Diagnosing Chicago's Noises with Ubiquitous Data

Lei Guo
Department of Computer Science
New York University
lg2681@nyu.edu

Zhichao Yu
Department of Computer Science
New York University
zy948@nyu.edu

Jincong Zhu
Department of Computer Science
New York University
jz2668@nyu.edu

Oukan Fan
Department of Computer Science
New York University
of394@nyu.edu

***Abstract*—Noise pollution situation in modern cities is getting worse. People suffer from it on working efficiency and mental health. 311 Services is a platform where citizens in major cities can complain about their daily issues, and noise is the third largest category of complaints in the platform. Each noise complaint contains information of location, date, and comments which could be used to identify the noise type. Through analytics on complaints data combining with check-ins data, traffic and POIs data, we modeled a noise map of Chicago based on noise locations. For each location, the type composition of noise complaints are also analyzed. We initially partitioned Chicago into about 360 small areas, however, due to the limit of the complaint dataset, we only got complaint counts for over 140 areas. By using MLlib in Spark, we investigated the correlations between noise complaints data and the other three datasets. Further, we filled in the missing noise complaints data, and inferred complaint counts for additional 23 areas. Finally, we derived the inferred dataset and visualized the map for noise distribution of Chicago. We also outline valuable insights for noise pollution, point out possible actions for pollution restraints and mention some of the study limitations.**

***Keywords*—Hadoop, Hive, Spark, noises analysis.**

I. INTRODUCTION

The rapid progress of urbanization has introduced severe noise pollution. In major cities like Chicago, noise pollution has become a major problem that not only affects the sleep quality and work efficiency, but also damages the mental health of citizens.

There's an increasing demand especially from citizens living in major cities on analytics which can model the citywide noise situation and the composition of noises based on different types, places or even dates. However, there are several challenges for citywide noise analytics. First, there are still not enough sensors distributed in major cities collecting actual noise data in decibels. Moreover, even though the data of noises measured in decibels could be acquired, it is still not enough to determine the actual situation of noise pollution, for example, people's tolerance to noises may changes over locations. Thus, the levels of noises varied by locations and noises' decomposition are worth investigating. So in order to diagnose citywide noise situation, single source of datasets is definitely insufficient.

Despite the fact that diagnosing urban noise situation is very complicated, many ubiquitous data sources are available and useful for analyzing urban noises. For example, In 311 Open Data platform, noise complaint is among the top three categories of citizens' complaints. Thus, we chose the Environmental Complaints dataset from Chicago Department of Public Health as our main data source for analytics. However, the complaints data is rather not sufficient and missing complaints data in some areas. So, we decided to introduce other three datasets: POIs, Check-ins and Traffic.

After appropriate preprocessing (ETL) of the four datasets, we stole the idea from Finite Element Analysis (FEA) and partitioned Chicago into about 360 small areas which are all identical rectangles. This had brought so much convenience for us to investigate the noises' geographical distribution. Also, we identified the type of each noise complaint by matching keyword in the comment field of the noise complaint dataset.

Further, we made use of the Linear Regression model with SGD in MLlib, Spark to illustrate the correlations between noise complaints data and the other three datasets. By making predictions based on the model, we filled in the unconvincing complaints counts in areas where the data in other three datasets are available.

Finally, we mapped the noise distribution mainly by noise complaints counts in different areas. And in a certain area, the type composition is also investigated.

For future work, we would like to analyze the noises distribution over dates. For example, we may divide the various dates information in each dataset into "weekdays" and "weekends" or "holiday" and "non-holiday".

II. MOTIVATION

Chicago, as one of the world's largest cities, has suffered from severe noise pollution for many years. Chicago City has

been taking actions in curbing noise pollution, an example could be the 311 services platform, which receives citizens' requests of services via mobile apps or phone calls. It has been found that the noise complaint is among the top three categories of complaints.

However, since there is still no detailed citywide analytics on noise pollution in Chicago, the progress on curbing pollution is rather slow and the effect from actions taken by relevant departments may not be desirable. For example, if we could know that a certain place in a certain duration of time is suffering most from some certain types of noises, then the city planners could take measures to deal with it much more efficiently.

