Package 'tsrobprep'

November 5, 2020

Title Robust Preprocessing of Time Series Data

Version 0.0.0.2

Date 2020-11-05

Description Methods for handling the missing values outliers are introduced in this package. The recognized missing values and outliers are replaced using a model-based approach. The model may consist of both autoregressive components and external regressors. The methods work robust and efficient, and they are fully tunable. The primary motivation for writing the package was preprocessing of the energy systems data, e.g. power plant production time series, but the package could be used with any time series data.

Depends R (>= 3.2.0)

License MIT + file LICENSE

Encoding UTF-8

Imports Matrix, quantreg

LazyData true

RoxygenNote 7.1.1

NeedsCompilation no

Author Michał Narajewski [aut, cre] (<https://orcid.org/0000-0002-3115-0162>), Florian Ziel [aut] (<https://orcid.org/0000-0002-2974-2660>), Jens Kley-Holsteg [ctb]

Maintainer Michał Narajewski <michal.narajewski@uni-due.de>

Repository CRAN

Date/Publication 2020-11-05 10:40:02 UTC

R topics documented:

`auto_data_cleaning` *Perform automatic data cleaning of time series data*

Description

Returns a matrix or a list of matrices with imputed missing values and outliers. The function automatizes the usage of functions model_missing_data, handle_outliers and impute_modelled_data. The function is designed for numerical data only.

Usage

```
auto_data_cleaning(
  data,
  tau = 0.5,
  S,
  extreg = NULL,
  no.of.last.indices.to.fix = S,
  indices.to.fix = NULL,
  outlier.lower.cap = 2 * log(dim(as.matrix(data))[1]),
  outlier.upper.cap = log(dim(as.matrix(data))[1]),
  margin = outlier.upper.cap,
  max.periods.to.smooth = 48,
  lags = c(1, 2, S, S + 1, 7 * S, 7 * S + 1, -1, -2, -S, -S - 1, -7 * S, -7 * S - 1),
  n.best.extreg = NULL,
  use.data.as.ext = FALSE,
  lag.externals = FALSE,
  consider.as.missing = NULL,
  whole.period.missing.only = FALSE,
  min.val = -Inf,
  max.val = Inf,
  digits = 3,
  ...
)
```

Arguments

`data`	an input vector, matrix or data frame of dimension nobs x nvars containing missing values; each column is a variable.
`tau`	the quantile(s) of the missing values to be estimated in the quantile regression. Tau accepts all values in (0,1), the default is 0.5.
`S`	a number of observations per period, e.g. per day.
`extreg`	a vector, matrix or data frame of data containing external regressors; each column is a variable.
`no.of.last.indices.to.fix`	a number of observations in the tail of the data to be fixed, by default set to S.

`indices.to.fix` indices of the data to be fixed. If NULL, then it is calculated based on the no.of.last.indices.to.fix parameter. Otherwise, the no.of.last.indices.to.fix parameter is ignored.

`outlier.lower.cap`
a number of observations that should fall above (below) the outlier lower cap, used for the calculation of the interquantile range. By default set to 2log(nobs).

`outlier.upper.cap`
a number of observations that should fall above (below) the outlier upper cap, used for the calculation of the interquantile range. By default set to log(nobs).

`margin` a value that indicates the number of interquantile ranges above (below) the quantile upper cap that is still considered not to be an outlier. By default equals outlier.upper.cap.

`max.periods.to.smooth`
maximum number of S periods to be smoothed in the detection of outliers in the differential series. By default set to 48.

`lags` a numeric vector with the lags to use in the autoregression. Negative values are accepted and then also the "future" observations are used for modelling. The default values are constructed under the assumption that S describes daily periodicity.

`n.best.extreg` a numeric value specifying the maximal number of considered best correlated external regressors (selected in decreasing order). If NULL, then all variables in extreg are used for modelling.

`use.data.as.ext`
logical specifying whether to use the remaining variables in the data as external regressors or not.

`lag.externals` logical specifying whether to lag the external regressors or not. If TRUE, then the algorithm uses the lags specified in parameter lags.

`consider.as.missing`
a vector of numerical values which are considered as missing in the data.

`whole.period.missing.only`
if FALSE, then all observations which correspond to the values of consider.as.missing are treated as missings. If TRUE, then only consecutive observations of specified length are considered (length is defined by S).

`min.val` a single value or a vector of length nvars providing the minimum possible value of each variable in the data. If a single value, then it applies to all variables. By default set to -Inf.

`max.val` a single value or a vector of length nvars providing the maximum possible value of each variable in the data. If a single value, then it applies to all variables. By default set to Inf.

`digits` integer indicating the number of decimal places allowed in the data, by default set to 3.

`...` additional arguments for the rq.fit.fnb algorithm.

