Processes, Potential Benefits, and Limitations of Big Data Analytics: A Case Analysis of 311 Data from City of Miami

Loni Hagen
School of Information
University of South Florida,
Tampa, Florida, USA,
lonihagen@usf.edu

Hye Seon Yi
Computer Science and Engineering
University of South Florida,
Tampa, FL, USA,
hsyi@mail.usf.edu

Siana Pietri
School of Information
University of South Florida,
Tampa, Florida, USA,
spietri@mail.usf.edu

Thomas E Keller
Genomics Program
University of South Florida, Tampa, FL, USA. tekeller@usf.edu

ABSTRACT

As part of the open government movement, an increasing number of 311 call centers have made their datasets available to the public. Studies have found that 311 request patterns are associated with personal attributes and living conditions. Most of these studies used New York City 311 data. In this study, we use 311 data from the City of Miami, a smaller local government, as a case study. This study contributes to digital government research and practices by making suggestions on best practices regarding the use of big data analytics on 311 data. In addition, we discuss limitations of 311 data and analytics results. Finally, we expect our results to inform decision making within the City of Miami government and other local governments.

CCS CONCEPTS

• Applied computing • Computers in other domains • Computing in government • E-government

KEYWORDS

311 data, big data analytics, information visualization, e-government

ACM Reference format:

Loni Hagen, Hye Seon Yi, Siana Pietri and Thomas E Keller. 2019. Processes, Potential Benefits, and Limitations of Big Data Analytics: A Case Analysis of 311 Data from City of Miami. In *Proceedings of dg.o 2019: 20th Annual International Conference on Digital Government Research (dg.o 2019), June 18, 2019, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3325112.3325212

dg.o 2019, June 18, 2019, Dubai, United Arab Emirates

ACM ISBN 978-1-4503-7204-6/19/06...$15.00
https://doi.org/10.1145/3325112.3325212

1 Introduction

Since its initiation in the 2010s, 311 call centers have enabled the direct interaction between citizens and governments concerning non-emergency information [3]. As part of an open government movement, a number of 311 call centers have made their data sets open to the public. As of January 2019, a total of 28 of 311 datasets are open to the public in the U.S. Scholars found interesting patterns by examining 311 data sets from major cities such as New York City (NYC), Chicago, and Washington D.C.. Using citizen's 311 requests data from New York City, Clark et al. (2013) found that poorer neighborhoods make fewer requests [3]. Kontokosta et al. (2017) found that households with "a higher proportion of male residents, of unmarried population, and minority population; a higher unemployment rate; and more limited English speakers" tend to use 311 requests less frequently even though they experience similar issues with heat and hot water problems as their counterparts [11]. Conversely, "those with higher rents and incomes and a higher proportion of female, elderly, and non-Hispanic White and Asian residents, with higher educational attainment" tend to report more frequently [11]. The study concluded that socioeconomic status, household characteristics, and language proficiency may have an effect on the frequency of 311 service requests. Similarly, using NYC 311 service request data, Wang et al (2017) found that education, income, race, and employment status are associated with 311 service request patterns [17]. Using NYC 311 data, Legesie and Schaeffer (2016) found that complaints about noise, drinking in public, or blocking the driveway are more frequent in areas where the boundaries of ethnic and racial groups are poorly defined [12].

Using 311 data as a measure of the service demands, these findings reveal individual attributes and living conditions may be associated with the frequencies of service requests in NYC. However, the studies present only a limited discussion on the limitations and difficulties of analyzing 311 data through big data analytic techniques. White and Trump (2016) warned that the 311 data should not be used as a generic measure of political

engagement or participation without controls for environmental variables [19].

We found three major research gaps from previous studies. First, the inferences from previous studies are limited to the context of the major cities such as New York City because previous studies explored 311 data from only a limited number of cities. We are motivated to understand citizen demands in smaller cities. Second, although there are pathways to conduct big data analytics using 311 data, the process of collection and preparation of data, and the interpretation of the results using information visualization, are not yet fully explored. Third, the potential merits and limitations of using 311 data for policymaking are not yet clear.

In an effort to address these research gaps, we explore 311 data from City of Miami using big data analytics and make inferences based on information visualization. We further discuss what the data reveals mainly by reflecting on the big data analytics strategies.

2 Literature Review

2.1 Data analytics for policymaking

Big data is data whose volume, diversity, and speed require new technical and analytical approaches to harness value from raw data [1]. Governments are embracing big data problems due to the deluge of data produced through new technologies such as mobile sensors and social media in addition to big administrative data. Some cases show successful governments' efforts to adopt big data analytics to extract intelligence from big data. One such an example is the All of Us program by National Institutes of Health, which gathers lifestyle, environment, and biological data from millions of volunteers in order to create "the world's biggest dataset for precision medicine" [9]. Big data and analytics are used to organize cities by informing long-term urban planning through analyzing energy, weather, and transportation systems [11]. Governments around the world also exerted efforts to publish government datasets for public use hoping to create economic innovations [21] and to improve government performance [17]. Clearly, some centralized initiatives and experiments show high expectations on big data and analytics. However, it is still puzzling specifically how (if possible) data analytics can inform policy decision-making by local governments. The major obstacle is analytics, rather than big data, because it is extremely challenging to gain useful insights from complex analytics results that are helpful for taking action [12].

As an initial effort to understanding big data analytics, it might be useful to discuss data analytics based on the major outcomes. Among many possible outcomes, we can classify data analytics outcomes as *descriptive* and *predictive* outcomes.

First, data analytics can provide description of data by extracting useful patterns from big data. By extracting handful of useful information from large volumes of data, human can process a small quantity of the extracted information for making inferences. This process is in a way producing a snapshot of big data so that human can cognitively process. Some popular algorithms are effective for descriptive purpose. For example, *clustering* algorithms are used to put together customers to groups based on similar taste and purchase patterns. *Association rule mining* is useful to find pairs of items that are frequently purchased together. We can also find extreme events or suspicious activities occurring on the web or in an institutional traffic data using *anomaly detection* algorithms. *Social network analysis* was used to identify the most influential figures during health emergency [8]. *Topic modeling* can extract naturally emerging topics from a large volumes of text data [7]. A handful number of patterns identified through these methods provide insights that would not be apparent using traditional approaches such as manual reading.

