Analysis of Crime in Pittsburgh

Aaditya Ramdas
MLD, CMU
adidas@cmu.edu

Abstract

Through the course of human civilization, law and order has been a constantly pressing issue, dealt by monarchs and democracies in varying ways. In this project, we look at predictability of yearly totals of crime (concentrating on burglaries) in neighborhoods of Pittsburgh, given census data detailing characteristics of ethnicity, education, employment, population, transport, environment (natural and built) for every neighborhood. One hopes that this may be of use to social scientists, criminologists, law enforcement agencies and city planners.

1 Introduction

1.1 The City of Steel

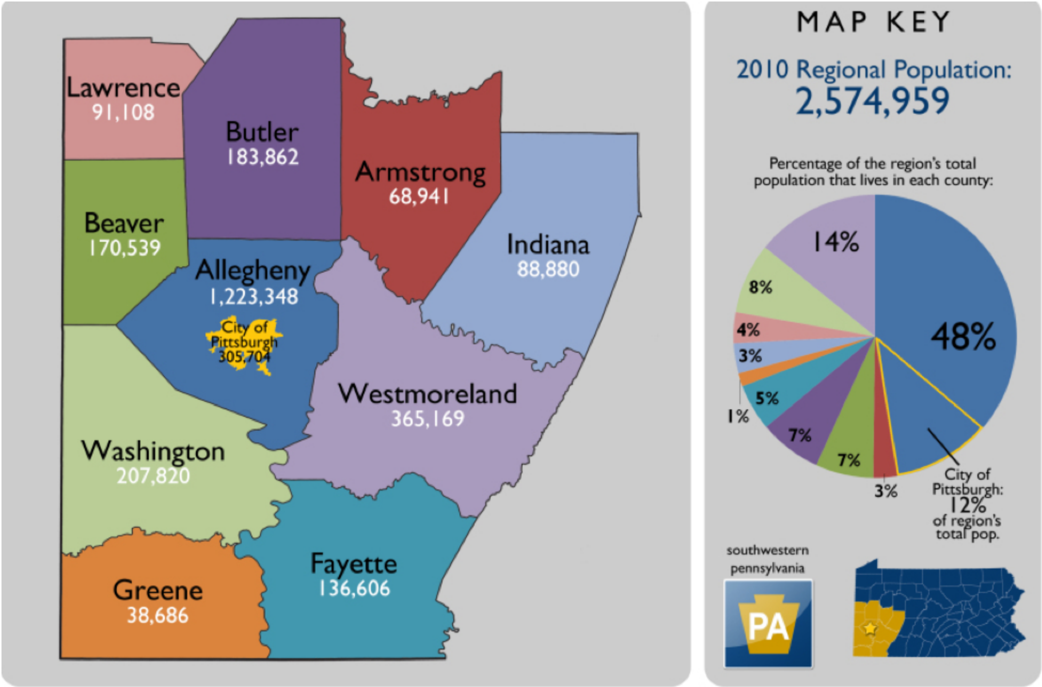


Figure 1: A map of Pittsburgh's region (left) and the region's population split among counties (right)

Pittsburgh is the biggest city and economic center of the Allegheny County in western Pennsylvania, being home to 305,704 (down from 676,805 in 1950) of the county's 1,223,348 inhabitants.

It differs quite drastically from the county on many aspects, some of which are predictable because of the large number of universities, but others may seem less intuitive.

For example, Pittsburgh has a larger percentage of residents with postgraduate degrees compared to Allegheny county (12.4% to 10.6%), but also a much higher level of poverty (21.7% to 12.6%). It has a much younger population between 5 and 34 years of age (47% to 37.6%) and almost double the percentage of blacks in its population (27.2% to 13.2%), as well as asians, hispanics, and foreign-borns. It has a higher percentage of renter-occupied units (47.9% to 35.3%) but also a higher number of vacant units (12.8% to 9.4%).

2 Dataset Source and Description

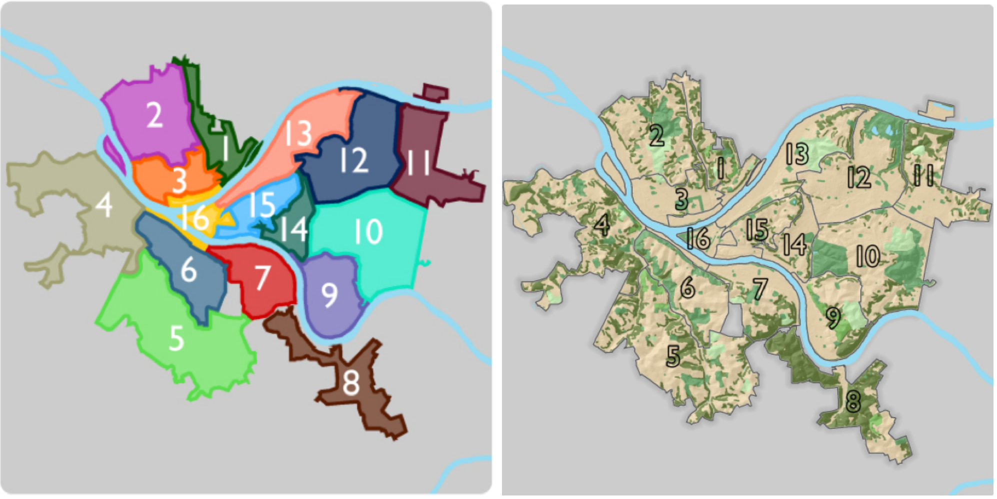


Figure 2: The 16 sectors of Pittsburgh (left), and the city's physical land use (right).

The raw census data from 2010 was collected, maintained and published by the City Government of Pittsburgh at the link *http://www.pittsburghpa.gov/dcp/snap/*.

Pittsburgh is divided into 16 sectors, which are then further subdivided into a total of 90 neighborhoods, each of which has its individual identity (for example, CMU is located in South Oakland). These neighborhoods are of vastly different sizes (the largest is 25 times the smallest) and populations with 5 of them having under 300, and 6 of them having over 10,000 (for example, I live in the largest one, South Squirrel Hill, whose population is 15,110).

For each of these neighborhoods, detailed information was collected regarding the ethnicity, education and employment details of its occupants, the median income and house value, the division of land space (both natural and built), the modes of transportation used, the amount of crime, etc. We shall describe these (as were available in raw form) here:

- Population - Decade-wise number of residents from 1940 - 2010, detailed breakup by ethnicity, division into broad age groups.
- Housing - Total number of units, occupied vs vacant units, renter vs owner occupied units, break up decade of construction, median home value, foreclosures and sales.
- Employment - Total number of employed residents and breakup by type of job, number of jobs located in neighborhood and breakup by type of job.
- Education+Income - Breakup of population based on levels of education attained, median income in 2000 and 2010, population under poverty.
- Crime - Major, minor and other crime reports, number of murders, rapes, aggravated assaults, robberies, burglaries, auto-thefts and drug violations.

