CAPP 30254 Final Project Report

Monitoring Mobility in the COVID-19 Pandemic

Steven Buschbach (sbuschbach)
Noah Selman (nselman)
Xiaoyan (Angela) Wang (xiaoyanw)

Executive Summary

With the advent of the COVID-19 pandemic, policymakers on local, state, and federal levels have all promoted policies to encourage 'social distancing' to prevent its spread. This report describes the use of publicly-available mobility data from Google to estimate county-level physical mobility trends, which may be useful to policymakers seeking to curb contagious disease. We fit a variety of supervised learning models to the data, with the intention of using current information to predict future mobility. We identify the random forest regressor as the best predictor. An initial iteration of this analysis done on data from February to early May performed fairly well at predicting out-of-sample observations. But surprisingly, refreshing it to include more recent data resulted in a best model that was not much more accurate than a 'global mean' baseline. We discuss potential explanations, including the fact that our models do not currently capture difficult-to-measure features like 'lockdown fatigue.' We then discuss ethical implications of a more successful version of our analysis and potential improvements that could make this tool more useful.

Background

During the COVID-19 pandemic, enforcing "social distancing" is imperative to public health. The easiest way for local policy makers to ensure social distancing is by setting blanket restrictions for public access to certain areas. For example, in early March 2020, Chicago mayor Lori Lightfoot prohibited entry to all lake-adjacent parks for an indefinite period, causing great distress to many Chicagoins. Governments across the nation have enacted restrictions and policies aimed at enforcing social distancing to varying degrees. These measures range from instructing law enforcement to issue fines for violation of social distancing policies to closing retails centers, restaurants and other public areas. While limiting the spread of the virus, these measures sacrifice economic productivity and quality of life for local residents.

Fortunately, technology companies such as Google release data describing how frequently different location categories are visited at a county-day level. This presents an opportunity to predict instances of high mobility or crowdedness for a particular type of location. Policymakers could apply a mobility-predicting model to refine social distancing enforcement mechanisms that continue to protect public health while minimizing restrictions of public areas and businesses. For example, in lieu of an indefinite blanket lakefront ban, Mayor Lightfoot could issue more targeted restrictions based on a model's predictions of high mobility while still achieving similar levels of social distancing. This might allow for public areas to remain open during low-traffic periods when the model predicts risks of virus spread are minimal. This has potential to reduce economic damage and increase community-member compliance with official social distancing rules.

Data

Our data is longitudinal with a temporal component. County-date combinations identify rows. We built our dataset from numerous sources, individually detailed below.

Google Mobility Data (Target Variable)

Our set of target variables is provided by Google. At a county-date level, they report change in baseline mobility for a variety of location types:

- Retail & Recreation
- Grocery & Pharmacy
- Parks
- Transit Stations
- Workplaces
- Residential

Values for each observation represent a percent deviation from an expected baseline mobility. For example, a value of 50 would represent a 50 percent increase in mobility from an expected baseline. The assignment of the baseline is not transparent, but is said to be based on day-of-week and time-of-year expectations. Given Google's eminent capacity to collect data on mobility trends, this seems a reliable benchmark. We created lagged moving averages of our target variable to use as features in our model. This is appropriated because policy-makers will have access to this data when predicting future mobility. More information can be found from the data source page.

While all of our analysis and code is built to handle any of these target variables, time and computing constraints limited our scope to just predicting Retail & Recreation mobility. Nonetheless, predicting any of our target variables would inform policy decisions aimed at reducing crowding. Our code can be readily adapted to train models for these other variables.

American Community Survey

From the American Community Survey we assembled a large array of demographic features at the county level. These included data on county age and income distribution, frequent means of travel, computer ownership, internet access, education, and disability status. In preparing the data, features were scaled by inverse county population when appropriate. Additional details about the data can be found at the Census website.

County Business Patterns and Non-Employer Statistics

From the County Business Patterns and Non-Employer Statistics we pulled data that describe the distribution of employees between different possible employment sectors as defined by NAICS codes. NAICS codes contain up to six digits, with each successive digit providing additional detail about the sector of work. The values in the data that we pulled represent the number of workers belonging to particular NAICS codes. When exploring the data, we found irreconcilable inconsistencies in the total number of workers by county when measuring at different numbers of digits. Because of this issue and apparent different categorization behavior between counties (wherein some counties appear to not assign employees to certain common NAICS codes), we elected to use only the first two NAICS digits. We rescaled the data by the inverse total number of workers. Additional information on the data can be found on the Census website.

