Base Model

For the training period, 7th March 2020 – 31st December 2020 (300 days), the binary vector of Covid-19 case anomalies, L, and the binary vector of search query anomalies $\hat{L}$, were calculated for each state. The two vectors, for each state, were then compared and matched to find the best lag-threshold combination that maximised the resulting F-score. The comparison algorithm uses unique matching whereby a Covid-19 case anomaly can only match to one symptom rate anomaly.

Using the best lag-threshold combination for each state, the number of forecasted Covid-19 case anomalies and symptom rate anomalies for the 300-day training period were plotted side by side.

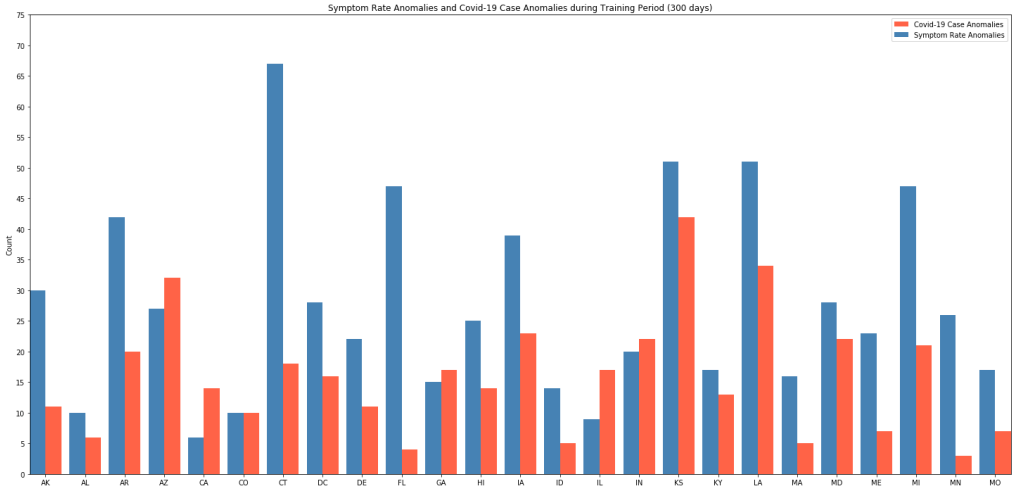

Figure 1 AK - MO

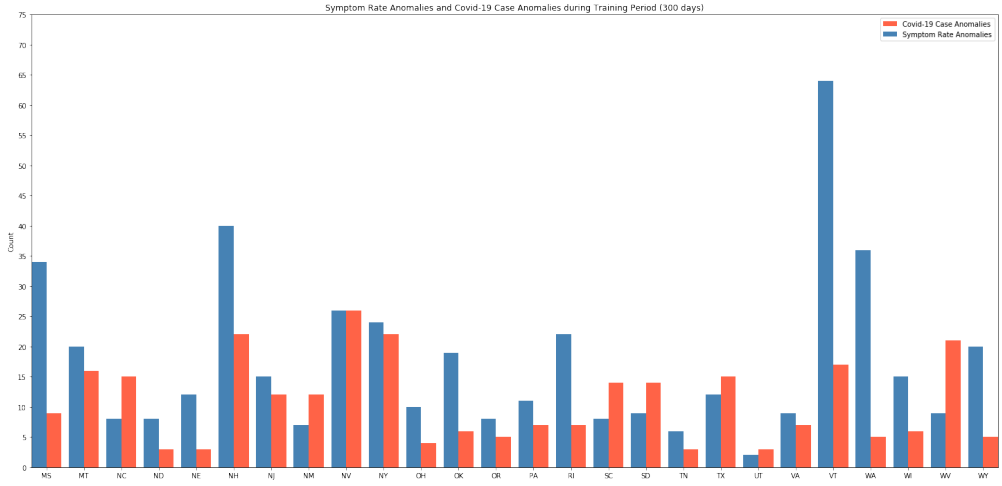

Figure 2 MS - WY

With the exception of AZ, CA, GA, IL, IN, NC, NM, SC, SD, TX and WV, all states had a higher or equal number of symptom rate anomalies than Covid-19 case anomalies.

