NYC Green Taxis with Neural Network

Brittany Nicholls
Department of Computer Science
and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, Maryland 21250
Email: brn1@umbc.edu

***Abstract*—We propose a methodology for determining which of Baltimore's neighborhoods a new resident would very likely want to avoid. Specifically, focus is placed on using the datasets provided by official local government agencies through Open Baltimore Beta to accomplish this goal. Additionally, we ensure our approach to solving this problem is scalable so that it could support "big data" amounts of data.**

The basic idea behind our approach to determining which neighborhoods a new resident would need to be wary of is to 1) generate features based off of a new inhabitants concerns 2) determine which neighborhoods are similar to each other based on the features and then 3) configure which neighborhoods are least desirable without requiring human intervention. Using this methodology, we successfully identify 20 neighborhoods in Baltimore City that new residents should avoid.

I. Introduction

These days, when people think of Baltimore City, many individuals refer to The Wire, a TV show created by HBO that ran from 2002-2008.

addresses a small subsection of Baltimore's society. Furthermore, in the past few years, Baltimore has received national attention for events such as the death of Freddie

Freddie Gray's death at the forefront of society's recent memory, people attribute a negative reputation to the entirety of Baltimore. While it is not a "politcally correct" thing to say, Baltimore is like any big city. There are certain neighborhoods in Baltimore that are perfectly enjoyable and safe and at the same time there exist other parts of the city that most people will want to avoid living in unless they are familiar with the neighborhood.

Unfortunately, people who are moving to Baltimore do not have any easy, reliable resources that will recommend neighborhoods they might want to avoid until they understand different neighborhoods in the city operate. If someone starts googling to try to find this information, they will end up on one of three subpar resources: 1) Forums, 2) Maps that contain crime statistics or 3) News articles, such as

A city should not be avoided simply because it ranks high in a list of the "most dangerous" cities in the United States; it just means one should be conscious of what part of the city one lives in. The goal of this paper is to use Open Source Data to determine which neighborhoods a new resident should avoid living in, while making sure that the methodology can scale well to support large amounts of data. As more government agencies and cities are releasing data, this would become a "big data" problem, and the current approach would be able to be extended to other cities.

A. Motivation

The motivation for this paper is inspired by both of the authors experience with moving to Baltimore. Both were aware that Baltimore is like any other big city, in which there are parts one should avoid living in, but looking online does not provide any obvious guidance and they just wanted one easy resource that does more than look at crime.

Currently, resources available to people looking to decide where to move in Baltimore are to visit a bunch of different neighborhoods, to read forums where opinions wildly differ from one another and can be confusing, to look at crime statistics provided by online tools, or to try to piece together what neighborhoods news articles are talking about. Additionally, new residents have multiple questions they are asking as they search for a place to live that go beyond the crime in the neighborhood.

The list of questions this approach would help new residents answer are as follow:

1) What neighborhoods are relatively safe?
2) What neighborhoods are people actually living in?
3) What neighborhoods are clean?
4) What neighborhoods are being invested in?

II. Related Work

A. Crime Reports

Looking at the data produced by Uniform Crime Reporting Program for the FBI,

the only cities where there were more murders, and compared to Baltimore's population of 620K, they have a population of 2.7 million and 8.5 million people, only does Baltimore have a high number of murders, it also has a lower population. As can be seen with the FBI report, reading crime stats can be intimidating and do not paint the "full picture" of the city. Thus, Baltimore should not be avoided just because it has more crime in certain areas than other cities do. This paper is trying to identify neighborhoods that residents who are unfamiliar with the city might want to avoid, whereas this crime report is looking at the city as a whole.

B. Using Machine Learning to Understand the Impact of Mixing on Neighborhoods

Very recently, machine learning was applied to understand how the impact of mixing people with different socioeconomic backgrounds into one neighborhood would affect the neighborhood as aplied a new technique, kernel regularized least squares, to census data from between 2000 and 2010 in order to determine how the change in a neighborhoods characteristics affected the average income in that neighborhood. While this is an example of someone applying machine learning techniques to neighborhood data, they are looking to solve a very different problem. However, potential future work of this paper would include analyzing the neighborhoods over time in order to build a better recommendation system, and the approach proposed by Hipp, Kbe useful in this future work.

