An Ethical Case Study on the Application of Smartphone Data Collection from Potholes and the Resulting Biases

COSC 4931 - Project 2

David Helminiak
Marquette University
Milwaukee, Wisconsin
david.helminiak@marquette.edu

Fayika Farhat Nova
Marquette University
Milwaukee, Wisconsin
fayikafarhat.nova@marquette.edu

Dawn Turzinski
Marquette University
Milwaukee, Wisconsin
dawn.turzinski@marquette.edu

Josephine Zucca
Marquette University
Milwaukee, Wisconsin
josephine.zucca@marquette.edu

ABSTRACT

This work sought to analyze a dataset describing the positions of potholes, within the city of Boston, relative to several data collection methods, including smartphone applications and traditional reporting techniques. The overview should be used as a singular case study as to how ethical considerations may be made by the data science community, as well as the significance of biases present within datasets created through emergent technologies. These considerations were made using an existing framework provided by DrivenData Labs: the deon ethics checklist. This analysis was performed using basic statistical methods and a multinomial logistic regression, implemented in the R coding language. The analysis was conducted with a particular emphasis on determining the ethical implications imparted by the biases present within the given dataset and the possible ramifications of conclusions drawn from the results.

CCS CONCEPTS

• **Social and professional topics → Codes of ethics**;

KEYWORDS

Ethics, Biases, Case Study, Machine Learning, R, Multinomial Logistic Regression, Smartphone Data Collection

ACM Reference Format:
David Helminiak, Fayika Farhat Nova, Dawn Turzinski, and Josephine Zucca. 2018. An Ethical Case Study on the Application of Smartphone Data Collection from Potholes and the Resulting Biases: COSC 4931 - Project 2.

1 GITHUB

Project code may be found at:
https://github.com/Yatagarasu50469/COSC4931-project2

.

2 INTRODUCTION

This project's research intends to answer the question: does the publicly sourced smartphone data collection for potholes, within the city of Boston, unfairly bias against protected sets and prove both directly and significantly disadvantageous to them? Potholes are an item of concern to major cities, causing general damage to cars resulting in the cost of repairs and thereby the overall cost of living in any given area. These are further problematic for the general aesthetic of a region, with their presence noteworthy to all, whether driving, biking, or simply walking within the city. Collecting the locations of potholes through a smartphone application intuitively suggests that there would be a strong likelihood of a bias relating to what population actually possesses a smartphone, a vehicle, as well as an inclination to install such an application. Throughout a number of published works, one particular smartphone application called StreetBump has been critiqued for potentially generating such biases and further marginalizing minorities through its generated data. If this is the case, a direct correlation between the number observed potholes with population densities that posses the economic means to afford means to send in this data should be observed. However, it may be considered that there might also be lower income populations, ordinarily incapable of obtaining these means, that have done so through third party aid programs. The local government itself may also make use of the application on their vehicles, such as from its Department of Public Works (DPW), the United States Postal Service (USPS), Boston police department vehicles, and/or contracted disposal services which would all regularly transverse lower income areas. Further, the actual effects of biases from any data collection method must consider the chosen implementation structure, since regardless of the actual information obtained, how people make use of it provides the effectual actions, that in turn affect people's lives.

The case study utilized the city of Boston's publicly available, pothole dataset to perform this evaluation, focusing on a comparison between traditional reporting methods and data harvested through a smartphone application: Citizen Connect. This program incorporates the aforementioned StreetBump with an additional, manual form submission method. Should the application work as intended, there should be similar geographic groupings between reports generated from traditional methods, including those sent in

over telephone lines and reported in person, as compared against those sent through the application.

It was shown that potholes are generally reported less through the Citizen Connect application for regions of lower socioeconomic backgrounds, different races, ethnicity, ages and means of daily transportation. The sole use of the Citizen Connect program would result in an omission of pothole reports from certain Boston neighborhoods as a whole, showing that a large degree of concern should be exhibited for the usage of smartphone application data collection as a whole. Also in line with the hypothesized results, an additional source of pothole reports was an application supplied specifically to city workers, which provided a fairly even distribution of reports. While the smartphone data collection does bias unfairly against several protected sets, the overall effect of combining the data with traditional methods and reports generated from city workers was found to improve Boston's ability to manage the repair of potholes throughout the city.

3 LITERATURE REVIEW

3.1 Smart Cities

More and more urban areas across the United States, and across the globe, are evolving into smart cities. Smart cities are classified as implementing new technologies in "innovative services for transportation, energy distribution, health care, environmental monitoring, business, commerce, emergency response, and social activities" [10]. Smart cities are possible because of the immense number of cellphones possessed by the general public. This makes it possible to distribute the technologies in an infrastructure-free way. Using "volunteers and their mobiles as a (sensing) data collection outlet is known as Mobile Crowd Sensing" [19]. Physically embedding the sensors located in smartphones, in the surrounding environment, "does not only entail an installation expense, but also significant maintenance costs" [10], making an infrastructure-based approach too costly to be reasonable. Smartphones are equipped with the capability to interpret a variety of signals using sensors including "gyroscope, compass, accelerometer, proximity sensor, and ambient light sensor, ... front and back facing cameras, a microphone, GPS and WiFi, and Bluetooth radios" [18]. With vast amounts of data becoming available from these sensors, the logical progression for governmental structures is to create applications that will use the generated information to improve the lives of the citizens. "Pressure is growing on city governments to leverage every opportunity to improve quality of life for inhabitants" [6], and moving towards smart cities is the clearest way to do so. Despite this, there are a number of concerns about using advanced technology, in systems that affect every day life for everyone, such as discrimination of lower income individuals, storing and privacy of individuals' information, and filtering and scheduling the storing of huge amounts of sensor data [4, 15]. Another concern is how smart cities affect developing cities which have less rigid infrastructure, where "road quality tends to be variable" and "the flow of traffic can be chaotic" [20], among other civic differences.

3.2 Potholes

Drivers in urban areas face potholes regularly on their commutes, regardless of the distance. Potholes are small craters in the pavement of the road way caused by the "expansion and contraction of ground water after the water has entered into the ground under the pavement" [3]. The water expands when it freezes in cold temperatures, and contracts when the ice melts, leaving gaps and cracks in the pavement that cave in under the weight of vehicles. Potholes cause extensive damage to cars and cost city and state transportation departments a significant amount of money to repair every year. It is estimated that "potholes cost the average American driver almost $400 a year in automobile repairs" [11]. Not only must potholes be repaired, they must be initially located, requiring "city employees traveling block to block, surveying streets and looking for potholes." [21]. "The City of Boston has over 800 miles of roads and repairs over 19,000 potholes every year" [21], making potholes a substantial concern. Methods for improving the detection and restoration of potholes are highly sought after due to their great impact. For example, Kansas City, Missouri has implemented a prediction algorithm based on traffic patterns, weather, and age to determine where potholes are most likely to form in order to repair the road before they even occur [13].

