Tracking Covid: Predicting Infection Through Mobility Data and Ensuing Privacy Concerns

Sydney Mason

2024-10-25

Imagine a world where your phone was used to determine your likelihood of infection by a virus. No- I'm not talking about how infrequently you clean it and the various pathogens likely clinging to its surface. I'm discussing how mobility data obtained from location history can be used to determine how likely you are to become infected with a pandemic or endemic virus.

In "Public Mobility Data Enables Covid-19 Forecasting and Management at Local and Global Scales", an article published in June of 2021 in Nature, the authors sought to determine if risk of infection was reduced by policies that restricted movement by individuals on a more local level. They further wished to forecast the next ten days of infection rates based on infection and behavioral information influenced by those policies. In the paper, they created two models, a behavior model and an infection model which were based off of mobility data, infection data, and policy data. The behavior model sought to contextualize human behavior during the pandemic by including variables such as restrictions and other Covid-specific policies, whereas the infection model estimated risk of infection given information on transmissibility of Covid.

The models were developed to fill the knowledge gap regarding local rates of infection, as the authors felt that although there was a wealth of modeling of global and large scale impacts of Non-pharmaceutical inerventions or NPIs on infection rates, local information was limited. The lack of local information made it more difficult for policymakers and stakeholders to determine when to relax NPIs as large-scale information was not always indicative of regional rates of infection. The four categories of NPIs that the authors developed, covering thirteen overall policies, were Shelter in Place, School Closure, Social Distance, and Travel Ban.

The authors found that reducing the mobility of persons through policies and restrictions in turn lowered the infections in an area. Their models were able to accurately forecast the infections in administrative areas with reduced standard error compared to the baseline model that didn't factor in mobility.

The data on NPIs was obtained from the Global Policy Lab and the Organization for Economic Co-operation and Development. The models were developed off of location and mobility information obtained predominantly from Google and Facebook (Meta). Other sources included Baidu, a Chinese tech company, and SafeGraph, a company that has been embroiled in scandal for everything from their investor list, to their data collection methods, to various contracts and datasets they've published including one that listed locations of places that provided abortion access. Although the data from Google and Facebook requires location to be on (with it being default off for Google), certain features are improved by turning on location such as GPS for Google Maps. The data was anonymized, and in the case of Google, if data was too limited to anonymize, it was not posted on their Covid-19 Mobility dashboard. Infection data was obtained from the 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository compiled by the Johns Hopkins Center for Systems Science and Engineering.

The two models were trained on 20 days of data, where all 20 days in an administrative region had to meet the criteria of, infections at time t were greater than 10, infections from the day before were greater than 0, and at a global scale only, mobility at time t was greater than or equal to 5% above baseline, and the 2 week rolling average of growth rate of infections needed to be flattening. These criteria ensured that the training data was from when there were active and consistent infections, meaning that Covid was fully in the community.

The behavior model at its core measured the change in behavior over the change in NPIs. In more specific terms, it's a reduced form model. Mobility is modeled as a a function of policies at a time t and X which represents all of the control variables including geographic information, time, etc. as a linear functional form. Policies and NPIs are represented as a value between 0 and 1 in the model, where 1 is the most intense level of policy enforcement and 0 is the least. The larger model is a summation of policies and their effects on mobility, that adds in a fixed effect for time, region, and error. The model is similar to many other models in econometrics. In order to properly run the behavior model, you first need to use the smaller mobility model to estimate mobility in the time period and then use ordinary least squares to check the model. After getting mobility measurements for each of the policies in a region, sum the policies along with the fixed effects to compute the final value for the effect of mobility policies in a region.

The information gained from the behavior model showed that overall movement between administrative units (classified as provinces/counties, regions, or countries) was reduced by 73% on average overall from the combined impact of NPIs. Time at home was estimated to have increased by 28% overall through mobility data obtained from Google which noted reduced time at work by 59.8% and retail spaces by 78.8%. Further analysis of the individual policies showed an overall reduction in mobility but policies such as travel bans were more effective than less extreme policies such as school closures.

The infection model is at its simplest is an analysis of change in infections over change in behavior using behavior as the independent variable. The model is a function of mobility and the same control variables present in the behavior model set equal to a logarithmic function which is the first-difference of log confirmed infections at a time t. The function of mobility and the control variables is a linear functional form. Since the dependent variable is the growth rate, it can withstand underreporting and other issues as long as the reporting remains constant overall. In applying the model, each policy is assigned coefficients that estimate the overall effect of the policy on infection rates and is paired with mobility over lag periods that allow for the model to catch up to any policy changes and mobility changes that may occur.