C. Baltimore Neighborhood Indicators Alliance

For the past 15 years, the Baltimore Neighborhood Indicators Alliance, of the Jacob Frances Institute, has released a Vital Signs report that relies on hard data about Baltimores Community Statistical Areas (CSA) in order to assess the quality of life in a CSA and work towards im-FJI Vital Signs are intended to be used in order to track the quality of life over time as well as provide hard data as valuable input into decision making when planning new programs and the future of t

While the BNIA-FJI work provides valuable insight into how a community is doing, the large areas that the CSAs cover do not allow for the more detailed approach

D. Unofficial Neighborhood Analysis

There are almost no published results of researchers analyzing neighborhoods; however, Ken Steif and others are doing urban spatial analysis and publishing informal white papers based on their ata. While predicting future home prices would be useful for a new resident who is looking to buy a house in Baltimore, this is too specific for our scope as we are trying to build a recommendation for any new resident, whether they are renting or buying. Again, this could fall into future work.

III. Data and Data Challenges

All of the data was pulled from official sources off of the Open Baltimore Beta

A. Victim Based Crime Dataset

The Baltimore Police Department posted an official dng Part 1 Victim Based Crimes from January 1, 2012 until April 8, 2017. The data stops on April 8, 2017 because that is the day of the last update of the data before the dataset was pulled down to be analyzed.

The crimes included in this dataset include:

- Aggravated Assault
- Arson
- Assault by Threat
- Auto Theft
- Burglary
- Common Assault
- Homicide
- Larceny
- Larceny from Auto
- Rape
- Carjacking (Robbery)
- Commercial Robbery
- Residential Robbery
- Street Robbery
- Shooting

The other information included in this dataset are approximate address, geocode, neighborhood, date and time, as well as a few details on whether the crime was indoors or outdoors and if a weapon was involved. The original dataset contains approximately 255K rows, which were filtered down to about 242K rows of data that were clean and unique.

It is important to note there are some limitations in this data due to legal concerns. The addresses provided in the dataset are approximate estimations of the true location of the crime and addresses that could not be geocoded were not included in the original dataset.

B. ECB Citations Dataset

The ECB Citations Dataset is provided by the Baltimore City Environmental Con Department, Department of Transportation, Department of General Services, Department of Public Works, Department of Housing and Community Development, and the Department of Recreation and Parks.

The citations are for various activities, violations, and problems that are not necessarily criminal and have to do with the community and common good. The violations range from trash accumulation and littering to issues with a building and yard, or even hunting/fishing violations. The dataset includes information such as the fine amount, balance amount, date of violation, agency writing the violation and violation description, as well as general information about the location of the violation.

C. Property Tax Dataset

Real Estate Property Tax information was provided by the Baltimore City Department of Find in the dataset. Fortunately, most of the data was pretty clean and included information such as the amount due in city and state tax, whether or not the property was a principal residence, as well as general information about the property. Most of the data that was filtered out was due to there being no tax information, but upon manual analysis it appears that most of those properties are either parks or seem to be vacant or condemned.

One thing to note is that a property was deemed either as a Principal Residence or Not a Principal Residence. A property is considered a principal residence if the property is owner occupied. So buildings that are not a principal residence include rentals as well as commercial property.

D. Vacant Buildings

A danformation contained in this dataset include the address of the vacancy, the date the vacancy notice was given, in addition to other information about the location of the vacancy.

This dataset was fairly clean and straightforward. One thing to note is that the date the vacancy notice was given is not necessarily the date the property was vacant. In one instance, looking at Google's Street View history for a property showed the property had been boarded up and was vacant years before the notice was given. For example, the house at 802 N Castle Street has as vacant date as of June 5, 2013; however, looking at Street View it is clear the house had not been lived in since November 2007 as it was boarded up.

