**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.

## Actual 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.

## Before Removing Outliers from Actual Case Counts

**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 an vector, C, of length p.

1. C’s standard deviation is calculated.
2. An outbreak anomaly is a defined as Cd that is larger than 2 standard deviations.

However, 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”. Outliers are also not removed from the data.

States such as Vermont (VT) show that there 108 days that have been identified to have experienced Covid-19 Case outbreak anomaly however the absolute value of new cases is below 50 (0.008% of the population in Vermont).

Chart, histogram

Description automatically generated

**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.

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

|  |  |  |
| --- | --- | --- |
| **STATE** | **Best F-SCORE** | **BEST LAG** |
| **AK** | 0.81 | 0 |
| **AL** | 0.69 | 11 |
| **AR** | 0.63 | 20 |
| **AZ** | 0.58 | 23 |
| **CA** | 0.63 | 29 |
| **CO** | 0.64 | 19 |
| **CT** | 0.77 | 0 |
| **DC** | 0.86 | 25 |
| **DE** | 0.72 | 25 |
| **FL** | 0.84 | 29 |
| **GA** | 0.76 | 26 |
| **HI** | 0.84 | 13 |
| **IA** | 0.70 | 27 |
| **ID** | 0.68 | 18 |
| **IL** | 0.51 | 27 |
| **IN** | 0.67 | 27 |
| **KS** | 0.52 | 9 |
| **KY** | 0.65 | 24 |
| **LA** | 0.75 | 0 |
| **MA** | 0.80 | 3 |
| **MD** | 0.74 | 11 |
| **ME** | 0.76 | 12 |
| **MI** | 0.46 | 11 |
| **MN** | 0.55 | 26 |
| **MO** | 0.66 | 0 |
| **MS** | 0.84 | 8 |
| **MT** | 0.68 | 5 |
| **NC** | 0.89 | 25 |
| **ND** | 0.76 | 26 |
| **NE** | 0.76 | 19 |
| **NH** | 0.83 | 30 |
| **NJ** | 0.40 | 0 |
| **NM** | 0.60 | 26 |
| **NV** | 0.65 | 30 |
| **NY** | 0.43 | 29 |
| **OH** | 0.72 | 30 |
| **OK** | 0.81 | 3 |
| **OR** | 0.69 | 12 |
| **PA** | 0.67 | 0 |
| **RI** | 0.69 | 27 |
| **SC** | 0.69 | 3 |
| **SD** | 0.75 | 27 |
| **TN** | 0.54 | 28 |
| **TX** | 0.63 | 23 |
| **UT** | 0.41 | 28 |
| **VA** | 0.65 | 28 |
| **VT** | 0.90 | 16 |
| **WA** | 0.50 | 20 |
| **WI** | 0.75 | 0 |
| **WV** | 0.63 | 13 |
| **WY** | 0.86 | 0 |
| **AVERAGE** | **0.685626327** | **17.07843137** |

The **best TH is always 0** and the **average best lag is 17 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 15 days**.

When TH increases, the number of symptom rate anomalies decreases within the period therefore:

* Number of **True Positives** decreases significantly.
* Number of **False Positives** decreases to zero.
* Number of **False Negatives** increases significantly

**When TH is increased from 0 to 1.0**

**TP** showed an average **84% decrease.**

**FP** decreased to 0 in 94% of cases.

**FN** showed **52% average increase.**

This trend causes the F-score to be highest when threshold is 0.

## After Removing Outliers from Actual Case Counts