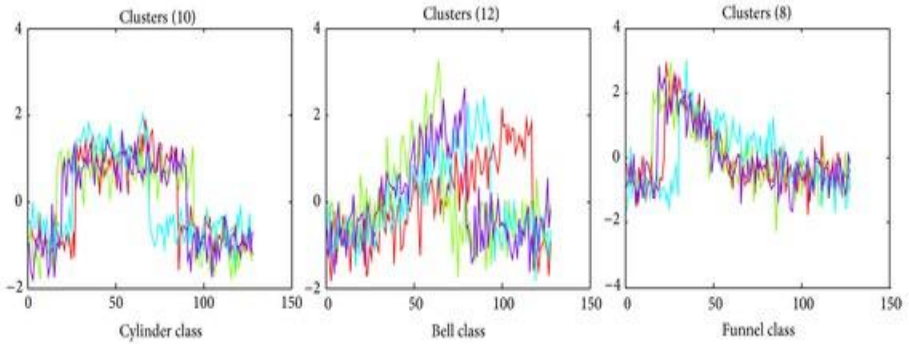
Clustering Time Series Data

Code: https://github.com/snehG0205/NCSA_genomics

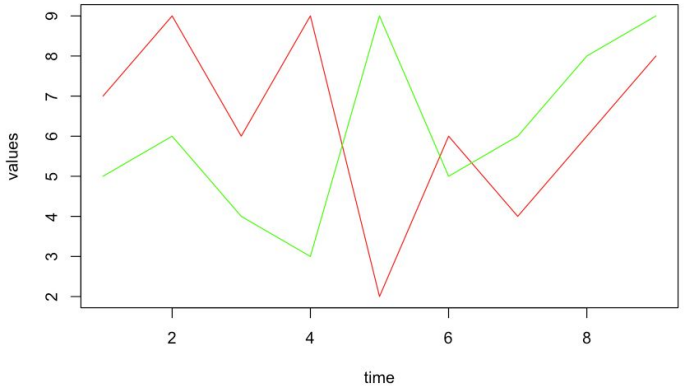
Datasets:

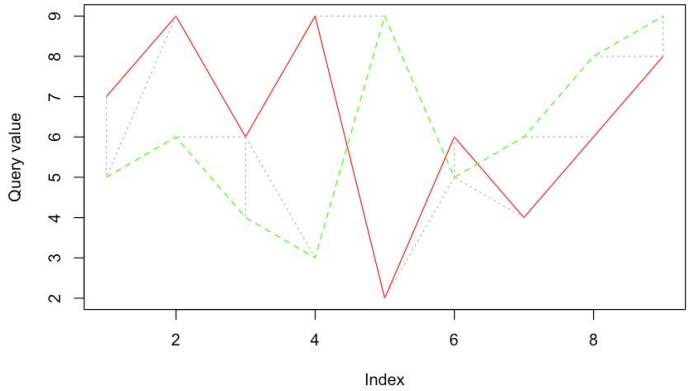
- Cylinder - Bell - Funnel dataset:
  - Dataset with time series of 3 differentiable shapes
  - Labels provided
  - Check efficiency of algorithms
- CGMAnalyzer dataset
  - Dataset with time series of 4 types
  - Classifies model as type-1, type-2, prediabetic and healthy
  - No labels provided



Distance Measures:

- Euclidean Distance :
  - ordinary straight-line distance between two point
- Dynamic Time Warping:
  - Dynamic Time Warping (DTW) is one of the algorithms for measuring the similarity between two temporal time series sequences, which may vary in speed





Clustering Methods:

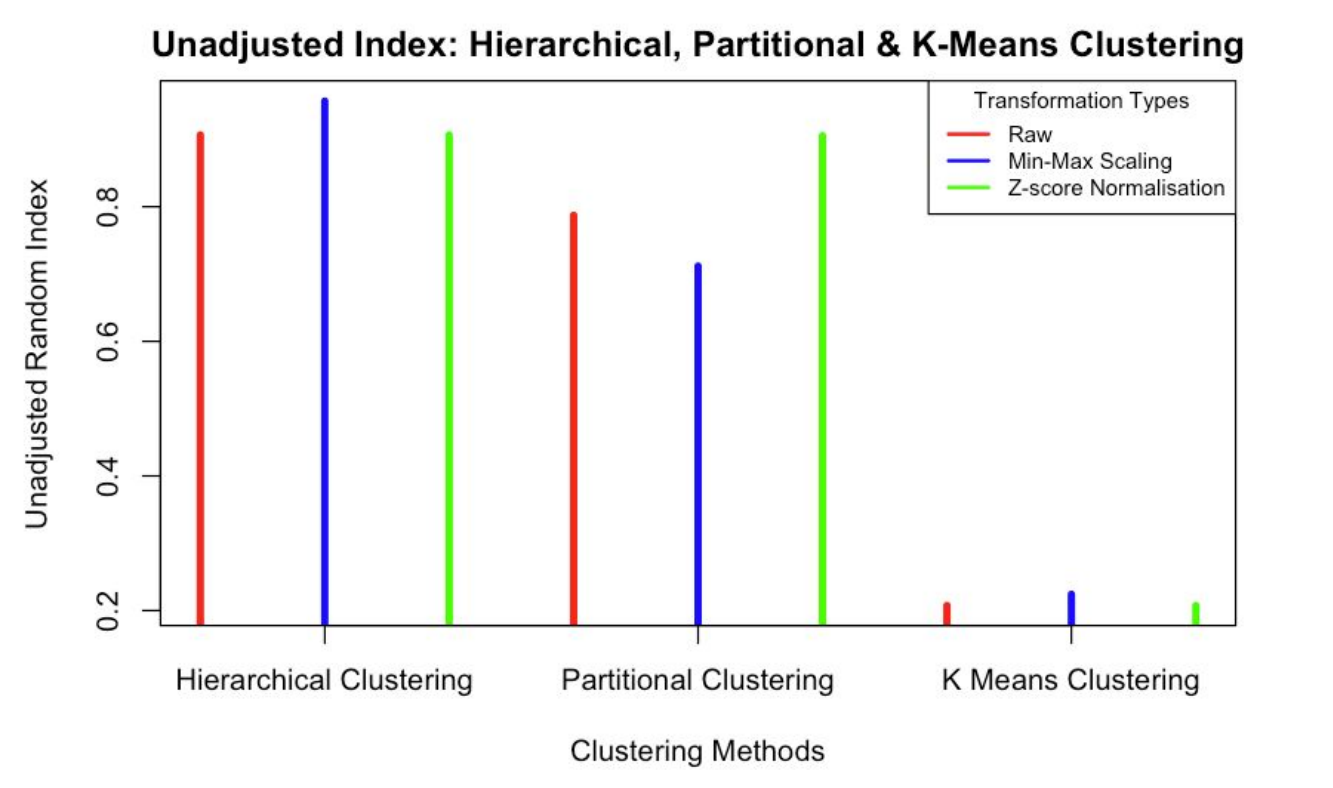
- Partitional (using DTW)
  - Partitional clustering is a method used to classify observations, within a data set, into multiple groups based on their similarity (in our case, DTW)
- Hierarchical(using DTW)
  - Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters.
- K-Means (using Euclidean distance)
  - Kmeans which is considered as one of the most used clustering algorithms due to its simplicity.

Transformation Methods:

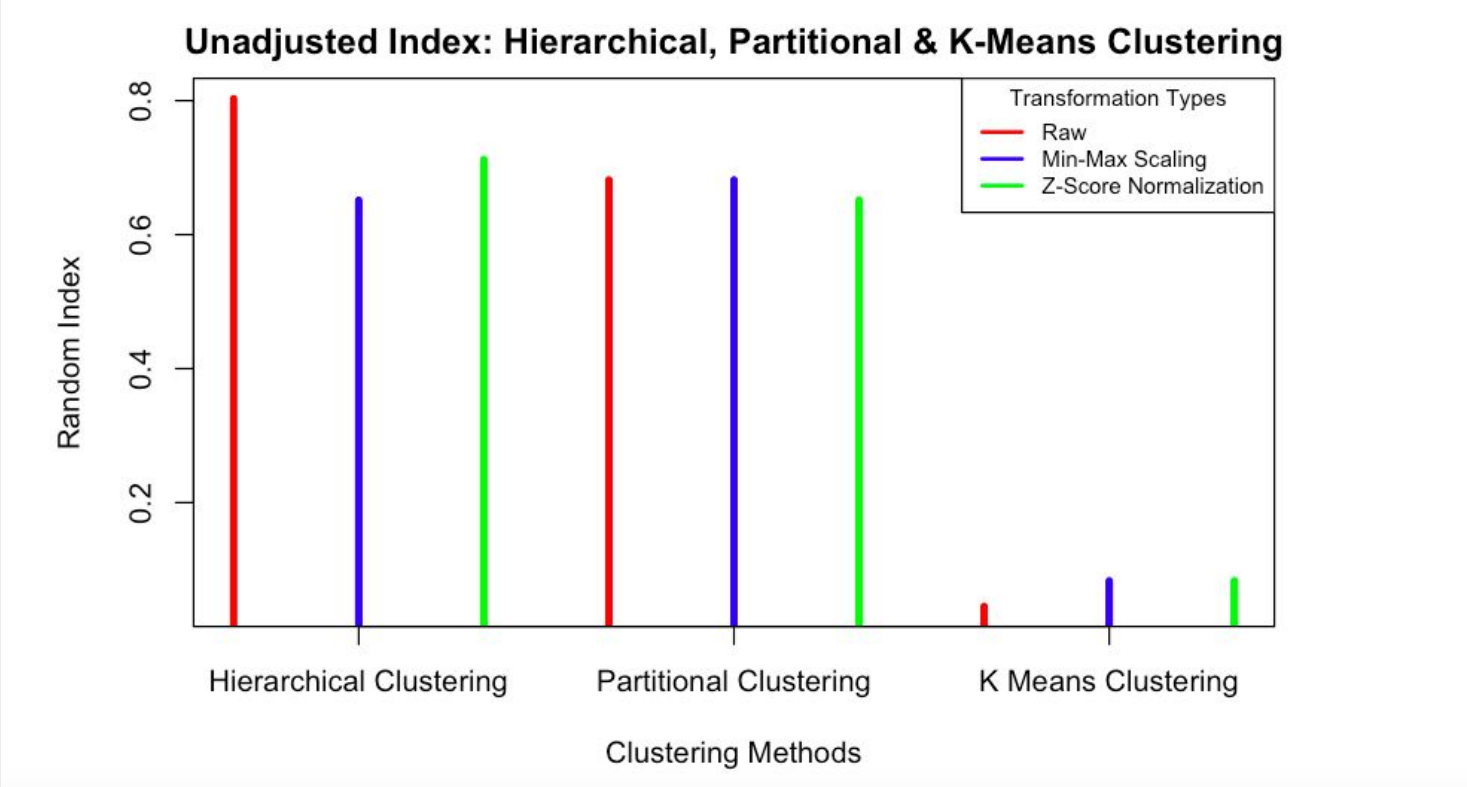
- Min-Max Scaling
- Z-Score Normalization
- Raw

Normalization Technique	Formula
Linear Scaling	$x' = (x - x_{min})/(x_{max} - x_{min})$
Z-score	x' = (x - μ) / σ

Results: CBF Dataset



Results: CGMAnalyzer Dataset



Conclusion

