# Outlier Detection in Dengue Fever Rates

Research carried out by Liam Reid as part of Final Year Project and Dissertation: ‘Proposing new methods of detecting outliers in cloud resource data’.

## Abstract

The aim of this experiment is to test the effectiveness of a newly implemented technique of detecting outliers. This technique involves detecting outliers using an ensemble of ‘weak’ classifiers that work together and vote on whether a datapoint is an outlier or an inlier. This technique is tested on Dengue Fever rates in regions of Vietnam. Observations of generated graphs show that this technique is effective in detecting outliers. Obvious outliers and some subtle outliers can be detected using this method but on rare occasions an outlier is missed and there are many false alarms. A comparison with a traditional classifier (KNN) proves that this method of detecting outliers is of good standard.

## Introduction

**Background Information**

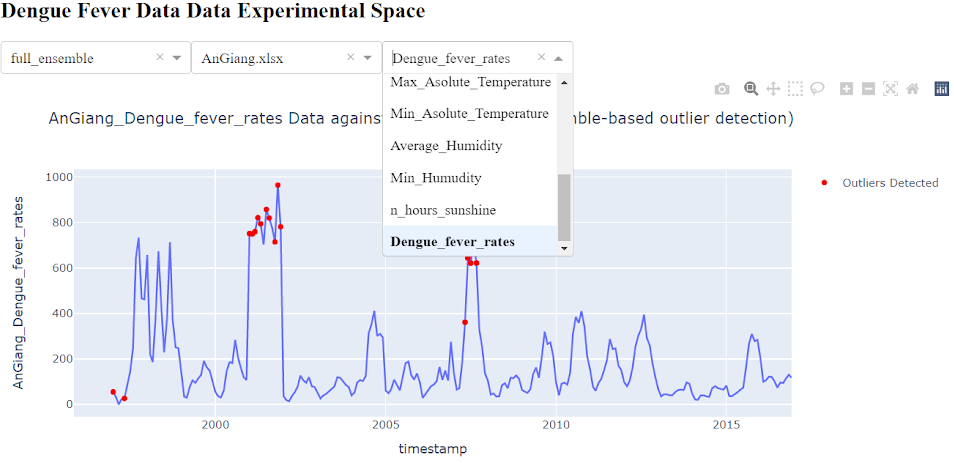
This document contains details of newly implemented outlier detection techniques on Dengue fever rates in the Bac Lieu and An Giang regions of Vietnam.

The Dengue Fever is a viral disease that is carried by mosquitos. It is widespread throughout tropical regions. Local environmental factors have an impact of the spread and severity of this virus and it is leading cause of hospitalisation and deaths in the areas that it affects [1].

The data consists of details about the climate of these regions over a 20-year period. Data such as average temperature, rainfall, humidity and the fever rate itself will be considered in this experiment. The data is unlabelled. The techniques used will be unsupervised and will not require any training. The ultimate goal of this experiment is to run outlier detection on these datasets and see if the outliers correlate between different datasets, if they do, then the outlier detection is working.

**How is the experiment carried out?**

This experiment is carried using a webapp developed using dash. The purpose of this application is to apply outlier detection algorithms to datasets defined by the user. The application plots the data and generate scores for the chosen method. A score is not generated for the detection methods on these datasets since they are unlabelled.



Screenshot of application used to perform detection and generate graphs.

The outlier detection methods in the ensemble are implemented using python. These methods work individually first to make a prediction with a confidence score. A voting system, also implemented using python, determines the final classification.

## Hypothesis

The implemented outlier detection method (ensemble) is an effective outlier detector for unlabelled datasets.

## Methods

The combined predictions of an ensemble of weak classifiers are used as the method to detect outliers in this experiment. These detectors generate a prediction for a piece of data (outlier or inlier), the combined predictions are combined to produce a final classification.

These techniques are unsupervised do not require training. This is beneficial for analysing this data since it is unlabelled [2]. Evaluating the performance of these techniques will be difficult since there is nothing to compare the classifications against.

