Predicting the number of COVID-19 cases by ZIP Code

A Technical Report
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

Larry Cai

April 23, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Larry Cai

Technical advisor: Mohammad Mahmoody, Department of Computer Science
Technical advisor: Luther Tychonievich, Department of Computer Science

Predicting the number of COVID-19 cases by ZIP Code

Larry Cai
University of Virginia
ljc2sh@virginia.edu

***Abstract*—The World Health Organization (WHO) declared coronavirus a pandemic on March 11, 2020 [1]. As of November 1, 2020, there are 46,386,903 COVID-19 cases globally and 1,199,500 COVID-19 related deaths [2]. In March, Guayaquil had a serious coronavirus outbreak. 778 people died on April 4, and corpses lay on the street [3]. This changed when Hector Hugo identified where COVID-19 was most concentrated by mapping 911 calls to locations [3]. Guayaquil mayor, Cynthia Viteri, dispatched medical officials to treat patients in coronavirus hotspots. Health officials can contain coronavirus by effectively predicting coronavirus hotspots.**

The U.S. Centers for Disease Control and Prevention's (CDC) dashboard displays the predicted number of coronavirus cases by county and state with multiple models [4]. It is hard to pinpoint coronavirus hotspots because some counties are extensive. New York University Langone Health's City Dashboard displays neighborhood coronavirus risk in cities but does not have data on many cities such as Fairfax, Virginia [5]. By predicting the number of COVID-19 cases by ZIP Code, public health experts, epidemiologists, and policymakers can identify COVID-19 hotspots and implement preventive measures such as email warnings to residents about local outbreaks.

Keywords—machine learning, linear regression, multiple linear regression, perception algorithm

I. Introduction

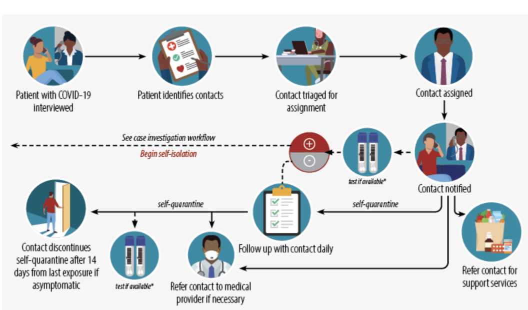

Figure 1: Contact Tracing Workflow from the CDC [6]

Currently, government officials are having a hard time containing the virus, because so many people have been infected. Daily cumulative COVID-19 cases continue to rise and current COVID-19 measures are not enough to control the virus. The government is currently using contact tracing to track the virus [6]. Contact tracing is keeping track of people and close contacts who have been tested positive for the virus [6]. Contact tracing is not effective right now because so many people are in contact with the virus; the government does not have enough people to keep up with the number of cases. In addition, people such as Dr Scott Atlas, former president Donald Trump's advisor, spread misinformation about the virus. Atlas' tweet "Masks work? NO" was removed from twitter on October 17, 2020 [7]. Therefore, it is up to local government officials to keep track of the virus and implement local measures to try to contain the virus.

My solution to this problem is to create a model to predict the future number of coronavirus cases by ZIP Code. Using this model, we can send health officials to COVID-19 hotspots and implement preventive measures such as email warnings to residents about local outbreaks.

Policy makers should be able to make informative decisions on problems that are affecting their governed region. Creating predictive models allows policy makers to visualize changes in the number of COVID-19 cases and make appropriate policies to deal with these issues. By modeling regions by ZIP Codes rather than counties, policy makers can spend less resources and implement more specific policies for each region. Counties are large, so a trend in a county may not be reflective of trends in one of its encompassing regions.

It is important for people to be aware of this issue because it is a global pandemic. Modeling the number of COVID-19 cases by ZIP Code will give people time to prepare. People can make plans and outings while avoiding COVID-19 hotspots.

