Prediction of Natural Gas Leak Events in New York City from Open Data

Christopher Prince, Nathan Weber, Dara M. Perl, Daynan Crull
NYU Center for Urban Science & Progress

***Abstract*—The dangers of natural gas distribution and aging infrastructure are well known, but determining the risk for incidents related to natural gas is a difficult statistical, engineering, and civic challenge. Furthermore, investigating and repairing infrastructure can be prohibitively expensive. These authors use publicly-available data and compare several statistical methods to identify an optimal predictive model. Given a broad approach to feature selection, a naive prediction based only on historical leaks performed better than more complex models, such as regularized linear regression, random forests, and multi-layer perceptron neural networks; however, these more complex models did demonstrate the ability to detect complex structures in the data, improve feature selection and possibly address more precisely defined objectives, such as prioritizing inspections over select geographic areas.**

I. Introduction

In New York City, there were more than 60,000 emergency calls made by the New York City Fire Department (FDNY) related to gas leaks between 2013 and 2015. While most of these calls resulted in no damage or injury, several incidents did end with fatalities. Infrastructure in Manhattan leaks three to five times more natural gas than cities with newer infrastructure [1] and of the 6,400 miles of gas main lines running under New York City's streets, 53% were installed prior to 1960. In 2012, Con Edison experienced 83 leaks for every 100 miles of gas main. Furthermore, replacing a main in NYC can cost from $2.2 million to $8 million per mile, so prioritizing investment is critical. [2]

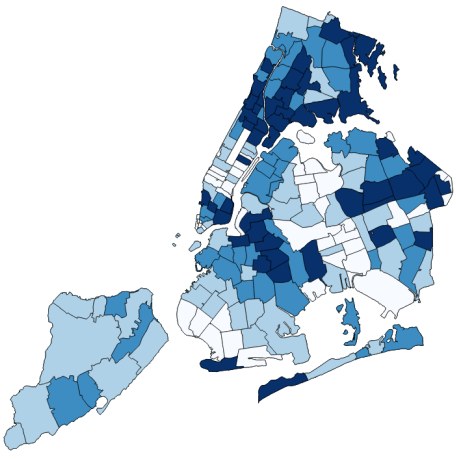

Figure 1. 2015 gas leaks per building unit. Coloring displays four quantiles with dark blue displaying the top quarter of gas leaks per building zip code, and white displaying the lowest quarter of gas leaks per building zip code.

While there have been successful attempts to use statistical analysis on large data sets to assess building risk in general, no analysis has been completed on gas leak data to date. In 2013, the New York City Mayor's Office of Data Analytics (MODA) worked with the Fire Department of New York (FDNY) to improve its Risk Based Inspection System. The system is designed to prioritize the inspections of buildings in the jurisdiction of FDNY so as to aid in the detection of severe violations. Prior to MODA's involvement, the FDNY model was based on limited data about the city's buildings and weighted simply by anecdotal evidence from firefighters. The improved model uses a statistical regression based on additional building data sets to increase the efficacy of inspections. The model performance went from not being much better than random selection to finding nearly three-quarters of severe violations in just the first one-quarter of inspections. [3]

In a similar fashion, these authors wish to determine if regression, or another machine learning method, can be used to predict the locations of natural gas leaks as reported to FDNY. The authors investigate if an effective predictive model for NYC natural gas leaks can be developed using only publicly available data sets and use various data selection strategies and machine learning methods to develop multiple models. Comparisons of the model performances with regard to their ability to predict 2015 gas leaks based upon 2013–2014 data results in the identification of some challenges and opportunities for further research.

II. Data and Methodology

A. Data

The primary data source is a dataset of incidents responded to by the New York City Fire Department (FDNY) between 2013 and 2015, as accessed from the New York City Open Data Portal. Incidents in the data categorized as a gas leak related to natural gas or liquid petroleum gas (LPG) are aggregated at the zip code and census tract level. In this data set, basic geographic identifiers including the zip code and road segment identifier (name) are provided. This data was integrated with data sets of complaints (929,000 records through 2015), violations (1.85 million records through 2015), and work permits (2.43 million records through 2015) from the NYC Department of Buildings (DOB) as well as the NYC Primary Land Use Tax Lot Output (PLUTO) data (2015 release, 850,000 records). Each of these datasets are available publicly through NYC Open Data or NYC.gov. Finally, demographic data from them American Community Survey (ACS) was used. Once integrated and inspected for a rough manual vetting of appropriate features, this integrated dataset yielded 735 features.

