Model Evaluation

Richa Jain and Estee Cramer

Introduction

Many of the models submitting to the COVID-19 Forecast Hub incorporate data about mobility. We look to compare these models with with models that do not incorporate such data to evaluate and compare their performances.

Mobility data may include different trends in a variety of areas including grocery shopping, parks, transit, retail, recreation, and workplaces. Mobility data tells us how much the population of a certain area is moving around and how visits to certain places are changing over time.

Methods

Step 1: Create Table of Model Characteristics

- Look for which models use social distancing data
- Data used by models (demographic data, hospitalization data)
- Model type (SEIR, Baysian, Statistical, etc.)

Step 2: Inclusion Criteria

- Locations: California (06), New York (36)
- Time period: December 5 2020 - December 19 2020 (California) & December 17 2020 (NY)
- Target: Incident Cases
- Horizons: 1 week ahead
- Models:
  - 2 models without mobility
    - LANL-GrowthRate, COVIDhub-ensemble, RobertWalraven-ESG
  - 2 models with only mobility
    - IowaStateLW-STEM, JHU_CSSE-DECOM, UVA-Ensemble
  - Baseline
    - COVIDhub-baseline
- Relative WIS = average_WIS_mobility/average_WIS_non-mobility
- Relative WIS (baseline) = average_WIS_after/average_WIS_before

Model Characteristics

Model	Case Data	Model Type	Social Distancing Assumptions?	Mobility Data?	Notes
COVIDhub-baseline	JHU CSSE	Median prediction at all future horizons	no	no	
LANL-GrowthRate	JHU CSSE	Statistical dynamical growth model	no	no	
COVIDhub-ensemble		Unweighted average or median of submitted forecasts	no	no	
RobertWalraven-ESG	JHU CSSE	SEIR model	no	no	
IowaStateLW-STEM	NYT, Johns Hopkins, Covid Tracking Project, USA Facts	Nonparametric space-time disease transmission model	no	yes	
UVA-Ensemble	CDC	AR, ISTM, SEIR model	no	yes (Baidu)	
JHU_CSSE-DECOM	JHU CSSE	Empirical machine learning model	no	yes (SafeGraph)	
UMich-RidgeTfReg	JHU CSSE	Ridge regression	no	yes	

These models were selected by first determining that we wanted to focus on case forecasts and finding the models that have submitted case forecasts. Then I separated models with and without mobility data. Finally, once a time period was determined, I went through the models and selected those which were submitting case forecasts during that time.

Step 3: Evaluation graphs

- After deciding on inclusion criteria, use covidhubUtils to score forecasts and determine which models are best.

California

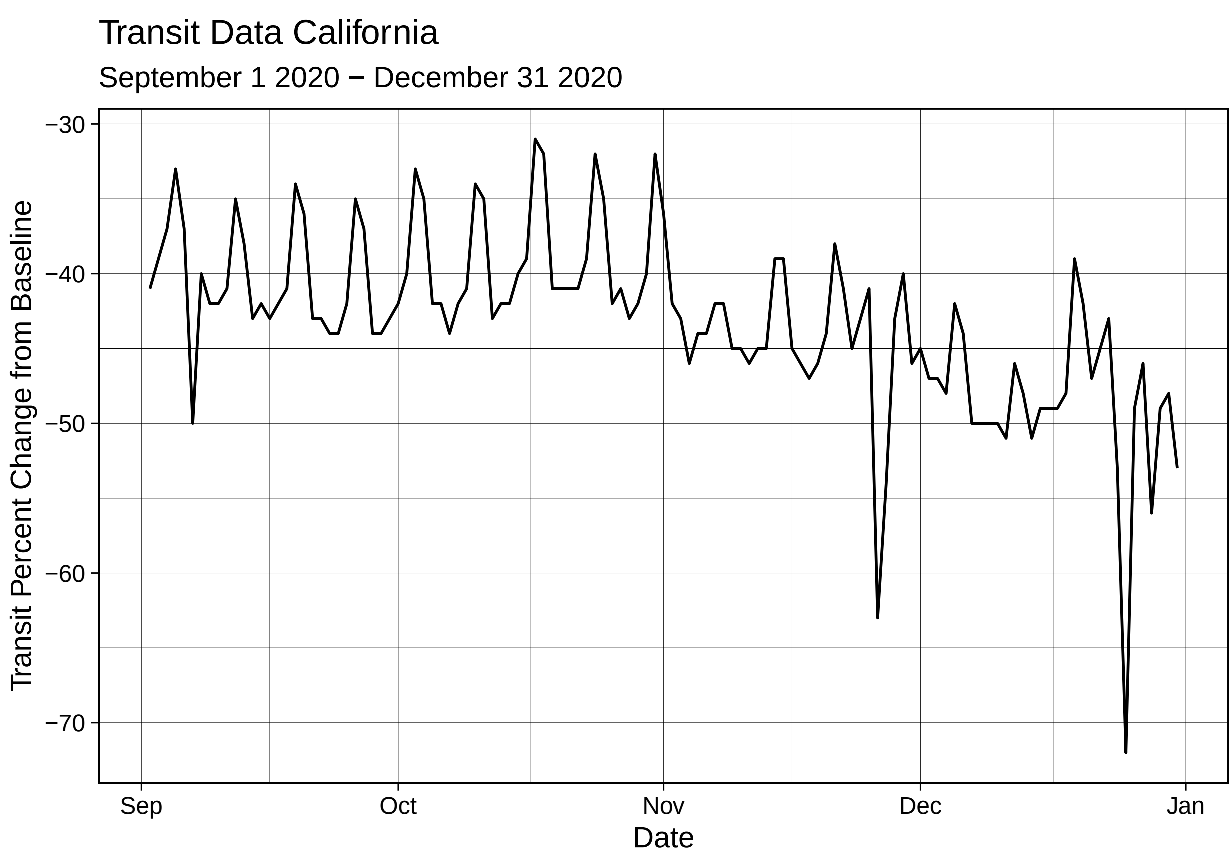
How did COVID-19 play out in California?