The same was then done for the 36-day testing period (1st Jan 2021 to 5th Feb 2021) using the best lag-threshold combination for each state.

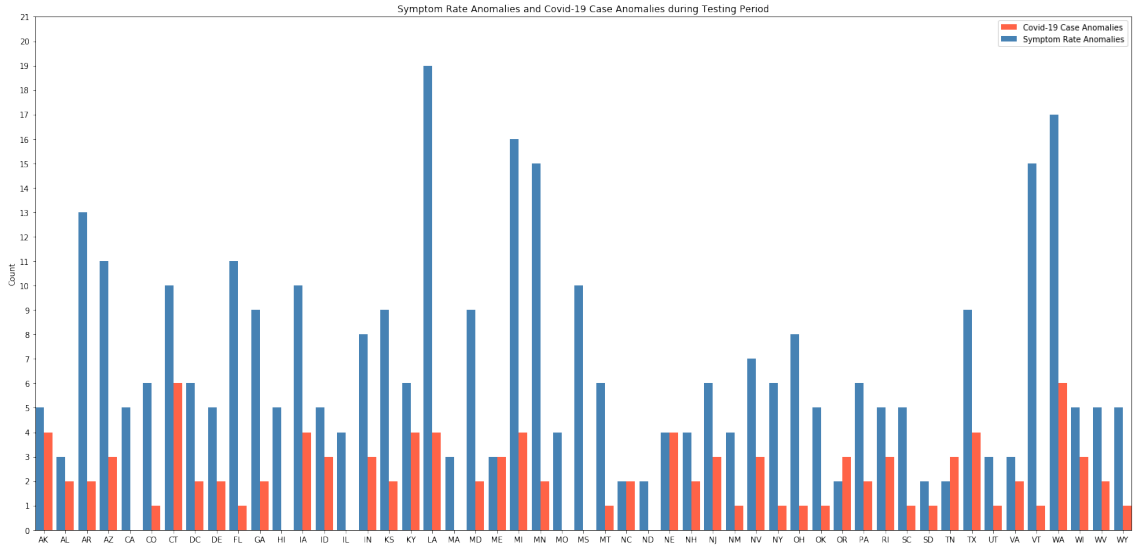

Figure 3 Outbreak Count in Testing Data

In the testing results, only OR and TN have a higher number Covid-19 case anomalies than symptom rate anomalies.

State CT

The training data for state CT has 67 symptom rate anomalies and 18 Covid-19 case anomalies. This is 272% more symptom rate anomalies than Covid-19 case anomalies. The best TH was 0.1 and best lag was 3 days.

For a lag of 3 days and threshold of 0.1, the following results were calculated:

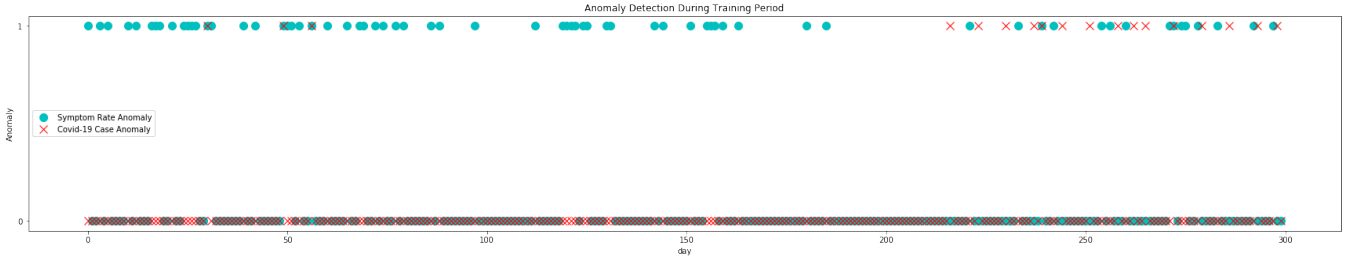

Figure 4 State CT's Anomalies (TH = 0.1, lag=3)