B. Allocation of gas leaks to geospatial boundaries

To use the geographical data for this research, the number of leaks per zip code were calculated for years 2013, 2014, and 2015. This zip code aggregation created 195 data points, one for each zip code. The authors also associated the gas leak data at the census tract level to improve the granularity of the analysis. The FDNY gas leak data contains zip codes but does not provide a specific street address location for each incident, nor census tract geographic identifiers. The road name and zip code from each incident was used to aggregate the data to a specific road segment. Next, a shapefile of all New York City roads was analyzed to create one or more road segments for each road in New York City. If a road was located in a single zip code, only one segment would be created. However, if a road crossed two or more zip codes, the road segment would be split at each zip code boundary, resulting in the number of road segments equal to the number of zip codes the road crossed. For each newly created road segment-zip code feature, a small buffer was created to allow for geographic association.

When the procedure was completed, 94% of the road name-zip code aggregations were associated with the road segment buffers mentioned above. The buffers were then overlaid on census tracts for New York City's five boroughs. A geo-spatial intersection was performed to determine the number of census tracts intersected by each road segment buffer. The gas leaks reported on a given segment are then allocated evenly among each of the census tracts with which its buffer intersects. Finally, summing the census tract frequencies results in the total number of gas leaks per census tract for the years of 2013, 2014, and 2015.

An example of this operation can be seen in Figure 2. Since multiple road segments can pass through the same census tract, there will be multiple frequencies associated with each tract. This aggregation process results in 2,163 census tracts having a gas leak count greater than zero. Each labeled census tract constitutes a data point in the analysis.

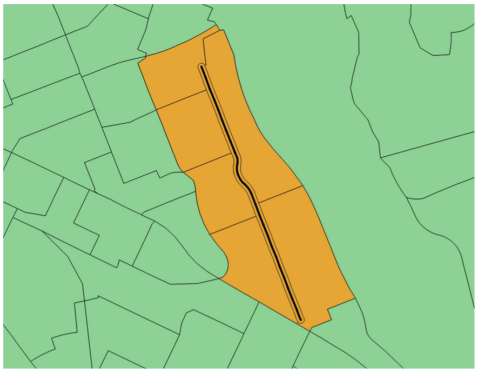

Figure 2. A segment of 110th Street, Forest Hills, NY 11375. The heavy black line denotes the road segment. The yellow shapes represent the buffer around 110th Street and the census tracts with which it intersects, and the green polygons are census tracts not intersected by the 110th Street buffer.

C. Methodology

The research idea is to investigate whether analytical models that are trained and calibrated on 2013–2014 data will more accurately predict 2015 gas leaks than a naive guess, using only historical trends to predict the frequency of future incidents. The working null hypothesis is that there will be no significant improvement with the given model than a naive method. For the naive method, the authors predicted that the average number of actual gas leaks (per zip code or tract) in the 2013–2014 data would remain constant. This annual average would then be the predicted number of gas leaks in 2015. The target prediction value is the number of leaks per building unit (per zip code or tract) in 2015. For final model evaluation, the Root Mean Squared Error (RMSE) of the predicted leaks per building unit relative to actual values was used:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_i (\hat{y}_i - y_i)^2}$$

with $\hat{y}_i$ and y_i being the predicted and actual values for the ith data point, respectively, and N being the number of observations.

For the zip code level data, the authors compared predictions to actual number of leaks; however since that data was not available at the tract level, they compared predicted leaks to the number of leaks calculated using the road segment approach described above. RMSE is scale-dependent, so while it would not be appropriate for comparing the effects of non-scaled variables, it is appropriate for comparing model performance on the same dataset. [4]

Several models were implemented: linear regression, linear regression with Lasso (L1) and Ridge (L2) regularization. Top features were identified through multiple techniques, including linear correlation, lasso regularization, maximal information coefficient (similar to linear coefficient, but can identify non-linear relationships), recursive feature elimination, and ridge regularization. A random forest model and a simple multi-layer perceptron (MLP) neural network were also implemented. All of the models were generated in python, using the scikit-learn machine learning packages, [5] except the neural network which was developed using tensorflow and its python bindings. [6] The codes for these models are available in the public code repository accompanying this paper. [7]