3.3 StreetBump

A number of applications have been developed to attempt to resolve the pothole problem. Solutions include sending geotagged pictures of potholes taken by users to city hall, using the phone sensors of Boston taxi drivers, studied via the Pothole Patrol app [14], placing similar phone sensors on public vehicles, such as garbage trucks, that cover the majority of a city's road network [7], and StreetBump. StreetBump has thus far been the most popular pothole phone application, with large amounts of research done on the process and accuracy. It uses the GPS and accelerometer sensors within a user's phone while he or she drives around the city to categorize bumps in the road and report potholes. This program is integrated with the Citizen Connect application, used in Boston, that is the source of the primary dataset. StreetBump, on its own merits, has been more thoroughly researched and analyzed than the Citizen Connect application as a whole. Essentially, this "army of mobile phone users driving the streets of Boston" [21] can allow the city to re-allocate workers from surveying the streets to repairing the potholes. The crowd sourced data records the latitude, longitude, speed, and direction of the vehicle when it hits a pothole, which is determined from the y- and z-axis phone sensor readings. "The raw data include both "actionable" and "non-actionable" bumps" [10]; actionable bumps are truly road damage, while non-actionable bumps are train tracks, speed bumps, manhole covers, and other road obstacles. The classification system is monitored by only acting on potholes that are reported by at least three separate users.

While this system is revolutionary and has enormous potential use in regard to sensors in smart cities, there are also a number of potential flaws, especially when considering scaling the system. "In the case of StreetBump, a driver might consciously slow down or drive around potholes to better preserve his or her car", which will modify the data from users [21]. Additionally, "sensor signals can vary according to the particular type of mobile device" [21], as well

as the location of the phone within the car, further altering the data. The ethical implications of StreetBump, and smart cities as a whole, may greatly affect lower income residents. "The user needs to have the app, an iPhone and a car. Although the app is free, not everyone has access to an iPhone and car. This may result in disproportionate amounts of data being generated about some roads, which can result in greater attention being given to road infrastructure in some locations" [21]. Put another way, "every bump from every enabled phone can be recorded. That is not the same thing as recording every pothole" [16]. Evolving into a technology-based society will eventually leave out those who cannot afford the technology, and because they cannot contribute to the data collection process, their interests may not be considered.

3.4 Smartphone Data Collecion Biases

A Pew Research Center study published in 2016, demonstrated the rapid rise in the adoption of smartphone technologies throughout the world, noting that 2013 had an average of 45% of populations from 21 emerging and developing countries use the Internet, or own a smartphone, with that number increasing to 54% after two years in 2015. In the case of the United States, its adult population has an 89% saturation rate for the same statistic and 72% owning a smartphone specifically [23]. The gap in smartphone ownership was addressed in a prominent work published by the Harvard Business Review in 2013, which cites Boston's potholes and the released SteetBump application as potentially problematic as the generated bias may be considered "particularly true of older residents, where smartphone penetration can be as low as 16%. For cities like Boston, this means that smartphone data sets are missing inputs from significant parts of the population - often those who have the fewest resources" [12]. The Royal Statistical Society additionally noted in 2014 that "what Street Bump really produces, left to its own devices, is a map of potholes that systematically favours young, affluent areas where more people own smartphones" [17]. An article, by Barocas in 2016, refers back to the 2013 study stating that: "systematic differences in smartphone ownership will very likely result in the underreporting of road problems in the poorer communities where protected groups disproportionately congregate" [5]. However, aside from this stated conjecture and reasoning, no studies were found that cite a first-hand investigation into the matter, nor that actually perform one.

4 DATASETS

As previously stated, the primary dataset used for this case study is the city of Boston's closed pothole dataset [24]. This dataset contains roughly 8620 rows of closed cases from 2014. There are 28 variables that are specific to the location and information of each pothole. One of the variables is Source, which is the originating technique that the particular pothole was reported through. These include the Citizen Connect app, the City Worker app, and others that will be grouped together as Traditional sources. This latter grouping of potholes can be reported with constituent calls to the city's office of public services, internally identified by employees, or entered through an online website form. The City Worker app is essentially Citizen Connect, but is used solely by city employees. Also of interest, is the time that a case remains open (i.e. the time it takes to respond to a report), the physical location, neighborhood,

Table 1: Summary of the pothole data set

Variable	Min.	Max.	Mean	Med.	Corr.
LATITUDE	42.23	42.39	42.32	42.31	-0.16
LONGITUDE	-71.18	-71	-71.09	-71.09	0
AGES 0 9 (#)	450	15545	5775	3845	0.19
AGES 10 19 (#)	284	17210	6582	3384	0.15
AGES 20 34 (#)	4590	29893	13246	12476	0.08
AGES 35 54 (#)	1816	30941	12518	9786	0.18
AGES 55 64 (#)	768	11366	4716	3873	0.18
AGES 65 UP (#)	624	10743	4778	3532	0.19
CASE OPEN (Days)	0	1113	2.228	0	-0.07
WORK TRANSPORT AT HOME (#)	254	1495	673.6	556	0.07
WORK TRANSPORT CAR TRUCK VAN (#)	1023	30312	12026	9248	0.17
WORK TRANSPORT DROVE ALONE (#)	976	24843	10068	7791	0.17
WORK TRANSPORT CARPOOLED (#)	47	5469	1958	1324	0.16
WORK TRANSPORT BUS TROLLEY (#)	179	9767	3664	2496	0.13
WORK TRANSPORT SUBWAY ELEVATED (#)	859	11046	4410	2138	0.17
WORK TRANSPORT RAILROAD (#)	0	623	267.8	258	0.15
WORK TRANSPORT BICYCLE (#)	20	1197	350.8	284	0.02
WORK TRANSPORT WALKED (#)	273	7329	1841	1826	-0.24
WORK TRANSPORT OTHER (#)	103	1279	388.6	300	0.03
WORK TRANSPORT TOTAL (#)	6458	54699	23623	17721	0.15
HOME OWNER (#)	810	14201	6516	5702	0.16
HOME RENTER (#)	3614	27036	11646	8287	0.12
HOME TOTAL (#)	5450	41237	18162	14844	0.14
REGION MEDIAN INCOME (USD)	26280	91468	55622	48541	-0.09
RACE WHITE (#)	1434	36761	15858	15051	0.04
RACE AFRICAN AMERICAN (#)	178	49144	15716	4462	0.15
RACE HISPANIC (#)	374	21419	9632	9140	0.24
RACE ASIAN (#)	386	10637	3554	1692	0.02
RACE OTHER (#)	156	9739	2854	1186	0.13
RACE TOTAL (#)	9023	114249	47614	35541	0.16
ELEVATION (ft)	-8.25	257.95	62.31	47.69	0.13

zipcode, and geocoded location. The remaining variables were not particularly relevant to the purpose of this research, and also did not include any demographic information. Additional datasets were added to investigate the correlation to Boston neighborhood demographic information.

