Our analyses consist of comparing and contrasting four pairs of data signals in the Covidcast API:

1. Estimated test positivity rate (percent) among people tested for COVID-19 in the past 14 days from CTIS vs. Percentage of Antigen tests that were positive for COVID-19 smoothed by pooling together the last 7 days of tests from Quidel
2. Percentage of positive NAAT([Nucleic Acid Amplification Tests](https://www.cdc.gov/coronavirus/2019-ncov/lab/naats.html)) in the past 7 days from Community Profile Report vs. Percentage of Antigen tests that were positive for COVID-19 smoothed by pooling together the last 7 days of tests from Quidel
3. Estimated percentage of people reporting that they or someone in their household experienced a sore throat in the past 24 hours from CTIS vs. The average of Google search volume for related searches of symptom set “Laryngitis, Sore throat, Throat irritation”, in arbitrary units that are normalized for overall search users, smoothed by a 7-day average.
4. Estimated percentage of outpatient doctor visits primarily about COVID-related symptoms from Change Healthcare claims data (CHNG) vs. Estimated percentage of outpatient doctor visits primarily about COVID-related symptoms from health system partners (Doctor-visits)

# **1. Methods**

For each pair of data signals, our analysis consists of four parts: Spearman rank correlation at state level or at county level, possible insights to spatial agreement / disagreement, Spearman rank correlation within a single state, and lagged Spearman rank correlation over time.

## **1.1 Spatial Spearman rank correlation**

Our analyses focused on finding a geographical pattern in Spearman rank correlation between any pair of data signals at state level and at county level. We plotted the Spearman rank correlation for a given pair of signals on a map of US states and used visual comparison. At the county level, we calculated this correlation for two most populous counties and two least populous counties in the United States.

### **1.1.1 Possible insights explaining high or low Spearman rank correlations at state level**

Given the spatial rank correlations calculated for each pair of data signals , we would like to know if there is any indicator associated with high / low rank correlations between any pair signals across states or across counties in the U.S.. First, we visualized and explored how the missing values are associated with the rank correlations across states / counties. Further, we explored some demographic indicators at state level: median household income, population density, and percentage of population aged 65 and above. We first visualized each of the demographic indicators and the rank correlation across states. Later, we created a regression model where the rank correlation being the response variable and others treated as covariates to see how we can predict rank correlation from these covariates indicated above.

## **1.2 Temporal Spearman rank correlation**

We focused on identifying any temporal pattern in Spearman rank correlation for six most populous states in the U.S. (New York, California, Pennsylvania, Texas, Florida, Illinois). We looked at the changes in the rank correlation over time in three ways. Firstly,we looked at the weekly rank correlation between a given pair of data signals. Secondly, we looked at the monthly rank correlation between a given pair of data signals. Lastly, we created a 31-day moving window and saw how the rank correlation changed over these 31-day moving windows.

### **1.2.1 Lagged temporal Spearman rank correlation**

When we visualized the data signals over time, we noticed that data signals followed a similar trend but one signal lagged behind the other. In this case, we would like to see how the rank correlations change if we account for these lags. First, we calculated the cross correlations between any pair of data signals specified above and found the lag where the rank correlations are maximized. Then, we adjusted one of the data signals given the number of lags and recalculated the rank correlations between adjusted data signals. At last, we checked to see how the lagged rank correlation changed over time with three different ways specified in 1.2.

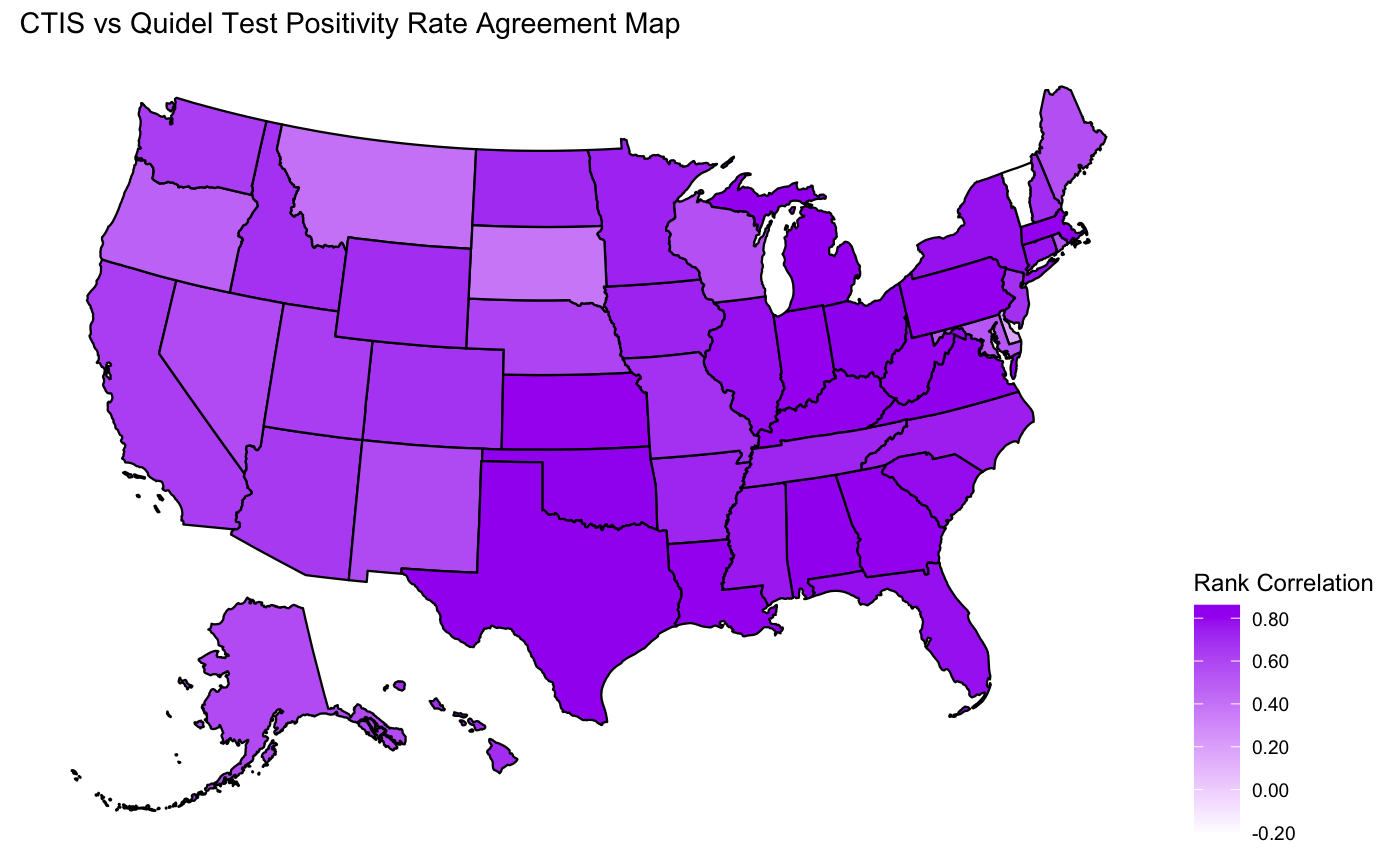
We also tried to use a different method to measure the agreement between any pair of data signals and we repeated each topic specified above with Pearson calculation calculated instead of Spearman rank correlation.

