COVID Data Science Project

Adi Gupta, Mary Kovic, Bailey Man, Ansh Motiani, Pietro Spini

Sep 27th 2020

Abstract

In this first brief we detail out a few aspects of our project. The first section introduces the problem from the perspective of a decision maker who wants to visits a point of interest (POI) but may be worried of becoming exposed to to COVID. We create a tool which, given a POI, offers the decision maker an estimate of the risk of exposure for the desired POI and suggests destinations that are similar but possibly with a lower risk of exposure. Section 2 introduces the methodology, Section 3 describes the data sources and their limitation. Section 4 explains the choices we made for the usability of the tool from a User Interface perspective Finally Section 5 briefly concludes and outlines directions for further development.

1 Motivation

Since the onset of the COVID-19 pandemic, people's everyday decisions, such where to shop for groceries, meet a friend for coffee or visit a salon, have been disrupted. Pre-COVID, a decision maker may have taken into consideration features like the distance to the desired point of interest, the price of the goods and services exchanged and the availability for parking. For many decision makers, the risk of the contracting COVID-19 adds a further dimension which needs to be incorporated in the decision of which places to visit and when. The risk faced by each decision maker is clearly heterogeneous: it depends on the safety protocol adopted by the vendor, on the number of people that may be visiting the same POI at the same time and on the idiosyncratic health condition of the decision maker. This latest dimension is the epidemiological risk of infection which may depend on health parameters like age, pre-existing conditions, co-morbidities and access to healthcare. In this project we abstract from the epidemiological risk and focus instead on the risk of exposure. In particular we focus on the risk associated to other visitors of the point of interest. A collection of online tools for risk assessment has been developed by various groups. A notable one is the COVID-19 Event Risk Assessment Planning Tool by researchers at Georgia tech. The model computes allows users to compute the risk of exposure in attending an event in a specified county assuming that the event is attended by a user specified number n of people. With respect to the COVID-19 Event Risk Assessment Planning Tool our tool main innovation consists in three aspects. First, we focus on risk assessment for visits of point of interest rather than events: these include shops, salons and sport facilities. Incidentally, this is a higher degree of granularity than the county level risk because it focuses on the specific risk at the point of interest that the decision makers intends to visit. Second, we do not require a user specified input of the number of people at the event. Instead we rely on traffic data about the point of interest to calibrate the "number of people present" parameter of the model. We see this as a substantial innovation since the exact number of attendees to an event is likely to be known only for small groups. Conversely, our estimate is informed by observed traffic flows. Third we emphasize the fact that infection rates display a high degree of geographic heterogeneity. From a modelling perspective, visitors who are residents of different geographic areas carry an area specific probability of having an active COVID case. We are able to capture this feature in the model by including aggregate data on visitors place of origin and merging such data with the area specific infection rates, as detailed out in Section 2.

2 Modelling Exposure

We focus on the event of "Exposure" or E for brevity, which we characterize as the probability that 1 or more people that visit your point of interest have active COVID-19 cases. Importantly, we do not address the epidemiological risk of infection and we limit the model to exposure as described above. We note, on the other hand, that the model could be augmented to consider the user specific epidemiological risk but it requires substantial epidemiological knowledge and may be dependent on the visitor specific demographic and health history information.

In this simplified model we assume that the user visit the point of interest and interacts with other visitors at the same point of interest. Each visitor possibly comes from a different geographical location. We stress the importance of a model that, while being local to the point of interest, can also capture mobility of visitors from neighboring districts. In particular each district $d \in D$ has a d-specific infection rate. The probability of the event E is computed as follows:

$$P(E|p,t) = 1 - \Pi_{d \in D} \left(1 - \frac{I_{dt}}{N_{dt}}\right)^{n_{pdt}}$$

where I_{dt} is the number of people with active COVID cases at time t in district. N_{dt} is the total population in district d. Finally n_{pdt} are the number of district d residents that are visiting the point of interest at time t. One way to look at the model is to think about the $P(E|p,d,t)$ as $1 - P(E^c|p,d,t)$ where $P(E^c|p,d,t)$ captures the probability that none of the other visitors we come in contact have active COVID-19 cases. Here we take the number of visitors coming from each district as given so n_{pdt} is known.