Six additional datasets were included, in this study, to enhance the primary information. One additional dataset is Age, which contains 26 rows with 14 variables, about each neighborhood in Boston, broken up by different age groups [22]. The transportation set consists of 26 rows with 21 variables specifying the most common means of transportation utilized in the Boston area neighborhoods [22]. These include walking, biking, public transportation, and personally owned vehicles. The next dataset, Housing Tenure, has 26 rows with 7 variables specifying the housing occupancy of the neighborhoods, particularly whether the homes are owned or rented [22]. Median Income, also with 26 rows, has only one variable, the median income of each Boston area neighborhood [22]. The Race and Ethnicity dataset consists of 713 rows with 5 variables, with all the race/ethnicity by decade for each of the neighborhoods [8]. The final dataset is the Boston Neighborhoods, which contains 26 rows with 8 variables, describing neighborhood features such as square miles, acres, and other information [9].

4.1 Data Cleaning

Substantial cleaning had to be done in order to prepare the data for further analysis and modeling. Hadley Wickham's rules for Tidy Data in RStudio were followed. As Wickham states, "Each variable forms a column. Each observation forms a row. And each type of observational unit forms a table" [25]. The geocoded locations were split into latitude and longitude, with the remaining sets of information transitioned into numeric values, where appropriate to do so, within the analysis. Values with zipcodes, but no neighborhoods, used the former to extrapolate the latter. About 50 rows were either missing variables, were incomplete, or missing so they were removed from the dataset.

The variables in the rows from four of the additional datasets (Age, Means of Transportation, Housing Tenure, and Race/Ethnicity) were combined with the melt function, to give one row per neighborhood. The Neighborhoods dataset included areas that were not in our primary closed potholes cases data, so they were removed from further consideration. Median Income had one neighborhood that is considered a vacation spot, and therefore has no reported income, so this row was also removed. Finally, all of the datasets were combined into one, where each pothole case was matched to its corresponding neighborhood. Overall, the only significant source (>5%) of missing data, in the final set, were the zipcodes from the starting list of pothole reports, with 36% missing from the full set, although this was compensated for by sole reliance on the neighborhood identifiers for manipulating and merging datasets. A summary of the initially generated, final data set, including minimum, maximum, mean, median, and the correlation value with respect to the Citizen Connect application as an originating source, is given in Table 1.

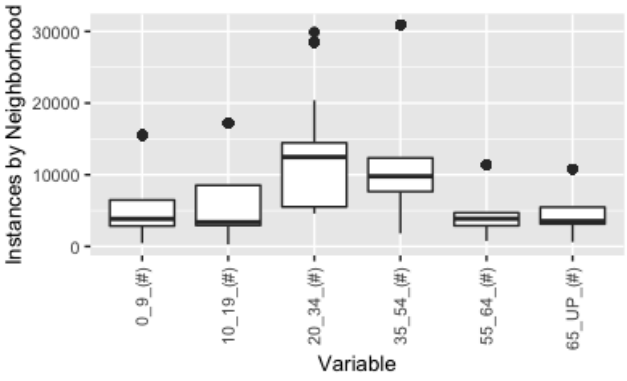

Figure 1: Box plot of the age ranges of the Boston population

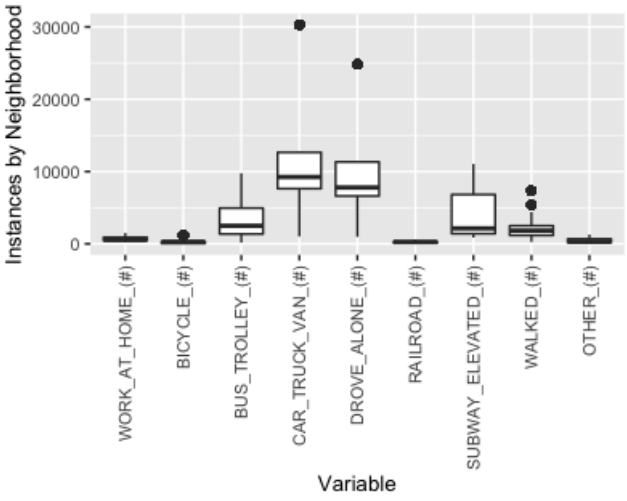

Figure 2: Box plot of the Boston area most common means of transportation

4.2 Exploratory Analysis

The data was analyzed according to a variety of demographic characteristics to create a more representative view of the city of Boston. Figure 1 shows that the most common age group, in the neighborhoods, is 20-34-years-old. This is not surprising because the age range represents the children of the baby boomer generation. Children and teenagers have approximately equal sized groups, while he middle-aged range of 55-64 and senior range of over 65 are also approximately equal. Breaking down the population by most common means of transportation results in Figure 2, where driving a car, truck, or van, as well as driving alone, are the most popular transportation methods. Carpooling is also a common style of commuting, but a significant number of people choose to walk instead. Public transportation is also available in the form of buses, trolleys, and elevated subway. The number of individuals who bike to work is relatively small, which remains logical given that the climate transitions through all four seasons. The rapid changes in weather