A quick observation shows that the time of year has an effect on the data. I.e., average temperature during summer months is higher than during winter months. To account for this, these datasets are split to months before the detection is performed. This means there will be no false positive predictions made when the time series moves from summer to winter, but the idea of concept drift is still considered year on year (due to factors like climate change) [3], [4].

There are 4 classification techniques in the Ensemble.

* **Moving Average**

This technique uses the average of the previous data points in the time series to classify the next. After the average is calculated, the standard deviation of the previous datapoints is calculated . The standard deviation is used as a threshold, if they next datapoint is less than or greater than the average calculated +/- the threshold then the datapoint is classified as an outlier [5]. The graph shows this technique in practice, the red lines represent the boundaries and the red dots are the outliers detected.

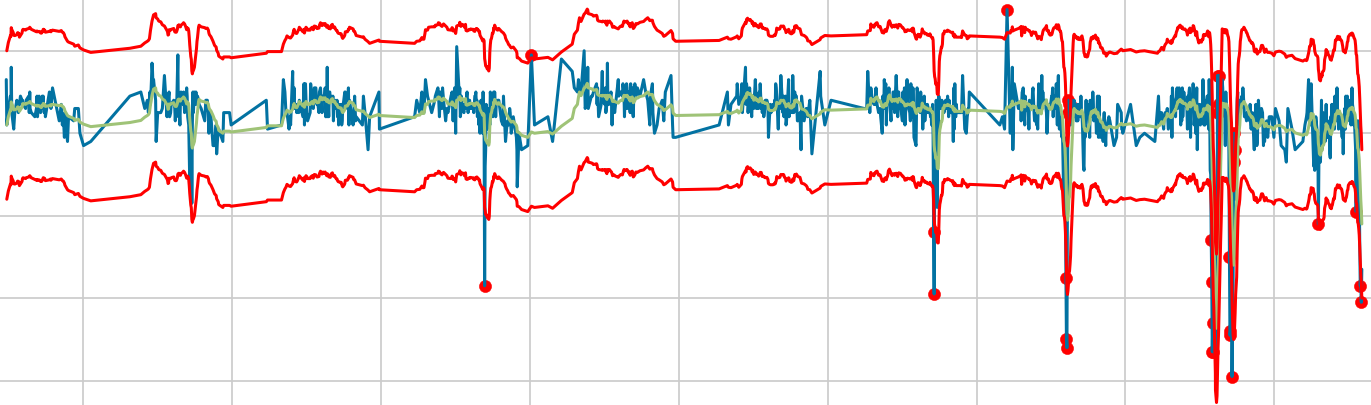


Fig. 1 Moving Average outlier detection showing boundaries

* **Moving Median**

This technique follows the same steps as the previous except a median is calculated instead of the average [6]. The graph below shows this technique in practice, observe how the boundaries are similar to moving average but different outliers have been detected.