There are also those that continue to ignore the dangers of COVID-19. Some of these individuals promote ‘anti-mask’ policies and fight against social distancing measures; They argue that government COVID-19 measures are just an excuse to exert control over citizens [8]. Molly McCann of the Federalist claims that “mandatory masking provides the foundation on which governments continue to justify emergency measures and rule by executive fiat, and it creates a national mood of consent that America will accept indefinite government expansion because we face a “new normal” [7]. Many people don’t understand the dangers of the virus when there are few cases in their area. By creating models of the virus, public health experts can show them the potential danger of the virus and the number of cases that will develop in the area should their current situation continue. They might be convinced to adopt COVID-19 preventative measures.

I took my data from the Virginia Department of Health (VDH) COVID-19 public use dataset which contains information on the number of COVID-19 cases per ZIP Code tabulation area. According to the United States Census Bureau “There are 898 ZIP Code tabulation areas (ZCTAs) in Virginia” [9]; ZCTAs were designated to “overcome the operational difficulties of creating a well-defined ZIP Code area by using Census blocks” [10]. I found that one of the ZCTAs was 'Not Reported', while the other was 'Out-of-State'.

II. Related Works

There already exist several public and free COVID-19 prediction models. The U.S. Centers for Disease Control and Prevention (CDC) predicts the number of cases by county and state with multiple models and displays several models together on a graph for the “National Forecast” [4]. The CDC github contains links to 65 prediction models [11]. The CDC doesn’t display models per ZIP Code. The smallest area they measure is by the county. Counties are large, making it difficult to pinpoint coronavirus hotspots and send health officials to areas for COVID-19 testing.

Data from New York University Langone Health’s City Health Dashboard (CHD) shows that the pandemic affects neighborhoods differently. CHD displays neighborhood coronavirus risk in cities and states based on 23 metrics, dividing neighborhoods by ZIP Code, but it does not have data on many cities including Fairfax, Virginia [5]. Different neighborhoods have different conditions such as racial breakdown, poverty, and local chronic conditions which affect infection rate [12].

Bin Yu from the University of Berkeley created one of the many models used in the CDC “National forecast” [11]. The group implemented 5 models to predict the number of COVID-19 deaths: a separate-county exponential predictor, a separate-county linear predictor, a shared-county exponential predictor, an expanded shared-county exponential predictor, and a demographic shared-county exponential [13]. The separate-county predictors create models based on the cumulative number of deaths in the specific county from the most recent 4-5 days [13]. The shared-county predictors predict information on a single county based on the aggregated data from multiple counties. Rather than use the number of cumulative deaths from the most recent five days, they use data starting from the third COVID-19 death in the county [13]. The expanded prediction takes into account factors such as the number of COVID-19 cases and deaths in neighboring counties while the demographic predictor includes county density and size, healthcare resources, and health demographic as variables [13].

There are many sources of information to find the number of COVID-19 cases. The Virginia Department of Health has COVID-19 cases conveniently mapped by ZIP Code [14]. The CDC, Johns Hopkins University School of Medicine (JHUSOM), and The COVID Tracking Project also have public databases. Information can be also extracted from 911 calls.

III. Running the Software

```
Enter a Virginia Zipcode: test
Not a valid number.
Enter a Virginia Zipcode: 22002
Not a valid number.
Enter a Virginia Zipcode: 22031
Enter the date you want to predict in the format "mm/dd/YYYY":04/20/2020
Before the VDH started recording cases:
Not a valid date.
Enter the date you want to predict in the format "mm/dd/YYYY":04/21/2021
Number of cases for 2021-04-21 :  2000
```

Figure 2. Program Inputs

In order to run the software with the most accurate results users must: 1. Download the latest ZIP Code public use dataset from the VDH website [14], 2. Place the dataset in the same folder as my python file (covidModel.py), 3. Open up the terminal and cd into the folder containing the python file, and 4. Run the command ‘python3 covidModel.py’. The program will prompt the user to “Enter a Virginia Zipcode” and ‘Enter the date you want to predict in the format “mm/dd/YYYY”. If the user enters an invalid input, then the program will return an error message. Once the user enters valid inputs, then the program will display the predicted number of COVID-19 cases for that ZIP Code.