- March 4 2020: state of emergency declared
- March 12 2020: cancel large events
- March 19 2020: stay at home orders
- May 8 2020: beginning of phase 2 reopening
- May 26 2020: phase 3 reopening
- June 18 2020: mask mandate put in place
- July 13 2020: 30 counties ordered to close indoor businesses
- July 22 2020: CA surpasses NY for confirmed cases
- October 10 2020: loosen restrictions on private outdoor gatherings
- November 19 2020: statewide curfew
- December 3 2020: new stay at home oder
- January 25 2021: no counties have stay at home order

Time Period Selection

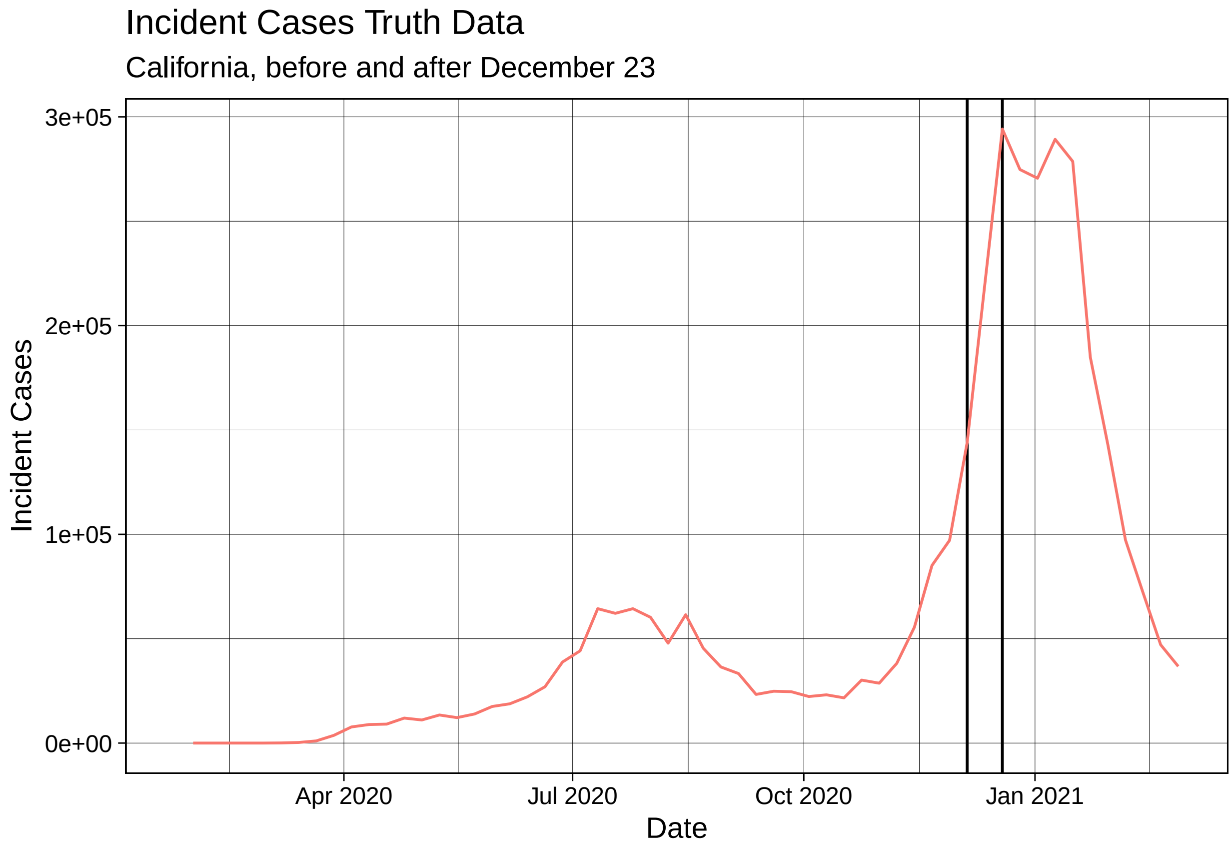
I decided to focus on the dates one month before and one month after December 5 2020 - December 19 2020. It is important to note that UVA-ensemble was only submitting forecasts starting three weeks prior to December 5 2020, but I think this will still give us a good idea of what was going on with mobility and non-mobility models.