Consider the following example, avoiding the t subscript for simplicity. For a given point of interest p, if there are only two districts A and B and $n_{pA} = 3$ and $n_{pB} = 5$ visitors respectively and the infection rates in the two districts are $\frac{I_A}{N_A} = 0.08$ and $\frac{I_B}{N_B} = 0.1$ respectively the risk is given by:

$$P(E|p) = 1 - (1 - 0.08)^3 \cdot (1 - 0.10)^5 = 1 - 0.7787 \cdot 0.59049 = 0.5598$$

The first factor is the probability that none of the $n_{pA} = 3$ visitors from district A have COVID-19. It assumes that visitors to the point of interest are sampled independently from the same population that has infection rate $\frac{I_A}{N_A} = 0.08$. The second factor is the probability that none of the $n_{pB} = 5$ visitors from district B, which has an infection rate of $\frac{I_B}{N_B} = 0.10$, has COVID-19. Hence the probability of being exposed to COVID-19 is given by 0.5598.

3 Data sources: Benefits and Limitations

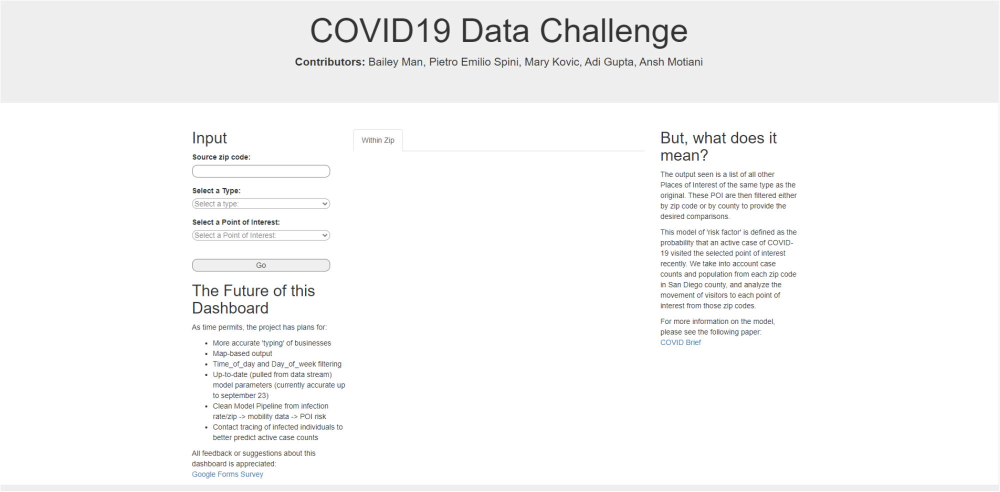
We rely on 4 primary data sources. First, Dataset 1 includes Zipcode specific COVID-19 Case counts from the Sandag open data portal. The epidemiological literature on COVID-19 has emphasized that Case Counts is likely to be a biased estimate of the total active cases in a region because the symptomatic rate is estimated to be around 79%. As a result, for every detected case it is likely that about 4 other cases are active yet undetected. For this reason we choose to mutliply the case count availble. Dataset 2 is the total population count per Zipcode available on the county's website. With these two datasets and the modelling assumption for asymptomaticity explained above we are able to compute the rate of infectious cases for any given zip code, given by $\frac{I_d}{N_d}$ in section 2. Dataset 3 is the SafeGraph Weekly Patterns data which we filter for the San Diego county area (FIPS code 06073-). The data contains, for each point of interest, the count of visitors whose home residence is located in a particular census block group, which we aggregate by census track. Hence, for each point of interest, we are able to obtain the sampled distribution of visitors across their location of origin. This data is very helpful to capture the heterogeneity across visitors that we detailed out in Section 1. In particular, this moment of the data estimates the n_{pd} described in Section 2. Finally we want to obtain location specific infection rates for each census track. The challenge is that census tracks are not univocally assigned to a zip code and hence we cannot simply assign the zip code level infection rates we compute with Dataset 1 and Dataset 2. Instead, we take advantage of the HUD USPS Zip code Crosswalk files which include, for each census tract, the percentage of resident population of the tract that lives in a given zip code. For example: if census tract 6073000100 splits over two zip codes, 92103 and 92110 with resident rates of 0.820296 and 0.179704 respectively. If, for example the infection rates in 92103 and 92110 are 0.12 and 0.05 respectively then we take the infection rate to be a mixture of the infection rates of 92103 and 92110 with the same weight as the fraction of the resident population. Namely:

$$I_{CT} = 0.82 \cdot 0.12 + 0.2 \cdot 0.05 = 0.106$$

This construction reflects the fact that a visitor from that census block group partially interacts with two regions with (possibly different) infection rates. We handle the few instances of tracts that split in three or more zip codes similarly. We do eliminate zip codes that have 0 residents (PO Boxes) and zip codes that contribute to any tract by less than 0.05.