In addition, many traditional noise detection methods require noise levels and ranges to be collected by sensors from road traffic, railways, major airports, industries, and venues covering, which is expensive. Many other traditional noise detection methods are also difficult to apply, and the level of noises in different locations and times may change dramatically. Also, people's tolerance to noises could also change over places and times. For instance, people's tolerance to noises in commercial districts is usually higher than their tolerance to noises in residential areas. Besides, many common sensors are not able to distinguish the composition of noises, which is sometimes important to tackling the noises. Thus, modeling a city's noises by analyzing ubiquitous datasets is not only a lot less expensive, but also closer to reality. The noise distribution varied by locations and the composition of noises in different areas can be acquired through our analytics.

III. Related Work

There are currently a few number of researches on the analytics of citywide issues in major cities, such as New York City or Beijing. Research by Lasse Korsholm Poulsen [1] has conducted data analytics on ride records for Green cabs and Uber in the outer boroughs of New York City. The research has shown that the performance of Green cabs in isolated zip codes differ significantly, and that Uber is growing faster than Green cabs in general and especially in the areas close to Manhattan. It provides a market insight for NYCTLC of what actions, if any, could be taken to preserve cabs' market share. We were also inspired by Jing Yuan's research on diagnosing regions of different functions in Beijing using POIs and human mobility [2]. The evolving pattern of Beijing could also be discovered. I. Schweizer et al. [8] propose methods and prototypes of noise maps based on participatory sensing, which is a solution to improve the data analytics for noise. Applying participatory sensing can make it less costly to create real-time noise maps. M. Gallo et al. [9] propose a model for estimating the noise produced by road traffic in urban areas. The model is proven valid in a typical middle-size city.

In addition, we studied the sequential POIs recommendation research conducted by Jitao Sang[10], which provides the user with recommendation of plan of activities by analyzing the user's personal check-in data and current status. And the probabilistic of recommendation applied Markov chain to value the probability of transferring from one POI to another. The correctness was proved by experimenting on massive user check-in data from social media. Another research, which is conducted by Piotr Mioduszewski[11], supported our work as well. The study was reported as a paper, which introduced the comparison of two assessment methods for environmental noise. One of them uses the concept of "Dynamic noise maps" while the other implemented "Static noise maps". The measurement station is equipped with microphones and power supply is provided by commercial telephone lines. Factors like daytime, nighttime, traffic, temperature are included in the study.

The Geo-Social media model developed by Hsun-Ping Hsieh[12] provided a more complex analysis towards noise pollution of New York City. They collected geo-social media data: geographical information (venue and location attributes), mobility footprints (check-in data), visual snapshots (consumer-contributed photos), and social interactions (social network). They used NYC 311 data for model training and inference validation and developed the model of Urban Noise Diagnotor (UND). There are four main multimodal features: Geographical Features are used to describe the category and interaction, Mobility Features are devised to describe mobility, visual features and social features.

The noise map implemented by Yilun Wang[13] revealed a visualized noise pollution. They highlighted two important components: location ranking and noise composition analysis. The map contains two main parts, the left is control panel which we can control specific data and time and also shows top 5 noisiest location list and results of noise composition analysis, the right is noise map of New York. Location Ranking shows noisiest regions in a specific setting. Noise Composition Analysis is used to take insight of the noise composition and main noise categories and distribution.

IV. Design

Our design patterns are shown in Fig. 1.

