The Dataset

For these analyses we used two population datasets from the US Census Bureau. The data was based on the US Census in 2010 and an estimation of the years from 2010-2018. We chose to do analyses focusing on changes in data patterns over time. We also used a dataset of counties provided by the maps R package to visualize data by county on a map of the US.

We chose to focus the INTERNATIONALMIG20XX columns and the POPESTIMATE20XX columns from the census dataset. We utilized these columns because they provided the most relevant information for the future. For the grouping of data, the CBSA and LSAD columns allowed for clustering of data by region/county.

Analysis 1:

Forecasting Population by Areas

Hypothesis

With so much planning effort by governments and businesses tied to the population numbers, it is crucial that we can forecast population changes. It was hypothesized that we could somewhat reliably forecast future populations exclusively using data points from past population estimates, using a model that can be repeatedly applied to different areas.

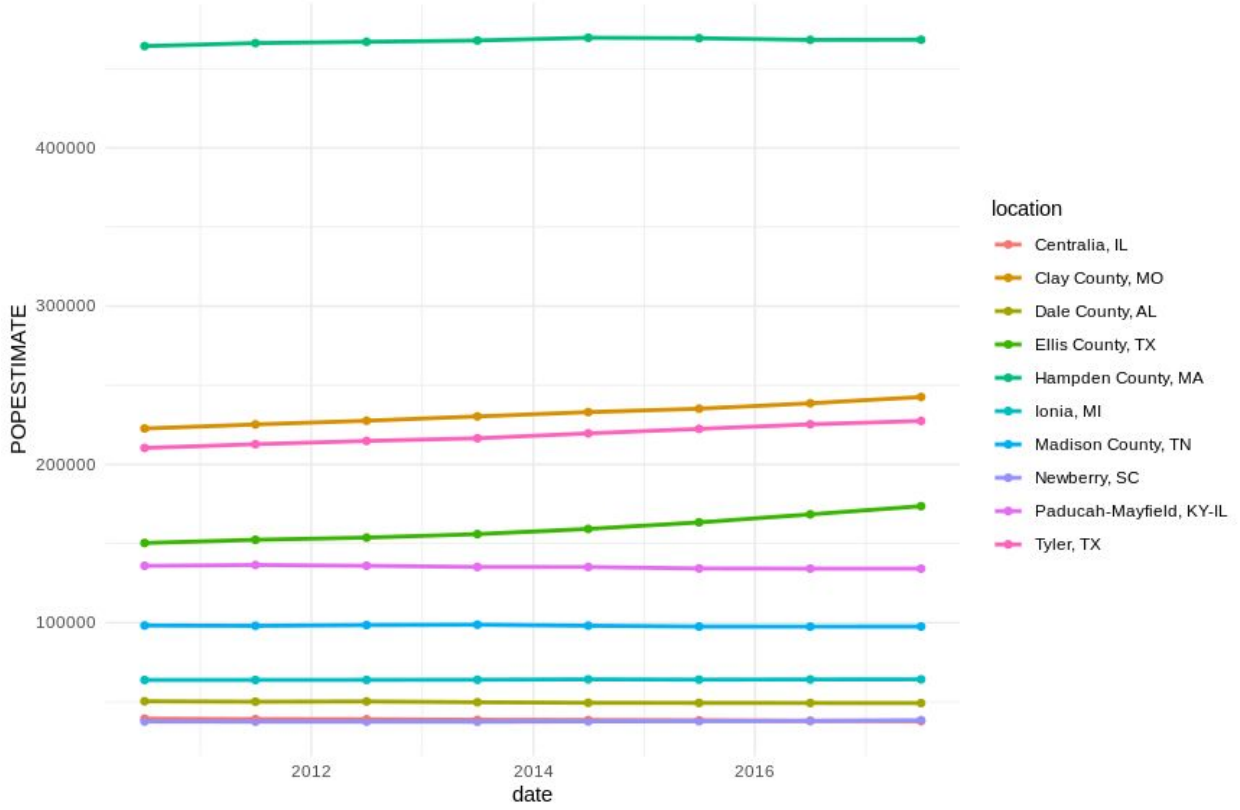
Finding

We utilized Holt's Additive Method, which is exponential smoothing with a linear trend, to forecast populations in the last 3 years of the dataset using data points from the first 5 years. We were able to achieve average mean absolute percentage error of 0.77% and average mean percentage error of 0.18% in forecasting.

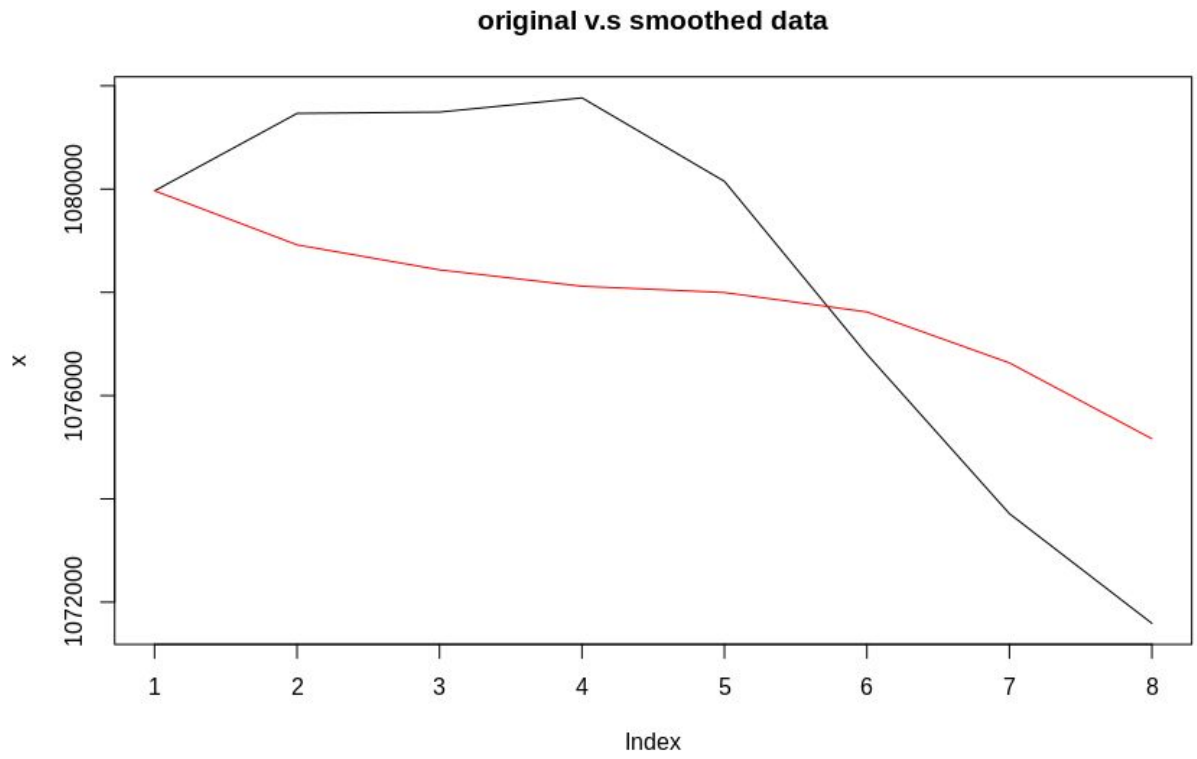
Investigation

The first thing we did was to visualize the population estimates over the years from a few locations. As seen below, the plot of populations from 20 random locations over the years shows us that there is great variance in populations and how they change. It also displays autocorrelation of trends and data points over time within locations, which suggest that we might

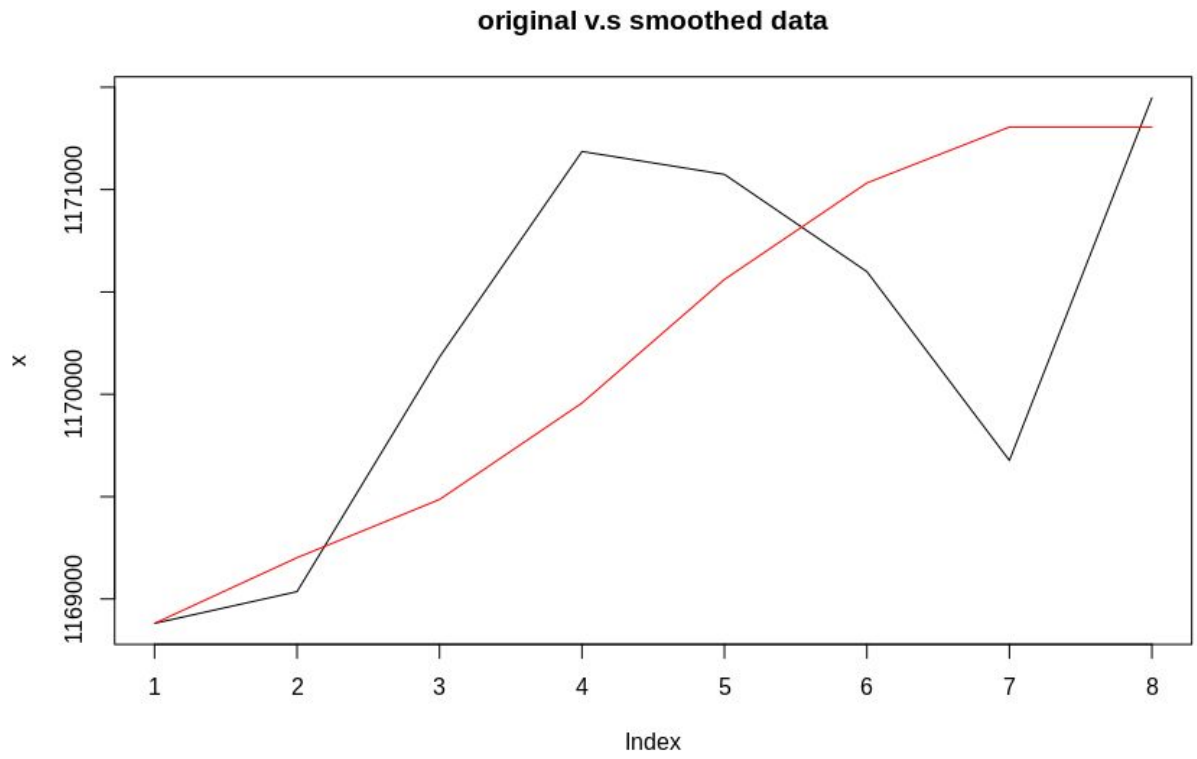
seek to use an autoregressive model.



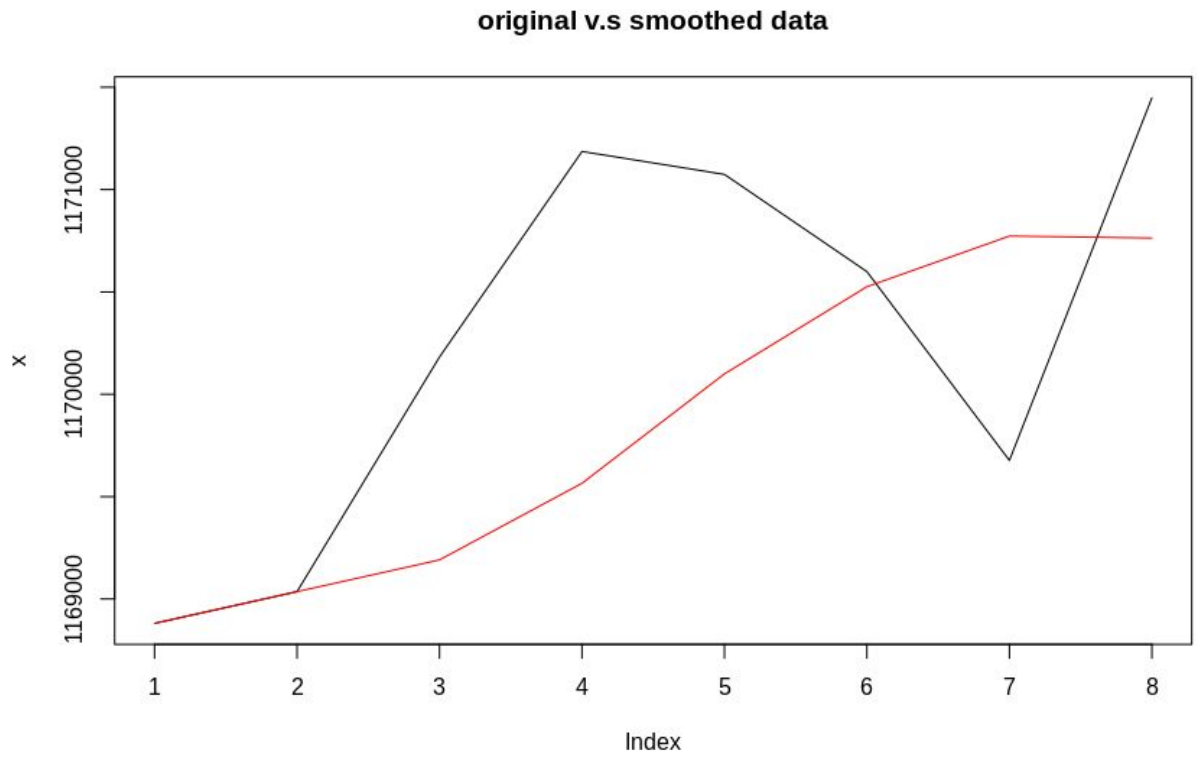
There are a couple models that we can use to fit and forecast such data, that exhibits linear or possible exponential relationship between its starting populations and future populations over time. We could use simple linear regressions using the model $y(t)=y(0)+bt$, or exponential regressions using the model $y=ab^t$, but they do not account for autocorrelations. We could use simple exponential smoothing, but they do not account for trends over time periods. We could try Holt-Winters, but the data displays no seasonality. To avoid using limited time on models that are not promising, we decided to go with Holt's method, which is exponential smoothing with linear or exponential trends over time. We were able to find a package in R that could carry out such smoothing and forecasting task using Holt's method (https://cran.r-project.org/web/packages/aTSA/aTSA.pdf). We fitted the smoothing model onto data of a few locations, and it seemed to be quite promising. Shown below are a few examples.



