

# Time Series

Knowledge Discovery and Data Mining 2 (VU) (706.715)

Roman Kern

ISDS, TU Graz

2018-03-22

# Time Series Basics

## Introduction

# Definition

## Definition

A time series is a set of observations taken at specific times, usually at equal intervals.

- Two- or more-dimensional data with “time” as one dimension
- $t \rightarrow x(t)$ ,  $t$  discrete
- Special properties
  - 1 By the directed nature of time, there is an inherent data order
  - 2 By causality, preceding series members may influence subsequent member - never vice-versa
  - 3 The data cannot be assumed to be i.i.d. (independent and identically distributed)

Note: There is one independent variable (time) and a number of dependent variables (data).

- Interpretation
  - ▶ e.g. Study a phenomenon
- Forecasts (weather, financial, ...)
- Control (e.g. heating)
- Simulation

# Recommended Literature

The screenshot shows the Springer website interface. At the top left is the Springer logo. A search bar is located below the logo. A navigation menu includes Home, Subjects, Services, Products, Springer Shop, and About us. A promotional banner reads: "+++ Save 50% on Print Books, eBooks & Journals in Medicine! Browse now >> +++".

The main content area features a link to "Statistics" and a section for "Springer Texts in Statistics". The book "Time Series Analysis and Its Applications" is highlighted. The book cover is shown on the left, with the title and authors' names. The text on the right of the cover reads: "© 2011", "Time Series Analysis and Its Applications", "With R Examples", and "Authors: Shumway, Robert H., Stoffer, David S.". Below the cover is a link to "Show next edition".

Below the book information, a paragraph states: "Presents a balanced and comprehensive treatment of both time and frequency domain methods with accompanying theory", followed by a link to "see more benefits".

On the right side, there is a "Buy this book" section. It shows the price for the softcover as 89,95 € (price for India (gross)). A "Buy Softcover" button is present. Below the button, there are three bullet points: "ISBN 978-1-4614-2759-9", "Free shipping for individuals worldwide", and "Usually dispatched within 3 to 5 business days.". Payment options for VISA, MasterCard, American Express, PayPal, and iNIBELC are listed. There are links for "FAQ" and "Policy".

At the bottom right, there is a section for "Zahlen und Fakten" (Numbers and Facts) with a link to "Zitate" (Quotes) and a count of 128.

# Recommended Literature

The screenshot shows the Amazon product page for the book "Knowledge Discovery from Data Streams (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series) 1st Edition" by Joao Gama. The page features a dark navigation bar with the Amazon logo, a search bar, and various account and navigation links. Below the navigation bar, the breadcrumb trail indicates the book's location: Books > Computers & Technology > Computer Science. The main title and author information are prominently displayed, along with a 5-star rating and a "Look inside" button. The product image shows the book cover with a grid of vertical bars in shades of green and yellow. The pricing section highlights the Hardcover edition at \$94.95, with a "Buy new" button and a quantity selector. A "More Buying Choices" section at the bottom indicates 11 new and 8 used copies available.

amazon Books ▾ Prime student 50% off Prime

Departments ▾ Your Amazon.com Today's Deals Gift Cards & Registry Sell Help Hello, Sign in Account & Lists ▾ Orders Try Prime ▾ Cart

Books Advanced Search New Releases Best Sellers The New York Times® Best Sellers Children's Books Textbooks Textbook Rentals Sell Us Your Books Best Books of the Month

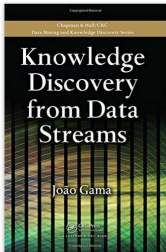
Books > Computers & Technology > Computer Science


## Knowledge Discovery from Data Streams (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series) 1st Edition


by Joao Gama (Author)

★★★★☆ ▾ 5 customer reviews

[Look inside ↴](#)



Kindle  **Hardcover** \$94.95 Other Sellers from \$67.19


Kindle  \$54.60 - \$79.76

**Buy new** \$94.95

**Only 2 left in stock (more on the way).**  
Ships from and sold by Amazon.com. Gift-wrap available.

**Want it Tuesday, March 21?** Order within **26 hrs 26 mins** and choose **One-Day Shipping** at checkout. [Details](#)

Qty: 1 ▾

 **Add to Cart**

[Turn on 1-Click ordering](#)

