## About the Data

Symptom Rate Query data begins on 7th March 2020 and ends on 31 Dec 2020.

Daily Increase in Covid-19 Case Count data begins on 7th March 2020 and ends on 31 Jan 2021.

The Daily Increase in Covid-19 Case Count data is extended 30 days to allow for a 30-day lag period when comparing outbreaks anomalies. Both datasets begin on 7th March 2020 because it is the first day with complete data on Daily Increase in Covid-19 Case Count for each state. The symptom rate signal used is a combination of Cough and Fever symptom rates.

## Covid-19 Case Counts Contain Negative Numbers

The Daily Increase in Covid-19 Case Count data contains negative values which contradicts its description:

*“The daily increase in API field positive, which measures Cases (confirmed plus probable) calculated based on the previous day’s value.”*

To handle this, all actual case counts < 1 are assigned the value 1.

## Highlighting methodologies and definitions:

**Defining Anomalies in Covid-19 Case Counts**

To calculate outbreak anomalies in Covid-19 Case Counts, the following steps are taken:

1. Relative increase from day d-1 to day d is calculated, Cd.

This results in a vector, C, of length p.

1. C’s standard deviation is calculated.
2. Outliers are removed (C with values greater than 3 standard deviations)
3. The standard deviation of C is recalculated.
4. An outbreak anomaly is a defined as Cd that is larger than 2 standard deviations.

**Comparing Anomalies of Symptom Rate and Anomalies of Covid-19 Case Counts**

During the comparison, **strict matching** is applied whereby a symptom rate anomaly must match to a unique Covid-19 outbreak anomaly.

## Training Results: Best Threshold and Best Lag

|  |  |  |  |
| --- | --- | --- | --- |
| **STATE** | **F SCORE** | **LAG** | **TH** |
| AK | 0.34146341 | 0.7 | 26 |
| AL | 0.375 | 1.1 | 0 |
| AR | 0.35483871 | 0.1 | 12 |
| AZ | 0.54237288 | 0.4 | 0 |
| CA | 0.6 | 1.3 | 5 |
| CO | 0.5 | 0.7 | 0 |
| CT | 0.35294118 | 0.1 | 3 |
| DC | 0.45454545 | 0.8 | 0 |
| DE | 0.42424242 | 1 | 18 |
| FL | 0.15686275 | 0.2 | 29 |
| GA | 0.5 | 0.6 | 12 |
| HI | 0.30769231 | 0.8 | 10 |
| IA | 0.51612903 | 0.2 | 3 |
| ID | 0.21052632 | 1 | 10 |
| IL | 0.61538462 | 1.1 | 11 |
| IN | 0.52380952 | 0.7 | 1 |
| KS | 0.25806452 | 0 | 26 |
| KY | 0.46666667 | 0.8 | 29 |
| LA | 0.47058824 | 0.1 | 28 |
| MA | 0.19047619 | 1.1 | 16 |
| MD | 0.56 | 0.4 | 2 |
| ME | 0.26666667 | 1.2 | 6 |
| MI | 0.32352941 | 0 | 20 |
| MN | 0.13793103 | 0.1 | 0 |
| MO | 0.33333333 | 0.7 | 5 |
| MS | 0.3255814 | 0.4 | 4 |
| MT | 0.5 | 0.9 | 6 |
| NC | 0.69565217 | 1.9 | 0 |
| ND | 0.18181818 | 1.7 | 4 |
| NE | 0.26666667 | 1.1 | 26 |
| NH | 0.58064516 | 0.6 | 14 |
| NJ | 0.44444444 | 1.3 | 0 |
| NM | 0.52631579 | 1 | 14 |
| NV | 0.53846154 | 0.5 | 1 |
| NY | 0.47826087 | 0.5 | 0 |
| OH | 0.28571429 | 0.2 | 0 |
| OK | 0.32 | 1.1 | 28 |
| OR | 0.46153846 | 1.9 | 19 |
| PA | 0.33333333 | 1.3 | 6 |
| RI | 0.34482759 | 0.6 | 9 |
| SC | 0.45454545 | 1.1 | 30 |
| SD | 0.52173913 | 2 | 10 |
| TN | 0.22222222 | 1.9 | 0 |
| TX | 0.59259259 | 0.5 | 1 |
| UT | 0.4 | 1.3 | 0 |
| VA | 0.375 | 2 | 0 |
| VT | 0.41975309 | 0.2 | 0 |
| WA | 0.14634146 | 0.2 | 13 |
| WI | 0.47619048 | 1.3 | 5 |
| WV | 0.4 | 0.8 | 10 |
| WY | 0.32 | 1.2 | 3 |
| **AVE** | **0.39990** | **0.8** | **9** |

**Under these conditions, the best threshold (TH) and lag is found for each state using the F-score**

The **average best lag is 9 days**. The F-Scores highlighted in red are below the average F-score for all 51 regions. If we consider these datapoints to be unreliable, we can find that the **average best lag** **is 8 days and the TH is 0.89**.

## Test Results for Base Model

Symptom Rate Query data for testing begins on 1st January 2021 and ends on 5th February 2021.

Daily Increase in Covid-19 Case Count data begins on 1st January 2021 and ends on 3rd March 2021.

Chart, bar chart

Description automatically generated

|  |  |
| --- | --- |
| **STATE** | **F Score** |
| AK | 0.666666667 |
| AL | 0.4 |
| AR | 0.266666667 |
| AZ | 0.428571429 |
| CA | 0 |
| CO | 0.285714286 |
| CT | 0.375 |
| DC | 0.5 |
| DE | 0.285714286 |
| FL | 0.166666667 |
| GA | 0.363636364 |
| HI | 0 |
| IA | 0.428571429 |
| ID | 0.25 |
| IL | 0 |
| IN | 0.363636364 |
| KS | 0.363636364 |
| KY | 0.6 |
| LA | 0.347826087 |
| MA | 0 |
| MD | 0.363636364 |
| ME | 0.333333333 |
| MI | 0.4 |
| MN | 0.235294118 |
| MO | 0 |
| MS | 0 |
| MT | 0.285714286 |
| NC | 0.5 |
| ND | 0 |
| NE | 0.75 |
| NH | 0.333333333 |
| NJ | 0.222222222 |
| NM | 0.4 |
| NV | 0.6 |
| NY | 0.285714286 |
| OH | 0.222222222 |
| OK | 0.333333333 |
| OR | 0.4 |
| PA | 0.25 |
| RI | 0.5 |
| SC | 0 |
| SD | 0.666666667 |
| TN | 0.4 |
| TX | 0.615384615 |
| UT | 0.5 |
| VA | 0.4 |
| VT | 0.125 |
| WA | 0.52173913 |
| WI | 0.25 |
| WV | 0.285714286 |
| WY | 0.333333333 |
| **AVE** | **0.32559** |

States where F Score is 0 is caused by True Positive value is 0.

## False Anomalies

A potential weakness is that this allows for changes in Covid-19 case counts, whereby the absolute value of those case counts is small, to be categorized as an “outbreak anomaly”. We categorise anomalies with absolute case count values of less than 50 as “false anomalies”.

Chart, bar chart, line chart

Description automatically generated

For states Alaska (AK), Hawaii (HI), Maine (ME), Montana (MT), North Dakota (ND), New Mexico (NM), Oregon (OR), Vermont (VT) and West Virginia (WV), the false anomalies account for at least 50% of the Covid-19 Case anomalies within the training period.