Second, another outcome we can expect from data analytics is making predictions. Often, future incidents are so complex and require millions of variables to make accurate predictions. For example, using *classification* algorithms, we can automatically detect spam emails from new incoming email communications. *Logistic regression* can be used to predict the probability of an applicant will default on the loan, based on a loan applicant's credit history. Using *time series analysis*, a clothing retailer can predict monthly sales based on seasonal impacts of the customers' purchasing decisions.

Now, the next question is once descriptive or predictive outcomes are produced using data analytics, in what ways these results can inform policy decision making. Traditionally, policy agenda setting depended on cognitive information gathered through human interaction, media reports, or citizen inputs. As citizens raise their voice more frequently using diverse channels, it is becoming impossible to manually understand the most pressing issues for citizens. Big data analytics is a useful tool for understanding reality by reducing massive data points to a small quantity of useful information [7]. It can also assist policy decision by making prediction of unknown future. However, we are still lacking real examples of using big data and analytics for actual policy decision-making.

2.2 Information visualization

In general, big data analytics results are composed of rows and columns of numbers, which are hard to interpret. Information visualization is highly effective for presenting big data analytics results. Preattentive processing theory, information theory, and psychological findings support the value of visualization for effective processing of information [10]. John Snow's analysis of deaths from cholera in London is a good example of the power of visualization. His analysis effectively demonstrated the reason behind the spread of cholera in specific areas of London was because of contaminated water pumps [6]. Visual presentation was the key to this finding. Information visualization is powerful because it can show known facts in combination with unexpected insights, as a whole. A commercial real estate database company Zillow is a good example of taking advantage of the power of the big data in combination of visual presentation [19]. Zillow present available information of houses for sale in combination with visualization on the map, which

makes obtaining insights extremely efficient (see Figure 1 for an example).

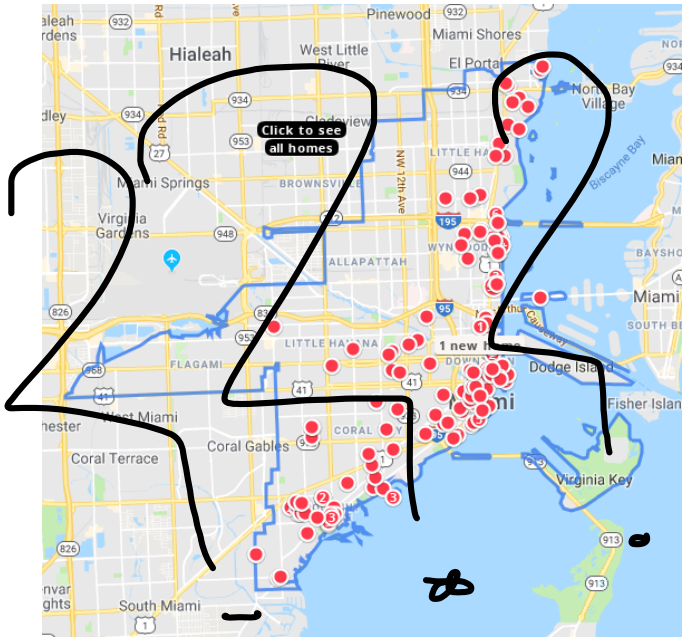


Figure 1: Zillow houses for sale ($800,000 and $950,000) in city of Miami area. We can see most of the houses in this price range are located following the coastal line.

3 Data and Methods

3.1 Analysis strategy

Before we introduce details of data and methods, it will be helpful to note our analysis strategy. We will improvise the data analytics lifecycle suggested by EMC Education Services [4]. Our data analytics strategy is composed of data collection, data cleaning/merging, and data analysis (see Figure 2). Each step informs the following as well as the previous steps, and frequently requires revisiting the previous step(s). In addition, we made several inquiries to government domain experts to validate each process and to understand the data thoroughly. We have carefully validated each step along the study. In preparation for the final report, a new programmer replicated the entire process in order to finally validate the process and the analysis. Below, we will explain details in data and the analytic methods.

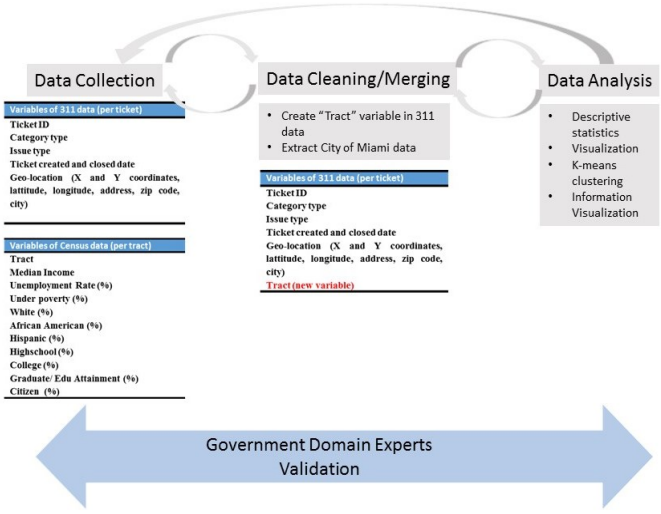


Figure 2: Data Analytics Strategy

3.2 Data collection, cleaning, and merging

We collected and processed 311 request data and Census data. Both of the datasets required new computational approaches to harness values mainly due to volume and diversity of the data. Below, we discuss the process of data collection and cleaning of the two datasets. We made the R script available to the public (https://github.com/311CityofMiami/311-city-of-miami).