**Ship to:**  
Select a shipping address: ▾

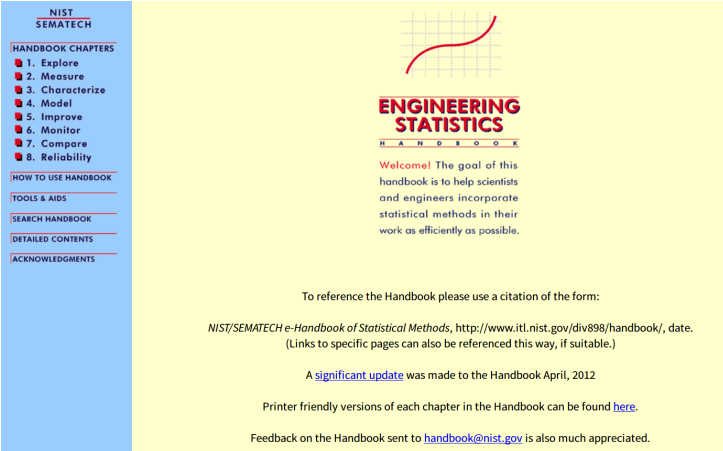
**More Buying Choices** 19 used & new from \$67.19

[11 New from \\$68.13](#) | [8 Used from \\$67.19](#)

[See All Buying Options](#)

ISBN-13: 978-1439826119  
ISBN-10: 1439826110

# Recommended Literature



**NIST  
SEMATECH**

**HANDBOOK CHAPTERS**

1. Explore
2. Measure
3. Characterize
4. Model
5. Improve
6. Monitor
7. Compare
8. Reliability

**HOW TO USE HANDBOOK**

**TOOLS & AIDS**

**SEARCH HANDBOOK**

**DETAILED CONTENTS**

**ACKNOWLEDGMENTS**

**ENGINEERING  
STATISTICS**  
H A N D B O O K

**Welcome!** The goal of this handbook is to help scientists and engineers incorporate statistical methods in their work as efficiently as possible.

To reference the Handbook please use a citation of the form:

*NIST/SEMATECH e-Handbook of Statistical Methods*, <http://www.itl.nist.gov/div898/handbook/>, date.  
(Links to specific pages can also be referenced this way, if suitable.)

A [significant update](#) was made to the Handbook April, 2012

Printer friendly versions of each chapter in the Handbook can be found [here](#).

Feedback on the Handbook sent to [handbook@nist.gov](mailto:handbook@nist.gov) is also much appreciated.

<https://www.itl.nist.gov/div898/handbook/>

# Challenges

- Trend (secular, growth)
- Seasonality (periodic)
- Cyclic Variation
  - ▶ Up/down, but not periodic in nature
  - ▶ e.g. economy, holidays, events
- Noise
  - ▶ Irregular patterns, not predictable
- Changes



## Stationary

- Needed by many forecasting methods
- Strictly stationary:
  - ▶ Probability of  $\{x_{t1}, x_{t2}, \dots, x_{tk}\}$
  - ▶ is equal to  $\{x_{t1+h}, x_{t2+h}, \dots, x_{tk+h}\}$
- Weakly stationary:
  - ▶ Mean is constant and does not depend on the time
  - ▶ Linear dependence between points does not depend on the time
    - ★ Autocovariance  $\gamma(s, t)$  depends on  $s$  and  $t$  only via  $|s - t|$

## Heteroscedasticity

- Variability change over time
- Opposed to homoscedasticity

## Autocorrelation Function (ACF)

- How much does a time series “repeat” itself
- Autocorrelation gives the peak at the best lag
- Partial autocorrelation function (PACF)
  - ▶ Leaves out the intermediate observations
- Cross-correlation
  - ▶ Extends this concepts to two time series

## Spectral Analysis

- Transform a time series into its frequencies
- e.g. Discrete Fourier Transform
  - ▶ Requires strict stationarity
  - ▶ Often applied on smoothed time series
- Wavelet analysis

# Modelling

Prediction, forecasting, ...

