Identifying and Prioritizing Fire Inspections: A Case Study of Predicting Fire Risk in Atlanta

Michael Madaio
Georgia Institute of Technology
North Ave NW
Atlanta, GA, USA
mmadaio@gatech.edu

Oliver L. Haimson
Department of Informatics
University of California, Irvine
Irvine, CA, USA
ohaimson@uci.edu

Wenwen Zhang
Georgia Institute of Technology
North Ave NW
Atlanta, GA, USA
wzhang300@gatech.edu

Xiang Cheng
Emory University
400 Dowman Dr
Atlanta, GA, USA
xcheng7@emory.edu

Matthew Hinds-Aldrich
Atlanta Fire Rescue Department
226 Peachtree Street, SW
Atlanta, GA, USA
mhinds-aldrich@atlantaga.gov

Bistra Dilkina; Duen Horng (Polo) Chau
Georgia Institute of Technology
North Ave NW
Atlanta, GA, USA
bdilkina@cc.gatech.edu; polo@gatech.edu

ABSTRACT

The Atlanta Fire Rescue Department (AFRD) attempts to reduce fire risk by inspecting buildings for potential hazards and fire code violations. This paper provides a case study exemplifying how data science can be used to help cities identify and prioritize potential property inspections, using machine learning, geocoding, and information visualization. As a result of this work, we generated a risk score for 5,000 buildings in the city, with an average of 73% accuracy for predicting future fires. We also identified 19,397 new potential properties to inspect, based on AFRD criteria, and developed an interactive map to allow AFRD staff to make informed decisions about fire inspections. The results of this study hold great promise for helping cities use data–driven techniques to make civic processes more efficient.

Categories and Subject Descriptors

D.4.8 [**Performance**]: Modeling and prediction

Keywords

Data science, government innovation, fire risk, predictive modeling

1. INTRODUCTION

The City of Atlanta Fire Rescue Department (AFRD), like many other fire departments, conducts regular property inspections to ensure that commercial properties comply with the city's Code of Ordinances for fire prevention and safety. The current process for AFRD's property inspections involves a legacy system of paper file records and inspections conducted on the basis of pre–existing permits, without a robust process for identification, selection, and prioritization of new properties to inspect. With an annual average of nearly 650 fires and 2,573 annual property inspections, the AFRD Assessment and Planning Unit wanted to ensure that the properties being inspected were those at greatest risk of fire. Knowing that the 2,573 current property inspections were not all of the commercial properties in the city of Atlanta, they also wanted to obtain a more complete list of commercial properties that potentially needed inspection.

AFRD, through this effort, is attempting to address the problem of how to conduct fire inspections in a more data-driven way, to more efficiently utilize their limited number of inspection personnel. This is one example of a civic problem which can be addressed through a careful application of data science techniques. Through a partnership between AFRD and the Georgia Institute of Technology with the Data Science for Social Good program, we discovered 19,397 potential new properties to inspect, and have provided AFRD with a method to prioritize those inspections using a fire risk score for buildings generated by predictive modeling machine learning classification techniques. By gathering data from a variety of government and commercial sources, we obtained a robust set of building information variables, in order to build a predictive model of fire risk based on the features of buildings[1] that had previously caught on fire. We then assigned the fire risk score generated by our model to the lists of current and potential properties to inspect, and created an interactive map of the city to provide a tool to visualize those results. We hope that our work will augment the Atlanta Fire Rescue Department's decision-making process to improve fire risk reduction, and help support a data-driven allocation of inspection personnel and other essential resources. Additionally, we hope our work can prove

[1]We will be referring to *buildings* and *properties* as two distinct ideas throughout this paper. The AFRD conducts property inspections and issues permits to the owners of those properties. However, it is the physical structure of buildings that catch on fire, and thus, when we build the predictive model, we do so with information about the buildings themselves, which may or may not contain properties.

Bloomberg Data for Good Exchange Conference.
28-Sep-2015, New York City, NY, USA.

useful as a model for the identification and prioritization of property inspections in other cities around the country, and, perhaps, the world.