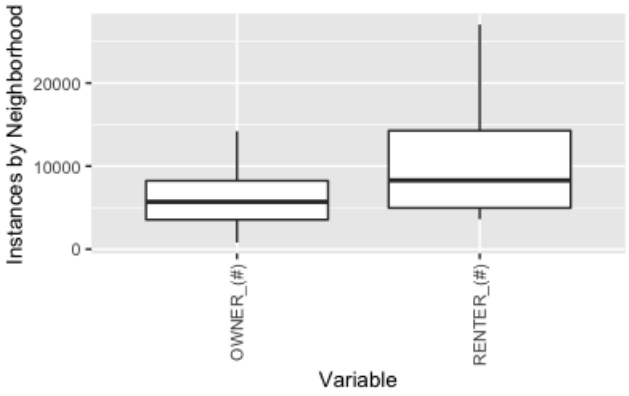

Figure 3: Box plot of the housing occupancy in the Boston area

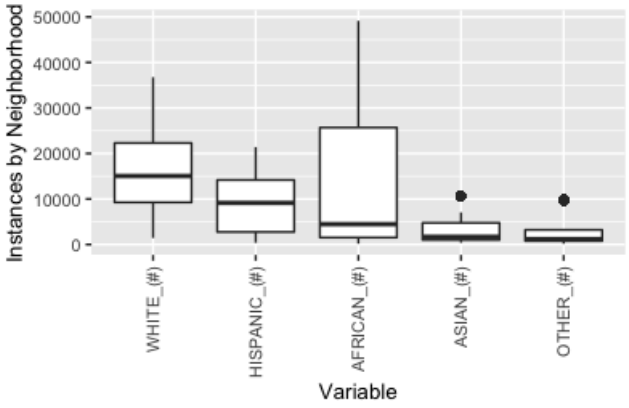

Figure 4: Box Plot of the Boston Area Race/Ethnicity

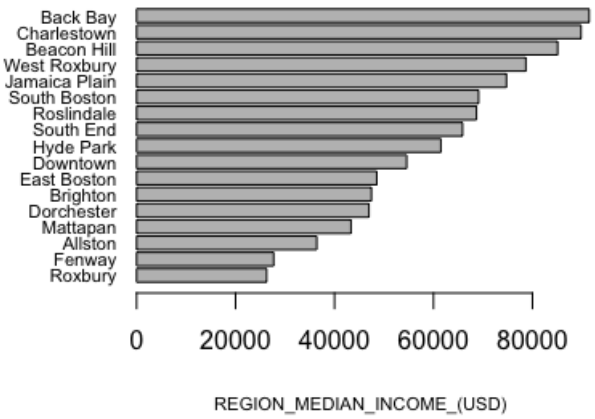

Figure 5: Bar chart of the Boston area median income levels by neighborhood

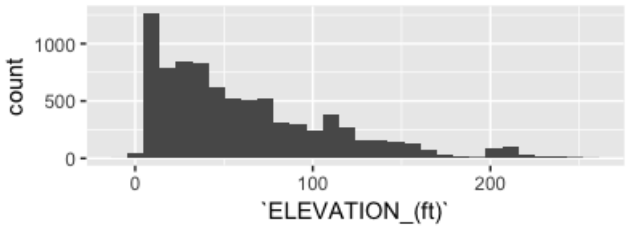

Figure 6: Elevation data subset showing decreasing numbers of reports relative to increasing altitude

conditions and heavy reliance on vehicles for daily commutes explains the high rate of pothole generation. Seen in Figure 3 the housing occupancy of Boston skews towards renting; accounted for by the overall population density. Ethnicity and race demographics are displayed in Figure 4, with the White population being the largest, followed in order by the Hispanic, Black/African American, and Asian/Pacific Islanders. Figure 5 details the median income levels for each of Boston's neighborhoods, with the lowest median incomes from the southern side of Boston. These are the neighborhoods that will be focused on later when determining if lower income areas are less likely to use the Citizen Connect application. Additionally, as is common in urban areas, the waterfront along Boston's southern edge has the highest median income level. The overall correlation between variables is shown in Figure 9. Given the high degree of interplay between them indicates strong potential for the construction of a predictive model. Looking at distributions of values for each of the variables in the set, only one produced a clear relationship with respect to pothole reports. This variable was Elevation and as seen in Figure 6, the number of pothole reports consistently drops with rise in altitude. This follows the logical expectation that since potholes form from subterranean water collection, and as water flows downwards in elevation, the higher the road, the less potholes are generated. Overall, these statistics provide a better background of the Boston area, to be considered with further analysis.

5 DATA ANALYSIS

The density of potholes was plotted by their geographic coordinates and originating sources, as seen in Figure 7, revealing the majority to be positioned in the North End and Financial District, as well as the region of Dorchester. Examining the quantity of reports provided by the Citizen Connect application, there is a clear disparity. The northern sections of Boston embody the vast majority of reports with the Dorchester region containing little to no representation. Comparatively, the reports sent through traditional methods are practically the inverse, with Dorchester being the most represented and northern areas, while still seeing some reports, are markedly less. It may thereby inferred that individuals who submit reports through the smartphone application are less likely to use traditional methods. The widespread adoption of the application implies that those still using the traditional methods must be otherwise unable to submit reports. Contrasting these two sources, the City Worker application generally shows a much more even distribution pattern, with the peaks generally correlating to

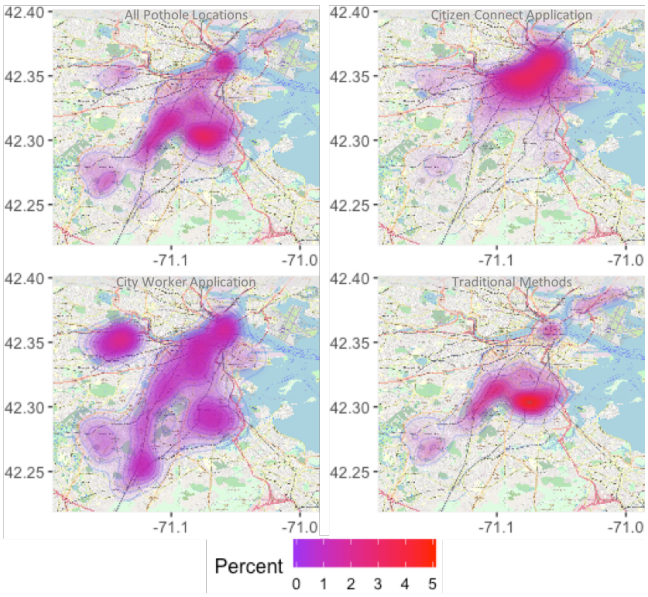

Figure 7: Percent pothole distribution throughout the city of Boston by originating source

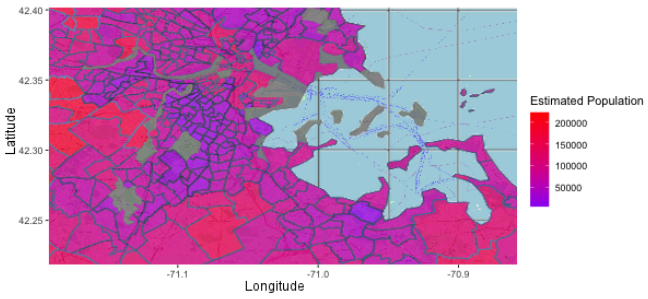

Figure 8: Estimated population distribution according to United States census data from 2011 to 2015 [1]

higher population densities as seen in Figure 8. In Figure 10, the correlation of each variable relative to the Citizen Connect application as the originating source were examined, with values arranged from more to less likely for use of the application. People who walk to work were the most likely, along with those who had higher regional income and lived further north. The disparities in values also show that white and asian races, as well as households with young adults, ages 20 to 34, correlate with higher app usage. Pothole reports were noted to remain open a longer period of time when the application was used to report it, than through traditional methods. However, this data was not normalized to compensate for the uneven distribution of variables relative to intended output classes, meaning that these results may be skewed by sociological and demographic distributions throughout Boston. Geographic and correlation results were further examined through the use of a multinomial logistic regression model analysis. Variables were fitted with a log-linear equation, set to have a maximum of 1,000

