A Case Study on using Unstructured Data Analysis Methods to identify local Covid-19 Hotspots

Nancy Joseph
Electrical and Computer Engineering
FAMU-FSU College of Engineering
Tallahassee, USA
nancy1.joseph@famu.edu

Shonda Bernadin, PhD.
Electrical and Computer Engineering
FAMU-FSU College of Engineering
Tallahassee, USA
bernadin@eng.famu.fsu.edu

Deidra Hodges, PhD.
Electrical and Computer Engineering
University of Texas at El Paso
El Paso, USA
drhodges@utep.edu

Praveen Sekhar, PhD.
School of Engineering and Computer
Washington State University
Vancouver, USA
praveen.sekhar@wsu.edu

***Abstract*—The purpose of this work is to use unstructured data analysis methods to identify Covid-19 hotspots within local communities using publicly-available health and socioeconomic data. Consequently, a detailed analysis showing which local communities are most impacted by Covid-19 in the North Florida region is conducted based on zip code profiling. This work contributes to the knowledge and discovery of the impact of Covid-19 on lower income communities.**

Keywords—unstructured data, Covid-19, health data, demographics, socio-economic data

I. Introduction

The Covid-19 pandemic has disrupted the lives of nearly every human being across the globe. Whether it has been through contracting the virus, knowing someone who has contracted the virus, having to quarantine, following stay-at-home orders, or pivoting to online education, this pandemic has taken its toll on humanity. The authors wanted to study how the Covid-19 pandemic disproportionately impacts the local communities that were under-resourced and distressed even before the pandemic began. How is Covid-19 affecting local communities, especially lower income communities, in and around, Leon County? What kind of data can be accessed to conduct this study?

Big data is everywhere and its availability in the public domain is a non-issue. However, determining which data to collect, cleaning the data, and preparing it for analysis is more challenging. In fact, most public data is unstructured for meaningful analysis. Thus, the problem becomes how do we use unstructured health and socioeconomic data from the public domain to obtain meaningful relationships and information. For this paper, a case study on using unstructured public socioeconomic data from Zip Data Map [1] for zip code profiling and Covid-19 data from Florida Department of Health [2] is described.

Motivation

The motivation of this case study is to see the disparity within the local communities of those impacted by Covid-19. Also, to verify the consensus of the national trends regarding the impact of Covid-19 on black and brown communities. Furthermore, to help public officials have a better understanding of which communities should get priority during the vaccine rollout. This method also presents a way to identify hotspots within Leon County by profiling different zip codes.

II. Background

Covid-19 is an airborne illness caused by a virus that can spread by contact [3]. The first outbreak occurred in Wuhan, China in 2019 and then spread across the globe by human carriers for 2020 [4]. The virus carriers have two to fourteen days to develop any symptoms. It can have the carrier to experience no symptoms to mild or extreme symptoms. The mild symptoms can be headaches, fatigue, loss of taste or smell, cough, and more. Severe symptoms include difficulty breathing, inability to stay awake, constant pressure on the chest area, and more [3]. The virus can have devastating effects on those individuals with pre-existing health conditions such as heart-disease and diabetes. Research shows that people of color are more impacted by pre-existing conditions, and subsequently, Covid-19 will impact communities of color disproportionately, as well.

Unstructured data is any data that can be found through mediums, such as the internet, newspapers, and more [5]. This type of data can be in the form of the content in documents, texts, or social media posts. Therefore, there is always a large volume of unstructured data available. This data is essential for business intelligence to make informed decisions that can impact a wide range of industries. This data can later be structured with the end goal in mind. According to the authors in [5], the process of mapping unstructured data is the following:

1) Extraction
2) Classification
3) Development of Repositories
4) Data Mapping

Extraction is the first process for the mapping of unstructured data by identifying the source and format of the data. There are two types of extraction: fact and entity [4]. Fact extraction gains the context and facts within the unstructured data. Entity extraction allows grabbing important categorical information. This includes but not limited to the following

SoutheastCon 2021 | 978-1-6654-0379-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/SOUTHEASTCON45413.2021.9401921

information: location, title, names, type, date, and more. Classification is used to group entities by content and format. This would eventually be used to create model predictions. Development of repositories is used to manage and store the classified data. It allows the birth of specific repositories classified by audio, text, video, and image [5]. Data mapping is the final step that converts unstructured data to structured data. The unstructured data that was classified and placed into its respective repositories is mapped by themes. This allows for better decision making after sorting and mapping all that data.