2. RELATED WORK

The clearest precedent for our research with AFRD is the recent work from the Fire Department of New York (FDNY) to build a "Risk-Based Inspection System" (RBIS). After conducting focus groups with firefighters and using data about building fires in New York City, the NYC Mayor's Office of Data Analytics built a "data-driven model to identify buildings at greatest risk for fires, to better prioritize FDNY's inspection process [1]. One key challenge faced by both the FDNY RBIS initiative, as well as our work with AFRD, was the difficulty of joining disparate sets of data about city buildings, gathered from various city departments with different building ID numbers, location formats, and varying levels of data completeness. One way to resolve this issue, as FDNY found, is to rely "less on technological expertise and more on strong political leadership from... senior figures in city and local government" [1]. As of the time of writing, the City of Atlanta's Office of Buildings, Office of Housing, and the Fire Rescue Department did not share a unified database of buildings, and thus, the process of joining building data sets became a more technologically difficult task than it might otherwise have been.

Other similar research has used predictive modeling in the context of forest fires in Europe, to better support the allocation of firefighting, fire prevention, and foliage recuperation resources to the areas of highest fire risk [2]. For instance, de Vasconcelos' work with spatial prediction of fires using logistic regression and neural networks is similar to the process we used in creating our predictive risk model [3]. However, their work, as well as others, has primarily dealt with fire risk at a regional level, using geographic and topographic features of Greek, Portuguese, and Spanish regions as variables in the model [3, 4, 5], instead of a more fine-grained unit of analysis of buildings in a city, as we employ here.

3. METHODOLOGY

In this paper, we will highlight and describe our general process for identification and prioritization of property inspections, emphasizing challenges and important steps in the process. Before we could discover new potential properties to inspect, or prioritize those inspections with a fire risk score, we first needed to join data from a variety of sources. This was done to construct as complete a picture as possible of the properties in Atlanta needing inspection, based on information about where fires had previously occurred in the city. After the data joining, we were able to identify 19,397 new potential properties to inspect, through a process of property discovery using AFRD and City of Atlanta fire code criteria, and geocoding techniques. Then, we built a predictive model of fire risk, using building information about fire incidents from 2011-2014, and evaluated the results of that model in several ways. Finally, the fire risk scores generated by that model were applied to 5,022 of the current and potential property inspections, and those results were visualized on an interactive map that AFRD staff will use to augment their inspection processes. See Table 1 for a summary of the different lists of property inspections and buildings we will be referring to throughout this paper.

3.1 Data Joining

The data used in this study came from a variety of sources, as tabulated in Table 2.

The historical fire incidents and inspection permit records were provided by AFRD. The majority of the commercial property data, which includes a variety of building features, such as year built, building material, number of floors and units, building condition and other information, were purchased from the CoStar Group, a commercial real estate agency, by AFRD. Other data, such as the Atlanta Fulton county and Dekalb county parcel information, data on parcel conditions, and business license information, were obtained from the City of Atlanta's Office of Buildings, and the Office of Housing's Strategic Community Investment (SCI) report. We also obtained socioeconomic and demographic data from the U.S. Census Bureau. All of these data sources contribute to developing our predictive model for commercial fire risk estimation.

A critical step of this study was to join different datasets together so that data from different sources about the same building or property could be unified to create the most complete picture of a given property. For instance, by joining fire incident and commercial property data together, we can obtain a general idea regarding which types of commercial buildings caught fire at the highest rates. Or, for example, by joining commercial property data with parcel data from the SCI report, we can generate a more comprehensive view regarding specific characteristics of buildings, including the structure and parcel condition, as well as vacancy information.