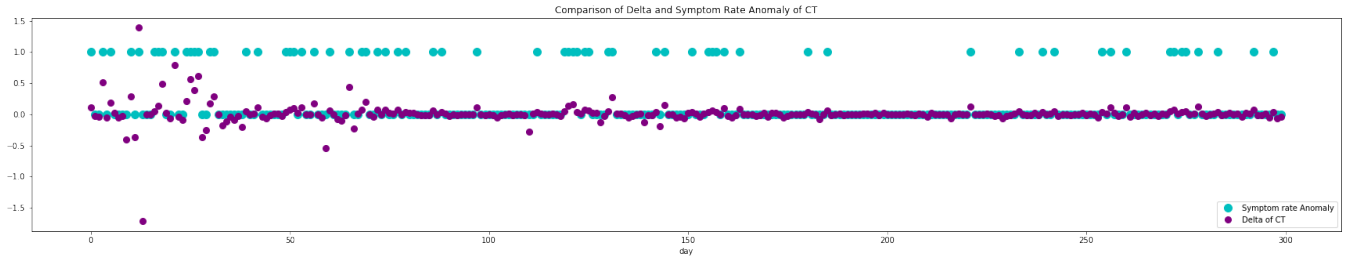

Figure 5 State CT's daily delta values compared with symptom rate anomalies (TH =0.1, lag = 3)

FP	TP
52	15
FN	**TN**
3	230

Precision = 0.224
Recall = 0.833
F-score = 0.352

Why is the best threshold not higher?

Given the same lag, when the threshold is set to 1.1, the number of symptom rate anomalies is 12. This is a more appropriate symptom anomaly forecast than 67 (TH=0.1, lag=3) however the F-score is lower (best F-score=0.352). Although the number of FP has decreased by 83%, this is offset by the decrease in number of TP (80%) and increase in FN (533%).

FP	TP
9	3
FN	**TN**
19	269

Precision = 0.25
Recall = 0.136
F-score = 0.1765

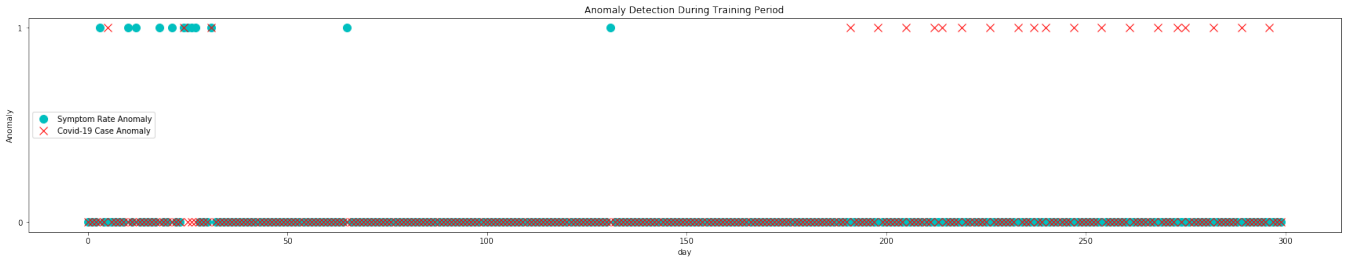

Figure 6 State CT's Anomalies (TH = 1.1, lag = 28)