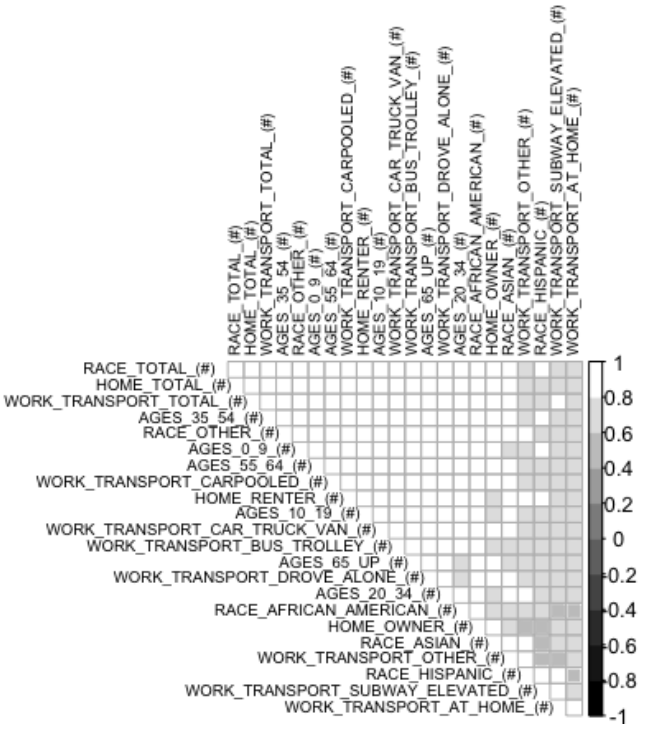

Figure 9: Absolute correlation values between above average dataset variables

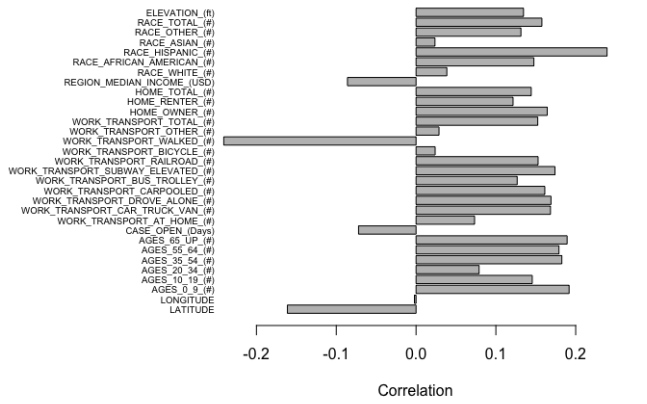

Figure 10: Correlation for the use of the Citizen Connect application, with categories with values further right less likely and values further left more likely

iterations without convergence, with an offset (indicated by the resulting y-intercept) using the R language's multinom function. The model's summary output yields a block of coefficients and their standard error values. Each of these blocks has a row of values that corresponding to a modeling equation. For example, the first row compares SOURCE = "Citizen Connect App" to the baseline SOURCE = "Traditional," and the second row compares SOURCE = "City Worker App" to the baseline SOURCE = "Traditional". With n

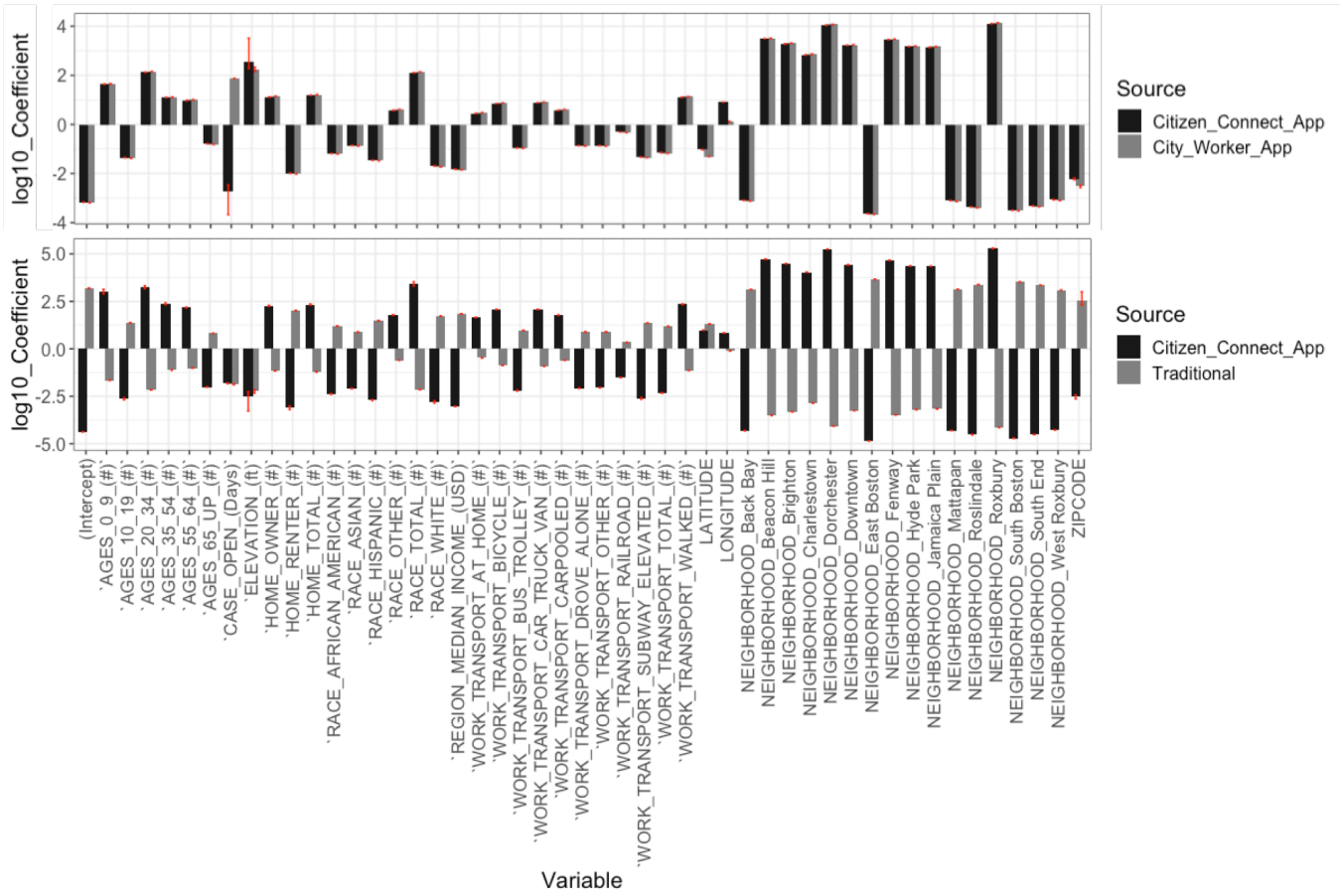

Figure 11: Multinomial logistic regression coefficient values, relative to Traditional methods on top and City Worker app on the bottom, plotted on logarithmic scale with each standard error in red

as the number of columns, the coefficients from the first row, are considered to be b[1,1:n] and the coefficients from the second are considered to be b[2,1:n]. So the following model equations can be written, where CCA stands for the Citizen Connect Application, CWA for the City Worker Application, and TM for Traditional Methods:

$$\ln\left(\frac{P(SOURCE = CCA)}{P(SOURCE = TM)}\right) = b[1,1] + b[1,2] + \ldots + b[1,n] \quad (1)$$

$$\ln\left(\frac{P(SOURCE = CWA)}{P(SOURCE = TM)}\right) = b[2,1] + b[2,2] + \ldots + b[2,n] \quad (2)$$