We joined the datasets together based primarily on the spatial location information. There are three types of spatial or location information in our datasets, including longitude and latitude, address information, and the parcel identification number, which is a unique ID number created by Fulton and Dekalb county for tax purposes. We performed a location join based on the above three types of location information. The final joins, and the variety of spatial information types, are illustrated in Figure 1. One obstacle we encountered was that spatial information had different formatting standards across the datasets. For example, the addresses from the CoStar Group were all in lowercase, with road names abbreviated instead of fully spelled out, while data from the multiple departments of the City of Atlanta tended to use a more consistent address format. Therefore, a spatial information cleaning process was conducted before joining the datasets directly. We used three tools, including ESRI ArcGIS, Google Geocoding API, and the US Postal Service's address validation API to clean up the location information and validate the coordinates, so the data could be in a more uniform format before being joined together.

3.2 Discovering New Properties to Inspect

To discover new potential properties to inspect, we first had to understand what types of properties currently required fire inspections, in order to find others of similar property types. Using a list of occupancy usage types from the current fire inspection permit database, we found more than 100 unique occupancy usage types that were currently being inspected. Then, by joining the currently inspected properties

Table 1: Summary of Inspection and Building Lists

Name	Count
Current Annual Inspections	2,573
Potential New Inspections (long list[2])	19,397
Potential New Inspections (short list)	6,096
Current And Potential Inspections (short list)	8,669
Current and Potential Inspections (short list) with Risk Score	5,022
Commercial Buildings used to build Predictive Model	8,224

Table 2: Data Sources Summary

Source	Name	Description
Atlanta Fire Rescue Department	Fire Incidents	Fire incidents from 2011 - 2015
	Fire Permits	All permits filed by AFRD
City of Atlanta	Parcel	Basic information for each parcel in Atlanta
	Strategic Community Investigation	Information regarding parcel conditions
	Business Licenses	All the business licenses issued in Atlanta
Atlanta Police Department	Crime	2014 crime in Atlanta
	Liquor Licenses	All filed liquor licenses by Police Department
Atlanta Regional Commission	Neighborhood Planning Unit	Boundary data for each Atlanta neighborhood
U.S. Census Bureau	Demographic	Household number, population by race and age
	Socioeconomic	Household median income
CoStar Group, Inc	Costar Properties	Commercial property information
Google Place APIs	Google Place	Information regarding places from Google Maps

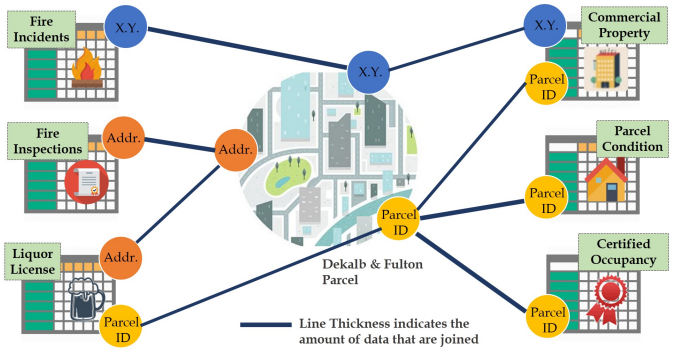

Figure 1: Data Joining. Six data sets were joined using three different spatial information types.

with the Atlanta Business License data, we discovered that, in addition to the 2,573 currently inspected properties, there were approximately 19,397 properties of the same property types in the city. For instance, the Fire Code of Ordinances stipulates that motor vehicle repair facilities needed inspection, yet only 186 of a total of 507 of those facilities in the city are currently inspected annually, suggesting many or all of the rest of those 507 should be inspected. Then, based on the most frequently inspected property types, we created a shorter list of 6,096 new potential property inspections (instead of 19,397), gathered from a variety of data sources, including the Atlanta Department of Finance Business License database, the liquor license database from the Atlanta Police Department, and other sources from the Georgia State Government, as well as the Google Places API.

Since many properties exist in multiple datasets, we had to ensure that the properties on our new potential list were unique and were not currently inspected, after the aforementioned datasets were joined together. Different approaches were used to ensure the uniqueness and newness of potential property inspections. The most reliable and efficient method was found to be joining them pairwise using geocoding and approximate ("fuzzy") text matching for the business names and addresses.