Holt's Linear Model on Rochester, NY

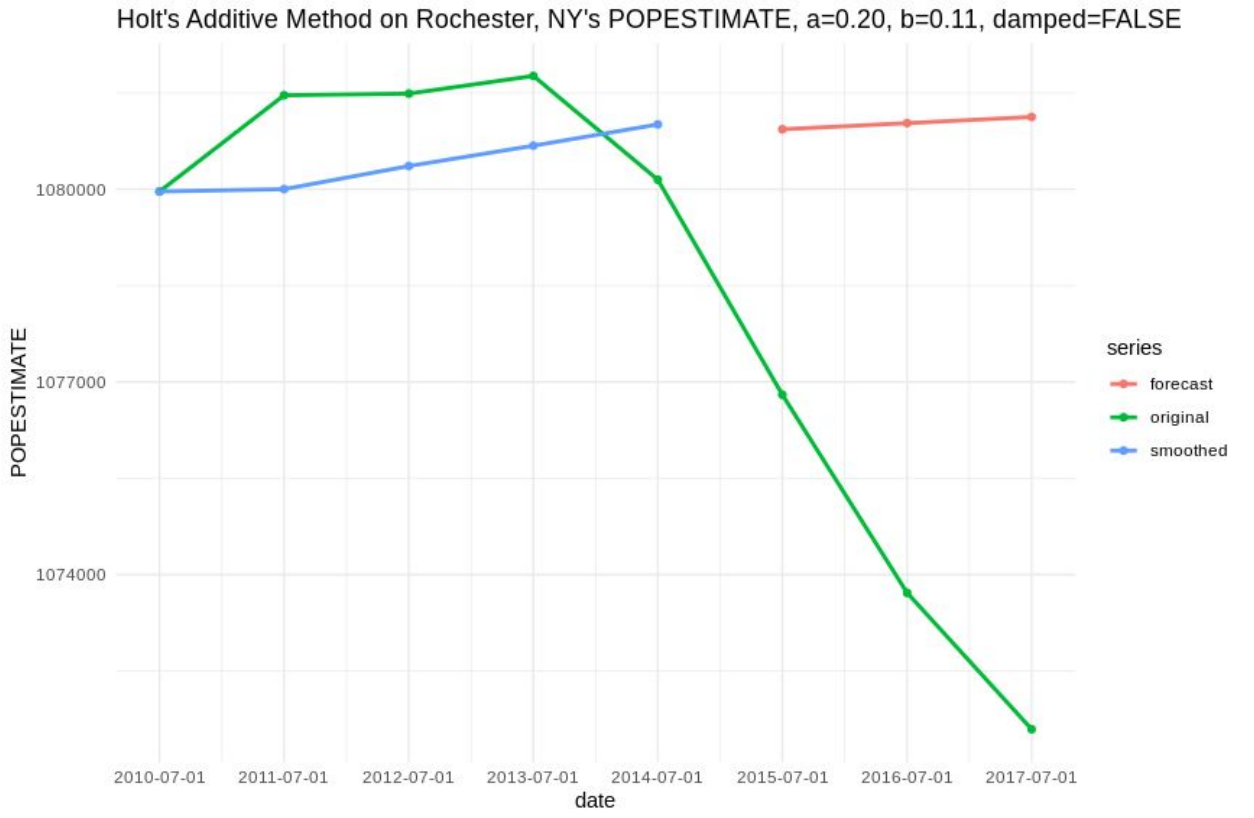


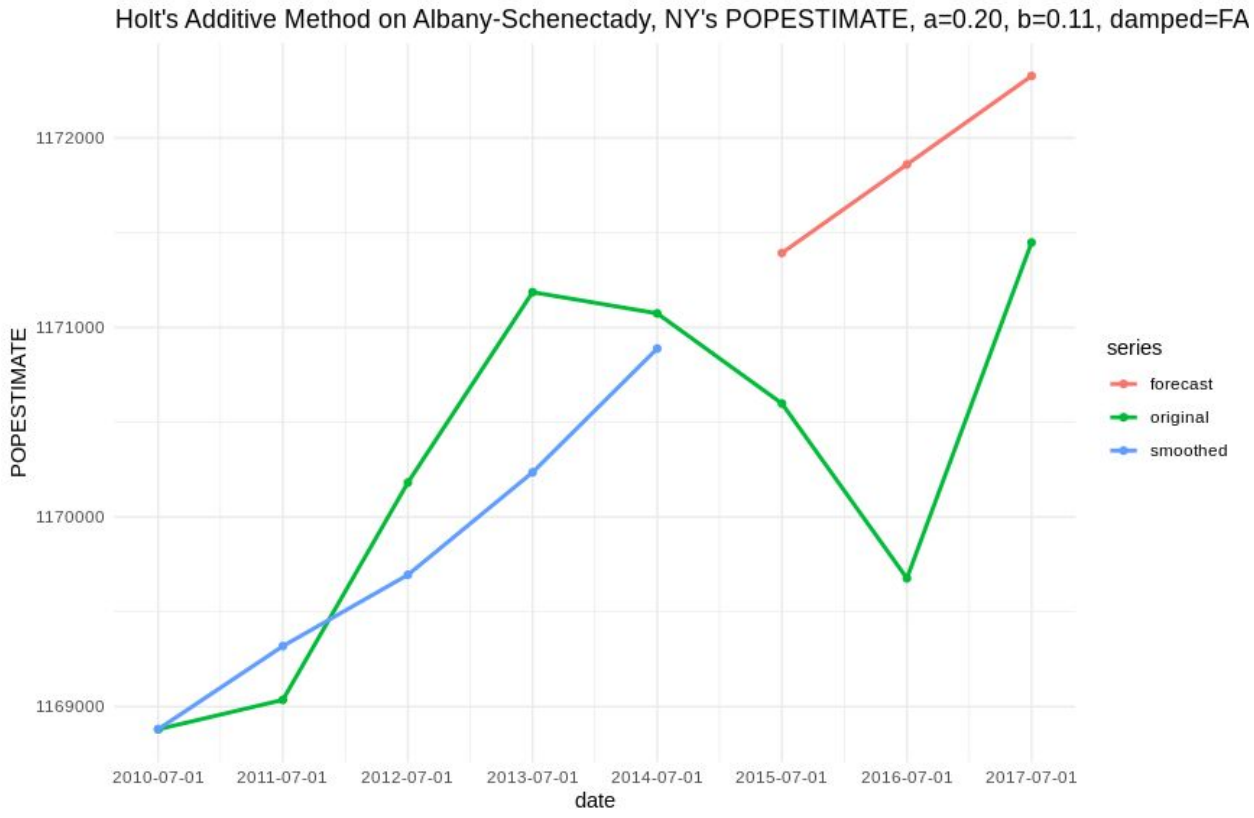
Holt's Linear Model on Albany-Schenectady, NY

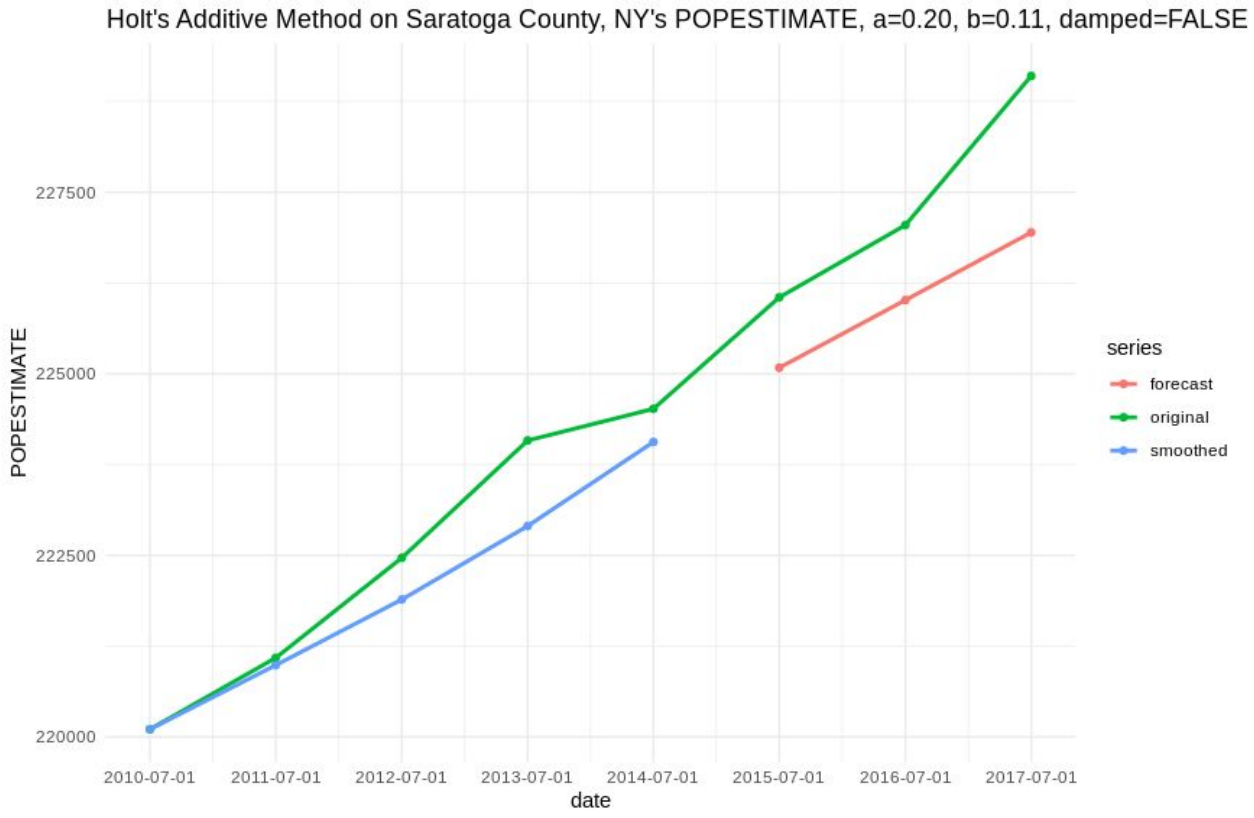


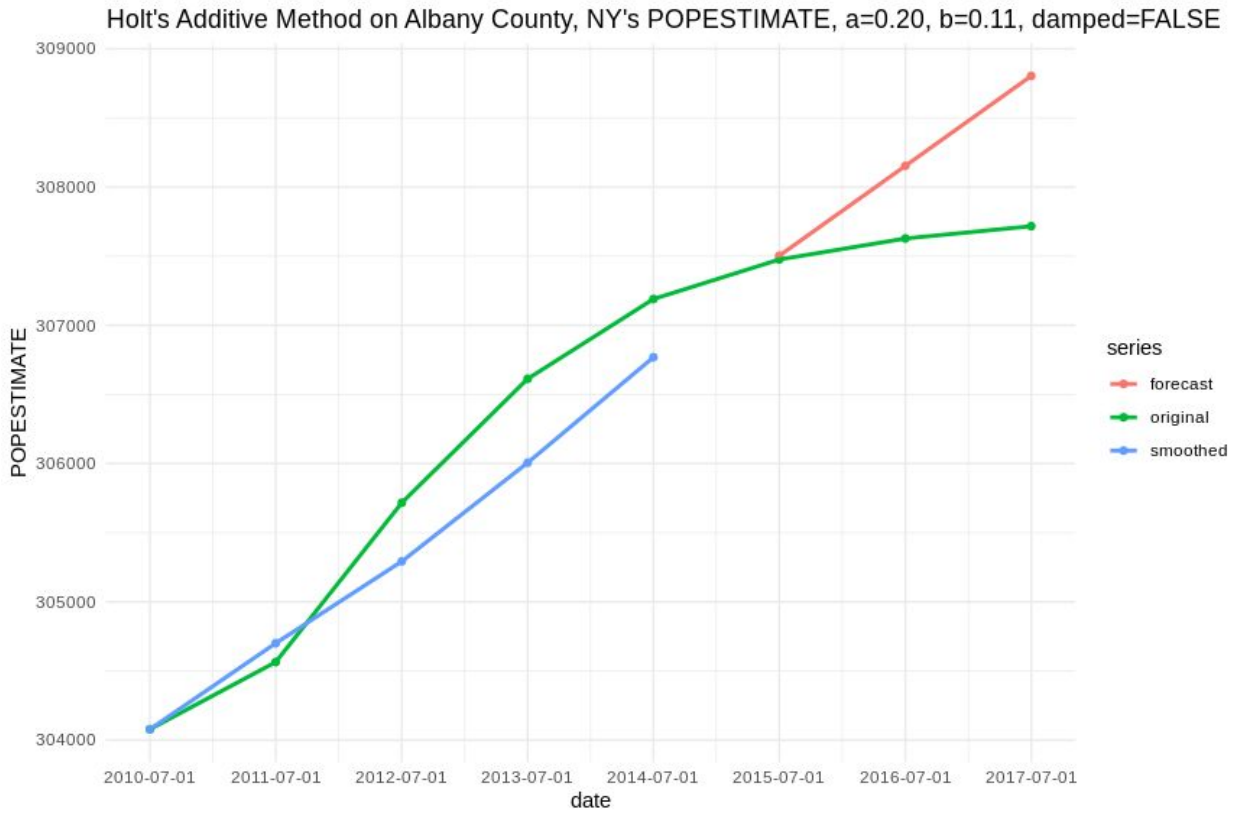
Holt's Exponential Model on Albany-Schenectady, NY

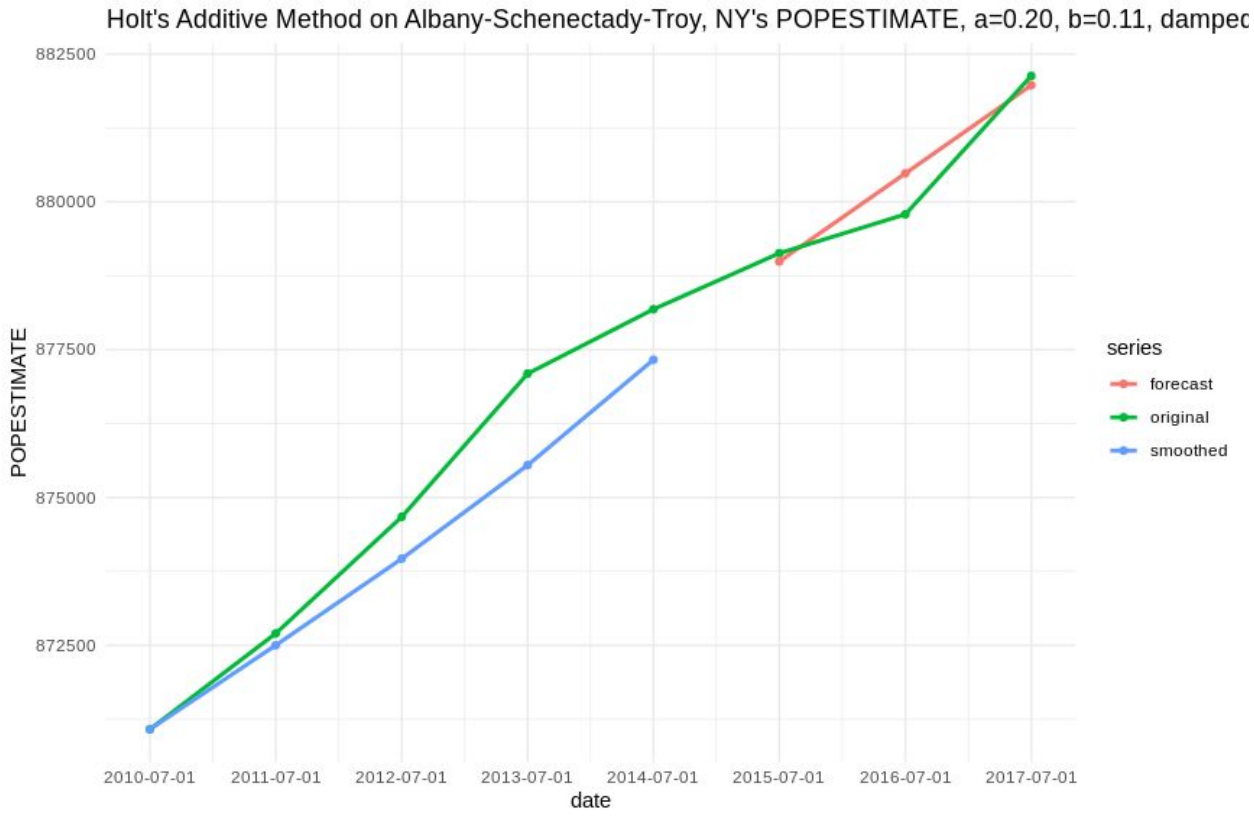
The pictures seem to confirm the validity of an autoregressive model. We also see that the default parameters for the models might be a little slow to pick up the changes over time, and might need higher values for the α and β parameters, to give more weight to the most recent updates and trends. On top of that, the default configurations did not include a forecasting period, and used all the data for smoothing and fitting the model. We therefore split the datasets into 2010-2014 and 2015-2017 to gain an understanding of the robustness of the model when used for forecasts. The results are visualized below for a selection of locations, as well as the aggregated error metrics in percentages.