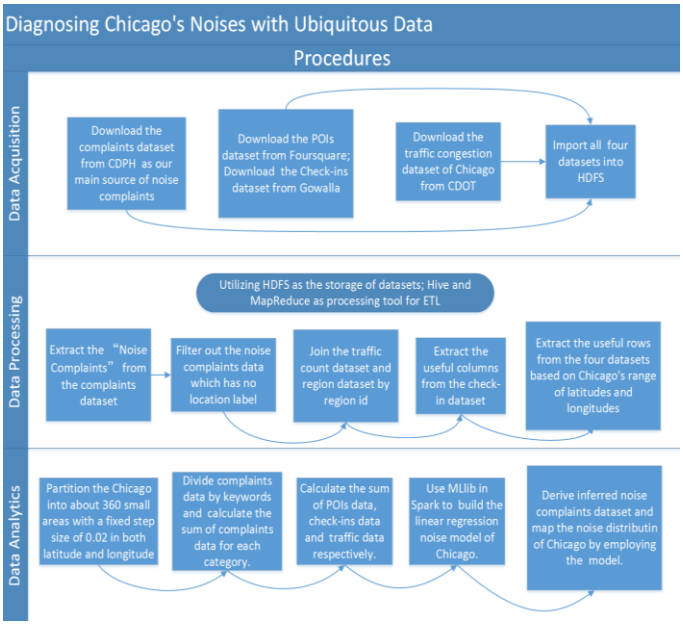


Fig. 1. Design diagram

A. Data Acquisition and Dataset Description

1) Environmental Complaints Dataset:

It contains a total of 152,795 environmental complaints since the year 1993. In addition, each row has fields: complaint id, complaint type, mapped location, street information, complaint date, complaint details and modified date, etc.

2) Check-in Dataset from Gowalla:

Our check-in dataset contains a total of 6,442,890 check-ins worldwide collected from Gowalla in the period Feb. 2009 - Oct. 2010. Each record has fields such as check-in time, mapped location, and user id, etc.

3) Points of Interests Dataset:

The POIs dataset was collected from Foursquare and has over 3,680,000 records. It contains venue data worldwide with 7 columns, including: venue id (Foursquare), mapped location, venue category (Foursquare), country code, etc.

4) Historical Traffic Congestion Dataset:

This dataset contains the historical estimated congestion for the 29 traffic regions in Chicago. It has total lines of 3,179,966 records since Jan. 18 2013 and each record has fields such as recorded time, region id, number of reads, etc.

B. Data Pre-Processing

Since the four datasets are unformatted and impossible for us to conduct analytics on, first we should clean up and format the datasets (ETL stage). In this stage, we mainly utilized analytic tools: MapReduce, HDFS and Hive.

1) Complaints Dataset Pre-Processing

- Extract the "noise complaints" from all types of evironmental complaints. We mainly used MapReduce as our processing tool in this step;
- Discard the noises records without location information so that the dataset could serve as the basis for our noise map. We mainly used MapReduce and Hive in this step;
- Extract four columns we want to conduct analytics on by using Hive, the columns include: latitude, longitude, date and complaint details.
- Since the noise complaints data in some remote areas is not convincing, we decided to focus on certain area with latitudes in the range [41.637916, 42.023684] and longitudes in the range [-87.862653, -87.522077]. Thus, we filtered out the noise complaints which were not reported in our interested range of areas. We mainly used MapReduce in this step.

2) Traffic Dataset Pre-Processing

- Inorder to replace the region IDs with its corresponding range of latitudes and longitudes, we utilized Hive as our processing tool to join the dataset with another region ID dataset on Region Id.
- By using Hive, we extracted three columns we want to conduct analytics on: (latitude range, longitude range), time and number of reads.
- Finally, we also cleaned the dataset to make sure that latitude and longitude of each record falls in the range [41.637916, 42.023684] and [-87.862653, -87.522077]. We mainly used MapReduce in this step.

3) Check-in Dataset Pre-Processing

- Clean the dataset to make sure that latitude and longitude of each record falls in the range [41.637916, 42.023684] and [-87.862653, -87.522077]. We mainly used MapReduce in this step.
- Filter out non-useful columns such as user ids by using Hive.

4) POIs Dataset Pre-Processing

- Clean the dataset to make sure that latitude and longitude of each record falls in the range [41.637916, 42.023684] and [-87.862653, -87.522077]. We mainly used MapReduce in this step.

C. Data Analytics: Tools and Methods

With our formatted and cleaned datasets, then we utilized MapReduce, Hive and Spark for further analytics.

First, we stole the idea from "Finite Element Analysis", and partitioned Chicago into about 20 × 18 small areas which are all fixed sized squares and each area has a 0.02 span in both latitude and longitude. And most importantly, we attribute all the data which falls in a certain area to the center of that area and labeled each area with a unique id in the form (i, j). The pattern of our partitioning method is shown in Fig. 2.