3.3 Building a Predictive Model

However, 19,397 new properties is far more than AFRD is able to inspect on a yearly basis, and, moreover, not all of those properties need to be inspected with the same frequency. We therefore created a predictive model to generate a fire risk score based on the characteristics of buildings that had previous fire incidents in Atlanta. This model was built using the R statistical programming language and used the SVM (Support Vector Machine) machine learning algorithm. The model uses 58 independent variables to predict fire as an outcome variable.

3.3.1 Data Cleaning

After joining various datasets together to obtain building information for buildings that historically had fire incidents, there was still significant data cleaning that needed to occur. The bulk of the data cleaning process involved finding the extent of the missing data and deciding how to deal with that missingness. Our missingness procedures were designed to

[2]We provided AFRD with two lists of potential properties: one longer list that was the most extensive that we could provide, and another shorter list that was more manageable to display on a map, which was refined using the most frequently inspected property usage types.

minimize deletion of properties with missing data, because a significant number of the properties in our model had NA values for many variables. For each variable with missing data, we used "NA" as a category, rather than removing properties with missing data. This required turning many continuous variables into categorical variables. For continuous variables with minimal missing data, we turned NA values into the median or the mean of the data, whichever was most appropriate, again to avoid getting rid of missing data.

3.3.2 Feature Selection

After merging the datasets, we had a total of 252 variables for each property. Our final model includes only 58 variables. We manually examined each variable to determine whether it may be relevant to fire prediction, and excluded many irrelevant variables in this initial process, such as the phone number of the property owner, or property ID numbers. We then used forwards and backwards feature selection processes to determine each variable's contribution to the model, and removed the variables that did not contribute to a higher predictive accuracy.

3.3.3 Model Selection

We built a series of models using several different supervised machine learning algorithms to find the most accurate model. The algorithms we tried included Logistic Regression, Linear Discriminant Analysis, Neural Network, C50 (Classification and Regression Trees), Gradient Boosted Machine, rPART, and finally Support Vector Machine (SVM). We decided to ultimately use SVM for the final risk score because it produced the most predictive results.

3.4 Evaluating the Predictive Model

We evaluated the performance of our predictive model in two ways:

First, we validated our model using a time-partitioned approach. Such a fire risk model would ideally be tested in practice by predicting which buildings would have a fire incident in the following year, and then waiting a year to see which ones actually did catch on fire. Because we wanted to effectively evaluate the accuracy of our model without waiting a year to collect data on new fires, we simulated this approach by using data from fire incidents in 2011 - 2014 as training data to predict fires in the last year of our data, 2014 - 2015. We used 10 bootstrapped random samples and took the average of each of them to calculate our results. This model performed very well, with an average accuracy of 0.77 and average area under the curve (AUC) of 0.75. See Figure 2a for a confusion matrix of the results. The most important metric in this case is the true positives - that is, how many buildings our model predicted would have a fire that actually did have a fire. Of the buildings in our dataset from 2014-2015 that did have a fire, our model was able to predict 73.31% of them. Considering how few fires occur (only about 6% of the buildings in our dataset had fires in 2011-2015), this is much better than guessing by chance which buildings would catch on fire.

We also validated our model using 10-fold cross validation, a more standard machine learning validation approach. This model also did quite well, with an average accuracy of 0.78 and average AUC of 0.73. See Figure 2b for a confusion matrix of the results. In this validation, we were able to predict true positives 67.56% of the time.

It is worth discussing here the implications of the false positives in this model. In both validation approaches, we had a substantial amount of false positives - that is, buildings that our model predicted would have a fire, but that did not actually have a fire. Though many predictive models try to maximize the specificity (the ratio of true negatives to all negatives) by increasing true negatives and reducing false positives, in the context of determining which properties to inspect, false positives are actually quite valuable. False positives represent buildings that share many characteristics with those buildings that did catch on fire. Thus, because they have these characteristics, these are buildings that may be at high risk of catching on fire, and likely contain properties that should be inspected by AFRD.