Distance Measure

Dynamic Time Warping

Clustering Method

Hierarchical Clustering

Normalization Method

Can be determined after all imputations

Imputing Time-Series Data using forecast and zoo libraries

Dataset:

CGMAnalyzer dataset

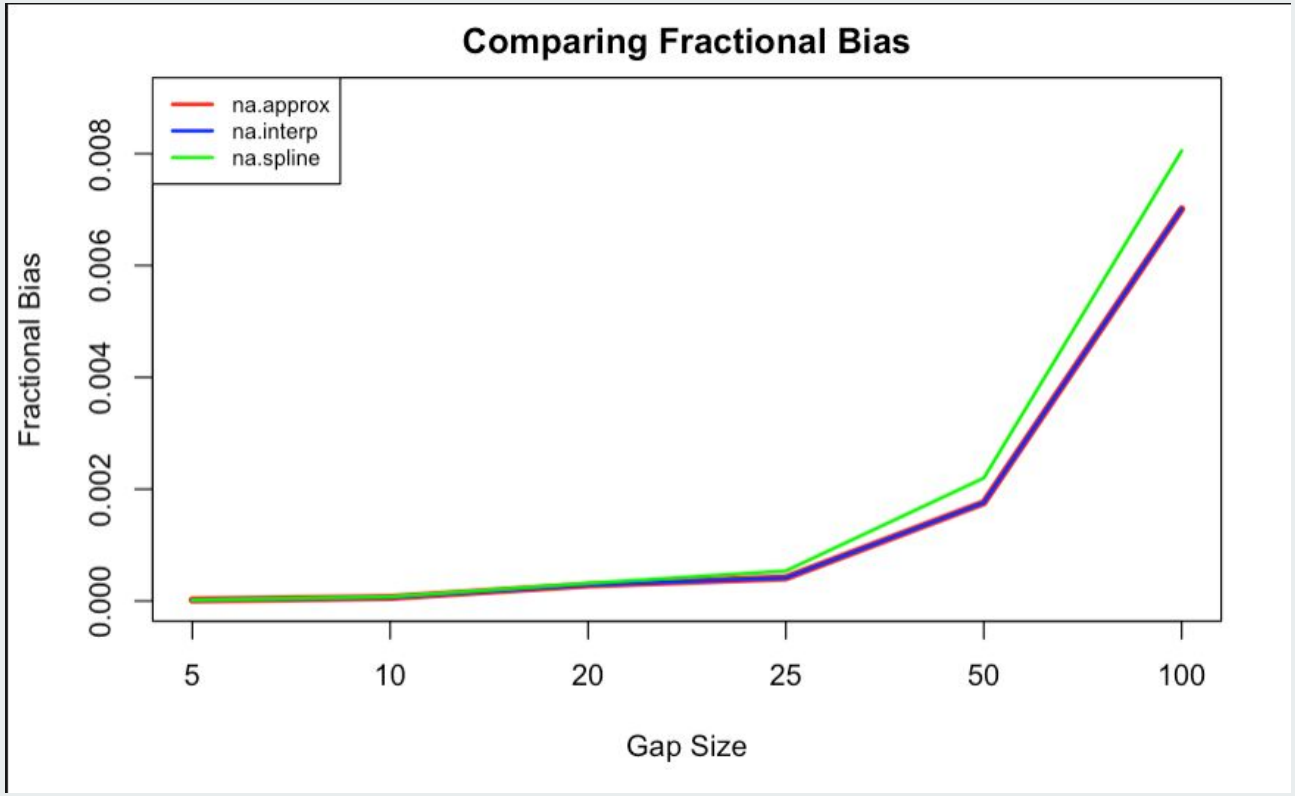
Methods:

- {forecast} -> na.interp : uses linear interpolation for non-seasonal series. For seasonal series, a robust STL decomposition is first computed. Then a linear interpolation is applied to the seasonally adjusted data, and the seasonal component is added back.
- {zoo} -> na.approx : Generic functions for replacing each NA with approximated values.
- {zoo} -> na.spline : Perform cubic (or Hermite) spline interpolation of given data points, returning either a list of points obtained by the interpolation or a function performing the interpolation.

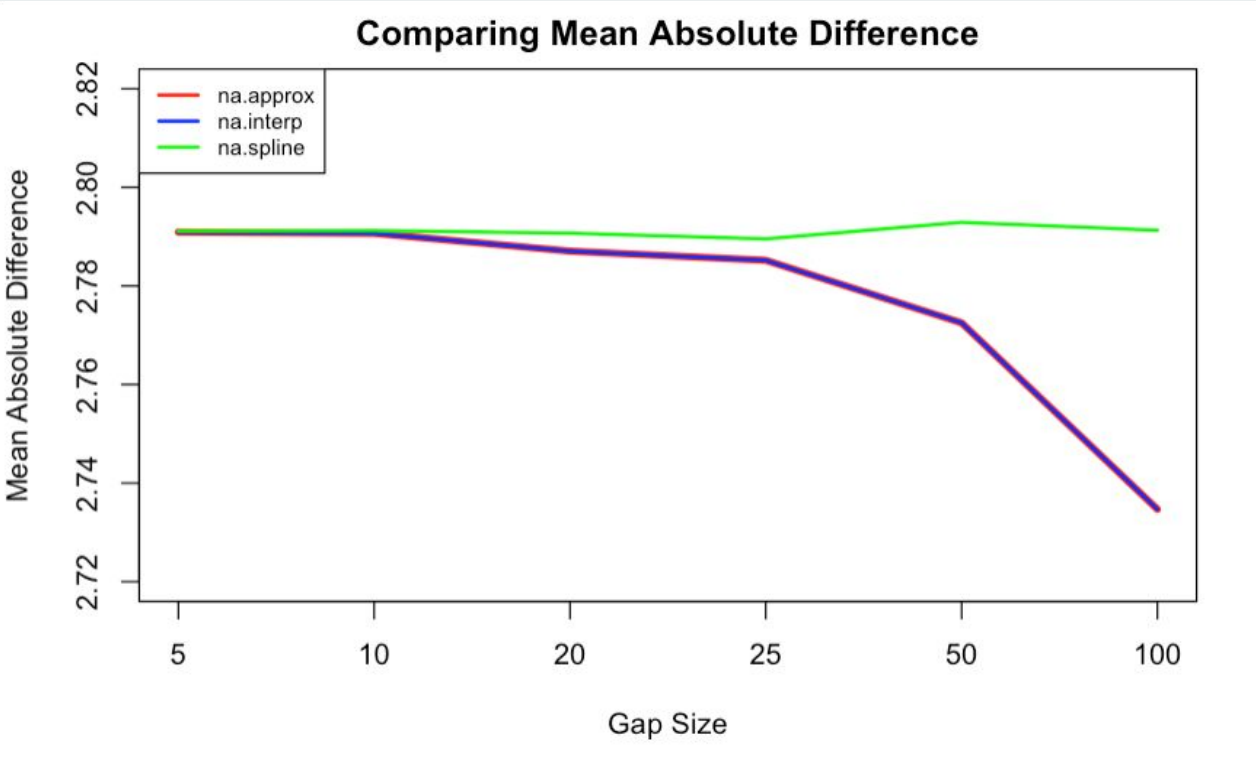
Comparing Metrics:

- Fractional Bias:
  - FB indicates whether predicted values are underestimated or overestimated compared to true values. A perfect imputation model has *FB = 0*
- Mean Absolute Difference:
  - The mean absolute difference of a dataset is the average distance between each data point and the mean. It gives us an idea about the variability in a dataset.
- Root Mean Square Error:
  - RMSE is the standard deviation of the prediction errors
  - Lower values of RMSE indicate better fit
- Index of Agreement:
  - The index of agreement represents the ratio of the mean square error and the potential error. The agreement value of 1 indicates a perfect match, and 0 indicates no agreement at all.

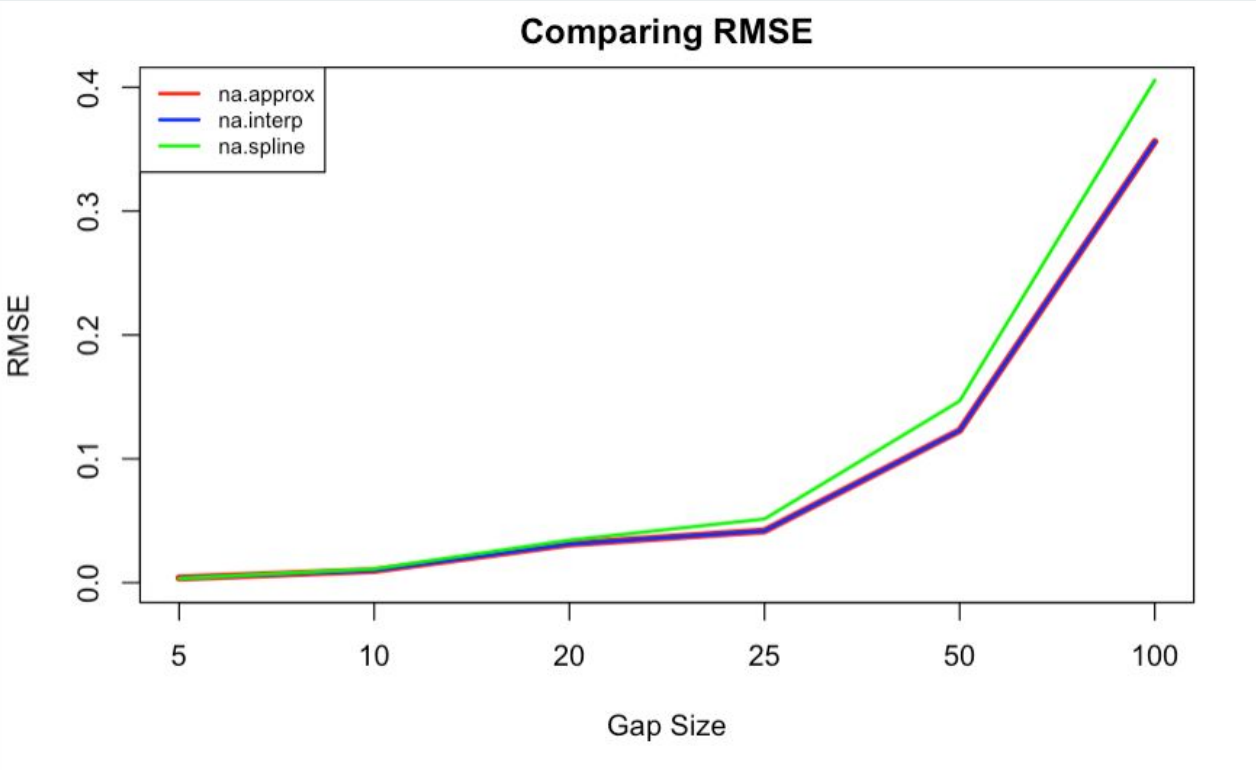
Results: Comparing Fractional Bias



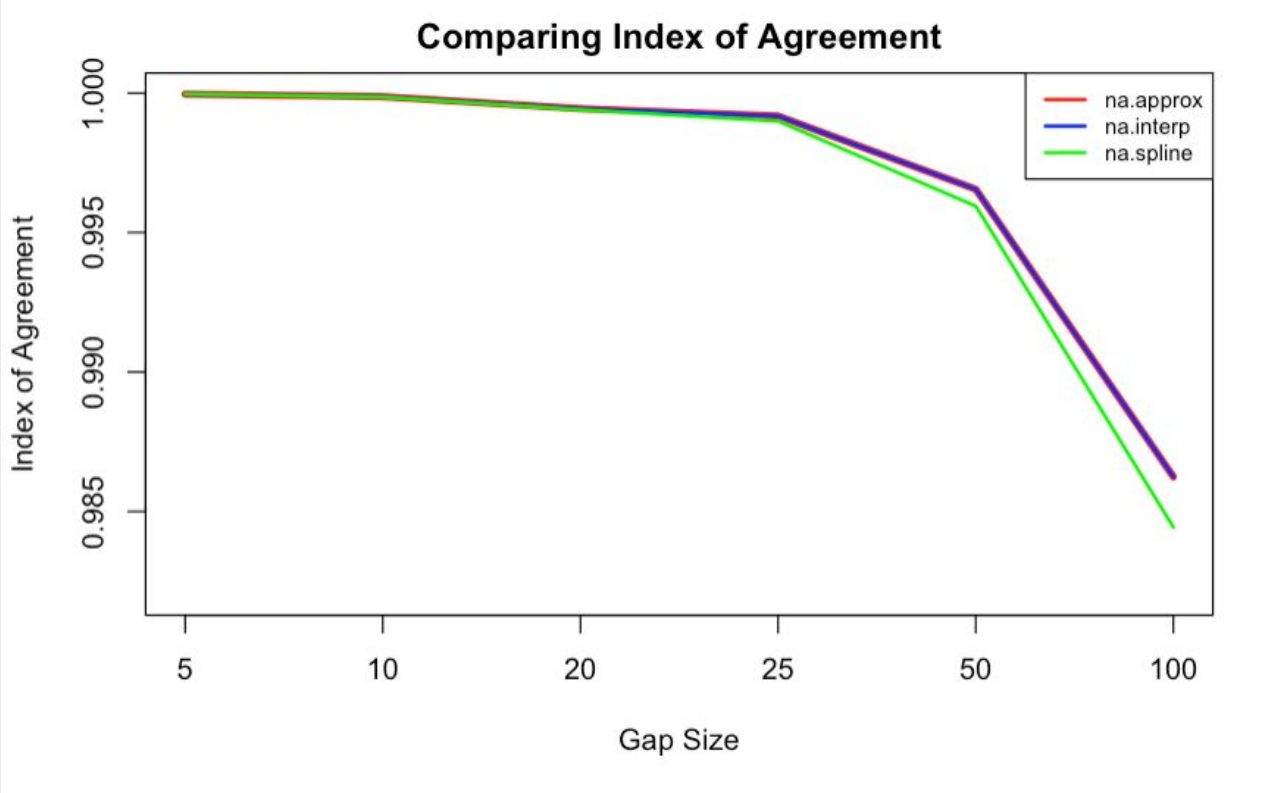
Results: Comparing Mean absolute Difference