E. Housing Permits

Baltimore Housing's Office of Permits and Building Inspections posted a dataset comprised of the submitted permits for both residential and commercial prhe permits are for things as simple as replacing a water heater, or for as difficult as constructing new medical buildings.

The data contains information such as the address of the property for the permit, a description and cost estimate of the work to be done, as well as general information about the location and permit. Approximately 370K permits were analyzed for the purposes of this paper.

F. Challenge of Working with Data

Since all of the data was provided through Open Baltimore Beta, it was fairly clean as it is meant for public consumption. One of the biggest problems encountered was data where some rows had important columns, such as the address, missing. As long as less than 10% of the records were missing valuable information, the dataset was used.

An additional problem faced is that the datasets were not thoroughly described on Open Baltimore Beta, although all of the datasets had easy to understand column names. In order to gain a basic understanding of what the datasets contained, each one had to be explored individually, with some additional help from outside resources in order to clarify some details in the data.

One of the datasets that really had to be cleaned up was the Part 1 Victim Based Crime dataset. In it, there were quite a few records that were duplicated. This was assumed to mean that multiple people were involved. For instance, there were about 50 duplicate

G. Other Datasets Considered

The Open Baltimore Beta iniative made other datasets available that could have been used in determining which neighborhoods new residents may not want to live in. Unfortunately, the following datasets were available, but were not used due to easy to find inaccuracies such as missing locations:

1) Grocery Stores: A potential list of grocery stores, including national-chain supermarkets as well as smaller local convenience stores.

2) Homeless Shelters: A potential list of homeless shelters, both emergency and transitional.

IV. Challenge of Determining Geography Method

There are a few different approaches that could be taken in order to deal with the geography of a city. The census breaks Baltimore city into 200 traddition, the BNIA-JFI previous research breaks the city into Community Statistical Areas which are clusters of census tr for the outlines of the CSAs. The reasoning for relying on the census tracts is the ability to measure the progress of an area over tunfamiliar with Baltimore's neighborhoods would prefer to avoid. Thus, the changes that occur in a constant section of a city do not matter for the purposes of this paper and thus, we did not continue with the use of CSAs.

An additional way to break a city in smaller chunks would be to look at zipcodes. Faccording to the population movement.

Thus, when working with new residents, breaking a city down into neighborhoods is the ideal approach . Neighborhoods are somewhat fluid and capture the changes that are occurring within the city. Neighborhoods change as their occupants move and as the buildings change. In addition, people naturally break a city down into neighborhoods, so it is a very intuitive way of breaking down a city. Lastly, neighborhoods are much smaller sections of the city, but large enough to capture trends especially as people who are similar tend to live near each other.

V. Identifying Neighborhoods to Avoid

By far, the largest challenge is finding hard evidence or research pointing to specific neighborhoods as places a person might want to avoid. It is not "politically correct" to point out which neighborhoods most people would want to avoid; however, it is common knowledge that every big city has areas that are not the best to live in if it can be avoided, especially if you are not familiar with the neighborhood and how it operates. One of the primary reasons why we avoid creating features based on demographics is so that the focus remains on objective features that would make a neighborhood "good" or "bad" to live in.

As pointed out in the Motivation section, forums reflect the opinions of locals in that a lot of people do not agree about a neighborhoods livability based on their own tolerances and preconceptions. Fortunately, there are a handful of neighborhoods that are notorious for not being the easiest to live in. These are the neighborhoods that this paper is trying to determine.