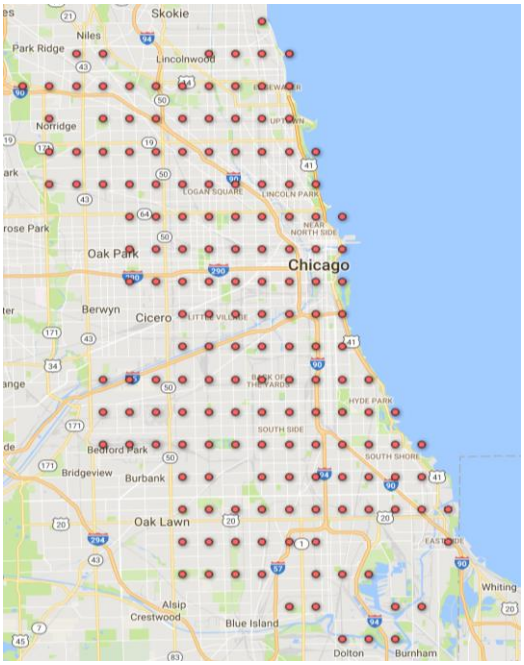


Fig. 2. Partitioning pattern

With the idea above, we first conducted analytics on our first dataset, the noise complaint data:

1) Complaints Dataset Analytics

- Through keyword matching in the field "complaint details", we categorize the noise complaint records into totally 8 categories: "Music/Party"; "AC/Ventilation"; "Banging/Pounding"; "Traffic"; "Construstion"; "Repair/Refurbish"; Hammer/Jack; and "Other". We mainly used MapReduce in this stage.
- Calculate the total counts of noise complaints for each area. Since we would like to get insights on type composition for each area of interest, we also calculate the number of records for each type as well as their percentage. Because the keyword matching is needed, we mainly used MapReduce as our analytic tool in this stage.

The analytics results we got from the step above have introduced us new challenges. We only fetched 174 lines of records for the 360 areas. One possible reason is that the 311 services might not cover the remote areas in Chicago. Even though it might seem sufficient since we only care about the noise distribution near the center of Chicago, complaints count in about 30 areas are rather small and unconvincing. And we could not simply discard these records since they take up about 20 percent in our total records.

We came up with an idea, that is suppose we could find the correlations between complaints counts and the other three counts (number of traffic; check-ins; POIs) in other three datasets, then we could infer the complaint counts in the missing areas and areas where complaints counts are unconvincing, as long as all the data we need in the other three datasets are available.

Finally, we decided to make use of Spark as our tools for further analytics on correlations. After looking into the records in the four datasets carefully, we found out that for a certain area, the complaint count is positively correlated with the other three. Thus, we decided to use Linear Regression to construct the correlation model. Fortunately, Spark provides an Machine Learning library which is called MLlib, and we found that the "Linear Regression with SGD" model is the most suitable for our analytics.

Before we could train the model, we should prepare the four datasets in the formats that the Linear Regression model in MLlib accept. The training and testing datasets should both contain only counts and we must make sure they are "joined" according to their unique area id. We used Hive and MapReduce to prepare the data:

2) Prepare the dataset for Linear Regression

- Use Hive to perform an inner join among taffic, check-ins and POIs datasets. This step output a table which stores useful information only for areas where all the three counts are available.
- Filtered out the unconvincing records in the complaint dataset where complaint counts are below 4, which takes up about 15% in our total records. This step was also processed using Hive and it output a dataset where all the complaint records are convincing. This dataset contains 144 lines of records and let's name it dataset 1.
- Use Hive to perform an inner join among all four datasets we currently have. This step output a dataset where each record contains valid data for all four fields. The dataset only contains 121 lines of record and name it dataset 2.
- Use Hive to perform an appropriate outter join among all four datasets we had after step 2. This step output a dataset which has "NULL" as complaint counts for some areas indicating that a much more convincing counts needs to be infered from our correlation model. The dataset contains 145 lines of records and name it dataset 3.

The two datasets dataset 2 and dataset 3 would be used for further Linear Regression and complaint counts predictions, respectively. Each line only contains four fields: total number of complaints, total reads of traffic, total number of check-ins, and total number of POIs.