Center for Disease Control

The CDC provides by-county COVID cases and deaths. Their website can provide additional details. Furthermore, the Behavioral Risk Factor Surveillance Survey (BRFSS) is conducted by the CDC and obtains state and metro-level data on risk behaviors and chronic health indicators, such as smoking frequency, incidence of diabetes or asthma, obesity, etc. We included these data because underlying health factors may affect people's mobility during a pandemic. Information is available here. We first merged the metro-level BRFSS data onto our counties. For any counties that were not represented in the BRFSS metro-level data (i.e., rural counties), we used state-level BRFSS data.

National Oceanic and Atmospheric Administration

From NOAA's FTP, we pulled daily weather station level measurements of temperature and precipitation. Each station was geolocated and measurements were aggregated by county. Upon merging to the master dataset, county-dates without a measurement were imputed using the state mean for that day. We created rolling temperature averages and a precipitation dummy as additional features.

State and Local COVID-related Policies

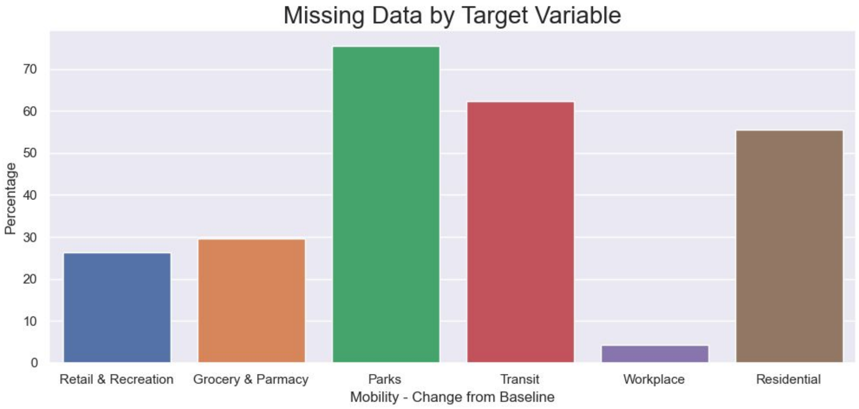
County level dates that local COVID-related policies were enacted were aggregated by a team of Johns Hopkins researchers and posted on Jie Ying Wu's Github page. These data were assembled from numerous news sources. Observations in the raw data represent an ordinal date on which a policy was implemented for a given county. From these columns, we created a dummy feature that signals 1 on days greater than or equal to the raw date. One notable caveat of this data source is it has not been updated for two months.

MIT Election Lab

The MIT Election Lab provides historic county-level election returns. We pulled data from the 2016 president election and calculated Democratic and Republican vote share. These data may indicate a dimension of local culture that could influence mobility. More information on the data can be found at the MIT Election Lab website.

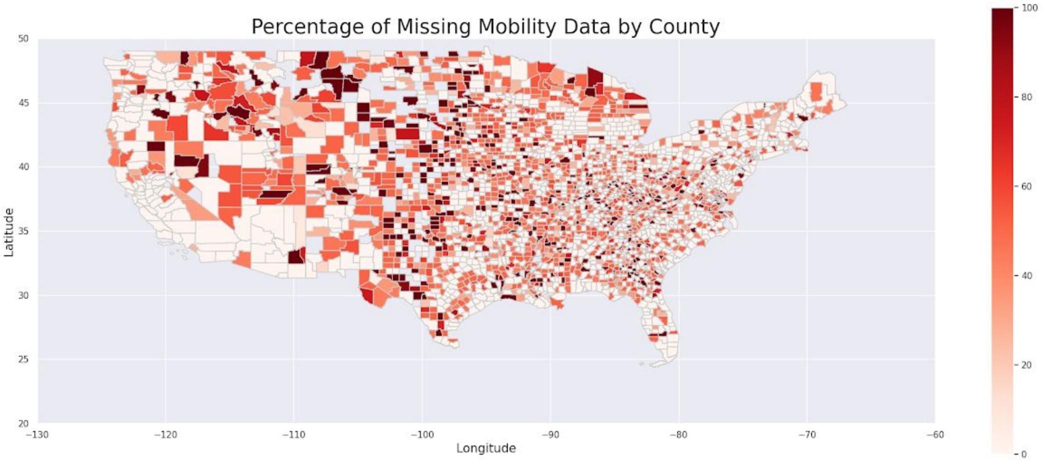
Missing Target Variables

Google does not report a mobility measure during instances in which they estimate they'd be risking users' anonymity. Consequently, our potential target variables are missing not-insignificant amounts of data. The 'Retail and Recreation' mobility that we use as the target variable in this version of our project has about 25% missing observations.