- Land Use - Total area, breakup into residential, commercial, industrial, institutional, open spaces, hillside and special uses.
- Built Environment - Number of parcels and structures and breakup according to their condition, residential and commercial building permits, code violations, condemned structures, demolitions and tax delinquent property.
- Natural Environment - Percentage of landslide-prone, undermined and flood plains, breakup into park space, woodland, greenways, cemeteries and street tree counts.
- Transport - Miles of major roads and total street miles, number of steps and treads, breakup by mode of commute to work into driving alone, carpool, public transport, taxi, motorcycle, bicycle, walk, other and work from home.

2.1 Further Descriptions of Data

We can find the detailed descriptions of each of the above covariates in [3]. We include a few of the confusing ones here for the sake of clarity and example.

Part 1 (Major Crime) Reports : This category lists the actual number of Part 1 crimes, which are incidents that include violent and property crimes. Aggravated assault, forcible rape, murder, and robbery are classified as violent while arson, burglary, larceny-theft, and motor vehicle theft are classified as property crimes–all are included in Part 1 crimes. Part 1 crimes are collectively known as Index crimes; this name is used because the crimes are considered quite serious, tend to be reported more reliably than others, and are reported directly to the police and not to a separate agency.

Part 2 Crime Reports : This category lists the number of Part 2 crimes. In Part 2, the following categories are tracked: simple assault, curfew offenses and loitering, embezzlement, forgery and counterfeiting, disorderly conduct, driving under the influence, drug offenses, fraud, gambling, liquor offenses, offenses against the family, prostitution, forgery, misdemeanors, public drunkenness, runaways, sex offenses, stolen property, vandalism, vagrancy, and weapons offenses. These directly affect the quality of life of residents and communities.

Approx. Num. Parcels : The approximate number of distinct tax parcels within a neighborhood (parcels are mapped by Allegheny County). This number is approximate because some parcels cross neighborhood boundaries–the parcel is included in the neighborhood where its centerpoint is located as calculated by GIS mapping.

Approx. Num. Unoccupied Parcels : This calculation is made based on parcels that do not contain buildings as modeled by GIS.

Approx. Num. of Structures : The approximate number of structures that exist, regardless of parcels. The estimate is made using GIS information about structures and their locations.

Num. Condemned Buildings : Total number of buildings which have been condemned as uninhabitable and eligible for demolition by inspectors.

Tax Delinquency : The number of properties that are delinquent on property taxes for more than two years. Also includes the percentage of a neighborhood's properties that meet this status.

Undermined Area : Percentage of neighborhood's land area that potentially has coal mine tunnels/shafts underneath the surface.

Num. Street Trees : Approximate number of street trees located in a neighborhood. Trees counted are those located along a public right of way.

Num. Sets of Steps : The total number of sets of public steps in each neighborhood, regardless of the length of the stairway. Most of these sets of steps are treated as public streets, and have names as such.

2.2 Dataset Cleaning and Altering

The neighborhood of Chateau had a population of 11 (down from 8267 in 1940) and a lot of missing data, and hence was deleted from the dataset. There were many columns which were exact sums of other columns (for example, total working population was split into several categories, total population was split into different ethnicities, total area was represented in square miles and acres, etc) and some of these were removed.

From the perspective of performing linear regression, many values represented as percentages could be misleading. For example, if the crime in one large neighborhood is twice a smaller one, but the percentage area used in park spaces remains the same, it might mislead the regression into thinking that crime and area used for park spaces are uncorrelated (while actually the park space area is indeed larger for the bigger neighborhood and lower for the smaller one, just as crime is). Hence, we convert all percentage values into numerical counts.

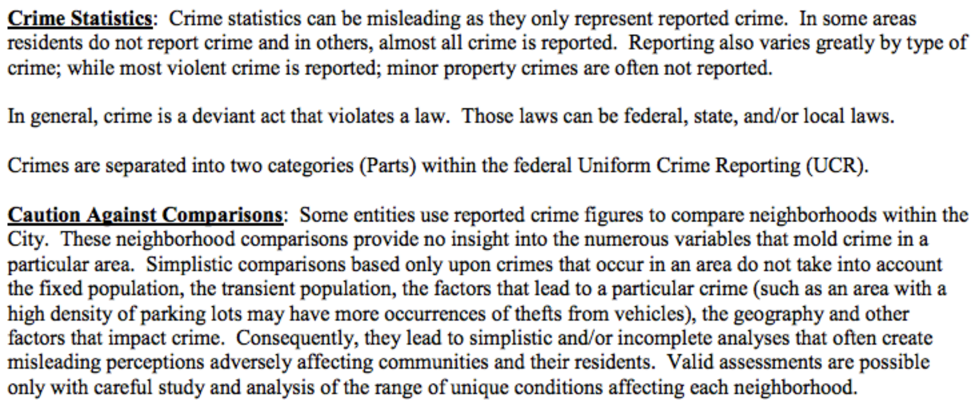
3 Analysing Crime Datasets

It is important to realize that a large number of useful covariates are missing. For example,

- Number and locations of 911 calls
- Number and locations of speeding car violations
- Drunk drivers charged after breath tests
- Abandoned vehicles
- New graffiti on walls
- Number of parking lots
- Illegal parking and towed vehicles
- Number of illegal firearm arrests.
- Number of bus stops
- Miles of bike lanes and trails

Of course, some of these will be naturally correlated in time and space with crimes (like 911 calls) and others could be predictive (large parking lots may have more break-ins). Also, the number of crime incidents is different from the number of accused involved in the incident, and that might also affect prediction. Furthermore, the ones that exist are highly correlated, and hence even if prediction works well, variable selection is hard and probably not robust.

This skepticism can be outlined in the following warning from the Police Annual Report:



Crime Statistics: Crime statistics can be misleading as they only represent reported crime. In some areas residents do not report crime and in others, almost all crime is reported. Reporting also varies greatly by type of crime; while most violent crime is reported; minor property crimes are often not reported.

In general, crime is a deviant act that violates a law. Those laws can be federal, state, and/or local laws.

Crimes are separated into two categories (Parts) within the federal Uniform Crime Reporting (UCR).