3) Linear Regression using MLlib in Spark

- Train the "Linear Regression with SGD" model by a random 80% split of the dataset which contains all convincing data. Since the model is very sensitive to stepsize, so we were really careful about choosing parameters for the LinearRegressionSGD() method.
- After choosing appropriate stepsize and number of iteration, we finally obtained a model with a RMSE of 16.749, which we believe is accurate enough for predictions of complaint counts.

Suppose that u stands for the total counts of noise complaints in a certain area (i, j), and x, y, z stand for traffic count, check-ins count and POIs count respectively. Then the Linear Regression model we derived is shown in the equation (1) below:

$$u = 1.000044 + 0.029078x + 0.032768y + 0.020688z \quad (1)$$

Notice that all the coefficients in the equation are rounded to 6 decimals.

4) Complaint count inference

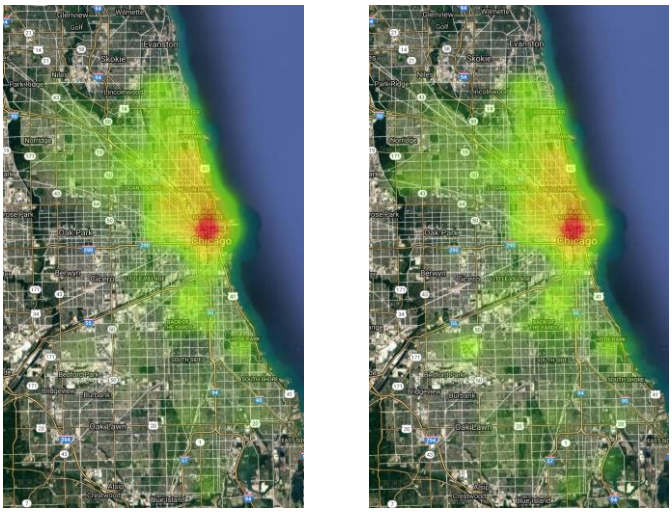
- After we obtained the equation, we used MapReduce to replace the unconvincing complaint counts with the predicted ones by employing the equation. Notice that in this step, the MapReduce job only output the predicted records.
- Finally, we obtained our complaint dataset which includes inferences after utilizing Hive to union the predicted records with dataset 2.

The final dataset we fetched contains totally 168 lines of records. And surprizingly the number of records in the original complaint count dataset is 174, which is very close to 168 in our final dataset. But remember that the original complaint count dataset has 30 lines of unconvincing records.

V. RESULTS

After our pre-processing of raw data, we ruled out the invalid data rows, such as the rows missing key information for our analytics. Also, we filtered out the unrelated data columns such as user-ID and date. Besides, we matched each area with the corresponding latitude and longitude for accurate analytics. Before the final step, we derived the linear regression model by passing the convincing training data into the model. In the end, we predicted the noise level of certain areas in Chicago where noise complaint data are missing.

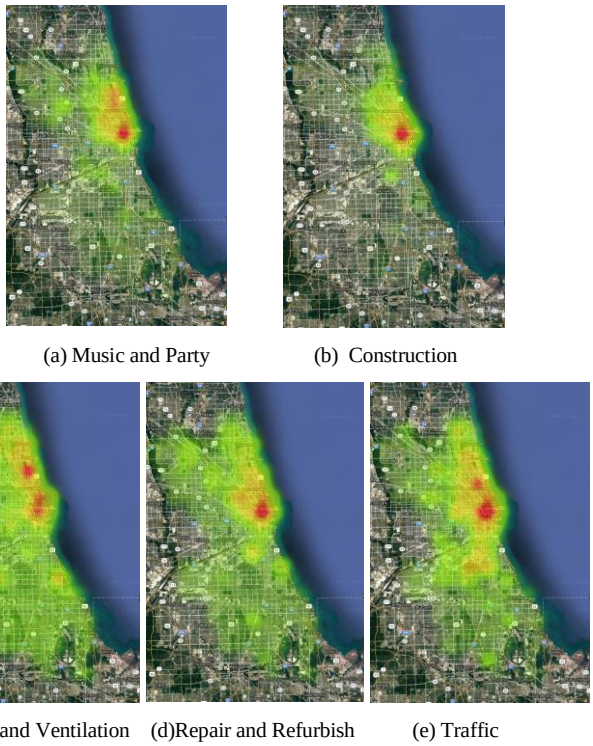
Our final inferred complaint dataset, which describes the noise distribution of Chicago, is shown below. By comparing the figure showing the noise distribution based on the original noise complaints data and the figure showing the noise distribution based on the inferred noise complaints data, we can tell that there are parts of the city where the noise level is believed to be higher than the level the complaints data represents. For example, the report of noise in the southwestern part of Chicago is not reflecting the real noise level according to the prediction of our model. By investigating the noise level of the area in Chicago, it is believed that the prediction of our model is convincing. In this way, government departments, firms or individuals may make better decisions with concerns about the noise level of the specific area.



