time was explored in conjunction with the random effects as well. Two models were fit such that the fixed effects and random effects would match; one model that had a linear parameterization of time and one model that had a quadratic parameterization of time.

Additional fixed effects that were explored were region (both with four levels and with nine levels), nested state and region, and political affiliation. Each of these were assessed individually and added to the chosen parameterization of year to assess increase in model fit, convergence, and interpretation. The three fixed effects for location allowed the model to control for some amount of spatial correlation, and also allowed us to make comparisons between locations in the rate of reported hate crimes.

Several models were explored but are not included here for parsimony and as they did not improve model fit or did not converge including interaction models and models with higher levels of polynomials and B-splines.

Once the parameterization of the outcome, covariates, and random effects was chosen based on the above criteria, the final model was fit using a zero-inflated Negative Binomial model. This model was then compared to the three other types of models explored at the beginning to compare the impact of allowing for overdispersion (Poisson) and zero-inflation (Negative Binomial).

**Final Model Specifications**

The final model has 3 fixed effects parameters and 3 random effects parameters on the city level. We are employing an intercept only dispersion model (i.e. where is the overdispersion parameter) and an intercept only zero-inflation model (i.e. ) where is the probability of an excess zero). The parameter specifications are identified in Equations 10 and 11. For Equation 10, is the set of region parameters where *m* takes values from 1-9 where one level is set to the reference (Level 1 in R, by default), and is defined as year within the study.

[10]

[11]

In this case, **G** is specified with an unstructured variance-covariance structure, and is defined in Equation 12, with the variances on the diagonal and the covariances on the off-diagonal. Using an unstructured method yields the most flexible covariance structure for random effects that is possible.

**Sensitivity Analysis**

[12]

For the 44 cities that never reported a crime during the entire study period, we performed a sensitivity analysis for the final model where we removed these cities entirely. We then compared the estimates obtained by the final model and the model fit using the subset data. Differences in beta estimates, confidence intervals, dispersion parameters, zero-inflation estimates, and random effects were compared between the two models.

**Extensions of Final Model**

To assess the model performance of the final model, both marginal and conditional pseudo R2 values were calculated to estimate the amount of variability captured by the fixed effects alone in comparison to the combination of random and fixed effects. These pseudo R2 measures are Nakagawa’s R2 measures that were developed for mixed models (Nakagawa & Schielzeth, 2012).

Cities were grouped by their temporal trajectories and compared. Predicted values were calculated for each city at each time point based on the contribution of both fixed and random effects. The population-level predictions were also calculated and were included on plots of city-level perditions for comparison.

**Grouping of cities by Random Effects and Predicted Values**

Given the high degree of heterogeneity across the cities and the flexibility in the model fit to capture various types of temporal trends, cities were grouped into intuitive trajectories. This allowed for a more in-depth exploration of how the differences between cities was expressed across time as temporal profiles. Given the parameterization of time as both a linear term and a quadratic term, there were nine discernable trends: increasing linear, flat linear, decreasing linear, increasing concave up, flat concave up, decreasing concave up, increasing concave down, flat concave down, and decreasing concave down. Cities were grouped using two possible sources of data; random effects mean and significance estimates, and the fitted values. Both of these sources were city specific, and the random effects estimated the strength and direction of the difference from the population average whereas the fitted values estimated city-specific values across time.

There are several cluster-based modeling techniques, such as Latent Class Growth Curves and Mixture Growth Models, that can also be used to cluster. Some of these methods are highlighted in the work of Mäkikangas, Bakker, and Aunola (2010) and Kim, Kang, and Morrow (2014). Whereas Mäkikangas et al. (2010) based their clustering on a latent growth curve analysis and Kim et al. (2014) utilized linear growth parameters within the model. In contrast, the current study used model-based predictions and random effects estimates in order to identify clusters of longitudinal trends. The groupings here are based on natural features of the fitted values (i.e. concave up, concave down, increasing, decreasing etc.) This project took inspiration from other clustering techniques but ultimately utilized results from the ZINB model to perform an informal analysis and to group cities.

Several methods were used, and they are outlined at length here. The conditional means of the random effects for each city were estimated from the final fitted model along with the estimated 95% confidence intervals using the ranef function in the lme4 package in R. Each city therefore had three means and confidence intervals estimated – one for each random effect. Five methods were used to isolate unique patterns of the predicted values based on the random effects estimates. First, the direction of the mean estimate for each random effect was considered individually. Next, the direction of both linear and quadratic year estimates was considered in conjunction with each other. Another method was using the significance of a random effect mean estimate; only cities with a significantly different trajectory from the population average were included in a group. This created very large groups for the non-significant estimates. Quantiles for each random effect estimate were calculated, and the 20th and 80th quantiles were used to group cities. Lastly, more arbitrary cutoffs were used based on the effect on the overall mean trajectory of that group and on the visualization of the predicted values.

The predicted values themselves were also used to identify trends. The difference between the 2019 fitted estimate and the 2005 fitted estimate were calculated for each city and depending on the value of this difference a city was classified as increasing, decreasing, or flat. Predicted values with a large variance were identified in order to capture quadratic trends.

The methods chosen to group the nine trajectories were evaluated based on the visual estimated smoothed mean lines for each group. The final method chosen to identify the three possible linear trends was to use the difference between the 2019 and 2005 predicted values to identify if a city was increasing, decreasing, or remaining relatively flat. The final method chosen to identify the three possible quadratic trends was based on the combination of random effect mean estimate for the quadratic year term. The three linear and three quadratic trends were combined to group cities into a total of nine possible trajectories.

Once cities were categorized into the nine possible trajectories, spaghetti plots were created for each group that included individual lines for each city based on the predicted values from the model, a smoothed mean curve for the group, and the population level smoothed mean curve. Descriptive statistics were calculated for each group for rates, counts, population size, and frequencies based on region. Welch’s two-sample t-tests were performed to assess if there was a significant difference in the average rate and average population size between each of the nine groups and all other cities not in that group.