3.2.1 request data from City of Miami. We have collected the 311 call data from Miami-Dade County (https://opendata.miamidade.gov/311/311-Service-Requests-Miami-Dade-County/), which contains data between 1-1-2013 and 2-25-2018—all the available data from the initiation of the data to the time of the initial analysis. After consulting to the 311 office of Miami-Dade County, we decided to select 311 reports from the City of Miami only because different administrative structure requires diverse policy requirements, which influence request patterns. City of Miami is selected because it is the biggest municipality among the 36 municipalities in Miami-Dade County.

In the process of cleaning, we dropped all instances containing missing values. In preparation for data cleaning and merging, we created the "Tract" variable in the 311 data to use it as a key to merge with Census data (see Data Cleaning/Merging in Figure 2). A unit of analysis is a "Census tract," which is "an area roughly equivalent to a neighborhood established by the Bureau of Census for analyzing populations" [13]. After cleaning and merging, the 311 requests data from City of Miami included a total of 64,252 observations with 109 tracts. We will refer to this data as 311 data for the duration of the paper. Each 311 request is categorized following two categorization scheme: "Category Type" and "Issue Type." Category Type reflects functional division of the government. Issue Type is a granular level of Category Type. There is a total of 10 Category Types and

119 Issue Types. Each Issue Type belongs to one of the 10 Category Types (see each year's requests by Category and Issue Type in Table 3).

3.2.2 Census data. To analyze 311 citizen requests in the context of socio-economic background, we collected 2016 U.S. Census American Community Survey (ACS) data—the most recent data available through Census at the time of initial analysis. We used R Census Application Programming Interface (details of Census variables are explained in this url: https://www.census.gov/data/developers/data-sets/acs-5year.2016.html). Using the R "censusapi" package [15], we collected the following variables for the analysis. We refer to this data as Census data hereafter. Table 1 describes variables and basic statistics.

Table 1: Descriptive statistics of Census variables (N = 109 Census tracts)

Census Variables	Mean	Sd	Min	Max
Median Income	40,054	27,468	0	156,938
Unemployment Rate (%)	10	7	0	35
Under poverty (%)	27	13	0	63
White (%)	14	16	0	60
African American (%)	21	30	0	87
Hispanic (%)	64	30	9	99
Highschool (%)	29	12	0	48
College[b] (%)	16	11	1	42
Graduate[a]/ Edu Attainment (%)	11	12	0	42
Citizen (%)	73	13	39	100

[a] "Graduate" includes graduate degree or professional degrees.

[b] "College" includes associates and bachelor degrees.

3.3 K-means clustering of 311 data

We took clustering approach to group together tracts that share similar 311 request patterns by implementing K-means clustering. In order to represent how people in each tract make distinct types of 311 service, we created vectors of service requests. We created a vector representing 311 service requests in a tract by using the Category Type field in 311 data. We followed Wang et al. (2017)'s approach to generate this vector [18].

Each t-th component of the vector will be the percentage of each type of requests t among all the requests made with a given area (census tract) c. Specifically, let the total number of service requests of each type t within an census tract c be s(c, t) and let $s(c) = \sum_t s(c,t)$ be the total number of service requests in the census tract c. Then a vector S(c) = (s(c, t)/s(c), t = 1..T), where T is the total number of service request types, serves as a signature of the location's aggregated 311 service request behavior [18].

One required input for K-means clustering is number of clusters to create. K-means clustering is an unsupervised machine learning method, thus it is not required to have a target variable (sometimes called gold standards or labeled data) to map out with the input variables. We only guess the best groupings of the instances based on their similarities [9]. In order to make a better "guess" regarding how many clusters to create, we calculated within sum of squares. Within sum of squares is calculated based on the smallest distance within cluster, and maximize distance between clusters. Lower within sum of squares indicate better performance in clustering. Within sum of squares values in Figure 3 suggest that between four and eight clusters create best groupings of the request patterns. After manually investigating clusters within this range, we decided to create six clusters because it is small enough to manually interpret the characteristics of clusters and big enough to distinguish clusters in a more informative way.

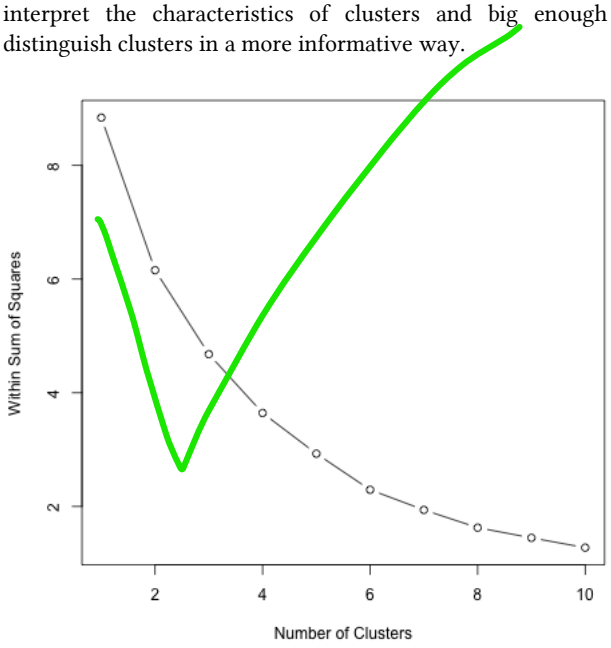


Figure 3: Within Sum of Squares by Number of Clusters

4 Findings

4.1 Phone call is the major methods for 311 requests

The sweeping majority of requests are submitted through phone calls. Departments of City of Miami also submit some of the major requests using internal systems. For example, Departments Service Request category explains 12% of the total requests. This type of service request is created by outside interfaces such as Building Department, which are then interfaced with 311 HUB. Requests through smart phones (iPhone and Android), web, email, or social media explains very small proportion of total service requests (see Table 2).

Table 2: Method 311 Requests Received

Method Received	Freq.	Percent
Phone	48,869	76
Departments Service Request[a]	7,526	12
City of Miami Neighborhood-Enhancement-Team	1,689	3
iPhone	1,450	2
Web	1,340	2
In-house	1,011	2
Email	978	2
Android	555	1
Others[b]	834	1

[a]Requests made through outside interfaces by public employees.

