Mapping Crime: A Statistical Study of NYPD Incident Data

Mario De Menech, Paulina Suchanek

Abstract

We present an investigation on temporal and spatial patterns of census and incident public data retrieved from the NYC Open data portal. We observe that the relative frequency of crime events has been quite stable over the years for all neighborhoods we considered, and that it is possible to cluster blocks over homogeneous regions to identify the city areas which should be considered more dangerous. The analysis is carried out using fused lasso methods, both for one- and two-dimensional data.

1. Introduction

In this report we describe our findings for the analysis of incident event data collected by the New York Police Department (NYPD) over a timespan of 18 years. Records relate to crime and accident events, including shootings (Shooting, 2020), arrests (Arrests, 2020), filing of complaints (Complaints, 2020) and motor vehicle accidents (Crashes, 2020). All records have a timestamp and are geo–referenced, and it is therefore possible to investigate their time evolution and to draw maps of their distribution; we aggregate them over different scales defined by administrative neighborhood areas, which allows us to correlate the distribution of the incident data to census data (Census, 2020).

The main challenge of these datasets is to identify the relevant patterns in the temporal and spatial distributions, and we approach this problem using fused lasso methods. We consider the generalized lasso problem

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|D\beta\|_1 \,, \tag{1}$$

where X is the identity matrix, y is the original signal (one- or two-dimensional) and β is the optimal approximation of y. The lasso term is defined by the penalty matrix D and the regularization parameter λ, with higher values of λ favouring clustering of β values. In the case of 1d time dependence analysis presented in the next section, the matrix D is simply a discrete difference operator. For the spatial correlations analysis of 2d maps, D is the incidence matrix of an underlying graph of the city areas.

. Correspondence to: Paulina Suchanek <paulina-maria.suchanek@studenti.unipd.it>, Mario De Menech <mario.demenech@studenti.unipd.it>.

Statistical Methods for High Dimensional Data, University of Padova, 2024-2025

1.1. Definition of neighborhood areas in New York City

New York City is comprised of five counties, also known as boroughs: Bronx, Brooklyn (Kings County), Manhattan (New York County), Queens and Staten Island (Richmond County). A census tract (CT) is a small geographic unit delineated for the presentation and analysis of decennial census data. Geographic files of Neighborhood Tabulation Areas (NTAs) are created by Department of City Planning, using whole census tracts as building blocks. They allow for more detailed analysis at the NYC neighborhood-level. Fig. 1 shows the graph with nodes corresponding to the centroids of the CTs (year 2020) while edges connect each CT with its neighbours.

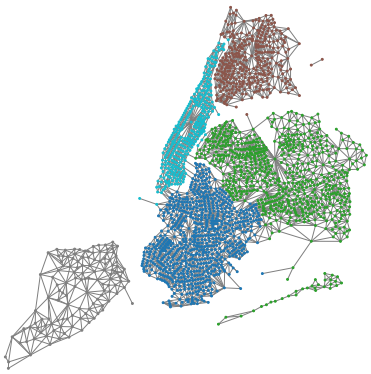

Figure 1. Graph of New York City Census Tracts (year 2020) built with nodes given by the centroids of the geographic units and the edges connecting neighbouring CT. The different colors refer to the five boroughs (Bonx, Brooklin, Manhattan, Queens and Staten Island). The graph has 2325 nodes and 7195 edges.

2. Temporal analysis

2.1. Crime in time

In the first part of the project we investigate which are the most dangerous areas in the city and how they were chang-

ing in time. Firstly, we look at the monthly sums of incidents (arrests and shootings) in the years 2006-2023 on the level of boroughs. Since the monthly counts are highly variable we use the 1d fused lasso trend filtering in order to denoise them. As a result we obtain plots (Fig. 2) indicating a seasonal pattern in the counts in each borough. There are fluctuations between summer with higher counts and winters with more rare incidents. For the arrests data the seasonal differences are visible in the years before 2015, while afterwards the number of arrests is definitely decreasing and the seasonal differences are not present anymore. The two years of covid pandemic (2020-2021) are the years with less arrests and higher number of shootings. From then the number of monthly arrests grows, while the counts of shooting incidents decrease. From the plots it follows