From December 5 - December 19 2020, the transit mobility percent change in New York was having a slight fluctuation so it seemed like an interesting time period to see whether mobility models were able to use that to their advantage.



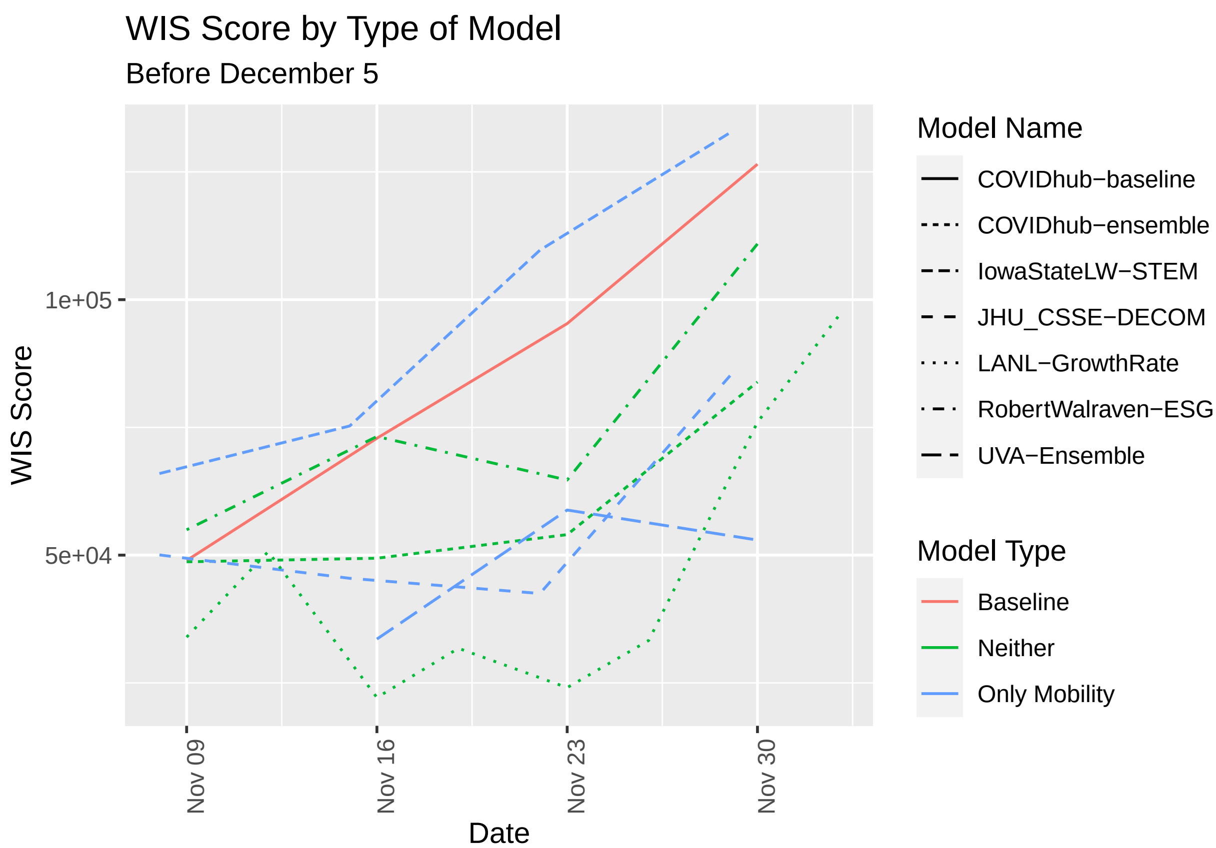
The graph, between Thanksgiving and Christmas, makes a U-shape which is the time period we are looking at. *A quick note: the baseline value is the media transit mobility value from January 3 - February 6 2020.*

