Mining Time Series

L. Torgo & T. Rudolecky

A Definition

Definition

- A time series is a set of observations of a variable that are ordered by time.
- E.g.,
  $x_1, x_2, \cdots, x_{t-1}, x_t, x_{t+1}, \cdots, x_n$
  where x_t is the observation of variable X at time t.
- A multivariate time series is a set of observations of a set of variables over a certain period of time.

Explanation

Obtaining a Time Series Model help us to have
a Deeper Understanding of the Mechanism
that Generated the Observed Time Series Data.

Forecasting

The Goal of Time Series Forecasting

- Given:
  $x_1, x_2, \cdots, x_{t-1}, x_t$ *The Past!*
- Obtain:
  a time series model
- Which is able to make predictions concerning:
  $x_{t+1}, \cdots, x_n$ *The Future!*

Time Series Data Mining

Main Time Series Data Mining Tasks

- *Indexing (Query by Content)*
  Given a query time series Q and a similarity measure $D(Q, X)$ find the most similar time series in a database **D**
- *Clustering*
  Find the natural goupings of a set of time series in a database **D** using some similarity measure $D(Q, X)$
- *Classification*
  Given an unlabelled time series Q, assign it a label C from a set of pre-defined labels (classes)

Summaries of Time Series Data

- Standard descriptive statistics (mean, standard deviation, etc.) do not allways work with time series (TS) data.
- TS may contain trends, seasonality and some other systematic components, making these stats misleading.
- So, for proving summaries of TS data we will be interested in concepts like trend, seasonality and correlation between sucessive observations of the TS.

Types of Variation

Seasonal Variation

Some time series exhibit a variation that is annual in period, e.g. demand for ice cream.

Other Cyclic Variation

Some time series have periodic variations that are not related to seasons but to other factors, e.g. some economic time series.

Trends

A trend is a long-term change in the mean level of the time series.

Stationarity

An Informal Definition

A time series is said to be stationary if

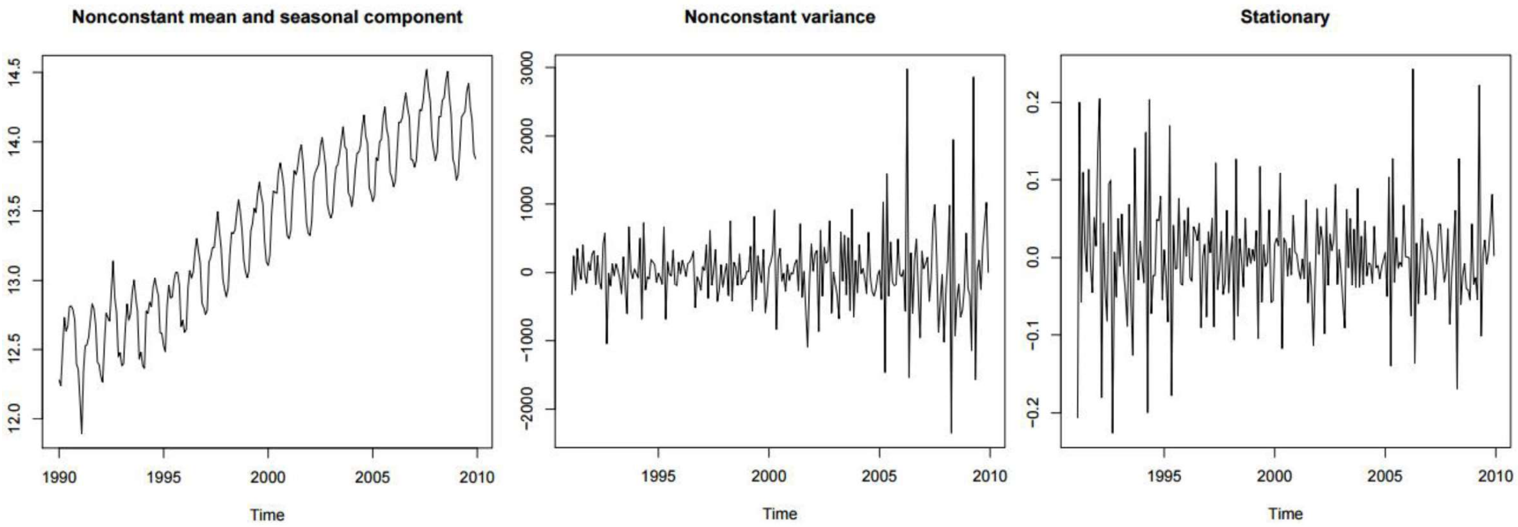
there is no systematic change in mean (no trend),