4 The User Interface Experience

The User interface allows to input a point of interest. Each figure is described below:



COVID19 Data Challenge

Contributors: Bailey Man, Pietro Emilio Spini, Mary Kovic, Adi Gupta, Ansh Motiani

Input

Source zip code:

Select a Type: Select a type:

Select a Point of Interest: Select a Point of Interest:

The Future of this Dashboard

As time permits, the project has plans for:

- More accurate 'typing' of businesses
- Map-based output
- Time_of_day and Day_of_week filtering
- Up-to-date (pulled from data stream) model parameters (currently accurate up to september 23)
- Clean Model Pipeline from infection rate/zip -> mobility data -> POI risk
- Contact tracing of infected individuals to better predict active case counts

All feedback or suggestions about this dashboard is appreciated:
Google Forms Survey

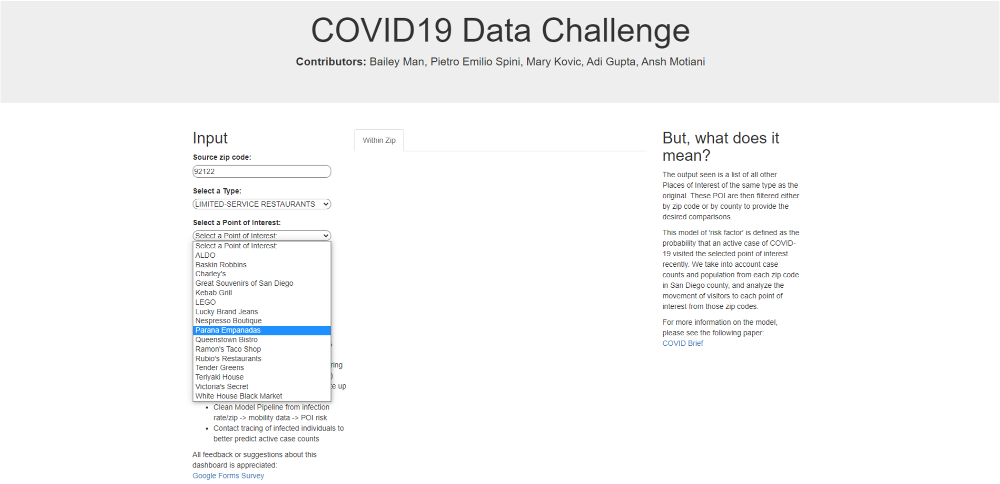
But, what does it mean?

The output seen is a list of all other Places of Interest of the same type as the original. These POI are then filtered either by zip code or by county to provide the desired comparisons.

This model of 'risk factor' is defined as the probability that an active case of COVID-19 visited the selected point of interest recently. We take into account case counts and population from each zip code in San Diego county, and analyze the movement of visitors to each point of interest from those zip codes.

For more information on the model, please see the following paper:
COVID Brief

Figure 1: Main Menu with input lines and explanation of the quantities computed by the tool



COVID19 Data Challenge

Contributors: Bailey Man, Pietro Emilio Spini, Mary Kovic, Adi Gupta, Ansh Motiani

Input

Source zip code: 92122

Select a Type: LIMITED-SERVICE RESTAURANTS

Select a Point of Interest: Select a Point of Interest:

- Clean Model Pipeline from infection rate/zip -> mobility data -> POI risk
- Contact tracing of infected individuals to better predict active case counts

All feedback or suggestions about this dashboard is appreciated:
Google Forms Survey

