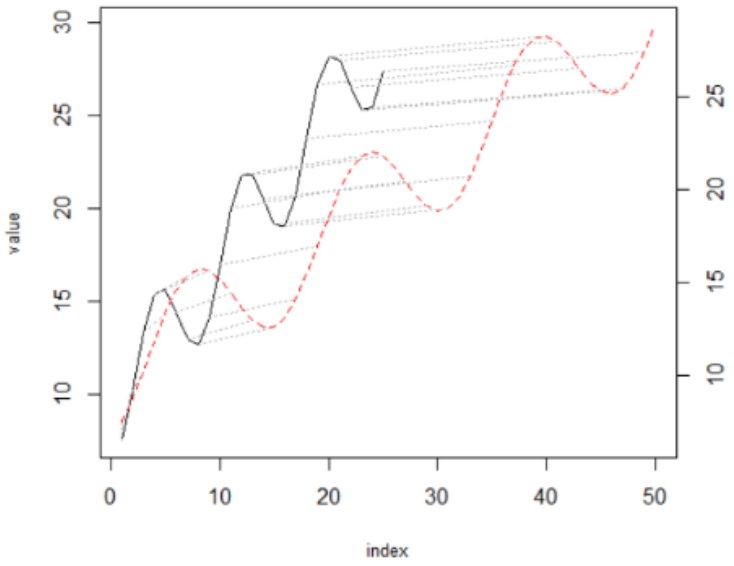
Home assignment 2 - Data Mining ITI8730. Urmas Pitsi, 192028IAPM.

Exercise 1: Distance function.

We have 2 timeseries 25 and 50 datapoints correspondingly. Chart 1.1. visualizes DTW in action by connecting corresponding datapoints in 2 timeserises by applying DTW.

Chart 1.1. Dynamic Time Warping



Exercise 2: Outlier detection. Local Outlier Factors (LOF) for all datapoints.

Data: 2 well separated clusters by gaussians. Additionaly 10 random datapoints with random proximity to the diagonal line separating the clusters.

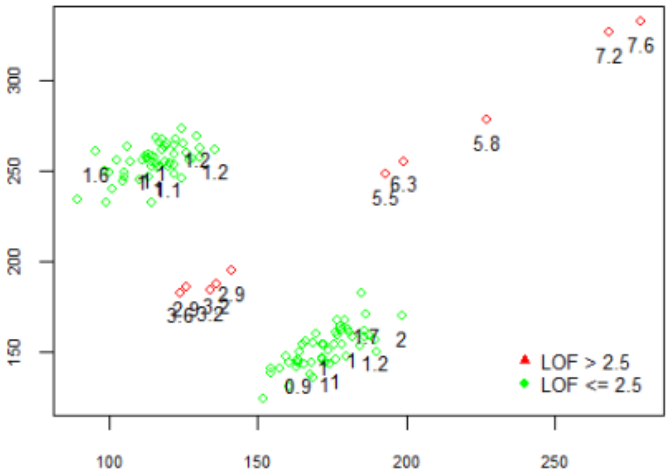
Hypothesis: LOF splits data into 2 clusters. Datapoints in the first cluster having LOF<=T and datapoints in the other cluster having LOF>T.

Conclusion: Hypothesis is true. By setting k=7 and threshold T=2.5, we could split our data into 2 clusters. LOF can be effectively used in detecting outliers. Assumes some insight of the data in order to find optimal hyperparameters k and T for neighbourhood size and threshold respectively.

Table 2.1: Average LOF.