[b]Others include, requests made through fax, walk-in and social media.

4.2 Requests by category types

Overall, total number of requests has been stable but with a decreasing trend. Number of requests made on 2017 has decreased by 9%, compared to that of 2013 (see Table 3 and Figure 4). Figure 4 shows that "Animal Service" and "Road and Bridges" related requests have declined since 2014. Requests on "Community Code Enforcement," on the other hand, has been increasing.

Table 3: Descriptive statistics of the 311 Data

Year	Total	Request	Request
2013	13,244	9	92
2014	13,792	9	91
2015	11,451	9	102
2016	12,086	9	104
2017	11,844	10[a]	119
Jan. 1 -	1,835	8	88
Total	64,252		

[a] A new category called Communications Department is created in 2017 right after hurricane Irma hit the area.

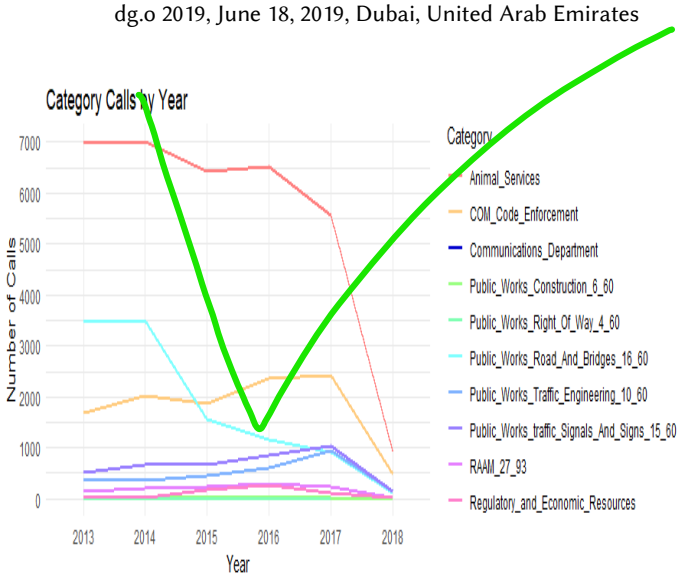


Figure 4: Frequency of requests per each categories by Year

In terms of Category Type, Animal Services includes the majority requests (52%). Code Enforcement and Public Works on Road and Bridges explain 17% each (Table 4).

Table 4: Top-ten the most frequent 311 requests by category

Category	# of rows	Percent
Animal Services	33,478	52
Community Code Enforcement	10,841	17
Public Works: Road and Bridges	10,745	17
Public Works: Traffic Signals and Signs	3,935	6
Public Works: Traffic Engineering	2,908	5
RAAM_27_93	1,115	2
Communications_Department	630	1
Regulatory_and_Economic_Resources	609	1
Public_Works_Construction_6_60	78	0
Public_Works_Right_Of_Way_4_60	13	0

In terms of Issue Type field, which is a granular level of category sub-division, illegal dumping and litter (17 %), stray (12 %), and pet account update (11%) are the most frequently requested items (Table 5).

Table 5: Top-ten the most frequent 311 requests by issue

Issue Type	Category	# of rows	Percent
Illegal dumping/litter	Community Code Enforcement	10,841	17
Stray/ dog-at-large	Animal Services	7,648	12
Pet account update	Animal Services	7,132	11
Graffiti on county property	Public Works Road and Bridges	5,204	8
Injured animal	Animal Services	3,667	6
Animal cruelty investigation	Animal Services	3,218	5
Stray dog pick up	Animal Services	2,526	4
Power road and bridge	Public Works Road and Bridges	1,979	3
Pothole	Public Works Road and Bridges	1,784	3
Animal bite to a person	Animal Services	1,770	3

We further analyzed number of requests by tract and visualized the information in Figure 5. Some areas in the northern part of the city make more requests per person ((right) Figure 5). Racial distribution in Figure 6 shows a high black population in the north, a high Hispanic population in most of the ~~area of~~ south, middle and west regions, and a high white population on the coast. To make an inference based on the information visualization in Figures 5 and 6, people living in areas with a high Hispanic population seem to make less 311 requests, which is similar to findings by Clark et al. [3].

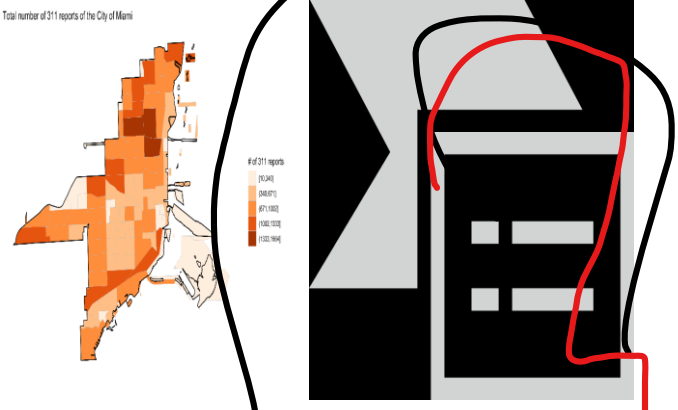


Figure 5: Frequency of 311 Calls per tract

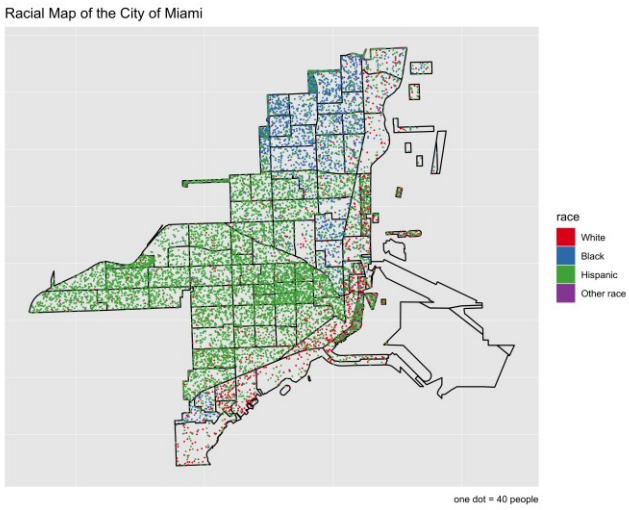


Figure 6: Geographical display of racial composition (dot density map)

Note: color of a dot represents each race and one dot explains 40 residents in the tracts.