Caution Against Comparisons: Some entities use reported crime figures to compare neighborhoods within the City. These neighborhood comparisons provide no insight into the numerous variables that mold crime in a particular area. Simplistic comparisons based only upon crimes that occur in an area do not take into account the fixed population, the transient population, the factors that lead to a particular crime (such as an area with a high density of parking lots may have more occurrences of thefts from vehicles), the geography and other factors that impact crime. Consequently, they lead to simplistic and/or incomplete analyses that often create misleading perceptions adversely affecting communities and their residents. Valid assessments are possible only with careful study and analysis of the range of unique conditions affecting each neighborhood.

Figure 3: A note of warning from [1]

3.1 Outline of Methodology

The future sections will outline our use of the following methods, in the same natural order that the author and analyst decided to try them, and some intuition is also provided as to how or why the next steps were arrived at. As a point of note throughout this report, two-step significance tests have been employed (ie a first step of variable selection followed by a second step of fitting a model to only the selected variables and measuring significance for the second step). While this is not theoretically justified, it is widely used in practice - hence the p-values should be taken with a pinch/bag of salt while the theory of one-step significance tests for sparse models gets developed.

- Brute Force - Lasso : Since we are working in a high-dimensional setting with the number of features p being larger than the number of data points n, it is natural to try to fit a sparse linear model, as done by the Lasso.

- Exploratory Data Analysis (EDA) : A two-step process of variable selection with lasso, followed by a simple linear regression fit does not yield many significant variables, encouraging us to look further into the data. This involves looking at distributions of variables, correlations between covariates and correlations of covariates with the response, possible outliers, etc.

- Log Transformed Lasso : EDA shows us that most of our variables (when transformed to counts instead of fractional percentages) are approximately log-normally distributed, signalling that performing using the Lasso estimator after $\log$-transforming the covariates and response might perform better. However, we find that this is not the case, and hence move on to our next model.

- GLM - Sparse Poisson Regression : $\log$-transformation ignores the fact that our responses are positive integer counts, and hence a generalized linear model like poisson regression (a method with a long and rich history in statistics) with a $\log$-link function might be a wiser choice. We also fit robust versions of these estimators to be more stable with respect to possible outliers.

- Generalized Additive Models : The non-satisfactory results of GLMs could stem from the inherent non-linear relationship between our response variable (crime) and the covariates - additive models are a flexible alternative to try - they have the advantage of generality (less parametric than a linear model) as well as simplicity since they fit univariate functions.

- Sparse Poisson Additive Models : In the previous step, we fit a GAM by first selecting the variables using a GLM model and then solve for an additive model using only the selected variables, and this two-step process is not very sound. The SpAM algorithm, which performed one-step sparse additive model fits, can be generalized to poisson regression to combine both these steps into a single algorithm.

- Negative Binomial : The model assumptions in the previous step were not well satisfied according to the relevant diagnostics. Hence we question some assumptions of poisson regression. As a start, we try to fit an overdispersed poisson or a negative binomial regression, which does not assume that the mean is equal to the variance like the poisson does. The data does look very overdispersed with a much larger variance than mean, but we do not have enough data to get a significant fit.

- ACE/AVAS : The $\log$-link function might be a common choice in the literature, but it may not be right for our setting. ACE and AVAS try to fit an additive model to some other smooth transformation of the response that is learnt from the data. Both of these actually do suggest that the $\log$-link function might be a bit aggressive.

- Mixed-Effect Models : The previous methods did not take any geography into account. Since each data point comes from a neighborhood in a city, it is reasonable to postulate that nearby areas have some shared or common effects. A mixed-effects model allows us to take sector-level interactions into account by adding variables to the linear model that are common to two neighbourhoods if they lie in the same sector.

We now spend time elaborating on the methodology used for each of the above methods, and briefly describe the results using some plots and tables.

4 Brute Force - Lasso

We have more covariates than observations, so we would need a sparse regression routine, and the natural first method of choice is the Lasso. We solve the optimization problem

$$\min_{\beta} \|\boldsymbol{y} - \boldsymbol{X}^{\top}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

where λ is chosen by cross-validation, and the ℓ_1 penalty encourages sparsity and is useful in situations where $p >> n$ (more features than data points). Alternatively, it is the MAP estimate for the following model

$$\begin{aligned} y_i &= \text{Normal}(m_i, 1) \text{ for } i = 1...n \\ m_i &= \boldsymbol{x_i}^{\top}\boldsymbol{\beta} \\ \beta_j &\sim \text{Laplace}(0, \lambda) \text{ for } j = 1...p \\ \text{Note that } \mathbb{E}[y_i|\boldsymbol{x_i}] &= \boldsymbol{x_i}^{\top}\boldsymbol{\beta} \end{aligned}$$

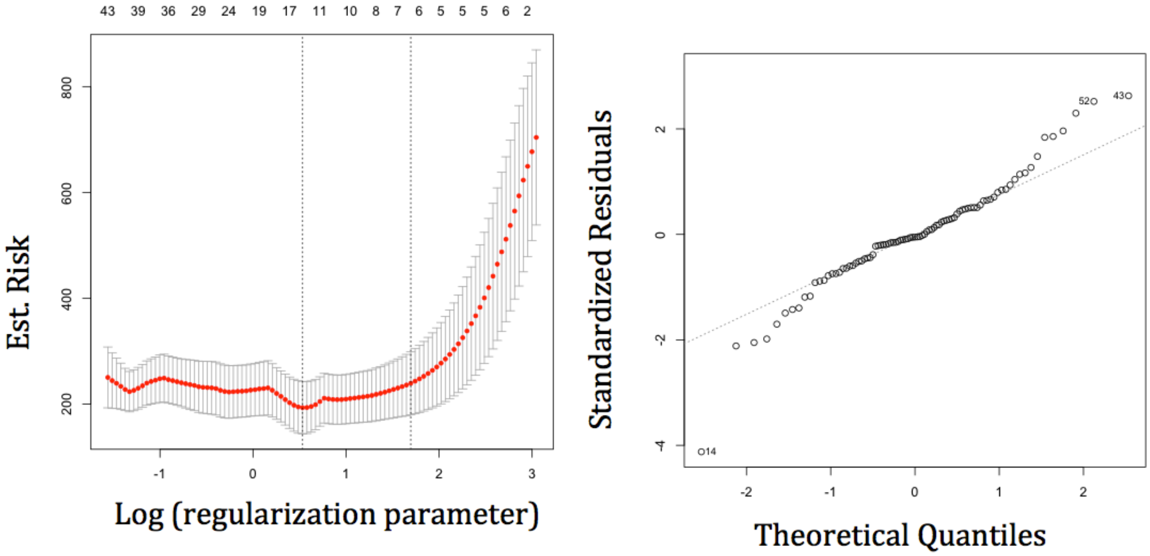


Figure 4: Cross Validation plot, we choose $\lambda = 2.5$ near the kink (left) and residual qq-plot (right)

In practice, while it is possible to solve for the lasso solution at a fixed λ using one of many specialized lasso algorithms like the prox-gradient method with soft-thresholding being the prox-operator, it is preferred to run a path algorithm to solve for the exact lasso solution along a sequence of λ values, warm-starting the search a new regularization paremeter using the solution to the old one. The figure here shows the estimated risk plotted against this sequence.