# **2. Results**

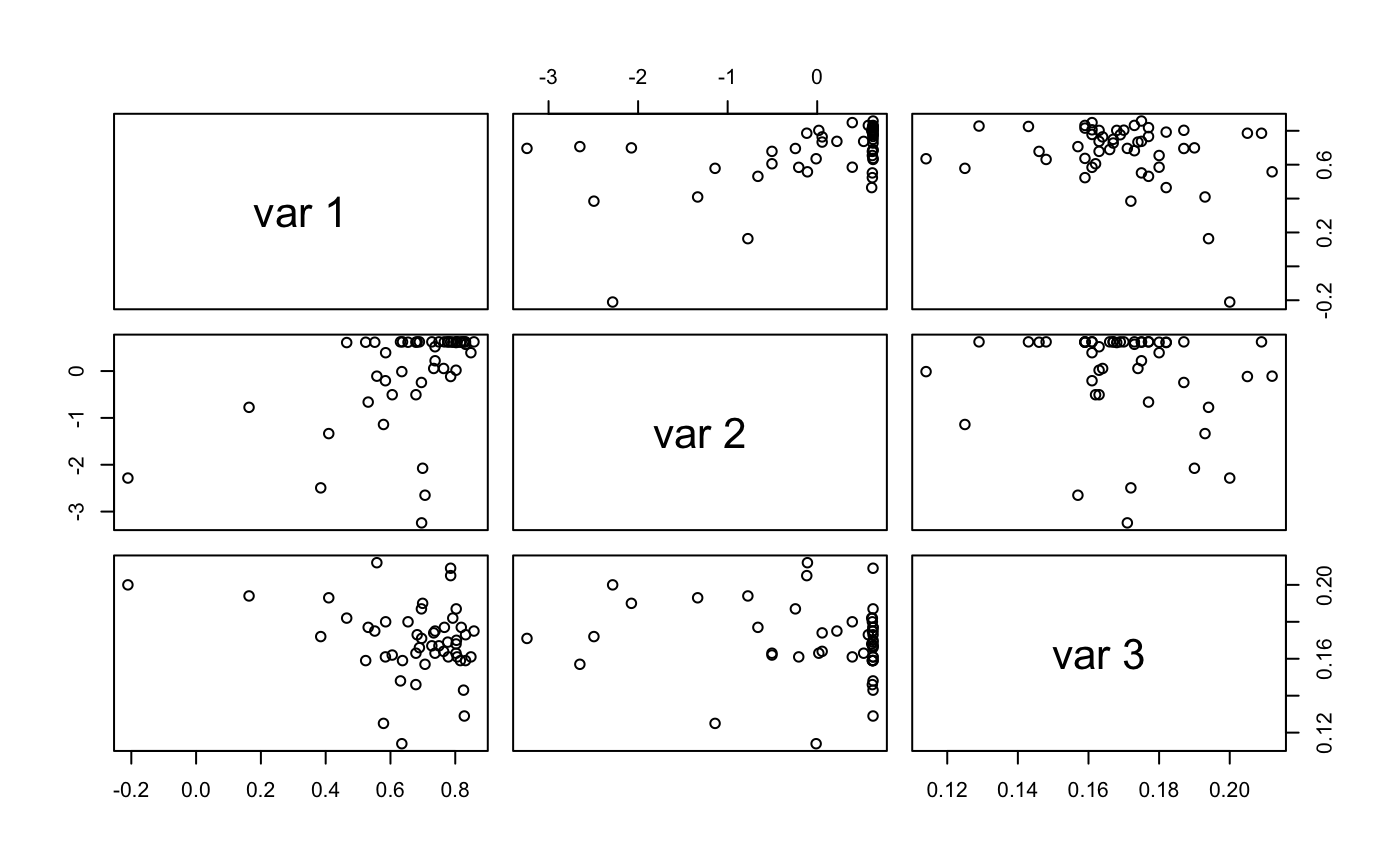
To answer our research question, we have looked at four pairs of signals that are supposedly getting the same idea. Our investigation mainly focuses on spatial correlation and temporal correlation.

**2.1 Past 14 Days Test Positivity Rate (CTIS) vs. Past 7 Days Test Positivity Rate (Quidel)**

For spatial correlation, we first investigate the rank correlation between two signals within each state in Figure 1. The agreement rate between two signals varies largely between states. States on the east coast have higher agreement rates than states on the west coast. For example, Ohio has the highest rank correlation of 0.86, while Vermont has the lowest correlation of -0.21. One possible cause behind the difference is the amount of missing values from CTIS. In Figure 2, we find a positive correlation between rank correlations across states and scaled CTIS observations, which means that the rank correlation increases as there are more CTIS observations in the state.



