Understanding the Relationship Between Mobility Data and COVID-19

January 13, 2021

Mentors: Nancy Reid, Samuel Perreault
Students: Allen Bao, Nicholas Martin, Saarthak Sangamnerkar, Stephen Brock

Introduction

The coronavirus disease 2019 (COVID-19) has caused a worldwide pandemic. At the heart of its spread is the contact people have with each other. Mobility could be a proxy to measure the number of interactions people will have, and this data is readily accessible from Google's mobility datasets. The main focus of this paper is to answer the question: how is mobility linked to the number of new COVID-19 cases each day? This question is addressed by investigating the relationship between Google's mobility data and the number of daily new cases of COVID-19.

An understanding of the link between mobility and the increase in COVID-19 cases will enable policy makers to leverage this information when deciding on important plans of action (e.g. closing schools may have a greater impact than closing pharmacies).

Previous Work

Given that we are only 10 months into the COVID-19 outbreak, very few studies have been published about the role of mobility and its effect on the pandemic. Bajardi et. al discussed the role of travel-related controls during the initial stage of the H1N1 outbreak in 2009 [1]. Although the travel restrictions led to a decline of about 40% in international air traffic to and from Mexico, these restrictions were unable to stop the virus from spreading at the rate of a pandemic [1].

Gatalo et. al examined mobile phone mobility data to study its correlation with the variation in COVID-19 cases [2]. However, the data was deemed ineffective to capture the behavioural components associated with social distancing and its effect on the spread of the virus [2]. It was observed that restricting the gatherings of individuals might also contribute to the reduction of new positive cases since the overdispersion of cases indicated that a small group can account for a large proportion of transmission [2]. Also, seasonal changes were found to be strongly correlated with the reduction in the infection rate [2].

Badr et. al. showed the strong correlation between mobility patterns and decreasing COVID-19 transmission rates for the most affected counties in the United States [3]. They observed that the Pearson correlation coefficients were above 0.7 for 20 of the 25 counties that were evaluated [3]. Additionally, they also demonstrated that it required around 9-12 days and potentially up to three

weeks for the effect of changes in mobility patterns to be perceptible, a range consistent with the virus' known incubation time [3].

Moritz U. G. Kraemer et. al. used real-time mobility data from Wuhan and detailed case data including travel history to examine the effect of imported cases in the spread of the coronavirus across cities in China [4]. They found that the initial spatial distribution of the cases in China could be explained well by the mobility data, but this relationship became weaker after control measures were implemented [4]. Strict mobility restrictions in China did appear to have a substantial effect on diminishing the spread of COVID-19 [4].

Datasets/Variables

The Google Community Mobility Reports are a large collection of datasets obtained from mobile device users with "Location History" enabled on their Google accounts [5]. The reports are aggregated and anonymized from collections of users in the same geographic regions, and are collected from all around the world [5]. The reports collected and used for this model were each from Ontario, split into major Ontario counties and regions. These reports collect location data over each day from the start of the pandemic for various categories of locations [5].

Each entry in the mobility dataset is shown by a percent change value, representing how the number of visits and length of stays at each location categories differed compared to a baseline [5]. This baseline is the median value of the 5-week period from January 3rd to February 6th 2020, to represent the time just before COVID-19 outbreaks began in most of the world [5]. Each baseline also corresponds to the specific day of the week to account for differences in mobility behaviour between days [5].

The mobility data is segmented into 6 categories: 1) grocery and pharmacy, 2) parks, 3) transit stations, 4) retail and recreation, 5) residential, and 6) workplaces [5]. Grocery and pharmacy mobility represents the mobility trends near food markets, food warehouses, farmer's markets, food shops, drug stores, and pharmacies. Parks mobility represents the mobility trend for parks, beaches, plazas, and public gardens. Transit stations mobility represents mobility trends for public transportation hubs such as subway, bus, and train stations. Retail and recreation mobility represents the mobility trends for restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theatres. Residential mobility represents the trends for residences. Workplaces mobility represents the mobility trend of places of work.

