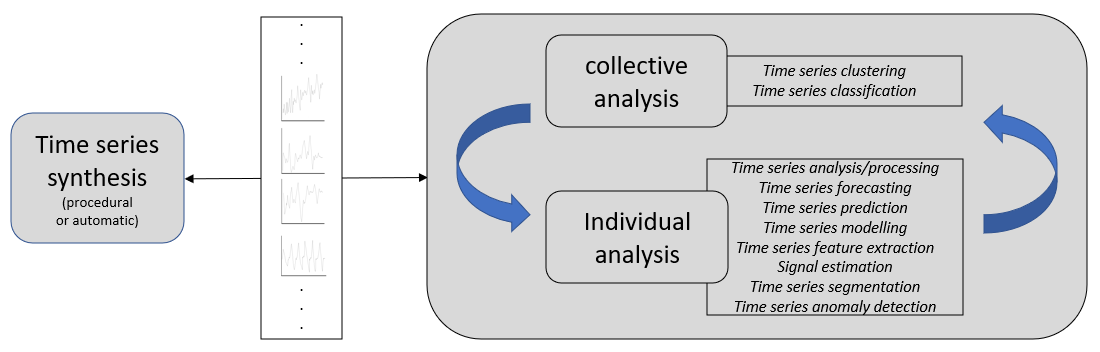
**Time series analysis and methods**

In time series domain there are several tasks that can be performed to analyze time series. Some of the tasks are Time series analysis/processing, Time series forecasting, time series prediction, time series modeling, time series synthesis, time series feature extraction, signal estimation, time series segmentation, time series clustering, time series classification, time series anomaly detection. This report is the survey of methods involved in the above-mentioned time series tasks and implementation of all the methods in a comprehensive python library. Several books, papers and existing time series libraries are already collected. The proposed comprehensive library contains the usage and extensions of existing libraries, or full implementation. The following are some of the time series domain tasks that can be performed:

*Time series analysis/processing, Time series forecasting, Time series prediction, Time series modelling, Time series synthesis, Time series feature extraction, Signal estimation, Time series segmentation, Time series clustering, Time series classification, Time series anomaly detection.*

The above tasks can be divided into two types: collective analysis based on a set of time series data and individual analysis based on a single time series. The following figure describes the flow of time series tasks that can be performed:



Variations in the methods

Variations in the data

**Analysis philosophy for time series:**

1. ***Procedural Data synthesis:*** Firstly, the data needs to be generated with all the time series components variations like Uniform, monotonic, erratic, sparse, seasonal, cyclic and even hybrids of them. The noise structure in the signals must be modelled too. The modeling must be procedural modeling as opposed to automatic data generation like deep learning methods.
2. ***Data analysis:*** After modeling all the components using robust methods, then we need to be able to detect the components in the data, extract major features in the data, decompose the time series for forecasting tasks, and perform clustering and classification analysis of the entire dataset. First perform collective analysis to form groups and within each group perform individual analysis.
3. ***Automatic data synthesis:*** Generate the data again automatically using learning-based mechanism from the analysis results.

Any complex time series have the following characteristics: trend, seasonality, cyclicality, outliers, noise, breakpoints, etc. and more. Create a separate library of methods to ***detect*** and ***quantify*** the above-mentioned characteristics.

*To understand the properties and visual appearance of the time series data, first looking at the real time series data from various domains and their characteristics. It is required that the domain is properly understood to check how the time series data may look in that domain.*

***Time series (Wikipedia):***

[***https://en.wikipedia.org/wiki/Time\_series***](https://en.wikipedia.org/wiki/Time_series)

Some examples of time series – heights of ocean tide, counts of sunspots, daily closing value of stock market prices. Time series are used in various domains like – statistics, signal processing, pattern recognition, econometrics (use of mathematical methods in economic systems), mathematical finance (concerned with mathematical modeling of financial markets), weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering and more. Some of the topics in time series domain are:

*Time series analysis:* methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

*Time series forecasting:* use of a model to predict future values based on previously observed values.

*Regression analysis:* relation between one or more different time series.

*Interrupted time series analysis:* detect changes in the evolution of time series from before to after some intervention which may affect the underlying variable.

Time series analysis is different from *cross sectional data*. The former involves ordering in the data while the latter does not have any ordering in the data. Time series analysis is also distinct from *spatial data analysis* where there is ordering in two dimensions.

A stochastic model for time series will most generally assume that observations close together in time will be more closely related than observations further apart. Time series can also be applied to discrete symbolic data (eg: sequence of characters, letters, words in English).