Since anonymity is likely correlated with population, we cannot assume the data are missing completely at random. Moreover, since our target is continuous, we cannot simply predict null as one categorical outcome. A brief discussion on how this could be addressed is provided in the 'Extensions for Future Work' section below. For now, we do not include observations with a missing target variable in our training or test set.

Missing data appears to be clustered in rural areas:



Data Preparation

The first step in the data preparation process was to clean the Google mobility data. We associated each county with a FIPS code to facilitate merging of data from other sources.

Afterwards, our data from each additional source were prepared separately, prior to merging onto the final data frame. Preparation steps often included transforming the data such that rows were defined by geographic levels and dates, sorting geographies into FIPS codes, imputing missing data, and creating basic features. The most frequent imputation procedure was to fill a missing observation with the state mean for that date. Most created features simply involved rescaling absolute measures by a relevant measure of size for that county. For example, to create variables regarding county-level participation in an industry, we divided the number of workers from that industry by the total number of workers.

Machine Learning Details

As stated previously, our code and methodology can be applied to any of the aforementioned mobility features as a target variable, but limited time and computing resources restricted us to analyzing just one target: change in Retail and Recreation Mobility.

Dimensionality Reduction

Our final dataset had 138 features. To reduce feature complexity and avoid problems associated with the curse of dimensionality, we investigated feature reduction techniques. We chose Principal Component Analysis (PCA), which identifies the linear combinations of our features that explain the most variance in our training data. PCA gave us a new dataset with about half as many features while explaining 95% of the variation in our original dataset.

Model Fitting

Since our target variable is continuous, we focused on supervised regression methods. We fit the following models to our PCA-reduced dataset for a number of possible parameters:

- Linear Regression
- Penalized Regression (LASSO and Ridge)
- Linear SVR (the regression analogue of support vector classifiers)
- Random Forest Regressor
- AdaBoost Regressor (with decision tree base regressor)
- K-Nearest Neighbors Regressor

This set of models represents a wide array of techniques, from traditional regression (e.g., OLS, LASSO, ridge) to nonparametric regression (KNN) to ensemble methods (random forest and AdaBoost). Because our main aim is creating an accurate tool to guide policy decisions, we were agnostic about which type of model we should use.

As a robustness-check that PCA feature-reduction would truly improve the predictive value of our final model, we fit a number of sparsity-encouraging models on the full, non-PCA dataset. These included:

- LASSO Regression
- Random Forest Regressor (with a random subset of features)
- AdaBoost Regressor

These models fit on non-PCA reduced data were also useful in identifying which features in our original dataset were most important in prediction, since identifying which principal components were important is not very informative to a policymaker, outside the predictive capacity of our final model.

Evaluation and Results

Cross Validation

Because our data has a temporal component, a traditional k-fold cross validation technique is not appropriate. We want to train a model that predicts future mobility, not one that interpolates mobility between dates. So we use a temporally-aware cross-validation technique.

Our technique is implemented as follows. First, we order all of our observations by date. We assign the first 90% of observations, representing the first 90% of data in time (approximately February through mid-May), as our training set. The final 10% was our test set, which we used to assess the final chosen model after validation. In both the validation analysis (described below) and test analysis, we used mean absolute error as our accuracy metric of choice, since it is more robust to outliers than mean squared error, and our target variable is very noisy and likely to contain outliers.

To validate our parameters, we use an incremental validation technique with 3 stages, where we train and validate our data on successively larger subsets of dates. For example, in the first validation fold, we train our models on the first 60% of the data temporally, and hold out the 60-70th percentile of dates for validation. In the second fold, we train our models on the first 70% of the data temporally, and hold out the 70-80th of dates for validation. In the last fold, we train our models on the first 80% of the data temporally, and hold out the 80-90th percentile of dates for validation. Only this last step makes full use of the non-test data.