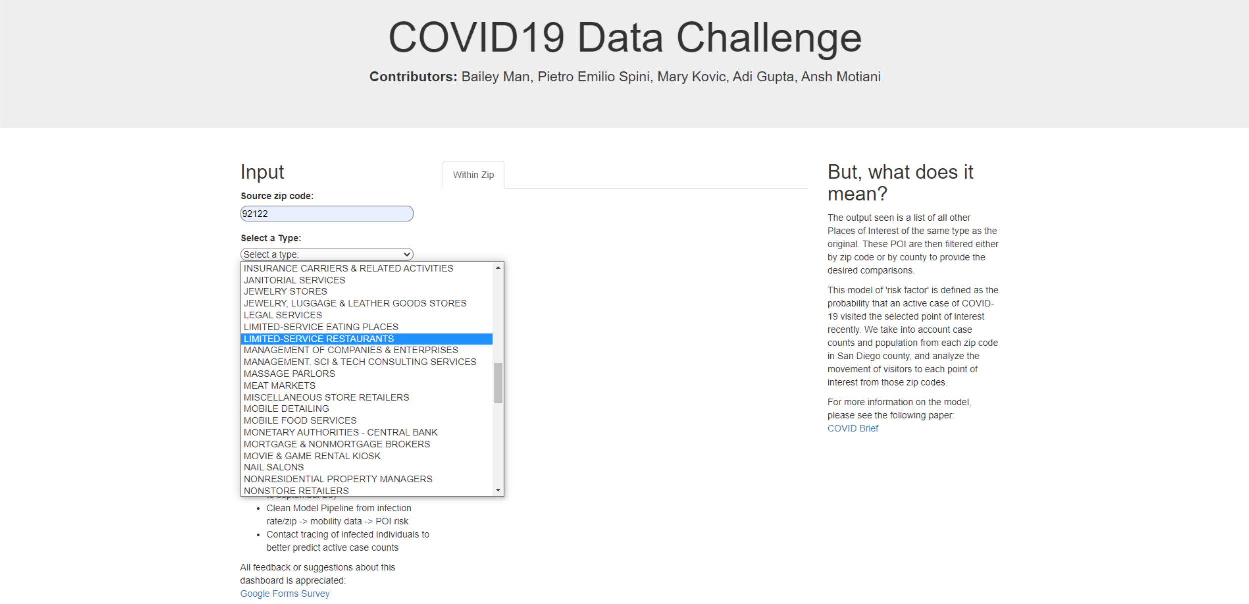
But, what does it mean?

The output seen is a list of all other Places of Interest of the same type as the original. These POI are then filtered either by zip code or by county to provide the desired comparisons.

This model of 'risk factor' is defined as the probability that an active case of COVID-19 visited the selected point of interest recently. We take into account case counts and population from each zip code in San Diego county, and analyze the movement of visitors to each point of interest from those zip codes.

For more information on the model, please see the following paper:
COVID Brief

Figure 2: Search by POI: user inputs a selected destination they wish to visit



COVID19 Data Challenge

Contributors: Bailey Man, Pietro Emilio Spini, Mary Kovic, Adi Gupta, Ansh Motiani

Input

Source zip code: 92122

Select a Type: Select a type:

Within Zip
INSURANCE CARRIERS & RELATED ACTIVITIES
JANITORIAL SERVICES
JEWELRY STORES
JEWELRY, LUGGAGE & LEATHER GOODS STORES
LEGAL SERVICES
LIMITED-SERVICE EATING PLACES
LIMITED-SERVICE RESTAURANTS
MANAGEMENT OF COMPANIES & ENTERPRISES
MANAGEMENT, SCI & TECH CONSULTING SERVICES
MASSAGE PARLORS
MEAT MARKETS
MISCELLANEOUS STORE RETAILERS
MOBILE DETAILING
MOBILE FOOD SERVICES
MONETARY AUTHORITIES - CENTRAL BANK
MORTGAGE & NONMORTGAGE BROKERS
MOVIE & GAME RENTAL KIOSK
NAIL SALONS
NONRESIDENTIAL PROPERTY MANAGERS
NONSTORE RETAILERS

- Clean Model Pipeline from infection rate/zip -> mobility data -> POI risk
- Contact tracing of infected individuals to better predict active case counts

All feedback or suggestions about this dashboard is appreciated:
Google Forms Survey