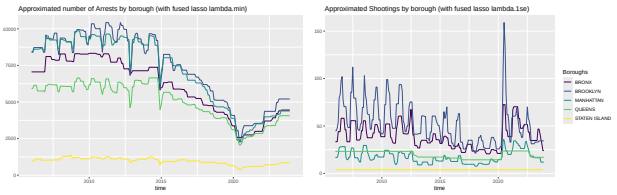

Figure 2. Approximation of monthly incident counts by borough. The plots are obtained using R package genlasso, equiped with cross-validation function calculating the optimal regularization parameter λ_{min} and λ_{1se} given by the one standard error rule.

that the relative number of incidents in different boroughs is rather stable. It can be confirmed by approximating the relative counts using 1-dim fussed lasso. Moreover, on the level of 77 police precincts existing in NYC, the percent of incidents for a given precinct does not vary much in time either. Fig. 3 shows the denoising property of the approximation with fused lasso penalty in the case of shootings counts.

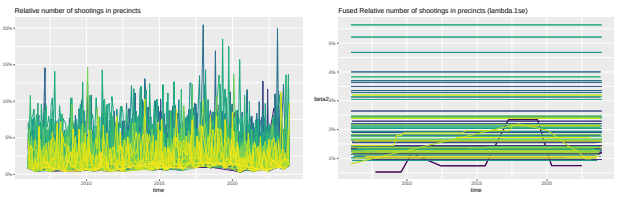

Figure 3. Monthly relative incident counts in 77 precincts

One can conclude that the most dangerous areas of the city are always the same ones and that the relative probability of a shooting in these places does not change in time. Similar results can be obtained for the arrest data, though there is slightly more variability present. The areas indicated by the above analysis as the most dangerous are visualised in Fig 4.

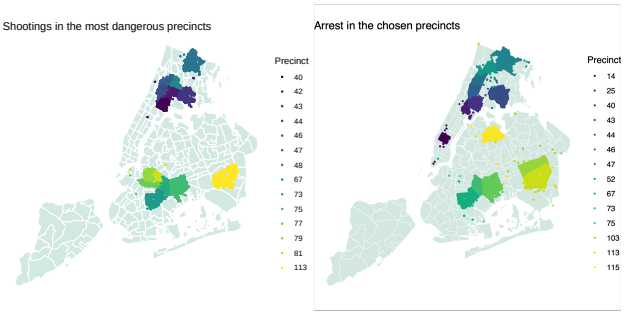

Figure 4. Precincts with the highest relative incident counts

2.2. Crime in NTAs

In order to gain a more quantitative understanding of the dangerous areas in the city we would like to model a map that shows the expected percentage of incidents (arrests/shootings) in 262 different neightborhoods (NTAs) in New York City. As we have seen, the relative number of crimes for a given area (precinct) does not change much in time. Thus we can calculate the relative number of incidents per year in a given neighborhood, and treat the counts for each year as independent samples from the probability distribution of crimes over the city. In this way we have a set of maps from different years, and we can model an optimal one. We use fused lasso regularization eq. (1) in order to encourage clustering of the areas of similar safety. The penalty matrix D is given by the incidence matrix of the graph of the NTAs. As the variable y we consider the matrix with rows defined by the counts of incidents in the NTAs for a given year. The matrix X is an appropriate matrix of ones since we look for one β vector optimal for all rows of y. In order to find the solution we use CVXPY library for complex optimization problems. In particular, the Splitting Conic Solver, a numerical optimization ADMM-based package is applied. The regularization parameter λ is chosen with the help of the 5-fold cross validation procedure, where in each fold a different subset of maps (rows of y) plays the role of the validation set. The optimal smoothing parameter λ is small