if there is no systematic change in variance and

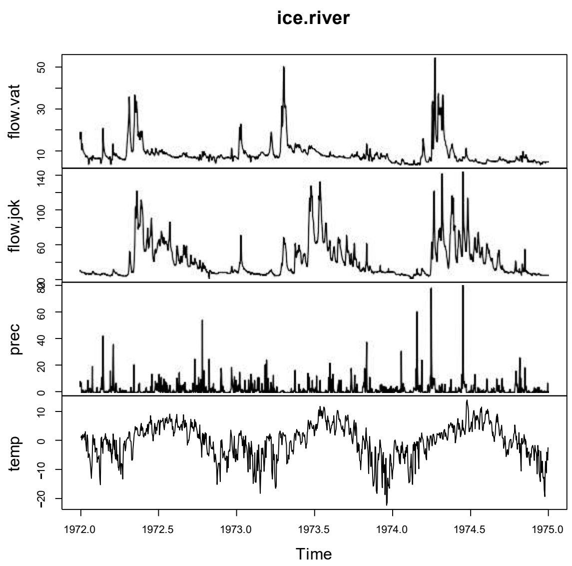
if strictly periodic variations have been removed.

Note that in these cases statistics like mean, standard deviation, variance, etc., bring relevant information!

Stationarity



Time Plots



- Ploting the time series values against time is one of the most important tools for analysing its behaviour.
- Time plots show important features like trends, seasonality, outliers and discontinuities.

Transformations - I

Plotting the data may suggest transformations :

To stabilize the variance

Symptoms: trend with the variance increasing with the mean.
Solution: logarithmic transformation.

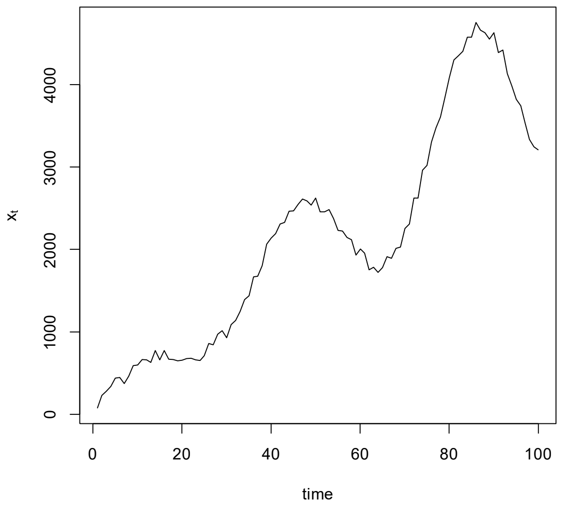
To make the seasonal effects additive

Symptoms: there is a trend and the size of the seasonal effect increases with the mean(multiplicative seasonality).
Solution: logarithmic transformation.