The second dataset that was used is "Confirmed positive cases of COVID19 in Ontario", found on the data catalogue on the official Ontario webpage [6]. Each row of the dataset represents an anonymized confirmed positive case. It contains various information about the case, such as the patient's gender, age group, and the public health unit that reported the case [6]. It also includes several dates, including the date of the case's reporting, the estimated date of symptom onset, the

test reporting date, and the specimen date [6]. The estimated date of symptom onset is of most interest to us, as we believe it to be the closest to the day on which viral spread can be linked to mobility. In this paper, case counts are assumed to be attributed to this estimated date of symptom onset rather than the date each case was reported positive.

These datasets were merged into a centralized dataset, segmented by region and by date. Additionally, a second dataset was generated with the same data, but with each of the mobility percentage changes replaced with simple moving averages and case count values replaced with rolling sums of the seven days leading up to and including the date of interest. The 7-day aggregations of counts and mobility allow us to accommodate for differences in the effects of specific days of the week on mobility or case counts (ex. people visiting parks more often on weekends naturally rather than because of the state of the pandemic). These differences can be seen in Appendix A. This can also help dilute the effects of unprecedented outliers. This centralized dataset uses the estimated date of symptom onset as its standard date metric to align new case reports with mobility data. The data spans the time from February 15th, 2020 to December 22nd, 2020.

Ten regions in Ontario were examined from the datasets: Durham, Halton, Hamilton, Middlesex, Niagara, Ottawa, Peel, Waterloo, Toronto, and York. Many other regions were taken out of our analyses due to missing data for an excessive number of days. A list of these regions can be found in Appendix C.

Exploratory Linear Model

Using a linear model to measure the relationship between the mobility data and the number of new cases is a great starting point because the model is simple, reveals any linear relationship, and provides a measure of significance (p-value) of the relationship. Our initial approach was to use a linear model to perform inference over new case counts and consequently, explain the relationship between the mobility data and new cases.

For the model variables, we decided to use population information - size, density, and area - based on Ontario region in addition to the mobility information. Since the mobility data is a relative change for each region, in order to create a model that generalizes to all regions, we need to scale the mobility factors by region population size. To do this, we introduced interaction terms between the population size and the mobilities. The model summary results are seen in Figure 4 of the Appendix B.

The linear model (Figure 4, Appendix B) has an R^2 value of 0.915 and is thus able to explain 91.5% of the variation in the response data. We can see that all the coefficients are statistically significant except for parks and residential mobility, which have p-values < 0.11. Although the park coefficient is not statistically significant, its interaction with population size is significant,

so we kept both in the model. We can see that mobilities for grocery & pharmacy, parks, transit, workplaces, and residential all have a positive relationship with new case counts, and retail & recreation mobility has a negative relationship. Furthermore, as the population increases, the strength of the relationship decreases for each of the mobilities. The exact impact of each mobility depends on the population size, which could make the coefficients hard to interpret. For interpretability, we created a table where each entry represents the impact for mobility on a particular region (with a particular population size).

Table 1: Linear Model Coefficients by Region

	Durham	Halton	Hamilton	Middlesex	Niagara	Ottawa	Peel	Toronto	Waterloo	York	Mean
population	2.117	1.811	1.744	1.536	1.455	3.1 23	4.682	9.005	1.808	3.587	3.087
new cases	-320187	-273819	-263669	-232330	-220014	-472237	-707999	-1361693	-273404	-542473	-466782
retail and recreation	0.386	0.168	0.12	-0.027	-0.085	1.101	2.209	5.282	0.166	1.431	1.075
grocery and pharmacy	-0.127	-0.096	-0.09	-0.069	-0.061	-0.228	-0.385	-0.819	-0.096	-0.275	-0.225
parks	-0.069	-0.041	-0.035	-0.015	-0.008	-0.163	-0.309	-0.712	-0.041	-0.207	-0.16
transit	-0.147	-0.101	-0.09	-0.059	-0.046	-0.3	-0.537	-1.195	-0.1	-0.371	-0.295
workplaces	-0.263	-0.136	-0.109	-0.023	0.01	-0.678	-1.321	-3.104	-0.135	-0.869	-0.663