Comparable Analysis Methods

The authors in [6] looked at methods that used unstructured text data (e.g. emails, survey responses, media comments, etc.) to analyze public health and safety for a smart city infrastructure in a developing nation. A smart city is a community that connects information and technology of physical devices as one entity [7]. It allows the people in the community to interact and monitor the conditions of various objects and buildings in the city. The most prominent examples of the duties carried out in smart cities are in transportation systems, utilities, public health, safety and more. The focus is on developing a smart infrastructure for the public safety aspect of a lower socioeconomic community in East London [6].

Natural language processing (NLP) techniques were applied to the unstructured text data. The data used crowdsourcing to collect in East London for a smart city project. Crowdsourcing is the establishment of sharing services and ideas from an online community [8]. A prime example of crowdsourcing would be Wikipedia, which allows users to add information about anything if it is well sourced. The outcome of the smart city project in East London is using video surveillance provided by the local authorities to extract crucial data that would be used. Furthermore, that unstructured data was mapped into structured data that would help to inform public safety decisions.

III. Methods

In this case study, the methodology used for extracting meaningful information from the unstructured data is as follows [9]:

1) Develop a goal for using the unstructured data.
2) Obtain the unstructured data.
3) Clean the data.
4) Structure the data.
5) Analyze the data.
6) Visualize the results.
7) Draw conclusions.

There needs to be a clear goal when dealing with unstructured data because there is so much information that it can be overwhelming. This is a good time to propose the questions that are going to be explored using the unstructured data. Thus, obtaining the correct source to answer the questions is essential. The initial question for this research was how does Covid-19 impact the local communities in Leon County? Also, what kind of data can be used to gather the important information for this study? With that in mind, the data is extracted from the Zip Data Map for socioeconomic information and FL Department of Health for the Covid-19 data. Excel was used to clean and structure the data into a single spreadsheet. As for the data analysis, a heat map was generated using the correlation between the data points. In order to visualize the results, graphs and charts were used between the zip codes and the different variables.

IV. Results and Discussion

The data variables that were collected from public datasets used for profiling are represented as the following:

- VAR_1:ZIPCODES for various counties
- VAR_2:CURRENT_POP = population (as of 2020)
- VAR_3:RACIAL_MAJ = racial majority
- VAR_4:RACIAL_MAJ_PERCENT = racial majority percentage
- VAR_5:PUB_SCHOOL_MAJ = public school racial majority
- VAR_6:PUB_SCHOOL_MAJ_PERCENT = public school racial majority percentage
- VAR_7:UNEMPL = unemployment
- VAR_8:MEDIAN_INCOME = median income
- VAR_9:SCHOOL_PERF = school performance
- VAR_10:AVG_COMMUTE_TIME = average commute time
- VAR_11:COVID_0222021 = COVID data as of February 22, 2021
- VAR_12:COVID_0222021_NORM = COVID data as of February 22, 2021 normalized to current population

Table 1 shows the 32301-32317 zip code profiles based on most recent data as reported by zip data maps. From table 1 zip code 32303 has the highest population of people closely followed by 32304. The racial majority in both zip codes is 'Caucasian', however, the racial majority in their corresponding public schools is 'African-American/Black' (AA/Black). In zip code 32304, AA/Black represents 84.4% of the public school population and about 50% of the current total population. Zip code 32317 is the least populated with a racial majority and public school majority of 'Caucasian'.