Each model was initially trained on 2013 data, using 2013 gas leaks as the target variable to properly weight model parameters. They were then cross-validated by predicting 2014 gas leaks using 2013 features, with the assumption that a true predictive model would have only the previous years' data to predict the next year's leaks. In this manner, each model was optimized by tuning hyper-parameters (such as specific features selected with linear regression or the learning rate and number of hidden units in the neural network). Once sufficiently optimized, each model was tested by predicting 2015 gas leaks using 2014 features and compared based on overall RMSE.

III. Results

Tables I and II and Figures 3 and 4 give visual and tabular performance indicators for each of the models. At the zip code level and the census tract level, the naive model boasted the lowest overall RMSE, with 0.002446 at the zip code level and 0.3173 at the census tract level.

Table I: Top Model Performance (Zip Code Level)

Model	Total RMSE
Naive	0.002446
Random forest	0.002946
Ridge regression	0.003661
Linear regression (select features)	0.003663
MLP neural network	0.005156

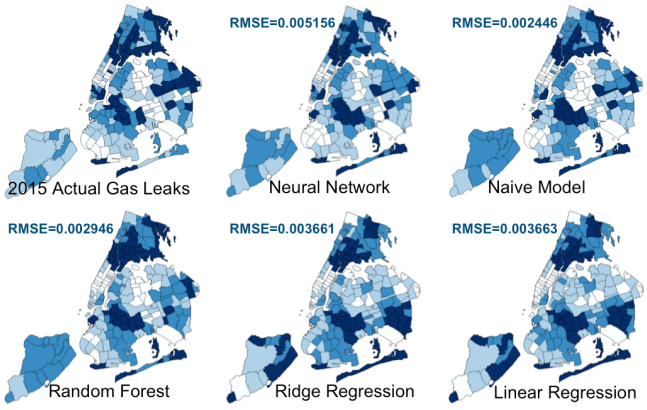

Figure 3. Performance of models at the zip code level

Table II: Top Model Performance (Census Tract Level)

Model	Total RMSE
Naive	0.3173
Random forest	0.4933
Linear regression (select features)	0.7343
Ridge regression	0.7447
MLP neural network	0.8998

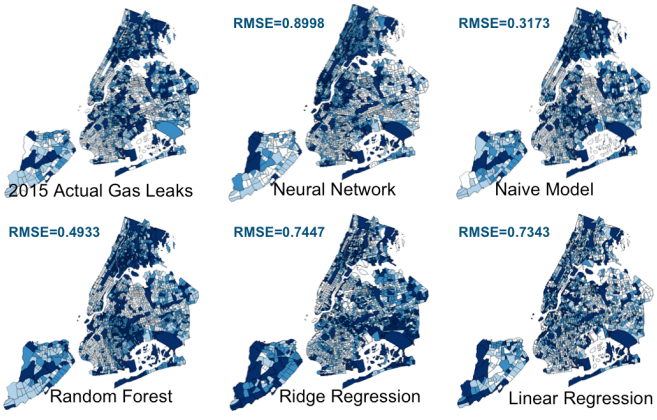

Figure 4. Performance at the census tract level

The MLP neural network and the random forest model did not perform as well as the naive model, but as seen by Figures 3 and 4, did demonstrate what appears to be more geographic sensitivity. These models picked up variations in Staten Island, Lower Manhattan, and Rockaway Peninsula, for example, that did not emerge in the linear regression approach, which is apparent at the census tract level. The tendency of these more complex models to over-fit might inhibit their predictive power but do indicate higher sensitivity, a characteristic that further analysis centered on interpolating possible correlations rather than prediction might exploit.

Additionally, several characteristics were found which were correlated with the likelihood of a gas leak at both the census tract and zip code levels. At the zip code level, these indicators were complaints being referred to the New York City Housing Authority (NYCHA), the percentage of the population that is Black or African American in that zip code, the number of vacant, open or unguarded buildings, alterations of an occupied building without valid permits and the household mean income. At the census tract level, these indicators were the amount of open space in the census tract, permits for construction equipment, failure to paint sprinkler piping, failure to notify DOB prior to the cancellation of earthwork, and the percentage of the population that is Black or African American in that census tract.