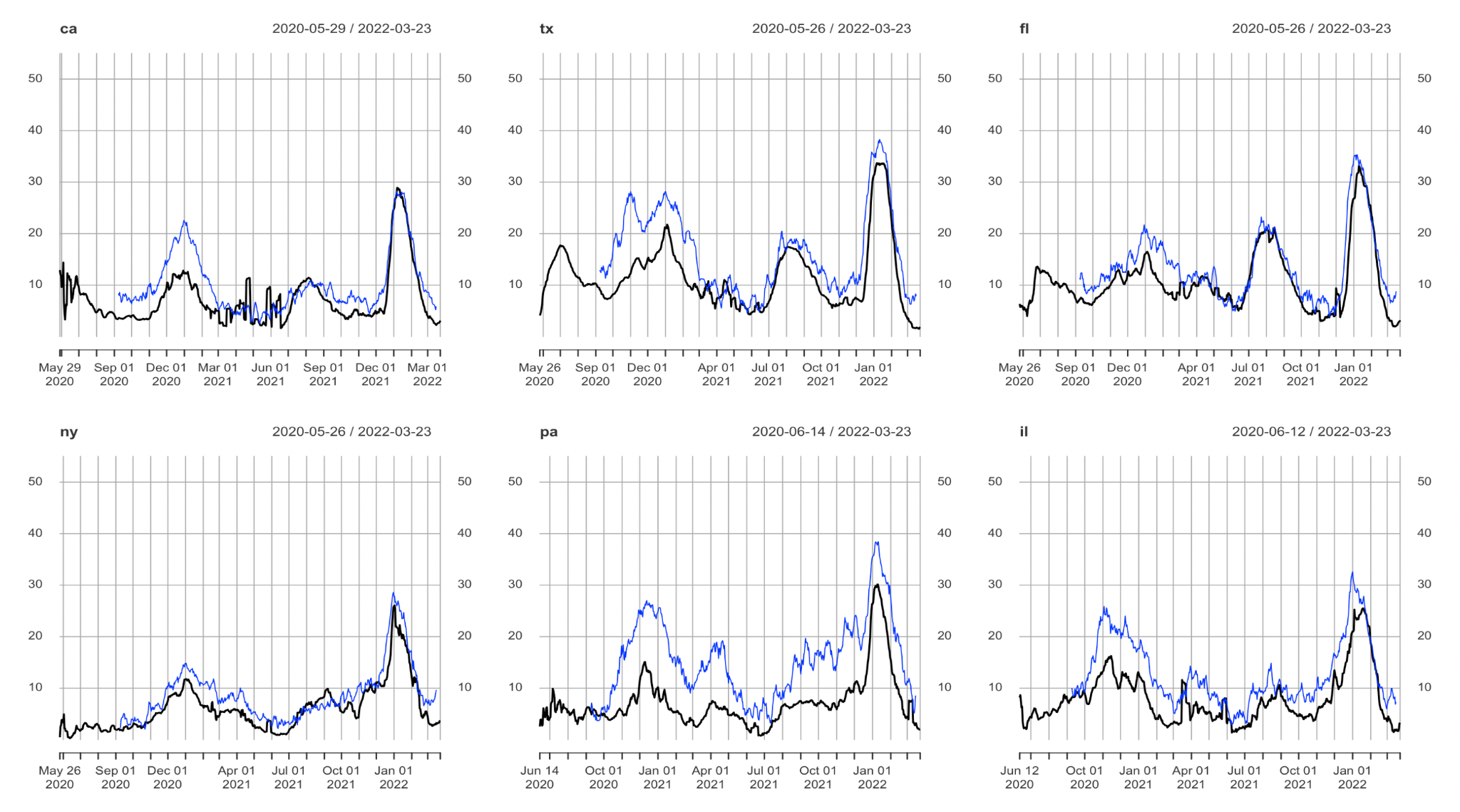
**Figure 1:** CTIS vs Quidel Test Positivity Rate Agreement Map



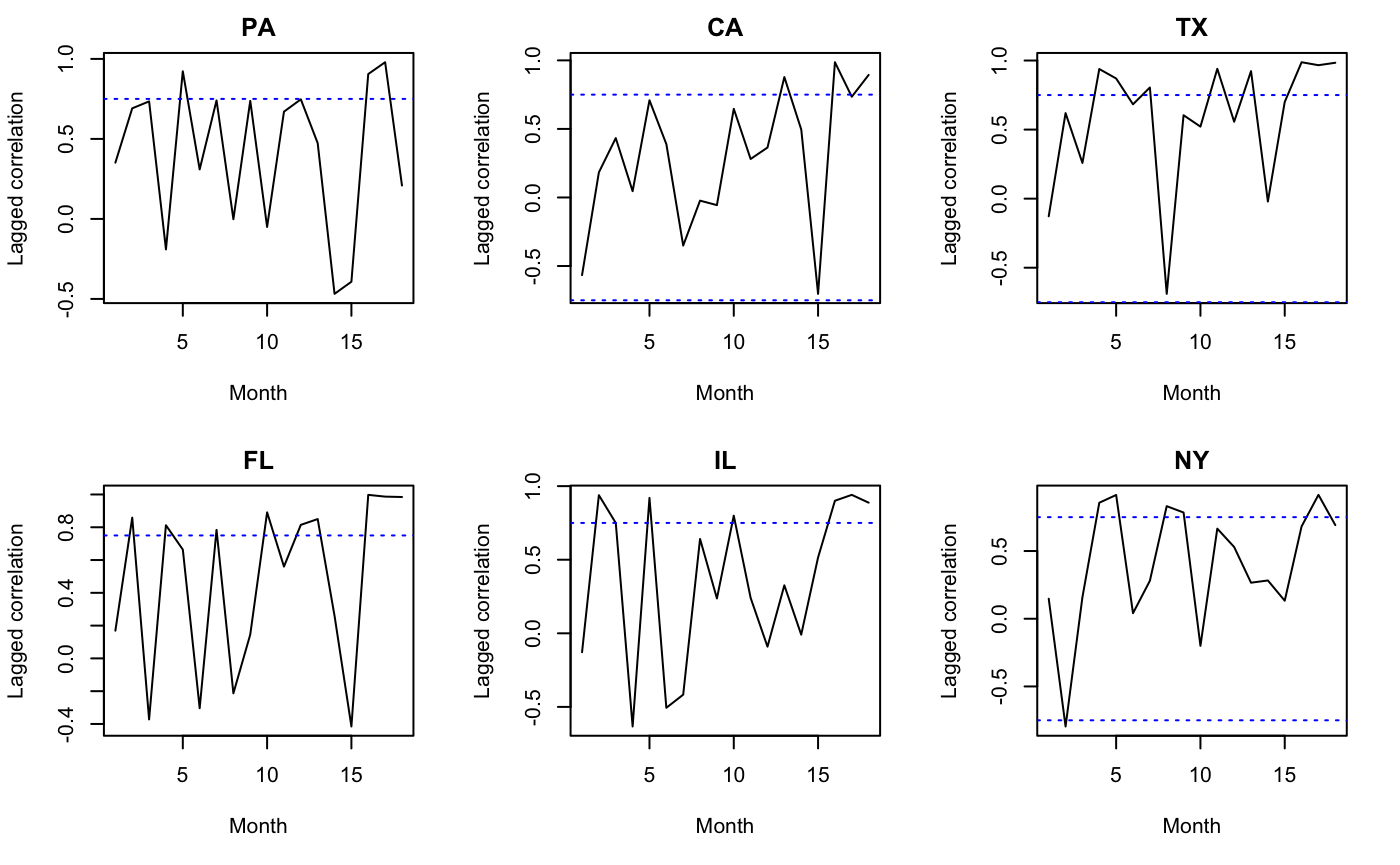
| Var 1 | Rank correlations across states |
| --- | --- |
| Var 2 | Scaled CTIS observations [0,1] |
| Var 3 | Population percentage of aged 65 and above |

**Figure 2:** Correlation Scatter Plot

For temporal correlation, we compare these two test positivity rates from 05/29/2020 to 03/23/2022 in six most populous states (CA, TX, FL, NY, PA, IL) to avoid the effect of missing values from CTIS. In Figure 3, we find that the test positivity rate from CTIS is lagged by 10 days, and that the test positivity rate from Quidel is constantly lower than that from CTIS, especially during the outbreaks around December 2020. One possible explanation is the different way these two test positivity rates are calculated. CTIS measures the percentage of people who report positive in the past 14 days through facebook survey, while Quidel measures the percentage of positive antigen tests taken in the past 7 days. Therefore, there should be a lag in CTIS compared to Quidel. In addition, a person can take multiple tests, which means that the denominator of the test positivity rate from Quidel (the number of antigen tests taken) is larger than the denominator of the test positivity rate from CTIS (the number of people reported). The test positivity rate with larger denominator should be smaller in value, which explains why the test positivity rate from Quidel is always lower. Lastly, we also compare the monthly rank correlation in these six states in Figure 4. The rank correlation is usually higher during the outbreak, as the correlation is high around February 2021 for all six states.