(a)Noise map without inference (b)Noise map with inference

Fig. 3 Noise maps

Besides, we analyzed noise distribution level of certain type of Chicago, so that we can compare the relative noise level of certain type of different areas. We selected five typical types of noises for comparison.



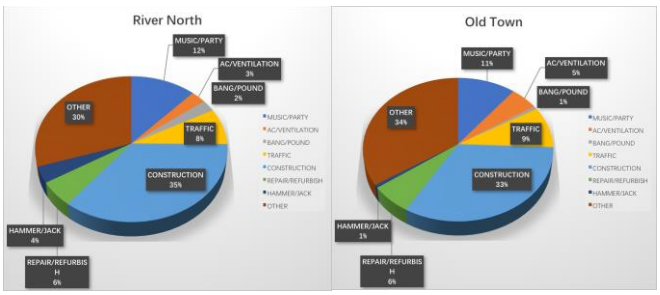
(a) Music and Party (b) Construction

(c) A/C and Ventilation (d)Repair and Refurbish (e) Traffic

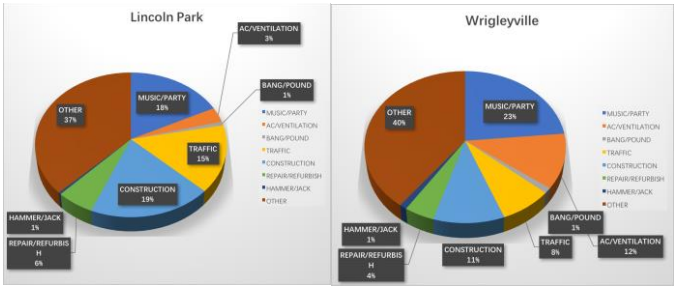
Fig. 4 Noise maps of main types of noises

Figure 4 shows the heat maps of the noise distribution in Chicago regarding the noise indicator in five different types. Among these, the heap maps for the category of traffic and air conditioner/ventilation show that there are specifically severe pollution in Near North Side and Lincoln Park. The communities affected by traffic noise also include Uptown and Back of the Yards, which are basically Chicago's central business district. The music/party noise heat map and the refurbish heat map both represent high level noise in Near North Side, which is a core entertainment district in Chicago. The difference between these two heat maps matches the fact that, unlike the music/party noise, the refurbish are also affecting people in Chicago suburb. The construction heat map basically matches the fact that most districts in Chicago has been highly-developed, therefore the high-level construction noise clustered around River North.

Furthermore, we analyzed six representative locations of Chicago to analyze their noise type composition. As the figures below illustrate, construction and music/party are the leading sources of noises.

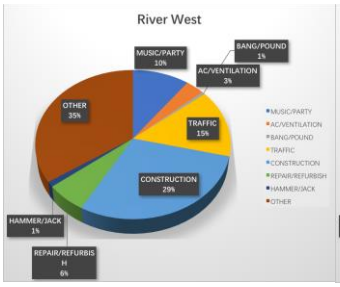


(a) River North (b) Old Town

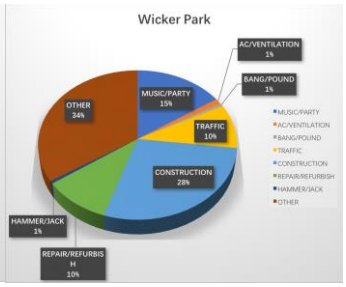


(c) Lincoln Park

(d) Wrigley Ville



(e) River West



(f) Wicker Park

Fig. 5 Type composition of typical locations

VI. Future Work

Our work can be improved in mainly three areas.