The complete.cases function was applied prior to calculation in order to ensure that all of the values in the data set were not null and unique. Char and string type variables, including the originating source and neighborhood, were factorized in preparation for the logistic regression with the function: as.factor. Before running the model, the baseline level of the outcome was selected and specified with the relevel function. The Traditional source was chosen initially as the baseline level for this project, with the City Worker app added subsequently, to be compared to the other sources in terms of the effective variables. The data was split into standard 70% training and 30% testing sets, with the model being generated from the former, and validated against the latter. The output coefficients and standard error, relative to each variable and their unique values, are given in Figure 11. Running the model with randomized splits 100 times yielded an average testing error rate of 26.9 %.

Given that, specifically for the geographic distribution of potholes, the City Worker application can be considered to be generally unbiased, and making an assumption that the city worker social composition is normally distributed for ages, race, home ownership, work transport, etc., coefficients that match closely between the remaining sources relative to Traditional, can be considered relatively insignificant for determining biases. In other words, since the City Worker application usage was determinable to approximately the same degree as the Citizen Connect application, relative to traditional methods and using identical variables, only two of them are significant. Using this metric potentially mitigates the error introduced by the uneven variable distribution for each of the desired output classes. Should the pothole case take less time to close, or is located at a lower longitude, then it was more likely to originate from the City Worker app and vice versa for the Citizen Connect app. Examining the Citizen Connect app and the Traditional relative to the City Worker App, however, the clear

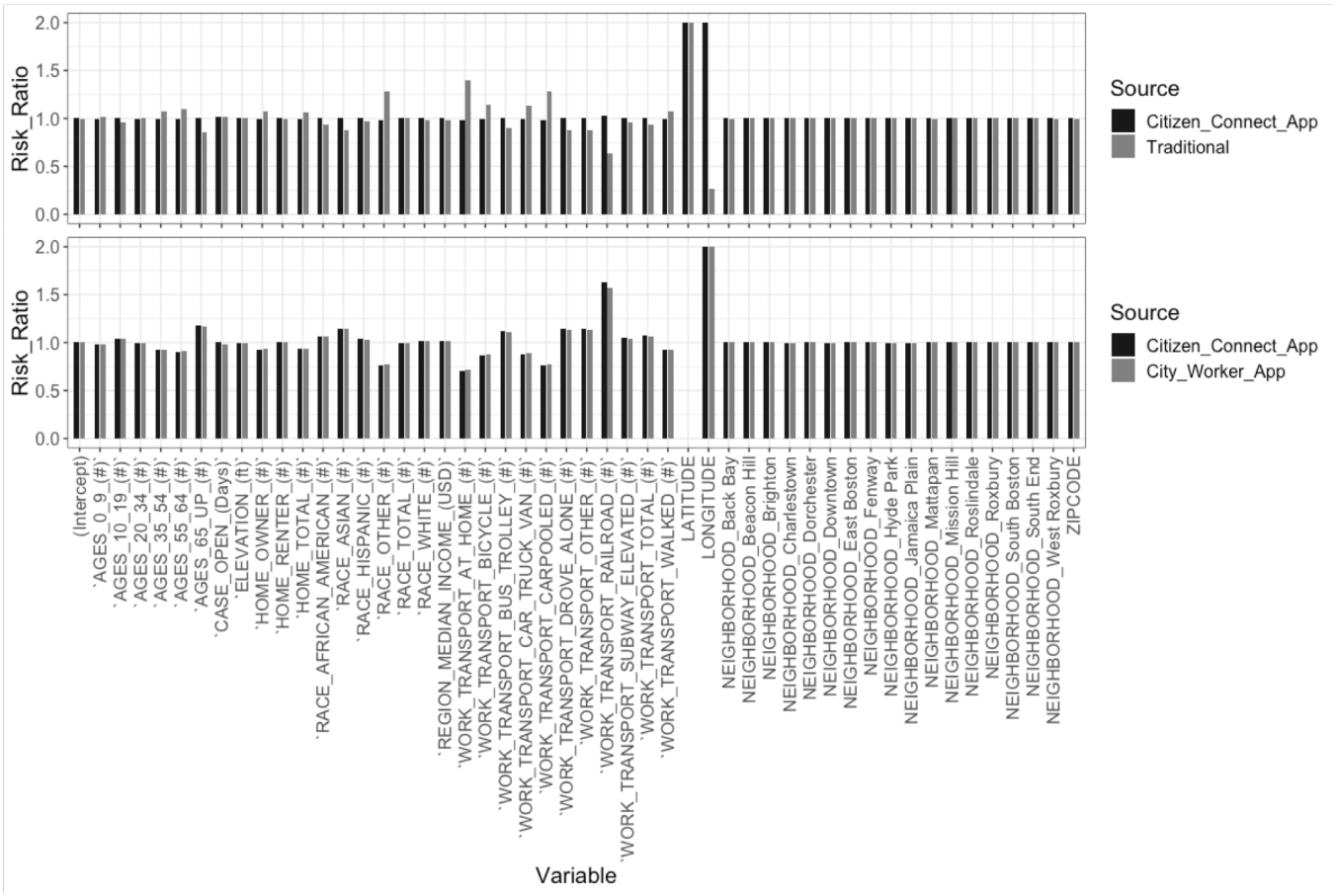

Figure 12: Multinomial logistic regression risk ratio values, relative to traditional methods on top and City Worker app on the bottom; values greater than 2 have been truncated

disparities previously identified by the correlation and geography mapping are seen again. Looking at the lower half of Figure 11, the sources' coefficient values are, for the majority, mirror images of one another. Segments of the population not accounted for by the smartphone application, are identified through traditional methods.