Additionally, because our training set and the data set that we ultimately apply the model to are the same (that is, a complete list of commercial properties in Atlanta), a perfect model with no false positives would do nothing more than tell us which buildings had previously caught on fire. While this is useful to know, it is data that AFRD already has, and which they already can use to inform their property inspections. False positives give us the added value of predicting which buildings have not caught on fire (within the five years of our fire incident dataset), but which are at risk of fire due to their building characteristics.

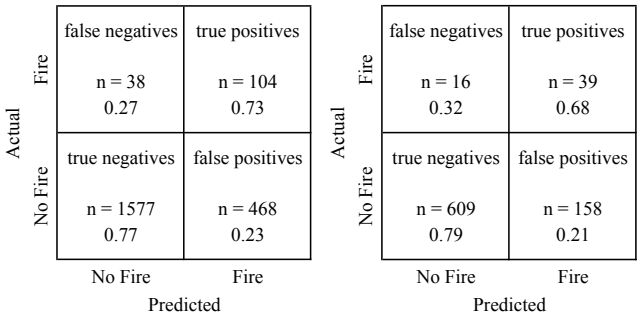

Figure 2: Confusion Matrices. (a) Results of Time Partitioned Approach. True positive rate was 0.73.; (b) Results of 10-fold Cross Validation. True positive rate was 0.68

3.5 Applying the Predictive Model

After we built the predictive model, we then applied the fire risk scores to the list of current and potential properties to inspect so that AFRD could focus on inspecting properties in buildings most at risk of fire. To do this, we first computed the raw output of our predictive model for the list of 8,224 commercial buildings we used to train and test the model. This generated a score between 0 and 1 for each building. To be more useful, we then translated those scores to a 1-10 scale. Then, we categorized the scores into low risk (1), medium risk (2-5), and high risk (6-10). We then needed to apply these risk scores to the 2,573 current and 6,096 potential properties to inspect. Unfortunately, be-