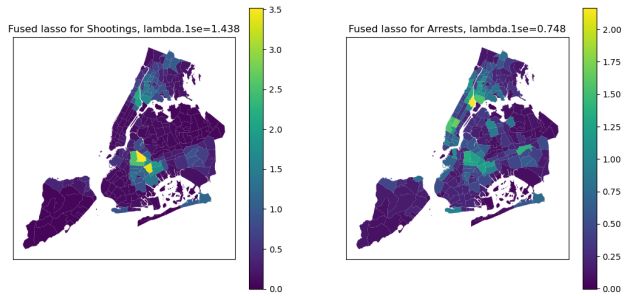

Figure 5. Maps of expected incidents percentage in NTAs

for both arrests and shootings. The areas of high incidents probability are well localised, often not far from much safer places and they do not vary much in time. This is a typical feature of the US cities. Larger values of λ give a map that is more blurred. Such a map is not an optimal solution to our problem of modeling the dangerous areas of the city, but it can show the general differences between boroughs, confirming what we arleady saw in the 1d time analysis of crime in the previous subsection. Finally, let us notice that there are few neighborhoods with very high relative number of incidents and many with much smaller, which makes it difficult to see from the map the differences between safer areas. In order to underline these differences we can look at the map of transformed values, for example in the logarithmic scale. It turns out that the map that describes better the safer areas can be obtained by first transforming the data and then modeling a map with fused lasso penalty. In this case the darker parts correspond to the safest neighborhoods, while the brighter to the neighborhoods with possible incidents. Since the transform is non-linear, it is more difficult to spot the most dangerous areas. The choice if to scale the data or not in this case depends on our aim: either we can get insight into the safest areas, or in the most dangerous areas, respectively.

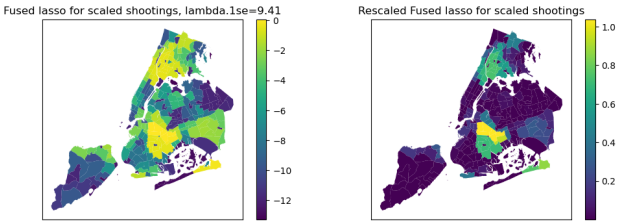

Figure 6. Maps of expected incidents percentage in NTAs

In the next section we will present the analysis of crimes on the level of census tracts (CTs). Because of the large number of the areas (10 times more that NTAs), the cross-validation procedure based on samples from different years is prohibitively costly. Instead, we consider data in the form of one map and model its optimal approximation with the help of fused lasso, fixing the regularisation parameter by masking some of the areas on the map.

3. Spatial analysis

The geo–localized events were aggregated over the CT units to define a discrete variable which can be visualized in a color map. Fig. 7 shows the density for the shooting incidents (Shooting, 2020). We note that shooting incidents are mostly concentrated in a limited number of CTs. We use the fused lasso method (1) to try and provide a more

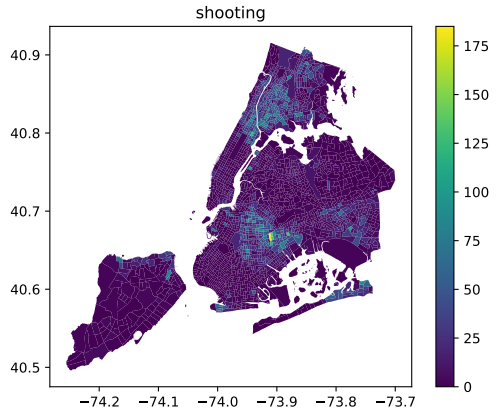

Figure 7. Map with shooting counts for the CT units.

interpretable representation of the events. This method was introduced for image denosising; neighbouring nodes are forced to have a similar value. The approximation of the graph variable y_i by the unkonwn variable β_i (the index i runs over the nodes of the graph) is obtained by solving the problem eq. (1), where D is now the adjacency matrix of the graph of CT (Fig. 1). The minimization problem can be solved using the Alternating Direction Method of Multipliers (ADMM) algorithm (Stephen Boyd & Eckstein, 2010). For a fixed penalty λ, the algorithm is quite effective and it is rather easy to get a map where the values of the variable β are clustered. The challenge is to define a cross-validation like procedure where the hyperparameter λ would be determined in a data driven way. To this end we would repeatedly randomly subset the graph nodes to be masked and train the model on the remaining nodes. The error of the model would be computed on the masked nodes. For this purpose eq. (1) was slightly modified introducing a diagonal matrix M with entries 0 or 1, where 0 indicates that the node is masked as in eq. (2).

$$\min_{\beta} \frac{1}{2}\|M(y-\beta)\|_2^2 + \lambda\|D\beta\|_1 \qquad (2)$$

We built our implementation of ADMM algorithm to solve this problem taking advantage of the sparsity of the matrix D. We found that the results of the CV procedure depend on the nature of the data (population, arrests, shootings, complaints, crashes) and on the normalization of the variable. We considered the scaling of the variable to the interval $[0, 1]$ both using the maximum and minimum values of the variable and the non linear transformation based on the empirical cumulative distribution function (CDF). In most cases our CV procedure led to the trivial solution $\lambda \approx 0$. Apparently the 'noise' level is mostly too low for these graph images, and therefore any attempt to smoothen the

data does not improve the capability of the model to predict the masked nodes. In general the normalization based on the CDF emphasizes the spatial fluctuations of the data (see fig. 8), and the CV points follow a more regular trend with a step-like behaviour as the regularisation parameter λ increases, as shown in fig. 9. The definition of a value of