## Model the time series

- Additive model
  - ▶  $y = a + b + c + d$
- Multiplicative model
  - ▶  $y = a \times b \times c \times d$
- Which to choose?
  - ▶ If the amplitude of the components jointly prefer the multiplicative model

Note: The components are usually: trend, seasonality, cyclic variation, noise

## Stationary issues

- White noise model
  - ▶ Known mean and variance for Gaussian white noise
- Random-walk model
  - ▶ Model by adding random movements
- Stationarity
  - ▶ Convert a non-stationary time series into a stationary
  - ▶ By differencing, i.e. change between consecutive observations

## Detrending

- Remove a long running, consistent change in the time series
- Different approaches
  - ▶ Linear regression
  - ▶ Hi-pass filter
  - ▶ Derivative
  - ▶ Smoothing
    - ★ Averaging smoothing, e.g. moving average, weighted mean, ...
    - ★ Exponential smoothing, e.g. exponential mean, ...
- Tools
  - ▶ Visual inspection
  - ▶ Regression analysis (curve fitting)

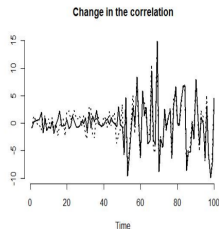
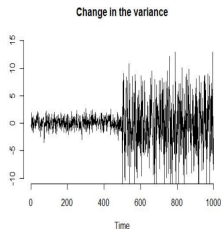
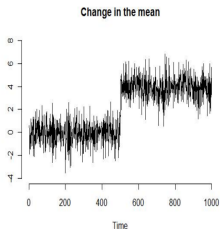
## Seasonality

- Periodic changes in the time series
- Different approaches
  - ▶ Introduce a season parameter
  - ▶ Autogression
  - ▶ Spectral methods
- Tools
  - ▶ Autocorrelation plot



## Change

- Drift or change detection
- Often tackled using sliding windows
- Link to burst detection



## Autoregressive Models

- Values depend on the previous values
- i.e. the history of the time series predicts its future
- $AR(p) : x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \dots + \alpha_p x_{t-p} + \epsilon_t$ 
  - ▶ where  $\epsilon_t$  is white noise
- The AR model is stationary
  - ▶ i.e.  $E[x_t] = 0$
- For  $AR(1)$  one can substitute the predecessor (until the beginning)
  - ▶  $x_t = \alpha x_{t-1} + \epsilon_t = \alpha(\alpha x_{t-2} + \epsilon_{t-1}) + \epsilon_t = \dots = \sum_k \alpha^k \epsilon_{t-k}$

## Moving Average Models

- Values depend on the previous deviations from the mean
- i.e. the history of the deviations predicts its future (lagged values of the forecast error)
- $MA(p) : x_t = c + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + \dots + \beta_p \epsilon_{t-p}$ 
  - ▶ ... where  $c$  is a constant ( $\mu$ )
  - ▶ ... where  $\epsilon_j$  are expected to be independent

## ARIMA

- ARIMA(p,d,q)
  - ▶ p is the number of autoregressive terms (AR)
  - ▶ d is the number of non-seasonal differences needed for stationarity (I)
  - ▶ q is the number of lagged forecast errors (MA)
- ARIMA(0,0,0) is a white noise model
- ARIMA(0,1,0) (without constant) is a random walk model
  - ▶ ARIMA(0,1,0) (with constant) is a random walk model with drift
- ARIMA(1,0,0) if the time series is stationary and autocorrelated

## Parameter Estimation

- Box-Jenkins methodology
- Strategy to identify parameters for ARIMA models
  - 1 Model identification
    - ★ e.g. autocorrelation plots
  - 2 Model estimation (parameters)
  - 3 Diagnostic checks on model adequacy
  - 4 If checks are positive
    - ★ Yes → use model for forecasting
    - ★ No → GOTO 1

See also: <https://people.duke.edu/~rnau/arimrule.htm>

## Alternative Approaches

- Many extensions to ARIMA proposed
- Exponential smoothing methods, e.g. Holt-Winters (triple exponential smoothing)
  - ▶ Only works for seasonal time series
  - ▶ Form of exponential smoothing
  - ▶ ... applied on level, trend and seasonal component
- Symbolic regression
- State space models
- Neural networks
  - ▶ e.g. recurrent neural networks (for example LSTMs)

## Alternative Approaches

- Prophet
  - ▶ Forecasting framework proposed by Facebook
  - ▶ Include insights from domain experts (analyst-in-the-loop)
  - ▶ Bayesian based curve fitting

Taylor, S. J., & Letham, B. (2017). Forecasting at Scale, 1-17.