4.3 Hurricane Irma and 311 requests

We investigated to what extent 311 service requests are influenced by natural disasters in the area. The City of Miami was heavily impacted by hurricane Irma, which hit the area on September 10th of 2017. Some of the city experienced powerful wind and rain and suffered from flooding [5]. The simple graph of 311 request frequency in Figure 7 shows a high influx of 311 requests right after Irma.

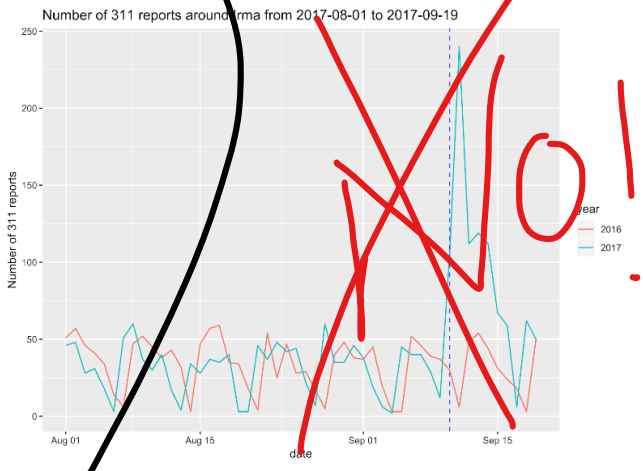


Figure 7: Comparing number of 311 requests on 2016 and 2017: the blue line is the date (September 10th, 2017) hurricane Irma arrived in the City of Miami

We further investigated frequency of 311 requests per category. The Communications Department category type was created the day Irma hit the area (September 10th, 2017), and the

majority of requests (among 630 requests total in the category) were made within the first five days after the hurricane. Figure 8 shows that the majority of the requests are regarding Communications Department during the hurricane Irma emergency.

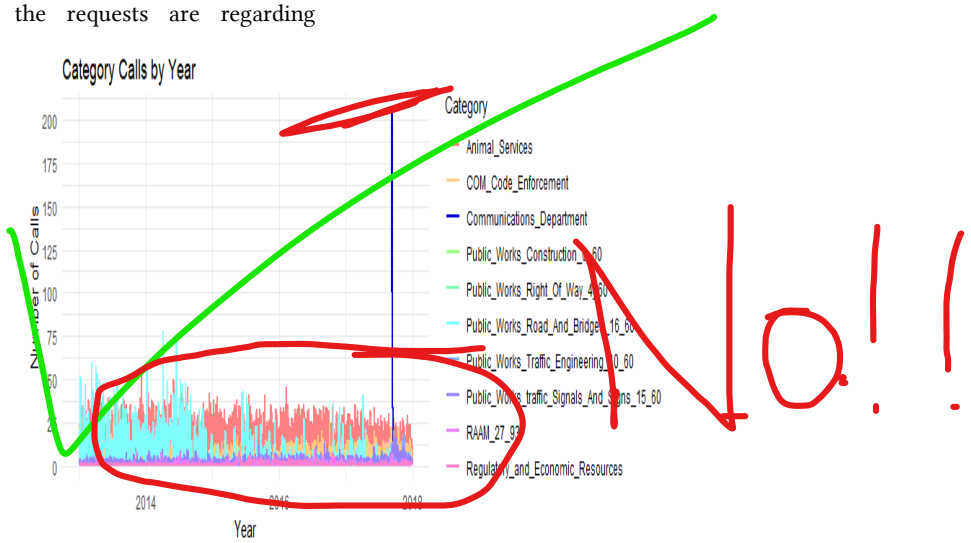
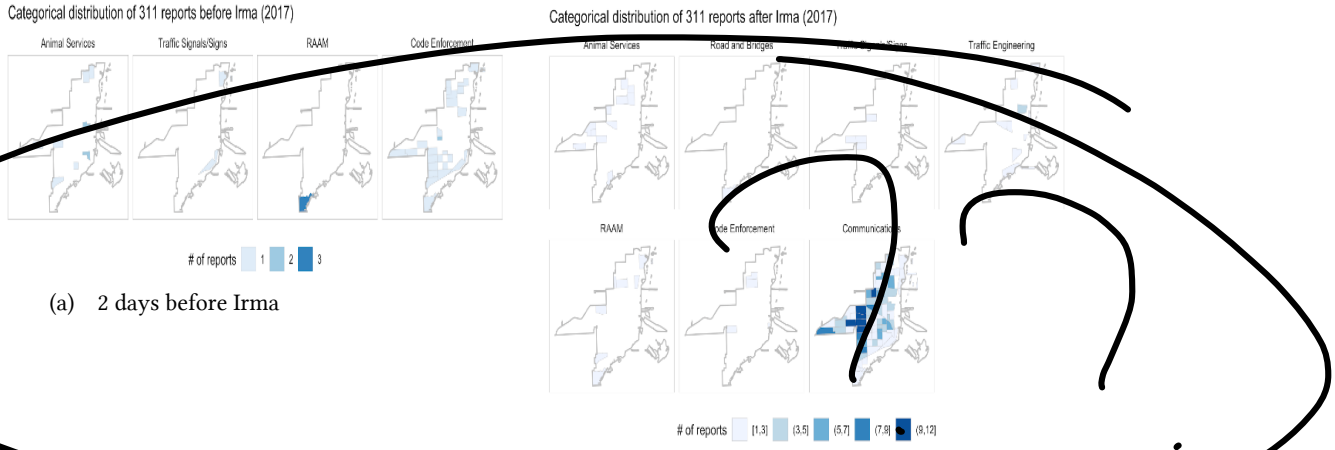


Figure 8: 311 Calls by year in relation with hurricane Irma (2017): Call frequencies per category

We then visualized the information on maps to understand if any specific areas had specific issues right after the hurricane. Figure 8(a) shows frequencies of 311 requests 2 days before the hurricane. Compared to the Figure 8(a), Figure 8(b) shows that Communications Department related requests were unusually high after Irma, in specific areas.