The counts and overall proportions of cities within regions stratified by grouping are presented in a table. This table also includes the overall proportions of city counts within regions along with a color indicating if the subgroup had a larger or smaller proportion than the average reporting. Fisher’s Exact Test for count data with simulated p-values (2,000 replications) was used due to the small counts per category. These tests were used to assess if there was an association between the proportion of cities within regions based on the trajectory grouping. These tests of association were performed between one group and the remaining groups (i.e. group 1 vs all other groups). Maps of the U.S. were constructed based on the latitude and longitude of each city and were color categorized based on the trajectory grouping of each city in order to perform a visual spatial comparison.

**Comparing Expected and Observed Counts and Rates**

Predicted values based on both fixed and random effects were calculated on the response scale and were compared to the observed outcome. The sum of square difference for each city was calculated across all years. The ten cities with the largest sum of square differences were identified and both observed and predicted lines were plotted together. The sum of square difference was also calculated stratified by year and region. The differences were plotted to highlight the years and regions where the difference was the greatest and when it was the lowest. In addition, line plots were created by region that show the negative and positive differences in expected values. For each region, the negative values and the positive values are both summed across all cities and then plotted on the same graph. These highlight the trends within regions and across years for smaller or larger expected values in relation to the observed values. The spaghetti plots of predicted trends by regions were superimposed on the observed spaghetti plots to highlight the smoothing performed by the model and to compare the estimates from the fitted values with those of the raw data.

All analyses were performed using R version 4.1.0 and SAS version 9.4 (kriging).

**CHAPTER III**

**RESULTS**

**Descriptive Statistics**

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedFor all study years, there was a large proportion of cities that did not report any crimes for any year with an average of 44.0% of cities not reporting across all years. The year with the highest number of cities reporting to the UCR was 2008 with 61.3% of cities reporting at least one crime. The year with the lowest number of cities reporting to the UCR was 2014 with only 50.9% of cities reporting at least one crime. 2019 had the largest mean count (4.72; SD = 23.6) and the largest range with a maximum of 410 crimes. The year with the highest average rate is 2006 (0.29; SD = 0.5) and the year with the lowest average rate is 2014 (0.17; SD = 0.3). Both averages counts and average rates vary across the 15 study years. Refer to Table 1 in the Appendix for all summary statistics by year.

Figure 4: Loess plots of average counts across time

The loess smoothed plot in Figure 4 show a smoothed mean line for average count across time and Figure 5 shows a smoothed loess curve for average rate across time. A non-linear unadjusted trend is illustrated, and we can see the impact of adjusting for population at risk for crimes in comparison with modelling raw counts. For average count by year, a roughly cubic trend with a low in 2014 and a high in 2019 is apparent. When we adjust for population, we can see a steadily decreasing linear trend from 2005 to 2014, followed by a slight increase. This curve could also be described as quadratic with an inflection point at 2014. For either outcome, these loess smoothed curves suggest a continuous parameterization of time with possible polynomials or splines.

Figure 5: Loess plots of average rates across time

Chart, box and whisker chart

Description automatically generatedThe region with the most cities was the Pacific region with 215 applicable cities. The region with the least cities was the East South Central region with only 31 applicable cities followed closely by the Middle Atlantic Region with 39 cities. The Middle Atlantic region had the highest mean count (10.31; SD = 48.2) and the West South Central region had the lowest mean count (1.79; SD = 4.3). The region with the highest mean rate per 10,000 was New England (0.41; SD = 0.6) and the region with the lowest mean rate was West South Central (0.1; SD = 0.2). We can see that in all cases of counts or rate, the standard deviation is large in comparison to the mean. The Middle Atlantic region has the largest range in count with a maximum of 410 counts and East North Central has the largest range in rate with a maximum of 6.78 per 10,000. Refer to Table 2 in the Appendix to see all means, standard deviations, and ranges for regions.

Figure 6: Boxplots of population adjusted rate stratified by region

Based on the boxplots in Figure 6 of rate by region, we can see the right-skewed, non-normal distribution of these data along with the large outliers. The variability across regions for rate of reported hate crimes is evident. Based on the spaghetti plots of observed rate (refer to Figure 1 in the Appendix), we can see that the variability extends to the city level as well. While many cities are clustered at the bottom of the graphs, the spaghetti plots emphasize the outliers for both cities and years. Different regions have different patterns of outliers, and for some regions it is difficult to discern an outlier pattern.

The 10 cities identified that had the largest standard deviation in rate were Taylor City (MI), Rapid City (SD), Flint (MI), Tigard (OR), Seattle (WA), Eugene (OR), Bloomington (IN), Revere City (MA), New Brunswick (NJ), and Brockton City (MA). Both Taylor City and Rapid City had a count of zero in in 2005, the highest count in 2006 (44 and 42 respectively) and then low counts for the remainder of the years. Seattle has a continuously increasing rate across the years, with the highest rate in 2019. Many of the other cities have multiple changes in direction, and it is evident that there is a high level of variability not only in the base rates but in the trajectory of rates across time at the city level. Refer to Figure 2 in the Appendix for plots of the outlying cities.

Overall, 44.0% of the total data was a zero count. Of the 11,970 possible city-year combinations, 5,188 of these had a zero count. Of the 786 cities in total, 44 of these never reported a crime during the 15 years. These cities make up approximately 5.6% of the zeros, and once they are removed, we still have approximately 38.4% zeros in the data. Looking at the histograms for both count and rate, we can see highly right-skewed data, along with a large count of zeros (refer to Figure 3 in the Appendix for counts and Figure 4 in the Appendix for rates). In addition, the mean of sums across all cities and years is 4.23, whereas the median is 1. For strict counts, we can see a steady decrease in frequency as we increase the count. For population adjusted rates, we can first see an increase, followed by a decrease for many of the years. We have less spread when using rates, but we still have a large number of zeros.