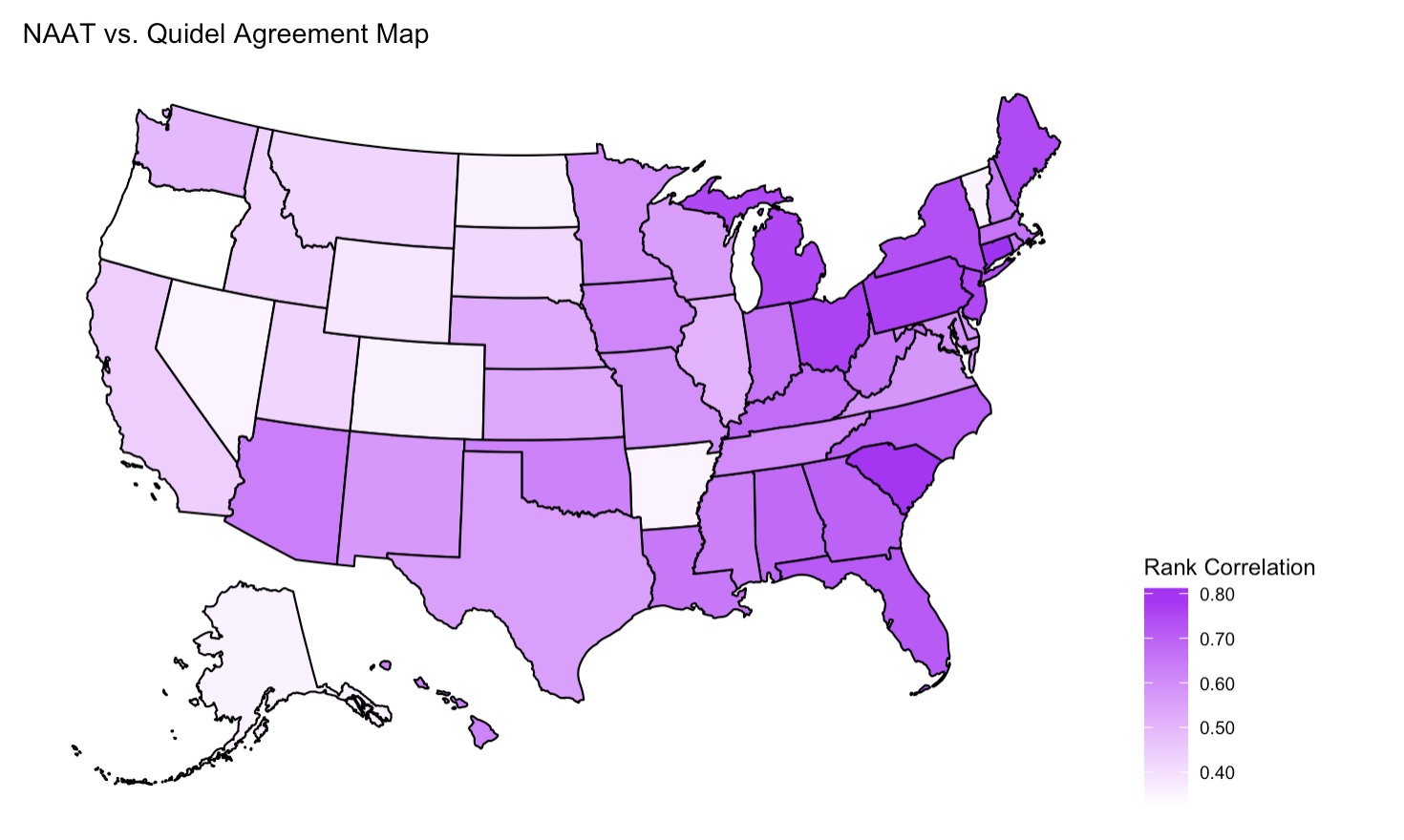
**Figure 3:** CTIS vs Quidel Test Positivity Rate Time Series Plots



**Figure 4:** CTIS vs Quidel Monthly Rank Correlation Plots

## **2.2 Past 7 Days Test Positivity Rate (Quidel) vs. Past 7 Days Positive NAAT Rate (Community Profile Report)**

From the rank correlation map in Figure 5, we find that the agreement rate between two signals varies largely across states. Similarly, states on the east coast have much higher agreement rates than states on the west coast. For example, Connecticut has the highest rank correlation of 0.81, while Oregon has the lowest correlation of 0.32. One possible cause behind the difference is the different tests these two signals are measuring. Quidel measures the percentage of positive antigen tests, while NAAT stands for nucleic acid amplification tests, which is also known as PCR test. Compared to antigen tests, PCR tests are more accurate but take more time. The popularity of these two tests are different across states, which also may contribute to the differences in agreement rates across states.