(a) 2 days before Irma

(b) 2 days after Irma

Figure 9: Before and After Irma by Category (2017)

4.4 service request patterns and socio-economic status

Using K-means clustering, we grouped tracts that share similar call patterns. As stated above in 3.3 K-means clustering of 311 data, we created six clusters: Cluster 1 included 6 tracts, Cluster 2 with 8, Cluster 3 with 45, Cluster 4 with 13, Cluster 5 with 11 and Cluster 6 with 22 tracts. Figure 11 shows the centers of each cluster, which is composed of frequency distributions of categories.

Using the K-means clustering results, we investigated to what extent 311 request patterns can reveal socio-demographic structures. As a result, we found that 311 service request patterns are indicative of underlying socio-demographic factors

within the area. In the following, we discuss some outstanding patterns. The findings below will be discussed based on the information visualization results in Figures 10, 11, and 12.

First, Cluster 1 is distinctive in terms of high rates of requests on **Road and Bridges** (over 50% of all requests initiated in the center of the cluster, see Figure 11). When combined with the Census data, we found that Cluster 1 includes tracts with high **African American populations** (46%, in comparison, they are 3% and 11% in Cluster 2 and 4 respectively) and the highest unemployment rate (15%). Overall, the clustering results indicate that tracts categorized in this cluster may include areas with high African American populations, high unemployment rate, and frequent issues with road and bridges.

Second, Clusters 4 and 5 includes areas with the highest median income (average of median income for Cluster 4 is $56,449 and Cluster 5 is $59,474). Although the value distributions for the center of Clusters 4, and 5 are similar as is presented in Figure 11, citizens in Cluster 4 make more requests on Roads and Bridges while citizens in Cluster 5 make more requests on Traffic Signals. As Figure 12 shows, overall characteristics of socio-economic status of the two clusters are similar to each other except for slight differences in the Hispanic population. Cluster 4 includes a higher portion of Hispanics compared to Cluster 5.

Third, Cluster 2 is unique in terms of frequent requests on Community Code Enforcement (see Figure 11, dark blue bar indicates about 40% of the calls are about Community Code Enforcement). Census data identifies that the areas included in Cluster 2 have high rates of those living under the poverty line (33%), and make the lowest average median income ($25,718). This is less than half of that of Cluster 5 ($59,474). In addition, Cluster 2 includes tracts with the highest Hispanic population (91%), the lowest citizen ratio (53%) and the lowest college graduate rates (10%). This indicates, income, poverty, race, citizenship, and education may be associated with living conditions that may attribute to high numbers of requests regarding Community Code Enforcement.

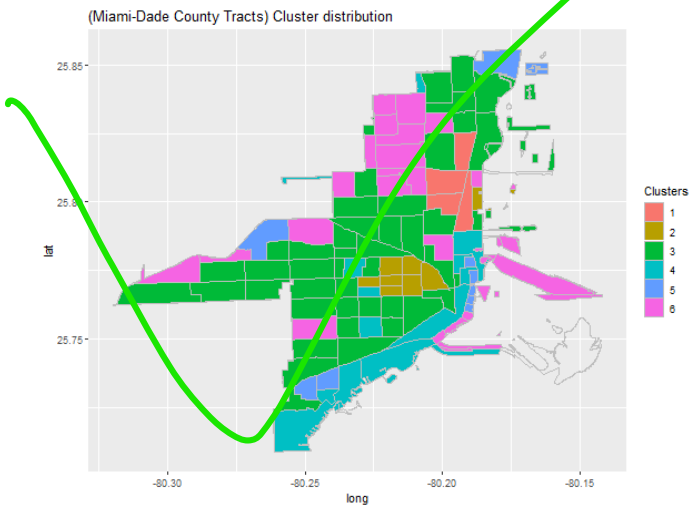


Figure 10: Geographical Presentation of Six Clusters based on category distribution of 311 requests

