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

Sydney Mason

2024-12-11

Introduction

Your phone tracks far more than you think. Everything from the sites you visit to the TikToks you watch to the Instagram profile you're definitely not stalking are subject to being mined for data by the various companies operating the apps. However, despite that being the data that most people consider when they think about data privacy, it only scratches the surface.
Google, Facebook, and various other apps not only have access to your activity, but also your location data. Therefore, anywhere you (and your phone) goes, tech companies and corporations have access to that data.

The companies make efforts to anonymize the data, and reports that use the data are limited to areas and data that is expansive enough that it should protect the privacy of individual users, at least in the case of companies like Google. However, that data still winds up being used, often without explicit consent. The data is used for a myriad of purposes, from mining to research projects.

In the paper "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 from Covid was reduced by policies that restricted movement by individuals on a more local level. They adapted a base SIR, or susceptible, infected, removed, model of infection to estimate growth rates, to include data on mobility and policies that limited mobility at various administrative levels, from individual counties to worldwide. The model used techniques common in econometrics and was based in more advanced algorithms, and was based upon mobility data obtained from various tech companies. The algorithm proved more accurate than the base models, but ethical considerations regarding how the data was obtained remain a concern. This paper seeks to evaluate and critique the statistical methods behind the paper as well as the ethical concerns, particularly regarding consent, that the mobility paper raises.

Analysis of Methods

Novel Analysis

For the novel analysis, I chose to implement a version of their baseline model to demonstrate the differences in error obtained from using the developed formula that minimizes error in prediction infection vs. a more traditional model. The baseline function used was $\log(\frac{I_{it}}{I_{i,t-1}}) = \gamma_i + \delta_t + \phi_{it} + \epsilon_{it}$ where I_{it} was the infection rate at time t, used in a logarithm with the lag for data regression and the $\gamma_i + \delta_t + \phi_{it} + \epsilon_{it}$ denoted fixed level effects, day to day effects, and error. This model is similar to the one that they used as a comparison for the errors on their model, estimating the growth rate using a simple logarithmic function based on the day before. I decided to use this model in my novel analysis because I thought it worthwhile to explore the errors of a more traditional model in a pandemic situation such as Covid and confirm the need for a more accurate model, even if it has issues with it.

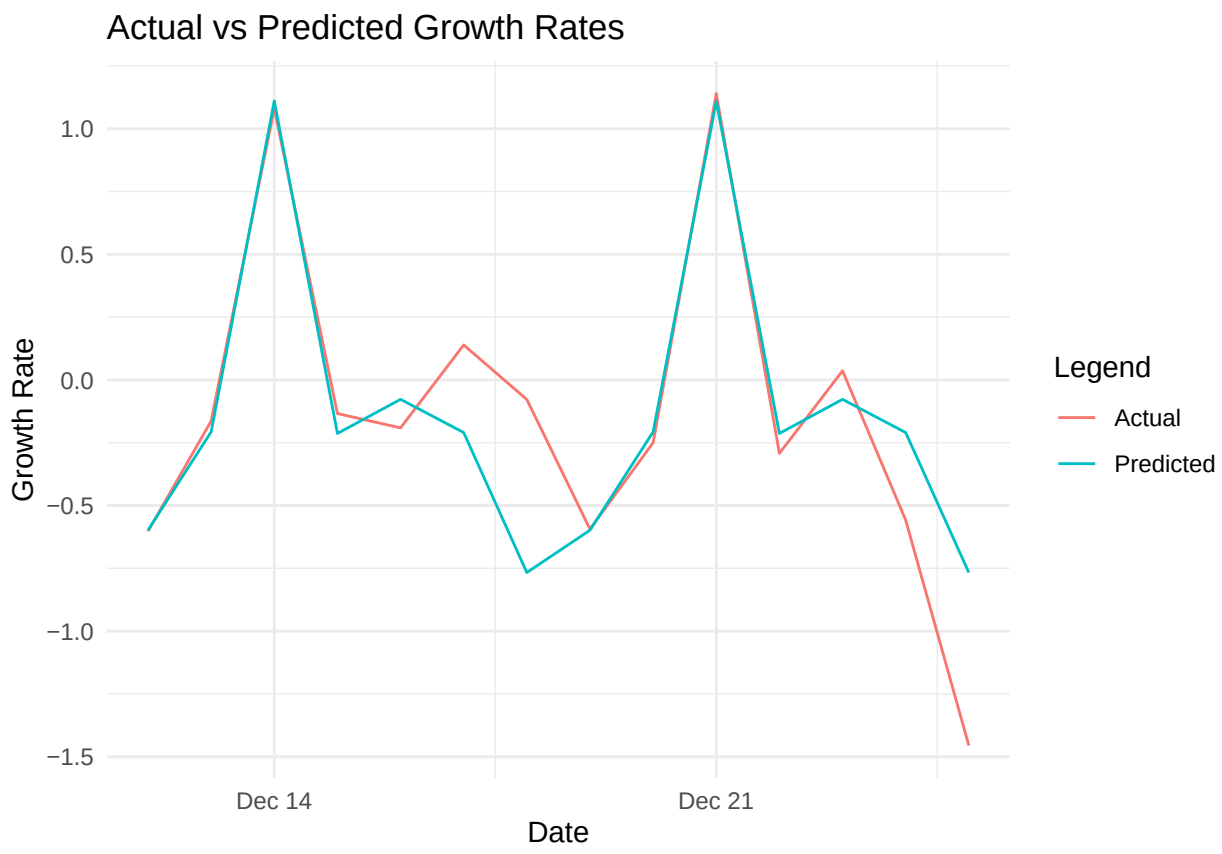
I used North Carolina Department of Health and Human Services, or NCDHHS data regarding Covid-19 cases from December 11- December 25th 2020, a two week period. The NCDHHS had information on positives obtained from antigen and PCR testing, the dates the data was collected, and cases that were reinfections, and deaths for the day in question. For the model, I only looked at the dates and information on positives obtained from PCR and antigen testing.