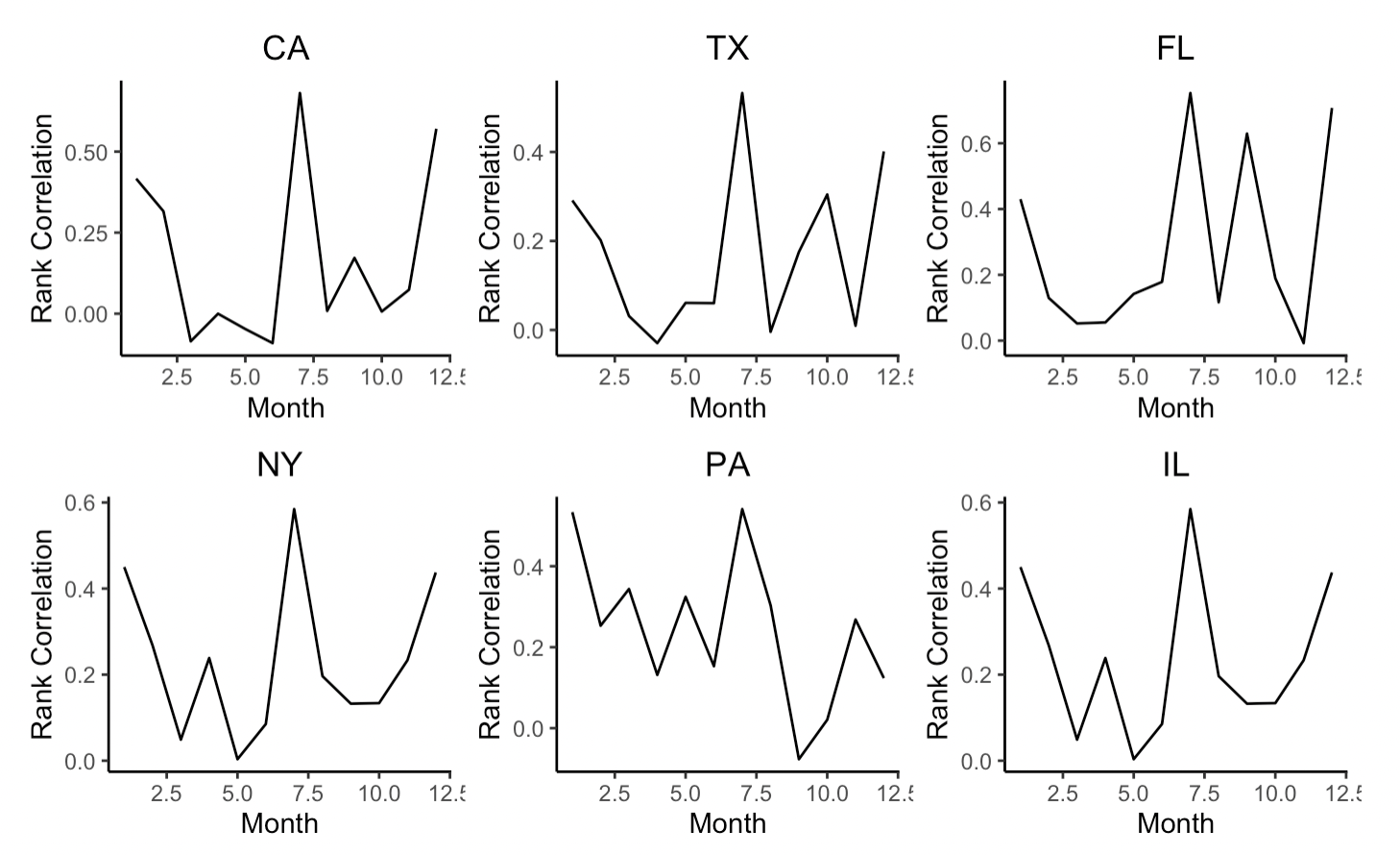


**Figure 5:** NAAT vs Quidel Test Positivity Rate Agreement Map

For temporal correlation, we start by comparing these two test positivity rates from 01/01/2021 to 12/31/2021 in six states. In Figure 6, there is no obvious lag as both signals measure the test positivity rate in the past 7 days. Overall, these two signals are in agreement for most of the time. California, Texas and Florida have a similar trend, as there is a peak around August for both test positivity rates in these three states. One major difference is that NAAT changes much more smoothly than Quidel over time. One explanation is that Quidel has fewer observations than NAAT. The test positivity rate fluctuates more when the value is lower, which also explains why Quidel varies much more from April to June. In addition, NAAT begins to grow more after the peak in August, which is after the Delta variant wave. One possible reason is that PCR tests are more frequently used to detect the Delta variant, and the test positivity rate tends to be higher due to Delta’s hyper-transmissibility. The monthly rank correlation plots in Figure 7 further strengthen our findings. The rank correlation is the highest in July for all six states, which reinforces our hypothesis that two signals agree more when there are more observations. Figure 7 also shows that Texas and Florida have a second peak around September, which might be due to the high positivity rate caused by the delta variant in August.

## 

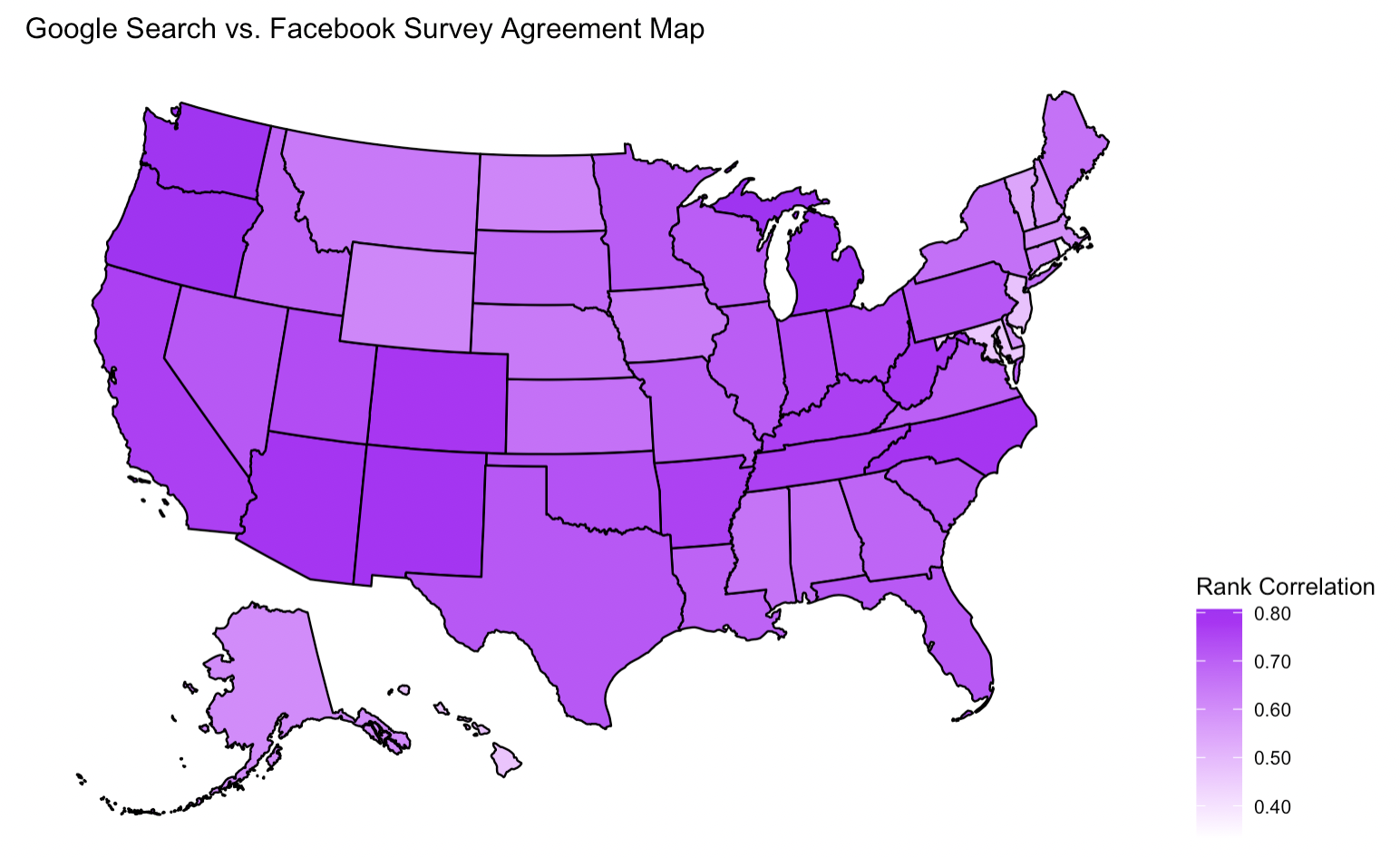
**Figure 6:** NAAT vs Quidel Test Positivity Rate Time Series Plots