To remove trend

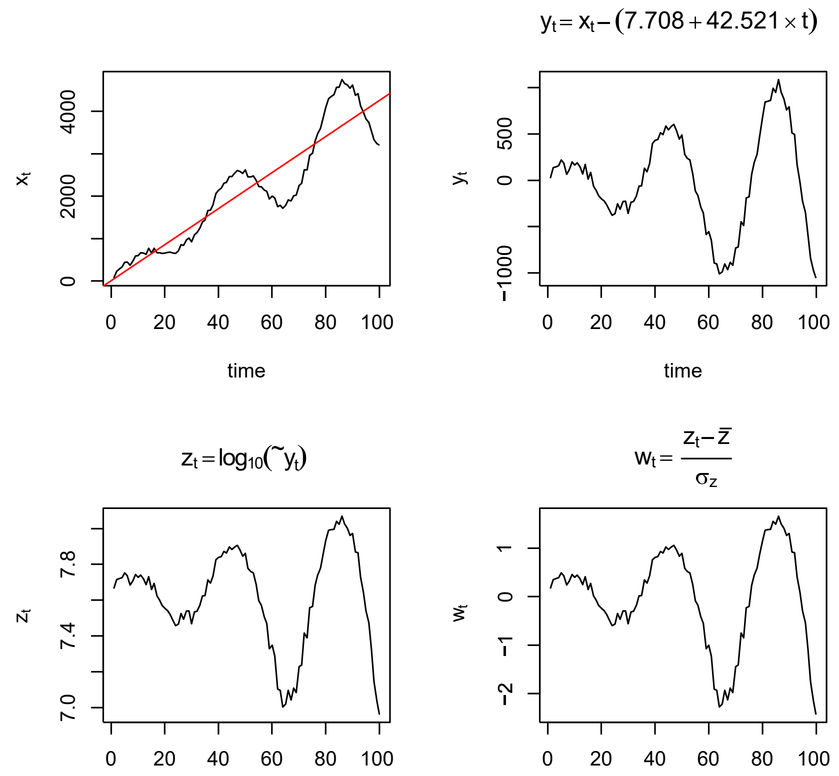
Symptoms: there is systematic change on the mean.
Solution 1: first order differentiation ($\nabla X_t = X_t - X_{t-1}$).
Solution 2: model the trend and subtractit from the original series ($Y = X_t - r_t$).

Transformations - a simple example (1)



An example time series with trend and a multiplicative seasonality effect.

Transformations - a simple example (2)



Tests of Randomness

Frequently we want to test the hypothesis that the observed time series is random.

A possible way is to inspect the correlogram.
An alternative, which is frequently used, is the runs test.

This test basically checks for things like the number of times the value of x_t is above (below) the median value of the series, or the number of times there is a sucession of monotonically increasing (decreasing) values of the series and so on.

Handling Real World Data

A Check List of Common Sense Things to Do (taken from Chatfield, 2004)

- Do you understand the context? Have the right variables been measured?
- Have all the time series been plotted?
- Are there missing values? If so, what should be done about them?
- Are there any outliers? If so, what should be done about them?
- Are there any discontinuities? If so, what do they mean?
- Does it make sense to transform the variables?
- Is trend present? If so, what should be done about it?
- Is seasonality present? If so, what should be done about it?

Measuring Similarity

Why?

Most time series data mining tasks require the similarity between series to be asserted (e.g. indexing, clustering, classification, etc.).

Types of matching

There are essentially two variants of similarity matching:

- *Whole matching*
  where the query time series is matched (as a whole) against all time series in the data base.
- *Subsequence matching*
  where all time series in the data base are searched for a subsection match against the query subsequence.

Defining a Distance Function

Distance (or dissimilarity) functions

Given any two time series s_1 and s_2 their distance (or dissimilarity) is denoted by $D(s_1, s_2)$.

Desirable Properties of a Distance Function

- Symmetry
  $D(X, Y) = D(Y, X)$
- Constancy of Self-Similarity
  $D(X, X) = 0$
- Positivity
  $D(X, Y) = 0 \quad \textbf{iff } X = Y$
- Triangular Inequality
  $D(X, Y) \geq D(X, Z) + D(Y, Z)$

Types of Distance Functions

- *Metric* - satisfy all properties
  e.g. Euclidean, correlation, etc.
- *Non-metric* - do not satisfy any of the properties
  e.g. time warping, LCSS, etc.