Time series methods can be divided into *frequency-domain methods (spectral analysis, wavelet analysis)* and *time domain methods (autocorrelation, cross correlation).*

Time series data can be thought of as 1-D version of panel data. A panel data has several records of data where in each single record is identified by a multidimensional identifier. If the identifier is only time, then the panel data is time series data, if the identifier has another variable other than time then it is panel data. If time is not included in the identifiers, then it is cross sectional data.

The goal of time series analysis is different in various domains:

* In the context of statistics, econometrics, quantitative finance, seismology, meteorology, and geophysics, the goal is to do time series forecasting.
* In the context of signal processing, control engineering and communication engineering it is used for signal detection.
* In data mining, pattern recognition and machine learning, it is used for clustering, classification, query by content, anomaly detection.

*Some of the tasks in time series analysis:*

|  |  |
| --- | --- |
| Exploratory analysis | * Autocorrelation analysis to examine serial dependence. * Spectral analysis to examine cyclic behavior. * Separation into components – trend, seasonality, slow and fast variation, cyclical irregularity |
| Curve fitting | * Interpolation, regression, Extrapolation |
| Function approximation | * Approximate to a function from a specified class of functions. |
| Prediction (statistical: from sample to population) and forecasting | * Given a sample of time series predict the model of the time series population. * Forecasting the future data points of time series by statistical inference about the population. |
| Classification | * To assign a category to each time series instance based on its characteristics. |
| Signal estimation | * Filtering of noise in the acquired signal to determine the original signal. (Kalman filter, estimation theory, signal processing, harmonic analysis, Fourier transforms) |
| Segmentation | * Representing a time series into sequence of individual segments, each with its own characteristic properties. Changepoint detection or Markov jump linear systems are an example. |

Time series models can be divided into following categories:

* ARIMA models
* ARFIMA (fractionally integrated models – which are generalizations of ARIMA models)
* Vector-based models of ARIMA (‘V’)
* Exogenous forcing time series-based models (‘X’)
* Non-linear dependence of level of the model with previous data points to produce chaotic time series – Non-linear auto regressive exogenous models.
* Non-linear models to represent changes of variance over time (heteroscedasticity), doubly stochastic models.
* Markov switching multifractal techniques (MSMF) to model volatility evolution.\
* HMMs, Markov models.

Much of the theory of time series is developed under the assumption of (i) stationarity and (ii) Ergodicity. But time series analysis methods can also be applied to seasonally stationary and non-stationary time series after modifications to the time series. Time frequency analysis is applied when amplitude of frequency components in the time series changes with time. *Time frequency representation* of time series is used in such cases.

Some of the tools used in time series domain are mentioned in the Wikipedia website and can become useful: <https://en.wikipedia.org/wiki/Time_series>

*Note: The real datasets are found in famous archives, but the characteristics present in those datasets are not tagged to them. Such datasets with characteristics attached can be found only in relevant textbooks and papers where they explain the data generation process also. First go through several datasets in different fields visually, then go through textbooks, papers etc., for characteristic datasets.*

***Dataset archives:***

*Monash time series forecasting archive: May 2021, Department of Data science and Artificial Intelligence Monash University, Department of Econometrics and Business Statistics Monash university, University of Sydney.*

Paper: <https://arxiv.org/abs/2105.06643>

Website: <https://forecastingdata.org/>

Codes: <https://github.com/rakshitha123/TSForecasting>

**Features:**

* Comprehensive time series archive for forecasting that contain datasets from similar sources to evaluate performance of new global forecasting methods. 20 datasets and 6 single very long datasets - 26 in total with 50 dataset variations in total based on frequency, missing values etc.
* Includes both real world and competition datasets.
* New time series data format called ***.tsf format*** based on Weka ARFF file format and also overcomes shortcomings in the .ts format of sktime time series repository. The new format also stores meta information of the time series.
* Different characteristics in the data like frequency, series lengths, inclusion of missing values.
* Feature analysis using ***tsfeatures*** and ***catch22featrues*** in the datasets to identify similarities and differences.
* Benchmarking with a set of standard baseline forecasting methods – traditional univariate methods and global forecasting methods across 8 error metrics.
* Dataset domains: tourism, banking, web, energy, sales, economics, transportation, health, nature.
* Contains datetime stamps.
* The archive size is 2.55 GB

**UCR repository:** time series classification archive – 128 datasets of non-normalized time series of varying lengths and missing values. Only univariate time series data.