Holt's Additive Method on Rochester, NY's POPESTIMATE, a=0.20, b=0.11, damped=FALSE
POPESTIMATE
1080000
1077000
1074000
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed

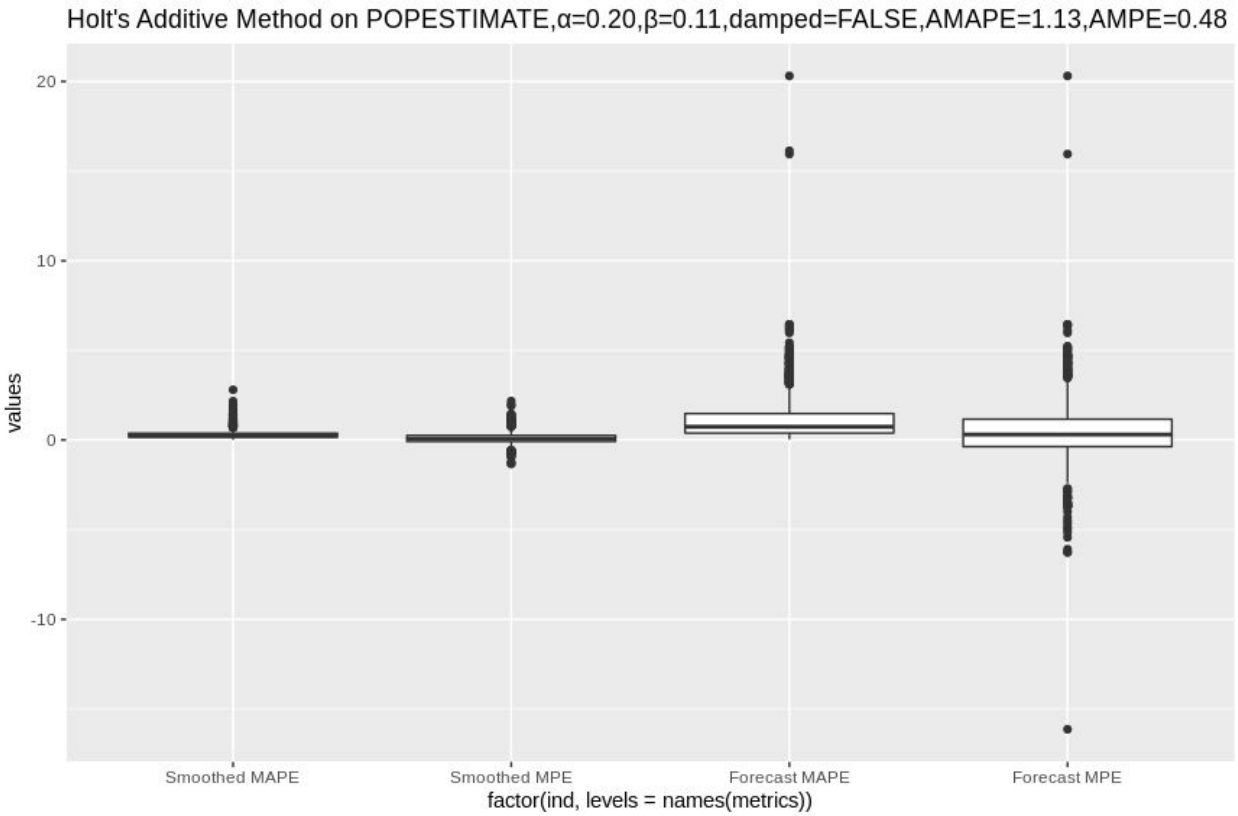

Holt's Additive Method on Albany-Schenectady, NY's POPESTIMATE, a=0.20, b=0.11, damped=FA
POPESTIMATE
1172000
1171000
1170000
1169000
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed


Holt's Additive Method on Saratoga County, NY's POPESTIMATE, a=0.20, b=0.11, damped=FALSE
POPESTIMATE
227500
225000
222500
220000
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed

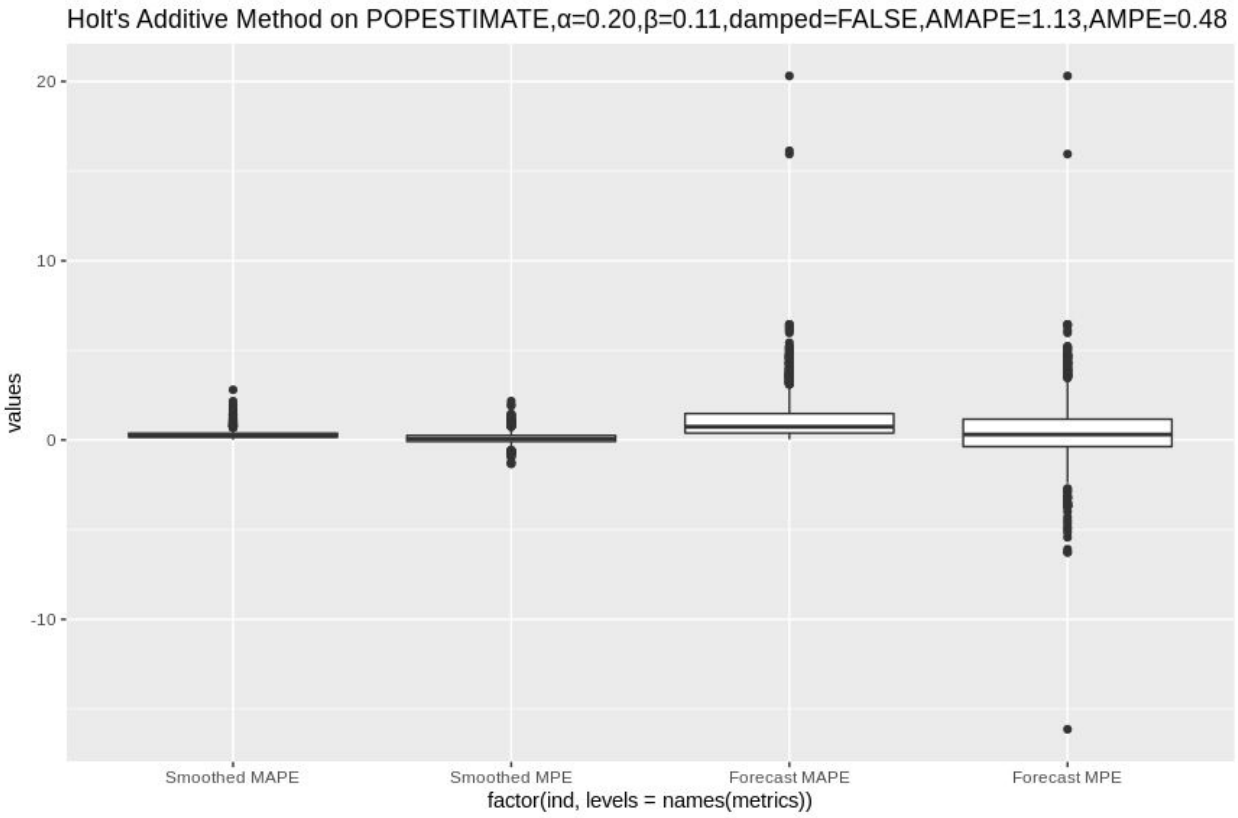

Holt's Additive Method on Albany County, NY's POPESTIMATE, a=0.20, b=0.11, damped=FALSE
POPESTIMATE
309000
308000
307000
306000
305000
304000
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed

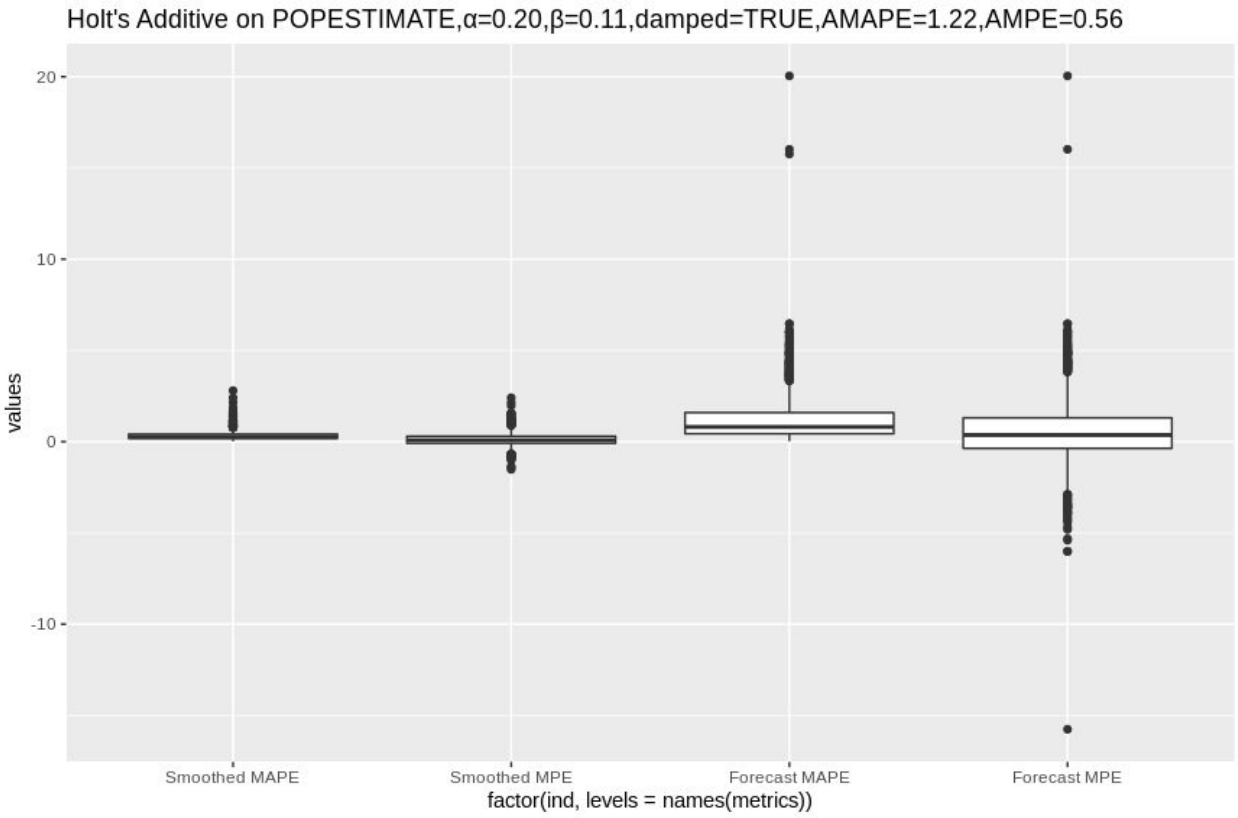


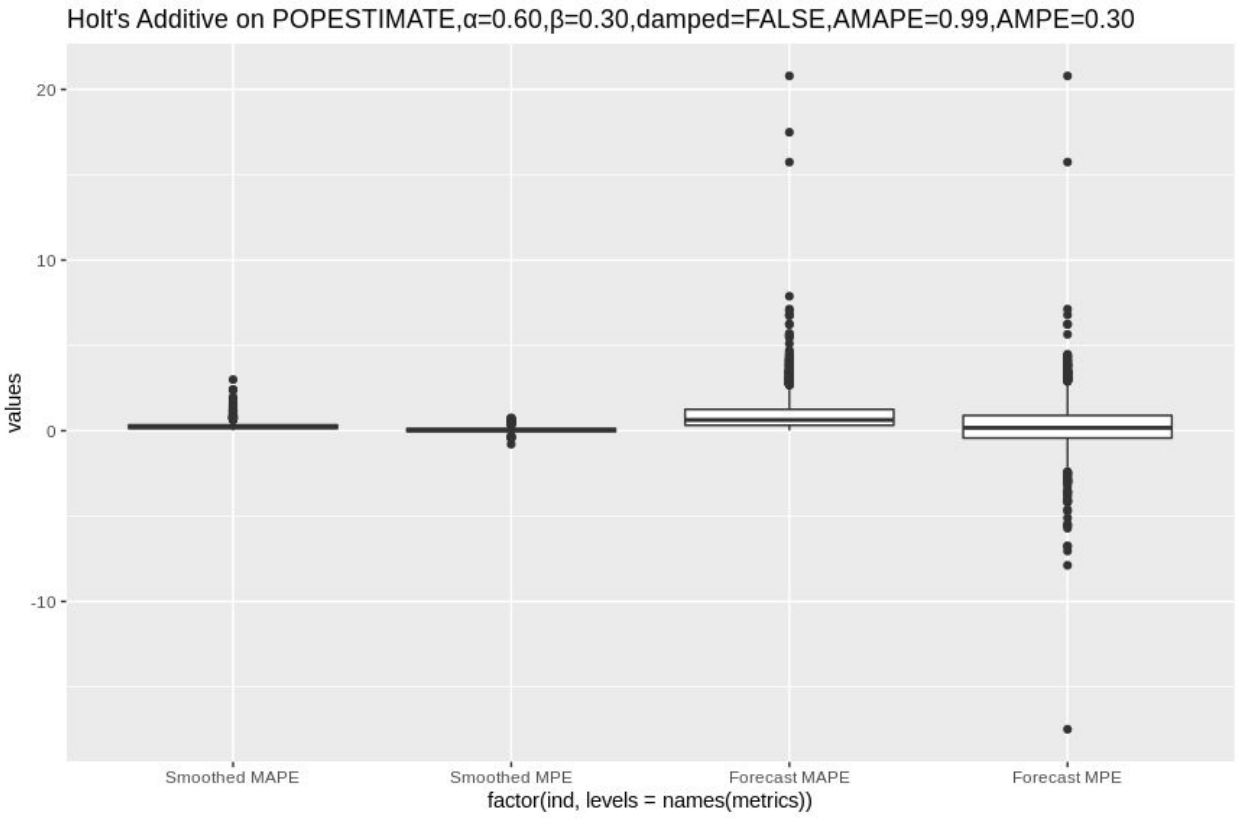
We can once again see from the plots a need for higher α and β in order to reduce the lag in forecast and smoothed trends. Not displayed are forecasts made with Holt's multiplicative method, because it was discovered that the package was not programmed correctly to forecast the exponential trend over time (it would produce a flat line as if done using simple exponential smoothing). We tried a few more combinations of α and β, and visualized the distribution of smoothing and forecasting errors using boxplots.