To actually implement the model, the model was trained over the 20 day period, and coefficients matching each of the policies were estimated. The model was then used to create a 10 day forecast and the percent error between the model estimates and the actual infections that day were generated.

In order to test the model, the authors used a basic logarithmic model that excluded mobility information to provide a baseline for infection rates. The model solely went off of past infections as a predictor for future infections. In a week-long period in April of 2020, models that neglected mobility data estimated new infections of 30,716, whereas the model created in the paper only forecasted 12,650, far closer to the observed cases of 10,496. Further, the error in the baseline model was larger than in the model developed including mobility variables, indicating that the model that included mobility variables was more accurate. When they analyzed the median percentage error (MPE), they found that in the US at a state level, the error was 88.73% in a model with no mobility data compared to an error of 21.82% in the model that included mobility data.

However, they did note some potential issues with their model. First, although it could be easy to state that mobility is the largest driving factor in infections, the authors stated that there should not be a causal relationship determined between mobility and infections, rather that mobility was able to show mixing that could allow other factors to take precedence. Second, although the model was effective in the circumstances they tested it on, it would not be as successful if the dynamics of the infection were changing rapidly, in which case a more traditional infection model would be more successful. Finally, the model depended on large-scale mobility data collected from large providers such as Google, which is not used in all countries. Therefore, there were some information gaps in mobility that would need to be addressed by a more traditional model. Overall, they advised that their model should be complementary to others.

The normative concern that the paper presents is issues of data privacy and by extension informed consent. The mobility data that the study uses is taken from large corporations who data mine on their own servers or obtain the data from third parties. Although the location data was anonymized, it still raises concerns about how much we as individuals and as a society are willing to share, especially in an increasingly virtual world.

Location data is continuously collected by apps and software on your devices, and most people agree to let it be collected when they click accept on Terms and Conditions. However, most people don't read Terms and Conditions, or if they do, it is written in a style that is not easily understood unless you have legal training. Therefore, although many times people provide consent, they don't fully understand what they're providing consent for, and don't know how their data will progress beyond the specific organization they're using. Many people are aware of the issues with sharing data with these companies and don't share their information, but certain applications are improved by sharing the data, so it becomes incentivized, adding a degree of coercion to the data collection. Location history in Google is utilized in order to provide better GPS information among other things, so although the default is off for many people- Google will prompt you to turn it on in order to improve your user experience, while neglecting to go in depth about how your location history might be used. Information is available on how location history is used if you do further research, where Google reserves the right to present the anonymized information to third parties. Beyond Google, Facebook and SafeGraph also collect data from its users without providing informed consent. SafeGraph in particular has come under fire for selling possibly dangerous datasets, including ones that were tracking the general locations people were traveling from and traveling after visiting Planned Parenthoods.

Although the data may be useful in creating models such as this, it prompts questions about how much of our lives we are willing to share with corporations, and how much privacy we truly have.

Covid-19 is now endemic, and with other viruses experiencing surges in the past years, developing models to determine how policies influence mobility and therefore infections will become increasingly crucial. This paper shows a future for how we approach policy making in times of crisis and provides a framework for understanding how to mitigate disasters when they do arise. Despite this, it is important to recognize the serious ethical implications that arise when models are based off of data that might not be obtained through informed consent, as well as the troubling underlying issue of how far are we willing to stretch these privacy issues in the name of progress? Overall, however, the paper provides an interesting outlook on future methodology to examine outbreaks at a local level and makes knowledge more accessible to the everyday policymaker or statistician.

References

Ilin, Cornelia, et al. "Public Mobility Data Enables COVID-19 Forecasting and Management at Local and Global Scales." Nature News, Nature Publishing Group, 29 June 2021, www.nature.com/articles/s41598-021-92892-8#data-availability.

Ossola, Alexandra. "Meet the Company Helping Scientists Study Covid-19 with Your Location Data." Quartz, Quartz, 23 Nov. 2020, qz.com/1934587/who-is-safegraph-the-company-giving-your-location-data-to-covid-researchers.

"Understand the Data - Community Mobility Reports Help." Google, Google, support.google.com/covid19-mobility/answer/9825414?hl=en&ref_topic=9822927&sjid=533892856493538
NA. Accessed 24 Oct. 2024.