Anomaly Detection Using FDA

With Applications to Sea Surface Temperature Data

Ryan Pownall

Supervisor
Edward Austin

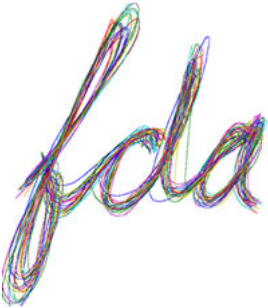
September 3, 2020

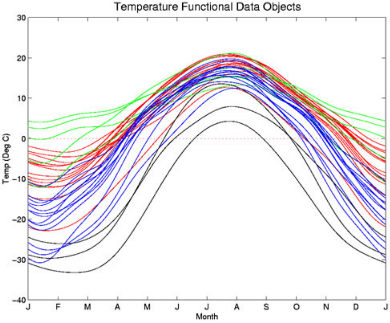
Overview

1. Introduction to FDA
2. The Dataset
3. Anomaly Detection

Functional Data Analysis

- Functional Data Analysis (FDA) involves the analysis of information on curves or functions.
- No assumptions (such as stationarity, low dimensionality, equally spaced observations ect.) have to be made about the functions or the data.





Functional Data Analysis - Basis Functions

- We use *basis function expansions* to model functions:

  $$x(t) = a_1\phi_1(t) + a_2\phi_2(t) + ... + a_k\phi_k(t) = \sum_{i=1}^{k} a_i\phi_i$$

  $\phi_i(t)$ is the i-th basis function
  a_i are constant coefficients

- Basis functions are essentially building blocks that make up our functions.

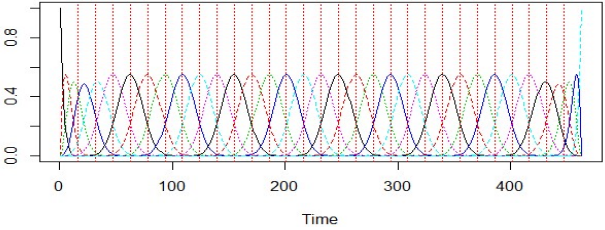
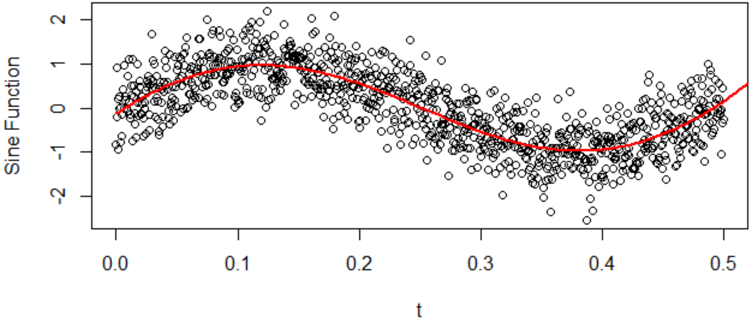


Figure: Order 6 B-spline basis function

Functional Data Analysis - Smoothing

- We want to eliminate high local variation within the data but retain a good fit.
- We minimise the Sum of Squared Error but add a *Roughness Penalty*.
- The smoothing parameter λ determines how much we smooth the data.



The Dataset

- The dataset consists of the Sea Surface Temperature recorded every month (for almost 40 years) at 2000 different locations in the North Atlantic Ocean.

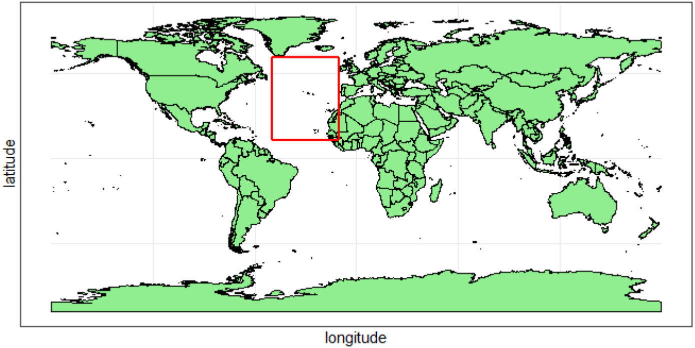
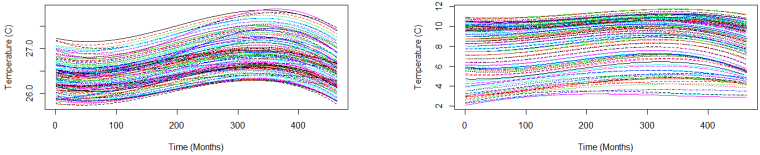


Figure: Dataset Location

The Dataset - FDA

- The discrete dataset is converted into functional form, with each curve representing one location.
- The data is smoothed using a smoothing spline with the smoothing parameter (λ) determined by GCV.



(a) Most Southerly 100 locations

(b) Most Northerly 100 locations

K-means Clustering

- I wanted to split the large dataset into smaller groups and K-means clustering seemed a sensible way to do this.
- The functions were clustered in this way to group locations with similar climates.
- The data was split into 15 clusters as this minimised AIC.