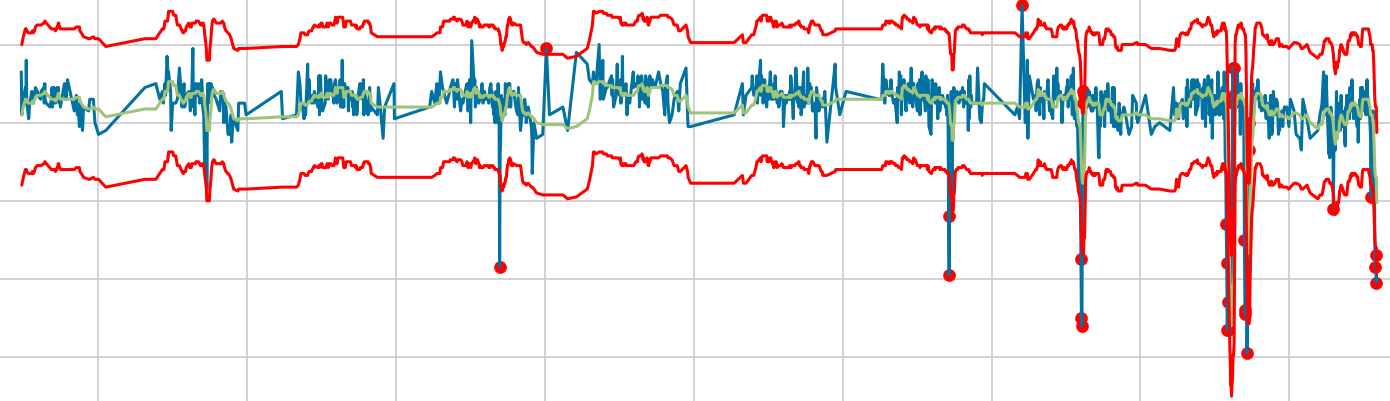


Fig. 2 Moving Median outlier detection showing boundaries

* **Moving Boxplot**

This technique takes a number of the previous data points and generates a boxplot. The interquartile range is combined with the upper and lower quartiles to produce a threshold (1.5 \* the inter-quartile range). If the next datapoint is outside the threshold then the datapoint is classified as an outlier [7].

1.5 \* IQR (Upper bound)

1.5 \* IQR (Lower bound)

3rd Quartile

1st Quartile

Fig. 3 Boxplot Outlier Detection Example

* **Histogram**

This technique plots histograms of subsets of the data. If a range in the histogram has a height less than a defined threshold then the range is said to contain outliers [8]. If a range has a height below the threshold, but the ranges beside it have a height higher than the threshold then it is considered a borderline inlier.

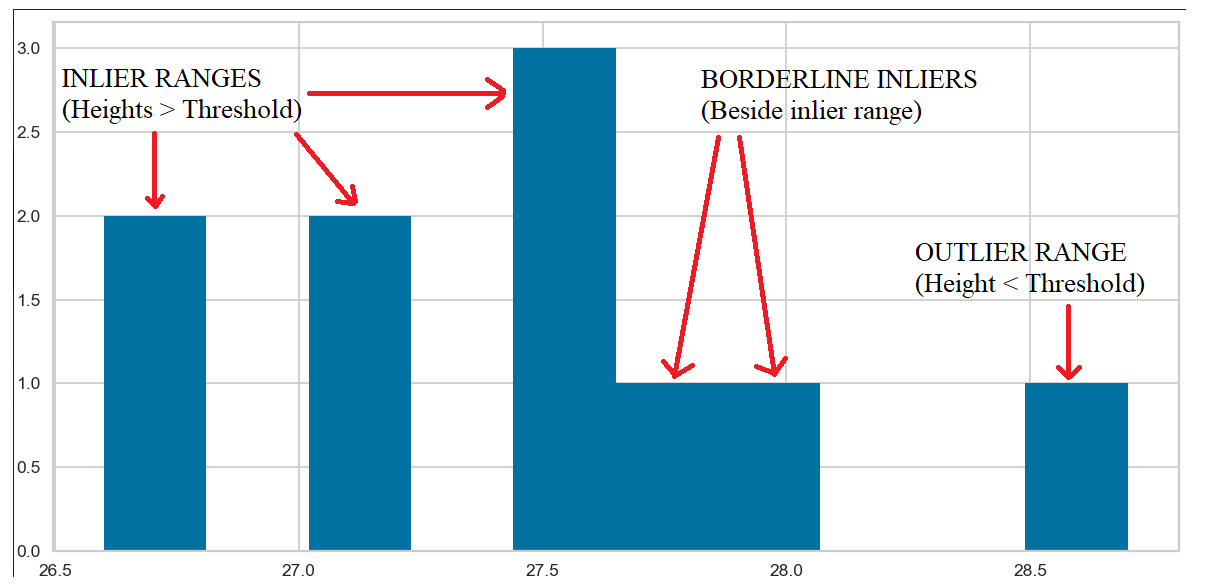
****

Fig. 4 Histogram Based Outlier Detection

**Voting on a Final Classification**

These techniques run individually first, generating a prediction and a ‘confidence’ score. The confidence scores are combined to generate a final prediction. The formula behind this voting mechanism is described below.