Using Wald z-tests, p-values were calculated for the regression coefficients. The coefficients were also further exponentiated to observe the risk ratios, where both sets of values, one for Traditional source, relative and another for City Worker app relative are seen in Figure 12. The ratio of the probability of choosing one outcome category over the probability of choosing the baseline category is often referred to as relative risk, or odds. The relative risk is the right-hand side of the linear equation exponentiated, leading to the fact that the exponentiated regression coefficients are relative risk ratios for a unit change in the predictor variable. A resulting value of 1 indicates that the source is not affected by the variable, while values greater than 1 show an increase in the odds that the source was responsible for the report, and for values less than 1: the odds of the source being identified as responsible are decreased. Examining the risk ratios confirmed separate analysis of the p-values. Therefore, no bias can be extracted here from the specific neighborhood, or zipcode a report originates from. The odds of using an application, in general are higher, than traditional for those ages 10-19, 65-up, African American, Asian, those who take a bus/trolley, drive alone, take the railroad, and elevated subway to work. These odds, relative to traditional, decrease for home-owners, those who work at home, take a car/truck/van, bicycle, carpool, or walked to work, and live further west. There is also a slight increase in the chance of the source being the Citizen Connect app for individuals who take the railroad to work relative to the City Worker app. Looking at the ratios relative to the City Worker app, show these distinctions more clearly, and further that Traditional reports more often originate for those in age groups 35-54 and 55-64, homeowners, and live further south in the city. Both reports from the Citizen Connect app and Traditional methods are more likely to be obtained from those living further east than those from the City Worker app. Interestingly, as seen in direct contradiction to the supposition proposed by the Harvard Business Review article, if an individual is of an age 65-up and decides to report the location of a pothole, they are actually more likely in Boston to report it with a smartphone.

6 DISCUSSION

The deon ethics checklist [2] for data scientists was used for the evaluation of the ethical ramifications of this study, particularly examining data collection, storage, analysis modeling, and system deployment. In order to limit the overall overview, only the ethical ramifications of the obtained pothole dataset and this study are considered.

In 2009, the city of Boston partnered with Connected Bits to allow citizens to submit photos of public service issues. This system contains the StreetBump application, but was eventually transitioned to the Citizen Connect application produced by Socrata, a third-party, private company, which also made use of the freely available StreetBump program. This code was created specifically to automate the collection and analysis of public service data for local government infrastructures with a focus on involving citizens with smartphones to augment the existing public services response system. The Citizen Connect application was replaced in late 2015 by a new application: BOS:311, though no information was publicly available as to the provenance, or developer of the application, nor why the change had been made.

The Citizen Connect application is still in production for both Android and iOS devices for numerous cities in the United States. The dataset of closed pothole cases herein investigated was provided through the National League of Cities (NLC) Open Data Portal, first uploaded in January of 2015 and last updated in April of the same year. This portal hosts data created and organized by the company Socrata, whose goal remains the promotion of making government data available and usable for both the government and people. Given Socrata's partial objective to make critical use of technology to improve data collection/analysis systems, there exists the possibility that the data was altered to highlight the effectiveness of the application and merits of the software, injecting a bias. This possibility is compounded by the fact that no information was made available as to how the data was stored, which means that Socrata may have hosted and been responsible for securing the data without any public audit.

All of the human subjects involved in the data collection gave consent through the acceptance of the application's terms of service, or through voluntary contact of Boston's public services. There were no apparent steps taken to remove sources of bias, though several public means of supplying information to the dataset were provided, including reports by phone call as well as through the Internet. It does not appear that any particular steps were taken to obfuscate Personally Identifiable Information (PII), but neither is any PII actually available inside of the set. Thereby, all subjects have also given permission for this study's usage of their provided data, with their PII also not available through the resulting work. No information was provided by the web portal as to its current security capability, nor as to how secure the data was prior to their obtaining of it, with the application itself having no publicly released security vulnerabilities. Access logs have not been made available for public assessment and there is no evidence of their existence. The site hosting the version of the dataset obtained for this study was run through the OWASP ZAP (Open Web Application Security Project Zed Attack Proxy) vulnerability scanner. This shows a low-risk possible opening through unsecured web channels to alter the contents and another through cross-site scripting. This study, its models, results, and code are posted publicly online through GitHub with no plans to protect and secure the data, nor schedule to delete the data after its use has expired. Since the data being posted was already within the public domain and has no PII, this has been considered to be a non-issue.

Though it may be suggested for future research, this analysis does not engage with the potentially affected communities, nor with subject matter experts, with the exception of published works mentioned under the Literature Review section. The data has been examined for possible sources of bias and found to possibly have taken some intentional steps to address these. The models generated herein were constructed primarily to detect these biases and evaluate their overall effects, and thereby have been designed to honestly represent the underlying data. While the models and code presented here are publicly available and fully auditable, the originating dataset does not share the latter attribute. The Citizen Connect application has been shown to strongly bias against certain protected sets, while traditional methods produce an approximately inverse set and the City Worker application produces the least overall bias. However, the overall model combines all three of these sources to produce minimal biases and an accurate representation of all supplied public service requests. Generating an actual evaluation of the biases would require a careful census of pothole distribution within the city of Boston at the time that the pothole dataset was obtained. Without this information, the modeling presented cannot accurately measure the overall biases in any of the given sources. Given the limited scope of this study, additional metrics, such as a comparative analysis against other cities' use of the Citizen Connect applications, and/or other smartphone data collection programs, were not performed. The decisions of the generated models from this study, showing which data source was most likely to correspond to a geographic region, were able to be explained by logical explanations and thereby communicate the shortcomings and limitations that this type of data collection presents.

The model was generated for a specific time period as a case study and has no plans to compensate for concept drift, prevent unintended uses and abuse of the model, or monitor the deployed model. The resulting analysis exists in the public domain; should users be harmed by its results, the code base is available for further evaluation and modification.

Models of pothole distribution could have been made using purely synthetic sample generation, which would guarantee a non-existent bias. However, the critical issue with doing so is that while it predicts where potholes may be found, it should always be preferential for workers to fix the ones that are known to, or have evidence for, actually existing. A possible alternative that would eliminate any potential bias would be to only address concerns identified through regular and evenly distributed land/road surveys. Ultimately the goal of identifying algorithm biases, is not the modification of the end results to marginalize non-protected sets for the sake of balance, but the acquisition of knowledge, which indicates the sources of discrimination that have entered into the dataset and algorithm. The goal of removing bias is not to remove from the bottom line of large corporations that have grown reliant on machine learning algorithms to make far-reaching, and

personally impactful decisions. Modifying an algorithm's construction to promote fairness over accuracy decreases its usefulness and guarantees that private companies will always have a motivation to prevent regulatory legislation from being forced upon them. Therefore, the onus is on data scientists, not to go out with their smartphones into marginalized communities and collect pothole data, but rather use the existent data to determine who can fix the root problems that prevent the people already in those communities from making use of these new tools to automatically report potholes. Possibly the best method for managing the bias that does exist with smartphone data collection applications is the continued usage of traditional reporting methods, at least until the technology becomes readily available and accessible to all currently protected sets being biased against by it.

7 CONCLUSION

Future work may consider requesting data directly from Socrata for usage of the Citizen Connect application that are still active and comparing the information against a collected, manual survey of pothole distribution. Alternatively, the data could be examined relative to other instances of smartphone data collection applications and their distribution of use, or against datasets obtainable from other cities with similar characteristics. The City of Boston should be contacted to ascertain details regarding their move away from the Citizen Connect application and the actual management of the collected data herein examined.

