CAPP 30254 Final Project: Project Proposal

Monitoring Mobility in the COVID-19 Pandemic

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

Project Summary

The present COVID-19 pandemic has forced policymakers to make many difficult economic and social decisions, often with little empirical research to guide them. One aspect that determines the spread of disease in a population is people's mobility (i.e., driving, foot traffic, transit rides). Reducing mobility and hence close contact with others is a key driver for the ongoing lockdowns in the United States and around the world. Our goal is to use published Google mobility data to predict car / foot traffic around the United States. To identify what drives mobility in different areas at different times, we will incorporate data from a variety of sources including CDC health data, Census/ACS demographic and economic data, weather data, and contemporaneous COVID-19 infection data. If time allows, we are also considering including some basic NLP features using Twitter text data.

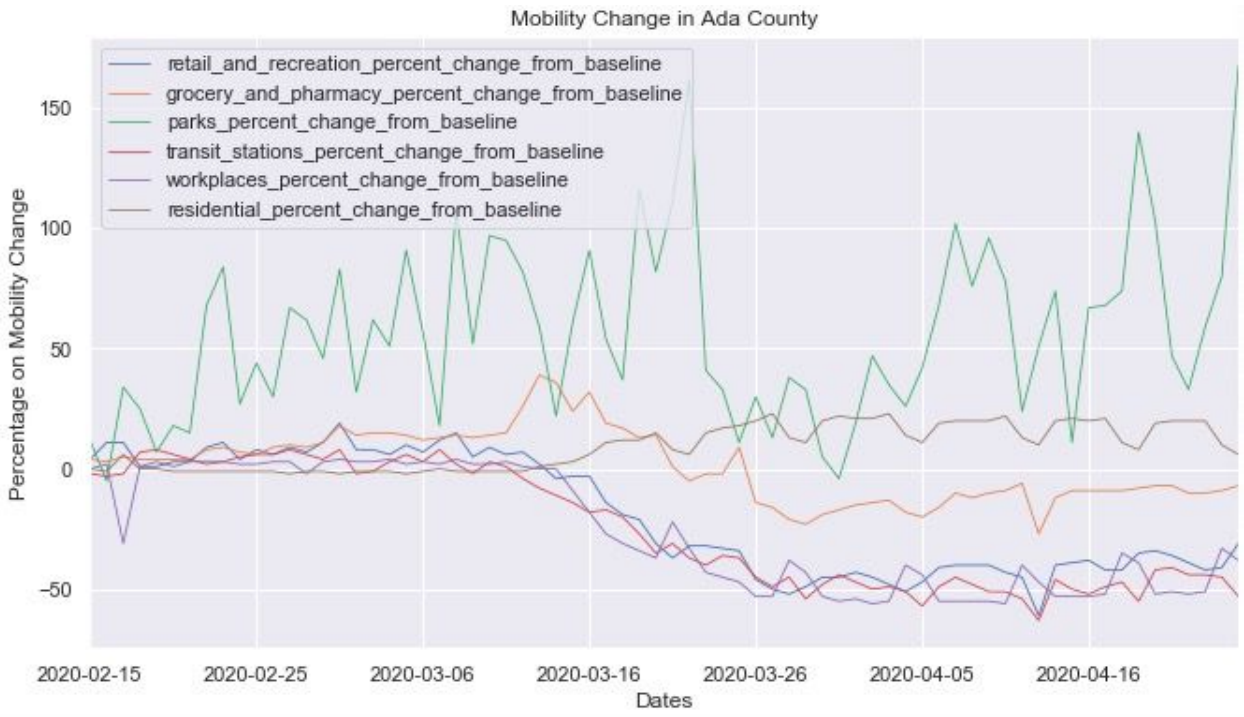
This tool could allow local officials to predict changes in resident mobility that may increase COVID-19 transmission rates and adjust stay-at-home policies and enforcement mechanisms accordingly. Health policy decision makers could find this model particularly useful. Lastly, our predictions may be of interest to statisticians and epidemiologists attempting to predict the spread of COVID-19, since future person-to-person interactions could be one of many components of their models. We will validate our results by testing predictions on future dates.