I first extrapolated the data from the file that was provided by the NCDHHS, isolating the two weeks in question and converting the date format so that I could manipulate and filter the data properly. I then calculated the growth rate as a logarithmic function of the total recorded infections at time t over the lag function used in linear regressions to obtain the actual growth rate. I then developed a linear model of the growth rate as a function of the day of the week.

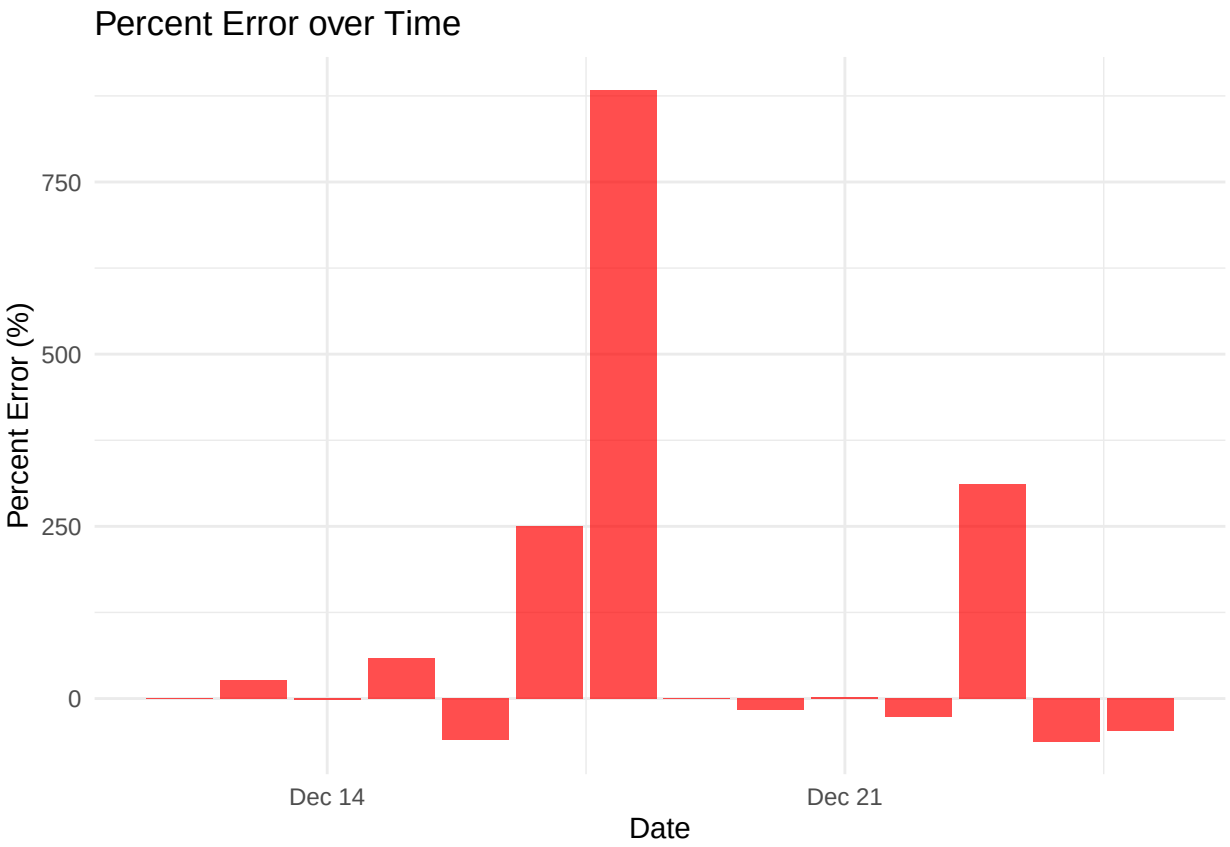
Then, using a similar method to the actual growth rate, I obtained a predicted growth rate by examining the data and using a logarithmic model to determine what an estimated growth rate would be based on past infections. I then calculated the error as a value and as a percentage to compare with the baseline information obtained in the paper to see how great the error was on average at the end of a two week period between the predicted growth rate and actual values.