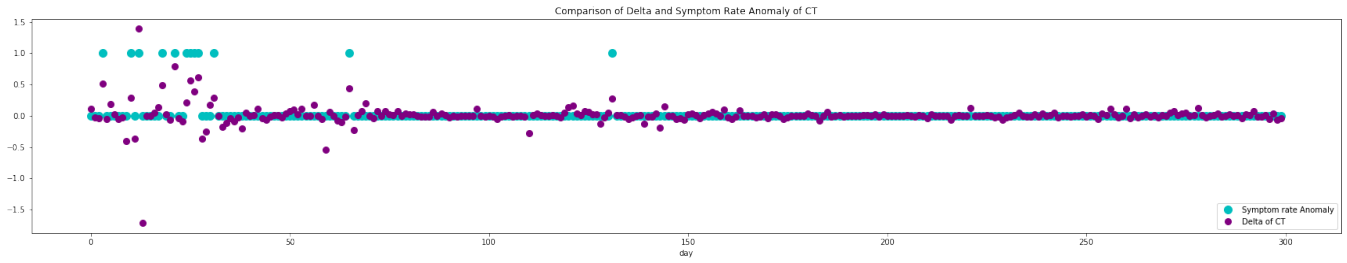

Figure 7 State CT's delta values compared with symptom rate anomalies (TH =1.1, lag = 28)

The following graph shows the F-Scores of state CT across a 0–30 day lag when different thresholds are applied. A general downward trend in F-scores (as TH increases) can be seen.

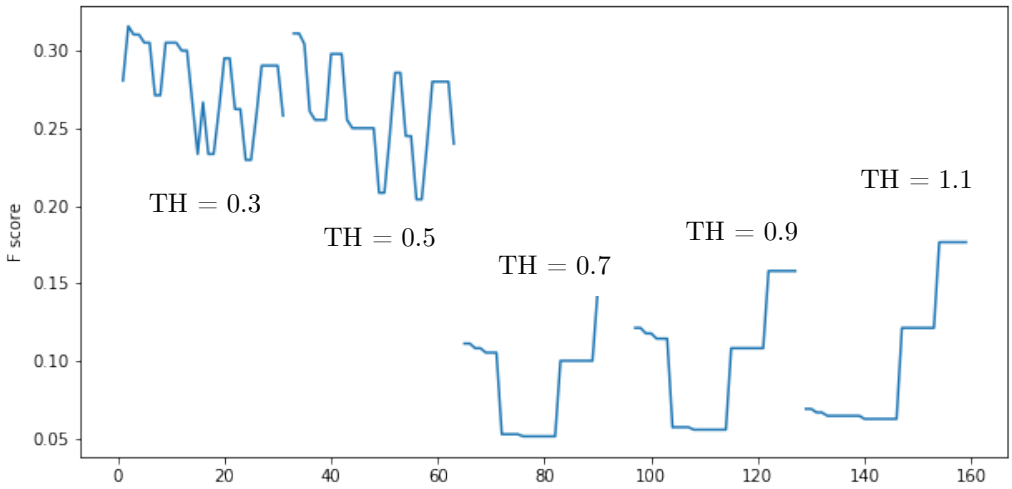

Figure 8 shows F scores (across 0-30 days lags) when TH is increased (0.3 – 1.1)

State VT

The training data for state VT has 64 symptom rate anomalies and 17 Covid-19 case anomalies. This is 276% more symptom rate anomalies than Covid-19 case anomalies. The best TH was 0.2 and best lag was 0 days.

For a best lag of 0 days and best threshold of 0.2, the following results were calculated.