IV. System Design

A. High Level Overview

The program works by extracting the ZIP Codes and dates from the VDH dataset. It then asks the user to select a ZIP Code which the program will predict the future number of COVID-19 cases for as well as predicted date. The program then sorts the data by date and finds the slope of the number of COVID-19 cases for the most recent week. It plugs the values into a linear equation to produce the model in Figure 3.

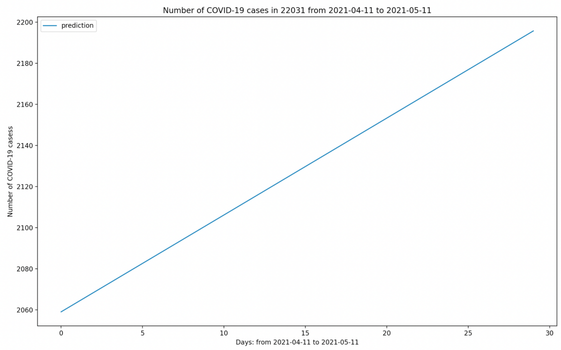

Figure 3. Model of the predicted number of COVID-19 cases from April 11, 2021 to May 11, 2021

B. Other Models

I attempted to predict the number of COVID-19 cases using multiple methods before ultimately deciding to stick with a linear model. I explored using reinforcement learning, but this required giving the model certain rewards based on actions it performed in a given state. I don't believe this model would have worked well considering I don't know the correlated reward associated with each action. I considered actions to include the government ordering lockdowns, self-quarantines, social distancing, mask wearing, etc. These actions usually occur over a period of time rather than an instance, so it probably would have been better to create a separate model to predict how these actions would impact the number of COVID-19 cases.

I also considered using the perceptron algorithm to classify an area as a hotspot. I learned about this supervised learning algorithm in machine learning, and it initially seemed to fit my needs. Hugo was able to fix the issue in Guayaquil by identifying where COVID-19 was concentrated and sending medical officials to treat the areas. I already had the number of COVID-19 cases by date and ZIP Code, so I just needed to define what a hotspot was based on certain criteria. The CDC defines a county as a hotspot if it meets "all four of the following criteria, relative to the date assessed: 1) >100 new COVID-19 cases in the most recent 7 days, 2) an increase in the most recent 7-day COVID-19 incidence over the preceding 7-day incidence, 3) a decrease of <60% or an increase in the most recent 3-day COVID-19 incidence over the preceding 3-day incidence, and 4) the ratio of 7-day incidence/30-day incidence exceeds 0.31. In addition, hotspots must have met at least one of the following criteria: 1) >60% change in the most recent 3-day COVID-19 incidence, or 2) >60% change in the most recent 7-day incidence" [15]. I didn't see the need for this. If I wanted to check where it was concentrated, there was no need for a model; I can just check which ZCTA has the most cases for the latest day.