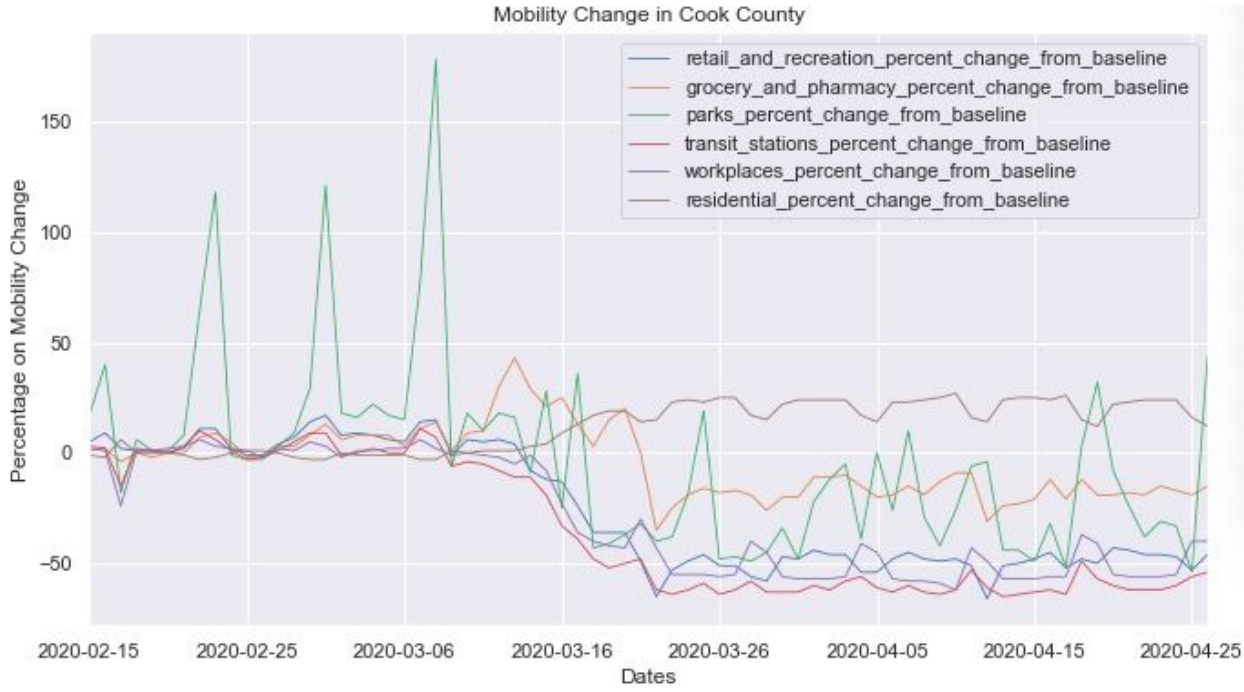
Data

Our target variables will come from Google's Mobility Data. They record percentage deviation from a baseline in visits to a type of location (e.g., residential, grocery and pharmacy, workplace, transit stations). The data is disaggregated at the county level daily since February 15th, and the data is generally well-populated, although there are some missing values.

We'll attempt to predict this variable using features listed in the table below. We observe substantial variation in the outcome both temporally and geographically, which should help us suss out any underlying signal. Temporally, for example, we find that workplace traffic was down on average -0.07 standard deviations from the baseline across all counties during the second half of February, but down -37.42 for the month of April. Geographically, we see that our outcome variable on any given day has a standard deviation of anywhere from 8-10 across counties in April, and 3-4 across counties in February. (Intuitively, this means that there was less heterogeneity in counties relative mobility in February, and much more by April as some counties began social distancing while others lagged). We have also included two plots of

mobility data for two counties below, as an example. Notice how mobility in parks in Ada County continues at a high rate, whereas it decreases in Cook County by a large amount.





To inform our understanding of the data, we'll calculate further summary statistics, look for patterns in missing or unusual values, and plot the data. While we have identified other data sources for our predictor variables, we must also go through a similar data exploration and

cleaning process for these as well. Through these strategies, we can verify our data is consistent with basic intuitions in time trends and relationships between variables.

Once we have explored and cleaned our data we'll identify models best-suited to predict google mobility. We'll probably try out a couple and see what works best on our validation data.