Each column represents the linear model's mobility coefficients for the particular region. The column at the end contains the average value across the regions. We can see that, in general, an increase in the mobility value was associated with a decrease in new case counts. Retail & recreation increasing new case counts makes sense as this includes restaurants and confined public spaces where people gather. Grocery & pharmacy is complementary to retail & recreation and can be viewed as a better alternative mobility instead of retail & recreation because grocery & pharmacy are unavoidable. Similarly, parks mobility could be considered a “good” mobility because in general, parks have a lot of space to move around and are not confined. Work and transit mobilities should intuitively be positive, which suggests that a different model structure would be more effective. Furthermore, it is not for certain that the relationships between the mobilities and the new case counts is linear in nature. As a result, we consider improvements on the explanatory and response variables we are investigating.

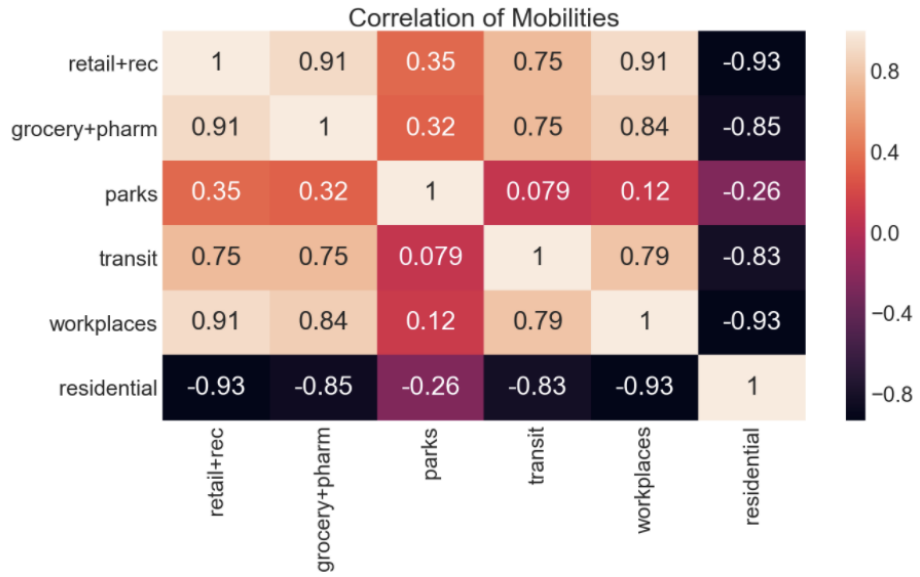

Figure 1: Correlation Matrix of the Mobility Data

The values in the correlation matrix of the mobility data (Figure 1) show that there are substantial linear correlations between the mobility variables. This motivates the use of uncorrelated, principal components in place of mobility variables in the model.

Principal Component Analysis

As shown in Figure 1, the mobility variables are strongly correlated with one another. In order to avoid the negative effect of such strong multicollinearity, such as instability in the estimation of the model parameters, we decided to use Principal Component Analysis (PCA). It was found that the first two components explained over 90% of the variation in the mobility data; therefore, we considered the first two components to represent the mobility data instead of the six mobilities. The principal components are linear combinations of the (standardized) original variables. Table 2 below provides the weights of each mobility in each principal component, where blank entries are 0 weighting.

Table 2: Individual Principal Component Weights

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
retail.rec	0.459			0.325	0.463	0.674
grocery.pharm	0.429	0.157	0.737	-0.478	-0.103	
parks	0.114	0.924	-0.329		-0.132	
transit	0.419	-0.288	-0.575	-0.633		

workplace	0.452	-0.161		0.397	-0.776	
residential	-0.461			-0.322	-0.391	0.721

Figure 2: (Linearly Scaled) First and Second Principal Component v. Mobility Variables. The first and second component are represented by the solid and dotted black lines respectively.