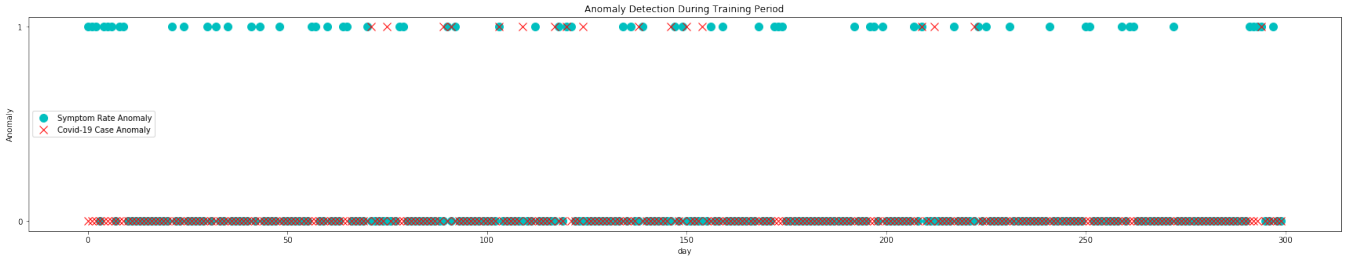

Figure 9 State VT's Anomalies (TH=0.2, lag=0)

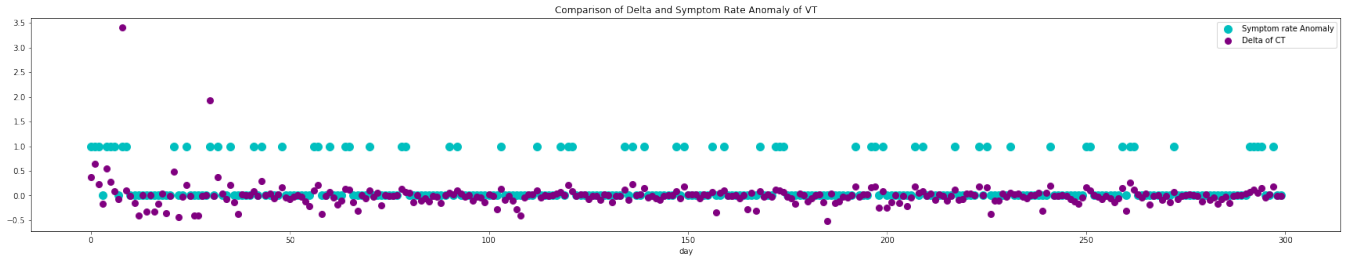

Figure 10 State VT's delta values and Symptom Rate Anomalies (TH=0.2, lag=0)

FP	TP
47	17

Precision = 0.266
Recall = 1.0
F-score = 0.420

FN	TN
0	236

State VT: TH = 1.1 and lag = 28 days:

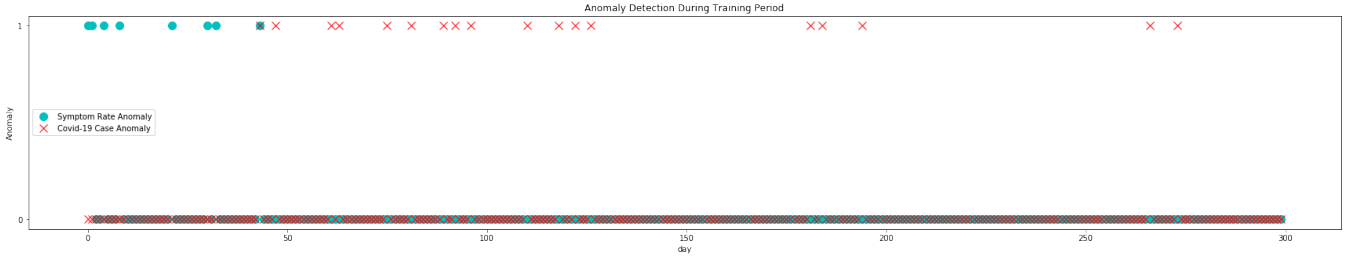

Figure 11

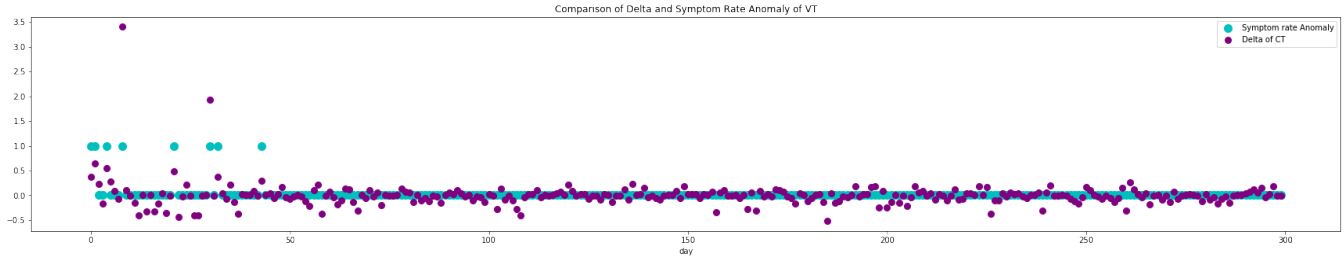

Figure 12

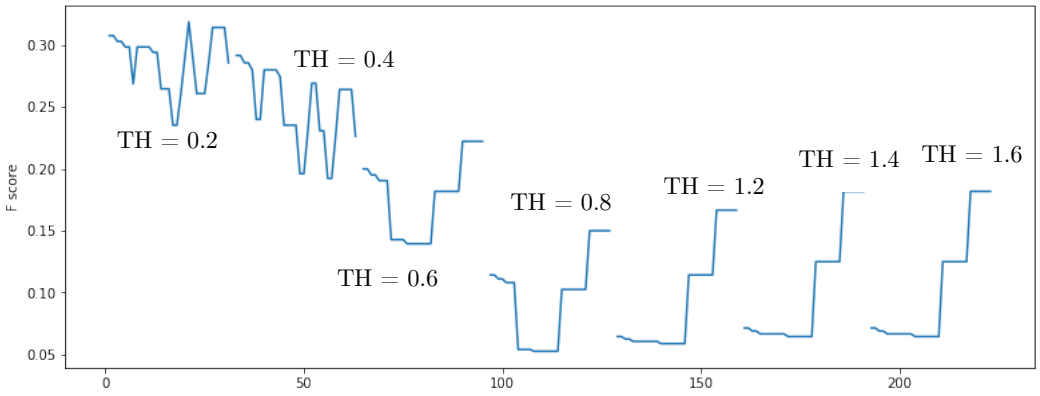

Figure 13 F scores (across 0-30 days lags) when TH is increased (0.2 – 1.6)

The comparison to find the best lag-threshold combination is repeated with the following change:

The threshold can take a value from 0-2.0 with 0.01 increments instead of 0.1 increments. This was expected to allow the model to give a more accurate definition of "symptom rate anomaly" in each state.
However, as seen in the Figure 12, the symptom rate anomalies are still over-forecasted.