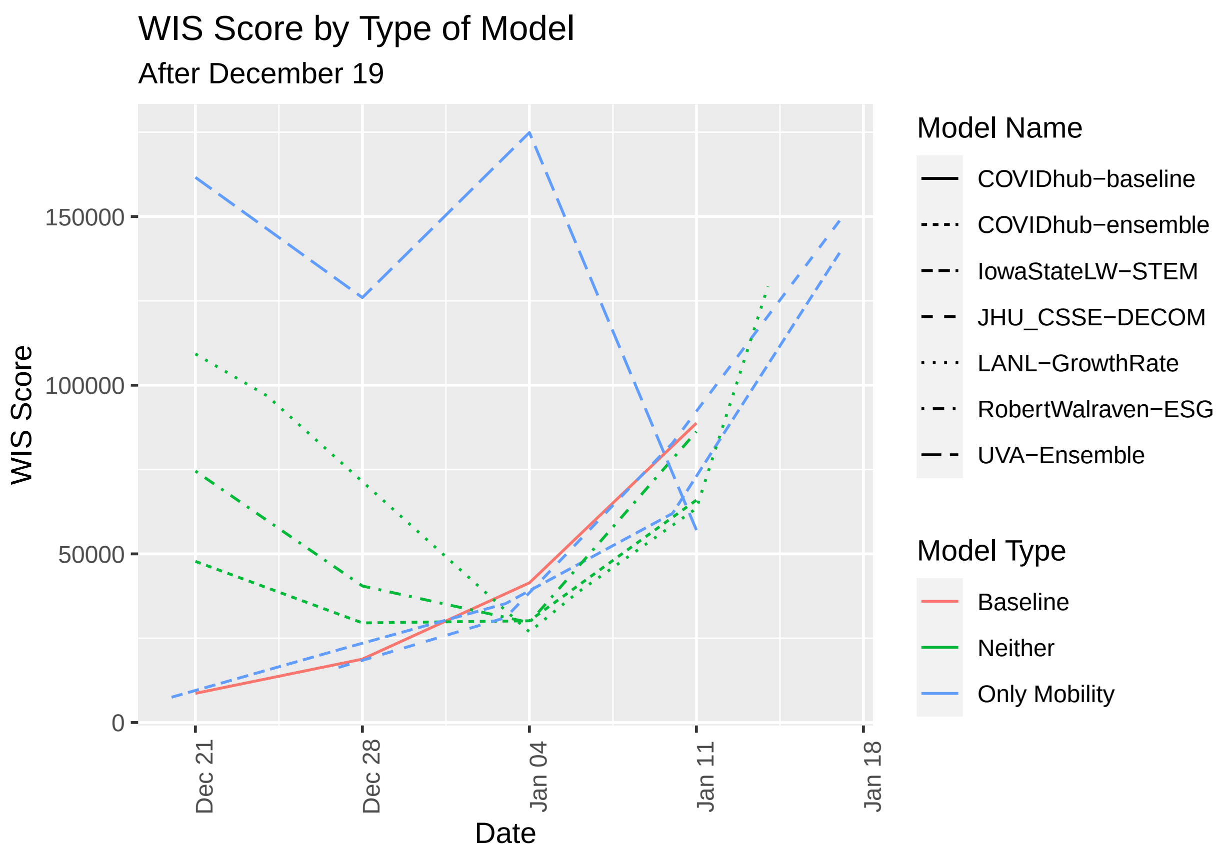
During the time period between December 5 and December 19, cases in California were increasing. Prior to December 5, cases were increasing; however, after December 19, cases began to decrease.



Results

I decided to take a look at the Weighted Interval Score (WIS) for each model before and after the time period selected.


WIS Score by Type of Model
Before December 5
WIS Score
1e+05
5e+04
Nov 09
Nov 16
Nov 23
Nov 30
Date
Model Name
COVIDhub−baseline
COVIDhub−ensemble
IowaStateLW−STEM
JHU_CSSE−DECOM
LANL−GrowthRate
RobertWalraven−ESG
UVA−Ensemble
Model Type
Baseline
Neither
Only Mobility



Looking at these graphs, we can see that mobility models had worse WIS after December 19 while models without mobility data had consistent WIS both before and after the selected time period.

I also decided to look at the average and relative WIS for each model and model type.

```
## # A tibble: 7 x 3
## # Groups:   model [7]
##   model              location wis_before
##   <chr>              <chr>         <dbl>
## 1 COVIDhub-baseline  06           85913.
## 2 COVIDhub-ensemble  06           58996.
## 3 IowaStateLW-STEM   06           95910.
## 4 JHU_CSSE-DECOM     06           55792.
## 5 LANL-GrowthRate    06           46102.
## 6 RobertWalraven-ESG 06           75946.
## 7 UVA-Ensemble       06           48450.
```

```
## # A tibble: 7 x 3
## # Groups:   model [7]
##   model              location wis_after
##   <chr>              <chr>        <dbl>
## 1 COVIDhub-baseline  06          39404.
## 2 COVIDhub-ensemble  06          43354.
## 3 IowaStateLW-STEM   06          53111.
## 4 JHU_CSSE-DECOM     06          69761.
## 5 LANL-GrowthRate    06          77375.
```

```
## 6 RobertWalraven-ESG 06          57724.
## 7 UVA-Ensemble       06         129886.
```

```
## # A tibble: 1 x 1
##   wis_before_mob
##            <dbl>
## 1         73204.
```

```
## # A tibble: 1 x 1
##   wis_after_mob
##           <dbl>
## 1        93315.
```

```
## # A tibble: 1 x 1
##   wis_before_neither
##                <dbl>
## 1             52854.
```

```
## # A tibble: 1 x 1
##   wis_after_neither
##               <dbl>
## 1            59625.
```

```
##   relwis_baseline
## 1       0.4586502
```

```
##   relwis_mob_neither_before
## 1                  1.385035
```

```
##   relwis_mob_neither_after
## 1                 1.565023
```

