# Overview

This report serves as a brief introduction to some fundamental methods for anomaly detection with the TROPOMI data. We will cover four separate methods for detection of anomalous methane mixing ratios across the contiguous United States. The time range considered is from 01 December 2018 to 31 March 2019 and the data is pulled from GES DISC with TROPOMI's low-resolution imaging.

A visualization of the mean methane ratio over the aforementioned time period is provided for reference:

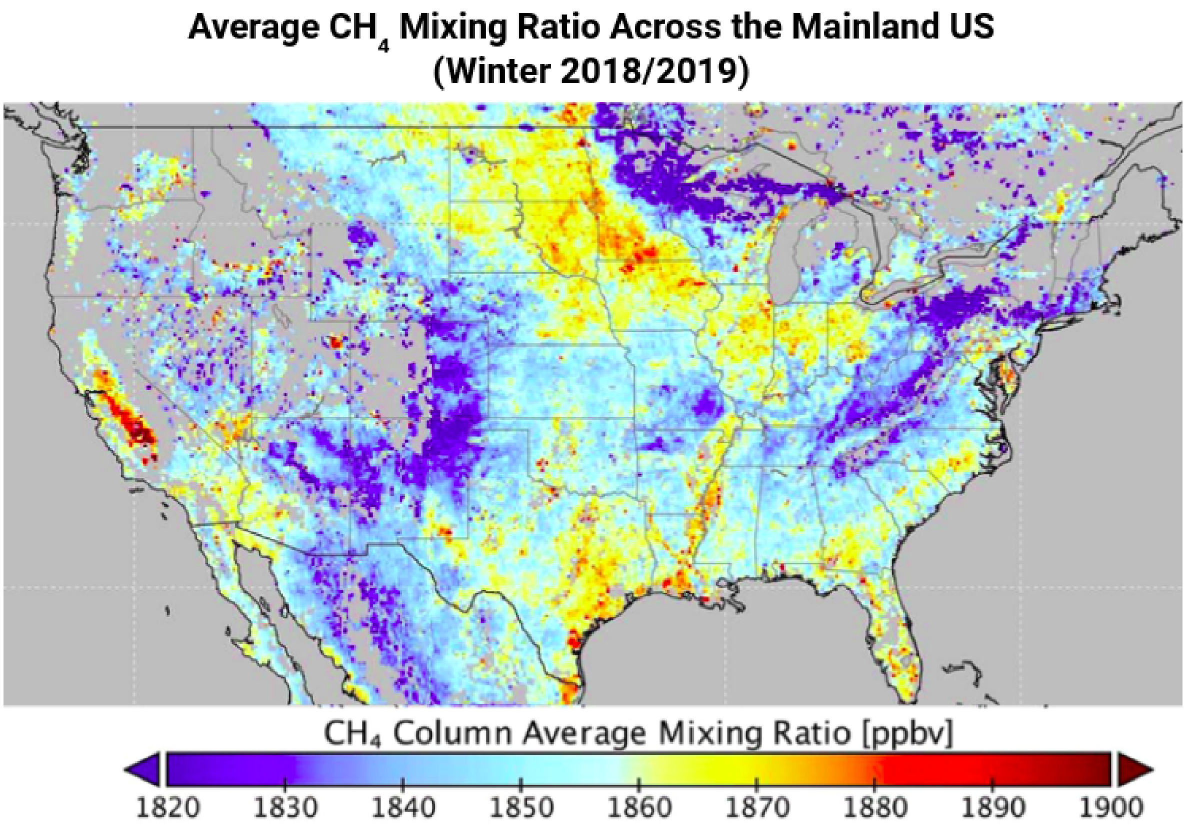
Average CH4 Mixing Ratio Across the Mainland US
(Winter 2018/2019)
CH4 Column Average Mixing Ratio [ppbv]
1820
1830
1840
1850
1860
1870
1880
1890
1900

Figure 1: Mean Methane Mixing Ratio between 01 December 2018 and 31 March 2019.

We will re-use this visualization to show where different methods detect anomalies in the methane mixing ratio. The following sections of this report describe some potential methods for anomaly detection.