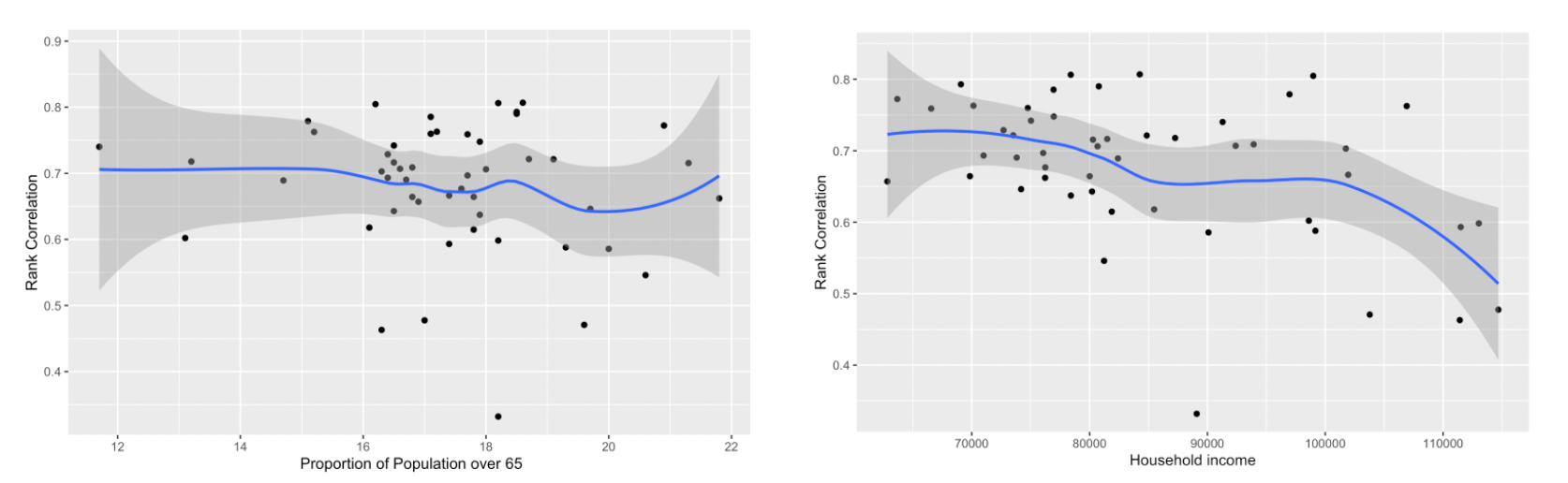
**Figure 7:** NAAT vs Quidel Test Positivity Rate Monthly Rank Correlation Plots

## **2.3 Percent of People Reporting Sore Throat Symptoms (CTIS) vs. Search Volume for Sore Throat (Google Search)**

When comparing two signals on sore throat symptoms, we find less agreement in New England / Midwest and more agreement in Northwest / Central in Figure 8. We notice that larger states have higher rank correlations and smaller states have lower rank correlations. For example, Oregon has the highest rank correlation of 0.81, while Rhode Island has the lowest correlation of 0.33. Different demographics is a possible reason for the difference in rank correlation across states. For example, states with older populations might report fewer COVID symptoms since they use Google or Facebook less frequently. However, Figure 9 shows that there is no obvious relationship between rank correlation and aged populations. The only correlated demographic information is household income. The scatterplot shows a negative correlation between rank correlation and household income, which means that the rank correlation between two signals decreases as the household income increases.