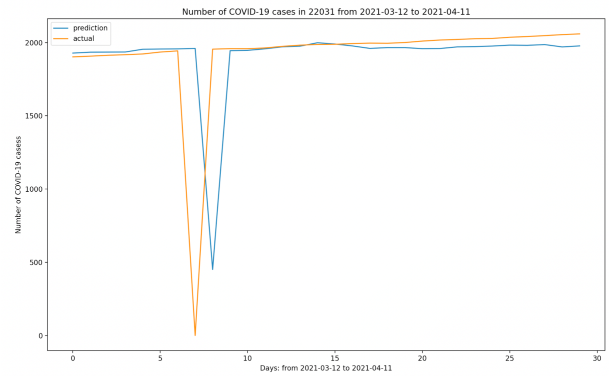

Figure 3: Multiple Linear Regression model based on the 25 nearest ZIP Codes

The last model that I considered before selecting the linear regression model was the multiple linear regression model. I thought that by having more variables, I would be able to more accurately predict the number of COVID-19 cases. The program would have worked similar to the linear regression model in that it asks users to enter a ZIP Code and date. Since I had the number of cases by ZIP Code, I decided to convert the ZIP Codes to latitude and longitude coordinates to find the closest points. CivicSpace Labs mapped ZIP Codes to states, latitude and longitude coordinates, and timezone using the ZIP Code gazetteers from the US Census Bureau from 1999 and 2000 [16]. The number of ZIP Codes provided by this dataset did not match the number of ZIP Codes provided by the USCB dataset; it has 1275 ZIP Codes listed in Virginia. I used this rather than the Google Maps API, because for the Google Maps API I needed to set up a key and then grant permission to use the API on the Google Cloud Platform. I wasn't sure how much the API charged and how many function calls I was going to perform, so I did not use the API. I then sorted and took the nearest 25 neighbors for my algorithm. I created a 2d-array to store the training data; I ran through the VDH dataset and ordered the rows by the date from the start date, the columns by the ZIP Code, and the value in the index as the number of COVID-19 cases at the specified date and ZIP Code. The Y dataset is the number of cases at the user specified ZIP Code the day after the corresponding row in the X dataset. If the X dataset goes from April 1 to 10, then the Y dataset is the data at the ZIP Code from April 2 to 11. I ran a multilinear regression function on the training dataset. I realized that my model had little practicality unless the user had a specific date in mind. This is because, with a linear model based on days, I can easily predict the number of cases by substituting a new value into the linear regression model. However, with the multiple linear regression model, it is based on the number of COVID-19 cases in neighboring areas, which I likely won't know. If it was a specific time, then I could model the Y value to be equal to the difference between the predicted date and last day in the VDH dataset. But I don't think this model would have been useful or effective.

C. Difficulties

It was hard to decide what data I wanted to use. The VDH dataset was the first one I found, and it had the number of COVID-19 cases by ZIP Code. It was hard to find other information such as population size, age distribution, wage gap distribution, etc. in different areas by ZIP Code in Virginia. These variables could have affected the number of COVID-19 cases, but most of the data did not come from the government or was only at the state-level. Data Commons shows important data separated by county, but it only showed a few nearby counties. I did not know how to extract all the county information into a single csv. They cite reliable sources such as fbi.gov (Federal Bureau of Investigation), bls.gov (United States Bureau of Labor Statistics), census.gov (United States Census Bureau), but the associated links did not link to the specific url which contained the information. As a result, I had a hard time finding the information. I didn't want to extract the information from each county individually, so I eventually decided to not use any additional variables. The New York University Langone Health's City Dashboard has a COVID-19 risk index for a couple of cities which contains the neighborhood risk level of COVID-19. I extracted their data as a csv file, but the issue I encountered is that they are missing information on some cities. They have information on Richmond, but no information on Fairfax.

I had some difficulties organizing the training data during the implementation. One was sorting the data by date. It's easy enough with a linear model, but I originally implemented a multiple linear regression based on the number of cases in the nearest 25 neighbors. I created a separate dictionary to map ZIP Codes to array indices. As for the actual dates, I was unsure whether the CDC public data actually contained the number of daily cases for each day after the initial recording. There are values for 327 days for each ZIP Code, but there are 331 days between the first day and the last day in the dataset. To fix this issue, I mapped the data starting from the number of days from the first record, and set the blank values as the number of cases from the day prior.