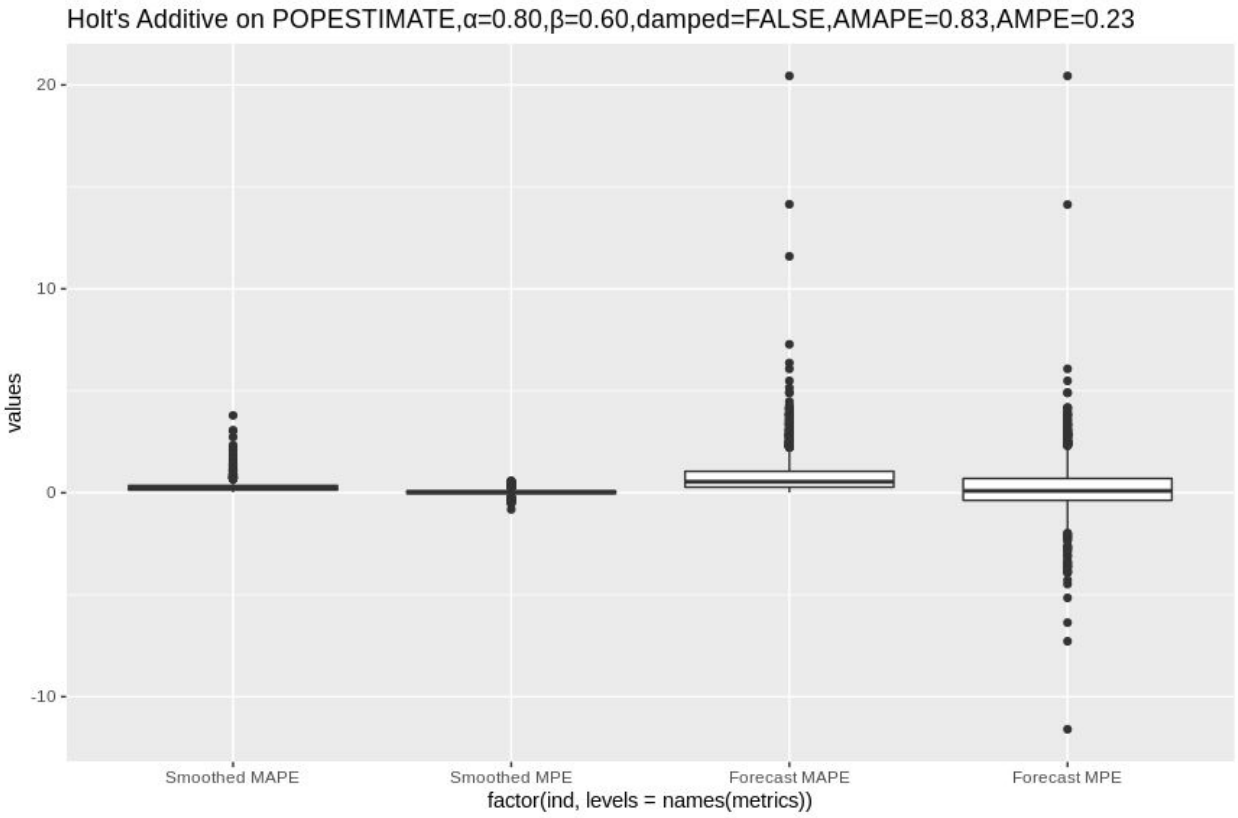


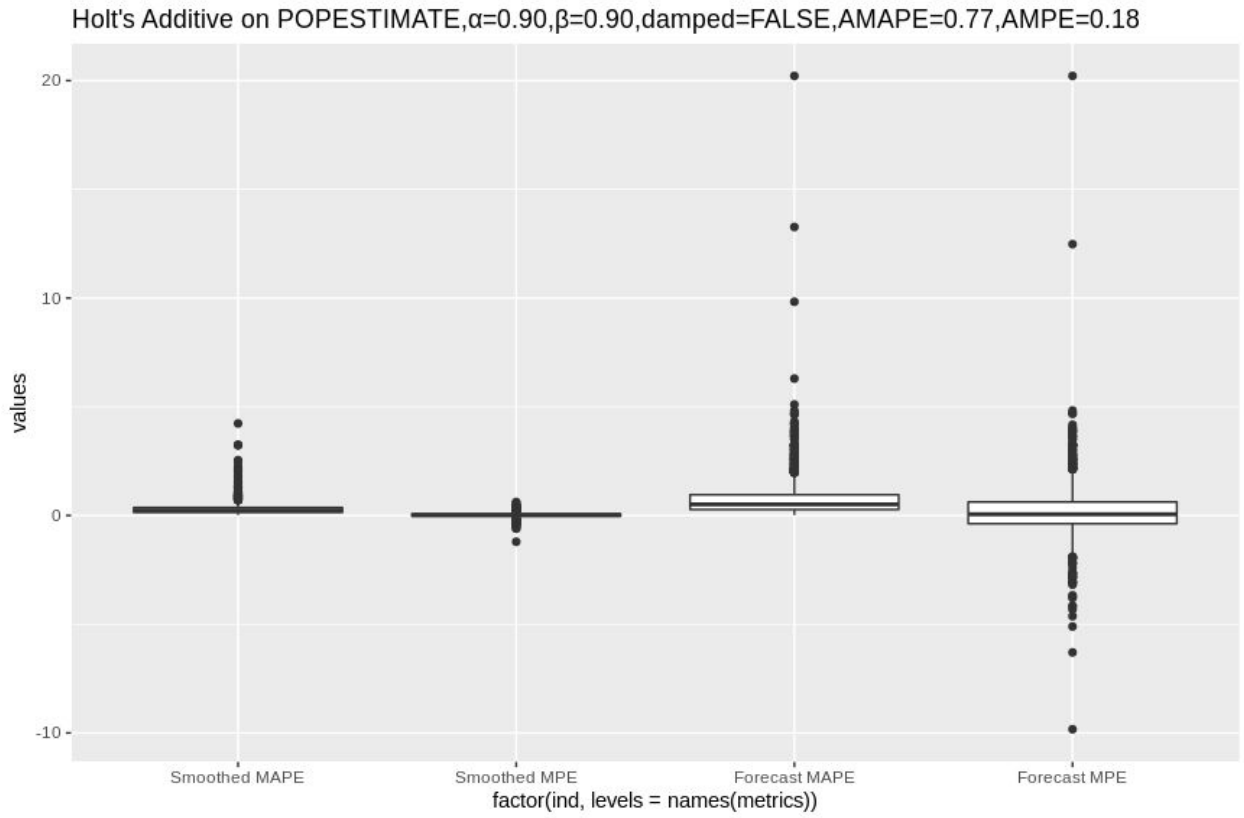
This is the boxplot of Mean Absolute Percentage Error (MAPE) and Mean Percentage Error(MPE). We can see that unsurprisingly, the forecasting errors are greater than the smoothing errors, and that MPEs are mostly positive for forecasting, which means that the model tends to overestimate future populations. We then seek to improve upon this benchmark by drawing from our earlier conclusion that α and β need to be higher to improve model sensitivity. Shown below are forecasting errors of several combinations of α and β.

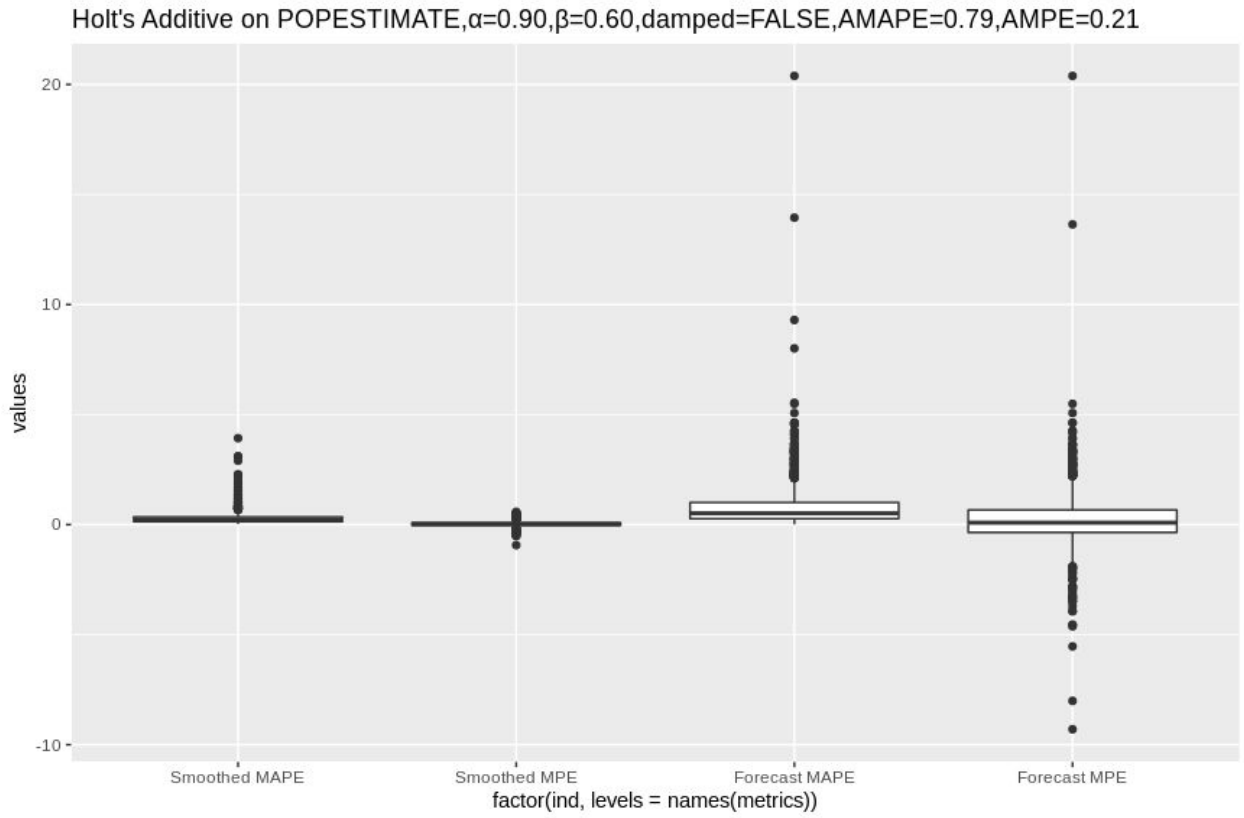

Holt's Additive Method on POPESTIMATE,α=0.20,β=0.11,damped=FALSE,AMAPE=1.13,AMPE=0.48
20
10
0
-10
values
Smoothed MAPE
Smoothed MPE
Forecast MAPE
Forecast MPE
factor(ind, levels = names(metrics))

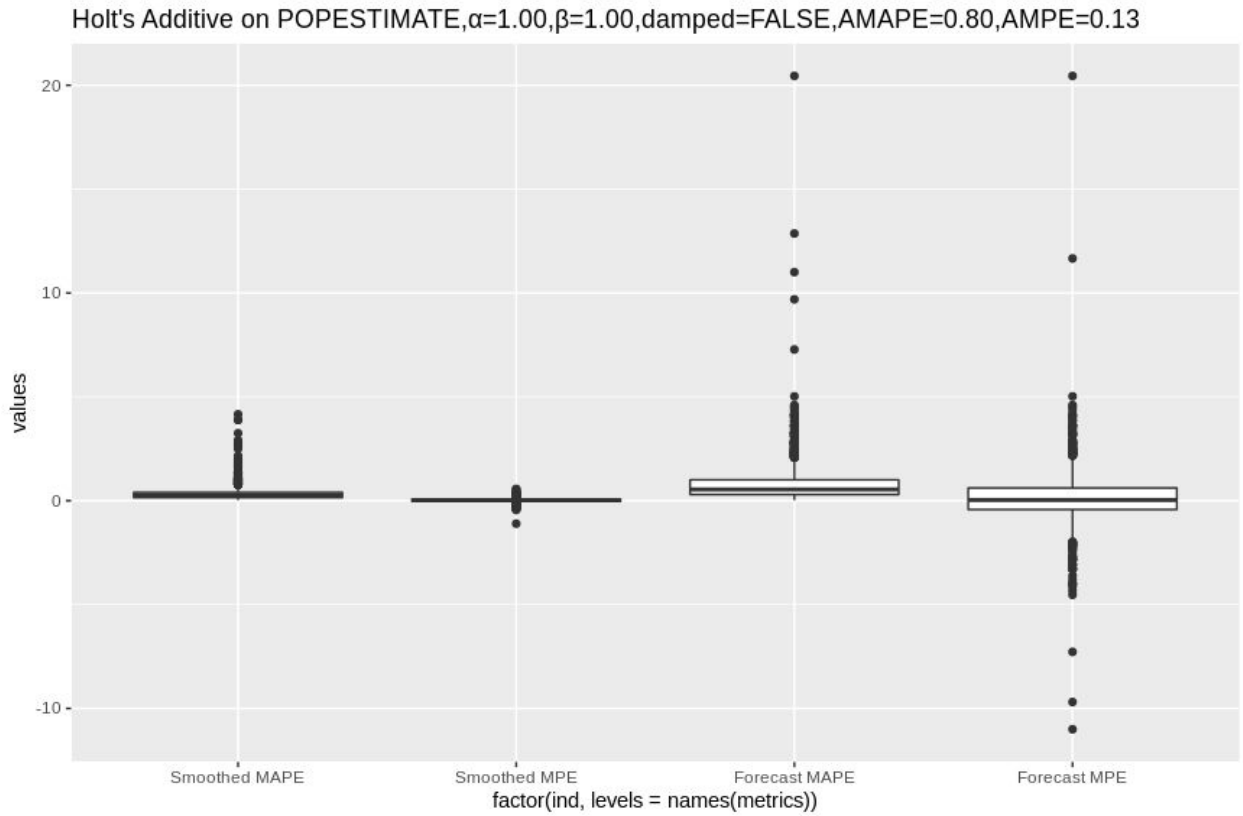

Holt's Additive on POPESTIMATE,α=0.20,β=0.11,damped=TRUE,AMAPE=1.22,AMPE=0.56
20
10
0
-10
values
Smoothed MAPE
Smoothed MPE
Forecast MAPE
Forecast MPE
factor(ind, levels = names(metrics))


Holt's Additive on POPESTIMATE,α=0.60,β=0.30,damped=FALSE,AMAPE=0.99,AMPE=0.30
20
10
0
-10
values
Smoothed MAPE
Smoothed MPE
Forecast MAPE
Forecast MPE
factor(ind, levels = names(metrics))

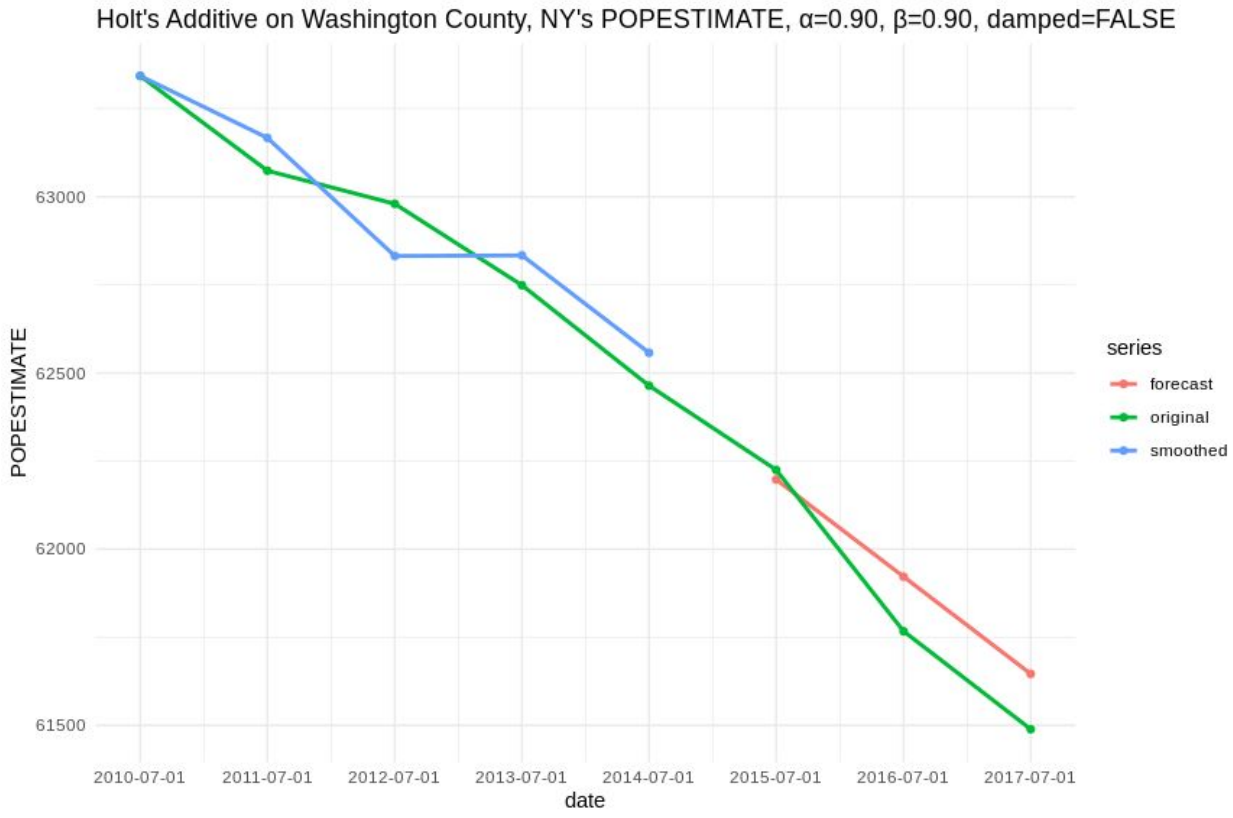

Holt's Additive on POPESTIMATE,α=0.80,β=0.60,damped=FALSE,AMAPE=0.83,AMPE=0.23
20
10
0
-10
values
Smoothed MAPE
Smoothed MPE
Forecast MAPE
Forecast MPE
factor(ind, levels = names(metrics))