For instance, the neighborhood of Sandtown-Winchester has received a lot of attention as the place where Freddie Gray lived and was arrhermore, a Washington Post article also points out that Sandtown-Winchester is affected by its history with crack and is still s

A Baltimore Sun article discussed the issues within the neighborhood of Coldstream Homestead Montebello, aka the Chum. In the Chum, there are a lot of gangs who have more or

less spread the message of ”no snitchi Interestingly, this article also pointed out one of the positive effects that a project that could be the future work of this paper could help with:

> “But many residents believe gun violence defines the city more than it should, pointing to multibillion-dollar waterfront developments, national attractions and major league sports teams...police to politicians, know the high homicide rate threatens economic vitality and efforts to draw new residents. And they are scrambling to stop

With this in mind, our approach is considered successful if it identifies the neighborhoods of Coldstream Homestead Montello, Harlem Park, and Sandtown-Winchester, as well as neighborhoods that are similar to these neighborhoods, as neighborhoods that should be avoided by new residents.

VI. Naive Approach

The first reasonable approach one might take in trying to determine which neighborhoods to avoid living in would be to look at the victim based crime levels in the neighborhood.y the Baltimore Police Departm red.

While it would make sense that someone would want to live in a low crime area, crime does not paint the whole picture about the livability of a neighborhood. Many factors affect crime counts. In the case of Downtown, there are a lot of tourists, so crime is naturally higher. Frankford, Belair-Edison and Brooklyn are among the top populated neighborhoods in Baltimore, so it makes sense that with more people there is more crimes.

VII. Determining Livability

As shown previously, looking purely at crime is not the best way of determining which neighborhoods should be avoided as there are many reasons why crime may be higher in one neighborhood than it is in another. Since this method of determining which neighborhoods someone should live in is targeting new residents looking to move to Baltimore City, it needs to take into account multiple factors that would affect whether or not someone should live in a neighborhood. This method is also agnostic of cost of living and other factors that may be specific to a user.

Very generally, features are generated using the datasets discussed above. Apache Spark 1.6.2 is used to manipulate the data from Baltimore is small enough to be worked with on a laptop. After the features are generated, they are run through the K-Means clustering algorithm that is predefined in Spark in order to determine which neighborhoods have similar features. At this point, each neighborhood will be assigned to a cluster along with other neighborhoods with similar features. To determine which cluster represents the place where a new resident would not want to live, the average of the scores for each feature for each cluster is summed up and the cluster with the highest score is ranked as the least desirable neighborhood.

A. Feature Selection

As mentioned in the Motivation section, this approach is trying to answer the following questions:

1) What neighborhoods are relatively safe?
2) What neighborhoods are people actually living in?
3) What neighborhoods are being invested in?
4) What neighborhoods are clean?

Therefore, the features were created to allow neighborhoods that look similar to each other based on how they would answer the above questions to be clustered together.

Note that all of the below features are normalized from 0 to 100 to ensure that each feature has equal weight.

To clarify, active properties are all properties that are currently paying tax, regardless of whether they are owner occupied or not.

1) Exceptional Violence Crime Count: For each neighborhood, the number of victim based crimes that have been classified as a shooting or homicide are counted. Of the victim based crimes, shootings and homicides are are indicative of exceptional violence and potentially fatal, so they are deserving of their own feature. The number of exceptional violence crimes are not normalized against the population or active properties because this is the type of crime that most residents do not want to have in their neighborhood, regardless of how large the population is.

2) Resident-Centric Crime to Property Count Ratio: A potential resident would care about crimes such as burglary, auto theft, residential robbery, shootings and homicide because those could directly affect them and are crimes that are hard to deter. The count of the number of these crimes is normalized against the number of active properties because these crimes are more likely to happen in neighborhoods where there are more houses.

3) Vacancy Count to Total Property Count Ratio: A high number of vacancies in a neighborhood is very likely indicative of people leaving the neighborhood and properties not being properly maintained. Once again, the number of vacancies in a neighborhood is normalized against the number of active properties because some neighborhoods have more properties than others.

4) ECB Citation Count to Property Count Ratio: The number of ECB citations issued are indicative of how dirty and regulation-abiding the residents are. The number of citations is normalized against the number of active properties in order to take into account property counts in a neighborhood. Neighborhoods with more properties will have more chances for a violation.

