**Moving Average Smoothing**

To minimise noise, moving average smoothing was applied to the Covid-19 Search Trend Symptoms Dataset, Q. This smoothing technique was applied to remove fine-grained variation between symptom queries each day and better expose the query signal.

A window, *w* is chosen, and the smoothing is applied as shown in equation 25. The scenario where d < w does not occur since the first recorded date in Q is 1st January 2020 however the training period only begins on 7th March 2020 therefore, there are ample days prior to 7th March 2020 to apply moving average smoothing. For each day symptom and region, we apply this smoothing technique to dataset Q which results

**Results: w = 3**

Figure 1 shows the comparison of delta, originating from Q (red) and sQ (green). The delta value is defined in equation 2.

**Shape

Description automatically generated with medium confidence**

Figure 1

After running sQ through the symptom rate outbreak prediction model, the results show that sQ does not perform better than Q. In many cases, sQ gives a higher number of symptom rate outbreaks than Q which is the opposite of what we are trying to achieve. Figure 2 and 3 shows the comparison between the number of symptom rate outbreaks which originated from the unsmoothed data Q (blue) and smooth data sQ (orange).

Chart, bar chart

Description automatically generated

Figure 2

Chart, bar chart

Description automatically generated

Figure 3

|  |  |  |  |
| --- | --- | --- | --- |
| **STATE** | **F-SCORE** | **BEST TH** | **BEST LAG** |
| AK | 0.48275862 | 0.8 | 30 |
| AL | 0.4 | 0.9 | 0 |
| AR | 0.30769231 | 0.2 | 6 |
| AZ | 0.48979592 | 0.2 | 2 |
| CA | 0.43478261 | 1.7 | 6 |
| CO | 0.47058824 | 1.3 | 0 |
| CT | 0.26086957 | 0.1 | 25 |
| DC | 0.52173913 | 1.2 | 3 |
| DE | 0.52631579 | 1.7 | 28 |
| FL | 0.12307692 | 0.2 | 16 |
| GA | 0.48387097 | 0.2 | 14 |
| HI | 0.22222222 | 0.3 | 1 |
| IA | 0.47826087 | 0.4 | 17 |
| ID | 0.20689655 | 0.6 | 10 |
| IL | 0.58064516 | 1.1 | 11 |
| IN | 0.48648649 | 0.9 | 7 |
| KS | 0.1025641 | 0 | 27 |
| KY | 0.42553191 | 0.1 | 8 |
| LA | 0.4 | 0.1 | 7 |
| MA | 0.12658228 | 0.2 | 16 |
| MD | 0.63157895 | 1.6 | 2 |
| ME | 0.33333333 | 1.1 | 5 |
| MI | 0.0952381 | 0.1 | 28 |
| MN | 0.125 | 0 | 6 |
| MO | 0.42857143 | 1.5 | 13 |
| MS | 0.35714286 | 0.7 | 8 |
| MT | 0.29885057 | 0 | 13 |
| NC | 0.63636364 | 1 | 14 |
| ND | 0.10526316 | 0.2 | 20 |
| NE | 0.28571429 | 0.9 | 5 |
| NH | 0.62295082 | 0.5 | 17 |
| NJ | 0.32786885 | 0.1 | 0 |
| NM | 0.6 | 0.9 | 20 |
| NV | 0.41176471 | 0.5 | 2 |
| NY | 0.41666667 | 0 | 0 |
| OH | 0.5 | 0.4 | 0 |
| OK | 0.22222222 | 0.6 | 8 |
| OR | 0.44444444 | 2 | 19 |
| PA | 0.19354839 | 0.4 | 30 |
| RI | 0.30769231 | 0.6 | 0 |
| SC | 0.4 | 0.7 | 29 |
| SD | 0.3 | 1.4 | 18 |
| TN | 0.25 | 1.8 | 0 |
| TX | 0.47368421 | 0.6 | 0 |
| UT | 0.33333333 | 1.4 | 0 |
| VA | 0.4 | 1.5 | 0 |
| VT | 0.31067961 | 0.3 | 2 |
| WA | 0.11764706 | 0.5 | 28 |
| WI | 0.4 | 1.8 | 0 |
| WV | 0.63829787 | 1.1 | 11 |
| WY | 0.22857143 | 0.8 | 6 |
| **AVE** | **0.36720** | **0.72941** | **10.54902** |

Table 1

The average F-score is lower than that of the base model (unsmoothed symptom rates) by 0.03. We take a closer look at state CT again to understand what is happening.