V. Results

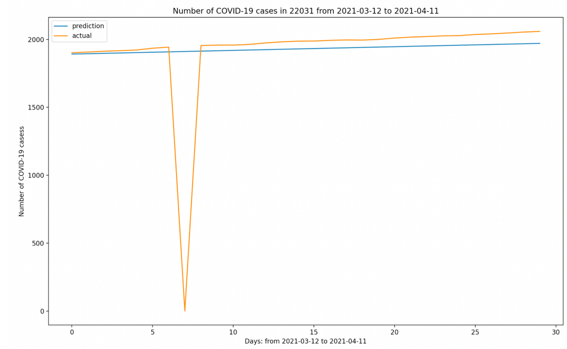

Figure 3. Model of the Prediction and Actual number of COVID-19 cases

The system works as intended and predicts the number of COVID-19 cases by ZIP Code with a reasonable accuracy. To test the accuracy of my model I used the number of COVID-19 cases for the 22031 ZCTA over the last 30 days from April 11, 2021 as a test set. For the training set, I looked at the week prior to March 12, 2021. I calculated the slope and used a linear model. According to the Berkley Yu group, they implemented many models including a linear one based on the most recent four days. They found that this model proved more accurate than the ones based on the most recent 5-7 days. When testing my model, I found that basing it on the number of cases over seven days more accurate than 4-6 days. There were 1902 cases of COVID-19 in 22031 on March 12, 2021. By April 12, 2021 there were 2059 confirmed cases, while my model predicted 1970 based on the past week. If I based the model on the most recent 6 days then I get 1962 cases, and if I based on the most recent 5 days then I get 1941 cases. I found a similar situation occurred when I set the ZIP Code to 22904. By April 11, 2021 there were 857 confirmed cases of the virus. Based on the most recent 5-6 days my model predicted there would be 838 COVID-19 cases by April 11, while my model based on the most recent week predicted 842.

Despite my tool satisfying the functional requirement of creating a model to predict the number of COVID-19 cases, it might still fail due to not satisfying certain non-functional requirements. My tool is a python program, so it is not intuitive to run to people who are not familiar with programming languages. There is also the problem with performance. I originally implemented a multiple linear regression model before switching my program over to a linear model. As time passes, the program will parse through more data. It is not currently optimized, so the program takes some time to complete. Finally, while it currently has a high degree of accuracy, the tool relies on the user to input data from the VDH public use dataset. This means that as time passes and the user fails to update the program's dataset, it will produce more inaccurate predictions. This makes running the program inconvenient as the user has to frequently manually update the dataset.

The intended user of my tool is policy makers; my program is intended to encourage policy makers to implement COVID-19 preventative measures such as social distancing in areas with a high number of predicted COVID-19 cases. Certain groups may be disproportionately affected as many neighborhoods are not diverse. The CDC found that people ages 18-29,30-39, 40-49, and 50-64 have twice as many cases as people ages 5-17 [17]. Thus the model should predict these areas to have a high number of COVID-19 cases compared to areas with lots of children and elders. Many low-income jobs cannot be performed while quarantined, so many individuals may risk contracting the virus for more money. Low-income neighborhoods will be devastated by COVID-19 preventative measures such as quarantines because people will be unable to earn money to support themselves. In addition, my program favors these regions which suppress their COVID-19 numbers by setting the number of COVID-19 cases for these areas to zero.

VI. Conclusion

I designed a system to predict the number of COVID-19 cases by neighborhoods. I conducted linear regression on the most recent COVID-19 cases for a ZIP Code. I tested the model using the week of March 5, 2021 to March 12, 2021 to predict the number of COVID-19 cases for the next 30 days. I then compared this to the actual data and found that by April 11, 2021 for the ZIP Code 22904, my prediction for the number of COVID-19 cases was 842, only 15 away from the actual number of COVID-19 cases.