Choosing $\lambda = 2.5$ near the kink of the cross validation plot leaves us with 10 variables. We then run a normal linear regression using lm and just these variables.

Even though the residuals are nearly normally distributed, none of the predictor variables are very significant - due to the additional bias of first choosing variables and then testing significance, variables tend to appear more significant than they actually are, and seeing such low values is surprising.

We now do some exploratory data analysis, to see if we can come up with a better model for our data.

5 EDA

5.1 Crime Variables (2010)

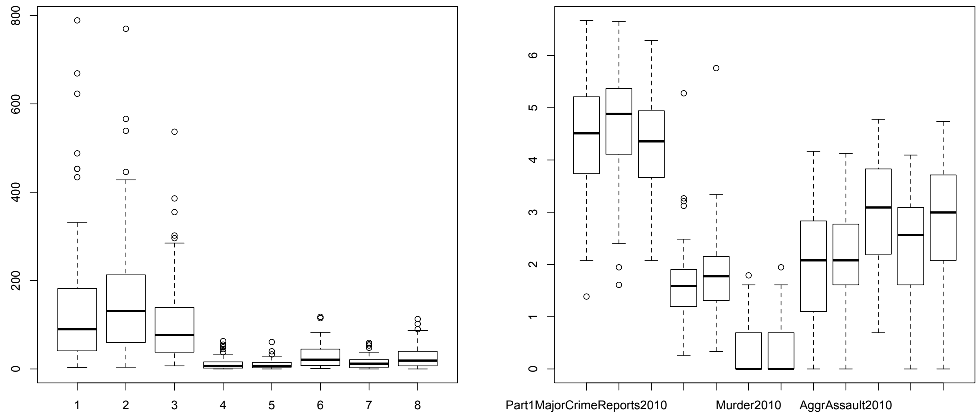


Figure 5: Boxplots for crime variables (left) and log(1+crime) variables (right)

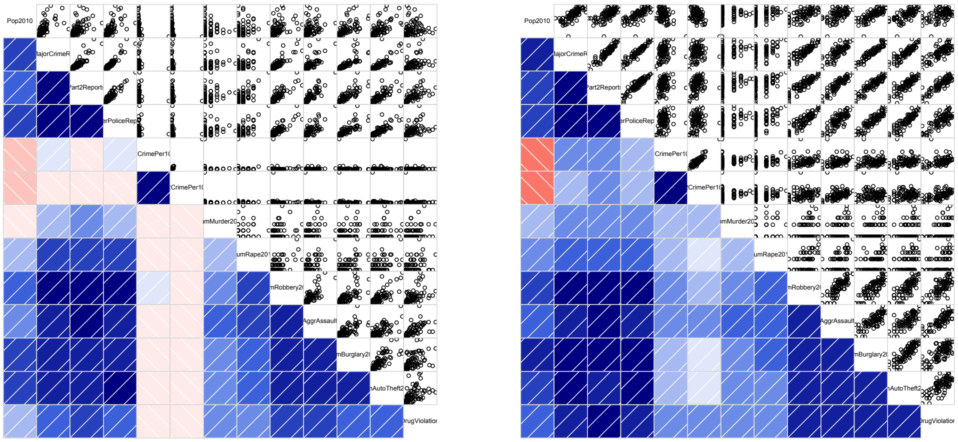


Figure 6: Correlations and scatter plots for crime variables (left) and log(1+crime) variables (right)

The numbers of murders and rapes are extremely low, with a vast majority of neighborhoods having zero occurrences of both, and when it is non-zero, it is always less than 5 - this might make it extremely difficult to predict because they are rare and occur only in exceptional circumstances. The two features that don't correlate well with the others are actually on a different scale, since they are the number of major/minor crime reports *per 100 persons* (and hence may not grow with population or other factors). For this project, we normalize everything to simple counts on the same scale in order to be able to make meaningful predictions (and hence don't try to predict these either).

Hence, the possibly reasonable variables to predict are the number of major crime reports, the number of minor crime reports, other police reports, robberies, aggravated assault, burglary, auto thefts and drug violations.

5.2 Predictor Variables (2010)

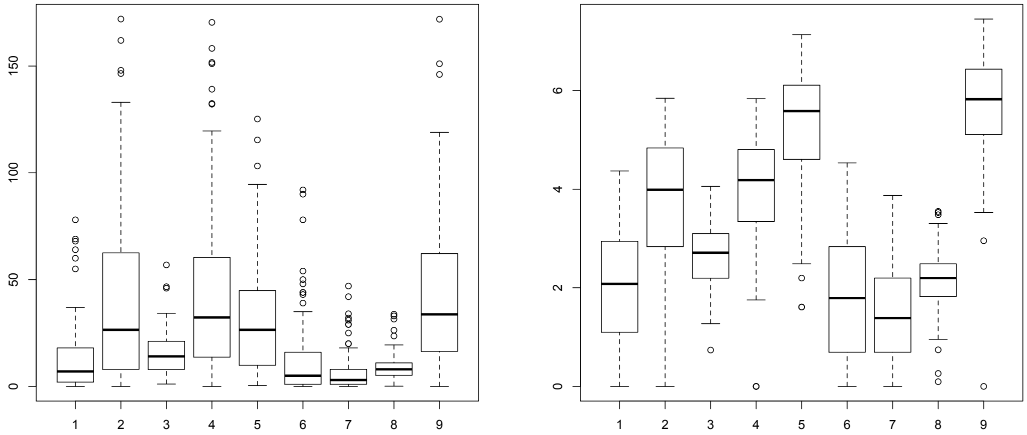


Figure 7: Boxplots for covariate variables (left) and log(1+covariate) variables (right) : Number of foreclosures, code violations, total street miles, population more than two races, unoccupied parcels, condemned structures, demolitions, population per acre and population who aren't high-school graduates