**Statistical Modeling**

Table

Description automatically generated The first set of models that were intercept only models with a random intercept for city added were used to explore which of the four proposed distributions fit the best. These models were all run with and without an offset. The primary differences between the models without an offset (the count models) and the models that included an offset (rate models) was the estimate of the intercept and the AIC. In the count models, the intercept is the estimate of the average counts, whereas in the rate models the intercept is the estimate of the average rate. Refer to Table 2 for all intercept-only model results.

For each of the two distributions and the two mixture distributions, the models that incorporated the offset all had a better fit as defined by a lower AIC, and smaller estimates of the random intercept. This makes intuitive sense as more variability is captured by the offset and therefore less was captured by the random effect. The models that included the population offset all had a significantly better fit (all Likelihood Ratio Test (LRT) for all comparisons; p < 0.0001).

Table 2: Intercept only generalized linear mixed model results with a Random Intercept

Clearly, the model with the largest AIC was the simple Poisson model with a value of 53,359. In addition, it is evident that we need to account for overdispersion as the Negative Binomial model greatly improves the AIC to 42,330. Allowing the Poisson to be zero-inflated (ZIP model) improved the AIC significantly to 49,550. Most importantly, the ZINB model has the best AIC (42,245) and this model estimates several useful parameters such as the overdispersion parameter and the zero-inflation estimate. Interestingly, the ZINB estimates slightly less overdispersion than the simple Negative Binomial estimates. However, the zero-inflated model is a significantly better fit than the model that does not account for zero-inflation (LRT; p < 0.0001), and the ZINB is a significantly better fit than the ZIP (LRT; p < 0.0001). While the models that do include a population offset have a lower AIC, these models demonstrate different and important information about the counts themselves. Based on the importance of considering the population at risk in each city, the models that include a population offset are emphasized. We can conclude that a ZINB with a natural log offset for population per city per year is the best fit for the data.

The exploration of the parameterization of time included 9 different parameterizations of year. Refer toTable 3 to see the AIC estimates and the significance of the year fixed effects from each of the random intercept model fits for different parameterizations of time. Each of these models were fit with only the parameterization of time in the fixed effects. The only model that did not converge was the cubic year model. We can see that in general the 4th degree Spline, the 4th degree orthogonal polynomials, and categorical year are the best fit based on AIC. This makes sense as these are the most complex models (with categorical being the saturated model) and the majority of the terms are highly significant. However, the quadratic year parameterization was Table

Description automatically generatedfairly close in AIC and offers a more interpretable beta coefficient. For these reasons, the quadratic form of year was chosen as the parameterization of slope.

Various fixed effects were explored, including four fixed effects for location. The first of these was large region (split into 4), region (split into 9), State (split into 48), and State within region. Models with state as a fixed effect would no longer converge when adding in additional random effects to the random intercept, or other fixed effects. For the two levels of regional covariates, the model with more specific regional breakdowns had a better AIC with a value of 42,151 in comparison to 42,164. The Likelihood Ratio Test (LRT) returned a p-value of 0.0003, which tells us this model improvement is significant. Therefore, the final model included a fixed effect for region with nine levels. In addition, this fixed effect offered useful interpretations about the comparison of rates between different regions.

Table 3: Parameterization of time results

The final parameterization of the random effects was chosen based on two criteria. The first of these was to have interpretable trajectories for cities. Therefore, if there was a linear parameterization of time in the fixed effects, there would be a linear parameterization of time in the random effects (similarly with quadratic). The Table

Description automatically generatedsecond was based on model fit as assessed through AIC. The best fitting model was time as quadratic in both the fixed and random effects and therefore the final model included a random intercept, a random slope for linear time, and a random slope for quadratic time (AIC = 40,907). Refer to Table 4 to see the exploration of random effects. Three different covariance structures were tested including unstructured, compound symmetric and variance components. The variance components model did not converge, and of the two remaining models the unstructured covariance structure had a better AIC (40,907 in comparison with 41,065). The resulting model allowed for the most flexibility along with the most interpretable estimates.

Table 4: Random effects exploration results

Table

Description automatically generatedAfter the model selection process was completed, we again fit all 4 distributions to compare the final fit between the Poisson, Negative Binomial, ZIP and the ZINB. In addition, the ZINB was fit with and without an offset to compare the count and rate models. The ZINB model had the lowest AIC (40,907) and the Poisson had the highest AIC (45,168). Refer to Table 5 for all AIC values and non-fixed effect parameter estimates from final models. Similar to the results for the intercept only distribution models, we can see that each added element for overdispersion and zero-inflation improved the model fit significantly. When we correctly account for zero-inflation through a mixture model we can see that there is a better model fit (LRT; p <0.0001). When we correctly account for the overdispersion by allowing the variance to have a quadratic relationship with the mean we have a significantly better fit (LRT; p <0.0001). Lastly, we can see that including a population offset that varies by year we increase the model fit as the model with no offset has an AIC of 41,294 (LRT; p <0.0001).

Table 5: Final model comparisons between distributions

**Interpretation of Final Model Parameters**

Table

Description automatically generated All parameter estimates are reported in Table 6. The beta estimate for the zero-inflation portion of the model was highly significant (p < 0.0001) and the value of -3.06 indicates that the probability of an excess zero is approximately 4.5%. While this is relatively low, the AIC of the ZINB is significantly higher than the NB which indicates that the mixture model that incorporates the probability of the excess zeros is a better fit to the data. The overdispersion parameter was estimated to be 4.55 which indicates that while overdispersion is present and we must account for it, the spread of the data was not very large after controlling for all other parameters.

Table 6: ZINB final model results

The combination of the two beta coefficients for year in the fitted quadratic model suggests an average decrease of approximately 3.7% per year (but with some flattening of the decrease over time). The decrease across time becomes less across the study period, and there is a very slight concave up pattern.

The fixed effects for region have East North Central as the reference group. The South Atlantic region and the West South-Central regions both have significantly lower rates of reported hate crimes in comparison to East North Central. New England and the Pacific Region both have significantly higher rates of reported hate crimes.