Finally, I pulled summary data on the growth rates including the average percent error and developed two charts, one showing the estimated growth rate as a line graph in comparison to the actual growth rate and the other a bar chart showing the percent error of the growth rate day to day.

Graph 1


Actual vs Predicted Growth Rates
1.0
0.5
0.0
−0.5
−1.0
−1.5
Growth Rate
Dec 14
Dec 21
Date
Legend
Actual
Predicted

Graph 2



A couple issues that I had were with data limitations. The NCDHHS didn't have in the dataset I pulled information on the days of the week or the positive rate compared to overall tests, two factors that they used in their baseline model. To remedy that, I manually altered the Excel file to add in the days of the week, solely doing it for the two week period I was analyzing. The fact that I didn't have specific positivity rates did alter the model, so the model I developed should be seen as comparable, not as an exact replica of their baseline model.

Overall, the percent error in growth rate for that two week period was on average 94.16 percent. I expected that the error would be high for this range in particular because of the spikes that were observed during the holidays in peak Covid times. The error was comparable with what was seen in ADM1 or the state range for the US of 88.73 percent with no mobility data included over a seven day period. The higher error is likely due to the fact that I further expanded my analysis to cover a more time, primarily because I knew that the holiday time I chose would lead to some more drastic changes and including a wider range of data would be helpful to get a more accurate error level.

Critique of Methods

The model developed in the paper, which was further broken down into two models, behavioral and infection, used data on non-pharmaceutical interventions through mobility limitations and infection data to create a reduced form model that predicted infections through mobility. I struggled to replicate the model, however, I feel the underlying statistical principles were sound.

Despite that, in attempting to replicate the model and deciding how I was going to conduct my novel analysis, one issue I did run into was that the study did not provide a GitHub for their work. Although one of the main goals they cited on their paper was to make an accessible model that was easily interpretable and usable by local decision makers, the lack of a GitHub showing their work or other repository was counterproductive to their goals. That said, the authors did provide an appendix that walked the reader through their methods. However, since there was no baseline code to implement, the time investment it would take in order to get the model running for a local government could be seen as enough of a hindrance that it wouldn't be used, reducing the overall impact that a more accurate model could have.

When critiquing the statistical methods used in this paper, my primary critique is of accessibility. Although machine learning is slowly becoming more accessible with the advent of AI and an expanding coding proficient workforce, I myself, with if not great knowledge likely more than the general populace, was unable to implement anything more than the baseline, error-ridden, model. If the authors had created a GitHub repository or made other efforts to make their work accessible, the methodology would have been easier to replicate and check. However, without that information it is too difficult to do so, raising questions about the overall methodology.