5) Total Residential Property Tax to Total Non-Residential Property Tax Ratio: In order to calculate if a neighborhood has a large amount of rentals/commercial properties or is primarily owner occupied the ratio of the owner occupied residential taxes to the number of non-owner-occupied properties is used as a feature. This helps group neighborhoods together that are full of owner-occupied properties, or are highly commercial, or are full of rentals.

Cluster	Count	EVC	RD	V	ECB	Tax	PC
1	62	L	M	L	L	L	H
2	82	L	L	L	L	H	L
3	75	L	L	L	M	M/H	L
4	43	M	L	M	M	M	L
5	20	H	L	H	H	M	L

TABLE I
AVERAGE OF THE FEATURES OF THE NEIGHBORHOODS IN THE CLUSTERS

6) Permit Cost to Property Count Ratio: The ratio of the total estimated cost from permits to the total number of properties helps identify neighborhoods where large amounts of money are being invested into it. The total permit cost to active properties ratio helps normalize for neighborhoods where large amounts of work are being done against many properties.

B. Algorithm

1) Determine Neighborhoods that are Similar: Use k-means in order to group together neighborhoods that are similar to each based on the features explained previously. With this dataset, 5 clusters provided the ideal groupings of neighborhoods. Tathe cluster index, the number of neighborhoods in that cluster, as well as the average of the clusters' neighborhoods score for the respective feature.

Note that:

- EVC=Exceptional Violence Crime Count
- RD=Residential Crime
- V=Vacancies/Properties
- ECB=ECB Citations/Properties
- Tax=Residential Tax/Non-Residential Tax
- PC=Permit Cost/Properties

Based off of the average of the features of the neighborhoods that were grouped together, the 5 resulting clusters separated neighborhoods into:

Lookin are neighborhoods that are heavily commercial or are filled with rentals and a lot of work is being done on the neighborhood. In addition, the crime is relatively low and there are not a lot of ECB citations, so those neighborhoods are clean. Neighborhoods in this cluster include Downtown and Inner Harbor and a lot of the parks or industrial areas, which makes sense.

Cluster 2 groups together neighborhoods that are highly residential and pretty clean and safe. This includes neighborhoods such as Canton, Blythewood and Ednor Gardens-Lakeside that are relatively middle-upper class.

Neighborhoods in cluster 3 are a lot like those in cluster 2 in that there are a lot of owner-occupied properties, but there is a bit more crime and issues with ecb violations. These neighborhoods are more working class or up-and-coming and are fairly safe. For example, some of the neighborhoods in this cluster are Woodmere, Belair-Edison, and Hampden.

Cluster 4 neighborhoods have more violence and crime in them than the previous groups of neighborhoods and are neighborhoods that might want to be avoided at night. There are also quite a few vacancies in these neighborhoods and not as many active properties are owner occupied. Examples of neighborhoods in this cluster are Hollins Market, Mosher, and Easterwood.

Cluster 5 is made up of neighborhoods that new residents should avoid unless they know what they are doing. These neighborhoods make up a good portion of Baltimore's homicides and shootings. The residential crime is low, but that does not mean a whole lot since the number of vacancies in these neighborhoods are high. These neighborhoods are also in disrepair, given the high amount of ECB violations and low number of estimated costs from permits. Neighborhoods in this cluster include Sandtown-Winchester, Coldstream Homestead Montebello, Harlem Park, Mondawmin, and Broadway East.

2) Filter for the Neighborhoods to Avoid: While K-Means works well for grouping neighborhoods together, it does not build any sort of recommendation for which neighborhoods a new resident should avoid if possible.

In order to determine which neighborhoods to be wary of, there needs to be a way to measure which neighborhoods do not meet the positive requirements listed earlier. To do this, take the average of each feature for all of the neighborhoods in a cluster. Then, score each cluster based on the sum of the averages of the features; the cluster with the highest sum will be the neighborhood that a new resident would like living in the least. Note that this method only works if a higher score for a feature correlates to being a negative factor for a new resident. For instance, a neighborhood with a high amount of exceptional violence crime should have a feature score closer to 1, because people do not want to live in an area with a lot of exceptional violence crime. In order to meet these requirements, the averages of the scores for some features had to be inverted by subtracting the calculated average from 100 (which is the highest value from the normalization).