The pseudo Marginal R2 value was 0.037 which indicates that approximately 3.7% of the variance in the outcome can be explained by the Fixed Effects. The pseudo Conditional R2 was 0.29 which indicates that the full model explains approximately 29% of the variance in the outcome. This means that approximately 25.3% of the variance can be explained by the random effects alone. Much more of the explained variability is captured by the random effects in comparison to the fixed effects.

**Sensitivity Analysis**

The sensitivity analysis that compared the full dataset to the dataset where the 44 cities who never reported a crime returned fairly similar results. Indeed, the only estimate that changed was the beta estimate for the Pacific region in comparison to the East North Central region; the increased rate was no longer significant. The Intercept estimate was slightly lower in the subsetted dataset (-10.794 vs. -11.083) and the random intercept estimate is also lower (2.08 vs. 2.41). The estimates for the zero-inflation parameter and the overdispersion parameter remained almost exactly the same. We can conclude that the 44 cities who never reported a crime during the study period did not significantly influence the results of the final model. Refer to Table 4 in the Appendix for all results from the sensitivity analysis.

**Groupings of Cities by Trajectories**

Table

Description automatically generated To further investigate the heterogeneity across cities, predicted values were grouped into various types of temporal trends. Nine possible trajectories were identified based on the parameterization of time (the nine combinations of linear and quadratic curves). Cities were successfully clustered into these trajectories and we were able to identify distinct differences in the temporal trends between cities. The primary finding of this analysis is that the city level heterogeneity can be clustered into meaningful trends across time that differ from each other. These trends allow for more interpretable city level differences based on random effects; we can categorize cities by simple trends across time based on fitted values from the model. Depending on the research question, different trends can be utilized. For example, if we want to identify cities who will likely have an increase in reported hate crimes in the following year, we can investigate the linearly increasing and concave up cities.

There were 232 cities with an increasing trend across time, 183 cities with a rate that was approximately the same at the beginning and the end, and 371 cities with a decreasing trend across time. Of the 232 cities with an increasing trend, 75 had a concave flat pattern, 72 had a concave up pattern, and 85 had a concave down pattern. Of the 183 cities that ended at approximately the same beginning and end value, 102 were concave flat, 46 were concave up, and 35 were concave down. Of the 371 cities that were decreasing across time, 140 were concave flat, 112 were concave up, and 119 were concave down (refer to Table 7 for counts by grouping). This means that the biggest group of cities followed a linearly decreasing pattern, and the smallest group of cities followed a pattern where the beginning and ending value were approximately the same but the rate dipped below in the middle of the study period. The various patterns are visualized in the Figure 7, and we can compare the overall population curve (dark blue line) with the averaged smooth line per group (black line).

Table 7: Sample size per trajectory grouping

Chart, diagram

Description automatically generatedDiagram

Description automatically generatedDiagram

Description automatically generated

Figure 7: Fitted trajectories grouped into the 9 trajectories, including a population curve (black) and a subgroup curve (blue)

Table

Description automatically generatedVarious tests of association were fit in order to compare variables across the 9 clusters. First, we tested to see if there were significant differences between the various trajectory groupings and the rate and population averages. The Welch’s two sample t-tests show clear differences in the average rate of reported hate crimes and the average population size in the subgroups of the cities in comparison to all remaining cities (refer to Table 8). The differences in population size highlights how larger or smaller cities may in general be grouped into different trajectories, and the differences in rates illustrates the relationship between average rate and trajectory. The three groups of cities that did not have a significantly different average rate in comparison to the rest of the cities were the flat/concave up group, the flat/concave down group, and the decreasing/concave down group. Otherwise, there were significant differences in the average rate between certain trajectories. This indicates that there are some underlying differences in the rates and population at risk that can potentially be used to distinguish certain groupings.

Table 8: Tests of association for continuous variables by trajectory group

Table

Description automatically generated

Table 9: Tests of Association for region by trajectory grouping along with descriptive statistics

When stratifying the groups by region, we can see that there are significant differences in the distributon of cities across regions for three trajectory groups. This finding highlights the potential regional (and potentially cultural) differences that are present in some groups. The decreasing/flat concave group, the increasing/concave up group, and the increasing/concave down groups all have significantly different proportions of counts in comparison to all other groups (refer to Table 9). The percentages in Table 9 are reported as row percentages. For these three trends, we can see that there may be association between the physical location and the type of trend. For example, the increasing/concave up group can be further analyzed by looking at how different regions may have different patterns across time. The biggest difference for this group is a much larger proportion of cities in the New England Region and in the East South Central region. There may be an interaction between region and type of trajectory across time.

Diagram

Description automatically generated The maps of the U.S. (shown in Figures 8 and 9) with groupings of trajectories are more translatable when using linear trends only or quadratic trends only instead of using all 9 categories. Interestingly, it is difficult to visually identify strict spatial trends based on the temporal trends. This finding makes sense in the context of the previous spatial analysis that found a spatial correlation of practically zero based on latitude and longitude. It seems like there are slightly more cities with a decreasing trends along the coasts, along with a clustering of flat cities around the great lakes, and more flat or increasing cities in the south and southeast. For the concave trends, we can also see a clustering of flat cities around the great lakes, and more flat or concave up cities around the Mississippi river and in the south. However, we can also see that these trends do not hold across all locations and in many areas of the U.S. the trends are indistinguishable. This further highlights the spatial heterogeneity and potential Diagram

Description automatically generatedlack of grouping within states or regions based on on temporal trajectories.

Figure 8: Map with latitude and longitude points color coded based on linear temporal trajectory

Figure 9: Map with latitude and longitude points color coded based on quadratic temporal trajectory

**Expected vs. Observed**

The ten cities with the largest sum of square difference between the expected rate and the observed rate were Eugene (OR), Boston (MA), Seattle (WA), Taylor City (MI), Santa Cruz (CA), New Brunswick (NJ), Bowling Green (KY), Washington D. C., Hoboken (NJ), and Buffalo (NY). Four of these cities (Eugene, Seattle, Taylor city, and New Brunswick) also were identified in the top ten cities with the highest raw variance in rates.