Gaussian 1	1.146
Gaussian 2	1.163
10 Outliers	4.816

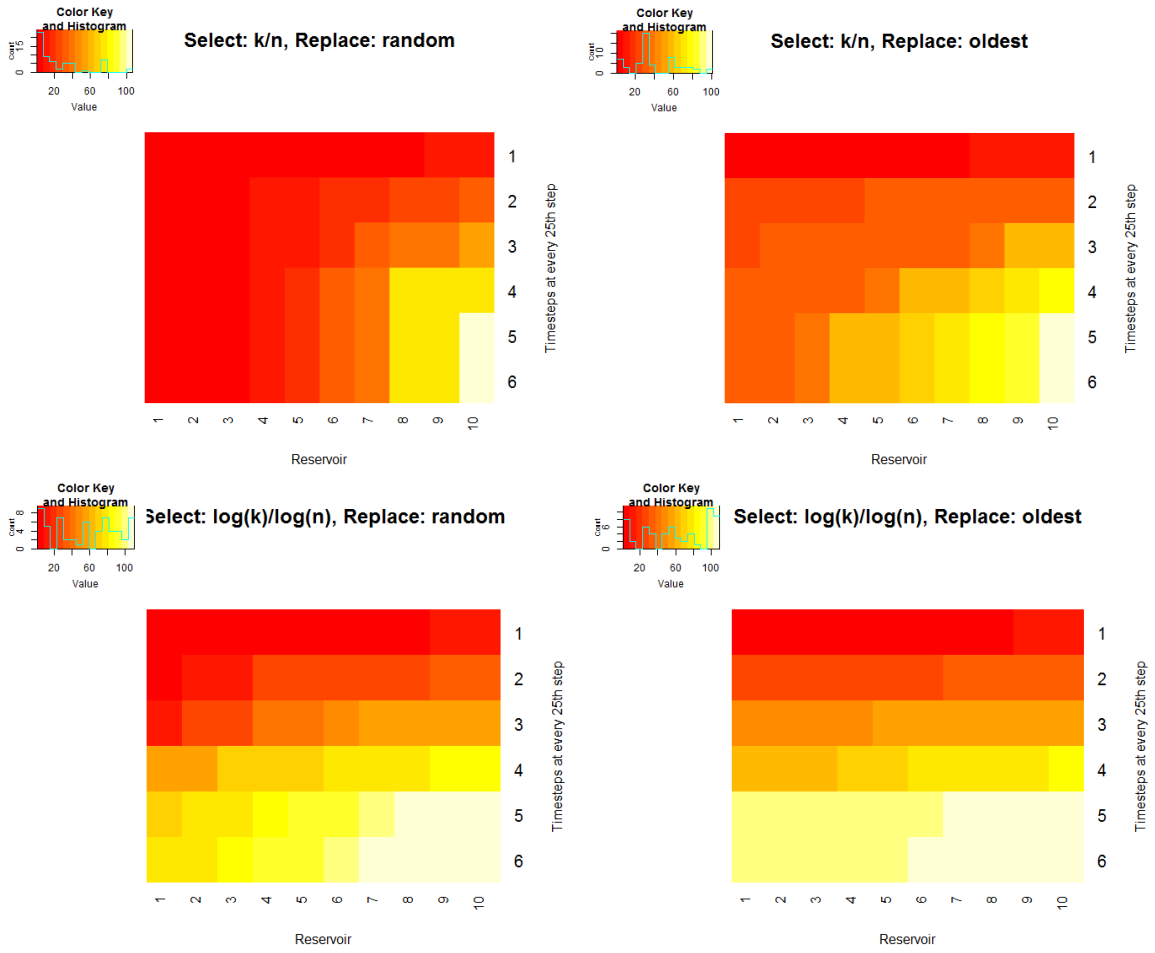
Chart 2.2: Local Outlier Factors, k=7, threshold=2.5



Exercise 3: Stream data mining: reservoir sampling comparison.

Selection strategies: (1) select with probability k/n, (2) with probability log(k)/log(n), where k=reservoir size and n=nth stream point. **Replacement strategies:** (1) random, (2) oldest. **Results:** reservoir turnover is faster with selecting log(k)/log(n) and replacing oldest, ie. datapoints in reservoir are more recent with the same number of timesteps. Lighter colors start to dominate sooner if we look at the charts.

Chart 3.1: Reservoir evolution through time. Horizontal slices depict reservoir on a particular time step. Redness means older datapoints, lighter means more recent.



Exercise 4: Time series forecasting.

Implemented linear regression forecasting with seasonality. Input data is 16 quarters of sales data as an example, predicted 4 quarters.

intercept	5.1
x-coefficient	0.15
R-squared	0.9208