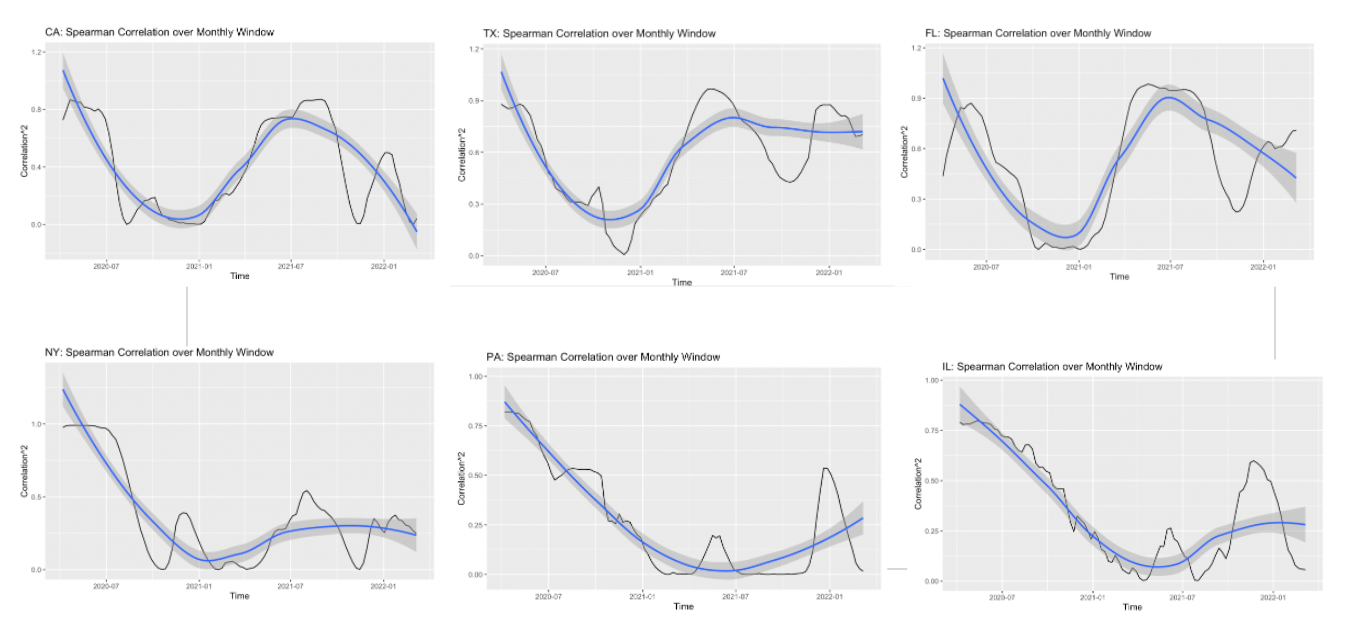


**Figure 8:** Google Search vs Facebook Survey Agreement Map



**Figure 9:** Google Search vs Facebook Survey Demographic Information Scatterplots

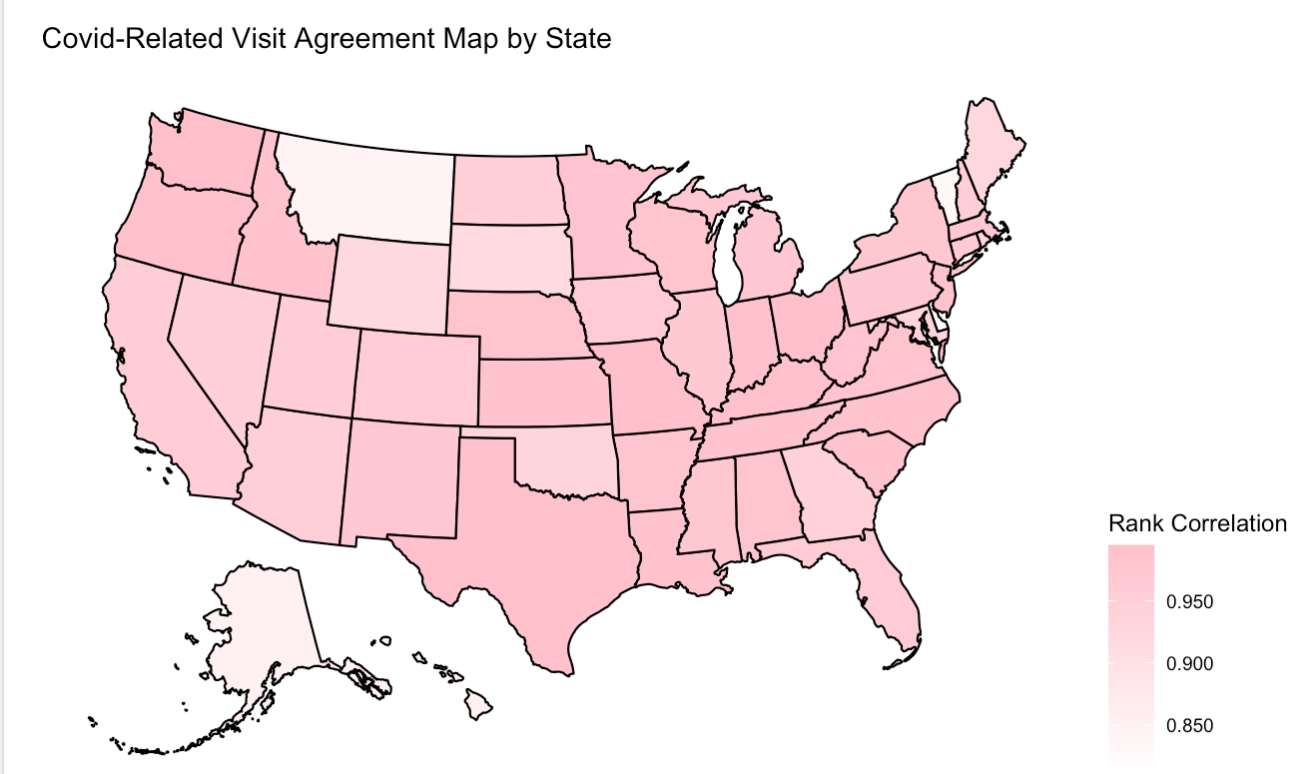
Figure 10 shows how rank correlation changes over time in these six chosen states. Although these plots look very different, the rank correlation is usually high during the outbreak. The correlation is high for all six states during the spring of 2020, which is the first COVID-19 outbreak. In California, Texas and Florida, the correlation changes similarly over time as the rank correlation arises again around the summer of 2021, which is the second outbreak caused by the Delta variant. Therefore, for this pair of signals on sore throat symptoms, we find a similar result that the agreement rate between two signals is higher during the outbreak.



**Figure 10:** Google Search vs Facebook Survey Correlation over Time

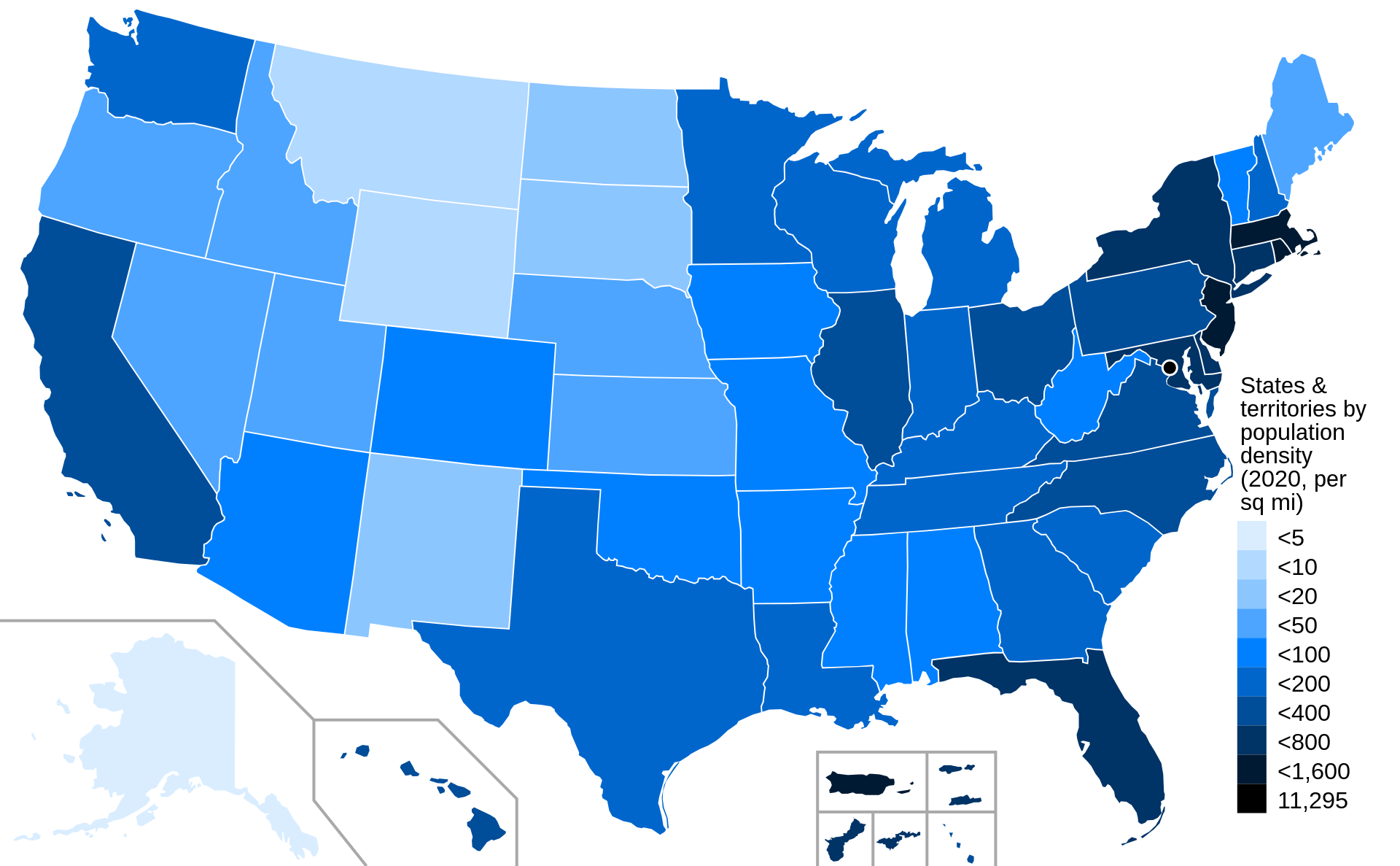
## **2.4 COVID Related Doctor Visits (CHNG) vs. Claims (Doctor-Visits)**

