Franco La Bruna
Yuchen Zheng
CSC 440: Data Mining
Class Project Proposal

Determining the Relationship between Crime Rates, Poverty, and Noise Complaints in New York City

Problem Statement

With the rise of data mining methodologies, we now can apply them to study a wide range of social problems. Inspired by Professor Luo, we would like to study the potential relationship between the enormous noise complaint data and crime rates & poverty in New York City area with implementations of various data mining techniques we have obtained from class.

It is one of our incentives that previous literature has seemingly done very limited or even none research on our topic. One paper that is most relevant to what we intend to do is by “Diagnosing New York City’s Noises with Ubiquitous Data” by Zheng, Y. et. al, in which they employed noise complaint data to build a model that recovers the noise situation throughout NYC to inform people and officials’ decision making. Though what they have accomplished is disparate from what we intend to do, one sentence in the introduction of that paper is genuinely enlightening, “As each complaint about noises is associated with a location, a time stamp, and a fine-grained noise category, … , the data is actually a result of “human as a sensor” and “crowd sensing”, containing rich human intelligence that can help diagnose urban noises.” For the project, our group is also going to use the data as a result of “human as a sensor” and “crowd sensing” to figure out the relationship between crime rates, poverty and noise in New York City.

Our project has roughly two major parts. The first step is to collect data of noise complaints, crime rates and poverty in New York City, as elaborated in Data Acquisition. We will then need to preprocess the data and implement classification and clustering techniques. In the preprocessing part, we mainly want to integrate the data, since different datasets need to be merged and hence create overlaps that result in redundancy (locations, etc) and data value conflicts, and reduce the data, since we want to avoid wasting time on, for instance, studying 1,000 similar verbose noise complaints from transportation at the same location within the same day. Upon getting more concise and clean data, we can further choose from assorted classification and clustering methods based on the data’s geo-locations and features to make it easier and more saleable for us to proceed to the next stage.

In the second step we will be mainly focusing on attempting to discover correlations and patterns with the utilization of different regression models, depending on knowledge of the response distribution, theoretical considerations and empirical fit to the data. Given the amount of data we have and will get, we are positive and optimistic that many statistically significant and interesting patterns can be found. Furthermore, we are going to try to study and explain any

causal relationship or hidden parameters behind these patterns such that we can set up prediction models to give insights for future applications and reference.

Data Acquisition
Noise complaint data can be readily acquired from New York City Open Data. The data provided there contains all noise complaints to the 311 non-emergency service hotline from 2010 until the present day, and for each noise complaint also lists the date of the complaint, the source of the complaint, a description of the offending noise, and location information by various metrics such as borough, street name, ZIP code, and GPS coordinates, if available.

Crime data is more nebulous, and will likely have to be compiled from multiple sources. NYC OpenData has databases listing firearm discharges, major felony arrests by date and borough, and a table listing emergency responses by location and the reason for the response. In the latter case, we could isolate requests for police intervention as an indicator of crime.

Poverty is also slightly nebulous, but we could determine it by pulling historical data about average income and poverty rates for each borough, and use that as a sort of weight. The NYC OpenData site has a listing of homeless shelters in the city by location; we could use a higher density of homeless shelters per population as an indicator of greater poverty in an area. While such data is only available from 2009 to 2012, the site also provides estimates of the number of homeless people by borough of NYC; we could also use this as an indicator of poverty. Using some geographical indicator, like ZIP code, as the centroid, we could assemble this data into a tensor to determine the relationship between these three data types.

Data Sources
https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
https://data.cityofnewyork.us/Public-Safety/Historical-New-York-City-Crime-Data/hqhv-9zeg
https://data.cityofnewyork.us/Public-Safety/NYPD-7-Major-Felony-Incidents/hyij-8hr7
https://data.cityofnewyork.us/Social-Services/Directory-Of-Homebase-Locations/ntcm-2w4k
https://data.cityofnewyork.us/Social-Services/Directory-Of-Homeless-Population-By-Year/5t4n-d72c
https://data.cityofnewyork.us/Social-Services/DHS-Daily-Report/k46n-sa2m
https://data.cityofnewyork.us/City-Government/Demographic-Statistics-By-Zip-Code/kku6-nxdu
https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Census-Tracts/37cg-gxjd
NOTE: Preliminary list. We are likely to add or remove data sources as our project proceeds.

Software Tools and Programming Languages
Software used will include the Python programming language in the Anaconda 3 distribution. Since our code will likely require us to assemble a tensor centered around location, given the fact we are dealing with multiple arrays of multi-dimensional sequences that are tied only by

region of occurrence, we will use code created by Zheng et. al. in "Diagnosing New York City's Noises with Ubiquitous Data" to build and decompose our data into a format that relates each set of attributes (noise complaints, crime, poverty) by location.

Methodologies/Algorithms choices:
At this point we do not know exactly what algorithms our project will require to get the necessary information. However, it is quite likely we will require the use of a pattern mining algorithm, like FP-Growth, modified to fit our data.

Citations

[1] Zheng, Y. et. al. "Diagnosing New York City's Noises with Ubiquitous Data." Ubicomp 2014. Published 17 September, 2014. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/Diagnosing20NYC27s20noise20using20ubiquitious20data-yu.pdf.

[2] NYC Information Technology and Telecommunications, and NYC Mayor's Office of Data Analytics. "NYC Open Data." NYC Open Data. The City of New York, n.d. Web. 01 Nov. 2016. https://data.cityofnewyork.us/.

[3] Gomez-Lievano, Andres, Oscar Patterson-Lomba, and Ricardo Hausmann. "Explaining the Prevalence, Scaling and Variance of Urban Phenomena." Cornell University Library. Web. 03 Nov. 2016. https://arxiv.org/abs/1604.07876.