Evaluating Results

Mean absolute error is our primary metric for evaluating models, which we have chosen over mean squared error because mean absolute error is less sensitive to outliers. For each model class, we calculate the mean absolute error on each validation set described above and average them. The model/hyperparameter set that has the lowest error is our 'best' model. Averaging across this 3-fold temporal validation process prevents overfitting on any one validation set.[1]

Ultimately, several random forest regressors topped our list of best models after validation (the best model having 1000 trees and log2 number of features), followed by k-nearest neighbors (9 neighbors being the best). Furthermore, models fit using the PCA-reduced data tended to have more accurate predictions, a sign that feature reduction may have successfully prevented overfitting.[2]

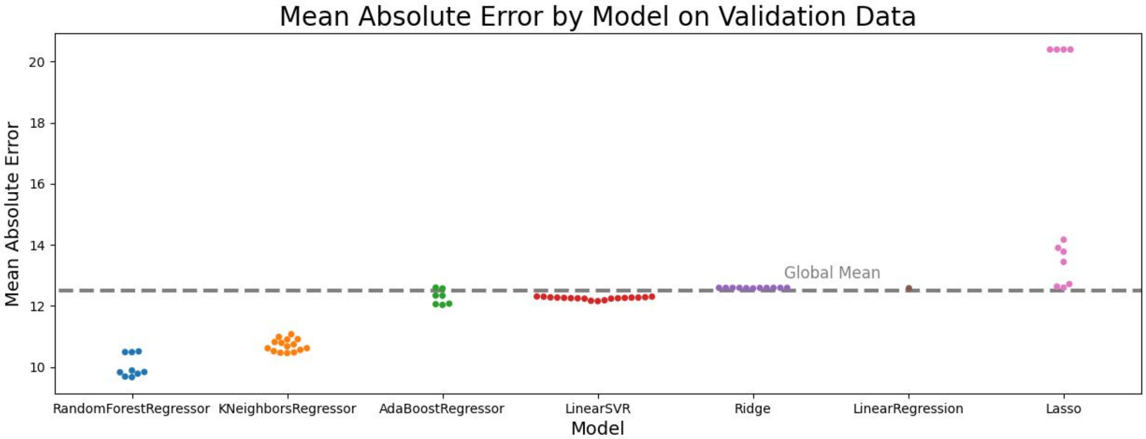
Validation Results

The following chart displays model MAE compared to an MAE calculated naive mean prediction.[3]

[1] An earlier presentation of this project used just one validation set, due to time constraints on fitting all of our models

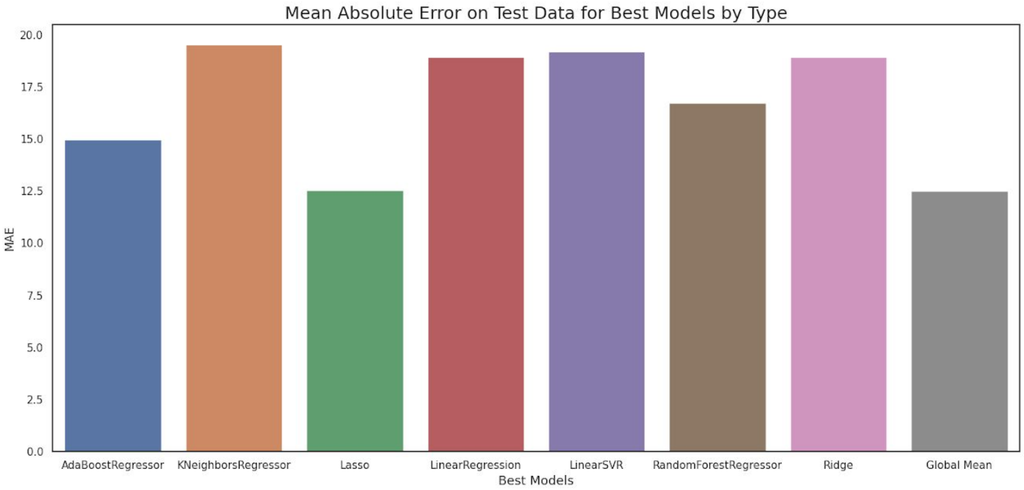
[2] We retrained models numerous times as new data became available. There were occasions in which non-PCA models performed better.