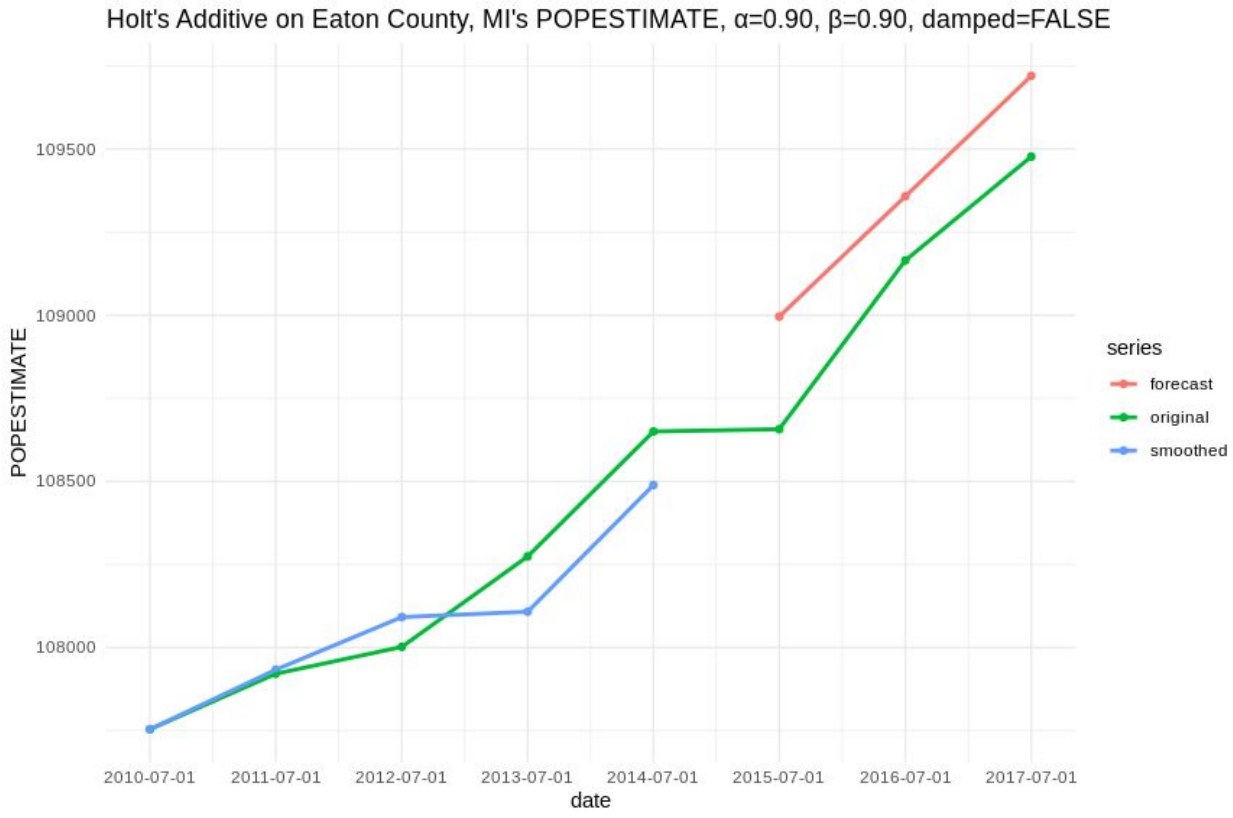

Holt's Additive on POPESTIMATE,α=0.90,β=0.90,damped=FALSE,AMAPE=0.77,AMPE=0.18
20
10
0
-10
values
Smoothed MAPE
Smoothed MPE
Forecast MAPE
Forecast MPE
factor(ind, levels = names(metrics))

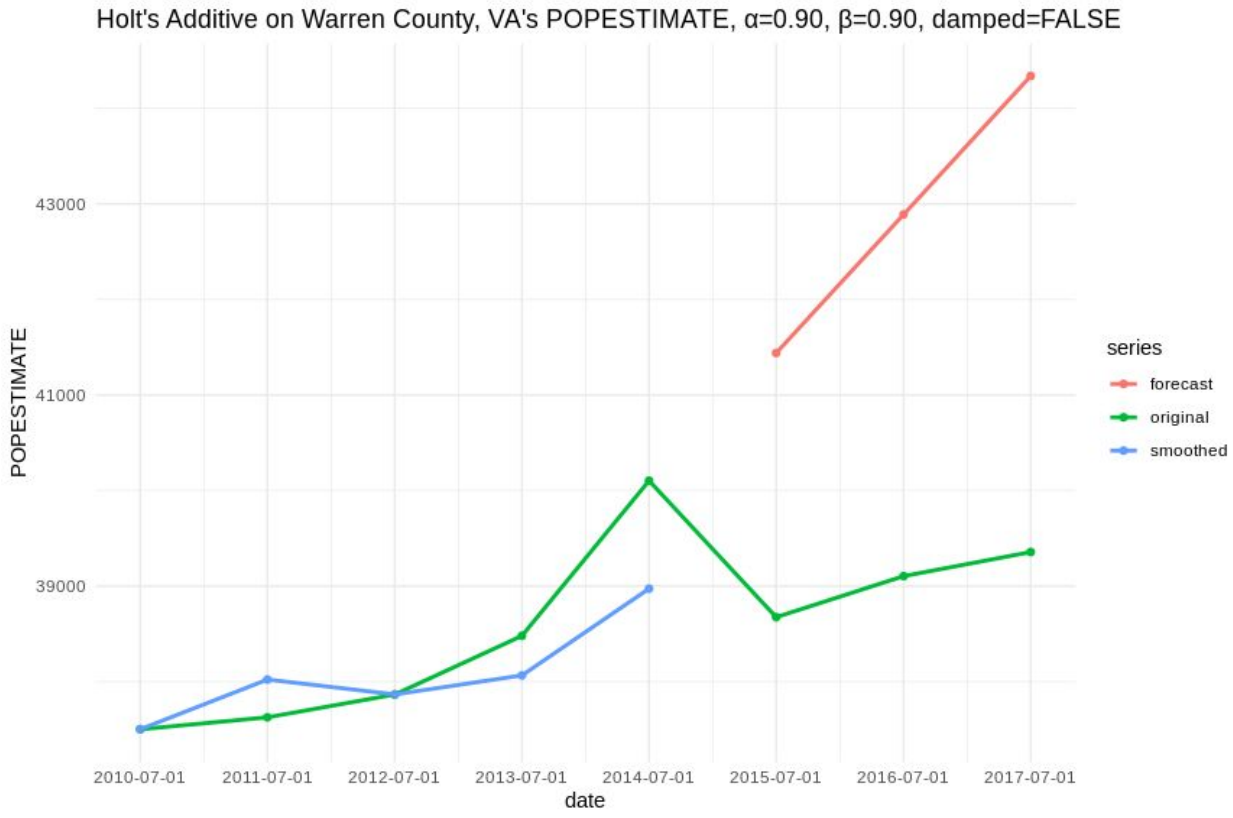

Holt's Additive on POPESTIMATE,α=0.90,β=0.60,damped=FALSE,AMAPE=0.79,AMPE=0.21
20
10
0
-10
values
Smoothed MAPE
Smoothed MPE
Forecast MAPE
Forecast MPE
factor(ind, levels = names(metrics))

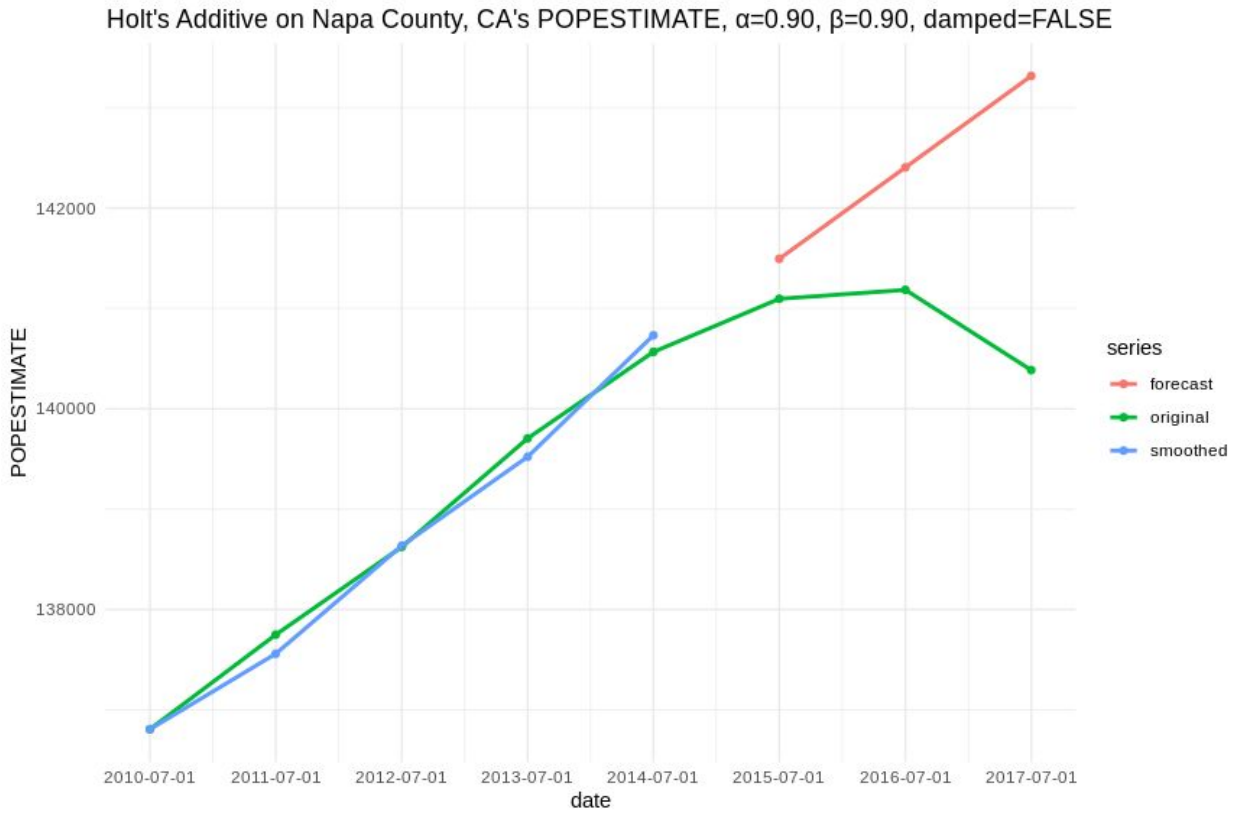


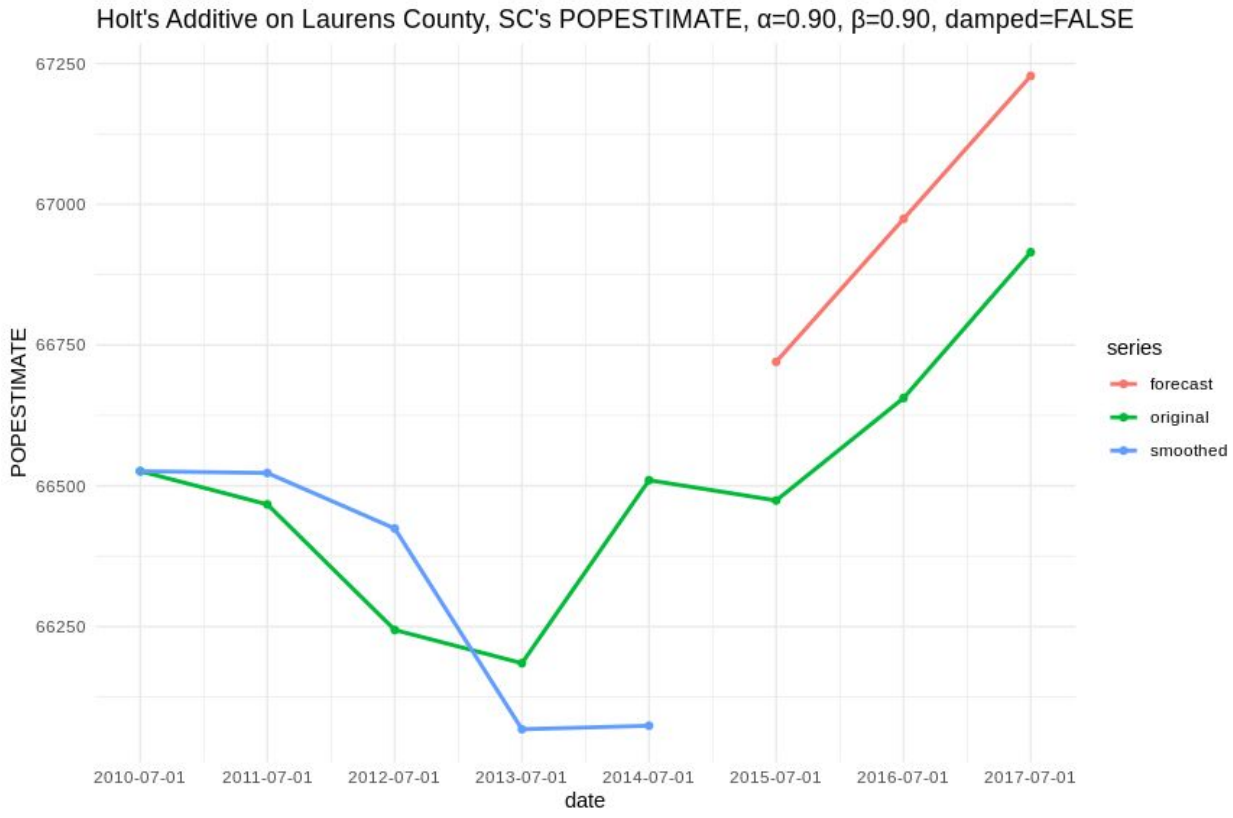
We found that the combination α=0.9 and β=0.9 yields the best result out of all of them. It was able to achieve average mean absolute percentage error of 0.77% and average mean percentage error of 0.18% in forecasting. Last plot shown above is the forecasting errors using parameters α=1 and β=1, which is equivalent to naively forecasting future population by adding up the difference between the last two data points. As it performs worse than using parameters α=0.9 and β=0.9, it is evidence that Holt's method is superior to naive forecasting. We have included below visualized examples of Holt's forecasting with α=0.9 and β=0.9, from ten randomly selected locations.