# 1 The Bootstrap

This method requires that we grid the region into small squares (subregions). For each of these subregions, we create a histogram of the response and use the bootstrap method to approximate the distribution of the response in this subregion. If there are sufficiently heavy tails (e.g., conforming to a Pareto distribution), we have found an anomaly! Otherwise, we have found no anomaly.

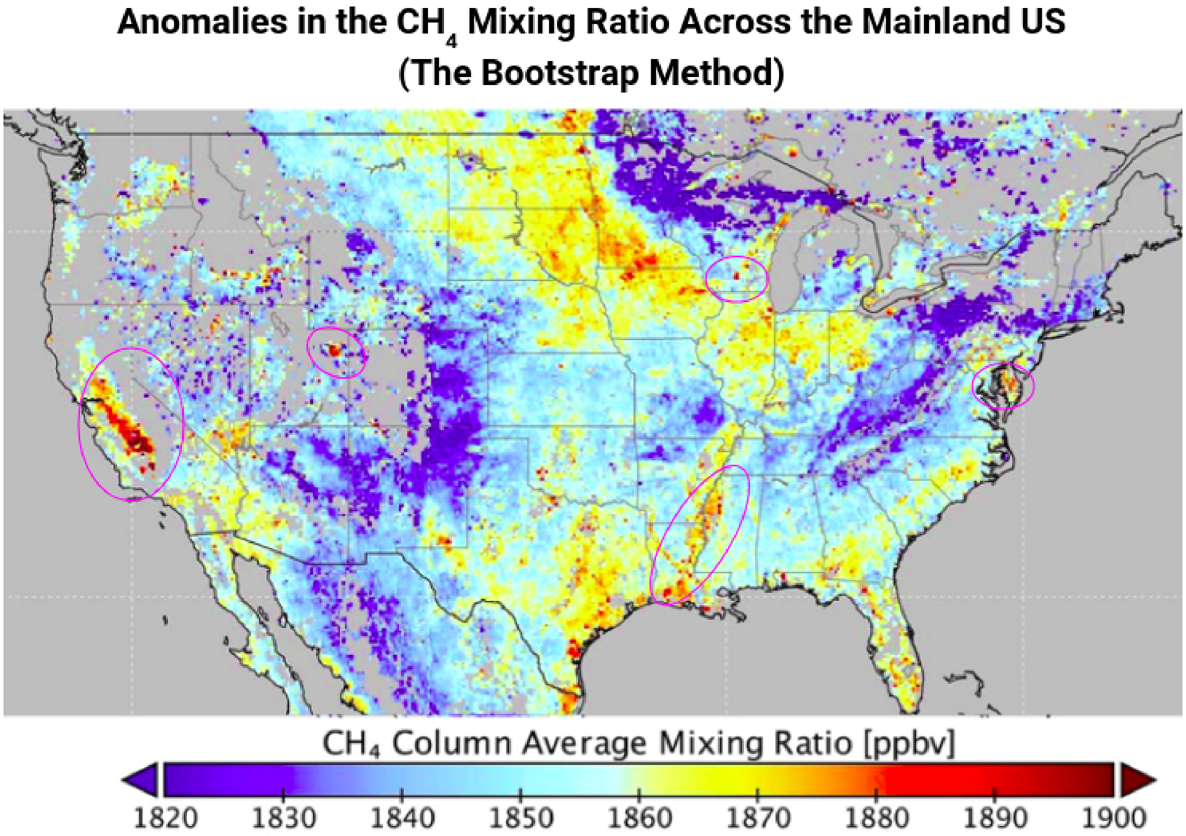
Anomalies in the CH4 Mixing Ratio Across the Mainland US
(The Bootstrap Method)
CH4 Column Average Mixing Ratio [ppbv]
1820
1830
1840
1850
1860
1870
1880
1890
1900

Figure 2: Bootstrap Anomalies on the Map.

This method highlights the fairly "obvious" results and misses some of the more subtle results. It is valuable mostly as a common-sense check. It tends toward Type II error.