3) Results: By following the methodology described, the clusters described in Tab

1) Cluster 5, whose neighborhoods are colored red in Fig
2) Cluster 4, orange
3) Cluster 1, yellow
4) Cluster 3, yellow-green
5) Cluster 2, green

Thus, the number 1 group of neighborhoods to avoid are in cluster 5. The 20 neighborhoods in cluster 5 are:

- Boyd-Booth
- Broadway East
- Carrollton Ridge
- Central Park Heights
- Coldstream Homestead Montebello
- Druid Heights
- East Baltimore Midway
- Franklin Square
- Harlem Park
- Johnston Square
- Middle East
- Midtown-Edmondson
- Mondawmin

- Oliver
- Penn North
- Penrose/Fayette Street Outreach
- Poppleton
- Sandtown-Winchester
- Shipley Hill
- Upton

The final results were successful in identifying Coldstream Homestead Montebello, Harlem Park, and Sandtown-Winchester as neighborhoods new residents should try to avoid. Upon analyzing the rest of the neighborhoods, they have similar features to the 3 that were trying to be grouped together in that the exceptionally violent crime is high, the vacancy rate is high, the number of ecb violations is somewhat high, and they have lower amounts of property taxes being paid on owner occupied properties. These are all features that are indicative of a neighborhood that is not in the best shape and that someone should move to if they are familiar with the neighborhood.

4) Scalability for Big Data: Apache Spark is a tool that can handle big data in a distributed fashion. Because the algorithm was implemented using the built in K-Means, that section is perfectly scalable. In addition, the averaging of the features of the clustered neighborhoods fits perfectly well with the Map Reduce paradigm that Apache Spark relies. In the end, it would not be unreasonable to collect all 5 of the final cluster scores back to the driver in order to perform a local sort and map the cluster to the appropriate fill color, then broadcast the cluster-color mapping to the executors in order to label the neighborhoods with their appropriate color. All of the information could then be saved off to, say, Elasticsearch, at which point one could modify the neighborhood outline geojson file so that the neighborhoods get filled with their appropriate color.

Simply put, this algorithm would be more than capable of handling big data should this methodology be used on open source data for other cities.

VIII. Future Work

Possibilities for future work are discussed in this section.

A. Use Technique on Other Cities

Explore whether or not these features and this techniques translates well to other cities.

B. More Features

New residents have other concerns beyond what was discussed in this paper. Parents might want to live in a good school district and young adults might want to live near restaurants and bars. Adding the ability to not only identify which neighborhoods should be avoided, but to actually recommend neighborhoods would an interesting problem to explore. Additionally, importing other datasets would allow for a more diverse and larger set of features. However, both of these additions would require acquiring and analyzing more data, which is a long process.

C. Analysis of Neighborhoods Over Time

Many of the datasets had over a years worth of data. Some future work could look at examining the neighborhoods over the past few years in order to identify neighborhoods that are becoming a more desirable place to live, or are becoming a less desirable place to live.

D. Use Technique for Other Use Cases

The primary context for this approach was to identify neighborhoods a new resident would not want to live in. Could the same technique be used, but change to a different context. For example, could neighborhoods be identified that are about to have a large population change due to deaths of the elderly or births.

IX. Conclusion

In this paper, we introduce a technique that will identify neighborhoods in Baltimore City that a new resident would want to avoid living in. We successfully identified a group of neighborhoods similar to Coldstream Homestead Montebello, Sandtown-Winchester, and Harlem Park, which were our targeted neighborhoods as they are not desirable neighborhoods to live in at the moment. Furthermore, we succeeded in choosing an approach that is easily scalable so that it could handle big data as more and more cities publish open source datasets that could be used in applications like this one. Ultimately, we generated a map of Baltimore, as seen in that a new resident can easily identify which neighborhoods they want to shy away from while looking for somewhere to live.