Project Report:
The Impact of Social Safety Factors with Considerations for Commercial Decisions

Paul Coen, Southern Illinois University
KirkLeroy J. Powell, Southern Illinois University

Abstract: *A look at the various clustering methodologies applied to social safety factors in determining locality specific geographic zones that are ideal for commercial decisions. Many independent repositories of safety data for large urban centers exist in disconnected formats. Our study makes a brief analysis of these factors inside Chicago, considered to be one of the least safest cities in the United States, to determine if a correlation exists within subspace dimensionality and if that correlation can be causal proof to influence commercial decisions by applying clustering techniques on several layers to identify specific regions and time frames that create the highest level of uncertainty in decision making and representing this analysis in a real-time data feed to create visually identifiable pockets of hostility, aggression, etc. This research will attempt to suggest improvements to techniques in processing data flow from urban communities and potentially reveal opportunities to improve upon known clustering methodologies.*

Keywords: Big Data, datasets, DBSCAN, dimensionality, K-means, MatPlotLib, Mean-Shift, meshing meta-algorithm, NumPy, Panda, Python, preferential geography, preprocessing, Ruppert's algorithm, Sow & Grow, Spectral clustering, subspace

1 Introduction

Our research topic originated from a consideration of how known factors might be influencing the patrons of commercial entities (i.e. restaurants) and how we can best relay that information to potential customers in a real-time data flow. We began with a rudimentary analysis of social factors that could be considered to affect eating habits; health codes, menu options, access to the internet, rental rates, condom distribution, family planning services, homeless shelters, abandoned buildings, political districts, etc. Our research and accompanying project are limited in scope and depth due to the natural constraints of class time and semester length. As such, we truncated our options for datasets based on their availability, scope of data, and plausible application within our project parameters. We derived that crime statistics, food safety, and school quality were optimal pools of information that satisfied our query requirements and provided rich, distinct, and timely datasets. Our datasets are all tied by their longitude and latitude, making the result sets optimal for image mapping on a hand-held device, such as within a cell phone app. The need for mapping meant that the clustering algorithms in traditional data mining would be best suited for our analysis. We applied a data preprocessing method that constituted reasonable arbitrary exclusions and implemented our criteria with a Python script. We then built a data processing tool in Python using open-sourced libraries. Armed with a treasure trove of Chicago's data and computational capabilities we were able to identify pockets of discernible quality within the data and found direct correlations among the datasets that prove our basic hypothesis, that preferential geography exists.

2 Roles

We met once a week to discuss our project research and implementation. We started by organizing a schedule, determined tools useful to our project, and broke up certain tasks into action steps for each person. We had several discussions about which algorithms we would implement and why. We then found datasets that met our criteria for research. We coordinated on the data preprocessing of these sets to ensure that they would interact well.

Kirk Powell
Kirk is especially skilled in presentation and writing. Kirk was responsible for writing the majority of the paper and the development of the presentation.

Paul Coen
Paul is experienced in data mining techniques and implementation. Paul developed the code and made the major decisions on implementation.

3 Tools

Our project is publicly available on Github:
https://github.com/IotaDraconis/Is_It_Safe_To_Eat_In_Chicago

Markdown is the best and simplest way to construct ReadMe files on a project. To view documents with the markdown language, you will need to install/enable a browser extension:
https://chrome.google.com/webstore/detail/markdown-viewer/ckkdlimhmcjmikdlpkmbgfkaikojcbjk?hl=en

Additionally, for a breakdown on encoding with Markdown:
https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet

We used Python, version 3.6.5, to build both data processing tools. Some of the essential libraries available through Python are only accessible in version 3 and above.
https://www.python.org/

We employed a useful open-source tool for analysis and structure of datasets within the Python language. These were easy to implement and resulted in stunning visual graphs.
Pandas 0.23.4:
https://pandas.pydata.org/

Matplotlib 3.0.0:
https://matplotlib.org/

NumPy 1.15.4:
http://www.numpy.org/

A set of known data mining algorithms is available at:
http://scikit-learn.org/stable/modules/clustering.html#clustering

For our amazing presentation, we used Office365 online through our student accounts at Southern Illinois University. However, this tool is available here:
PowerPoint Online
https://www.office.com/launch/powerpoint?ui=en-US&rs=US&auth=2

We used different text editors, which is largely just personal preference and familiarity. Any such text editor that can handle CSV files without additional encoding is suitable.
Sublime Text 3
https://www.sublimetext.com/3