Looking at the first two figures, we can see that IowaStateLW-STEM's model had a better WIS while JHU_CSSE-DECOM and UVA-Ensemble had worse worse WIS. The non-mobility models got better with the exception of LANL.

Looking at the next four figures, we can see that the average WIS for models with mobility data got worse while average WIS for models without mobility data only got slightly worse. Mobility models went from a WIS of 73k to a WIS of 93k while non-mobility models went from a WIS of 52k to 59k.

Finally, in the last three figures, we can see that since the relative WIS increased after December 19, mobility WIS also increased. We can also see that the baseline performed well and had a relative WIS of less that 1.

New York

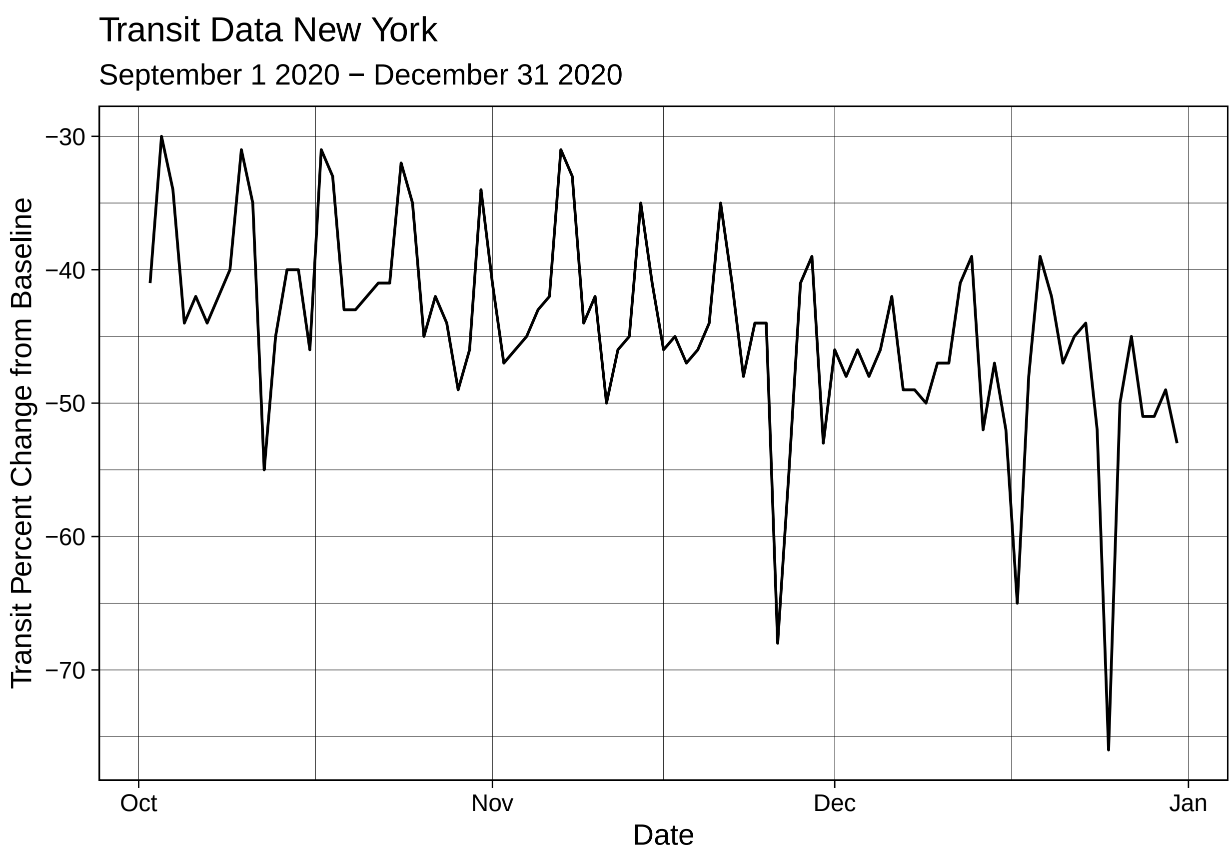
How did COVID-19 play out in New York?