Figure 11 shows how the rank correlation between these two variables vary across states in the U.S.. In Figure 11, the rank correlation ranges from 0.534 to 0.970. The rank correlation is the highest in Texas (TX) and the lowest in Maine (ME). We can tell that the rank correlation is quite high as the average rank correlation is 0.869, which denotes that these data signals agree with each other quite well at state level.

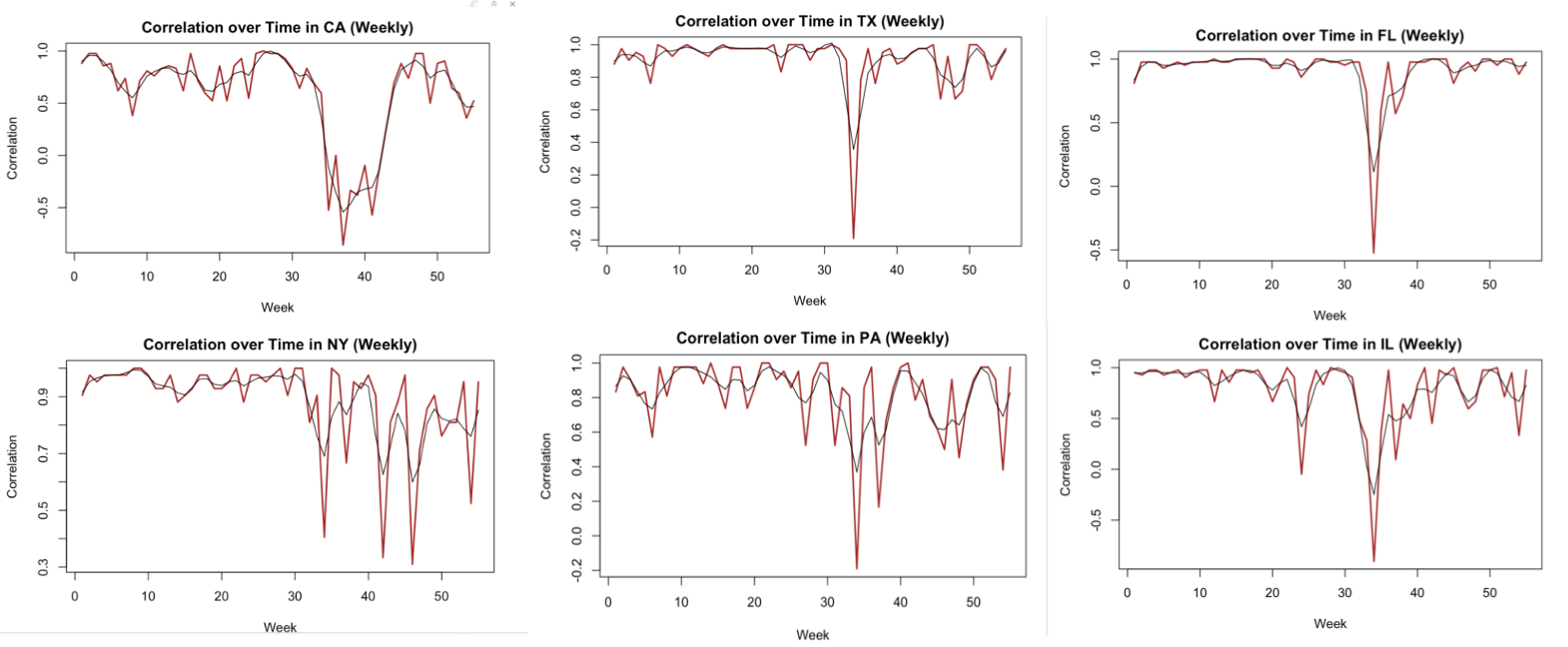


**Figure 11:** CHNG vs. Doctor-Visits COVID Related Doctor Visits Percentage Agree Map

Furthermore, we explore how population density at state level is associated with geographical rank correlation between these two variables. The darker the color indicates a higher population density in a state. Given Figure 11 and Figure 12, we find that the rank correlations are higher in southern and western states in the U.S. where population densities are also high. We also notice that the rank correlations are lower in northern states in the U.S. where population densities are also low. This indicates that there is a geographical pattern in the rank correlations between these two variables.



**Figure 12:** Population Density across States in the U.S.



**Figure 13**: CHNG vs. Doctor-Visits Weekly Rank Correlation Plots

Figure 13 above shows how the weekly rank correlations change from 1/10/2021 to 3/10/2022 for those six states denoted above. We can tell that most of the rank correlations between these variables are high in general. The rank correlations mostly range from 0.6 to 1.0 especially in Texas (TX) and Florida (FL). We also see a significant drop in rank correlations between week 30 and week 40 in general. We find some of the rank correlations within this period of time reach negative values. It could be the case that the weekly segments are more likely to capture variations in the doctor visits percentages from either data source. Moreover, the rank correlations are more sensitive to the ups and downs in the data. In this case, we would like to see negative rank correlations between these two data signals.