The Minkowski Metrics

$$D(X, Y) = \left(\sum_{i=1}^{k} (x_i - y_i)^p\right)^{\frac{1}{p}}$$

- City Block ($p = 1$)
- Euclidean ($p = 2$)

$$D(X, Y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Correlation between two time series

$$\rho_{x,y} = \frac{\sum_{t=1}^{N}(x_t - \bar{x})(y_t - \bar{y})}{N\sigma_x\sigma_y}$$

For normalized series ($\frac{x_t - \bar{x}}{\sigma_x}$), we have

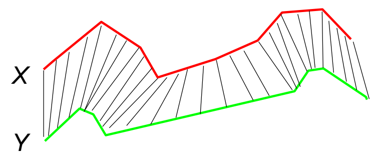
$$\rho_{x,y} = \frac{1}{N}\sum_{t=1}^{N} x_t y_t$$

Dynamic Time Warping - introduction

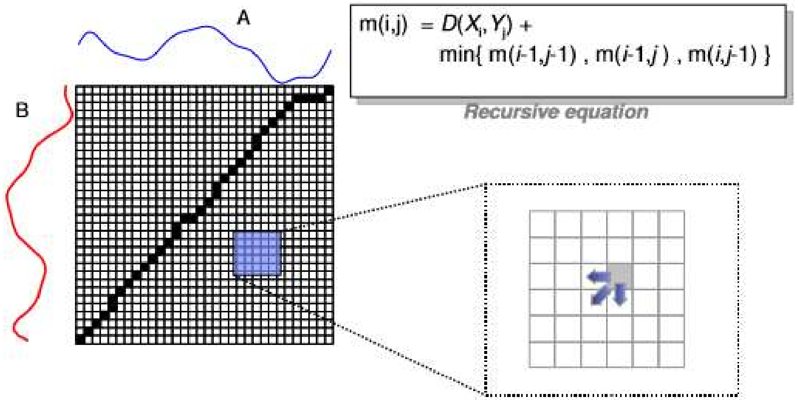
Dynamic Time Warping (DTW) is a non-metric distance function.

Main Ideas of DTW

- Allow for local deformations (stretch and shrink) along the time axis.
- Able to handle series of different lengths



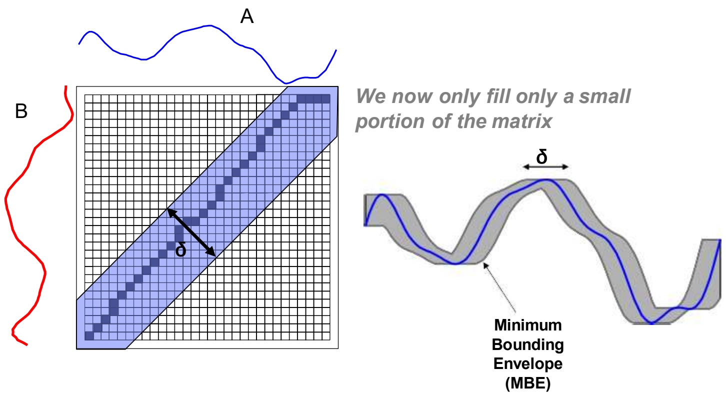
Dynamic Time Warping - how to calculate?



taken from Michalis Vlachos tutorial on Time Series

Find the path on the matrix that ensures the “best results”. It is implemented using dynamic programming techniques.

Dynamic Time Warping - how to speed-up calculations?



taken from Michalis Vlachos tutorial on Time Series

Restrict the set of paths (warping paths) that are considered to find the "best results". Several methods exist to carry out this restriction (e.g. Sakoe-Chiba band, Itakura parallelogram, etc.)

LCSS – Longest common subsequence

$$A = ((a_{x1,1}, \ldots, a_{xp,1}), \ldots, (a_{x1,n}, \ldots, a_{xp,n})),$$

$$B = ((b_{x1,1}, \ldots, b_{xp,1}), \ldots, (b_{x1,m}, \ldots, b_{xp,m})).$$

For a trajectory A, let $Head(A)$ be the sequence:

$$Head(A) = ((a_{x1,1}, \ldots, a_{xp,1}), \ldots, (a_{x1,n-1}, \ldots, a_{xp,n-1})).$$