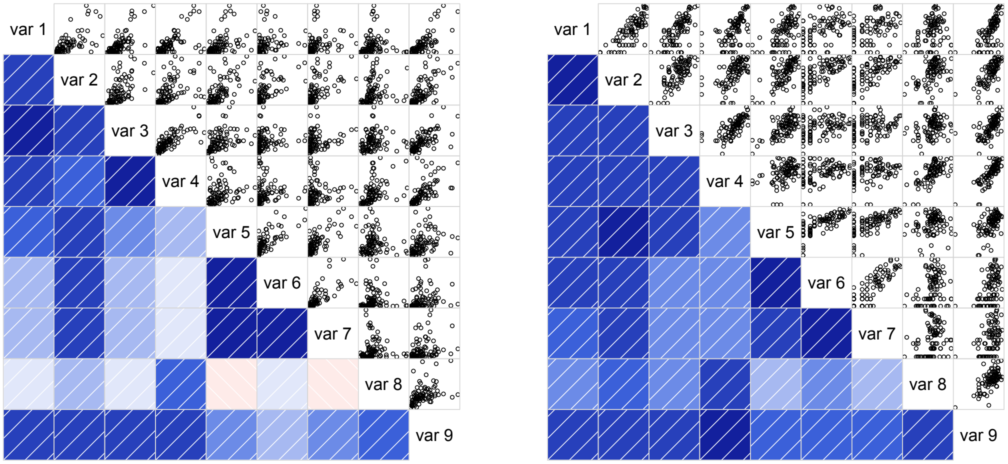


Figure 8: Corrgram for covariate variables (left) and log(1+covariate) variables (right) : Number of foreclosures, code violations, total street miles, population more than two races, unoccupied parcels, condemned structures, demolitions, population per acre and population who aren't high-school graduates

Almost all predictor variables also seem to be normally distributed from their boxplots. For this reason, we shall log-transform our data and proceed. However, many of these covariates are also highly correlated, and hence variable selection might be a tricky process, and this is a situation where sparse estimators are known to sometimes perform poorly.

6 Lasso after Log transformations

Here, we do the natural next step of log-transforming both the crime variables and the covariates, and then applying the lasso for sparsity. The new model is (here $\log$ of a vector is elementwise $\log$):

$$\begin{aligned}
\log(y_i) &= \text{Normal}(m_i, 1) \text{ for } i = 1...n \\
m_i &= \log(\boldsymbol{x_i})^\top \boldsymbol{\beta} \\
\beta_j &\sim \text{Laplace}(0, \lambda) \text{ for } j = 1...p \\
\text{Note that } \mathbb{E}[\log(y_i)|\boldsymbol{x_i}] &= \log(\boldsymbol{x_i})^\top \boldsymbol{\beta}
\end{aligned}$$

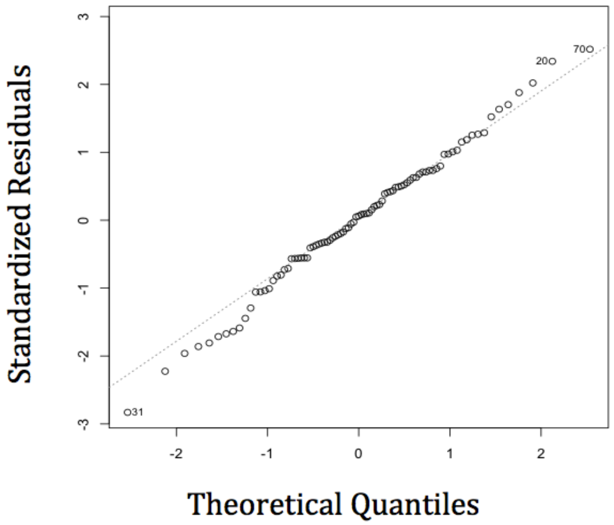


Figure 9: Residual qq-plot for Log-Lasso

The residuals seem to look very close to normally distributed, and so the model assumptions aren't violated for this fit. However, the regression yields very few variables as being significant (summarised in the table below). As mentioned earlier, the two-step regression normally overestimates the significance of variables, due to the added bias of selecting variables before estimating significance.

Covariate Name	Significance ($< 0.001^{***}$, $< 0.01^{**}$)
Jobs in n'hood (retail/trade)	$**$
Intercept	$**$

Table 1: Significant variables as estimated by two-step Log-Lasso

The above table still does not seem satisfactory - the only variable that is picked is an odd one, and the complete absence of other expected variables might make the analyst skeptical. Note that we are presently ignoring the fact that our response is essentially count data (the number of crimes in a neighborhood) and hence might be better modeled using a sparse poisson regression, which is what we shall do next.

7 Sparse Poisson Regression

The canonical link function for the poisson family is $\log$, and the major use of this family is to fit surrogate poisson log-linear models to what is actually multinomial frequency data. Stimulus factors have their marginal totals fixed in advance (or for the purposes of inference). The main interest lies in the conditional probabilities of the response factor given the stimulus factors.

It is well known that the conditional distribution of a set of independent poisson random variables, given their sum, is multinomial with probabilities given by the ratios of the poisson means to their total. This result is applied to the counts for the multi-way response within each combination of stimulus factor levels, allowing models for multinomial data with a multiplicative probability specification to be fitted and tested using poisson log-linear models. The model can be summarised as:

$$\begin{aligned} y_i &= \text{Poisson}(m_i) \text{ for } i = 1...n \\ \log(m_i) &= \log(\boldsymbol{x_i})^\top \boldsymbol{\beta} \\ \beta_j &\sim \text{Laplace}(0, \lambda) \text{ for } j = 1...p \\ \text{Note that } \log(\mathbb{E}[y_i|\boldsymbol{x_i}]) &= \log(\boldsymbol{x_i})^\top \boldsymbol{\beta} \end{aligned}$$

We find the CV value of lambda that gets minimum error using $cv.glmnet$. We regress again using $glmnet$ with this value of lambda, choose the relevant non-zero covatiates, and regress again using glm and only the relevant variables.