[3] The mean prediction is simply the global mean of all observations in the training data. We also tested our data against a county based mean which is equivalent to a linear regression with only county fixed effects. This mean outperformed our best model. However, our models are county-agnostic in that they do not incorporate county fixed effects due to our focus on features informed by an intuition of the mechanism by which they'd influence mobility. Moreover, the MAE for the county-mean model exhibited a peculiar pattern: the error sharply dropped for each temporally successive cross validation slice. Looking only at the final slice, the null model scores an MAE of about 8, far outperforming any of our models. Since a policymaker would use all



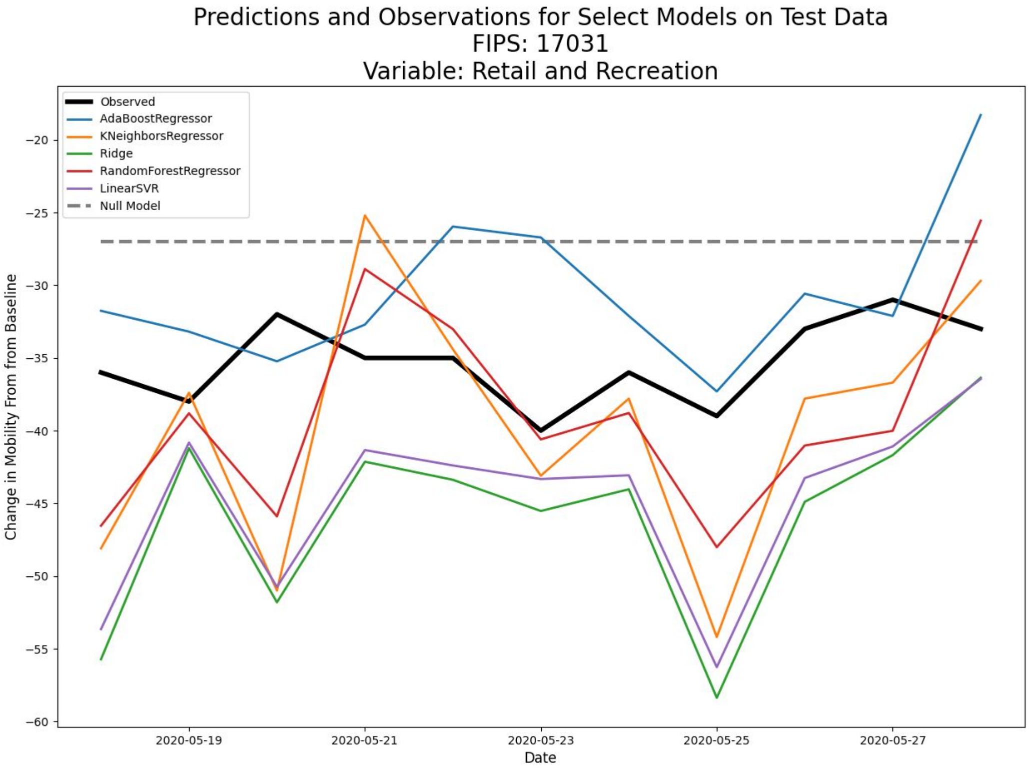
Test Results

Now, we compared each model with fitted hyperparameters on our test data, and then we compared these results to the baseline model of predicting the global mean of the training data for every observation. Surprisingly, our models do not provide any predictive benefit above the naive ‘global mean’ model for the test data. This is especially interesting because in an earlier iteration of this analysis on data that ended in early May, our fitted models tended to outperform the global mean baseline. It was only in our most recent data pull that our fitted models seem to have become worse at prediction than the baseline. This suggests pecuralities in the data and/or inadequate feature design. While further analysis needs to be done, right now our best guess is that our models did a poor job at capturing ‘lockdown fatigue,’ which became more apparent in the data from mid-May onwards.



available data for a naive prediction, this may be more practical comparison. However, because of high variance in our null model’s MAE, it may begin to perform worse as data is added. This model is omitted from the chart since its use of county fixed effects makes interpreting the comparison difficult.