Confidence is calculated by the distance between a datapoint and the threshold.

For predictions, -1(outlier) and 1(inlier), the above equation computes a minimum prediction of -n and a maximum of n. By visualising possible outputs on a spectrum, it can be said that an outlier score < 0 is likely to be an actual outlier.

N (max inlier)

-N (min outlier)

0

## Results – An Giang Data

Graphs generated when applying the outlier detection algorithm on the An Giang data. Red dots indicate outliers detected.

|  |  |
| --- | --- |
| Graphical user interface, chart  Description automatically generated | Graphical user interface  Description automatically generated |
| I. Detection result for An Giang Dengue Fever Rate | II. Detection result for An Giang Average Humidity |
|  |  |
| Graphical user interface  Description automatically generated | Chart  Description automatically generated |
| III. Detection result for An Giang Total Rainfall | IV. Detection result for An Giang No. Raining Day |
|  |  |
| A picture containing different, various, colorful, several  Description automatically generated | Graphical user interface, application  Description automatically generated |
| V. Detection result for An Giang Average Temperature | VI. Detection result for An Giang Total Evaporation |

**Evaluation of An Giang Data**

The graphs above show that outliers have been detected in peaks and troughs throughout the time series, indicating that (possible) actual outliers have been detected. Most notably, a cluster of detections were made in graph I in the spike between 2000 and 2005. This correlates with the outlier detected in graph III indicating rainfall caused a spike in the fever rate. More notably, outliers were indicated in graph IV between 2010 and 2015 where there seems to be no raining days. In graph III, a spike in rainfall is detected, which correlates with the spike in evaporation detected in graph VI around the same time.

A great number of detections have been made in graph V compared with the rest, the detection technique may be ineffective against such unstable data. But a cluster of detections are shown in the 2005 to 2010 period. There is a trough in the time series which correlates with the detection made in graph II where an outlier is detected in average humidity. The outlier detected here does not look irregular when plotted since it is not major peak or a trough but could be expected to be an actual outlier because of the irregular temperature.

## Results – Bac Lieu Data

Graphs generated when applying the outlier detection algorithm on the Bac Lieu data. Red dots indicate outliers detected.

|  |  |
| --- | --- |
|  |  |
| I. Detection result for Bac Lieu Dengue Fever Rate | II. Detection result for Bac Lieu Average Humidity |
|  |  |
|  |  |
| III. Detection result for Bac Lieu Total Rainfall | IV. Detection result for Bac Lieu No. Raining Day |
|  |  |
|  |  |
| V. Detection result for Bac Lieu Average Temperature | VI. Detection result for Bac Lieu Total Evaporation |

**Evaluation of Bac Lieu Data**

The graphs show that more outliers are detected here compared with the An Giang data, especially in graph V. Besides this, the outliers detected are in the peaks and troughs of the data. Detections were made around the spikes in fever rates in graph I, but the detector is failing to correctly classify the top of some peaks.

Similarly to graph V in the An Giang Data, the detector has marked many data points as outliers, solidifying the fact that this detector may be ineffective against unstable data. Besides graph V there are a number of correlations between the graphs. The initial spike in dengue fever rates in graph I correlate with the average humidity in graph II and the drop in no. raining days in graph IV. These correlations are marked as outliers by the detectors. Again, the trough in graph IV is picked up by the detector as well as a sudden spike in total rainfall in graph III.

## Comparison with a Traditional Classification Technique – KNN

To test the effectiveness of this newly implemented ensemble, KNN outlier detection has been applied to the datasets [9].