TABLE I. Leon County Socioeconomic Data as Reported by Zip Data Map

COUNTY	VAR_1	VAR_2	VAR_3	VAR_4	VAR_5	VAR_6	VAR_7	VAR_8	VAR_9	VAR_10
LEON	32301	29321	AA/BLACK	47.32	AA/BLACK	79.6	5.8	36207	BELOWAVG	15.1
LEON	32303	47359	CAUCASIAN	58.39	AA/BLACK	58.4	5.8	47325	AVERAGE	18.9
LEON	32304	46145	CAUCASIAN	49.92	AA/BLACK	84.4	5.8	16916	BELOWAVG	15
LEON	32305	20073	AA/BLACK	58.07	AA/BLACK	65.7	5.8	36648	BELOWAVG	23.5
LEON	32308	21586	CAUCASIAN	63.09	CAUCASIAN	46.6	5.8	58834	ABOVEAVG	17.1
LEON	32309	30159	CAUCASIAN	76.18	CAUCASIAN	75.9	5.1	76857	EXCELLENT	22.6
LEON	32310	17402	AA/BLACK	47.91	AA/BLACK	65.6	5.1	28375	BELOWAVG	24.9
LEON	32311	17493	CAUCASIAN	61.62	CAUCASIAN	45.1	5.8	51675	ABOVEAVG	20.9
LEON	32312	31869	CAUCASIAN	77.61	CAUCASIAN	71.5	5.8	93121	EXCELLENT	23.5
LEON	32317	14067	CAUCASIAN	73.17	CAUCASIAN	59.9	5.8	81191	EXCELLENT	24.6
CLARK, WA	98686	17385	CAUCASIAN	80.95	CAUCASIAN	62.5	7	75639	AVERAGE	23.1
EL PASO, TX	79902	21236	HISPANIC	81.59	HISPANIC	91.8	8.9	29000	AVERAGE	20.7

Fig. 1 gives a heatmap of correlations drawn between the Covid-19 data and economic and demographic profiles by zip code profiling in Leon County.

	VAR_1	VAR_2	VAR_3	VAR_4	VAR_5	VAR_6	VAR_7	VAR_8	VAR_9	VAR_10	COVID	COVID_N
VAR_1	1.00											
VAR_2	-0.30	1.00										
VAR_3	0.29	-0.26	1.00									
VAR_4	0.59	-0.30	-0.07	1.00								
VAR_5	0.17	0.24	0.83	-0.29	1.00							
VAR_6	0.24	0.33	0.56	0.08	0.64	1.00						
VAR_7	0.79	-0.18	0.58	0.53	0.51	0.41	1.00					
VAR_8	0.07	-0.23	-0.58	0.69	-0.81	-0.35	-0.17	1.00				
VAR_9	-0.12	-0.17	-0.52	0.67	-0.70	-0.30	-0.12	0.86	1.00			
VAR_10	0.17	-0.58	0.03	0.48	-0.27	-0.19	-0.04	0.48	0.37	1.00		
COVID	0.58	0.16	0.62	0.31	0.72	0.66	0.88	-0.42	-0.24	-0.25	1.00	
COVID_N	0.00	0.77	0.20	-0.22	0.63	0.62	0.32	-0.58	-0.36	-0.68	0.67	1.00

Fig. 1. Heatmap of correlations with Covid-19 data (as of 2/22/21) Economic and Demographic Profiles by Zip Code in Leon County, 32301-32317, Clark County, WA and El Paso County, TX

Regression analysis was performed to determine which, if any, variable achieved statistical significance at $p \leq 0.05$ with current Covid cases. Notice that in table 2, the variables that had p-values more than 0.05. More analysis should be conducted to further investigate these results.

TABLE II. REGRESSION ANALYSIS ON COVID-19 CASES IN LEON COUNTY ZIPCODES

Variable	*p-value*
ZIPCODE	0.4840874
CURRENT_POP	0.5212518
RACIAL_MAJ	0.7292359
RACIAL_MAJ_PERCENT	0.9149744
PUB_SCHOOL_MAJ	0.9943149
PUB_SCHOOL_MAJ_PERCENT	0.5396035
UNEMPL	0.4278719
MEDIAN_INCOME	0.2813724
SCHOOL_PERF	0.3270075
AVG_COMMUTE_TIME	0.9096922

There is a strong positive correlation of 0.77 between the population size and the normalized number of Covid-19 cases as shown in Figure 1. In addition, there is a strong negative correlation of -0.68 between the normalized number of Covid-19 cases and average commute time to work.