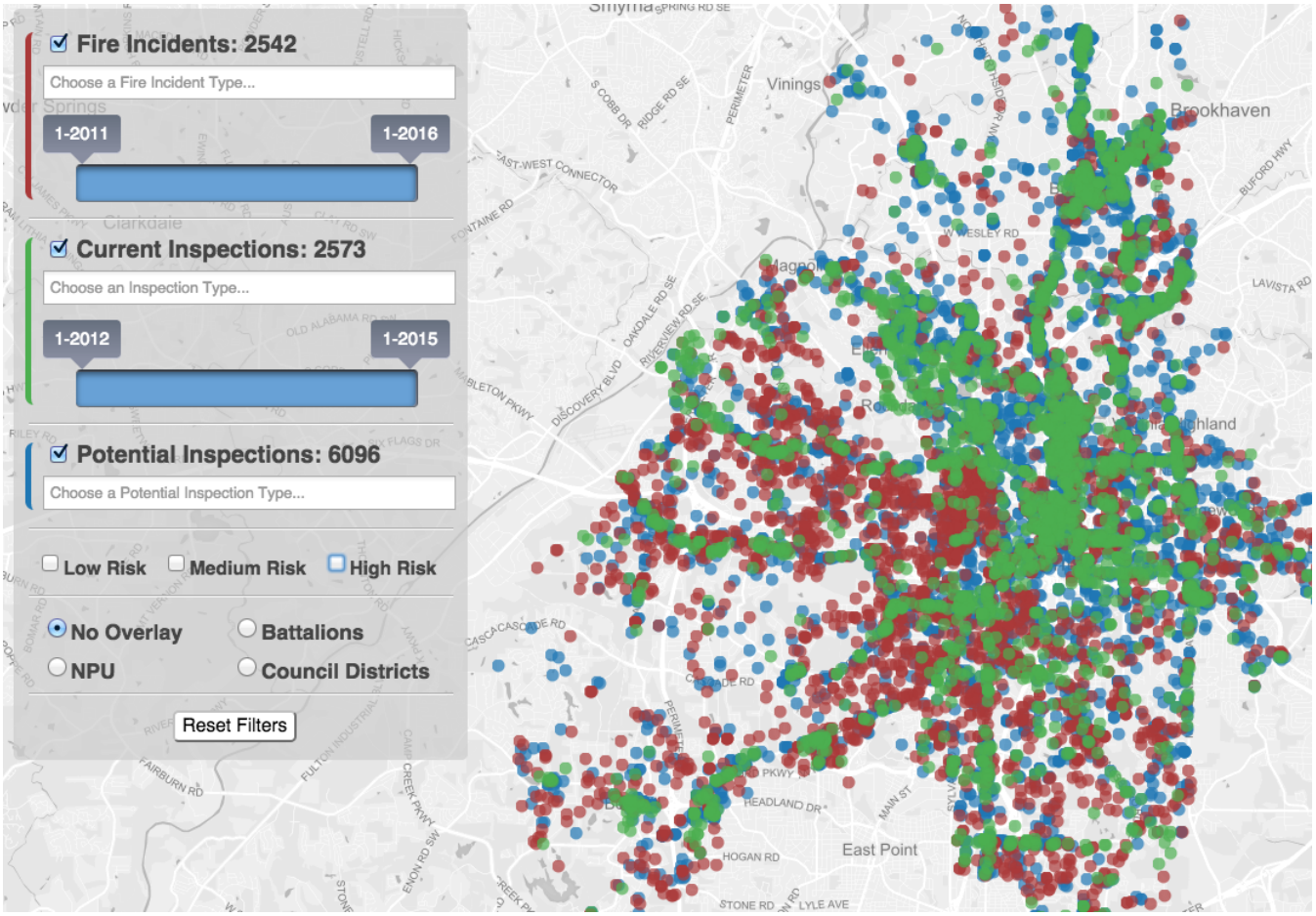

Figure 3: Interactive Map of Fires and Inspections. Allows inspectors to filter properties based on property usage, date, and risk score.

cause the predictive model was built using a dataset of 8,224 commercial *buildings*, some of those buildings had multiple properties inside them (a mall, for example), and many of the commercial buildings in our dataset did not contain any of the potential *properties* to inspect. We therefore had to discover which of the 8,669 current and potential properties were contained in the 8,224 commercial buildings with a risk score, and assigned those properties the risk score associated with the building they were contained within.

3.6 Visualizing Property Inspections

We ultimately provided the AFRD Assessment and Planning Unit with lists of potential properties to inspect, both short (6,096) and long (19,397), along with their associated building and parcel information in the form of a CSV file, but we also wanted to provide an interactive map for the fire inspectors and AFRD executive staff to use to augment their policy and decision-making processes. The map in Figure 3 was made using the open source map-making tools Mapbox and Leaflet to create the map base layer. Then, using the Javascript visualization library D3.js, we displayed differently colored circles on the map to represent fire incidents, currently inspected properties, and potential property inspections using their X/Y coordinates. We also incorporated a user interface developed through discussions with the AFRD Assessment and Planning Unit, and refined by incorporating feedback from fire inspectors and AFRD executive staff. The map includes a panel for displaying property information when hovering over a circle on the map, such as its business name, address, occupancy usage type, date since last inspection (or date of fire, if applicable), and fire risk score, if available.

The user interface panel also included methods to filter the fire incidents, and the current and potential property inspections according to their property usage type, the date of fire or inspection, and their risk score. Finally, we incorporated a set of regional overlays requested by the AFRD executive staff, including the AFRD battalions, the Atlanta Neighborhood Planning Units (NPU), and Council Districts, of which the latter two are political subdivisions of the city. We also included dynamically generated counts and percentages of fire incidents, and current and potential inspections for each of those regional overlays, so the AFRD executive staff could make decisions at a battalion, NPU, or Council District level. We believe that this map could be used as a powerful tool for supporting data-driven conversations about AFRD personnel and resource allocation, inspection staffing decisions, and may even inform decisions regarding community education programs for fire safety and preven-

tion.