Another critique I have is with the mobility data that was obtained. Beyond ethical concerns regarding it, there were issues in the data obtained. Depending on the country and administrative level being included in the analysis, there were vast differences in the amount of mobility data they had. The amount of data they had ranged from just over a month in countries like France to nearly two months in China. The differences in data meant that the percent errors calculated and amount of data the model was able to train on in different regions was wildly different. This means that the model that predicted data for China using their mobility data would likely be far more accurate than the model for France or the US, both of which had far less data. A further issue is that out of the countries listed, two of the five only had mobility data, and infection data was not included. This meant that when the behavioral models were created, those models were not able to be applied to two of the countries in question to develop the infection model, further reducing the amount of information that the study was going off of.

Overall, although the statistics appear sound, the lack of easy replicability and the discrepancies in the data that the authors were able to obtain to train their model are causes for concern in the overall paper.

Analysis of Normative Concerns

In this paper, the primary normative consideration raised is that of informed consent as it relates to data privacy. The location data that the study used was obtained from various sources, including Facebook, Google, SafeGraph, and Baidu. The user data was extracted through enabling location services in the various apps. However, the extent that the data was used was not always explicitly stated to the user.

In ideal circumstances, consent should be both informed and non-coercive. Informed means that the user in question is aware of all of the potential consequences and benefits of sharing their data. Non-coercive means that there is no way in which a person should feel forced to share their data either by circumstance or other influences. When it comes to informed consent, many companies technically do inform you of the data you may be sharing. However, the degree of privacy you give up is often hidden by legal jargon or the sheer volume of information you have to sift through.

In the case of Google, when looking at their location services policy which is a subset of their larger privacy policy that goes into data sharing, the privacy policy itself is over 5500 words. Considering the average speed for reading is approximately 240 words per minute, it would take the average adult at least 23 minutes to get through the total privacy policy, a time commitment that most would be unlikely to give. Therefore, when Google asks if someone agrees to their privacy policy, in all likelihood, they wouldn't know fully what they are sharing, making the consent uninformed.

In terms of being coercive, when it comes to location services, Google's default is to have it off. The catch is that when you go to use GPS through Google Maps, Google alerts you saying that your directions will be more accurate if you give location sharing permissions. When someone is in a position of needing to use a navigation app, odds are they are in a more stressful scenario and are more likely to give their consent than they normally would because they feel pressured to give it in order to get to their destination. Although it isn't overtly coercive, Google does take advantage of the circumstances in order to obtain more data from their users in a manner that isn't ethically sound.

Facebook operates similarly in their data mining policies, where location sharing is optional, but encouraged. There is limited information available on Baidu's policies and SafeGraph obtains data through third-parties such as gaming apps on phones. The same concerns are applicable to all of them however, as consent is not explicit in any of their terms and conditions.

In summary, although the overall usage of the data is positive as it is aiding a public health crisis, it brings into question how knowledgeable people are about the degree of data they are sharing with tech companies and if they are in fact consenting to that data sharing. It is impossible to fully say whether the usage of data in this manner should be allowed, or at what point we should allow it. Kant under deontology would require us to make a blanket determination of when privacy can be violated and informed consent made tacit, allow it or not. I think it is a much more nuanced argument however, because despite the fact that the data is being obtained in manners that aren't ethically sound, the data is being used to help people in the long run. Whether or not it should be allowed is up to the individual to

decide, but despite personal feelings on the matter, everyone should be aware of how easily their data can be accessed and used.

Conclusion

Although the algorithm developed is a useful tool for local decision-making in pandemic scenarios, the methodology with data collection brings up ethical concerns. From a paternalism standpoint, meaning the decision to do something that may violate some ethical boundaries is made with the end goal of helping a moral agent, the decision to use user data on location makes sense. In a crisis scenario, where traditional models are inaccurate, it is easy to reach for models like this one where the promise of more accurate results can cause one to ignore issues with development, from issues with replicability to concerns with the data it is based upon. This model is a useful tool, but understanding the critiques of it are vital to implementing it responsibly, as it is with any model. In the future, I would be curious to see if the code for their model becomes open-source and if becomes a tool with endemic viruses such as Covid and the flu, or, as much as we would like to pretend otherwise, a likely future pandemic.

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.

Curcic, Dimitrije. "Reading Speed Statistics." WordsRated, 24 May 2023, wordsrated.com/reading-speed-statistics/.

"Privacy Policy – Privacy & Terms." Google, Google, policies.google.com/privacy?hl=en-US. Accessed 11 Dec. 2024.

"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.