Details

The function calls handle_outliers to detect outliers, removes them and applies model_missing_data function. For details see the functions' respective help sections.

Value

A matrix or a list of matrices with imputed missing values or outliers.

See Also

model_missing_data, handle_outliers, impute_modelled_data

Examples

```
autoclean <- auto_data_cleaning(data = GBload[,-1], S = 48, tau = 0.5,
             no.of.last.indices.to.fix = dim(GBload)[1], consider.as.missing = 0,
             min.val = 0)
```

GBload *The electricity actual total load in Great Britain in year 2018*

Description

A dataset containing the electricity actual total load (MW) in Great Britain in year 2018 presented in half-hour interval. Each data point regards 30 minutes of electricity load starting at given time. The data consists of both missing values and outliers.

Usage

```
GBload
```

Format

A data frame with 17520 rows and 2 variables:

Date date indicating the delivery beginning of the electricity

Load actual electricity load in MW ...

Source

https://transparency.entsoe.eu/

handle_outliers *Detect and model unreliable outliers of time series data*

Description

Returns an object of class "tsrobprep" which contains the original data and the modelled values to be imputed. The function handle_outliers detects unreliable outliers given specific threshold and suggests replacement using model_missing_data function. The function analyses both absolute level of the data and the differential series. The function is designed for numerical data only.

Usage

```
handle_outliers(
  data,
  tau = 0.5,
  S,
  no.of.last.indices.to.fix = S,
  indices.to.fix = NULL,
  outlier.lower.cap = 2 * log(dim(as.matrix(data))[1]),
  outlier.upper.cap = log(dim(as.matrix(data))[1]),
  margin = outlier.upper.cap,
  max.periods.to.smooth = 48,
  extreg = NULL,
  use.data.as.ext = FALSE,
  min.val = -Inf,
  max.val = Inf,
  ...
)
```

Arguments

data	an input vector, matrix or data frame of dimension nobs x nvars possibly containing unreliable outliers; each column is a variable.
tau	the quantile(s) of the replaced values to be estimated in the quantile regression. Tau accepts all values in (0,1), the default is 0.5.
S	a number of observations per period, e.g. per day.
no.of.last.indices.to.fix	a number of observations in the tail of the data to be fixed, by default set to S.
indices.to.fix	indices of the data to be fixed. If NULL, then it is calculated based on the no.of.last.indices.to.fix parameter. Otherwise, the no.of.last.indices.to.fix parameter is ignored.
outlier.lower.cap	a number of observations that should fall above (below) the outlier lower cap, used for the calculation of the interquantile range. By default set to 2log(nobs).

`outlier.upper.cap`
: a number of observations that should fall above (below) the outlier upper cap, used for the calculation of the interquantile range. By default set to log(nobs).

`margin`
: a value that indicates the number of interquantile ranges above (below) the quantile upper cap that is still considered not to be an outlier. By default equals outlier.upper.cap.

`max.periods.to.smooth`
: maximum number of S periods to be smoothed in the detection of outliers in the differential series. By default set to 48.

`extreg`
: a vector, matrix or data frame of data containing external regressors.

`use.data.as.ext`
: logical specifying whether to use the remaining variables in the data as external regressors or not.

`min.val`
: a single value or a vector of length nvars providing the minimum possible value of each variable in the data. If a single value, then it applies to all variables. By default set to -Inf.

`max.val`
: a single value or a vector of length nvars providing the maximum possible value of each variable in the data. If a single value, then it applies to all variables. By default set to Inf.

`...`
: additional arguments for the model_missing_data function.

Details

The function recognizes two types of outliers: on the absolute level and in the differential series. As outliers are considered all observations that in any of the above mentioned series fall above (below)

interquantile_range * margin

where margin is given by the user and

interquantile_range = upper_quantile - lower_quantile.

The upper and lower quantiles are derived from the data based on the outlier.upper.cap and outlier.lower.cap arguments, respectively. In the case of outliers in the differential series, the function uses also smoothing, i.e. if after replacing of the outliers in the differential series, the series still exhibits outliers, the neighbouring observations are also replaced in order to smooth the "jump". The outlier values are replaced using the model_missing_data function.

The modelled values can be imputed using impute_modelled_data function.

Value

An object of class "tsrobprep" which contains the original data, the indices of the data that were modelled, the given quantile values, a list of sparse matrices with the modelled data to be imputed and a list of the numbers of models estimated for every variable.

See Also

model_missing_data, impute_modelled_data, auto_data_cleaning

Examples

```
outliers.handled <- handle_outliers(data = GBload[,-1], S = 48, tau = 0.5,
                    no.of.last.indices.to.fix = dim(GBload)[1], min.val = 0)
outliers.handled$estimated.models
outliers.handled$replaced.indices
new.GBload <- impute_modelled_data(outliers.handled)
```

`impute_modelled_data` *Impute modelled missing time series data*

Description

Returns a matrix or a list of matrices with imputed missing values or outliers. As argument the function requires an object of class "tsrobprep" and the quantiles to be imputed.