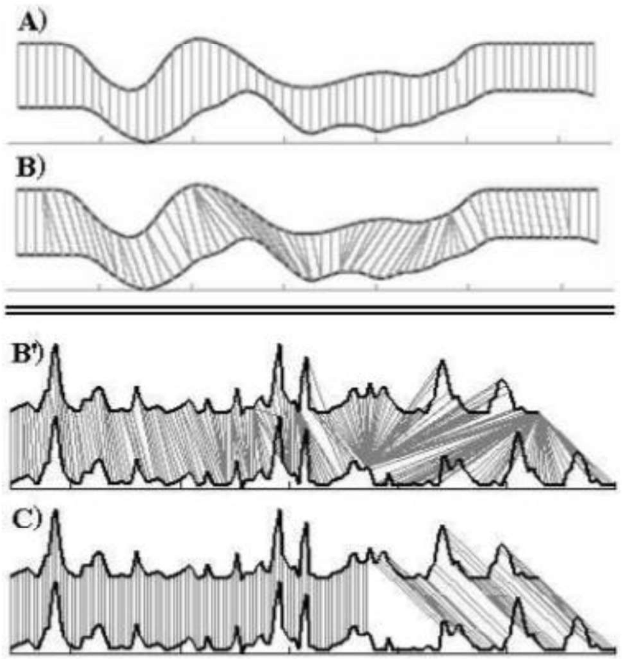
Given an integer δ and a real number $0 < \epsilon < 1$, the similarity function $LCSS_{\delta,\epsilon}(A, B)$ is defined using the recurrent algorithm (4) [32]. N and M are the size of the sequences A and B respectively at the first step of the recurrent algorithm.

$$LCSS_{\delta,\epsilon}(A,B) = \begin{cases} 0 \quad \text{if } A \text{ or } B \text{ is empty,} \\ \\ 1 + LCSS_{\delta,\epsilon}(Head(A), Head(B)), \\ \quad \text{if } d\left(a_{x_k,n}, b_{x_k,m}\right) < \epsilon, \forall 1 \leq k \leq p, \\ \quad \text{and } |n - m| \leq \delta \text{ and } |N - n - M + m| \leq \delta, \\ \\ max\left(LCSS_{\delta,\epsilon}(Head(A), B), LCSS_{\delta,\epsilon}(A, Head(B))\right) \\ \quad \text{otherwise.} \end{cases} \tag{4}$$

LCSS – Longest common subsequence

	A	S	D	G	H	J	P	L
W	0	0	0	0	0	0	0	0
S	0	1	1	1	1	1	1	1
F	0	1	1	1	1	1	1	1
D	0	1	2	2	2	2	2	2
H	0	1	2	2	3	3	3	3
P	0	1	2	2	3	3	4	4
L	0	1	2	2	3	3	4	5
J	0	1	2	2	3	4	4	5

Comparison of distance functions



The graphs represent the results of associations when computing (A) Euclidian distance, (B & B') Dynamic Time Warping (DTW) distance, and (C) distance based on the longest common subsequence (LCSS).

Goals of an Evaluation Method

- The golden rule:

  The data used for evaluating (or comparing) any models cannot be seen during model development.

- The goal of any evaluation procedure:
  - Obtain a reliable estimate of some evaluation measure.
    High probability of achieving the same score on other samples of the same population.
- Evaluation Measures
  - Predictive accuracy.
  - Model size.
  - Computational complexity.

Obtaining Reliable Estimates

- The usual techniques for model evaluation revolve around resampling.
  - Simulating the reality.
    - Obtain an evaluation estimate for unseen data.
- Examples of Resampling-based Methods
  - Holdout.
  - Cross-validation.
  - Bootstrap.

Time Series Data Are Special!

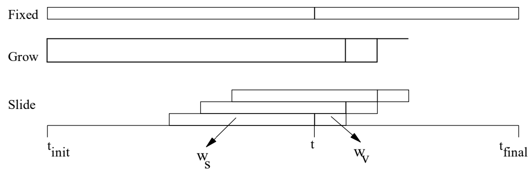
Any form of resampling changes the natural order of the data!