Smartphone data collection for the reporting of potholes has been shown in this study to generate unfair biases against protected sets. If the Citizen Connect application was the only way to report potholes, rather than in combination with traditional methods, certain groups would be disadvantaged. However, in this instance, the bias can be considered to have been effectively mitigated by the inclusion of traditional reporting methods and the prioritization, whether intentional or not, of reports submitted through these alternative techniques. Data scientists should continue to consider the bias promoted by the development of smartphone applications that replace existing systems. This study does warn that concrete evidence should always be required to support claims that individuals will be disadvantaged by a given implementation. Failure to do so will only cause unnecessary harm to the progression of society and the overall conversation.

ACKNOWLEDGMENTS

The author would like to thank Socrata and the City of Boston for the public donation of their dataset.

REFERENCES

[1] [n. d.]. Census Data API Discovery Tool. Retrieved December 14, 2018 from https://www.census.gov/data/developers/updates/new-discovery-tool.html.

[2] 2018. An ethics checklist for data scientists. Retrieved October 12, 2018 from http://deon.drivendata.org/#data-science-ethics-checklist.

[3] 2018. How Do Potholes Form? Retrieved December 15, 2018 from http://www.summitengineer.net/resources/learning/52-potholes.

[4] A. Alkhelaiwi and D. Grigoras. 2016. Scheduling crowdsensing data to smart city applications in the cloud. In *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*. 395–401. https://doi.org/10.1109/ICCP.2016.7737179

[5] Solon Barocas and Andrew D. Selbst. 2016. Big DataâĂŹs Disparate Impact. *California Law Review* 104 (2016). https://doi.org/10.15779/Z38BG31

[6] K. Benouaret, R. Valliyur-Ramalingam, and F. Charoy. 2013. CrowdSC: Building Smart Cities with Large-Scale Citizen Participation. *IEEE Internet Computing* 17, 6 (Nov 2013), 57–63. https://doi.org/10.1109/MIC.2013.88

[7] Umang Bhatt, Shouvik Mani, Edgar Xi, and J. Zico Kolter. 2017. Intelligent Pothole Detection and Road Condition Assessment. *CoRR* abs/1710.02595 (2017). arXiv:1710.02595 http://arxiv.org/abs/1710.02595

[8] Analyze Boston. [n. d.]. Boston Neighborhood Demographics. Retrieved November 1, 2018 from https://data.boston.gov/dataset/boston-neighborhood-demographics.

[9] Analyze Boston. [n. d.]. Boston Neighborhoods. Retrieved November 1, 2018 from https://data.boston.gov/dataset/boston-neighborhoods.

[10] T. S. Brisimi, C. G. Cassandras, C. Osgood, I. C. Paschalidis, and Y. Zhang. 2016. Sensing and Classifying Roadway Obstacles in Smart Cities: The Street Bump System. *IEEE Access* 4 (2016), 1301–1312. https://doi.org/10.1109/ACCESS.2016.2529562

[11] F. Carrera, S. Guerin, and J. B. Thorp. 2013. BY THE PEOPLE, FOR THE PEOPLE: THE CROWDSOURCING OF "STREETBUMP": AN AUTOMATIC POTHOLE MAPPING APP. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-4/W1 (2013), 19–23. https://doi.org/10.5194/isprsarchives-XL-4-W1-19-2013

[12] Kate Crawford. 2013. The Hidden Biases in Big Data. Retrieved December 14, 2018 from https://hbr.org/2013/04/the-hidden-biases-in-big-data.

[13] Skip Descant. [n. d.]. How Kansas City Mo. is snuffing out potholes before they appear. Retrieved November 1, 2018 from http://www.govtech.com/fs/How-Kansas-City-Mo-Is-Snuffing-Out-Potholes-Before-They-Appear.html.

[14] Jakob Eriksson, Lewis Girod, Bret Hull, Ryan Newton, Samuel Madden, and Hari Balakrishnan. 2008. The Pothole Patrol: Using a Mobile Sensor Network for Road Surface Monitoring. In *Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services (MobiSys '08)*. ACM, New York, NY, USA, 29–39. https://doi.org/10.1145/1378600.1378605

[15] D. Estrin, K. M. Chandy, R. M. Young, L. Smarr, A. Odlyzko, D. Clark, V. Reding, T. Ishida, S. Sharma, V. G. Cerf, U. HÃűlzle, L. A. Barroso, G. Mulligan, A. Hooke, and C. Elliott. 2010. Participatory sensing: applications and architecture [Internet Predictions]. *IEEE Internet Computing* 14, 1 (Jan 2010), 12–42. https://doi.org/10.1109/MIC.2010.12

[16] Tim Harford. 2014. Big data: A big mistake? *Significance* 11, 5 (2014), 14–19. https://doi.org/10.1111/j.1740-9713.2014.00778.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2014.00778.x

[17] Tim Harford. 2014. Big data: A big mistake? *Significance* 11, 5 (dec 2014), 14–19. https://doi.org/10.1111/j.1740-9713.2014.00778.x

[18] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications Magazine* 48, 9 (Sept 2010), 140–150. https://doi.org/10.1109/MCOM.2010.5560598

[19] Giovanni Merlino, Stamatis Arkoulis, Salvatore Distefano, Chrysa Papagianni, Antonio Puliafito, and Symeon Papavassiliou. 2016. Mobile crowdsensing as a service: A platform for applications on top of sensing Clouds. *Future Generation Computer Systems* 56 (2016), 623 – 639. https://doi.org/10.1016/j.future.2015.09.017

[20] Prashanth Mohan, Venkata N. Padmanabhan, and Ramachandran Ramjee. 2008. Nericell: Rich Monitoring of Road and Traffic Conditions Using Mobile Smartphones. In *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems (SenSys '08)*. ACM, New York, NY, USA, 323–336. https://doi.org/10.1145/1460412.1460444

[21] Daniel O'Leary. 2013. Exploiting Big Data from Mobile Device Sensor-Based Apps: Challenges and Benefits. *MIS Quarterly Executive* 12 (12 2013), 179–187. https://pdfs.semanticscholar.org/4815/250d4be82d81eea0970eac445215d8b533af.pdf

[22] Boston Planning and Development Agency Research Division. [n. d.]. Boston in Context: Neighborhoods. Retrieved November 1, 2018 from http://www.bostonplans.org/getattachment/6f48c617-cf23-4c9f-b54b-35c8a954091c.

[23] Jacob Poushter. 2016. Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies. Retrieved December 14, 2018 from https://data.nlc.org/dataset/Closed-Pothole-Cases/jf2j-9tf6/.

[24] Socrata. [n. d.]. Closed Pothole Cases. Retrieved November 1, 2018 from https://data.nlc.org/dataset/Closed-Pothole-Cases/jf2j-9tf6/.

[25] Hadley Wickham. 2014. Tidy Data. *Journal of Statistical Software, Articles* 59, 10 (2014), 1–23. https://doi.org/10.18637/jss.v059.i10