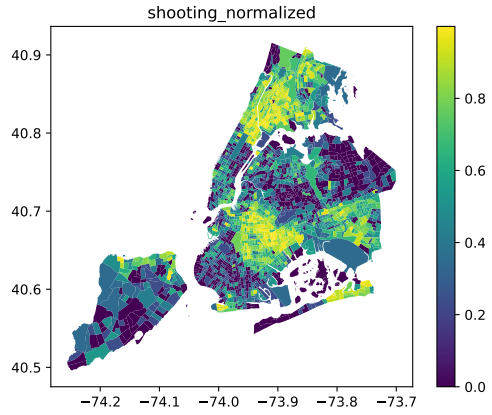

Figure 8. CDF normalized shooting counts.

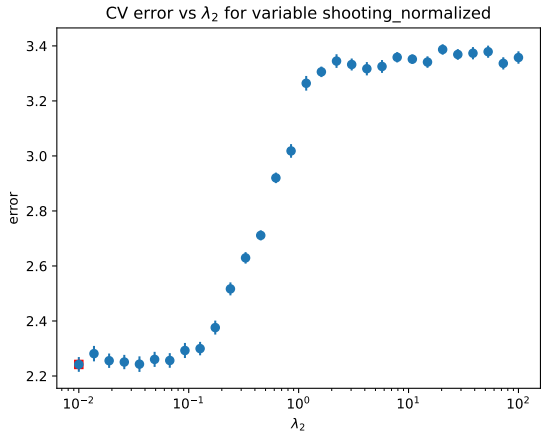

Figure 9. CV plot for the normalized shooting count on CT. Model is trained on 95 percent of nodes and the CV error was computed on the remaining 5 percent. Each point with its error bar is the estimate of 40 CV runs.

λ which is most suitabile for the clustering of the data can be therefore related to the shape of the CV curve, and an automated procedure could search for the largest value of λ which keeps the CV error low, i.e. marks the the transition from the low to the high error regime (see for example fig. 10).

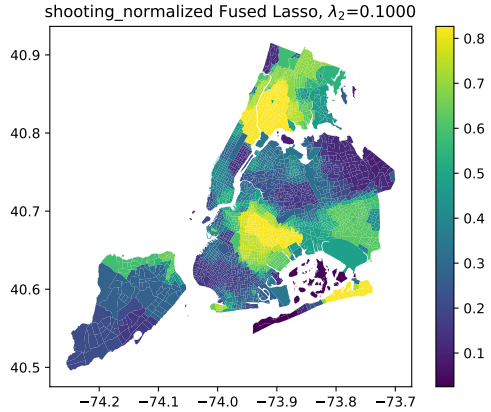

Figure 10. Clustering of normalized shooting counts: the penalty parameter was set to $\lambda = 0.1$, which allows for a low CV error while favoring the sparsity of the cluster data.

4. Conclusions

We conclude that the fused lasso regularization methods can be very useful in the analysis of crime in NYC. Although the data is very rich, when it comes to visualizing crime, we have to cope with a large number of features and then sparsity induced by fused lasso is crucial. With the help of 1d trend filtering we got indications about the location and stability in time of the most dangerous areas in the city. Moreover, using fused lasso penalty we were able to obtain maps that give insights into the most dangerous areas, or into the safest areas if we appropriately scale the data. The most challenging part of the analysis was to design a procedure to choose an optimal regularization parameter λ when the penalty matrix D was the incidence matrix of a graph. In the first approach, considering the smaller graph of NTAs, we used the cross-validation procedure applied to the dataset viewed as a set of samples of maps from different years. In the second approach, the less standard cross-validation procedure was worked out, where the randomly chosen nodes of the graph were masked in order to serve as the validation sets in different folds.

References

Arrests. Nypd arrests data (historic). https://catalog.data.gov/dataset/nypd-arrests-data-historic, 2020.

Census. Nyc planning - 2020 census. https://www.nyc.gov/site/planning/planning-level/nyc-population/2020-census.pag, 2020.

Complaints. Nypd complaint data (historic).

https://catalog.data.gov/dataset/nypd-complaint-data-historic, 2020.

Crashes. Motor vehicle collisions - crashes. https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes, 2020.

Shooting. Nypd shooting incident data (historic). https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic, 2020.

Stephen Boyd, Neal Parikh, E. C. B. P. and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 1:1–122, 2010.