## 2 The Local Outlier Factor

This method works by removing all of the data corresponding to response values that are within one standard deviation of the mean. The remaining response values are plotted on a spatial map and the points of unusually high or low density are classified as anomalies.

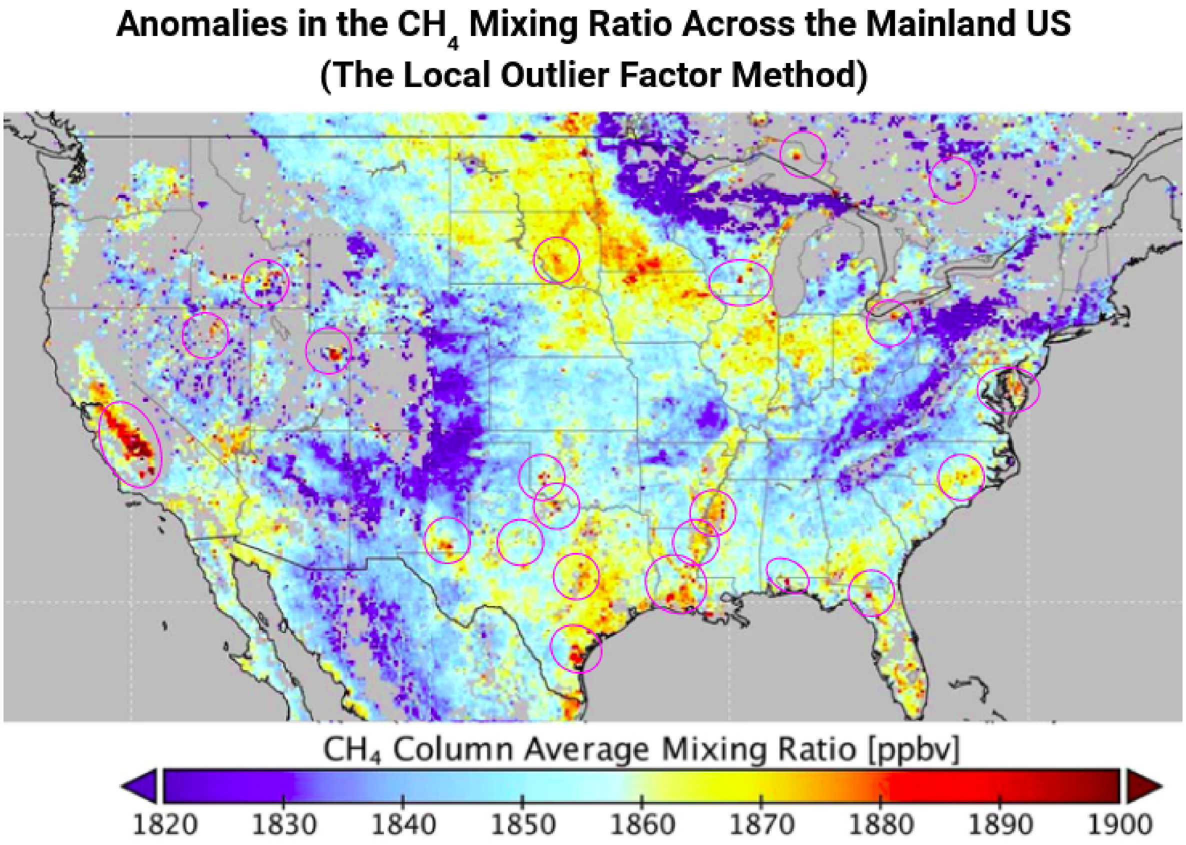
Anomalies in the $CH_4$ Mixing Ratio Across the Mainland US
(The Local Outlier Factor Method)
$CH_4$ Column Average Mixing Ratio [ppbv]
1820
1830
1840
1850
1860
1870
1880
1890
1900

Figure 3: Local Outlier Factor Anomalies on the Map.

This method does a great job of finding anomalies, but tends to classify some normal points as anomalies. It tends toward Type I error.

**References:** For Python, see the function in the `scikit-learn` package. For `R`, see the function in the `dbscan` package.