The sum of square difference between expected and observed rates was also calculated by year and region. The year with the greatest sum of square difference was 2005 and this was due to the large difference in the Middle Atlantic region. Upon further investigation, the largest difference in expected and observed counts was due to New York City. In 2005 NYC reported 0 hate crimes, followed by 274 in 2006 and reaching 410 in 2019.Chart, line chart

Description automatically generated

Figure 10: Difference in expected and observed Counts by year and region

For Figure 10 NYC was removed. Once NYC was removed, we can see that the Pacific Region had the largest peaks in 2008 and in 2017. The Middle Atlantic Region and the West South Central region tended to have the lowest variability. The regional variability in expected vs. observed counts is distinguishable across time. Figure 10 includes the directionality of the difference between expected and observed counts illustrates the large heterogeneity within regions across years. A striking element of this figure is the outliers for the pacific region. This is likely due to the fact that the Pacific region has the highest number of cities in comparison to the other regions, and many of these cities are quite large. This difference in underlying population for the pacific region is likely inflating the cumulative difference between expected and observed which is why we see stronger negative peaks in the Pacific region.

On the region level, we can see that in general there are larger negative peaks than there are positive peaks which indicates more outliers for larger observed values in comparison to expected values. The large overdispersion and outliers in the raw data are likely influencing this trend. Figure 10 indicates that there are more city year combinations that have more observed counts than expected counts based on the fitted model. This finding highlights the impact of the excess zeros and low reporting as the city-level averages are biased towards low counts and are not effective at predicting the true higher counts in the data.

Graphical user interface

Description automatically generatedGraphical user interface, chart

Description automatically generatedGraphical user interface, application

Description automatically generatedWhen looking at Figure 11, which shows both the expected and observed curves by region, we can see the smoothing of the raw counts that is performed by the model fit. The blue fitted lines follow the possible linear or quadratic curves discussed in the trajectory groupings section. The years and cities that have the highest counts also have the greatest difference as the overall average is pulling the smoothed expected curve down. This further illustrates what we see in the sum of differences in Figure 10. We can see the impact of the low counts in biasing the fitted curves to be lower in general in comparison to the actual higher counts. This bias toward low counts is visually identifiable based on the clustering on the blue lines below many of the true counts represented by the green lines in Figure 11.

Figure 11: Observed (green) and expected (blue) rates for each city across time and stratified by region

**CHAPTER IV**

**DISCUSSION**

**Model-Based Conclusions**

All of our findings must be within the context that we cannot disentangle the reporting of hate crimes from the true counts of hate crimes from this data. Hate crime data in general is quite messy and requires complex analytical tools along with complex conceptual understanding. In addition, the changing definitions, laws, cultural elements, and number of law enforcement agenciues who report at all have fluctuated and changed throughout the study period. The recognition, classification, and reporting of hate crimes all have many different social, psychological, and bias based elements. This means that it is probable that not only the rates themselves are dynamic, but the amount of underreporting is as well. Centering this complicated interpretation of the outcome is incredibly imporant when conceptualizing the results and grounding the current findings in other data sources.

Keeping in mind this interpretation complexity, the final ZINB model offers several interesting insights into the reported hate crimes in the U.S. from 2004 to 2019. The slight linear decrease over time cannot be interpreted as a strict decrease in the true occurrence of hate crimes in this country. Indeed, other data sources indicate that there is either an overall increase or a non-significant increase in the rate of hate crimes across time (BJS, 2017). Based on the low amount of variability captured by the fixed effects in comparison to the high variability captured by the random effects, we argue that this slight linear decrease does not give us the full picture.

Based on the large Conditional R2 value in comparison to the Marginal R2 value, we know that a large portion of the variance in the rate of reported hate crimes is attributable to the city level in comparison with the national level. This model can be used to identify either areas that need to increase their reporting practices or areas that have higher rates than average and therefore need more funding for law enforcement and victims. Future studies and interventions can therefore be targeted more locally rather than nationally. In addition, future studies can investigate additional fixed effects that can quantify the causes for these differences.

Based on the fitted model, we must remember that the fit is using averages, which are of limited application. We know from other sources of data that there is a large amount of underreporting in the UCR database. However we can assert that there is little to no overreporting in this dataset. Based on the stringent requirements for a crime to be considered bias-based, and the fact that the source of the data is law enforcement, it is reasonable to assume that most if not all of the counts are true counts. This means that the fitted values will be biased towards the underreporting, rather than an average of the true counts. This further highlights the importance of addressing underreporting, both from a community perspective and from a law enforcement perspective. It also underscores the lack of the quality of the data in the FBI UCR database for drawing conclusions about the true nature of hate crimes.

When assessing the grouped trajectories, it may be of interest to concentrate more heavily on the cities that had an increasing trajectory or a concave up pattern. Both of these may indicate an increase in the next year for reported hate crimes. It would also be helpful to see if the cities that had an overall decreasing trend had changed any law enforcement practices or had changing populations. It may be that single outlying years can cause a concave pattern and investigating these years to see if they were truly outliers in terms of counts could give some insight into what causes a flare up of hate crimes. The utility of grouping cities by their trajectories should not be understated, as these groupings highlight distinct differences between cities that was not captured by fixed effect variables. Treating rates as dynamic processes that follow different trajectories allows us to not only model differences between locations, but also to isolate patterns that can indicate where future research should concentrate.

The tests of association between different grouped trajectories indicate that further investigation is needed into the underlying root causes of the differences. Noticeably, the cities differ from other cities based on the group that they are in, and that the trajectory groupings may have unique features that define them. The differences in average rates and population sizes indicate that there are likely other confounding factors (such a demographics of the cities, political/social factors etc.) that may also be associated with the type of trend. The differences in regional distribution for the three categories that have a significant difference indicate that the unique features of these trajectory groupings may also have a spatial component. All of these results together illustrate that there are underlying differences in cities that are associated with a number of unique trajectories across time in the rates of reported hate crimes.