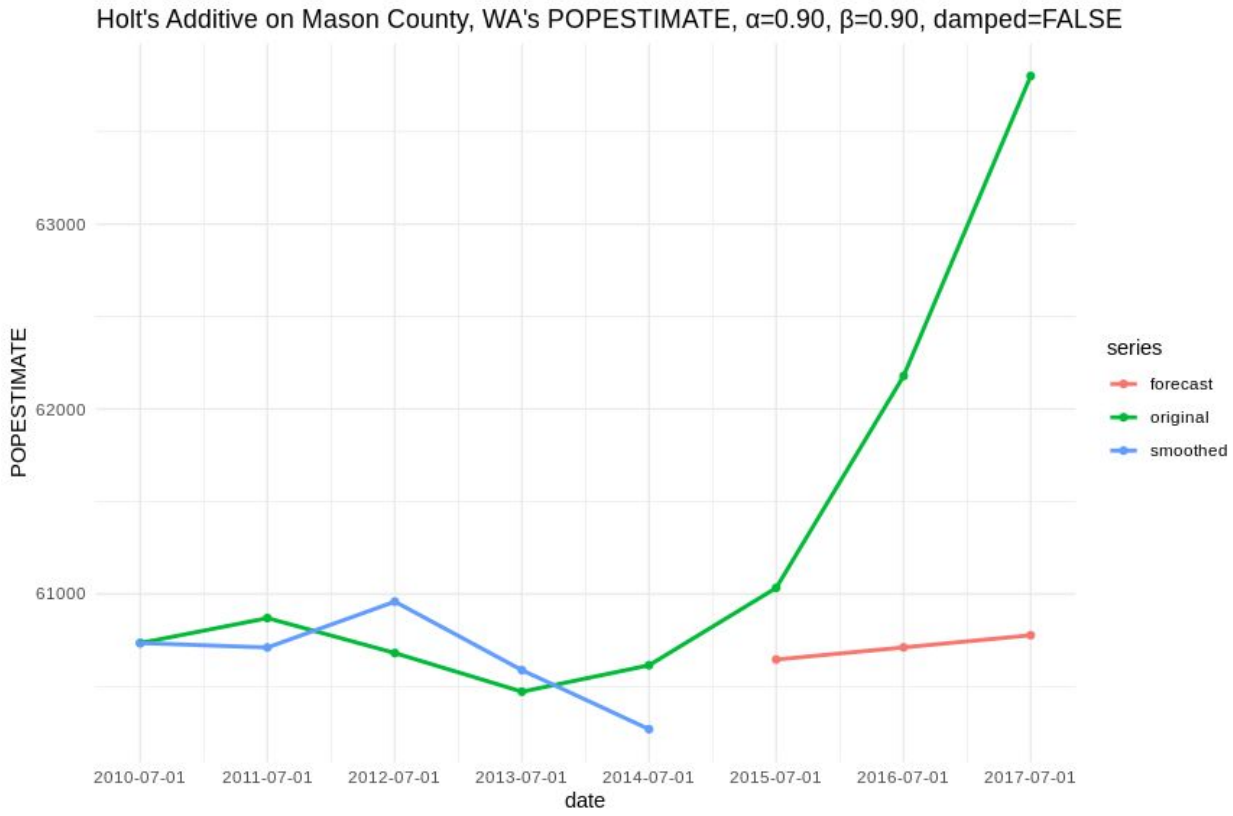

Holt's Additive on Washington County, NY's POPESTIMATE, α=0.90, β=0.90, damped=FALSE
POPESTIMATE
63000
62500
62000
61500
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed

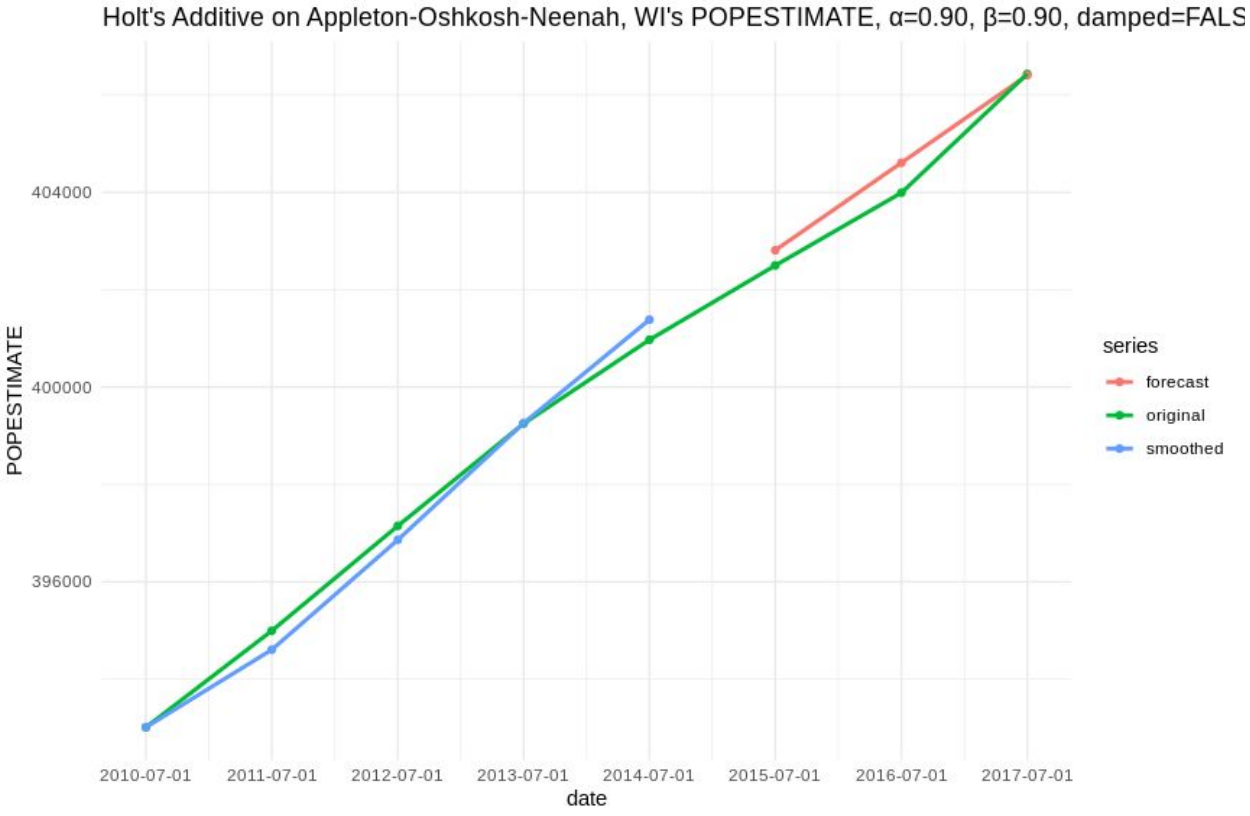

Holt's Additive on Eaton County, MI's POPESTIMATE, α=0.90, β=0.90, damped=FALSE
POPESTIMATE
109500
109000
108500
108000
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed

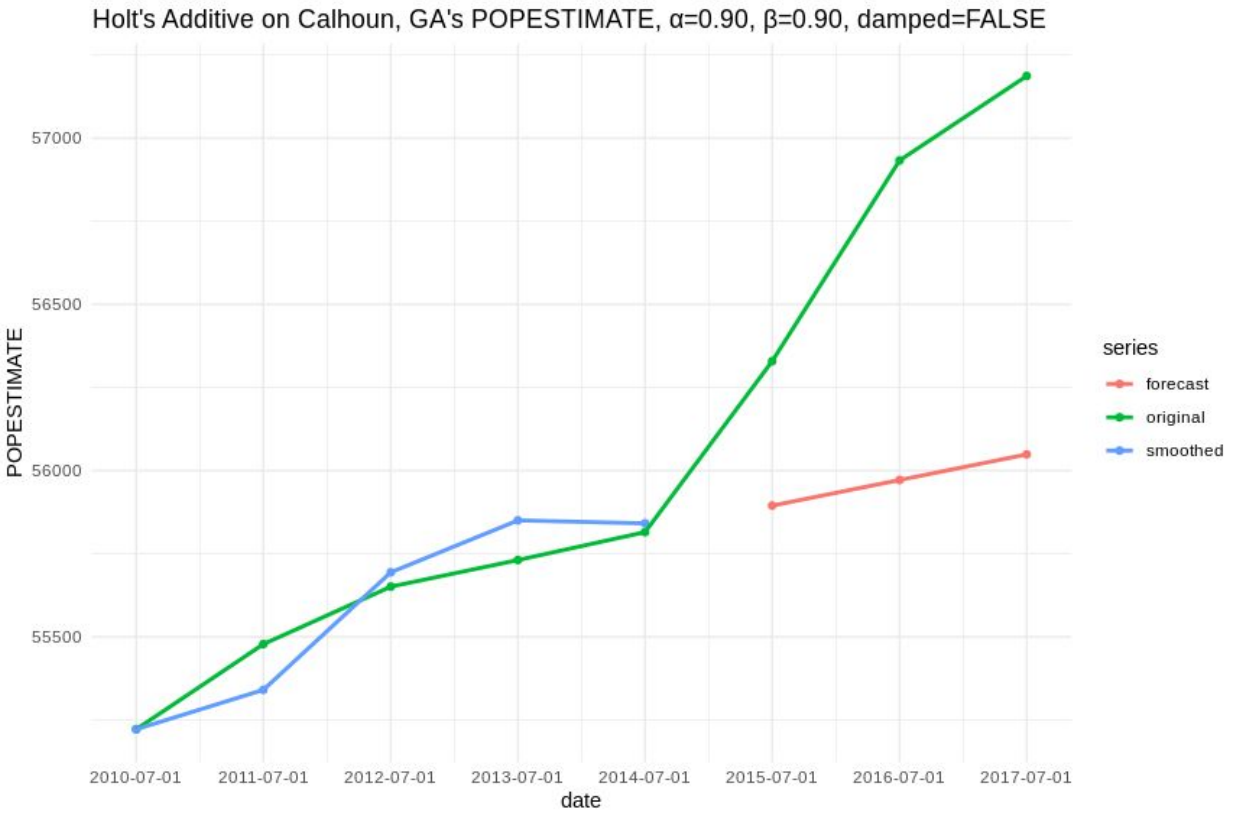

Holt's Additive on Warren County, VA's POPESTIMATE, α=0.90, β=0.90, damped=FALSE
POPESTIMATE
43000
41000
39000
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed

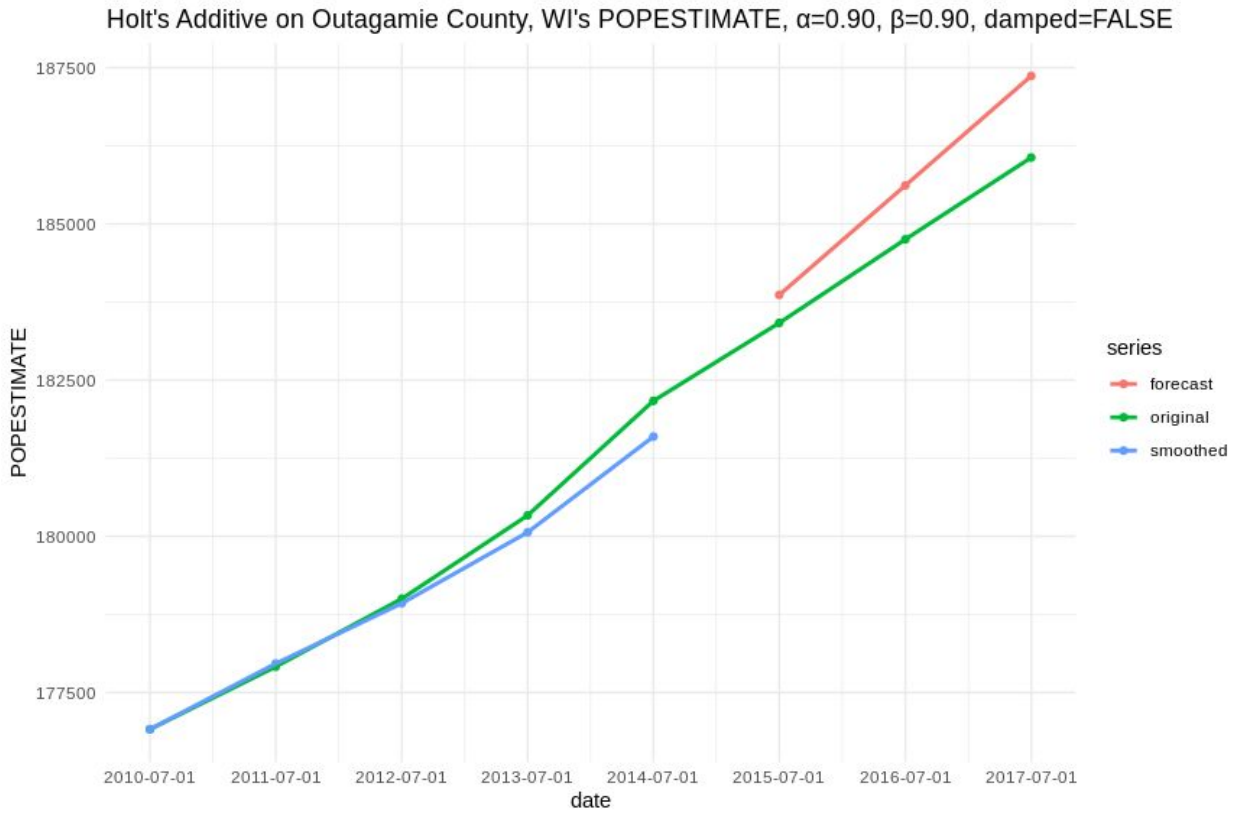

Holt's Additive on Napa County, CA's POPESTIMATE, α=0.90, β=0.90, damped=FALSE
POPESTIMATE
142000
140000
138000
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed

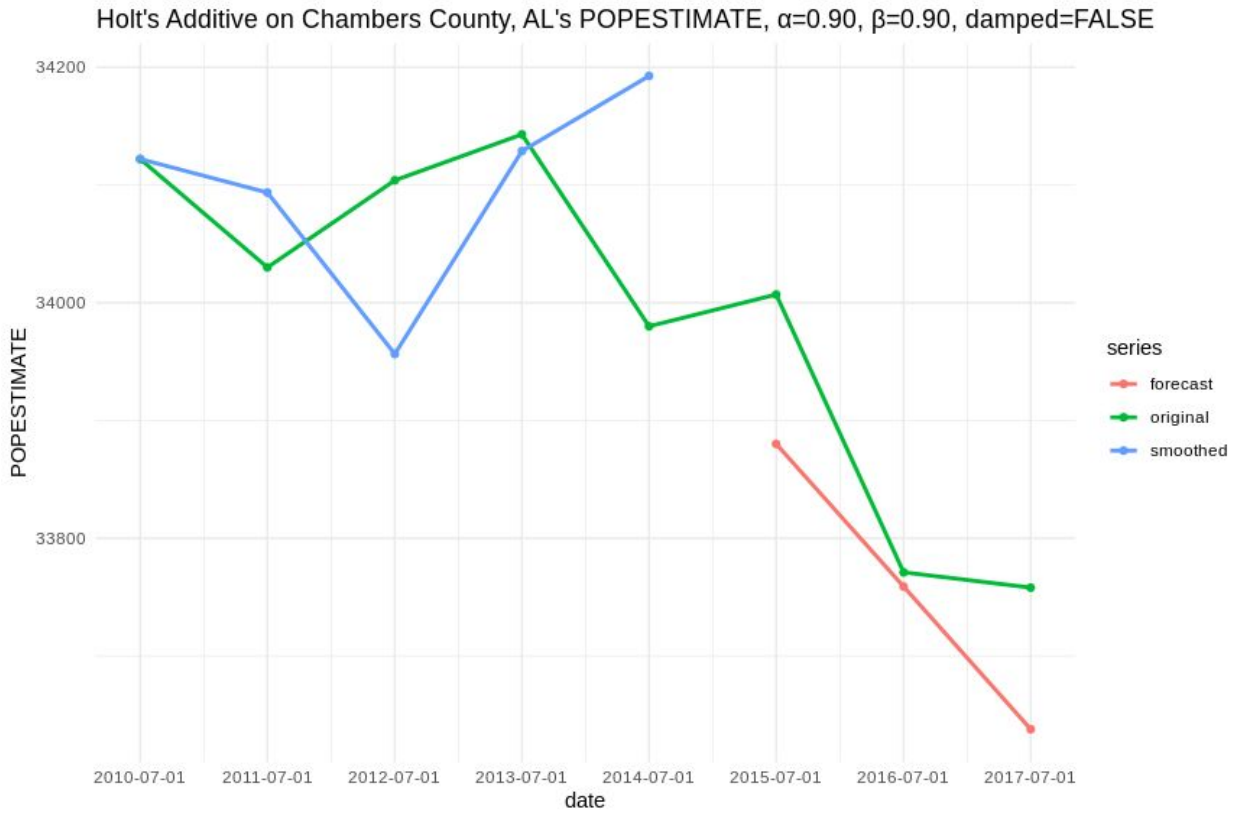

Holt's Additive on Laurens County, SC's POPESTIMATE, α=0.90, β=0.90, damped=FALSE
POPESTIMATE
67250
67000
66750
66500
66250
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed


Holt's Additive on Mason County, WA's POPESTIMATE, α=0.90, β=0.90, damped=FALSE
POPESTIMATE
63000
62000
61000
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed


Holt's Additive on Appleton-Oshkosh-Neenah, WI's POPESTIMATE, α=0.90, β=0.90, damped=FALS
POPESTIMATE
404000
400000
396000
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed


Holt's Additive on Calhoun, GA's POPESTIMATE, α=0.90, β=0.90, damped=FALSE
POPESTIMATE
57000
56500
56000
55500
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed


Holt's Additive on Outagamie County, WI's POPESTIMATE, α=0.90, β=0.90, damped=FALSE
187500
185000
182500
180000
177500
POPESTIMATE
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed


Holt's Additive on Chambers County, AL's POPESTIMATE, α=0.90, β=0.90, damped=FALSE
34200
34000
33800
POPESTIMATE
2010-07-01
2011-07-01
2012-07-01
2013-07-01
2014-07-01
2015-07-01
2016-07-01
2017-07-01
date
series
forecast
original
smoothed

Limitations and Future Work

The forecasting performance of exponential models such as Holt's Multiplicative Method and exponential regressions were not investigated, nor were simpler methods such as linear regression. By exploring a larger variety of standalone and ensembled models, we could potentially further improve the results. We could attempt to establish statistical prediction intervals to help stakeowners better manage the risks from using such forecasts, by investigating the distributions of forecasting errors and their relationships with other factors such as population size and income.

Analysis 2:

Immigration Patterns

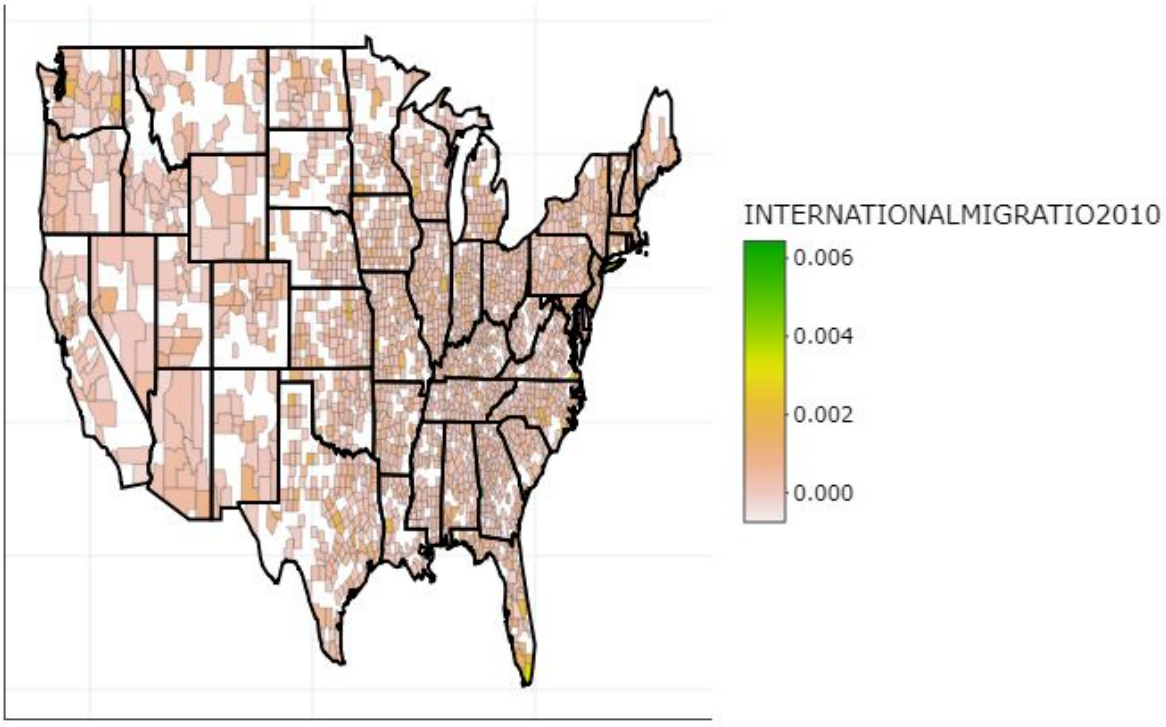
Hypothesis

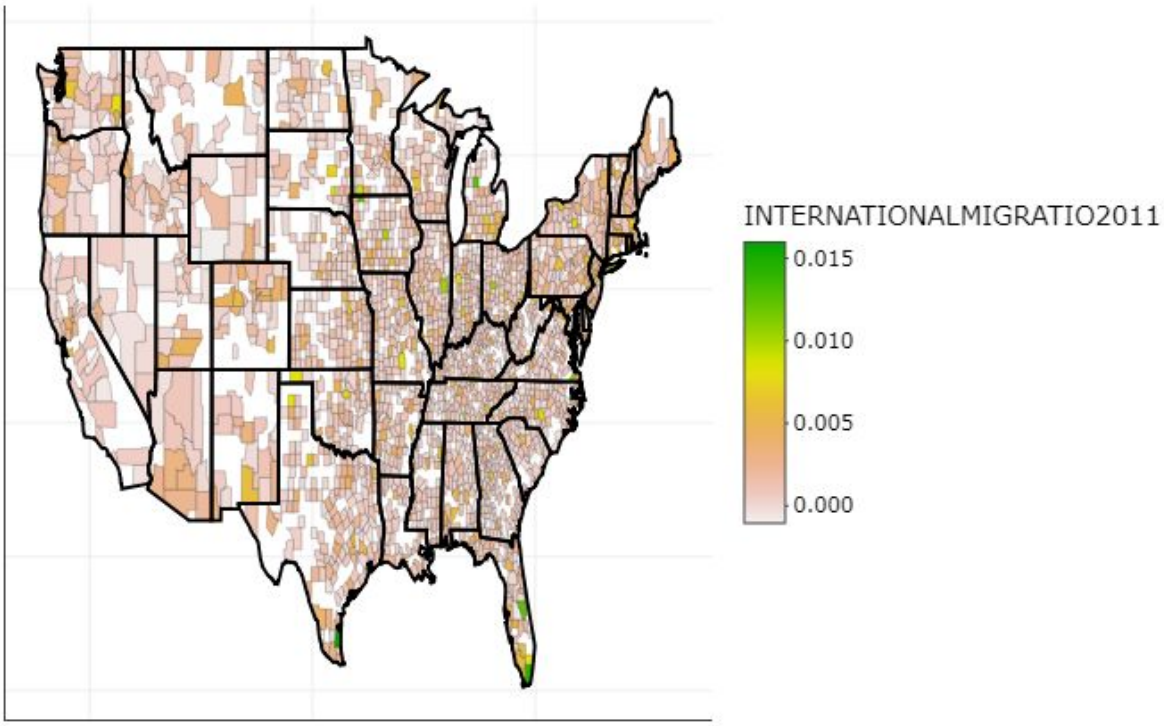
Most of our original thoughts came from common sense ideas or stories from the media. We hypothesized that immigration numbers would be higher closer to the Canadian and Mexican borders due to the low distance giving easier access. However, we also thought that immigration numbers, especially near the Mexican border, would decrease after the administration change in 2016 due to the new administration stance on immigration. The final idea we had was that after 2016 a spike would occur in migrations out of the US due to the results of the election. This would lead to lower immigration numbers especially in the northern part of the country to population being lost to international migration.

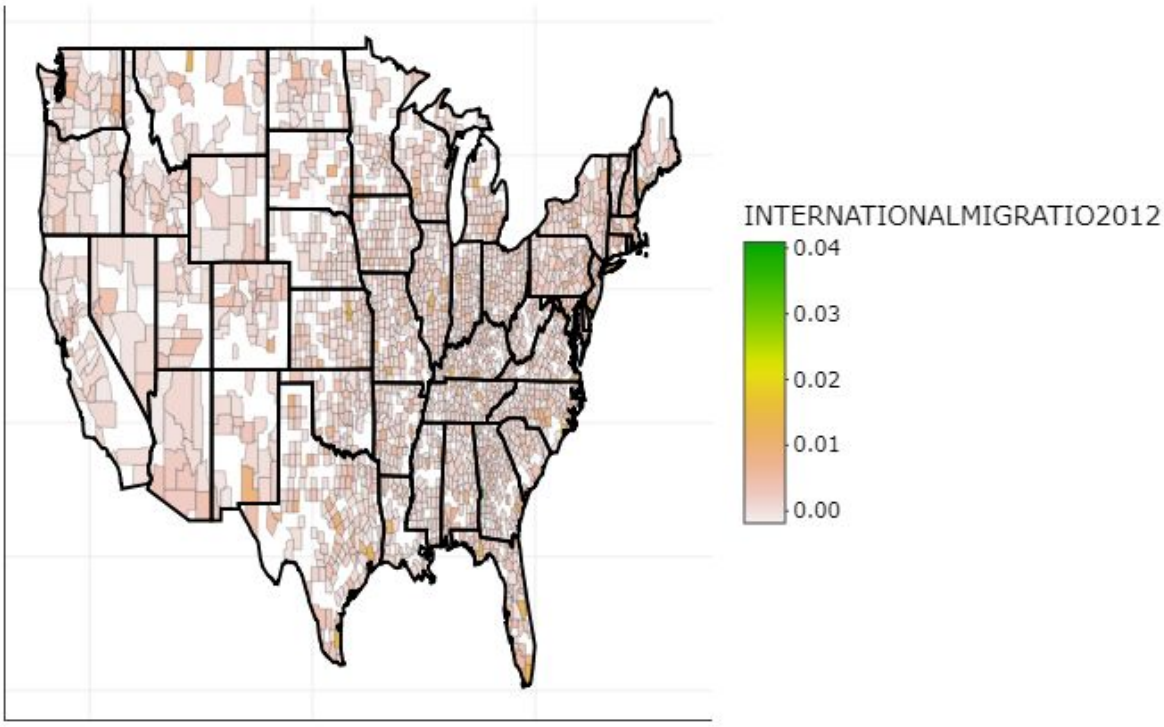
Investigation

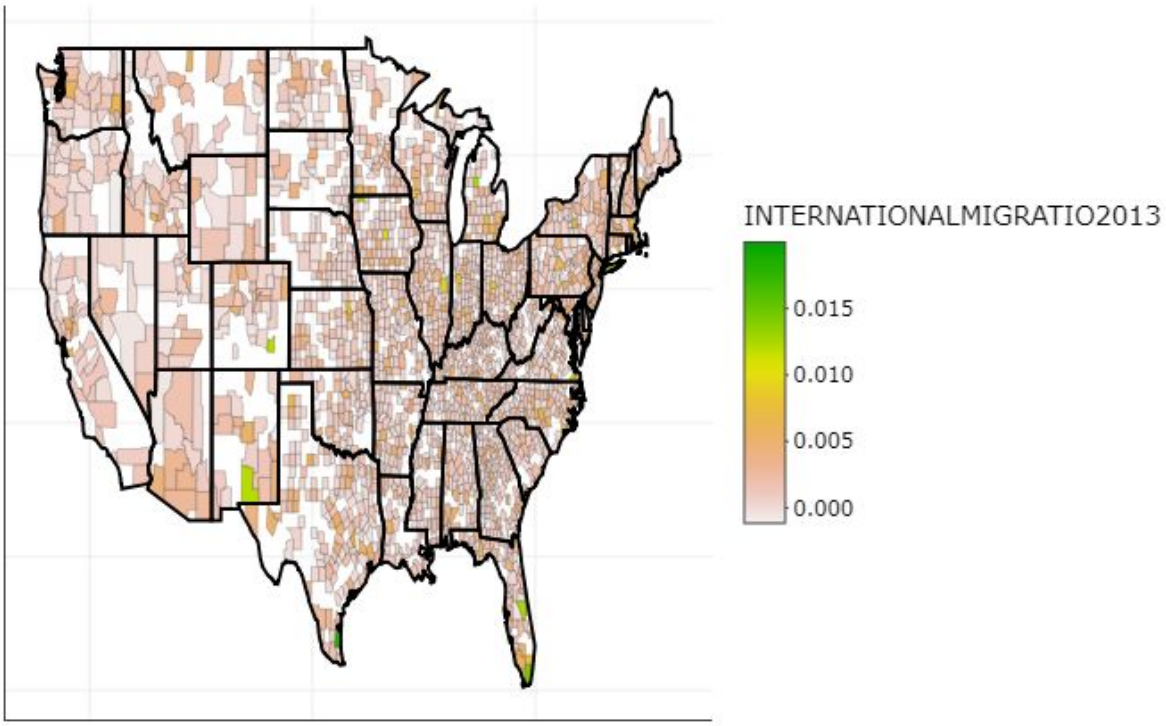
The first thing that jumped out was the stark contrast that different areas had in terms of immigration numbers. Cities like Seattle and Miami tended to have much higher immigration numbers than non-urban areas. They retain this advantage even when the numbers are adjusted for population of the county. However, some new areas pop up at the top as well. Osceola County, FL is one of those newcomers. This is likely due to the proximity to Miami-Dade County which is home to Miami. Also, Cook County, MN seems to have a very high flat immigration. However, they have a very low ratio which does not make any sense because Cook County has a population of 5,000 according to Wikipedia. This may be caused by the poor mapping of CBSAs to county lines or just a mistake in the data. Because the data had a few very large outliers, it was difficult to see differences in counties. To solve this issue we decided to normalize the data to make differences more clear.