Atom
https://atom.io/

4 Data Sets

All of our datasets were handled as single CSV files. One can access all the datasets used in this project at:

Chicago Crime Dataset:
https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2 saved as *crime.csv* in the */datasets* directory

Chicago Food Inspections Dataset:
https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5
saved as *food_inspections.csv* in the */datasets* directory

Chicago Public Schools Dataset:
https://data.cityofchicago.org/browse?category=Education
saved as schools_{:*begin_year*}-{:*end_year*}.csv in the */datasets* directory

schools_2011-2012.csv
schools_2012-2013_Elem.csv
schools_2012-2013_HS.csv
schools_2013-2014_Elem.csv
schools_2013-2014_HS.csv
schools_2014-2015.csv // not available
schools_2015-2016.csv
schools_2016-2017.csv
schools_2017-2018.csv

Datasets for the Chicago Public Schools are somewhat more complicated to implement than the other two. First, there is no dataset for 2014/15. For 2012/13, and 2013/14, the datasets are broken into High School and Elementary. After 2014/15, the datasets are single sets that can be downloaded as a CSV file.

5 Data Preprocessing

The */code/prepare.py* file is designed to automate the data preprocessing step in alignment with the expectations of the */code/*main.py file used to process datasets. There is no ideal minimum size [data points] nor minimum & maximum file size requirements for the datasets. It is reasonable to assume that as the size of a dataset increases that the sampling of the information will likewise increase in value as well. However, application of the clustering algorithms produced definitive consistency across numerous random samples. The intent of preprocessing was to isolate the data attributes necessary for useful clustering and reduce the computational costs of processing large datasets. Our initial theory of breaking down Chicago into a grid for a predetermined number of centroids was assisted by the discovery that the latitude & longitude of the crimes was available in the crime data set. This truncated block-by-block reference was sufficient for a generalized understanding

of geographical locations within and about the city. We quickly adopted this as a prerequisite for subsequent data sets in order to link them in a truly meaningful way.

As we worked through the massive crime data set provided by the city, we found that several of the columns were redundant or irrelevant. The data is trimmed by isolating the geographical point, the types of crimes (as identified by a police code, if an arrest was made, and its complimentary classification as 'domestic'), and the date the crime occurred. Some of the data wasn't directly interesting, such as "if an arrest was made", but could prove useful in a future extension of this research work. We wrote a python script to perform this action on our dataset preparing it for the clustering algorithms. Additionally, to compare the different subspaces of datasets, we applied a range modification to balance the data. We used a range from 0 to 1.

A similar redundancy was found within the food dataset. This set needed to be parsed in such a way that we only kept the following: Who the business was doing business as (their recorded name), the local name, the type of facility it was, the risk of the restaurant, the last inspection date, the last inspection type, the results and violations, and latitude and longitude. We felt that these columns would give us the best look at how 'safe' various restaurants are around the city. The algorithms mentioned further ended up using the latitude, longitude, and risk for each set. Risk was the most important factor in the set due to it giving a range from 0 to 2 of how risky it is to eat at a given restaurant.

Collection of the Public Schools data was less pleasant than the crime data. In mining Big Data, as researchers, we needed to be prepared for amalgamations of information collected from different sources, using disparate techniques, with oscillating objectives, and varied results. It was determined that the layer of public-school safety data was valuable to our research work because it provided some contrast against the safety of the general public (each dataset was independent of the other). The appearance of longitude and latitude of the schools further solidified its usefulness in our application. The data for the public-school safety is reduced to location (latitude & longitude), school classification, safety events, and a date.

The limited history of the data required that the crime data be truncated to 2011 to present to match the public-school data. The school classification is needed to maintain clarity across the data sets since they are split after 2012 through 2014 and then consistent through 2017. There is a gap in the data for 2014/15 that could present a problem in the conclusions reached, however, given the sampling process across a large spectrum (nearly a decade of data) that that issue is not likely a concern. The safety events from the dataset chosen include: Safety Score, Family Involvement Score, Environment Score, and Rate of Misconduct. The breadth of this set provides increased texture to the mapping of correlations. To be sure, the dataset for public schools contained a much wider scope of information that could be useful or interesting, but this data needed to be trimmed to limit computational costs and preserve the focus of our research.