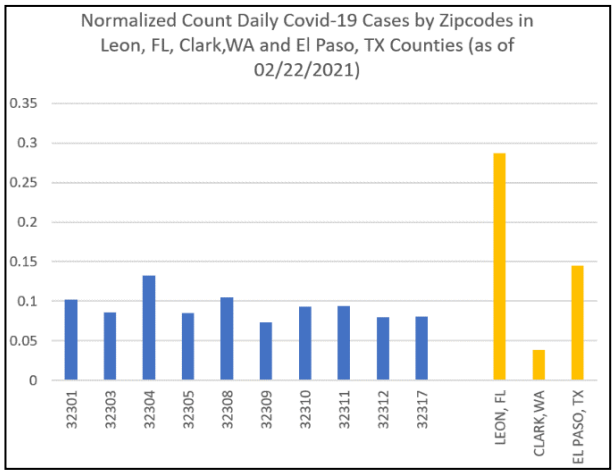

Fig. 2. Normalized Daily Covid-19 Cases data (as of 2/22/21) by Zip Code in Leon County, Clark, WA, and El Paso, TX

According to Fig. 2, zip code 32304 has the highest number of Covid-19 cases. Meanwhile, zip code 32309 has the lowest number of cases in Leon County. Zip Code 32317 has the second lowest number of cases. When comparing Leon county to El Paso county and Clark county, Leon county has the highest number of normalized Covid-19 cases.

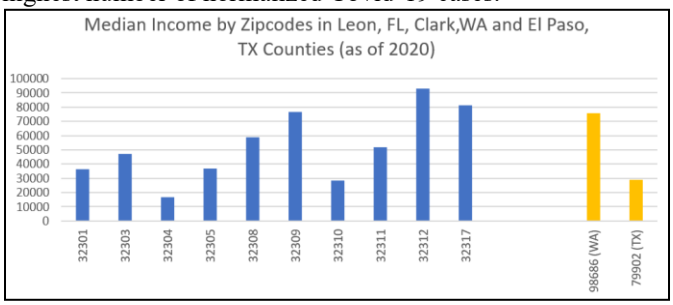

Fig. 3. Median Income Data by Zip Code in Leon County, Clark, WA, and El Paso, TX

According to Fig. 3, zip code 32304 has the lowest median income. In contrast, zip code 32317 has the second highest median income. Zip code 32309 has the third highest median income. The highest median income is 32312. The zip code in Clark county has a higher median income in contrast to the zip code in El Paso county.

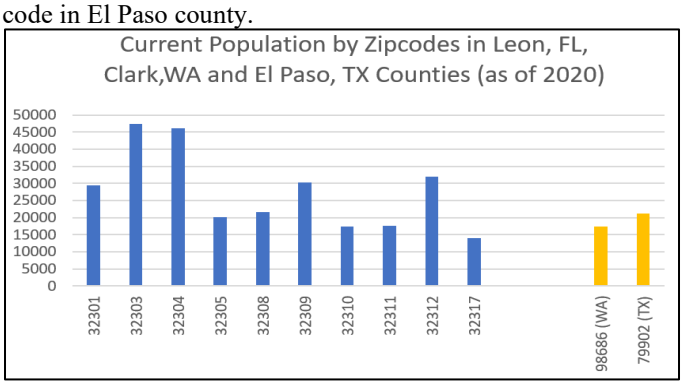

Fig. 4. Current Population by Zip Code in Leon County, Clark, WA, and El Paso, TX

As shown in Fig. 4, the zip codes with the highest populations are 32303 and 32304. The zip codes with lowest populations are 32317, 32310, and 32311. The zip code 79902 in El Paso has a higher population than zip code 98686 in Clark, WA.