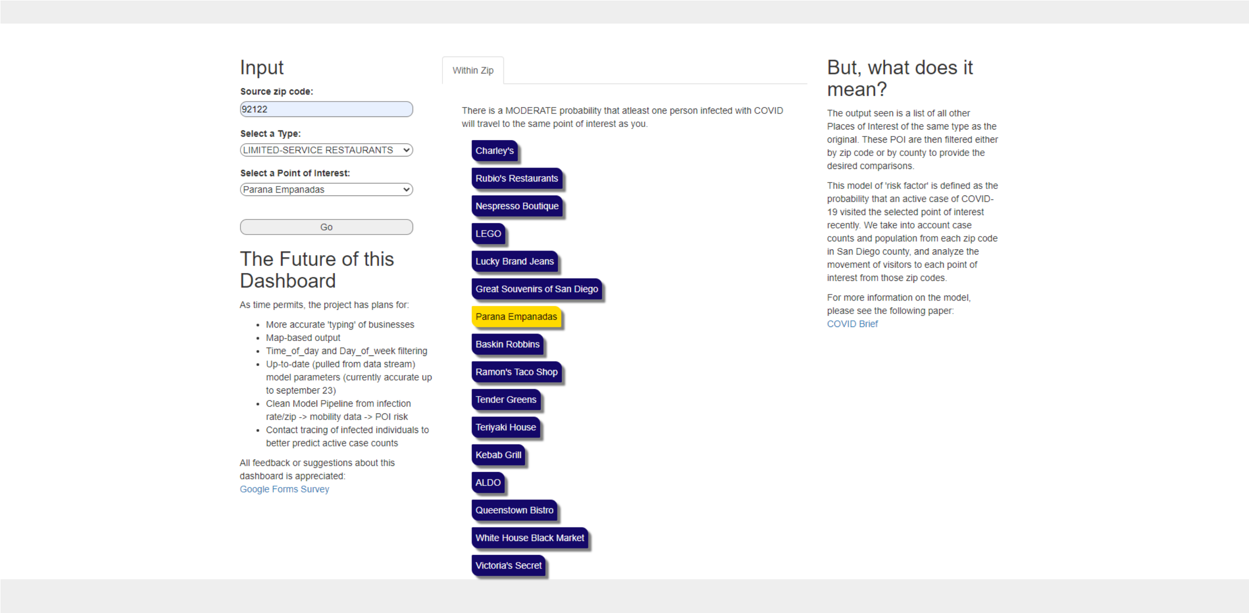
But, what does it mean?

The output seen is a list of all other Places of Interest of the same type as the original. These POI are then filtered either by zip code or by county to provide the desired comparisons.

This model of 'risk factor' is defined as the probability that an active case of COVID-19 visited the selected point of interest recently. We take into account case counts and population from each zip code in San Diego county, and analyze the movement of visitors to each point of interest from those zip codes.

For more information on the model, please see the following paper:
COVID Brief

Figure 3: User selects a type of POI: in the example Limited Service restaurants



Input

Source zip code: 92122

Select a Type: LIMITED-SERVICE RESTAURANTS

Select a Point of Interest: Parana Empanadas

The Future of this Dashboard

As time permits, the project has plans for:

- More accurate 'typing' of businesses
- Map-based output
- Time_of_day and Day_of_week filtering
- Up-to-date (pulled from data stream) model parameters (currently accurate up to september 23)
- Clean Model Pipeline from infection rate/zip -> mobility data -> POI risk
- Contact tracing of infected individuals to better predict active case counts

All feedback or suggestions about this dashboard is appreciated:
Google Forms Survey

There is a MODERATE probability that atleast one person infected with COVID will travel to the same point of interest as you.

Charley's
Rubio's Restaurants
Nespresso Boutique
LEGO
Lucky Brand Jeans
Great Souvenirs of San Diego
Parana Empanadas
Baskin Robbins
Ramon's Taco Shop
Tender Greens
Teriyaki House
Kebab Grill
ALDO
Queenstown Bistro
White House Black Market
Victoria's Secret

But, what does it mean?

The output seen is a list of all other Places of Interest of the same type as the original. These POI are then filtered either by zip code or by county to provide the desired comparisons.

This model of 'risk factor' is defined as the probability that an active case of COVID-19 visited the selected point of interest recently. We take into account case counts and population from each zip code in San Diego county, and analyze the movement of visitors to each point of interest from those zip codes.

For more information on the model, please see the following paper:
COVID Brief

Figure 4: Results are displayed with ranking of probability level: in the example moderate

5 Further Extensions of the Risk Assessment Tool

We see multiple major lines of expansion for the project at hand. First, we would like to account for the duration of a visit in a place of interest. We have information about the median dwell of visitors that can be helpful to better calibrate the model. We would like to refine the choice of asymptomatic rates (in our simple model we considered an 4 to one asymptomatic rate as described by Section 3). Ideally we could get area specific asymptomatic adjustments to capture the geographical variation in asymptomatic rates, which is likely related to the demographics and health variables of the population that resides in that area (these include insurance status, testing availability and the aforementioned demographic variables). We would like to incorporate a time dimension for the visitors counts. This would ultimately result in a further detailed input to the decision maker who can now query a specific time of the day when she wishes to plan her visit. On the side of the user interface we think the tool could benefit from a clickable map that automatically inputs the zip code of interest. The tool could also benefit from a more transparent description of the busyness typing which can improve the user experience by increasing the usability of the tool. Finally, contact tracing data in place of the aggregate regional data could dramatically improve the accuracy of the tool replacing case counts with an appropriately anonymous active case count. This last point raises issues of privacy and data ethics that would need to be explored before releasing the tool.
Any feedback on these dimensions is gratefully appreciated.