We can use the mappings of trajectories to their spatial location as a tool to identify outlying cities based on their trajectory grouping. For example, if the majority of cities within an area have a decreasing trend, and one city has an increasing trend, we can identify this city as an outler and further investigate that city. In addition, we can investigate areas that tend to have a more identifiable trend to compare them to areas that lack an identifiable trend. The combination of visual spatial analysis with temporal profiles allows the user to not only identify outliers and trends across space, but also to ground the temporal groupings within space.

The comparison between expected and observed counts highlights the impact that the underreporting, or low counts, in the data has on the model. There is a lot of variability in the data that is not estimated exactly by the model, and we can see that these estimates are pulled more towards the low counts. This can also be used as a tool to identify cities, year, and regions that have a large difference in what we would expect to see based on the model and what actually occurs. This can help to isolate years where we have a greater than expected increase in reported hate crimes which could indicate either a spike in crimes, or underreporting in the other years. It could be interesting to compare the difference between expected and observed in other datasets to see how the estimates change.

This data and subsequent models highlight the importance of secondary analysis; both through the lens of the federal government and through the lens of community led efforts. The ZINB model highlights certain problems, and the outlier cities that should be further investigated. For example, were there truly zero hate crimes in New York City in 2005? Are the peaks in certain cities more representative of underreporting in other years or did something happen that sparked an excess of hate crimes? The differences in expected and observed counts can be utilized to isolate cities that diverge from the fitted predictions. The cities with very low counts should be investigated further to assess if these are true low counts or low reporting. If they are true low counts, the subpopulations can be studied to understand why the risk of hate crimes is so much lower.

**Barriers to reporting**

A major barrier to reporting hate crimes to law enforcement is the relationship between law enforcement and the citizens that they serve. The Bureau of Justice Statistics released a report in 2017 based on the NCVS data that indicated that 54% of those who were victimized by a hate crime do not report this victimization to the police. Individuals cited several reasons including they handled the crime through a different resource, they did not believe that the police wanted to get involved, they believed the police would be ineffective, or they believed the police would further harm the victim (BJS, 2017).

The amount of voluntary participation from law enforcement agencies in reporting to the UCR has grown since the inception of the Hate Crime Statistics Act. At the beginning, 2,771 agencies reported data, and by 2012 14,500 (approximately 75% of police agencies in the US) were reporting data (Anti-Defamation League, 2013). In a report published by the Bureau of Justice Statistics, using the NCVS, Wilson (2014) estimates that 293,800 violent and property hate crime victimizations occurred in the U.S. in 2012. Of these, an estimated 60% were not reported to police. However, this still leaves a large gap in the estimated crimes that were reported to police, and the 5,796 hate crimes reported to the FBI UCR that same year (FBI UCR). Wilson (2014) reports that in the time period 2004-2012 there was an average of 269,000 bias-motivated victimizations, and an average of 105,890 victims claimed to have reported the incident to the police. However, in this same time-period, the UCR reports an average of only 8,770 hate crimes. Pezzella et al. (2019) note that while the number of reporting law enforcement agencies has drastically increased since the inception of the program, the counts reported have only marginally increased. In addition, they note that as the definition of hate crime has broadened to include new groups (i.e. sexual orientation, anti-Islam, anti-Asian etc.) the resulting counts have not increased as would be expected. Importantly, Wilson (2014) also noted that the percentage of hate crimes involving violence increased from 78% in 2004 to 90% in 2011 and 2012.

While some methodological discrepancies between the UCR and the NCVS can explain the discrepancies, there is still a clear gap between the incidents of victimization and the incidents that the UCR reports yearly.In a report titled “Improving the Identification, investigation, and Reporting of Hate Crimes” published by the U.S. DOJ Hate Crimes Enforcement and Prevention Initiative (2020), suggest many reasons for underreporting to the UCR including lack of training for law enforcement, strained police-community relations, inter-departmental culture, and legal complications (non-uniform laws, difficulty in proving bias etc.). Nolan and Akiyama (1999) performed a survey and factor analysis to identify important factors that affect the participation of law enforcement in hate crime reporting. They identified 10 factors on the department level and ten factors on the individual level that affected underreporting, some of which included hate crimes not being deemed important, perceptions that no problem exists, insufficient infrastructure/time/support, and many others (Nolan & Akiyama, 1999). Nolan and Akiyama include that the most commonly cited reason among LGBTQ+ individuals to not report a crime to law enforcement is the risk of secondary victimization. Secondary victimization is defined as insensitivity and abuse by police. Their conclusions support the overall barriers on the part of civilians and law enforcement on creating and sustaining accurate reporting and data.

Pezzella, Fetzer, and Keller (2019) offer two potential hypotheses as to the underlying mechanisms of underreporting. Using stepwise logistic regression, and controlling for potential confounders, they found that victim perception of police legitimacy and an absence of confidence in law enforcements ability to discharge justice largely explains underreporting (Pezzella, Fetzer, & Keller, 2019). Indeed, Pezzella et al. highlight the history of strained relations between communities most at risk for hate crimes and law enforcement. Particularly, the history of police brutality against those in the African American community and the LGBTQ+ community is important to note as these are the two groups most targeted by hate crimes. In addition, a report published by the Bureau of Justice Assistance notes that further barriers exist including language, citizenship status, and cultural expectations (1999). The conclusions of the multiple papers cited here support the knowledge that there are many barriers on the part of civilians and law enforcement to creating and sustaining accurate reporting and data.

In a summary report prepared by the U.S. DOJ Community Oriented Policing Services (COPS), data and arguments are presented that address this issue specifically (2020). The three recommendations that the COPS report (2020) presents are “1. Develop comprehensive training for law enforcement on identifying and reporting hate crimes”, “2. Support law enforcement efforts to develop strong community bonds through systemic hate crime education and outreach”, and “3. Reward innovative, effective practices to improve law enforcement identification and reporting of hate crimes”. The summary report includes many great resources and recommendations that concentrate on the relationship between communities (particularly communities at risk for hate crimes) and the law enforcement that serve them. Interventions targeted at this level could have a manifold impact; not only would there be more hate crimes reported to law enforcement and to the FBI UCR database, but there would likely be interventions that could decrease the number of true hate crimes. The final model here indicates that these local level interventions would allow the interventions to be targetted to local factors and that more data are needed in order to more properly identify these factors.