4. IMPACTS AND FUTURE WORK

Our goal in this work was to help AFRD improve the quality, completeness, and efficiency of their commercial property inspections in Atlanta. Though there are many more properties to inspect than they currently have the personnel capacity to support, future inspections may be prioritized to target the properties most at risk of fire, leading to a reduction in the frequency and severity of fire incidents in the city of Atlanta. In addition, it is our hope that the results of this work can help inform AFRD's personnel and resource allocation decisions, as well as support community education for fire risk prevention targeted at a particular battalion, Neighborhood Planning Unit, or Atlanta city council district.

In the future, we hope AFRD or other fire departments and city organizations interested in applying a data analytic approach to their property inspections can use the methods outlined here for identifying and prioritizing new property inspections.

In addition, future research should seek to refine, expand, and further validate our prediction model. Due to missing or erroneous entries in the data sources, we were only able to provide risk scores for 5,022 of the short list of 8,669 current and potential inspections which we provided to AFRD. A future version of this research might train the model on a dataset that has fewer building information variables, but may be applicable to more properties. Other research could improve on the accuracy of the model, perhaps by incorporating other sources of data, such as violations of prior fire inspections, data from the Department of Health and Wellness inspections, information from the Certificates of Occupancy, or other, more behavioral sources, such as sanitation or noise violations, rather than the building and structural data that we used. In addition, more research needs to be done on the usefulness and usability of the interactive map, and how exactly AFRD inspectors and executive staff are using it to inform their day to day planning, decisions, and operations.

One step that cities can take towards this process is to generate a unique Building Identification Number (BIN), used by relevant city departments, such as the Office of Buildings, Office of Housing or city planning departments, as well as the Fire and Police Departments. This would allow for an easier joining of various disparate sources of data, without the need for extensive data cleaning, address validation, text matching, and other complex, and potentially error-generating processes.

Identification, selection, and prioritization of risky properties for inspection can be very difficult for cities that do not have an integrated data platform, because buildings and properties may have relevant information that is isolated from other data sources, and which may not have a regular, timely process for updating information. Our work can be a model for the complex process of new property inspection identification and prioritization. Our experience joining isolated data sets from different government departments, commercial data, and open data sources could be invaluable for many cities that want to begin utilizing data science for a smarter city, without requiring a significant financial investment. We hope the impact from our work may further promote the beneficial use of open public sector data in the city of Atlanta, and elsewhere.

5. ACKNOWLEDGMENTS

We want to thank our partners at the Atlanta Fire Rescue Department for their support and willingness to re-examine their existing processes, especially Chief Baker, Chief Rhodes, Chief Day, and the staff of the Assessment and Planning Unit. We also want to thank the Georgia Institute of Technology and the Data Science for Social Good program for their funding and academic support for this research.

6. REFERENCES

[1] E. Copeland. Big data in the big apple. *Capital City Foundation.*, 2015.

[2] A. Alonso-Betanzos, O. Fontenla-Romero, B. Guijarro-Berdiñas, E. Hernández-Pereira, M. I. P. Andrade, E. Jiménez, J. L. L. Soto, and T. Carballas. An intelligent system for forest fire risk prediction and fire fighting management in galicia. *Expert Systems with Applications*, 25(4):545–554, 2003.

[3] M. P. de Vasconcelos, S. Silva, M. Tome, M. Alvim, and J. C. Pereira. Spatial prediction of fire ignition probabilities: comparing logistic regression and neural networks. *Photogrammetric engineering and remote sensing*, 67(1):73–81, 2001.

[4] L. S. Iliadis, A. K. Papastavrou, and P. D. Lefakis. A computer-system that classifies the prefectures of greece in forest fire risk zones using fuzzy sets. *Forest Policy and Economics*, 4(1):43–54, 2002.

[5] L. S. Iliadis. A decision support system applying an integrated fuzzy model for long-term forest fire risk estimation. *Environmental Modelling & Software*, 20(5):613–621, 2005.