6 Clustering Methods

Our slate of algorithms started with K-means, DBSCAN, Mean-shift, and Spectral clustering. We eventually began to consider using Sow & Grow as well. However, the progression of our project followed from K-means to DBSCAN and ultimately the meshing meta-algorithm.

K-means

We initially conceived of the clustering algorithms using K-means because we could define the number of centroids across a grid, which was essentially a map. However, as we applied this algorithm it didn't yield a useful output because the clusters were to consistent. Yes, there is a lot of crime, food safety issues, and public schools in Chicago, however that doesn't resonate an impactful perception.

DBSCAN

The meaningful understanding of the clusters was best suited to DBSCAN. This required significant difficulty in tuning the data to produce results that created vibrant clusters. It also, unfortunately, raised the computational costs. However, taken solely as a linear map, the clusters did not present significant distinctions to reflect the value of densities.

Mesh meta-algorithm

Finally, we engineered a meta-algorithm using meshing techniques. Ultimately, this layered application developed rich graphical representations of the city of Chicago that could be easily interpreted. While tuning the mesh took effort, the overall computational cost was insignificant.

7 Conclusions & Recommendations

Benefits of Known Algorithms

Designing an optimal algorithm to analyze large datasets is like re-inventing the wheel. The primary benefit of an improvement in design would be to reduce the computational cost. However, the likelihood of discovering a more efficient methodology is rare. As such, our project directly benefited from the use of known algorithms and we were able to isolate the useful ones and eliminate the others.

The rich development of software tools for the purpose of data mining meant that we spent little to no time developing a methodology for analysis. Leaving implementation as the only real problem to solve.

Implementation was clearly the single largest benefit to using a known algorithm for clustering. Using a grid system (such as latitude & longitude) enabled us to analyze the data in a way that was consistent with known parameters for the city of Chicago. In fact, the three-dimensional modeling was so purely obvious that an overlay map of the city wasn't needed for a cursory interpretation.

Weaknesses of Known Algorithms

Building an application that implements the available tools was difficult. Each dataset had to be 'handled' by a preprocessing script. This meant specific analysis of the data to determine its usefulness. Again, in applying DBSCAN, a manual tuning of each parameter was necessary to obtain

useful information. The application of meshing also required manual tuning to produce insights of value.

The easy access to established tool sets hinders one's empirical knowledge of the analysis. In implementation, where one simply selects different parameters as filters, the value of what is being done becomes almost unknown to the user. This black box of libraries is enormously helpful in getting results but is dangerously close to plug-n-play technology that minimizes the importance of a scientific approach.

K-means was not useful because it uses a structured set of centroids. DBSCAN produced only two dimensional maps of clusters. Visually, this only represented some of the information of each cluster. The need to tune DBSCAN for each cluster was frustrating at times. DBSCAN has a significant computational cost that literally crashed either of our laptops[1]. The meshing meta-algorithm was limited by its inputs and the need to tune the mesh on each subsequent set. Several other algorithms were discarded for lack of suitability, such as Spectral clustering and Mean-shift.

Confirmation of Hypothesis

A confirmation bias exists in our research work. As natives to Illinois, we are more directly familiar with the violence associated with the city of Chicago. We live and work in a Southern, a somewhat rural part of Illinois, that seems significantly safer than our perceptions of Chicago. An analysis of relative crime, food quality, and schools with alternative metropolitan areas is needed to put perspective on our conclusions.

However, as it relates specifically to Chicago, we are able to show definitively that certain portions of the city are less safe than others. This proved an old adage of real estate, "Location, location, location!" which essentially means that preferential geography exists.

Errors & Limitations

Real-time data flow is both impossible and unrealistic. The city of Chicago doesn't provide a real-time flow of data. Construction of such a portal would be expensive and require cooperation within the city. Additionally, the amount of data flow presents a serious conundrum. The data has to be preprocessed to limit its scope to the relevant attributes. Then the processing of the data requires a computational magnitude that is impractical to implement at this time. We presume real-time data flow is necessary for accuracy. However, since the dataset is itself a sample, there is no benefit to the accuracy of our statistical models achieved by real-time data flow (outside of anomalies, outliers, and noise). It is for these reasons that we conclude that real-time data flow is not ideal.

Our first attempt to process data suffered from computational limits. We grossly underestimated the cost for processing even just the sample sets from the data. We were able to compute the clusters of with DBSCAN using 50,000 data points in subsets of the data. The need for speed limited the number of times we could run the script and continue to tune our parameters for distance and minimum points.