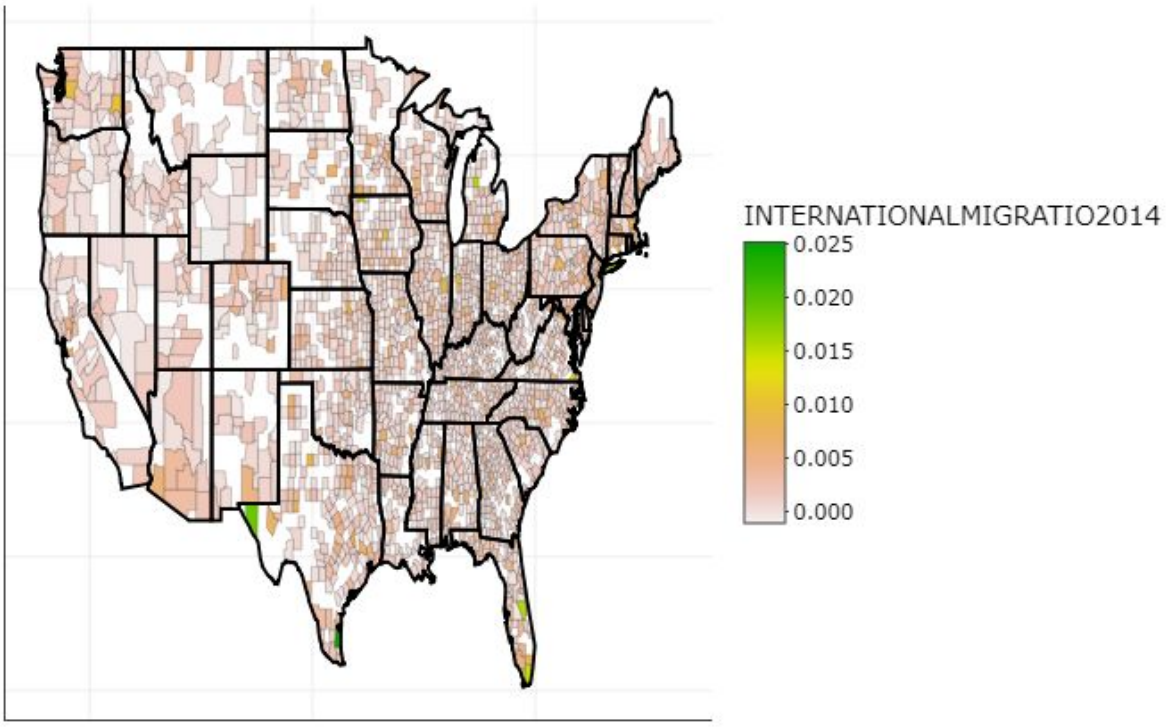
Here are normalized (between 0-1) heatmaps of international migration (listed plots for years 2010-2018)

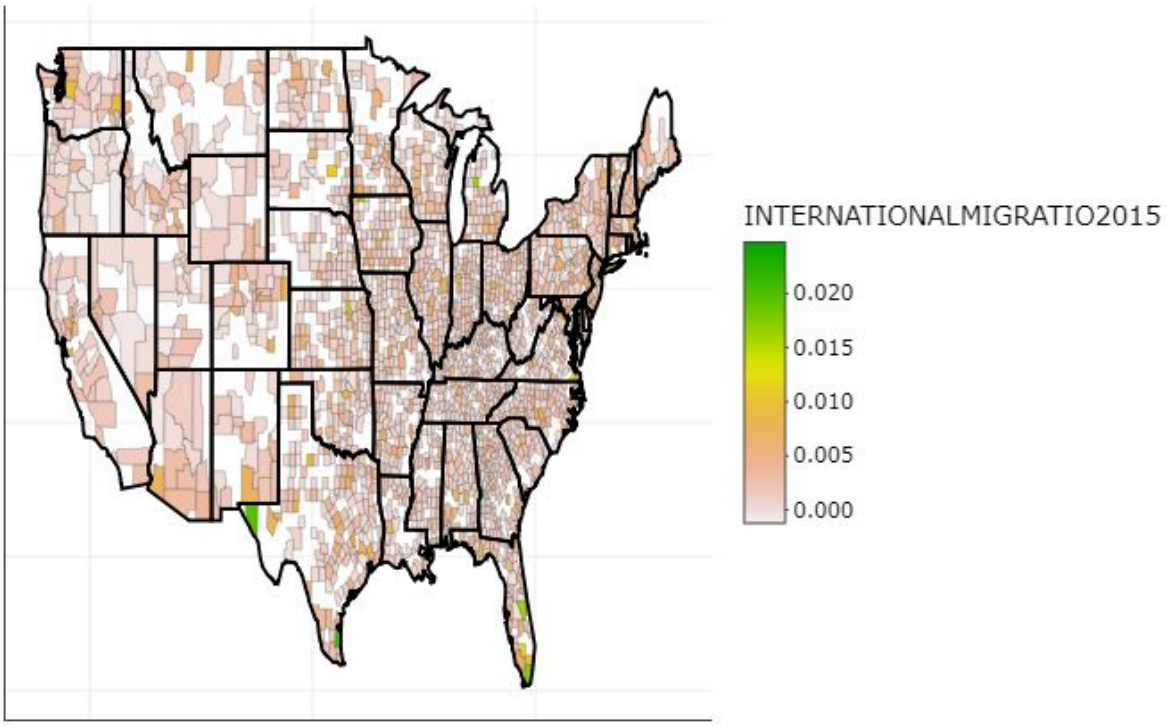


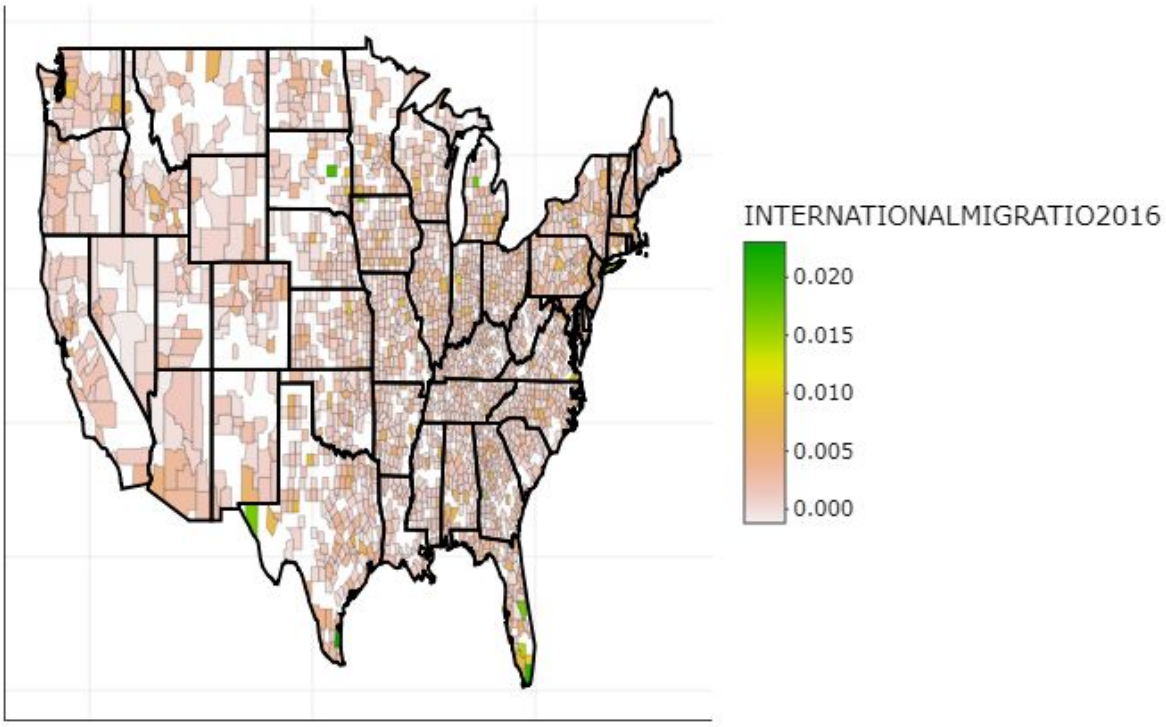

INTERNATIONALMIGRATIO2011
0.015
0.010
0.005
0.000

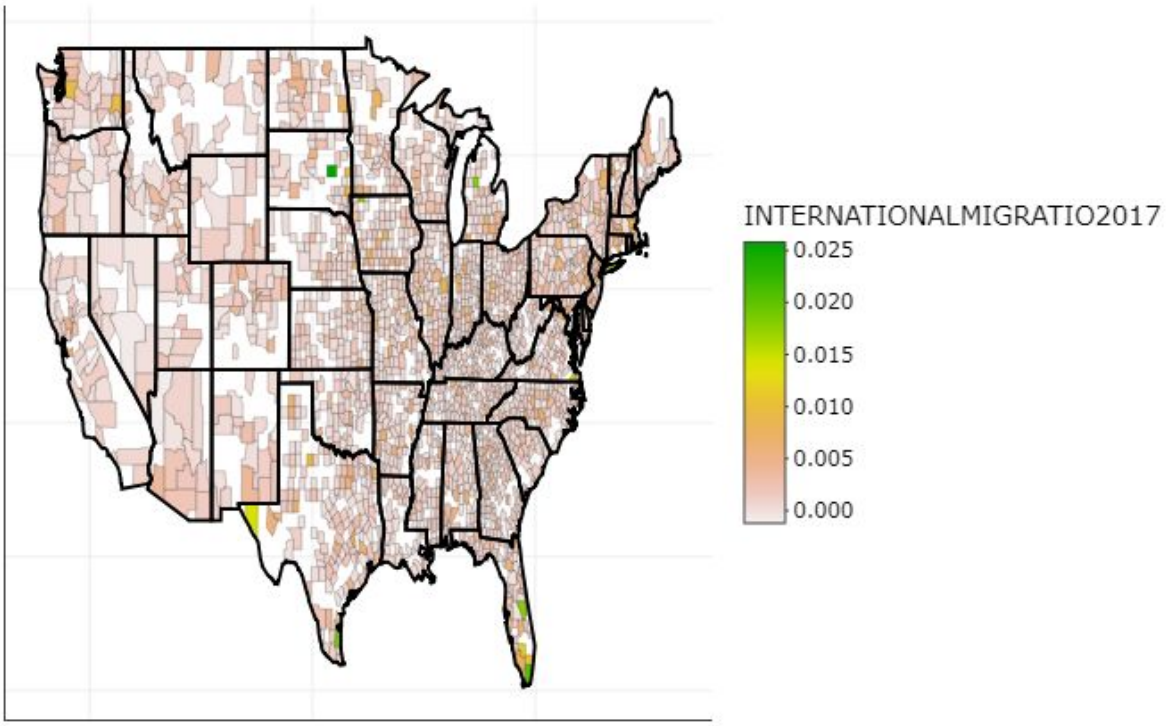

INTERNATIONALMIGRATIO2012
0.04
0.03
0.02
0.01
0.00

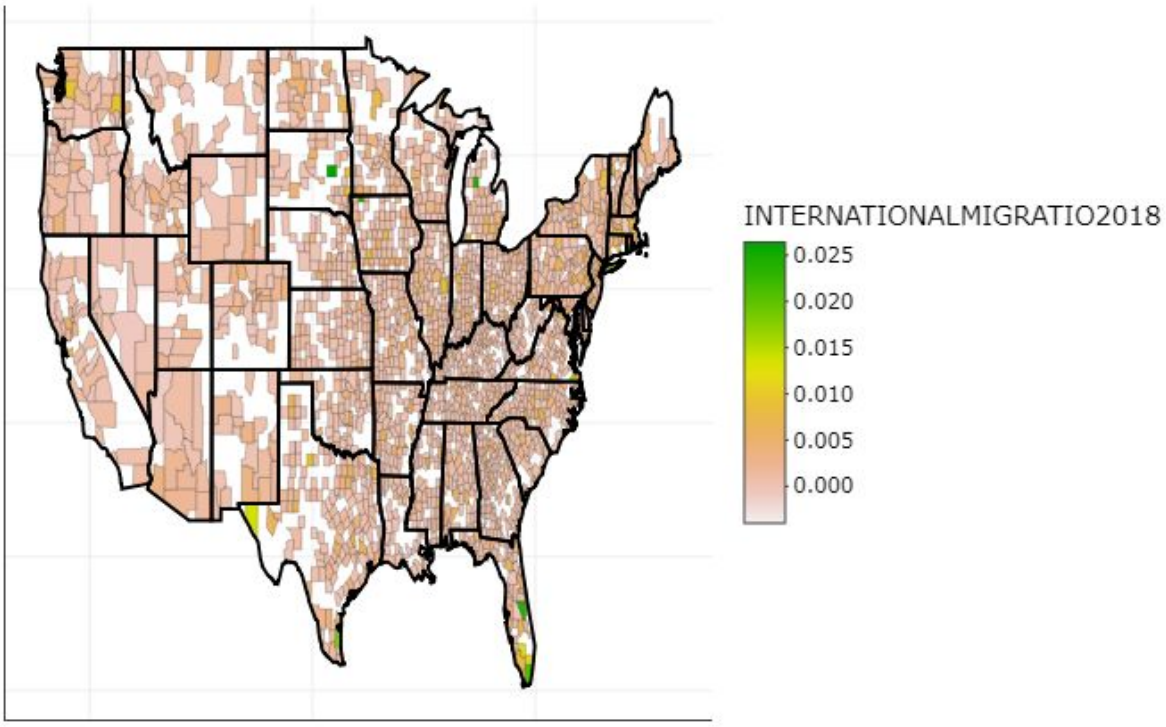

INTERNATIONALMIGRATIO2013
0.015
0.010
0.005
0.000


INTERNATIONALMIGRATIO2014
0.025
0.020
0.015
0.010
0.005
0.000


INTERNATIONALMIGRATIO2015
0.020
0.015
0.010
0.005
0.000


INTERNATIONALMIGRATIO2016
0.020
0.015
0.010
0.005
0.000





Of note is that in more recent years, international migration leveled out to what can be originally observed in 2010, whereas the highest peak migration is in 2012. It is interesting that 2012 is

the year that the reelection of former president Barack Obama occured. Of the major events in the United States in 2012, one that could provide a big enough incentive for migration based on moral issues could be the legalization of same-sex marriage, which could be attributed to Obama's supportive stance. While we initially speculated that 2016, the year of Trump's election would have higher migratory patterns due to people's statements about leaving the country, it is clear that this is not visible through the census data. Our thoughts about most immigration occurring near the Canadian and Mexican borders is somewhat supported by the visualizations. Many of the counties with the highest immigration were in states close to the border. However, once past the highest few counties, the distribution appears more random.

Future Work

Finding a dataset that has county delimited data would allow a better mapping to take place. This data used statiscal areas which sometimes are equivalent to counties, however, they sometimes include multiple counties which makes it impossible to plot the data completely on a county base map. Another option would be to create a CBSA based map. Another interesting angle we could take would be to look at domestic migration as well. Further Analysis could be done by clustering states together to allow for greater understanding of which states have the most migration (likely those on the border) as well as determining whether regions that are typically considered conservative or liberal are more likely to migrate (likely those in liberal areas). However, this would require another dataset with the political lean of a state to calculate (perhaps from 538.com).