Correct Evalution of Time Series Models

- General Guidelines
  - Do not “forget” the time tags of the observations.
  - Do not evaluate a model on past data.

- A possible method
  - Divide the existing data in two time windows
    - Past data (observations till a time t).
    - “Future” data (observations after t).
  - Use one of these three learn-test alternatives
    - Fixed learning window.
    - Growing window.
    - Sliding window.

Learn-Test Strategies



Fixed Window

A single model is obtained with the available "training" data, and applied to all test period.

Growing Window

Every w_v test cases a new model is obtained using all data available till then.

Sliding Window

Every w_v test cases a new model is obtained using the previous w_s

Some Metrics for Evaluating Predictive Performance

Absolute Measures

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2$$

- Mean Absolute Deviation (MAD)

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|\hat{x}_i - x_i|$$

Relative Measures

- Theil Coefficient

$$U = \frac{\sqrt{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}}{\sqrt{\sum_{i=1}^{n}(x_i - x_{i-1})^2}}$$

- Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(\hat{x}_i - x_i)}{x_i}\right|$$

Assumptions of “Classical” Linear Approaches

- *Linearity*
  The model of the time series behaviour is linear on its inputs.
- *Stationarity*
  The underlying equations governing the behaviour of the system do not change with time.

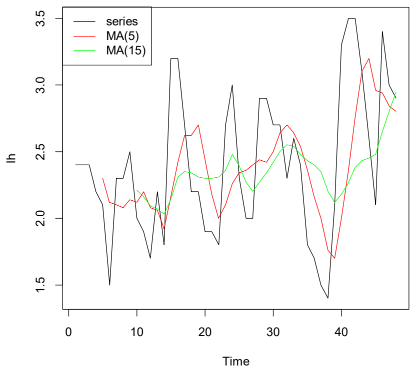
Most “classical” approaches assume stationary time series, thus one usually needs to transform non-stationary time series into stationary ones before using these tools.

Moving Average Models

Definition

A moving average of order q, MA(q),is a time series given by

$$Y_t = \sum_{i=0}^{q} \beta_i X_{t-i}$$



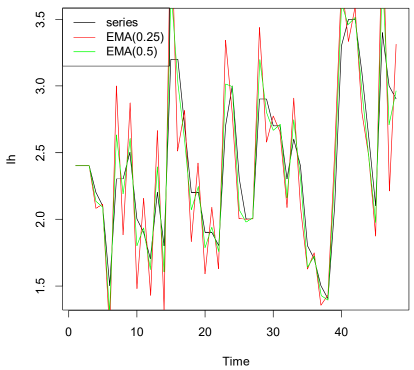
Exponential Moving Average Models

Definition

An exponential moving average is a series given by

$$\begin{aligned} Y_t &= \alpha \times X_t + (1 - \alpha) \times \mathrm{EMA}_\alpha(X_{t-1}) \\ Y_1 &= X_1 \end{aligned}$$

where $\alpha \in [0..1]$ is a smoothing parameter.