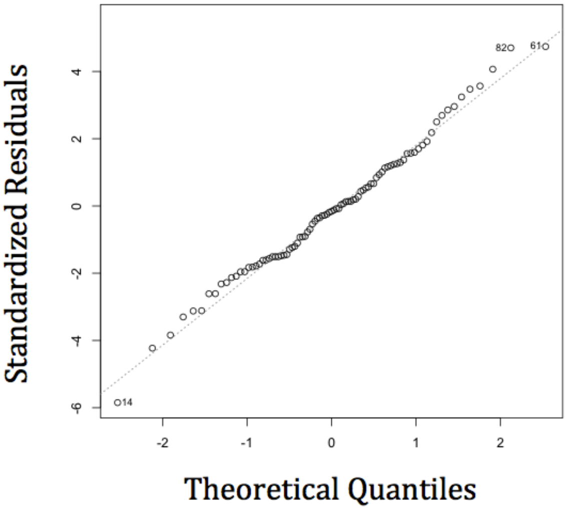


Figure 10: Residual qq-plot for sparse poisson regression

Remark (Residuals): Once we move to generalized linear models for regression, the normal residuals $r_i = y_i - \hat{y}_i$ are not the object of study (even if we are using $\log$ link functions, $\log(y_i) - \log(\hat{y}_i)$ are not the right quantities). Instead we need to use estimates called *deviance residuals*, which in the case of poisson regression is defined as $d_i = \text{sign}(y_i - \hat{y}_i)\sqrt{2y_i \log(y_i/\hat{y}_i) - 2(y_i - \hat{y}_i)}$. We expect these instead to be approximately normally distributed because the square of the deviances can be shown to be approximately chi-squared distributed. Hence, qq-plots for deviance residuals are still meaningful. Note that functions in R often return standard residuals under $obj\$resid$, but return deviances under $obj\$deviance$ or something similar - plotting residuals can be very misleading, pointing in the direction of heteroskedasticity against individual variables and non-existent outliers.

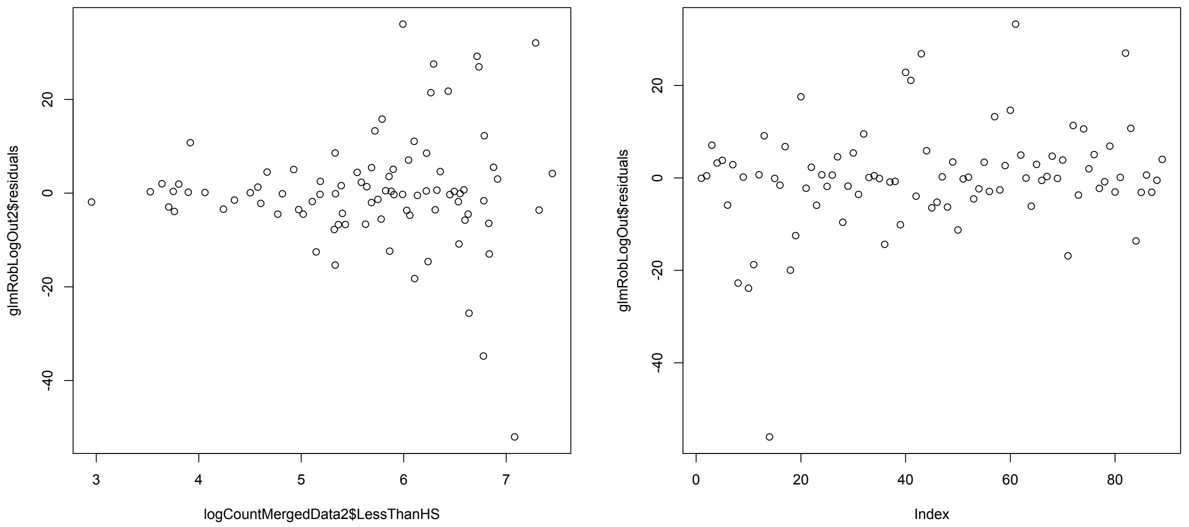


Figure 11: Scatter plot of wrong residuals $y_i - \hat{y}_i$ vs Population with less than high school degree (looks seemingly heteroskedastic) and wrong residual plot (it looks like there are a few non-existent outliers) - with reference to Remark (Residuals).

Covariate Name	Significance ($< 0.001^{***}$, $< 0.01^{**}$)
Pop. less than high school degree	$* * *$
# Foreclosures (2008)	$* * *$
Street density (per sq. mile)	$* * *$
Special land use	$* * *$
# Code violations	$**$
# Renter occupied	$**$
Jobs in n'hood (finance/real-estate)	$**$
Pop. in 1940	$**$
Intercept	$* * *$

Table 2: Significant variables as estimated by two-step sparse poisson regression

Discussion of some significant variables In many experiments, the population of 1940 turned out to be a significant predictor. This could either be because of an interesting sociological phenomenon where populations moved and areas that were occupied are now deserted, or because of spurious correlations. Similarly, special land use and the number of finance/real-estate jobs seem to be curious variables to have been chosen. However, one might expect that the number of burglaries is indeed related to the street density (and hence the number of houses lining those streets), the number of forclosures or code violations (an indication of the housing conditions), the population with less than high school degree and the number of renter occupied units (education and income related pointers).

Robust Estimators For better dealing with outliers, it is possible to fit robust estimators which re-weight data points appropriately and use penalties that are more robust to outliers than the squared loss or the ℓ_1 regularizer (these are often capped penalties or non-convex ones like SCAD fitted by approximate local-minimizers). In our experiments, these did not give qualitatively different results and hence we do not delve into too many details.

8 Poisson (Generalized) Additive Models

In this section, we relax the linearity assumption, and instead use general smooth one-dimensional functions for each variable - this is called an additive model. The formal description is :

$$\begin{aligned} y_i &= \text{Poisson}(m_i) \\ \log(m_i) &= \sum_j s_j(\log(x_{ij})) \end{aligned}$$

where s_j is a learnt univariate function. We fit the model to the data using a two-step process. First, we choose the variables using cross validation and poisson linear models (as in the previous sections) - hence the selected variables are the same as those in the table in the previous section (with a few more non-significant ones, totaling ten variables). Then, we fit an additive model to only the selected variables and plot what the curves look like.

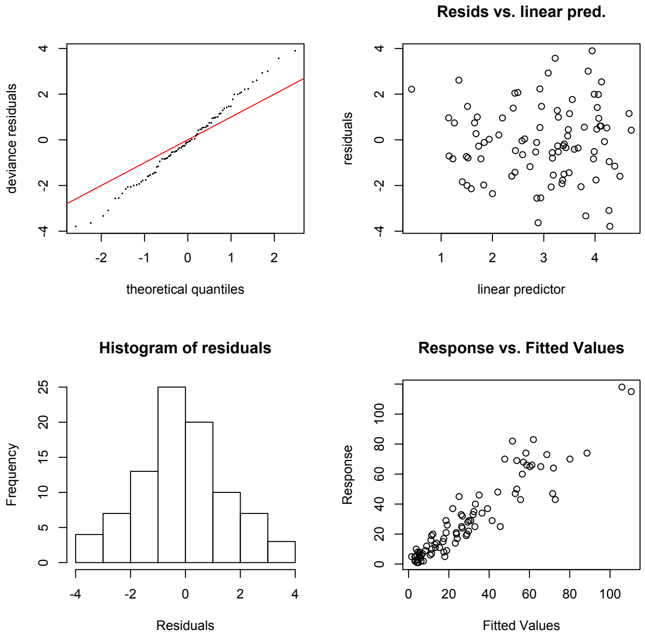


Figure 12: Routine checks for the fit of the Poisson GAM model assumptions.