- March 7 2020: state of emergency declared
- April 15 2020: mask mandate put in place
- April-May 2020: stay at home orders
- May 15 2020: slight reopening
- June 8 2020: NYC phase 1 reopening
- June 22 2020: NYC phase 2 reopening
- July 6 2020: phase 3 reopening
- July 19 2020: phase 4 reopening
- November 13 2020: new restrictions
- December 1 2020: slight reopening of schools
- December 11 2020: slight reopening of gyms and salons
- December 23 2020: full reopening of gyms
- February 15 2021: NYC middle schools go back to in person
- March 22 2021: NYC high schools go back to in person

Time Period Selection

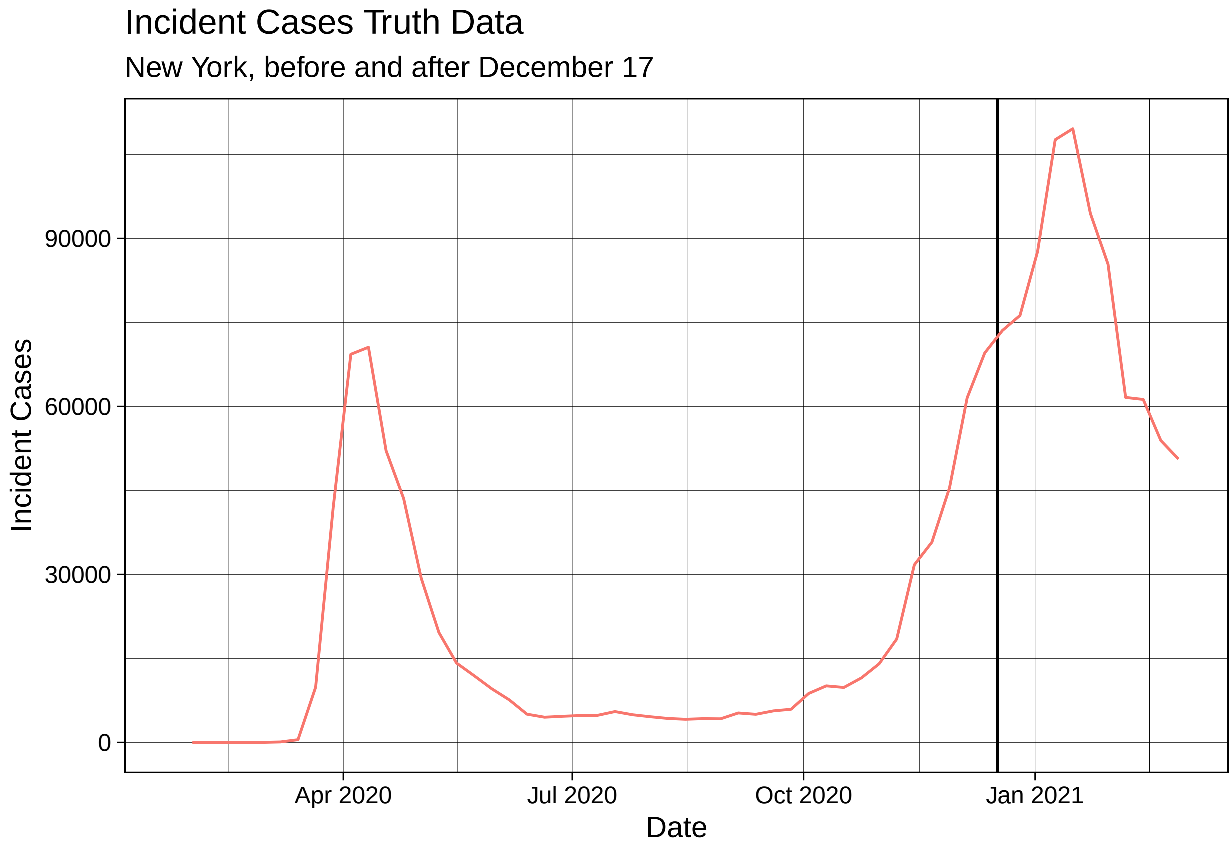
I decided to focus on the dates one month before and one month after December 17 2020.

There was a huge decrease in the transit mobility percent change in California on December 17 2020 so it seemed like an interesting time period to see whether mobility models were able to use that to their advantage.