# 3 The Isolation Forest

This method works by removing all of the data corresponding to response values that are within one standard deviation of the mean. The remaining response values are plotted on a spatial map and, for each remaining point, the spatial subregion holding that point is cut in half until only that point is contained within. The points that are isolated with the fewest number of cuts are considered anomalous.

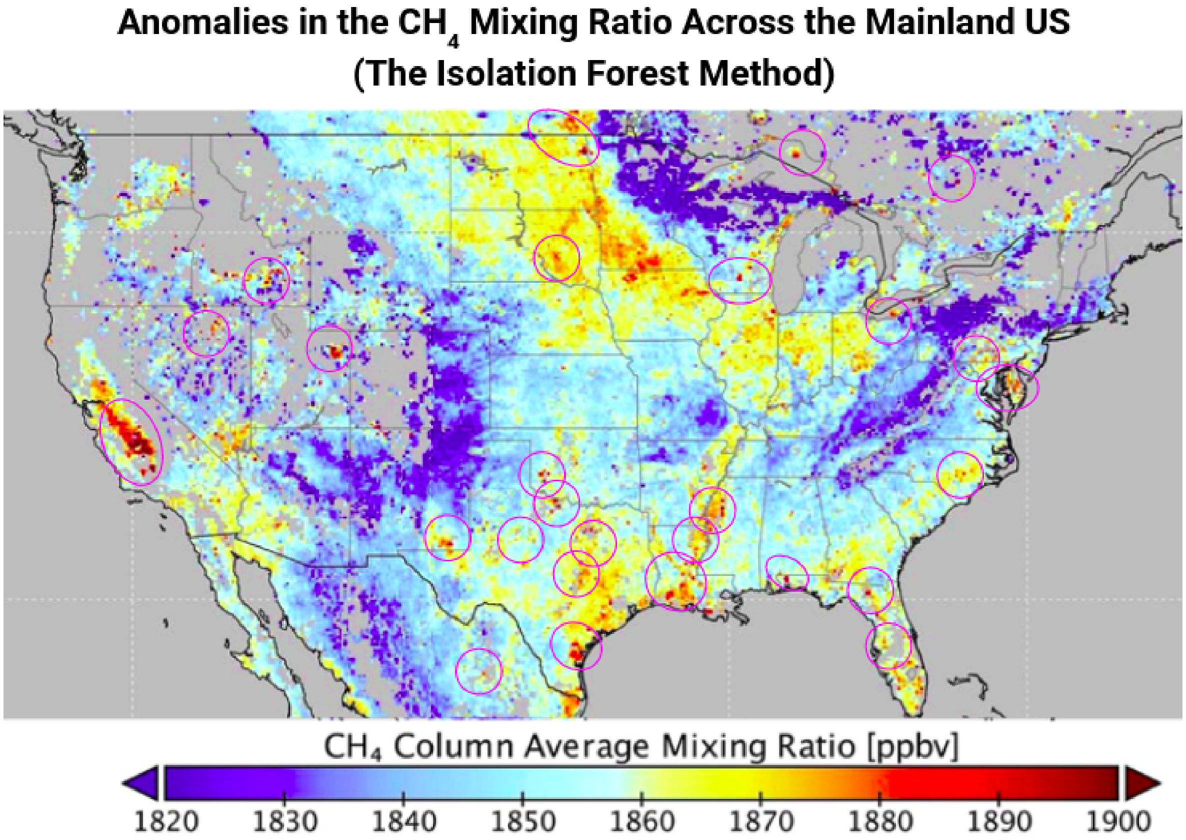
Anomalies in the $CH_4$ Mixing Ratio Across the Mainland US
(The Isolation Forest Method)
$CH_4$ Column Average Mixing Ratio [ppbv]
1820
1830
1840
1850
1860
1870
1880
1890
1900

Figure 4: Isolation Forest Anomalies on the Map.

This method behaves very similarly to the local outlier factor. It tends toward Type I error.

**References:** For Python, see the function in the `scikit-learn` package. For `R`, see the function in the `isotree` package.

# 4 The Autoencoder Neural Network

This method works by trying to predict the entire set of response values using only a small, encoded subset of the features in the data. The response values that are most unlike the others tend to be excluded from the encoded features, and they are classified as anomalies.