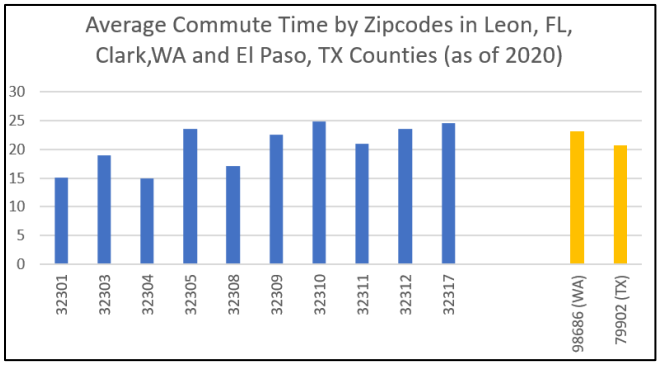

Fig. 5. Average Commute Time by Zip Code in Leon County, Clark, WA, and El Paso, TX

Fig. 5 shows the average commute time is broken down by zip code. The zip codes with the highest average commute time are 32317 and 32310. The zip codes with lowest average commute time are 32304 and 32301. There is a relatively high average commute time in Clark, WA compared to El Paso, TX.

One limitation of this study is that the public data that was used for daily Covid-19 cases were available only as a cumulative count for each zip code. They were not disaggregated according to race, age or gender, which makes the analysis more challenging when investigating vulnerable communities. Consequently, to adequately study more detailed issues of how Covid-19 impacted communities of color in the north Florida region, more disaggregated Covid-19 datasets are required. Thus, more robust data collection methods may be used, or perhaps, adequate disaggregated Covid-19 datasets may not be easily available in the public domain. Another limitation of this study is that the daily Covid-19 case count is only current up to the day the data was pulled from the public domain, which was on February 22nd, 2021 for this study. A more automated method is needed to continuously collect Covid case updated data from the Web.

V. Conclusion

The median income and average commute time have a negative correlation with the number of daily Covid-19 cases. This may suggest those people who live closer to work and have a lower income have a shorter commute time because they are essential workers, or work at a job where time-off during the pandemic was not an option for them. This may also indicate an increase in the number of cases in that zip code because essential workers may be in constant contact with more individuals making it easier to contract the virus. Workers in the higher income communities may have the option of working from home, thus, minimizing the number of potential contacts and reducing their chances of contracting Covid-19. This was demonstrated in the zip code in Clark, WA and the two zip codes in Leon county that had higher median income, but a lower number of normalized Covid-19 cases. In future work, data collection of more disaggregated Covid-19 data will be conducted to gain deeper insights to the impact of the pandemic on various zip codes. In addition, more automation will be used to keep Covid-19 case counts current and up to date.

References

[1] "Zipdatamaps - Interactive ZIP Code Maps and Data," zdm. . https://www.zipdatamaps.com/. (Accessed on 01/08/21).

[2] Florida's COVID-19 Data and Surveillance Dashboard. [Online]. https://experience.arcgis.com/experience/96dd742462124fa0b38ddedb9b25e429. (Accessed on 01/08/21).

[3] "Coronavirus Disease 2019 (COVID-19)," Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/. (Accessed on 01/01/21).

[4] "How Coronavirus Spreads," Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html. (Accessed on 01/01/21).

[5] M. F. Abdullah and K. Ahmad, "The mapping process of unstructured data to structured data," *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)*, Kuala Lumpur, 2013, pp. 151-155, doi: 10.1109/ICRIIS.2013.6716700.

[6] A. J. Currin, "TEXT DATA ANALYSIS FOR A SMART CITY PROJECT IN A DEVELOPING NATION," Ph.D. dissertation, Dept. Information System, Univ. of Fort Hare, City of Univ., Alice, Eastern Cape, South Africa, 2015.

[7] A. FOUNOUN and A. HAYAR, "Evaluation of the concept of the smart city through local regulation and the importance of local initiative," Presented at 2018 IEEE International Smart Cities Conference (ISC2), Kansas City, MO, USA, 2018, pp. 1-6.

[8] "Crowdsourcing," *Merriam-Webster*. https://www.merriam-webster.com/dictionary/crowdsourcing. (Accessed on 02/20/21).

[9] "How to Structure and Analyze Unstructured Data in Real Time," MonkeyLearn Blog, 10-Sep-2020. https://monkeylearn.com/blog/unstructured-data-analysis/. (Accessed on 01/01/21).