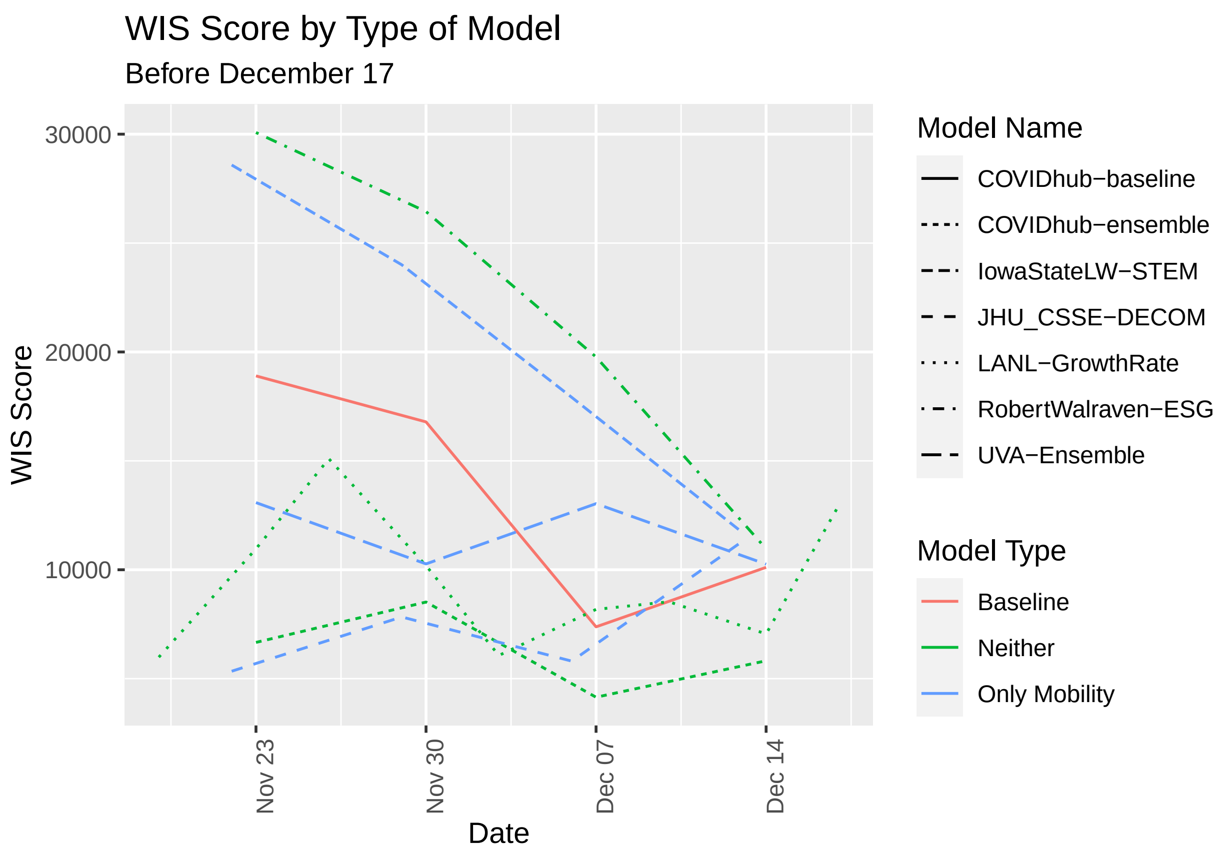
The graph, between Thanksgiving and Christmas, has a decrease to about -65% which is the date we are looking at. *A quick note: the baseline value is the media transit mobility value from January 3 - February 6 2020.*

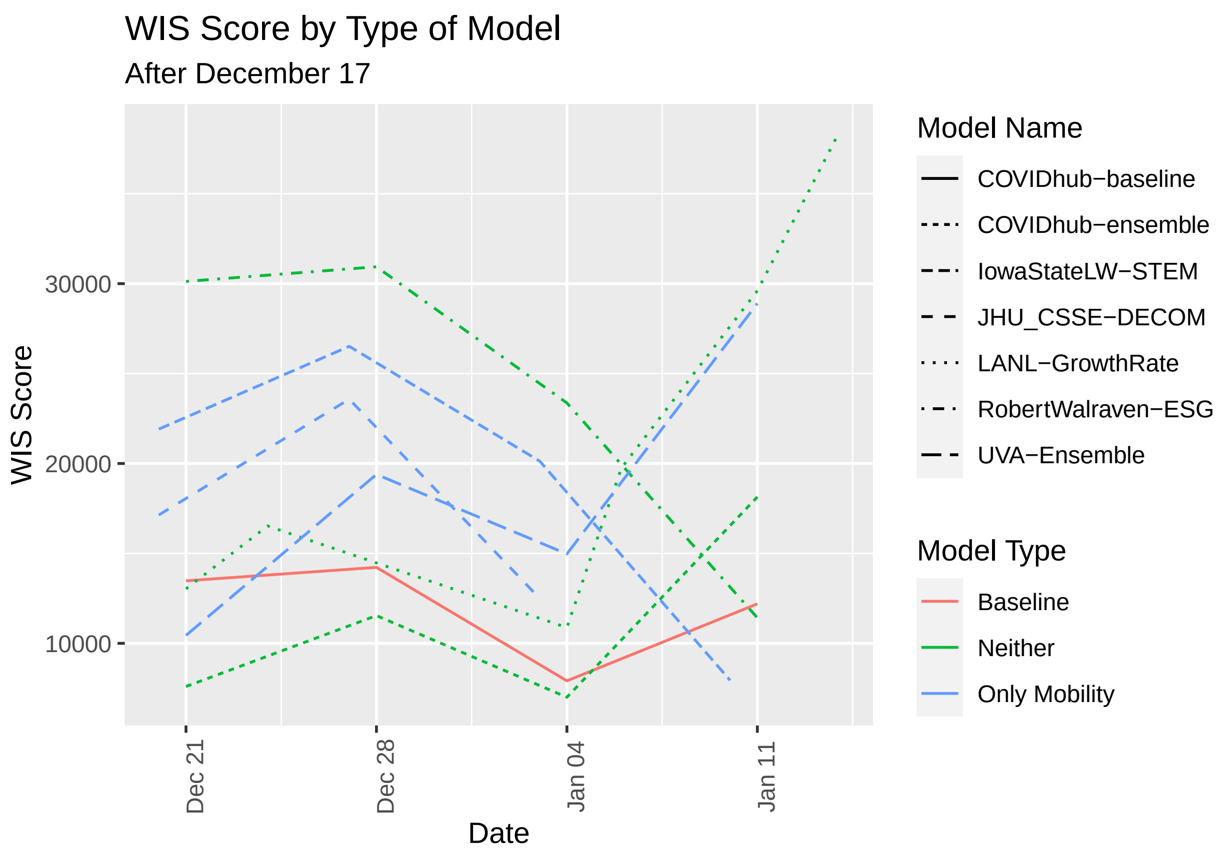
Cases in New York were increasing before and after December 17 2020:



Results

I decided to take a look at the Weighted Interval Score (WIS) for each model before and after the time period selected.


WIS Score by Type of Model
Before December 17
WIS Score
30000
20000
10000
Nov 23
Nov 30
Dec 07
Dec 14
Date
Model Name
COVIDhub−baseline
COVIDhub−ensemble
IowaStateLW−STEM
JHU_CSSE−DECOM
LANL−GrowthRate
RobertWalraven−ESG
UVA−Ensemble
Model Type
Baseline
Neither
Only Mobility



Looking at these graphs, we can see that mobility models got a little bit better after December 17 with the WIS being below 3000. Models without mobility data only got a little better with LANL getting much worse.

I also decided to look at the average and relative WIS for each model and model type.

```
## # A tibble: 7 x 3
## # Groups:   model [7]
##   model              location wis_before
##   <chr>              <chr>         <dbl>
## 1 COVIDhub-baseline  36           13296.
## 2 COVIDhub-ensemble  36            6288.
## 3 IowaStateLW-STEM   36           20560.
## 4 JHU_CSSE-DECOM     36            7566.
## 5 LANL-GrowthRate    36            9448.
## 6 RobertWalraven-ESG 36           21813.
## 7 UVA-Ensemble       36           11661.
```

```
## # A tibble: 7 x 3
## # Groups:   model [7]
##   model              location wis_after
##   <chr>              <chr>        <dbl>
## 1 COVIDhub-baseline  36          11953.
## 2 COVIDhub-ensemble  36          11072.
## 3 IowaStateLW-STEM   36          19125.
## 4 JHU_CSSE-DECOM     36          17736.
```

```
## 5 LANL-GrowthRate    36          21364.
## 6 RobertWalraven-ESG 36          23962.
## 7 UVA-Ensemble       36          18427.

## # A tibble: 1 x 1
##   wis_before_mob
##            <dbl>
## 1         11883.

## # A tibble: 1 x 1
##   wis_after_mob
##           <dbl>
## 1        19604.

## # A tibble: 1 x 1
##   wis_before_neither
##                <dbl>
## 1             12160.

## # A tibble: 1 x 1
##   wis_after_neither
##               <dbl>
## 1            20118.

##   relwis_baseline
## 1       0.8990317

##   relwis_mob_neither_before
## 1                 0.9771831

##   relwis_mob_neither_after
## 1                0.9744733
```

Looking at the first two figures, we can see that IowaStateLW-STEM's model had a better WIS while JHU_CSSE-DECOM and UVA-Ensemble had worse worse WIS after December 17. All of T=the non-mobility models got worse after.

Looking at the next four figures, we can see that the average WIS for models with mobility data got slightly worse going from 11k to 19k while models without mobility data got much worse going from 8k to 20k.

Finally, in the last three figures, we can see that since the relative WIS decreased after December 17 which means that mobility model WIS decreased as well and performed better than non-mobility models. We can also see that the baseline, again, performed well.

Discussion

If models with mobility data performed better, we would have expected the relative WIS after the specified dates to be less than 1. In California, the relative WIS after December 19 was greater than 1 and greater than the relative WIS before December 5. In New York, the relative WIS after December 17 was less than 1.

Models with mobility data did not perform better than models without mobility data in California, but they did perform better in New York.

From here, we can either conclude that models with mobility data are not more or less likely to perform better than models without mobility data or we can say that we need to look into more locations and dates to see more variations before making a definitive conclusion.