## Forecasting

- Confidence intervals
  - ▶ Based on the assumptions of the model
  - ▶ i.e. it is not the true probability (only if all assumptions are met)
- Variance of forecasting risk = variance of intrinsic risk + variance of parameter risk



## Error Analysis

- Study the residuals (errors of the model)
  - ▶ There should be no information left in the residuals
- Evaluate the performance (goodness of fit)
  - ▶ Due to their nature, time series cannot be randomly split
    - ★ i.e. traditional cross-validation will not work
    - ★ replace with walk-forward sliding window testing
  - ▶ often via 10%-20% held out data
  - ▶ e.g. root mean square error, mean absolute error

## State Space Models

- Describe the system via states variables
- State variables cannot be directly measured
  - ▶ ... but inferred from the input/output information
- e.g. particle filter

## Kalman Filters

- Linear-Gaussian signal observation filter
- Requires a model based prediction
  - ▶ e.g. a physical model
- Combines the prediction with observations (i.e. noisy measurements)
- Updates the estimate of the state

# Time Series Feature Engineering

SAX, Windows & Friends

## Window

- Sliding window
  - ▶ Fixed size
  - ▶ Overlapping
    - 1 Size depends on the number of observations
    - 2 Size depends on the time span each window captures
- Tumbling window
  - ▶ Fixed size
  - ▶ Not overlapping
- Landmark window
  - ▶ Flexible sized windows, i.e. growing size
  - ▶ Starting at a fixed position (landmark)
- Variable sized windows
  - ▶ Window grows (and shrinks) according to some criteria
  - ▶ e.g. keep the mean of sub-windows lower than a threshold (ADWIN algorithm)

Perng, C.-S., Wang, H., Zhang, S. R. R., & Parker, D. S. S. (2000). Landmark: a new model for similarity-based pattern querying in time series databases.

## Window Based Feature Generation

- Build features from each window
  - ▶ Mean, Variance, Skewness, Number of Peaks, ...
  - ▶ → each window is transformed into a feature vector
- Optionally, compute the correlation between features
  - ▶ ... for feature selection/transformation

Have a look at: [https://github.com/blue-yonder/tsfresh/blob/master/tsfresh/feature\\_extraction/feature\\_calculators.py](https://github.com/blue-yonder/tsfresh/blob/master/tsfresh/feature_extraction/feature_calculators.py)

## Comparison of time series

- Euclidean distance
  - ▶ Compare two time aligned time series
- Dynamic Time Warping
  - ▶ Allow deviations in time

## Piecewise Aggregate Approximation (PAA)

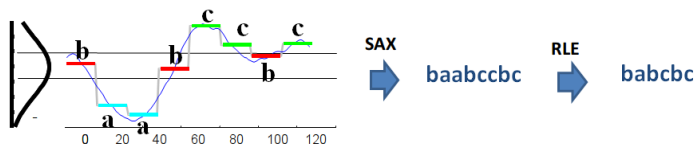
- Quantise the time series on time intervals
- ... by typically taking the average of the values within the interval
- Thereby effectively compressing the time series



# Time Series Feature Engineering

## Symbolic Approximation (SAX)

- Discretise the values of the time series
- ... using a fixed vocabulary of symbols
  - ▶ Size of bins may be fixed or computed based on some criteria
- Good representation for indexing of time series
- Good representation for novelty detection in time series (outlier detection)
- Good representation for frequent pattern mining in time series



## Piecewise Linear Approximation

- Replace the original time series by a series of linear approximations
- ... using a little linear pieces as necessary
- Basic approaches
  - ▶ Top down: start with a single linear approximation and iteratively split
  - ▶ Bottom up: Merge small linear approximations iteratively into larger pieces



Keogh, E., Chu, S., Hart, D., & Pazzani, M. (2001). An online algorithm for segmenting time series.

## Knowledge Discovery Algorithms

- Clustering of time series
  - ▶ Often based on distance measures
- Outlier detection
  - ▶ Identify anomalies in the data
- Classification of time series
  - ▶ e.g. k-NN with Euclidean distance

# Tools

## Practical aspects

- Python

- ▶ Prophet
- ▶ TS-Fresh
- ▶ Pandas, NumPy, scikit-learn, Statsmodels
- ▶ Orange (extension)

- R

- ▶ <http://www.statmethods.net/advstats/timeseries.html>
- ▶ <https://cran.r-project.org/web/views/TimeSeries.html>
- ▶ <https://github.com/robjhyndman/>

- MatLab

- ▶ Extensive documentation available

- Java

- ▶ JMotiv, Weka, ...

# The End

Next: Q & A Session I