**State CT**

In the base model, CT forecasted **67 symptom rate anomalies**. With moving average smoothing applied to the symptom rate dataset, the **model forecasts 93 symptom rate anomalies**.

Figure 4 shows the occurrences of symptom rate anomalies and Covid-19 case anomalies, .

**Text

Description automatically generated**

Figure 4

Figure 5 shows and CT’s delta value, from smoothed dataset sQ.

**A picture containing line chart

Description automatically generated**

Figure 5

Figure 6 shows a comparison of from the unsmoothed dataset Q (purple) and from smoothed dataset sQ (pink).

**Chart, line chart

Description automatically generated**

Figure 6

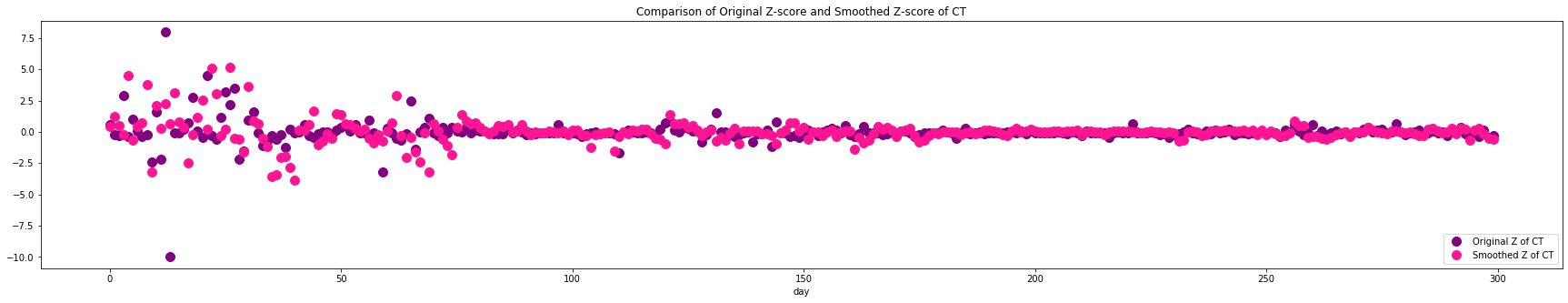
To understand why a greater number of symptom rate anomalies are being predicted, we must look at what the Z-scores look like. This is because a symptom rate anomaly is defined like so:

Therefore, there must be a higher number of Z scores in the smoothed dataset which are greater than 0.1 (the best threshold for both the smoothed and unsmoothed dataset). The following tables and figures compare the difference of Z-scores originating from the original dataset Q and smoothed dataset sQ.

|  |  |  |
| --- | --- | --- |
|  | **Q** | **sQ** |
| **Average** | 0.0106 | -0.0017 |
| **Std Deviation of** | 0.173 | 0.0514 |

Table

After the Z-scores have been calculated, we see in Figure 7 that the **smoothed Z-scores (pink) look very similar to the original Z-scores (purple).**



Figure

Figure 8 shows two histograms of Z-scores. On the left is a histogram of original Z-scores and the smoothed Z-scores on the right. The red line signifies the threshold level of 0.1.

Icon

Description automatically generated with medium confidence

Figure 8

Figure 9 shows the F-scores for different thresholds, across 0-30 day lags, using the smoothed dataset sQ.

Text

Description automatically generated

TH = 0.1

TH = 0.3

TH = 0.5

TH = 0.7

TH = 0.9

TH = 1.1

TH = 1.3

TH = 1.5

TH = 1.7

TH = 1.9

Figure 9

**Results: w = 7**

Results do not improve as the moving average window increases. When w is 7, the average F-score across the dataset is 0.334. In most states, symptom rate anomalies when w=7 is greater than symptom rate anomalies when w=3.

|  |  |  |
| --- | --- | --- |
| Original | Smoothed (w=3) | Smoothed (w=7) |
| 0.399 | 0.367 | 0.334 |

Table : Comparison of average F-scores of different models

**Chart

Description automatically generated**

Figure : Comparison of delta for different MA window