A. *Notes on performance*

While total error was orders of magnitude higher for census tract than zip code, these were much more granular geographical areas and that added precision may have value if applied to prioritizing inspections. Initial analysis revealed that when the tract-level model predictions were aggregated back to the zip code level and compared, they fared better and indicate an area of further research – particularly to address whether increasing the granularity of observations in this matter can counter-act the over-fitting that occurs with a small number of observations (such as 195 zips) and large dimensionality (735 features). Additionally, this evaluation metric of total RMSE penalizes extreme values, which influenced models that accurately predicted zips and tracts with little or no gas leaks (a large majority) while drastically under-estimating the top zips or tracts. If the primary goal is to prioritize top tracts, it may be better pursue more precise objectives, like optimizing models based only on specific zips or tracts with a high likelihood of incidents.

Figure 5 demonstrates the tendency of the Multi-Layer Perceptron neural network (in green below) to over-fit on the historical data (which is also what comprises the naive prediction, in red below) and how both approaches under-estimate the top zip codes, although the MLP neural network to a lesser extent. More data and better prediction objectives may reveal promising opportunities for neural networks.

IV. Conclusion

Building a model to predict the number of gas leaks in a given area in New York City, could save lives and help streamline inspections. These authors set out to use publicly available data sets to find predictors of gas leaks and the model which best predicts the gas leaks in a given area based on the previous years. After a review of a series of complex models, it was found that, using a Root Mean Squared Error measure

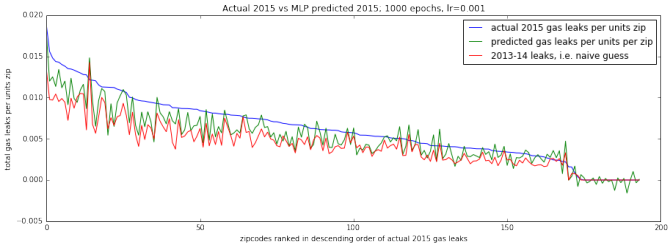

Figure 5. Multi-Layer Perceptron Neural Network Overfitting on Historical Data

of accuracy, none of these models performed as well as the naive model at predicting the number and location of gas leaks across New York City at both the census tract and zip code levels. The more complex models were, however, able to find predictors for these leaks including population dynamics, characteristics of the built environment and construction in the area.

It is important to note that with such a complex set of factors—including the physical state of the gas distribution network, streets, and buildings, as well as (mostly unobserved) human behaviors that introduce risk—it would be extremely difficult to assign any sort of causation to gas leaks. A conventional statistical analysis might focus on correlative relationships, but that is complicated when there is a large number of features and feature selection is non-trivial. Our research intuition is that building predictive models certainly would not address cause and may not shed deep insight into correlative factors for gas leaks, but could add value as applied to prioritizing more deterministic methods, such as on-site inspections. These probabilistic models would prove their worth in terms of accuracy and add value by optimizing the inspection effort. Additionally, as predictive models become more powerful and research with these datasets continues, a predictive model approach could lend insight into important correlations that might be used to develop risk profiles for characteristics of the built environment, such as building type.

REFERENCES

[1] M. E. Gallagher, A. Down, R. C. Ackley, K. Zhao, N. Phillips, and R. B. Jackson, "Natural Gas Pipeline Replacement Programs Reduce Methane Leaks and Improve Consumer Safety," *Environmental Science & Technology Letters*, vol. 2, no. 10, pp. 286–291, oct 2015. [Online]. Available: http://dx.doi.org/10.1021/acs.estlett.5b00213

[2] A. Forman, "Caution Ahead: Overdue Investments for New York's Aging Infrastructure." *Center for an Urban Future*, 2014. [Online]. Available: http://eric.ed.gov/?id=ED555648

[3] N. MODA, "NYC by the Numbers: Annual Report 2013," NYC Mayor's Office of Data Analytics, Tech. Rep., dec 2013. [Online]. Available: https://assets.documentcloud.org/documents/1173791/moda-annual-report-2013.pdf

[4] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, oct 2006. [Online]. Available: https://doi.org/10.1016%2Fj.ijforecast.2006.03.001

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[7] "Project code repository." [Online]. Available: https://github.com/cmprince/USI_social_impact