**Type of Bias and Crime**

In a release by the DOJ Office of Justice Programs, they noted that while there was a non-significant increase in the total number of hate crime sin 2012 (293,800) in comparison with 2004 (281,700), the types of biases changed significantly (DOJ BJS, 2014). In 2012, 51% of hate crime were motivated by “ethnicity bias”, whereas only 22% were “ethnicity bias” motivated in 2004 (DOJ BJS, 2014). The occurrence of religious bias tripled, and the occurrence of gender bias doubled (DOJ BJS, 2014). The NCVS found that nearly 90% of all hate crimes in 2011 and 2012 were violent, and that this was a growing trend (DOJ BJS, 2014). Based on this evidence, the type of bias that motivates the crime must be considered in future studies along with the overall rates.

**Limitations**

The primary limitation of this study is the quality and quantity of the data. As discussed in great length in the discussion, the conclusions that we can draw are limited to the domain of voluntarily reported hate crimes from law enforcement to the FBI. Therefore, the applications of this data through the ZINB models are limited in their functionality and in their application. Further limitations of this is that all count and population estimates were performed in aggregate, rather than identifying types of bias and demographic differences in population. This means that the interpretation of our “population at risk” used in the offset is not necessarily accurate. For example, cities with different demographic characteristics are likely to experience different rates and types of bias-motivated crimes. A working paper summary by Cikara, Fouka, and Tabellini (2021) summarizes their current research by saying that as a proportion of a minoritized group increases, so too does the likelihood that members of that group will be targeted by hate crimes. Therefore, understanding the migration of populations within the U.S. and understanding how the changing demographics influence those who are at risk is important to include. The aggregated population and bias-types proves to be a limitation in this project.

In addition, this sudy does not incorporate a measure of uncertainty for each hate crime. It is reasonable to assume that the amount of overreporting in this dataset is minimal, based on the origins in law enforcement. However, data that is closely tied to many social and psychological latent variables should have some level of uncertainty in the outcome. This may be especially true for other data sources, such as media reports and self-reported data. The combination of many data sources may help us to estimate the level of ucnertainty in each data source.

**Future Directions**

While this project only includes intercept only models for both the overdispersion parameter and the zero-inflation parameter, these models can be extended to include covariates. This allows for the overdispersion and zero-inflation to depend on levels of a variable. Equations 13-15 are based on Brooks et al (2017). The zero-inflation estimate uses a logit link which is shown in Equation 13, an the dispersion estimate uses a log link. The zero-inflated intercept as estimated by the model when back transformed estimates the overall probability of zero-inflation. First, define equation 13:

[13]

Where an SZ is a structural zero (or an excess zero). The extension to this model allows overdispersion and zero-inflation to depend on covariates and can be expressed in Equations 14 and 15.

[15]

[14]

In addition, there are many interesting covariates that could be included such as demographics by city, local and state laws surrounding hate crimes, law enforcement biases, community/law enforcement trust, political affiliation and more. All of these variables are attempting to explain the discrepancies between the different data sets and the city-level heterogeneity. In addition, this data is an excellent candidate for latent variable analysis. Many of the factors that likely have a big influence on this data are difficult to measure, and therefore a Latent Variable Analysis could provide a large amount of insight into the trends of hate crimes and the groupings of data that are influential.

It is of great importance to consider more than one type of data source in order to gather a fuller picture of the true nature of hate crimes. It is reasonable to assume that each source of data will have certain biases and restrictions (UCR data, law enforcement data, national surveys, community level data, media data etc.). A study by Vergani et al. (2020) used Chile as a case study and attempted to combine various sources of data on hate cirmes. A major conclusion that the authors draw is the importance of disaggregating the data as the conclusions often differ depending on the source of the data. In comparison with other sources of data in the U.S. (NCVS, local organizations, media etc.) the UCR database offers a fairly restrictive view on the true nature of hate crime trends.

These data illuminate the need for more investigation on a local level. Resources need to be allocated to not only responding to the hate crimes that are being reported, but at improving the reporting structures and the relationship between law enforcement and those that they serve. If we target interventions only at the national level, the model would indicate that less resources are needed (i.e. slight linearly decreasing trend). However, if we look at the individual cities, we will better be able to target cities that have either increasing rates, low levels of reporting, or inconsistent data. Through the combination of many data sources and flexible modeling techniques, we can better understand the true nature of hate crime rates in the United States. Armed with this knowledge, we can endeavor to decrease the rate and impact of this specific kind of crime that has such far reaching effects.