First, we did not split Chicago into areas in a more reasonable way, such as dividing Chicago based on the bound of communities or geographical features. Given the spreading features of noise, in this paper, we just simply split Chicago into equal hundreds of areas.

Secondly, we did not analyze the change of noise level over time and date. Considering that areas with more POIs inside normally receive more noise complaints in weekends and holidays, research based on time could reveal the noise distribution at the time dimension.

Finally, the types of noises should be analyzed so that more specific categories of noises can be the target of our work. Due to the limit of lack of natural language processing tools and the poor description in the noise complaints detail, the "other" noise category takes too much percentage in our work. By analyzing the details of the complaints with NLP tools, more specified categories of noises can be derived.

VII. Conclusion

In the process of trying to predict the complaint counts for areas where the complaint counts are missing or unconvincing, we derived a Linear Regression model using MLlib. Our final model has a RMSE of 16.74. Since the data range of our final noise complaint dataset goes from 0 to over 540, we believe the model of our research is accurate enough for predictions of complaint counts. As for the accuracy of the type composition of our work, according to news report of press, the annoying construction noise is damaging the quality of life of residents around the neighborhood of Wrigley Ville [14].

In this way, the model we derived for predicting the noise complaints in the areas where the data is missing, and the type composition analytics of various areas is convincing.

Acknowledgment

We would like to acknowledge Professor Suzanne K McIntosh and the TAs of "Real-time and Big Data Analytics" course for giving us support on selecting project topic, fixing technical issues and the techniques in analytics. We would also like to acknowledge the technicians of NYU HPC for the guidance on using the Hadoop cluster, Dumbo and Mercer.

References

[1] Lasse Korsholm Poulsen, Daan Dekkers, Nicolaas Wagenaar, Wesley Snijders, Ben Lewinsky Raghava Rao Mukkamala and Ravi Vatrapu. Green cabs vs. Uber in New York City. Big Data, IEEE International Congress on, June 2016.

[2] Jing Yuan, Yu Zheng and Xing Xie. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.

[3] City of Chicago, Data Portal. https://data.cityofchicago.org/Environment-Sustainable Development/CDPH-Environmental-Complaints/fypr-ksnz

[4] T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.

[5] A. Gates. Programming Pig. O'Reilly Media Inc.,Sebastopol, CA, October 2011.

[6] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implemenation, 2004.

[7] C. Olston, B. Reed, U. Srivastava, R. Kumar, A. Tomkins. Pig Latin: A not-so-foreign language for data processing. In proceedings of SIGMOD, June 2008.

[8] I. Schweizer, R. Bartl, A. Schulz, F. Probst, M. Muhlhauser. NoiseMap - Real-time participatory noise map. In Proceedings of 2nd International Workshop on Sensing Applications on Mobile Phones (PhoneSense '11), 2011.

[9] M. Gallo, O. Mascolino, G. Mazza. A Model for Estimating Road Traffic Noise in Urban Areas. Environment and Electrical Engineering, IEEE 16th International Conference on, June 2016

[10] Jitao Sang, Tao Mei, Jian-Tao Sun, Changsheng Xu, Shiping Li. Probabilistics Sequential POIs Recommendation via Check-In Data. Proceedings of the 20th International Conference on Advances in Geographic Information Systems, 2012.

[11] Piotr Mioduszewski, Jerzy A. Ejsmont, Jan Grabowski, Daniel Karpin´ ski. Noise map validation by continuous noise monitoring. Applied Acoustics, Elsevier, 2011.

[12] Hsun-Ping Hsieh, Tzu-Chi Yen, Cheng-Te Li. What Makes New York So Noisy?: Reasoning Noise Pollution by Mining Multimodal Geo-Social Big Data. Proceedings of the 23rd ACM international conference on Multimedia, 2015.

[13] Yilun Wang, Yu Zheng, Tong Liu. A Noise Map of New York City. Proceedings of the 16th ACM International Conference on Ubiquitous Computing, 2014.

[14] https://www.dnainfo.com/chicago/20150806/wrigleyville/neighbors-say-wrigley-field-construction-getting-too-noisy