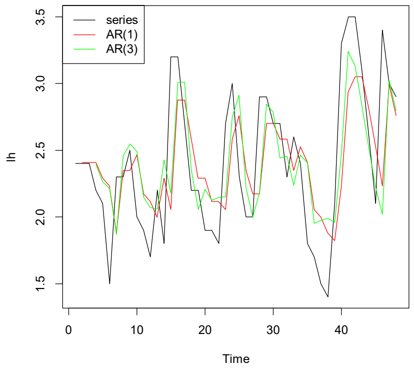


Autoregressive (AR) Models

Definition

An autoregressive model of order p is a series given by

$$Y_t = \sum_{i=0}^{p} \alpha_i Y_{t-i}$$

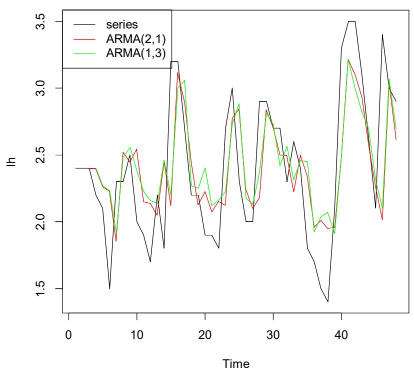


Mixed Autoregressive and Moving Average Models

Definition

A mixed ARMA model of order *p, q* is a series given by

$$Y_t = \sum_{i=0}^{p} \alpha_i Y_{t-i} + \sum_{i=0}^{q} \beta_i X_{t-i}$$



Integrated ARMA (or ARIMA) Models

Definition

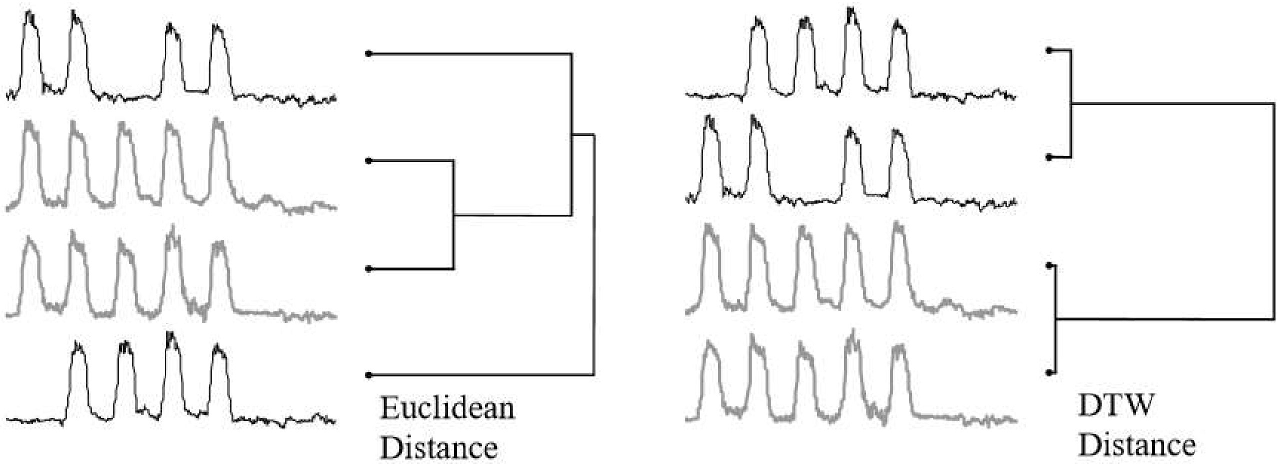
An integrated ARMA (or ARIMA) model of order *p, d, q* is a series given by

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$$

where y'_t is the differenced series (it may have been differenced more than once). The "predictors" on the right hand side include both lagged values of y_t and lagged errors. We call this an **ARIMA(p, d, q) model**, where

$p =$ order of the autoregressive part;
$d =$ degree of first differencing involved;
$q =$ order of the moving average part.

Clustering



[1]

Clustering

Whole time-series clustering is considered as clustering of a set of individual time-series with respect to their similarity. Here, clustering means applying conventional (usually) clustering on discrete objects, where objects are time-series.

Subsequence clustering means clustering on a set of subsequences of a time-series that are extracted via a sliding window, that is, clustering of segments from a single long time-series.

Time point clustering is another category of clustering which is seen in some papers. It is clustering of time points based on a combination of their temporal proximity of time points and the similarity of the corresponding values. This approach is similar to time-series segmentation. However, it is different from segmentation as all points do not need to be assigned to clusters, i.e., some of them are considered as noise.