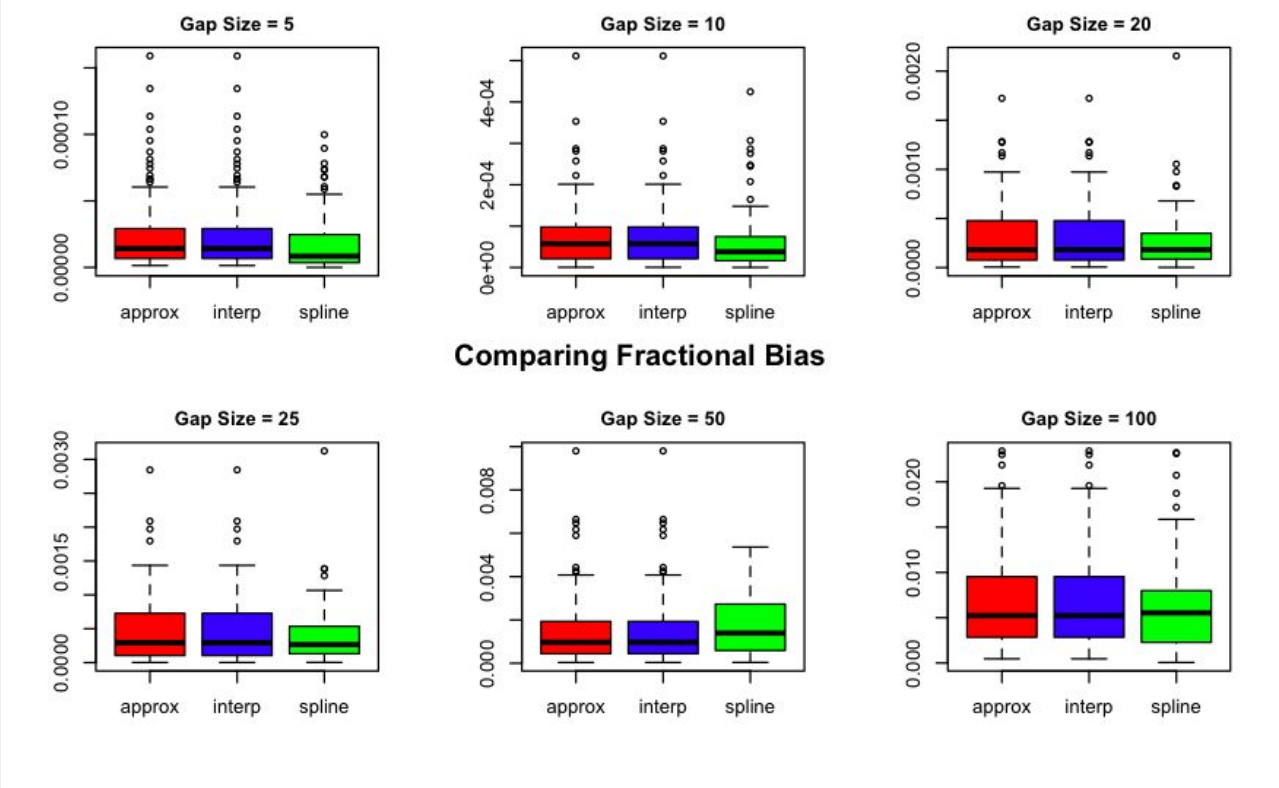
Results: Comparing Root Mean Squared Error