The principal component, once put on a scale similar to that of the original variables, seems to follow the trends of retail & recreation, grocery & pharmacy, transit, and workplaces very closely. The same is true for residential mobility, albeit reflected in the x-axis. This is expected since residential contributes a negative weight on the first principal component. The first principal component is a good proxy for all of the mobility variables except parks.

A similar analysis shows that the second principal component, on the other hand, seems to fit on the "park" variable very well. To sum it up, the first two principal components efficiently summarize all six mobility features.

Final Model Selection

We wish to conduct statistical inference on the effect of mobility on daily new case counts. However, since changes in mobility behaviours cannot correspond to immediate changes in new case counts, we instead study the relationship between average mobility over the last 7 days and 7-day rolling sums of new case counts over the future 7 days. This allows for a 14 day lag between mobility and onset of symptoms [3]. A generalized additive model (GAM) was fit to the nonlinear time series data to estimate the effect of mobility on the residuals. We applied a GAM for our investigation because the object of study is the impact of mobility on daily case counts; we are not interested in prediction and so smooth effects of recent case counts allow us to fit our model to the data and study mobility without accounting for all covariates relevant to new case counts.

The response, future 7-day new case count rolling sum, is assumed to be drawn from a Negative Binomial distribution. Originally we modelled the 7-day rolling sum with a Poisson response. However, the diagnostics revealed a poor distributional match as well as overdispersion. The overdispersion is what led us to applying a Negative Binomial fit. Furthermore, we decided on using the identity as our link function due to it having the highest percentage of deviance explained when compared to variations of the model fit using log and square root links. The model can be described explicitly as follows:

$$E[Y_i|X] = X^T\beta + f_1(N_{0i}) + f_2(N_{2i}) + f_3(N_{4i}) + f_4(N_{6i}) + f_5(N_{7i}) \qquad (\star)$$

The response, Y_i, denotes the 7-day rolling sum of new case counts from day i+1 to day i+7.

Fixed effects in ($\star$) are represented by X and includes Ontario municipal regions, as well as the results of our PCA on mobility over days i - 6 to i by Ontario region as stated in our section on PCA. It is important to mention that in our model PCA fixed effects interact with regional population. This seems intuitive as regions with larger populations should expect larger changes in new cases due to changes in mobility than regions with smaller populations.

Smooth effects are described by N_{ji} which details the rolling sum of new case counts from days *i-j*-6 to *i-j* in Ontario, for $j = 0, 2, 4, 6,$ and 7. This was designed under the assumption that recent daily case counts should impact new daily case counts, so we include multiple rolling sums of case counts over previous days to account for lags in new cases. Finally, β is the vector of parameter coefficients to be estimated.

Results

Table 3: Principal Component Model Coefficients

	Estimate	Standard Error	*p*-value
PCA1:Population	4.41×10^{-6}	3.70×10^{-7}	$< 2 \times 10^{-16}$
PCA2:Population	-2.89×10^{-6}	5.10×10^{-7}	$< 1.42 \times 10^{-8}$

As can be seen in the table above, the estimates for both PCA components used in the model were significant. While the estimates are small, their interaction with population can allow for significant effects. Population sizes lie between 479,183 and 2,965,713 so a unit increase in the first component corresponds to a predicted increase of expected 7-day rolling sum case counts between 2 and 13, all other variables held fixed.

Table 4: Coefficient Estimates of Standardized Mobility Variables from PCA (multiplied by10^6)

Retail & Recreation	2.04
Grocery & Pharmacy	1.5
Parks	-2.74
Transit	2.49
Workplaces	2.47
Residential	-2.04

By applying the principal component weights we can retrieve the coefficient estimates for the standardized mobility variables, as captured in Table 4. Most of the estimates for regions in Ontario were found to be significant. Similarly, the smooth effects with indices $j = 0, 2,$ and 4 were also significant.