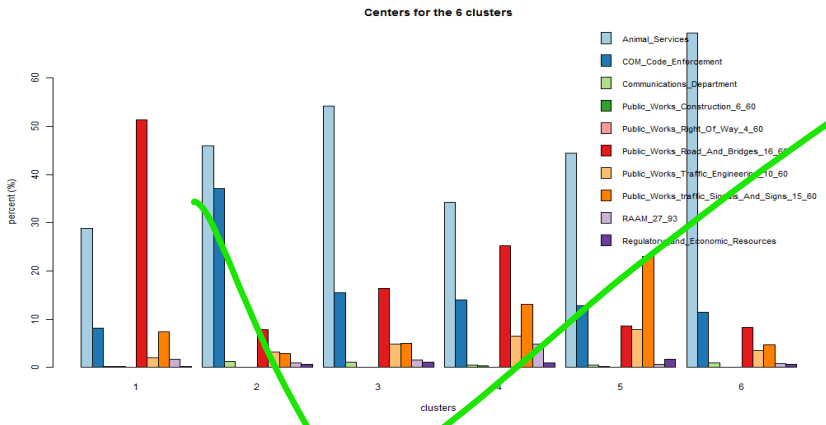


Figure 11: Cluster 5 includes outstanding patterns of Public

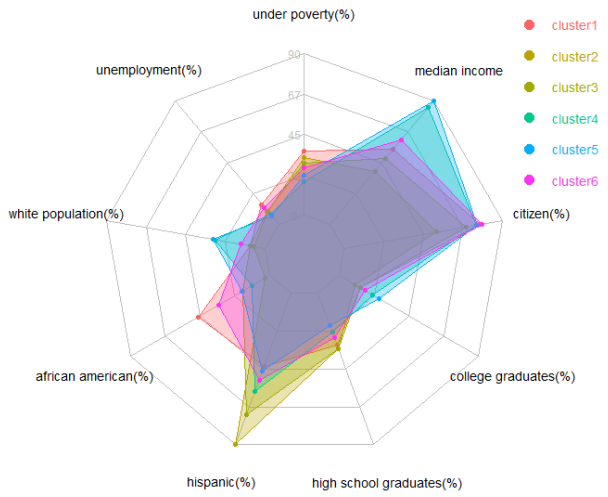


Figure 12: Socio-demographic and economic characters of the six clusters

Note: in order to visualize the average median income by cluster, we transformed the raw median income value to fit the ranges between 0 and 91. Median income of 91 corresponds to $59,474 average median income of Cluster 5.

5 Discussion

The focus of the discussion will be about knowing the potential benefits and limitations of using 311 data and big data analytics.

Before the discussion, we stress that the purpose of the study is to describe, not to predict the future nor to test hypotheses. We reported underlining patterns and correlations among variables for a descriptive purpose only. We cannot tell causations between socio-economic status and 311 request patterns, nor can we make a prediction. Rather, the analysis is intended to provide insights to policy makers by informing them about the underlying living conditions and socio-economic structures that might be associated with 311 requests.

5.1 What does the data tell?

The analytics results using information visualization provide us with many new insights about 311 requests in the City of Miami. First, although the expectation on use of social media and new technology adoption in government administration has been high, actual use of these new technologies for 311 requests is extremely low in the City of Miami. The majority of service requests were made through landline phone communication (76% of all requests), and requests made by internet and smartphone explain only 2 and 3 percent of total requests, respectively. These are puzzling results because Boston 311 requests data showed distinctively higher rates of new technology use (38 percent and 5 percent of the entire requests are made through internet and smartphone, respectively) [3]. The rates of new technology use between 2013 and 2018 is significantly lower in the City of Miami compared to that of Boston several years ago. This shows the big discrepancies in adoption of new technologies depending on the size of cities. Our discussion with the domain experts in local governments revealed that resource restriction and a lacking level of infrastructure, hinders active adoption of new technologies for 311 requests.

Second, based on the descriptive statistics and information visualization, we found that the frequency of service requests have been declining. In addition, we found that Animal Service, Code Enforcement, and Road and Bridges related requests explain 86% of total requests. These requests, especially Animal Service related requests, may be related with the regulation, which requires all pet owners to register pets and update shots records annually. In a more granular level using the Issue Type field, we found that illegal dumping/litter, stray dog, and pet account update are the top three issue types. Together, these issue types explain 40% of total requests. Graffiti on county property, road and bridge, as well as pothole are the three major requests categorized in Public Works Road and Bridges Category. These results show a contrast from the results from NYC where streets, trees, sidewalks, noise, and graffiti were the major requests [14]. This highlights the importance of conducting analysis of 311 data by local government to make decisions that are relevant to the local context. 311 requests patterns reflect the space where the public is situated in. By analyzing City of Miami 311 data and providing detailed descriptions of the most highly requested items, we contributed new knowledge regarding the population and its use of the Miami 311 service. This new knowledge shows that citizen engagement with the 311 service is on a decreasing trend).

Third, the information visualization of 311 requests and hurricane Irma unveils an influx of communication-related requests right after the hurricane in specific areas. This can inform the local government to be prepared for emergencies by knowing possible high demands on specific requests in specific areas.

Fourth, from the analysis using K-means clustering, we learned that 311 request patterns might be able to explain living conditions and socio-economic status of populations. For example, relatively poor and foreign Hispanic populations seem to live in areas, where high volumes of Community Code Enforcement requests are made. Also, in areas where high numbers of infrastructure (road and bridges) related requests are made, we observed high Black populations and high unemployment rate.

The major takeaway from these findings is that data captures reality and thus, the data is a reflection of current social and living conditions. People who live in a poor community composed of newly immigrated population may end up living in cheap houses where high foreclosure and high code enforcement activities take place [16]. Our findings show the value of descriptive outcomes in providing new insights about citizen demands and living conditions as a whole.

Although we found some useful insights, there are pitfalls of 311 data and the analytics results. As more 311 data opens, it is important to understand the pitfalls of the data and analytics results in order to improve the future data quality and to make better informed decisions based on analytics results.

5.2 What does the data not tell?

Theoretically, 311 request patterns are driven by both individual attributes of callers as well as the contextual conditions of the area. One major limitation of 311 request data is ***a lack of individual level data*** [20]. This makes it almost impossible to find specific reasons why the pattern occurs based on the findings only. Although we combined 311 data with Census data to make inferences per tract, what we have learned is an aggregate value. Therefore, the results do not fully answer to why the patterns occurs. In order to answer to this question, we need field studies including citizens and government workers.

In addition, ***the data cannot tell us contextual information***, in which the data is collected. Contextual information such as, regulative environments and residential environments, partially define the content and structure of the data. However, it is not possible to capture complete information in a dataset. In order to understand contextual information better, we have closely communicated with domain experts, directors of the 311 department in the City of Miami as well as Miami-Dade County, in the process of analysis. We learned that regulations and government structures heavily influence to the nature of data. For example, we learned that each municipality under Miami-Dade County has different regulations, which are reflected in patterns and frequencies of 311 requests in each municipality. Also, we learned that lacking cutting edge technical infrastructure is the major reason why 311 requests are mainly received through phone. The patterns we captured are interesting, but following field studies can fill many gaps in order to explain why we see these patterns.