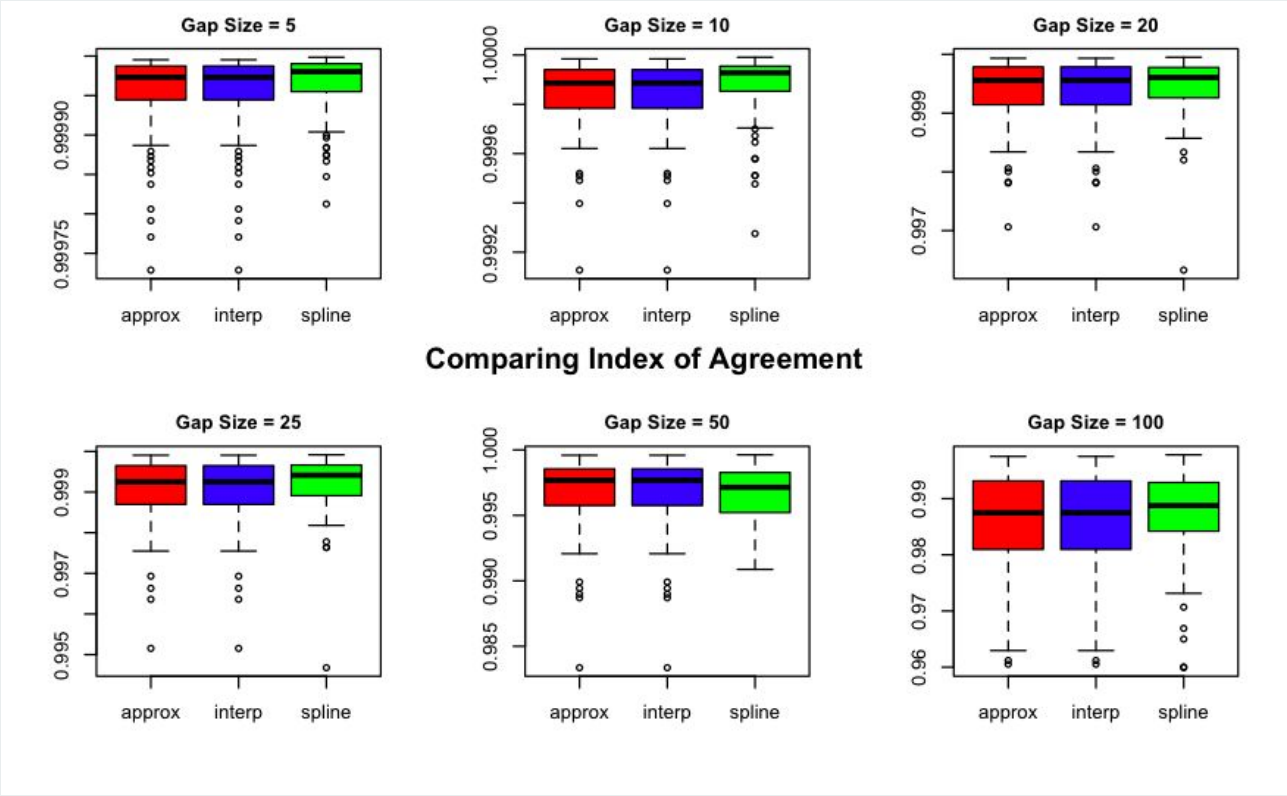
Results: Comparing Index of Agreement



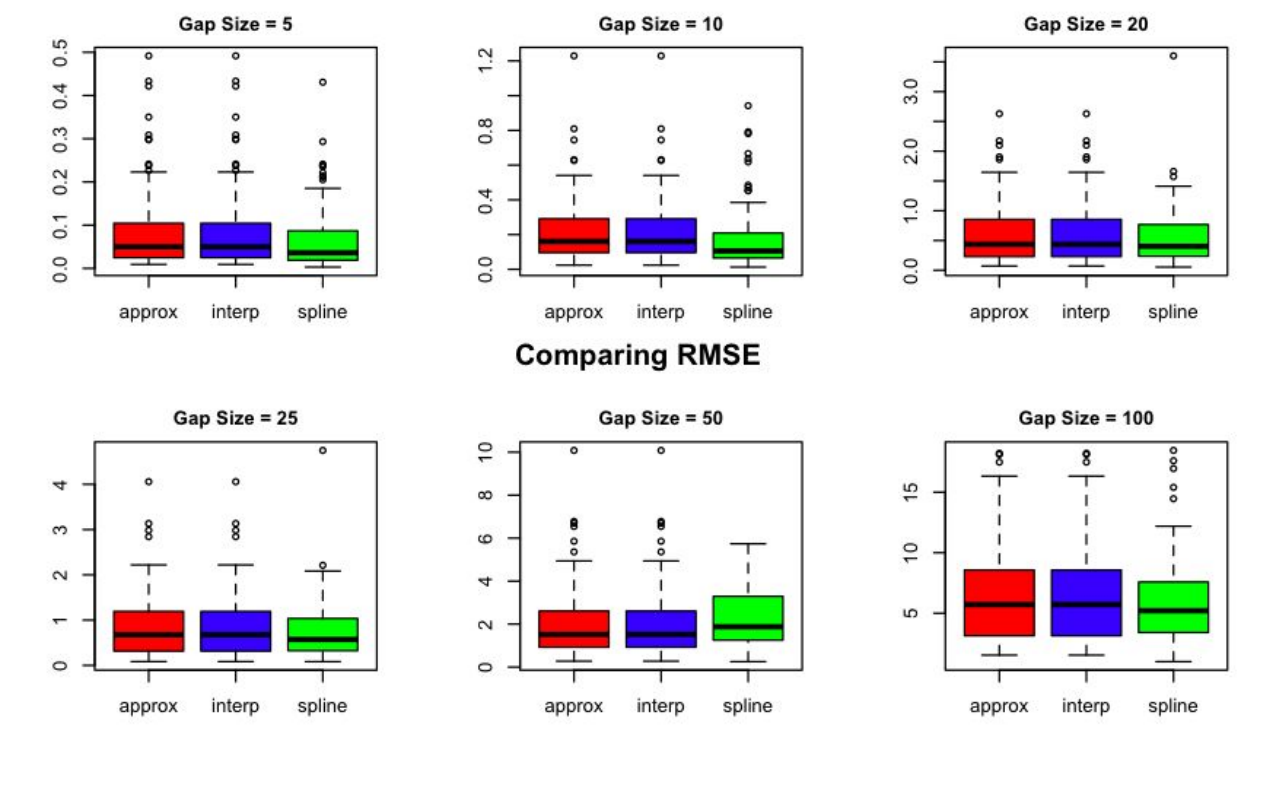
Results: Comparing Fractional Bias



Results: Comparing Index of Agreement



Results: Comparing Root Mean Squared Error



Results: Comparing Mean absolute Difference