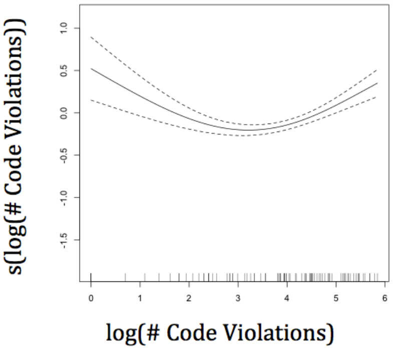


Figure 13: An example of the learnt smooth function for Code Violations.

9 Poisson Sparse Additive Models (SpAM)

Instead of performing a two-step procedure like above, it is possible to generalize the SpAM algorithm to generalized additive model settings and perform variable selection at the same time as fitting the data to a poisson additive model. This one-step model-fitting is done by the new R package called SAM.

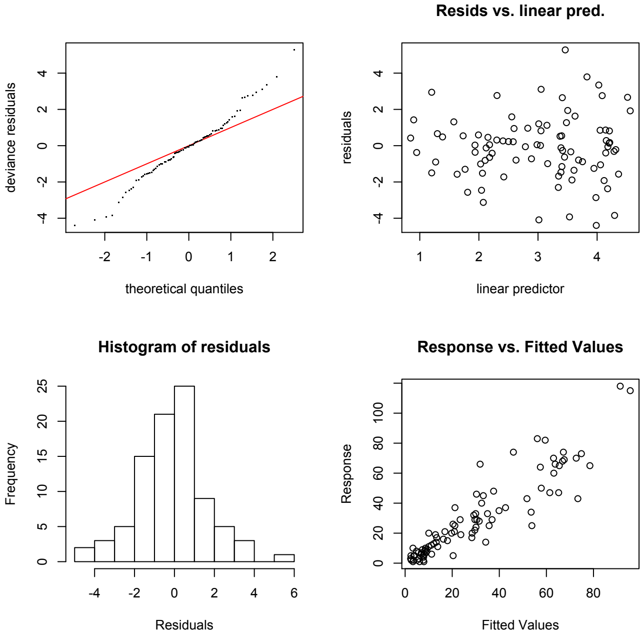


Figure 14: Routine checks for the fit of the Poisson SpAM model assumptions.

We then use the selected variables by SAM to re-fit a GAM on only these variables. This removes the shrinking bias of the SpAM algorithm, but introduces a bias on the estimated significance of these variables. We can run similar model-checking diagnostics as in the previous section as given in the above figure. We summarise the variables chosen by SAM and their estimated significance as given by the second step of fitting a GAM below:

Covariate Name	Significance ($< 0.001^{***}$, $< 0.01^{**}$)
Pop. less than high school degree	$***$
Pop. under poverty	$***$
Total street miles	$***$
# Units built before 1939	$***$
# Code violations	$***$
# Units (2000)	$**$
Intercept	$***$

Table 3: Significant variables as estimated by Poisson SpAM followed by a GAM fit.

We note that the deviance residuals are quite far from being approximately normal (with large quantil deviations as seen by the y-axis scaling of the qq-plot) and the histogram of residuals is visibly skewed. This perhaps indicates that our model assumptions or not being met well.

In the next two sections, we shall question both the poisson assumption as well as the $\log$-link function, which are used in this Poisson SpAM model.

10 Overdispersion and Negative Binomial Models

A poisson distribution has its mean equal to its variance, and poisson regression makes a similar such assumption about the data. If the data has much larger variance than mean, it is said to be *overdispersed*.

We use *odTest* (library *pscl*) to determine if there is overdispersion or not (when fitting a poisson GLM to certain selected variables), and it is overwhelmingly certain that the data is overdispersed. A common reason is the omission of relevant explanatory variables, or dependent observations. Under some circumstances, the problem of overdispersion can be solved by using a negative binomial distribution instead.

A negative binomial distribution has an extra parameter compared to a poisson, called the overdispersion parameter, which allows it an extra degree of freedom in fitting data with larger variance (this parameter equals 1 for a poisson), not constraining it to be equal to the mean.

$$\begin{aligned} y_i &= \mathrm{NegBinom}(m_i, \theta) \text{ for } i = 1...n \\ \log(m_i) &= \log(\boldsymbol{x_i})^\top \boldsymbol{\beta} \\ \beta_j &\sim \mathrm{Laplace}(0, \lambda) \text{ for } j = 1...p \\ \text{Note that } \log(\mathbb{E}[y_i|\boldsymbol{x_i}]) &= \log(\boldsymbol{x_i})^\top \boldsymbol{\beta} \end{aligned}$$

It does choose the overdispersion parameter to be 7.9 instead of 1. However, there is not enough data to fit a more general model confidently, and only the intercept turns out to be significant with p-value < 0.01.

11 ACE and AVAS

The poisson models typically choose a $\log$ link function between the mean of the response and the covariates. However, this might not be true for the data at hand. ACE and AVAS try to fit additive models between the covariates and some learnt function of the response. In other words, one can think of them as trying to fit

$$\mathbb{E}[g(y_i)|\boldsymbol{x_i}] = \sum_j \beta_j s_j(\log(x_{ij}))$$

where s_j is a univariate smooth function of the j-th predictor variable and g is a smooth transformation of the response (burglaries). What we see from the learnt transformation of y is that the $\log$ link function might be too aggressive for our purpose (square-root seems better). It is possible to see this using the following figures from fitting ACE (and similar for AVAS).

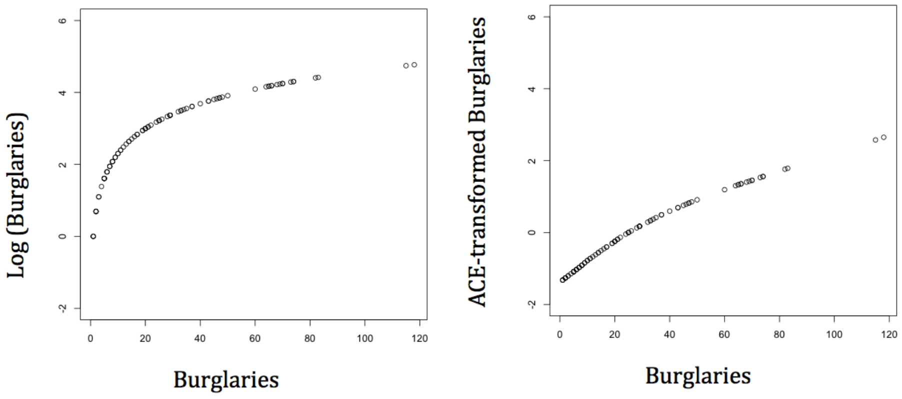


Figure 15: Comparision of the ACE-transformed response to a log-transformation.