## Results – KNN Outlier Detection on An Giang Data

Graphs generated when applying KNN on the An Giang data. Red dots indicate outliers detected.

|  |  |
| --- | --- |
|  |  |
| I. Detection result for An Giang Dengue Fever Rate | II. Detection result for An Giang Average Humidity |
|  |  |
|  |  |
| III. Detection result for An Giang Total Rainfall | IV. Detection result for An Giang No. Raining Day |
|  |  |
|  |  |
| V. Detection result for An Giang Average Temperature | VI. Detection result for An Giang Total Evaporation |

## Results – KNN Outlier Detection on Bac Lieu Data

Graphs generated when applying KNN on the Bac Lieu data. Red dots indicate outliers detected.

|  |  |
| --- | --- |
|  |  |
| I. Detection result for Bac Lieu Dengue Fever Rate | II. Detection result for Bac Lieu Average Humidity |
|  |  |
|  |  |
| III. Detection result for Bac Lieu Total Rainfall | IV. Detection result for Bac Lieu No. Raining Day |
|  |  |
|  |  |
| V. Detection result for Bac Lieu Average Temperature | VI. Detection result for Bac Lieu Total Evaporation |

## Comparison of Results

KNN failed to detect some of the obvious outliers (peaks/troughs) that are detected by the ensemble method. In graph IV of the An Giang data, the ensemble method correctly identifies outliers in the 3-4 year stretch where there were no raining days, but KNN fails to detect this data as anomalous. Similarly, KNN detects two spikes in dengue fever rates in the An Giang region and misses a major peak in around 2007, the ensemble technique detects this spike.

KNN performs better for graph V in both regions. The ensemble method detects a large number of outliers whereas KNN detects a few in areas of the graphs that (appear to be) actual outliers. KNN appears to detect the top of peaks better than the ensemble method.

Something important to note when comparing these results is that the ensemble method takes much less time to run than KNN. It took KNN ~6 seconds to process each dataset and it took the ensemble ~1 second using the same hardware.

## Conclusion

The ensemble method of detecting outliers is effective in detecting outliers in unlabelled datasets. It is difficult to say to what extent it is effective since accuracy, precision, recall and f1 are impossible to calculate without labels. But, by comparing this method with a traditional classifier, and observing peaks and troughs within the datasets, it can be said that the ensemble is detecting outliers in the correct places. For most datasets the ensemble appears to be performing the same or even better than the traditional KNN detector but it is clear that the ensemble needs more work around unstable datasets. Another important thing to note is that with similar results, the ensemble performs detection up to 6x faster than KNN.

## References

[1] World Health Organization. (2022, Jan. 10). *Dengue and Severe Dengue* [Online]. Available: <https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue>

[2] Campos, G.O., Zimek, A., Sander, J. et al. (2016). *On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study*. Data Min Knowl Disc 30, 891–927 <https://doi.org/10.1007/s10618-015-0444-8>

[3] Tsymbal, Alexey. (2004). *The problem of concept drift: definitions and related work*. Computer Science Department, Trinity College Dublin 106.2: 58. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.58.9085&rep=rep1&type=pdf>

[4] Naish, S., Dale, P., Mackenzie, J.S. et al. (2014). *Climate change and dengue: a critical and systematic review of quantitative modelling approaches*. BMC Infect Dis 14, 167. <https://doi.org/10.1186/1471-2334-14-167>

[5] Dr. Dataman. (2021, Apr, 18). *Anomaly Detection for Time Series. (1) Simple Moving Average.* [Online]. Available: [https://medium.com/dataman-in-ai/anomaly-detection-for-time-series-a87f8bc8d22e](%20https:/medium.com/dataman-in-ai/anomaly-detection-for-time-series-a87f8bc8d22e)

[6] Anomaly. (2016, Jan. 12). *Detecting Anomalies with Moving Median Decompsition*. [Online]. Available: <https://anomaly.io/anomaly-detection-moving-median-decomposition/index.html>

[7] Andrea, Kliton & Shevlyakov, Georgy & Smirnov, Pavel. (2013). *Detection of outliers with boxplots*. 141-144. <https://www.researchgate.net/publication/261173084_Detection_of_outliers_with_boxplots>

[8] Goldstein, Markus & Dengel, Andreas. (2012). *Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm*. Available: <https://www.goldiges.de/publications/HBOS-KI-2012.pdf>

[9] Harrison, O. (2018, Sep. 10). *Machine Learning Basics with the K-Nearest Neighbors Algorithm.* [Online]. Available: <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>