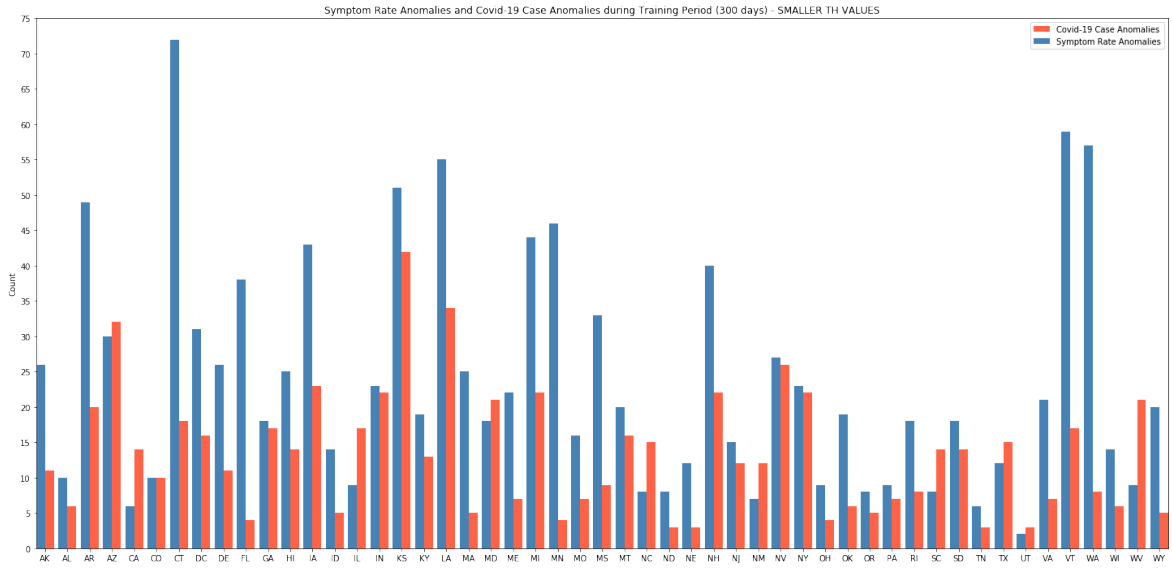

Figure 14

More importantly, as seen in the Figure 13, this change does not produce a lower symptom rate anomaly forecast (compared to the original implementation). Only 11 states see a marginal improvement in symptom rate anomaly forecasts.

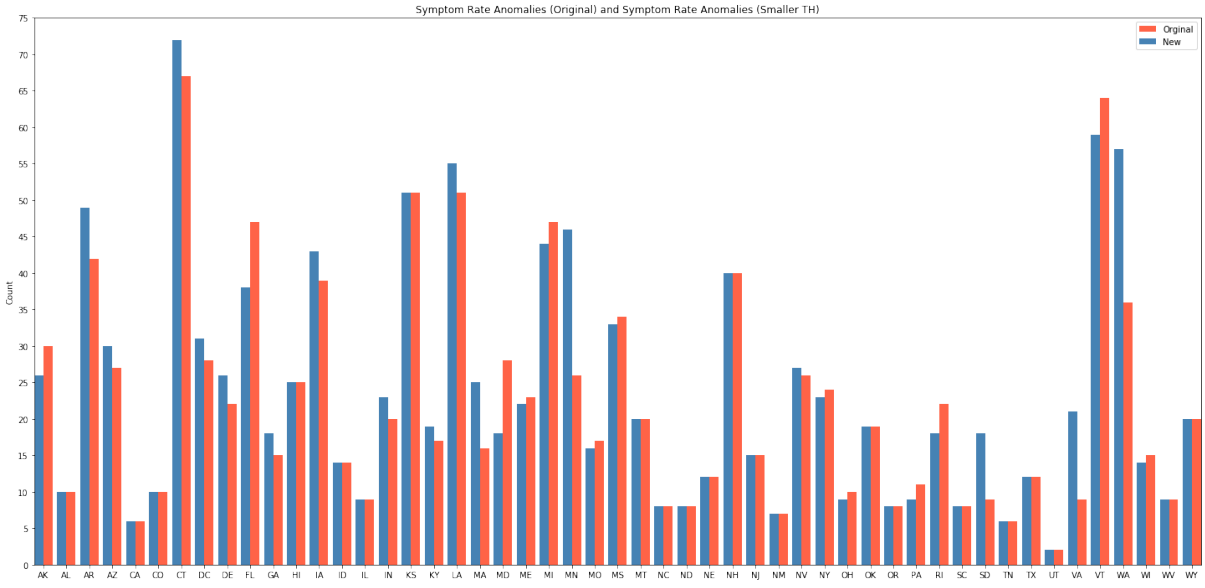

Figure 15

Thoughts

There is a lot of noise in the symptom rate data causing false positive anomalies. This could be due to search queries that are made out of concern or curiosity instead of signalling a positive Covid-19 case diagnosis. At first, I thought that the best TH had to be a larger value so that the number of symptom rate anomalies would be lowered to a more appropriate level. However when the TH was increased, the number of true positive counts decrease while false negatives increase causing the overall f-score to be lower.

Next Steps: Moving average and exponential smoothing will be applied to the symptom rate data to minimise noise and see if results can be improved.