VII. Future Implementation

If I was given more time, I would have implemented a web application using a three-tier architecture. I would upload my data to the UVA mysql database which would be based on the public use data set from the VDH. Users would interact with a react frontend on their mobile phone or computer by entering in the ZIP Code of the area they want to forecast. Information would then be sent over to the server. The server interacts with the database to pull out the number of COVID-19 cases for that area over the past week. If the selected area is not one of the designated ZCTAs, then I send the information back to the client and display an error message on their device. Otherwise, it then creates a linear model based on the number of COVID-19 cases over the past week and sends the model to the client's device. The client can then view the predicted number of COVID-19 cases over the next 30 days.

References

[1] Cucinotta, D., & Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta Bio Medica: Atenei Parmensis*, *91*(1), 157. https://doi.org/10.23750/abm.v91i1.9397

[2] JHUSOM (2020, October 08). Johns Hopkins University School of Medicine. *COVID-19 Map.* October 08, 2020, https://coronavirus.jhu.edu/map.html

[3] Dube, R., & Córdoba, J. (2020, June 30). Ecuador City Beat One of World's Worst Outbreaks of Covid-19. WSJ. https://www.wsj.com/articles/ecuador-city-beat-one-of-worlds-worst- outbreaks-of-covid-19-11593532974

[4] CDC (2020, November 1). Centers for Disease Control and Prevention. COVID-19 Forecasts: Cases. November 01, 2020, https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/forecasts-cases.html

[5] CHD (2020, November 1). City Health Dashboard. City Health Dashboard. November 01, 2020, https://www.cityhealthdashboard.com

[6] CDC. (2020, February 11). *Contact Tracing for COVID-19*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/contact-tracing.html

[7] Ankel, S. (2020, October 21). *"Masks work? NO": Twitter removes tweet by White House coronavirus adviser that says face coverings are not effective against COVID-19*. Business Insider. https://www.businessinsider.com/tweet-removed-dr-scott-atlas-masks-dont-prevent-covid-19-2020-10?international=true&r=US&IR=T

[8] McCann, M. (2020, June 1). *Mandatory Masks Aren't About Safety, They're About Social Control*. The Federalist. https://thefederalist.com/2020/05/27/mandatory-masks-arent-about-safety-theyre-about-social-control/

[9] USCB (2018, June 25). The United States Census Bureau. *Virginia*. https://www.census.gov/geographies/reference-files/2010/geo/state-local-geo-guides-2010/virginia.html

[10] USCB (2013, June 5). United States Census Bureau. *Census Bureau Geography and Maps Frequently Asked Questions*. http://www.census.gov/geo/www/tiger/tigermap.html

[11] velmalopez. (2021, April 13). *COVID-19-Forecasts*. GitHub. https://github.com/cdcepi/COVID-19-Forecasts/blob/master/COVID-19_Forecast_Model_Descriptions.md#Yu_Group

[12] Ducharme, J. (2020, July 22). *These Maps Show How Drastically COVID-19 Risk Varies By Neighborhood*. Time. https://time.com/5870041/covid-19-neighborhood-risk/

[13] Altieri, N., Barter, R. L., Duncan, J., Dwivedi, R., Kumbier, K., Li, X., … Yu, B. (2021). Curating a COVID-19 Data Repository and Forecasting County-Level Death Counts in the United States. Harvard Data Science Review. https://doi.org/10.1162/99608f92.1d4e0dae

[14] VDH (2021, April 12). Virginia Department of Health. *COVID-19 in Virginia*. https://www.vdh.virginia.gov/coronavirus/

[15] Oster, A. M., Kang, G. J., Cha, A. E., Beresovsky, V., Rose, C. E., Rainisch, G., ... & Villanueva, J. (2020). Trends in number and distribution of COVID-19 hotspot counties—United States, March 8–July 15, 2020. *Morbidity and Mortality Weekly Report*, *69*(33), 1127.

[16] CivicSpace Lab. (2018, February 9). *US ZIP Code Latitude and Longitude*. OpenDataSoft. https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/information/?location=4,38.09998,-93.12012&basemap=jawg.streets

[17] CDC (2020a, February 11). *Cases, Data, and Surveillance*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html