Model diagnostics were sound for the most part, with only the constant variance assumption being uncertain. This could in part be due to possible 0 inflation resulting from some regions

having 0 new case counts for a long duration during the beginning of the pandemic before infections reached the area.

Discussion and Application

One option to validate our model's results could be to compare them to those obtained from other regions. An example of this would be to use mobility and case count data from certain regions in the United States.

Descriptive statistics

The descriptive statistics performed in this report illustrate on a high level the relationship between the Google mobility data and the increase in case counts by region. When each mobility is compared by itself against the new case counts, the relationships for the mobilities are statistically significant. Policymakers can use the direction and strength of the relationships to better understand which mobilities are most influential. Furthermore, by analyzing the descriptive statistics over time, policymakers can see how effective messages and policies are at influencing the individual mobilities.

Linear Model

The linear model is illustrative because it considers all the mobilities simultaneously and therefore, is investigating the isolated impact of each mobility, given the other mobilities are held fixed. Using an intuitive understanding of what the mobilities represent is useful for better interpreting the coefficients in the model. It is clear, however, that the linear model is not all-encompassing - there is some behaviour causing spread that is not captured in the mobility data. For example, the model coefficients for workplace and transit mobilities are negative, which means that as workplaces and transit stations are visited less frequently, cases go up. This motivates the use of a model that uses the mobility data to understand the deviation from the trend in cases. This structure of analysis can be used on any data that has a similar structure; therefore, this model can be extrapolated to data from the United States, Europe, or elsewhere.

Generalized Additive Model

Table 4 provides information on the effect individual mobility variables have on new case counts. Note that they all have the same order of magnitude, and all estimates are positive except for parks and residential. This is inline with our intuition because an increase in residential mobility suggests more people are staying home and abstaining from unnecessary travel. Similarly, due to the large open spaces provided by parks, it's reasonable to assume an increase in park mobility would lead to an increase in physical distancing and hence a decrease in new case counts. Conversely, confined spaces and close proximity of people are recurring themes of the other mobility types. It is then no surprise that our model indicates their positive association with new case counts.

In particular, the parks coefficient estimate has the largest magnitude, which suggests to policy makers that encouraging individuals to spend more time in parks might aid public health efforts. Transit and workplace mobility were approximately tied for the largest positive effects on new case counts. Therefore, new measures to reduce public transit and onsite workplace attendance should be prioritised over restrictions on retail and recreation.

The investigation has potential limitations as well. Many regions have missing data in the Google mobility reports. This is especially prevalent in areas of low population, many of which were taken out as a result. A consequence of this is the potential overrepresentation of urban regions, where data collection is more frequent and consistent.

Furthermore, formulating our model we were aware of the potential to overestimate the standard errors of mobility estimates due to the data focusing solely on Ontario. Provinces issue regional lockdowns and social restrictions, hence a possibility of mobility behaviour in Ontario being highly correlated across regions as many Ontario regions undergo simultaneous restrictions. To account for this concern, the model can be applied to datasets whose regions undergo a diverse array of social distancing implementations.

References

[1] Bajardi P, Poletto C, Ramasco J.J, Tizzoni M, Colizza V, Vespignani A. (2011). Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic. PLoS ONE 6(1): e16591. https://doi.org/10.1371/journal.pone.0016591.

[2] Gatalo O, Tseng K, Hamilton A, Lin G, Klein E. (2020). Associations between phone mobility data and COVID-19 cases. Lancet Infect Dis 2020; published online Sept 15. https://doi.org/10.1016/S1473- 3099(20)30725-8.

[3] Badr H.S, Du H, Marshall M, Dong E, Squire M.M, Gardner L.M. (2020. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. Lancet Infect Dis 2020; published online July 1. https://doi.org/10.1016/S1473-3099(20)30553-3.

[4] Kramer M.U.G. et al. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. Science; Vol. 368, Issue 6490, pp. 493-497 DOI: 10.1126/science.abb4218.