Data quality and interoperability are still challenging for big data analytics, and are not fully highlighted in the report. We spent about 75-80% of efforts and time on data cleaning and merging, which is always the most time consuming and challenging process for big data analysis [9]. Especially, data merging is necessary for developing rich analysis because one dataset can only present very limited information. We used Census data and 311 call data, as well as City of Miami shape files for the analysis and visualization. Each dataset included pre-defined fields designed to achieve the goal of the original data collection. For example, 311 call data include address information, but Census data do not have specific address available. As such, some semantically identical fields are defined differently, which can make two datasets not interoperable. In addition, the basic unit of data is often different (i.e., Census data use tract as the base unit, while 311 call data is organized by individual request). In our case, we created a "Tract" variable in 311 data as a workaround in order to merge Census and 311 data. We used location information included in 311 data such as Latitude and Longitude to match with the Census Tract variable. Again, coming up with a final version of cleaned and merged data set required multiple rounds of trial and error until we came up with a final clean dataset for analysis. From the initiation to the analysis, it took about one year to clean, analyze, and validate the results.

Validation of the process. One programmer programmed the entire process initially. In order to validate the process, another programmer replicated the process to make sure the process was flawless.

6 Conclusion

The major contribution of the study is in making suggestions on best practices of using 311 data to extract useful information and to visualize that same information for policy decision making. Best practices involve suggesting the data analytics strategy and providing interpretation of the descriptive outcomes. In addition, we discussed pitfalls of 311 data and limitations of analytics results. We further made some resourceful solutions to deal with some limitations of 311 data. Finally, we expect the insights provided here can help policy decisions taken by the City of Miami government.

As stated above in the discussion, this study includes some limitations that require future study. Future study will include some statistical tests (t-test, or analysis of variance) to make the findings more robust. In addition, interview studies with residents and government workers can explain some causations this study could not explain. Adding these additional components to a 311 study, will produce results that are more useful for decision makers. For technical communities, new technical initiatives to support local governments will be highly beneficial to local governments because often local governments do not have a capacity to purchase high-cost analytics software required for collecting and analyzing 311 data [2].

7 ACKNOWLEDGMENTS

Loni Hagen was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2017S1A3A2066084).

REFERENCES

[1] Big data: The next frontier for innovation, competition, and productivity | McKinsey & Company: *http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation*. Accessed: 2013-08-15.

[2] Chatfield, A.T. and Reddick, C.G. 2018. Customer agility and responsiveness through big data analytics for public value creation: A case study of Houston 311 on-demand services. *Government Information Quarterly*. 35, 2 (Apr. 2018), 336–347. DOI:https://doi.org/10.1016/j.giq.2017.11.002.

[3] Clark, B.Y. et al. 2013. Coproduction of Government Services and the New Information Technology: Investigating the Distributional Biases. *Public Administration Review*. 73, 5 (Sep. 2013), 687–701. DOI:https://doi.org/10.1111/puar.12092.

[4] EMC Education Services 2015. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. John Wiley & Sons, Inc.

[5] Flechas, J. and Smiley, D. 2017. Irma hit downtown Miami — and turned its biggest streets into rivers. *miamiherald*.

[6] Gilbert, E.W. 1958. Pioneer Maps of Health and Disease in England. *The Geographical Journal*. 124, 2 (1958), 172–183. DOI:https://doi.org/10.2307/1790244.

[7] Hagen, L. 2018. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management*. 54, 6 (2018). DOI:https://doi.org/https://doi.org/10.1016/j.ipm.2018.05.006.

[8] Hagen, L. et al. 2017. Crisis Communications in the Age of Social Media: A Network Analysis of Zika-Related Tweets. *Social Science Computer Review*. (Aug. 2017), 089443931772198. DOI:https://doi.org/10.1177/0894439317721985.

[9] Kelleher, J.D. and Tierney, B. 2018. *Data Science*. The MIT Press.

[10] Kerren, A. et al. eds. 2008. *Information Visualization: Human-Centered Issues and Perspectives*. Springer-Verlag Berlin Heidelberg.

[11] Kitchin, R. 2014. The data revolution: Big data, open data, data infrastructures and their consequences. Sage.

[12] Lavalle, S. et al. 2011. Big data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*. 52, 2 (2011), 13.

[13] LibGuides: Finding Census Tract Data: About Census Tracts:*//libguides.lib.msu.edu/c.php?g=96120&p=625755*. Accessed: 2019-01-17.

[14] Minkoff, S.L. 2016. NYC 311: A Tract-Level Analysis of Citizen–Government Contacting in New York City. *Urban Affairs Review*. 52, 2 (Mar. 2016), 211–246. DOI:https://doi.org/10.1177/1078087415577796.

[15] Recht, H. 2018. *Package 'censusapi.'*

[16] Schilling, J. 2009. Code enforcement and community stabilization: the forgotten first responders to vacant and foreclosed homes. 2, (2009), 64.

[17] Ubaldi, B. 2013. Open Government Data: TOWARDS EMPIRICAL ANALYSIS OF OPEN GOVERNMENT DATA INITIATIVES. *OECD Working Papers on Public Governance; Paris*. 22 (May 2013), 0_1,1,4-60.

[18] Wang, L. et al. 2017. Structure of 311 service requests as a signature of urban location. *PLOS ONE*. 12, 10 (Oct. 2017), e0186314. DOI:https://doi.org/10.1371/journal.pone.0186314.

[19] Where does Zillow get information about my property? 2018. *http://zillow.zendesk.com/hc/en-us/articles/213218507-Where-does-Zillow-get-information-about-my-property-*. Accessed: 2018-07-07.

[20] White, A. and Trump, K.-S. 2018. The Promises and Pitfalls of 311 Data. *Urban Affairs Review*. 54, 4 (Jul. 2018), 794–823. DOI:https://doi.org/10.1177/1078087416673202.

[21] Zuiderwijk, A. et al. 2014. Innovation with open data: Essential elements of open data ecosystems. *Information Polity*. 19, 1 (Jan. 2014), 17–33. DOI:https://doi.org/10.3233/IP-140329.