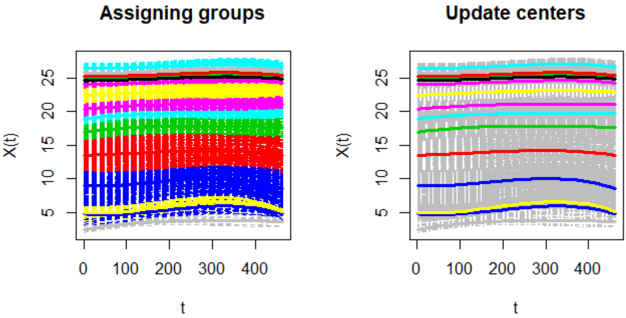
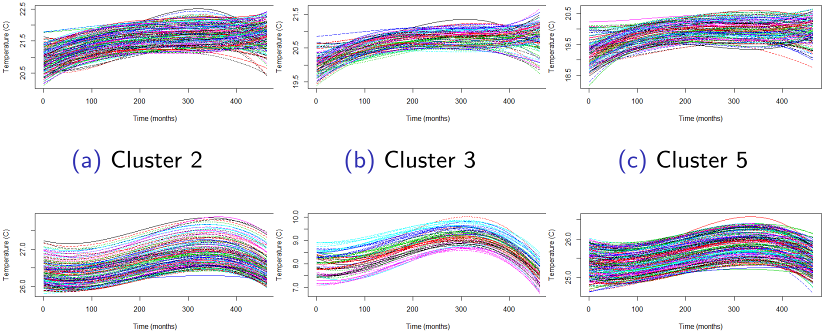


Figure: K-means Clusters

K-means Clustering - Continued

- The set is split into 15 clusters and I have used anomaly detection methods for functional data in order to detect outlying curves within each cluster.



(a) Cluster 2

(b) Cluster 3

(c) Cluster 5

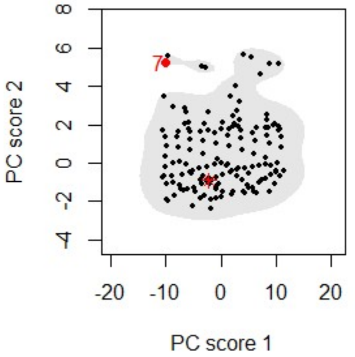
(d) Cluster 6

(e) Cluster 10

(f) Cluster 14

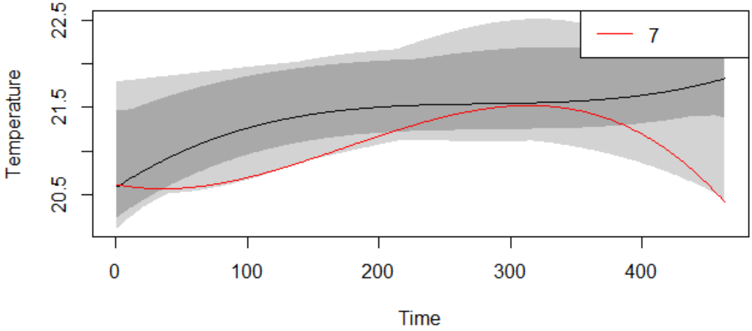
Anomaly Detection

- A *principle component analysis* is used to decompose functional data into the first two principle components and their principle component scores.
- The first two principle component scores are used to make the anomaly detection multivariate rather than functional.



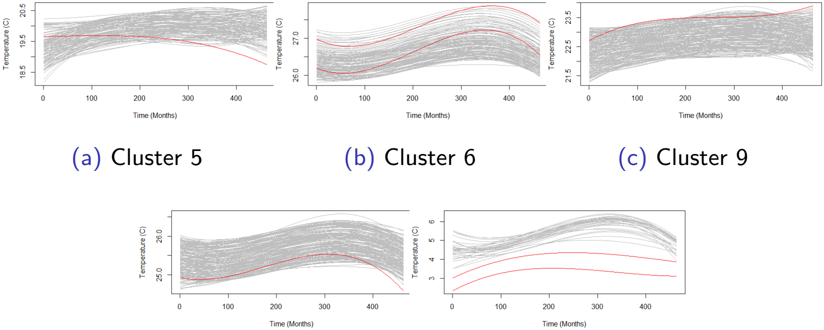
Anomaly Detection - Continued

- Outliers in functional data can be identified as outliers in the *bivariate score space*.
- The score space here is the *highest density region*.
- Any curve outside this region can be considered anomalous.



Anomaly Detection - Sea Surface Temperature Data

- I used this method to detect outliers within the Sea Surface Temperature dataset.
- I focused on curves outside of the 99.9% highest density region; within each cluster.



(a) Cluster 5

(b) Cluster 6

(c) Cluster 9

(d) Cluster 14

(e) Cluster 15

Anomaly Detection - Locations

- The anomalous locations are almost exclusively on the coast of Africa and Greenland.

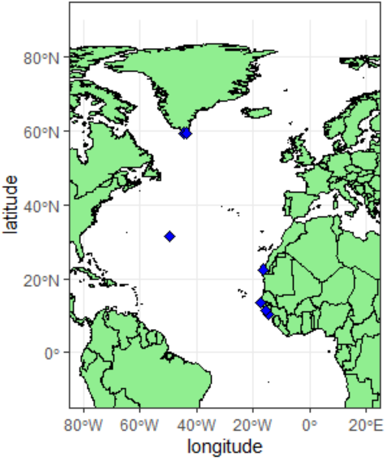


Figure: Locations of Outliers

References

Rob Hyndman and Han Lin Shang.
Rainbow plots, bagplots, and boxplots for functional data.
Journal of Computational and Graphical Statistics, 19, 12 2008.

James Ramsay.
Functional Data Analysis.
Springer, Dordrecht, 2nd ed. edition, 2006.