[5] Google LLC. (2020). Google COVID-19 Community Mobility Reports. Retrieved from https://www.google.com/covid19/mobility/

[6] Government of Ontario. (2020). Confirmed positive cases of COVID19 in Ontario. Retrieved from https://data.ontario.ca/en/dataset/confirmed-positive-cases-of-covid-19-in-ontario/resource/455fd63b-603d-4608-8216-7d8647f43350

Appendix A

Table 4: Mean Mobility Data Comparisons Between Weekdays and Weekends.

Mobility Type	**Weekdays**	**Weekends**	**t-statistic**	**p value**
Retail + Recreation	-38.4	-36.5	-0.89	0.375
Grocery + Pharmacy	-11.8	-12.8	0.66	0.515
Parks	29.8	59.6	-4.42	0.000
Transit	-54.8	-40.0	-6.61	0.000
Workplaces	-50.7	-18.6	-15.03	0.000
Residential	18.4	6.7	14.18	0.000

Table 5: Total Case Counts per Day

Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
8661	8263	8130	7656	8051	7404	6700

Appendix B

```
model summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:         new_cases_lag7   R-squared (uncentered):                   0.915
Model:                            OLS   Adj. R-squared (uncentered):              0.911
Method:                 Least Squares   F-statistic:                              295.7
Date:                Wed, 30 Dec 2020   Prob (F-statistic):                   1.02e-196
Time:                        09:27:38   Log-Likelihood:                         -1551.6
No. Observations:                 401   AIC:                                      3131.
Df Residuals:                     387   BIC:                                      3187.
Df Model:                          14
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
population                 -2.718e-05    1.8e-05     -1.511      0.132   -6.25e-05    8.19e-06
pop_density                    0.0379      0.009      4.338      0.000       0.021       0.055
retail+rec                   -11.6845      2.811     -4.157      0.000     -17.211      -6.158
grocery+pharm                 11.8354      3.142      3.767      0.000       5.658      18.013
parks                          1.2603      0.363      3.469      0.001       0.546       1.975
transit                        4.0938      2.490      1.644      0.101      -0.801       8.989
workplaces                     7.0906      2.549      2.782      0.006       2.079      12.102
residential                   20.9556     12.202      1.717      0.087      -3.034      44.946
retail+rec*population      2.194e-05   4.72e-06      4.646      0.000    1.27e-05    3.12e-05
grocery+pharm*population  -2.039e-05   5.25e-06     -3.886      0.000   -3.07e-05   -1.01e-05
parks*population          -2.522e-06   6.21e-07     -4.058      0.000   -3.74e-06    -1.3e-06
transit*population        -8.808e-06   4.21e-06     -2.095      0.037   -1.71e-05    -5.4e-07
workplaces*population     -9.765e-06   4.35e-06     -2.245      0.025   -1.83e-05   -1.21e-06
residential*population    -2.674e-05   2.05e-05     -1.302      0.194   -6.71e-05    1.37e-05
==============================================================================
Omnibus:                       27.519   Durbin-Watson:                   0.132
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               31.217
Skew:                           0.651   Prob(JB):                     1.66e-07
Kurtosis:                       3.418   Cond. No.                     1.21e+09
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 1.21e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
mean response: new_cases_lag7    30.674687
dtype: float64
performance (rmse): 9.030690770958303
```

Figure 4: Linear Model Summary

Appendix C

List of Ontario regions excluded: Algoma, Brant, Brantford, Bruce, Chatham-Kent, Cochrane, Dufferin, Elgin, Essex, Frontenac, Sudbury, Grey, Haldimand, Haliburton, Hastings, Huron, Kawartha Lakes, Kenora, Lambton, Lanark, Leeds and Grenville, Lennox and Addington, Manitoulin, Muskoka, Nipissing, Norfolk, Northumberland, Oxford, Parry Sound, Perth, Peterborough, Prescott and Russell, Prince Edward, Rainy River, Renfrew, Simcoe, Stormont, Dundas, and Glengarry, Thunder Bay, Timiskaming, and Wellington.