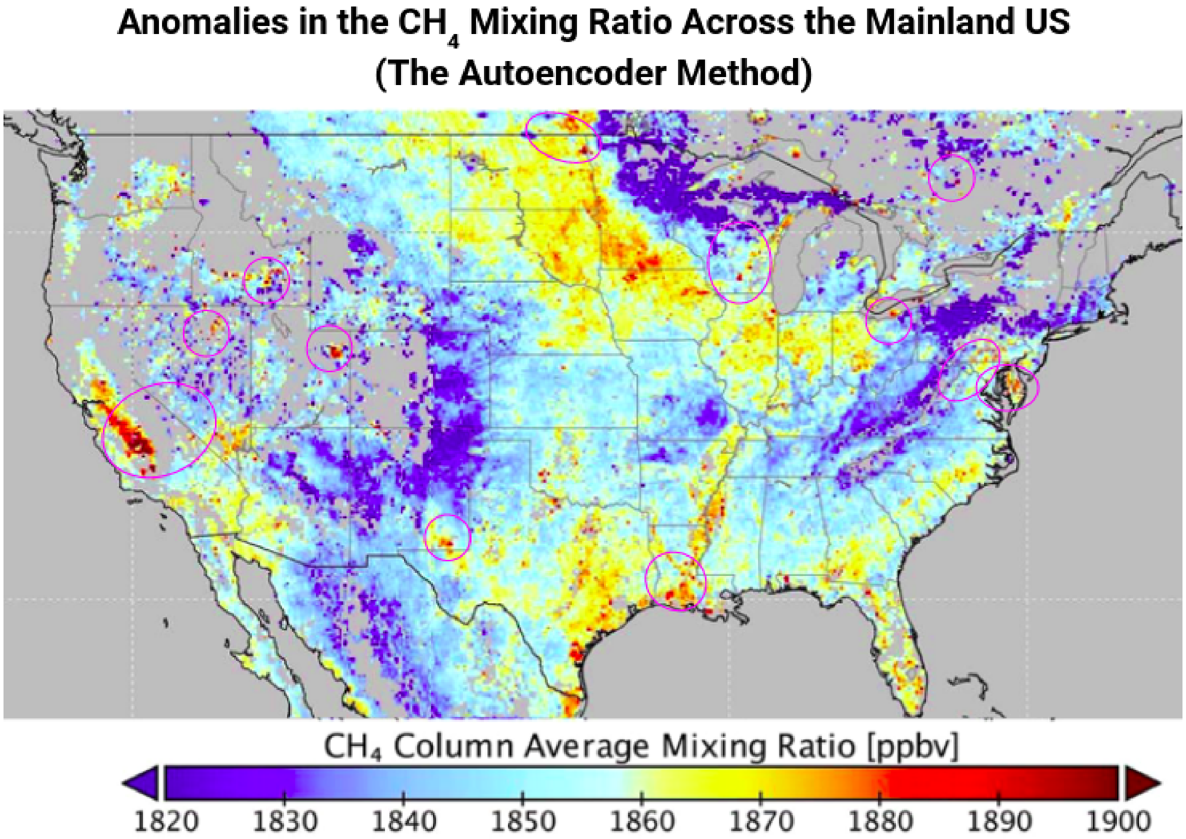
Anomalies in the $CH_4$ Mixing Ratio Across the Mainland US
(The Autoencoder Method)
$CH_4$ Column Average Mixing Ratio [ppbv]
1820
1830
1840
1850
1860
1870
1880
1890
1900

Figure 5: Autoencoder Anomalies on the Map.

This method requires extensive hyperparameter tuning to be effective. In general, it tends toward Type II error.

**References:** For Python, see the function in the Keras package. For R, see the function in the R interface to the Keras package.

# Some Helpful Graphics

The following graphics intend to depict the high-level concepts supporting some of the more confusing anomaly detection techniques.