[5]

Clustering of time series subsequences

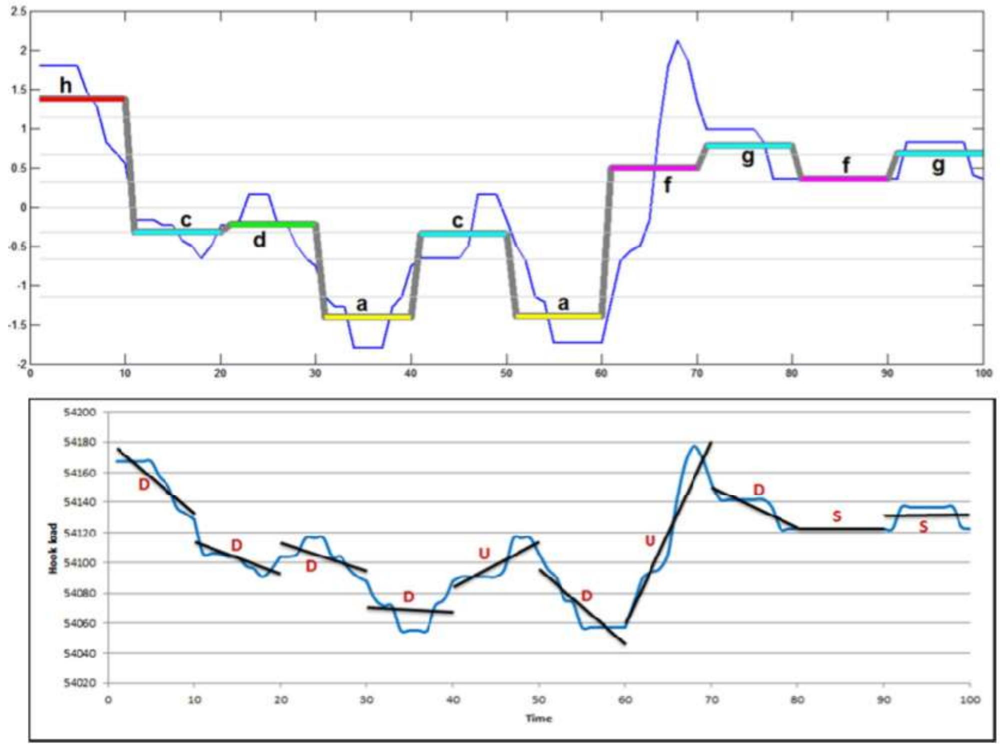
Subsequence Clustering: Given a single time series, sometimes in the form of streaming time series, individual time series (subsequences) are extracted with a sliding window. Clustering is then performed on the extracted time series subsequences.

Abstract

Given the recent explosion of interest in streaming data and online algorithms, clustering of time series subsequences, extracted via a sliding window, has received much attention. In this work we make a surprising claim. Clustering of time series subsequences is meaningless. More concretely, clusters extracted from these time series are forced to obey a certain constraint that is pathologically unlikely to be satisfied by any dataset, and because of this, the clusters extracted by any clustering algorithm are essentially random. While this constraint can be intuitively demonstrated with a simple illustration and is simple to prove, it has never appeared in the literature. We can justify calling our claim surprising, since it invalidates the contribution of dozens of previously published papers. We will justify our claim with a theorem, illustrative examples, and a comprehensive set of experiments on reimplementations of previous work. Although the primary contribution of our work is to draw attention to the fact that an apparent solution to an important problem is incorrect and should no longer be used, we also introduce a novel method which, based on the concept of time series motifs, is able to meaningfully cluster subsequences on some time series datasets.

[6]

Discretisation: SAX, PAA, TVA



[2]

Resources

[1] Selina Chu, Eamonn Keogh, David Hart and Michael Pazzani. *Iterative Deepening Dynamic Time Warping for Time Series*
[2] Bilal Esmael, Arghad Arnaout, Rudolf K. Fruhwirth and Gerhard Thonhauser. Multivariate Time Series Classification by Combining Trend-Based and Value-Based Approximations
[3] Nuno Castro and Paulo Azevedo. Multiresolution Motif Discovery in Time Series
[4] Florence Duchene, Catherine Garbay and Vincent Rialle. Mining Heterogeneous Multivariate Time-Series for Learning Meaningful Patterns: Application to Home Health Telecare
[5] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi and Teh Ying Wah. Time-series clustering - A decade review
[6] Eamonn Keogh and Jessica Lin. Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Researc