**UEA repository:** multivariate time seriesclassification archive – 30 datasets of equal length multivariate time series without missing values.

**Monash, UCR, UEA repository:** time series extrinsic regression archive – 19 multidimensional time series datasets.

**Examples of sets of time series from similar sources:**

* All product sales from a particular store
* All smart meter readings in a particular city or state.

**Time series benchmarking archives for traditional univariate forecasting:**

* Time series data library (Hyndman 2018)
* ForeDeCk (2019)
* M3 and M4 competitions

**File format (.tsf):**

* Each data file contains (i) attributes that are constant throughout the whole dataset. (ii) attributes that constant throughout a time series (iii) attributes that are particular to each datapoint.
* Tags are used to describe the meta information of the corresponding dataset like

@frequency (seasonality)

@horizon (expected forecast horizon)

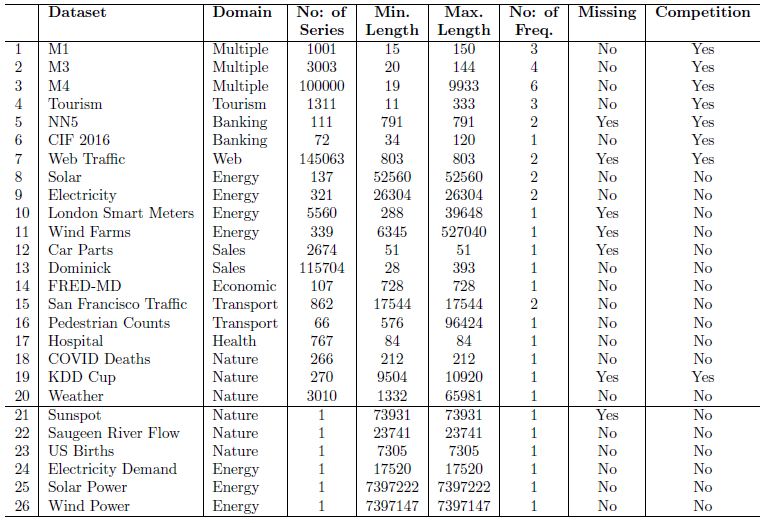
@missing (contains missing values or not)

@equallength (whether series have equal lengths)

@attribute (*series\_name; start\_timestamp*)

@data

* Each of the 26 datasets have a .tsf file associated with it in the archive.

***Datasets in the archive:***

*UCR time series archive: Sep 2019, Department of computer science and engineering – UCR USA, School of computing sciences – UEA UK.*

Paper: <https://arxiv.org/abs/1810.07758>

Wesite: <https://www.cs.ucr.edu/~eamonn/time_series_data_2018/>

Supporting webpage: <https://www.cs.ucr.edu/~hdau001/ucr_archive/>

Codes: <https://drive.google.com/drive/folders/11V1G94I-7KVTSXkl9HdP2QxS8pQMEXAh>

**Features:**

* 128 time series datasets for time series classification.
* Both already normalized data and raw data are present in the archive.
* Uses Matlab for experiments
* Briefing document includes the password to open the archive and example plots of time series in different datasets.
* The data contains train and test data in .tsv file format. Timestamps are not included in the data, just the ordering of the data.
* Excel file includes the description of all the domains from which data are gathered.
* Does not contains datetime stamps
* The archive size is 853 MB

*UEA multi-variate time series archive: Oct 2018, School of computing sciences – UEA UK, Department of computer science and engineering – UCR USA.*

Paper: <https://arxiv.org/abs/1811.00075>

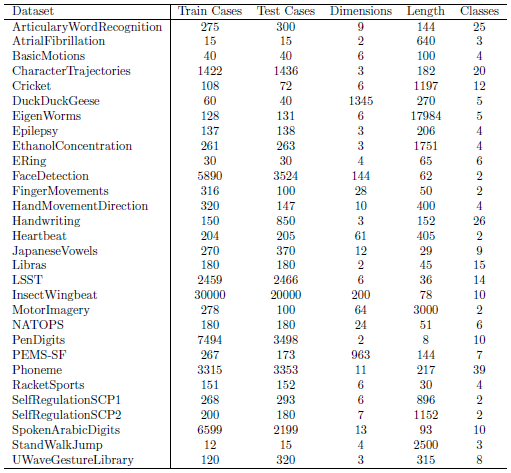
Website: <https://www.timeseriesclassification.com/> (contains UCR archive data also in WEKA format and sktime .ts format)