REFERENCES

1. Cameron, C., Trivedi, P., 1998. Regression Analysis of Count Data. Cambridge University Press, New York.
2. Fang, R., Wagner, B., & University of Colorado Denver Anschutz Medical Campus, degree granting institution. (2013). *Zero-inflated Negative Binomial (ZINB) regression model for over-dispersed count data with excess zeros and repeated measures, an application to human microbiota sequence data*.
3. Ver Hoef, J. M, & Boveng, P. L. (2007). Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data? *Ecology (Durham),* *88*(11), 2766-2772.
4. Greene, W. (2008). Functional forms for the Negative Binomial model for count data. *Economics Letters,* *99*(3), 585-590.
5. Bouguila, N., & Fan, W. (2020). *Mixture models and applications*. Springer.
6. Verbeke, G., & Molenberghs, G. (2011). *Linear mixed models for longitudinal data*. Springer.
7. Lord, D., Washington, S. P, & Ivan, J. N. (2005). Poisson, Poisson-Gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis and Prevention,* *37*(1), 35-46.
8. Rider, P.R., (1961). Estimating the Parameters of Mixed Poisson, Binomial and Weibull Distributions by Method of Moments. Bulletin de l'Institut International de Statistiques 38, Part 2.
9. Cohen, A.C., (1963). Estimation in Mixtures of Discrete Distributions. In Proceedings of the International Symposium on Discrete Distributions, Montreal, Quebec.
10. *Zero-Inflated Negative Binomial Regression*. NCSS Statistical Software. (n.d.). https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Negative\_Binomial\_Regression.pdf.
11. Brooks, M., E, Kristensen, K., Benthem, K,J., Magnusson, A., Berg, C., W., Nielsen, A., . . . Bolker, B.,M. (2017). GlmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. The R Journal, 9(2), 378.
12. Hardin, Hilbe, & Hilbe, Joseph M. (2007). Generalized linear models and extensions (2nd ed.). College Station, Tex.: Stata Press.
13. R: General-purpose Optimization. (n.d.). https://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html.
14. Kelley, C. T. (2003). *Iterative methods for optimization*. SIAM, Soc. for Industrial and Applied Math.
15. The United States Department of Justice. (2021, May 3). *Learn About Hate Crimes*. https://www.justice.gov/hatecrimes/learn-about-hate-crimes.
16. FBI. (2018, September 10). *UCR Publications*. FBI. https://www.fbi.gov/services/cjis/ucr/publications#Hate-Crime%20Statistics.
17. Levin, B., Reitzel J. D. (2018). *Report to the Nation: Hate crime rise in U.S. cities and counties in time of division and foreign interference*. Center for Victim Research. https://csbs.csusb.edu/sites/csusb\_csbs/files/2018%20Hate%20Final%20Report%205-14.pdf  
    <http://hdl.handle.net/20.500.11990/975>
18. The United States Department of Justice. (2021, July 14). Facts and Statistics. https://www.justice.gov/hatecrimes/facts-and-statistics.
19. Bureau of Justice Statistics, Masucci, M., & Langton, L. (2017, June). *Hate Crime Victimization, 2004–2015*. US Department of Justice. <https://www.bjs.gov/content/pub/pdf/hcv0415.pdf>
20. *ADL: 2012 Hate Crime Statistics Report "Seriously Flawed"*. Anti-Defamation League. (n.d.). https://www.adl.org/news/press-releases/adl-2012-hate-crime-statistics-report-seriously-flawed.
21. Pezzella, F. S., Fetzer, M. D., & Keller, T. (2019). The Dark Figure of Hate Crime Underreporting. *American Behavioral Scientist*, 000276421882384. https://doi.org/10.1177/0002764218823844
22. Bureau pf Justice Statistics, Wilson, M. (2014). Hate Crime Victimization 2004-2012 – Statistical Tables. US Department of Justice. <https://bjs.ojp.gov/content/pub/pdf/hcv0412st.pdf>
23. U.S. DOJ Hate Crimes Enforcement and Prevention Initiative (2020). Improving the identification, investigation, and reporting of hate crimes; A summary report of the law enforcement roundtable. <https://cops.usdoj.gov/RIC/Publications/cops-w0895-pub.pdf>
24. U.S. DOJ; Bureau of Justice Assistance (1999). A policymaker’s guide to hate crimes. <https://www.ojp.gov/pdffiles1/bja/162304.pdf>
25. Ghojogh, B., Ghojogh, A., Crowley, M., and Karray, F. (2020). Fitting a mixture distribution to data: tutorial. arXiv preprint arXiv:1901.06708, 2019a.
26. Vergani, M., Navarro, C., Freilich, J. D., & Chermak, S. M. (2021). Comparing Different Sources of Data to Examine Trends of Hate Crime in Absence of Official Registers. *American Journal of Criminal Justice,* *46*(3), 445-460.
27. DOJ BJS, McCarthy, K., (2014). U.S. residents experienced about 293,800 hate crime victimizations in 2012 – Unchanged from 2004. <https://www.ojp.gov/sites/g/files/xyckuh241/files/archives/pressreleases/2014/ojppr022014.pdf>
28. Cikara, M., Fouka, V., Tabllini, M. (2021). Hate crime increases with minoritized group rank, *Harvard Business School Working Paper*.
29. Hu, Mei-Chen, Pavlicova, Martina, & Nunes, Edward V. (2011). Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. *The American Journal of Drug and Alcohol Abuse,* *37*(5), 367-375.
30. Youngdeok Kim, Minsoo Kang, & Morrow, James R., Jr. (2014). Longitudinal Trajectories Of Step-Count Measures In Women Using A Latent Class Growth Curve Modeling. *Medicine and Science in Sports and Exercise,* *46*(5), 788.
31. Mäkikangas, Anne, Bakker, Arnold B, Aunola, Kaisa, & Demerouti, Evangelia. (2010). Job resources and flow at work: Modelling the relationship via latent growth curve and mixture model methodology. *Journal of Occupational and Organizational Psychology,* *83*(3), 795-814.
32. Kristensen, K., & McGillycuddy, M. (2021, July 13). *Covariance structures with glmmTMB*. Covariance structures with glmmtmb. https://cran.r-project.org/web/packages/glmmTMB/vignettes/covstruct.html.
33. Nakagawa, S. and Schielzeth, H. (2013), A general and simple method for obtaining *R*2 from generalized linear mixed-effects models. Methods Ecol Evol, 4: 133-142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

APPENDIX

Table

Description automatically generated

Table 1: Descriptive Statistics by year

Table

Description automatically generated

Table 2: Descriptive Statistics by Region

Chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, line chart

Description automatically generated

Figure 1: Observed Spaghetti plots by region

Chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Figure 2: 10 cities with the largest standard deviation in raw rate

Diagram, engineering drawing

Description automatically generatedGraphical user interface, diagram

Description automatically generated

Figure 4: Histogram of rates

Figure 3: Histogram of counts

Table

Description automatically generatedTable

Description automatically generated

Table 3: Sensitivity Analysis Findings