## The Local Outlier Factor

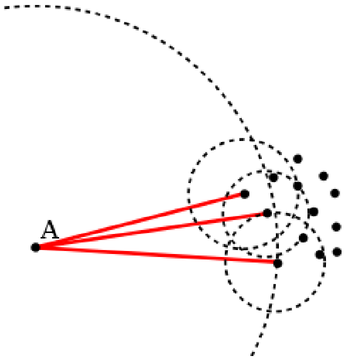

Figure 6: The Local Outlier Factor Finds Points with Unusual Density.

## The Isolation Forest

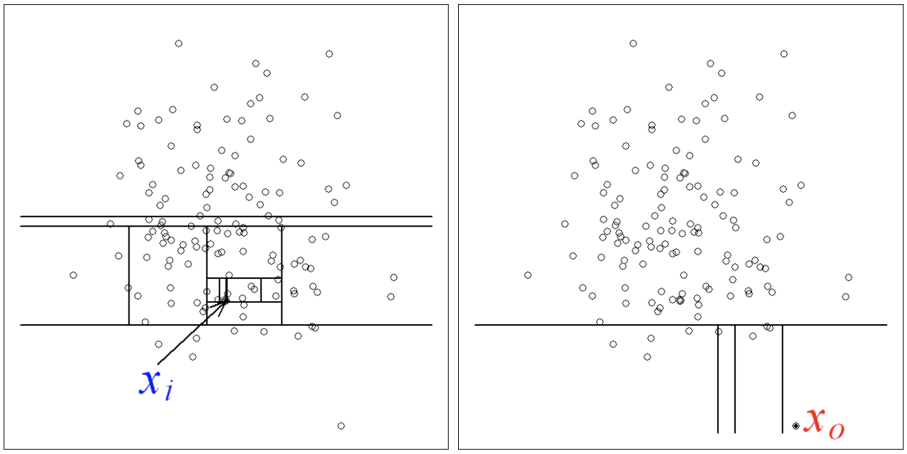

Figure 7: The Isolation Forest Partitions Spatial Regions to Find Outliers.

## The Autoencoder Neural Network

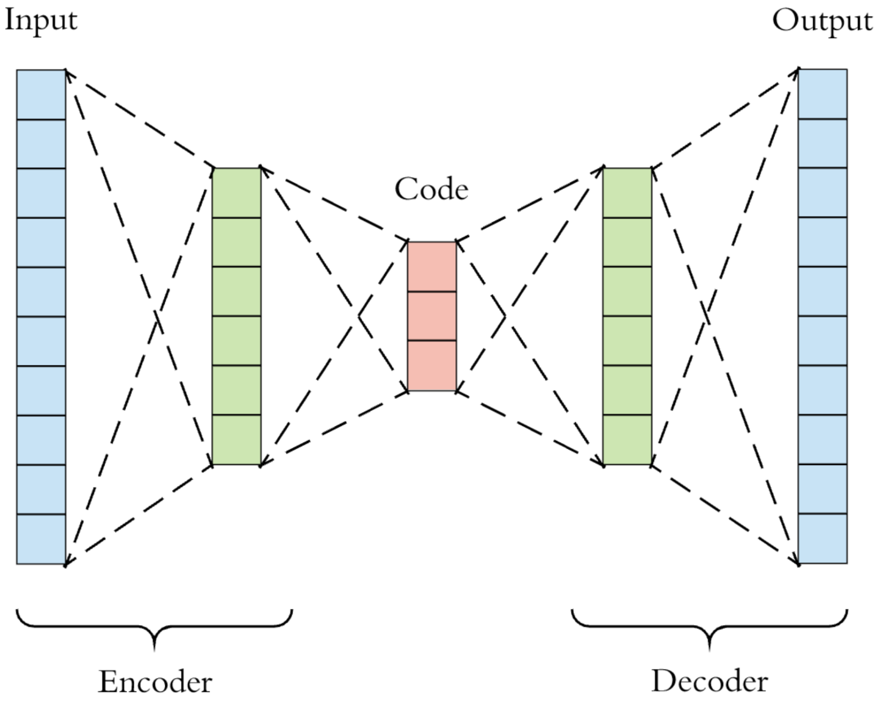

Figure 8: The Autoencoder Estimates Input Features from a Small Encoded Subset.