Code: <https://bitbucket.org/TonyBagnall/time-series-classification/src/master/> ; <https://github.com/uea-machine-learning/tsml> ; <https://github.com/alan-turing-institute/sktime>

**Features:**

* Multi-variate time series classification archive where more than one time series are associated with each class label.
* Extension of Mustafa Baydogan’s archive (<http://www.mustafabaydogan.com/index.php>) of 12 multivariate datasets which has drawbacks of being very small, non-independent.
* Includes 30 datasets with train/test splits. The datasets are of equal length and no missing values are present.
* The datasets are in Weka multi-instance format. Code to split the multivariate ARFF is also available.
* The datasets are available at the page: <https://www.timeseriesclassification.com/dataset.php> ; [www.timeseriesclassification.com/Downloads/MultivariateTSCProblems.zip](http://www.timeseriesclassification.com/Downloads/MultivariateTSCProblems.zip)
* Sources for this dataset include UCI machine learning archive, series of Brain computer interface competitions, some made by UEA. Some of the domains of the datasets include Human activity recognition, Motion classification, ECG classification, EEG/MEG classification, Audio spectra classification, and others.
* The data source, description and download links are provided separately in a table format for each dataset in the archive on the website: <http://www.timeseriesclassification.com/dataset.php>
* Does not contains datetime stamps.
* The archive size is 5.33 GB

***Datasets in the archive:***



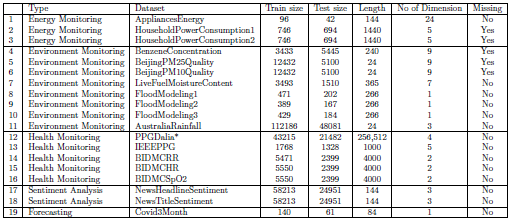
*Monash, UEA, UCR time series extrinsic regression archive: Oct 2020, Faculty of Information technology, Monash university.*

Paper: <https://arxiv.org/abs/2006.10996>

Website: <http://tseregression.org/>

Code: <https://github.com/ChangWeiTan/TSRegression>

**Features:**

* Includes 19 time series datasets from different domains with varying number of dimensions, unequal length dimensions and missing values.
* Datasets are from 5 application areas – Energy monitoring, Environment monitoring, Health monitoring, Sentiment analysis and Forecasting.
* The datasets are formatted with ***.ts format*** used in tsml and sktime time series machine learning repos.
* Forecasting tasks can also be tackled using extrinsic regression.
* Different regression algorithms are tested like FPCR, FPCR with B-spline, Grid-search optimized SVR, RF with 100 trees, XGBoost with 100 trees, NN-ED, NN-DTW, FCN, ResNet, Inception network, Rocket.
* Contains datetime stamps.
* The archive size is 4.09 GB

*Note: Use a time series database to access the dataset quickly from a group and plot, do analysis on the data on the fly in a GUI. HDF5 is a good database used by many scientific research organizations.*

Two datasets in hand now:

* Monash forecasting and extrinsic regression archive (real datasets)
* Procedural generation-based dataset (synthetic dataset)

Perform time series analysis on both datasets and note down the differences. Also note down the characteristics in the real datasets and apply it in synthetic datasets.

**Characteristics in real time series datasets:**

|  |  |
| --- | --- |
| ***Time series Component*** | ***Characteristics*** |
| 1. Types of trends: | * 1. Linearly increasing, decreasing, and remaining constant   2. Steep increases or decreases   3. Exponentially increasing, decreasing, and remining constant   4. logarithmically increasing, decreasing, and remaining constant   5. Polynomial trends with increasing, decreasing, and remaining constant   6. constant trend   7. heavy jumps in the trend curve at few points   8. heavy jumps in the trend curve at periodic intervals   9. jump regions   10. sparse regions |
| 1. Types of noise: | 1. High variance noise with short/long term memory/no memory 2. Random walk behavior around trend curve 3. One or two outliers (peaks/dips) 4. Regular outlier (peaks/dips) (non-periodic) 5. Periodic outliers (peaks/dips) 6. Changes in variability – volatility 7. Erratic noise with high range up and down variations 8. Periodic changes in variability around trend curve |
| 1. Types of periodic variations: | * 1. High frequency or low frequency variations   2. Peak/mountain variations   3. Triangular variations   4. custom pattern variations   5. varying amplitudes variations   6. multiple frequency variations (very high, moderate, and low combinations)   7. exponentially increasing amplitudes |

A complex time series can now be modelled using combinations of different models above.