Our best validated models appear to perform much better for counties with high population density. Below, we show how various models performed at predicting Cook County mobility. It is clear that our models still fail to capture a lot of the underlying variance in our target variable



From our non-PCA reduced models, we were able to extract which features proved to be most predictive in assessing mobility. It turns out that our model predicts that mobility did substantially decrease as a result of policies like closed schools, restaurant dine-in bans, etc. Other variables that proved to be important included COVID-19 case-related variables (e.g., cases per capita and raw number of cases), as well as weather variables (people going out more when there is a higher maximum temperature).

Policy Recommendations

A specific policy recommendation is not appropriate. Our model as-is cannot be accurate enough to make such consequential policy decisions, but it can be built upon and refined to eventually come closer to that goal. The ultimate purpose of training this model was to provide a tool which could predict instances of high mobility so that policymakers could enact more narrowly-tailored measures to enforce social distancing without employing blanket, indefinite bans of certain spaces. While not achieving this goal, our model takes the necessary first steps. Policymakers would be wise to heed our note below in the 'Extensions for Future Work' section and invest in development of such a tool. Generally, specific policy recommendations should not be founded on results from students' first attempt at machine learning-based modeling.

Ethics

There are a few ethical concerns policy-makers must mind while using a model of this type. First, since large portions of data are missing for many counties, use of this model may produce errant predictions if data is not missing completely at random. Secondly, it may be the case that our error isn't uniform by county. Such a model could encourage policy-makers to enact too strict or lax social distancing enforcement, possibly leading to health or economic problems.

Furthermore, this model does not consider heterogeneous consequences of social distancing between counties.Imagine a hypothetical county with many lower SES, 'essential jobs' workers that still must work during the pandemic. This model could justify economic restrictions and shutdowns because of these county characteristics. These shutdowns could disproportionately harm such a community.

One way to go about doing a more thorough audit for potential bias would be to do a more systematic investigation of where our model is predicting high mobility, and where it is predicting low mobility. What do these counties have in common? Can we simulate hypothetical counties that have high or low mobility, using our model? If for any given county, we change key variables in our model, such as ACS racial makeup or SES indicators, how does this change our result? These are potential avenues to pursue a more thorough audit on any potential bias in our model.

Caveats

One major limitation we had was missing data. About 28% of county-date observations were missing for mobility data. This was especially true for small, rural counties, but the missing data was not necessarily exclusive to these areas. In fitting our models, we dropped the observations that were missing this target variable. One potential way to better-account for this is by performing a separate classification step before we fit a regression model. In the classification step, we model whether a county-date observation is 'missing' or 'not missing,' and then we add in our current model that estimates mobility for nonmissing observations. We did not do this for 2 reasons: (1) time constraints on this project, and (2) knowing whether mobility data will be 'missing' may not be so useful to a policymaker. But a more robust model would be able to identify why some of the data is missing in the first place, since the source of this missingness may affect the generalizability of our current regression model.

A major caveat is the relatively short time frame of our sample, giving uncertainty in its generalizability for future use. There are reasons to believe there may be a structural break between the period used to train the model and the coming months:

- Social movements in reaction to horrific police brutality may have recalibrated people's willingness to socially distance
- Temperature is an important feature and will vary over a different range during other seasons that are not present in our current data
- People may react differently to future policies due to "lockdown fatigue"

Nonetheless, our results could represent a small data point in a policymaker's informational toolkit to help control the pandemic.

Extensions for Future Work

Limited time, computational resources, and available data prohibited us from carrying out our analysis to a more rigorous extent. First, we would have been more creative in how we addressed missing observations in our target variable. Below are a few work-arounds we would have liked to try:

- Implement a two-stage mode in which we first perform a classification step to predict whether a county-day will be null, and then conditionally add it to our current model (described above).
- Bucketing our target variable and including null observations as a category. Bucketing the target variable is defensible because our predictions are only actionable for policy makers if mobility reaches a particularly high level, which could be a particular bucket.
- Adding county-level fixed effects
- Adding time-variant features that capture public willingness to comply with social distancing rules[4]
- Using mobility from another source, such as Facebook which may be better populated

Once caveat of those approaches is knowing an observation will be "missing" is not useful to a policymaker. Generally, we would have liked to conduct a deeper analysis into our missing data, perhaps trying to predict it using other features. We also would have liked to test if our error is correlated with any of our features.

[4] This could include features like days-since-stay-at-home-policy. We had expected access to Twitter data which would have allowed the creation of features that track public interest in quarantine. These features may have added valuable variance.