12 Poisson (Generalized Linear) Mixed-Effect Models

A mixed-effect model is a kind of small-area estimation technique. The broad idea behind small-area estimation is that we may not have enough information about small neighbourhoods to make substantive claims with high confidence. However, if we take geography into account and use the fact that the city can be divided first into 16 sectors, which are further subdivided into 90 neighbourhoods, then neighbourhoods in the same sector can borrow strength from each other.

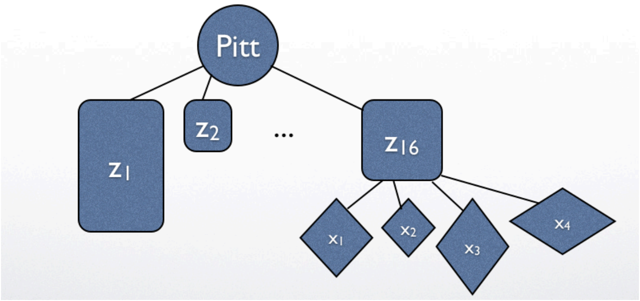
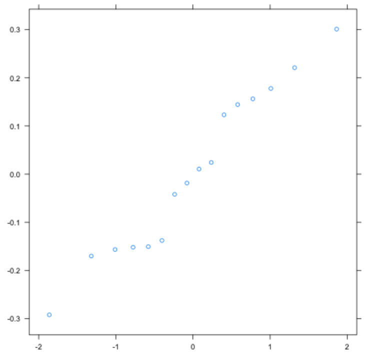


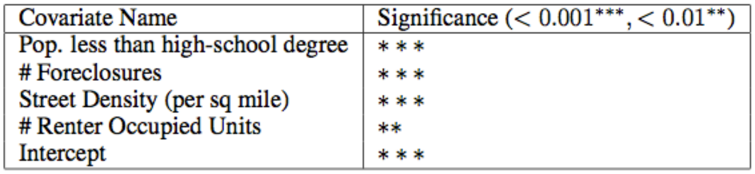
Figure 16: A graphical description of the hierarchy of variables.

This is done by introducing a new Gaussian random variable z_s for each sector and this variable is common in the regression to every neighbourhood i in the sector s. If $s(i)$ denotes the sector corresponding to neighbourhood i, the poisson mixed-effect model can be summarised as

$$\begin{aligned} y_i &\sim \text{Poisson}(m_i) \text{ for } i = 1...90 \\ z_s &\sim \text{Normal}(0,1) \text{ for } s = 1...16 \\ \log(m_i) &= x_i^\top \beta + z_{s(i)} \text{ for } i = 1...90 \\ \text{Alternatively } \log(\mathbb{E}[y_i|x_i, z_s]) &= x_i^\top \beta + z_{s(i)} \end{aligned}$$



QQ plot for Random Effects



Covariate Name	Significance (< 0.001***, < 0.01**)
Pop. less than high-school degree	* * *
# Foreclosures	* * *
Street Density (per sq mile)	* * *
# Renter Occupied Units	**
Intercept	* * *

Table 2: Significant variables as estimated by the Poisson-GLME model.

Figure 17: The 16 inferred z_s variables (left) and summary of important variables (right).

The model assumptions don't seem to be violated in residual plots or the quantile-quantile plots for the inferred random effect variables. The significance test should again be taken with a pinch of salt, but the variables seem to be believable - it does seem like all four make sense as predictors of the number of burglaries in a neighbourhood. More complex GLME models, including those where the sector-level effect is not just gaussian noise but is a linear function of the covariates, can also be tried.

13 Conclusion

We find that most of the existing models have different pros and cons in their flexibility and assumptions. Since we have very few data points, we need to enforce *sparsity*. Since we are dealing with counts, we would like a *poisson* model, but in fact the data is overdispersed and hence a *negative binomial* might be a better choice. The dependence on covariates is captured better by *non-linear additive* models, and probably with a *non*-log link-function. There seem to be outliers, and hence we would prefer a more *robust* and unequally-weighted fitting procedure. To take advantage of the geography and get better inferences for smaller neighborhoods, it seems natural to consider *mixed-effect models* or other forms of small-area estimation. Combining all of these wishes would lead to too general a model with not enough data to get confident estimates of parameters (apart from lack of theory describing how to do so in the first place). Hence, one would have to stick to some combination of these requirements, as deemed necessary by analyst.

13.1 Significant/Important Variables

In summary, even though we do not have a single satisfactory model, most models came up with fairly similar sets of important or significant variables. These included population with less than a high school degree (2010), population under poverty (2010), number of foreclosures (2008), number of condemned structures (2010), number of demolitions (2010), number of unoccupied parcels (2010) and number of code violations (2010).

13.2 Ethical Considerations

The focus of this project was on predictive models for crime, and *not causative* ones. While this may be clear to statisticians, it may not be as clear to the public who may have access to this report and analysis. It should be emphasized that there are no qualitative judgments being passed on what the causes for crime are, since that is a complex sociological issue and is the subject of other books and research. Students, researches, journalists and others who study this data should be aware of the possible impact it may have on the public mindset, and other economic factors like the costs of houses, the cost of insurance, and care must be taken to emphasize results in the right manner and context.

13.3 Disclaimer

”The data made available here has been modified for use from its original source, which is the Government of the City of Pittsburgh. Neither the City of Pittsburgh Government nor the Departments of City Planning (DCP) and City Information Systems (CIS) make any claims as to the completeness, accuracy or content of any data contained in this application; makes any representation of any kind, including, but not limited to, warranty of the accuracy or fitness for a particular use; nor are any such warranties to be implied or inferred with respect to the information or data furnished herein. The data is subject to change as modifications and updates are carried out. It is understood that the information contained in the web feed is being used at one's own risk.”

13.4 Acknowledgements

I'd like to thank my advisor, Larry Wasserman, for meeting with me every week, and not growing tired of my rookie questions ranging from what exploratory data analysis is, to philosophical musings about doing statistical inference after model selection.

13.5 Relevant Links

[1] *http://www.pittsburghpa.gov/police/files/annual_reports/10_Police_Annual_Report.pdf*

[2] *http://www.pittsburghpa.gov/dcp/snap/raw_data*

[3] *http://apps.pittsburghpa.gov/dcp/PGHSNAP_Data_Dictionary.pdf*