We did not use training data and no error analysis was done. A more thorough work would have done both. It is our recommendation that anyone on this path make such steps essential in a more serious approach (that has less time constraints and more qualified data scientists).

[1] We did manage to get past this by using Paul's desktop computer. This was running python through Anaconda and my be the reason it was able to plot successfully with DBSCAN.

Opportunities for Future Work

Increasing the pool of datasets would provide a more in-depth correlation of safety and commercial decisions. Choosing just three sets of data was a necessary convenience given the time frame to complete our research. It was clear from the beginning of our analysis that there exists a wide spectrum of plausible subjects (datasets) that could influence and even skew assumptions in an alternative manner. For instance, a high concentration of identity theft crime in downtown Chicago might be better explained by population concentration or wealth distributions.

Applying the same methodology to additional metropolitan areas would be a gargantuan task. But one cannot ignore that the referencing factors for balancing the data would likely improve dramatically when compared to all major metropolitan communities throughout the world. The classifications of safety, and conversely the factors that combine into communities to make them safe, are sufficiently self-evident to warrant a much larger expedition into other societies.

To identify classifications and their confidence ratio's might be the single most important opportunity in data mining public safety and its inherent effect on commercial decisions. An anecdote of grandma being mugged can be confirmed by mountains of data. Likewise, the presumption of poverty driving criminality could be seriously challenged on sociological scales. Anyone undertaking such a task would be well-served to enlist the fields of Sociology and Criminal Justice departments in their work.

Another possibility for research is to investigate all possible parameters with DBSCAN and/or MESH and apply a visualization to the clusters generated based on just latitude, longitude, and quantity. This would mean going through and giving DBSCAN 28 dimensions of data to work under with around 6.7 million elements for just the crime set. The clusters generated from this would need to be projected into a graph purely based on geographic location to get a good visualization of how other aspects of our data effects the clusters.

Ultimately, the most interesting area of research worth pursuing would be a bridge between DBSCAN and Ruppert's algorithm for meshing. Essentially, a great deal of manual fine tuning work was needed on DBSCAN and subsequently the application of a meshing meta-algorithm to create the visual representations of these datasets. Finding a mathematical bridge between the two would reduce computational costs and make implementation significantly more efficient.

Commercial Applications

As a tool, the rudimentary and callow structure we designed, could serve as a basis for a functional purpose in commercial decision-making. Even the static, dated, and limited information of this research provides compelling evidence of preferential geography. This information is likely both useful and interesting to populations making commercial decisions, such as deciding on a place to eat. We can extend that analysis to include the business owners decision on where to open a restaurant. And ostensibly where populations chose to reside and congregate.

8 References

[1] Jyoti Agarwal, Renuka Nagpal, and Rajni Sehgal 2013. Crime Analysis using K-Means Clustering. International Journal of Computer Applications (0975 – 8887) Volume 83 – No. 4, December 2013.

[2] Mostofa Ali Patwary, Diana Palsetia, Ankit Agrawal, Wei-keng Liao, Fredrik Manne, and Alok Choudhary. 2012. A new scalable parallel DBSCAN algorithm using the diskjoint-set data structure. In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '12). IEEE Computer Society Press, Los Alamitos, CA, USA, Article 62, 11 pages.

[3] Gary L. Miller, Steven E. Pav, and Noel J. Walkington 2003. When and Why Ruppert's Algorithm Works. Only available online: https://www.researchgate.net/profile/Steven_Pav2/publication/2920965_When_And_Why_Ruppert's_Algorithm_Works/lins/556dfa9408aefcb861db95ec.pdf

[4] Junhao Gan and Yufei Tao. 2017. Dynamic Density Based Clustering. In Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17). ACM, New York, NY, USA, 1493-1507. DOI: https://doi.org/10.1145/3035918.3064050

[5] Qinghao Hu, Jiaxiang Wu, Lu Bai, Yifan Zhang, and Jian Cheng. 2017. Fast K-means for Large Scale Clustering. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17). ACM, New York, NY, USA, 2099-2102. DOI: https://doi.org/10.1145/3132847.3133091

[6] Anne L. J. Ter Wal, Ron A. Boschma 2008. Applying Social Network Analysis in Economic Geography: Framing Some Key Analytic Issues. The Annals of Regional Science, Volume 43, Issue 3, pp 739–756, July, 2008.
DOI: https://doi.org/10.1007/s00168-008-0258-3