Data Set	Source	Features of Interest
Google Mobility Data	**https://www.google.com/covid19/mobility/**	**Change in density at various types of locations - target variable**
NYTimes Case Data	**https://github.com/nytimes/covid-19-data**	**COVID Infections and deaths**
COVID Tracking Project	**https://covidtracking.com/**	**Testing and ICU data at state level**
CDC Data	**https://wonder.cdc.gov/** **Chronic Disease Indicators: https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-/g4ie-h725** **Behavioral Risk Factors: https://chronicdata.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factor-Surveillance-System-BRFSS-P/dttw-5yxu** **Obesity: https://chronicdata.cdc.gov/Behavioral-Risk-Factors/BRFSS-Table-of-Overweight-and-Obesity-BMI-/fqb7-mgjf** **Provisional COVID Deaths by County: https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-in-the-United-St/k**	**Health indicators by state. Responses to particular survey questions define rows. We can identify questions of interest and aggregate results by state.**

	n79-hsxy	
ACS / Census	https://www.census.gov/data/developers/data-sets/ACS-supplemental-data.html	**County demographic data is available. Population density and socioeconomic information will be for interest.**
Weather Data (NOAA)	https://www.ncei.noaa.gov/support/access-data-service-api-user-documentation	**Weather features like temperature and precipitation effect decisions to leave home**
School Closure Data	https://www.edweek.org/ew/section/multimedia/map-coronavirus-and-school-closures.html	**Date public schools ordered to close by state**
Apple Mobility Data	https://www.apple.com/covid19/mobility	**Number of routing requests - potential target variable**

Machine Learning

The primary focus of our project is regression, since the predictor of interest is mobility in a given county on a given day. Because there is a temporal aspect to this data, we expect the final model to enable predictions up to one week out (e.g., use data from the past week or more to predict mobility changes in the coming days). We believe that this question in particular will be useful to policymakers. But we have also considered an adjacent regression problem: namely, can we use time-invariant features to predict which counties will most likely be mobility 'hotspots'? This question could be of broader use to policymakers because it allows for generalization beyond a fixed window of time.

We try a variety of regression models, the simplest being penalized and unpenalized linear models. We will then extend to more complex models, such as decision tree / random forest regressors, as well as neural nets. We expect these more complex models to perform better, since there is likely a large degree of nonlinearity and complicated interactions that goes into predicting mobility.

Furthermore, the fact that there will be hundreds of potential features related to county demographic / health / economic / weather data, we will likely have to perform some sort of feature reduction like PCA. We have also considered incorporating basic NLP statistics from Twitter for each county, which would add to our feature count. Some of these reduced 'features'

may not lend to intuitive interpretations of linear regression coefficients, which would reduce the advantage of simpler linear models (i.e., their interpretability).

Evaluation

We plan on using k-fold cross validation to tune any hyperparameters in each of our models. For validating the correctness of each model, we are considering two options: the first is to use k-fold cross validation, similar to hyperparameter tuning. But given that the ultimate application of this project would be to predict future changes in mobility behavior, a more likely option for us to evaluate models would be to take advantage of the constantly updating COVID-19 data available. We would do this by stopping data collection at a given date, training our models, and then obtaining the most recent data (e.g., the two weeks after we pulled the training data) and using this as testing data to compare our models.

Ethics

Predicting traffic mobility has a few general ethical advantages to society. In current times in particular, when hospitals are overloaded with patients suffering from COVID-19, access to regular healthcare is limited, visiting hospitals can also be a risky endeavor. So, with these predictions at hand, specific communities of people (e.g., patients with medical conditions, patients currently in rehabilitation, expectant mothers, etc.) can take advantage of this data to plan their future commute to the hospitals for essential medical appointments.

In Machine Learning, it is important to think whether our models might suffer from unintended biases. Hypothetically, in our models, if there was a specific event or circumstance in the most recent week that caused a large spike in traffic, then that might serve as a large outlier in our training dataset and might affect future predictions. Such events are not unexpected because there have been various religious gatherings and other rallies in recent times. But this specific bias is easy to handle by simply manually choosing not to train our models on specific data points which we suspect are outliers. One potential future work is to make our models robust by having it automatically detect these scenarios without human intervention.