Usage

```
impute_modelled_data(object, tau = NULL)
```

Arguments

`object`	an object of class "tsrobprep" that is an output of functions model_missing_data and handle_outliers.
`tau`	the quantile(s) of the missing values to be imputed. tau should be a subset of the quantile values present in the "tsrobprep" object. By default all quantiles present in the object are used.

Value

A matrix or a list of matrices with imputed missing values or outliers.

See Also

model_missing_data, handle_outliers, auto_data_cleaning

Examples

```
model.miss <- model_missing_data(data = GBload[,-1], S = 48, tau = 0.5,
              no.of.last.indices.to.fix = dim(GBload)[1], consider.as.missing = 0,
              min.val = 0)
model.miss$estimated.models
model.miss$replaced.indices
new.GBload <- impute_modelled_data(model.miss)
```

`model_missing_data` *Model missing time series data*

Description

Returns an object of class "tsrobprep" which contains the original data and the modelled missing values to be imputed. The function model_missing_data models missing values in a time series data using the quantile regression implemented in quantreg package. The model uses autoregression on the time series as explanatory variables as well as the provided external variables. The function is designed for numerical data only.

Usage

```
model_missing_data(
  data,
  tau = 0.5,
  S,
  no.of.last.indices.to.fix = S,
  indices.to.fix = NULL,
  lags = c(1, 2, S, S + 1, 7 * S, 7 * S + 1, -1, -2, -S, -S - 1, -7 * S, -7 * S - 1),
  extreg = NULL,
  n.best.extreg = NULL,
  use.data.as.ext = FALSE,
  lag.externals = FALSE,
  consider.as.missing = NULL,
  whole.period.missing.only = FALSE,
  min.val = -Inf,
  max.val = Inf,
  digits = 3,
  ...
)
```

Arguments

`data`	an input vector, matrix or data frame of dimension nobs x nvars containing missing values; each column is a variable.
`tau`	the quantile(s) of the missing values to be estimated in the quantile regression. Tau accepts all values in (0,1), the default is 0.5.
`S`	a number of observations per period, e.g. per day.
`no.of.last.indices.to.fix`	a number of observations in the tail of the data to be fixed, by default set to S.
`indices.to.fix`	indices of the data to be fixed. If NULL, then it is calculated based on the no.of.last.indices.to.fix parameter. Otherwise, the no.of.last.indices.to.fix parameter is ignored.

`lags` a numeric vector with the lags to use in the autoregression. Negative values are accepted and then also the "future" observations are used for modelling. The default values are constructed under the assumption that S describes daily periodicity.

`extreg` a vector, matrix or data frame of data containing external regressors; each column is a variable.

`n.best.extreg` a numeric value specifying the maximal number of considered best correlated external regressors (selected in decreasing order). If NULL, then all variables in extreg are used for modelling.

`use.data.as.ext` logical specifying whether to use the remaining variables in the data as external regressors or not.

`lag.externals` logical specifying whether to lag the external regressors or not. If TRUE, then the algorithm uses the lags specified in parameter lags.

`consider.as.missing` a vector of numerical values which are considered as missing in the data.

`whole.period.missing.only` if FALSE, then all observations which correspond to the values of consider.as.missing are treated as missings. If TRUE, then only consecutive observations of specified length are considered (length is defined by S).

`min.val` a single value or a vector of length nvars providing the minimum possible value of each variable in the data. If a single value, then it applies to all variables. By default set to -Inf.

`max.val` a single value or a vector of length nvars providing the maximum possible value of each variable in the data. If a single value, then it applies to all variables. By default set to Inf.

`digits` integer indicating the number of decimal places allowed in the data, by default set to 3.

`...` additional arguments for the rq.fit.fnb algorithm.

Details

The function uses quantile regression in order to model missing values and prepare it for imputation. In this purpose the rq.fit.fnb function from quantreg package is used. The function computes the quantile regression methods utilizing the Frisch-Newton algorithm for user-specified quantile values. The modelled values can be imputed using impute_modelled_data function.

Value

An object of class "tsrobprep" which contains the original data, the indices of the data that were modelled, the given quantile values, a list of sparse matrices with the modelled data to be imputed and a list of the numbers of models estimated for every variable.

See Also

impute_modelled_data, handle_outliers, auto_data_cleaning

Examples

```
model.miss <- model_missing_data(data = GBload[,-1], S = 48, tau = 0.5,
              no.of.last.indices.to.fix = dim(GBload)[1], consider.as.missing = 0,
              min.val = 0)
model.miss$estimated.models
model.miss$replaced.indices
new.GBload <- impute_modelled_data(model.miss)
```

Index