

Package ‘StatMatch’

September 13, 2009

Type Package

Title Statistical Matching

Version 0.8

Date 2009-09-11

Author Marcello D’Orazio

Maintainer Marcello D’Orazio <madorazi@istat.it>

Depends R (>= 2.7.0), proxy, lpSolve

Suggests optmatch

Description This package provides some R functions to perform statistical matching between two data sources sharing a number of common variables. These functions can also be used to impute missing values in data sets through hot-deck methods.

License GPL (>= 2)

LazyLoad no

Repository CRAN

Date/Publication 2009-09-13 09:08:37

R topics documented:

create.fused	2
fact2dummy	3
gower.dist	5
mixed.mtc	7
NND.hotdeck	12
RANDwNND.hotdeck	17

Index	21
--------------	-----------

create.fused	Creates a matched (synthetic) dataset
--------------	---------------------------------------

Description

Creates a *synthetic* data frame after the statistical matching of two data sources at *micro* level.

Usage

```
create.fused(data.rec, data.don, mtc.ids,  
             z.vars, dup.x=FALSE, match.vars=NULL)
```

Arguments

data.rec	A matrix or data frame that has been used as <i>recipient</i> in the statistical matching application.
data.don	A matrix or data frame that has been used as the <i>donor</i> in the statistical matching application.
mtc.ids	A matrix with two columns. Each row must contain the name or the index of the recipient record (row) in <code>data.don</code> and the name or the index of the corresponding donor record (row) in <code>data.don</code> . Note that this type of matrix is returned by the functions <code>NND.hotdeck</code> and <code>RANDwNND.hotdeck</code> .
z.vars	A character vector with the name of the variables available only in <code>data.don</code> that should be “donated” to <code>data.rec</code> .
dup.x	Logical. When <code>TRUE</code> the values of the matching variables in <code>data.don</code> are also “donated” to <code>data.rec</code> . The name of the matching variables has to be specified with the argument <code>match.vars</code> . To avoid confusion, the matching variables added to <code>data.rec</code> are renamed by adding the suffix “don”. By default <code>dup.x=FALSE</code> .
match.vars	A character vector with the names of the matching variables. It has to be specified only when <code>dup.x=TRUE</code> .

Details

This function allows to create the synthetic (or fused) data set after the application of a statistical matching in a *micro* framework. For details D’Orazio *et al.* (2006).

Value

The `data.rec` data frame with the `z.vars` filled in and, when `dup.x=TRUE`, with the values of the matching variables for the donor records.

Author(s)

Marcello D’Orazio (madorazi@istat.it)

References

D’Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley, Chichester.

See Also

[NND.hotdeck](#) [RANDwNND.hotdeck](#)

Examples

```
lab <- c(1:15, 51:65, 101:115)
iris.rec <- iris[lab, c(1:3,5)] # recipient data.frame
iris.don <- iris[-lab, c(1:2,4:5)] # donor data.frame

# Now iris.rec and iris.don have the variables
# "Sepal.Length", "Sepal.Width" and "Species"
# in common.
# "Petal.Length" is available only in iris.rec
# "Petal.Width" is available only in iris.don

# find the closest donors using NND hot deck;
# distances are computed on "Sepal.Length" and "Sepal.Width"

out.NND <- NND.hotdeck(data.rec=iris.rec, data.don=iris.don,
  match.vars=c("Sepal.Length", "Sepal.Width"), don.class="Species")

# create synthetic data.set, without the duplication of the matching variables
fused.0 <- create.fused(data.rec=iris.rec, data.don=iris.don,
  mtc.ids=out.NND$mtc.ids, z.vars="Petal.Width")

# create synthetic data.set, with the "duplication" of the matching variables
fused.1 <- create.fused(data.rec=iris.rec, data.don=iris.don,
  mtc.ids=out.NND$mtc.ids, z.vars="Petal.Width", dup.x=TRUE,
  match.vars=c("Sepal.Length", "Sepal.Width"))
```

fact2dummy

Transforms a categorical variable in a set of dummy variables

Description

Transforms a factor or more factors contained in a data frame with a set of dummy variables.

Usage

```
fact2dummy(data, all=TRUE, lab="x")
```

Arguments

<code>data</code>	A factor or a data frame that contains one or more factors (columns whose class is “factor” or “ordered”) that have to be substituted by dummy variables.
<code>all</code>	Logical. When <code>all=TRUE</code> (default) the output matrix will contain as many dummy variables as the number of the levels of the factor variable. On the contrary, when <code>all=FALSE</code> , the dummy variable related to the last level of the factor is dropped.
<code>lab</code>	A character string with the name of the variable to be pasted with its levels. This is used only when <code>data</code> is a factor. By default it is set to “x”.

Details

This function substitutes categorical variables in the input data frame (columns whose class is “factor” or “ordered”) with the corresponding dummy variables.

Value

A matrix with the dummy variables instead of initial factor variables.

Author(s)

Marcello D’Orazio (madorazi@istat.it)

See Also

[gower.dist](#)

Examples

```
x <- runif(5)
y <- factor(c(1,2,1,2,2))
z <- ordered(c(1,2,3,2,2))
xyz <- data.frame(x,y,z)
fact2dummy(xyz)

fact2dummy(xyz, all=FALSE)

#example with iris data frame
str(iris)
ir.mat <- fact2dummy(iris)
```

gower.dist

*Computes the Gower's Distance***Description**

This function computes the Gower's distance (dissimilarity) among units in a dataset or among observations in two distinct datasets.

Usage

```
gower.dist(data.x, data.y=data.x, rngs=NULL, KR.corr=TRUE)
```

Arguments

- `data.x` A matrix or a data frame containing variables that should be used in the computation of the distance.
- Columns of mode `numeric` will be considered as interval scaled variables; columns of mode `character` or class `factor` will be considered as categorical nominal variables; columns of class `ordered` will be considered as categorical ordinal variables and, columns of mode `logical` will be considered as binary asymmetric variables (see Details for further information).
- Missing values (NA) are allowed.
- If only `data.x` is supplied, the dissimilarities between rows of `data.x` will be computed.
- `data.y` A numeric matrix or data frame with the same variables, of the same type, as those in `data.x`. Dissimilarities between rows of `data.x` and rows of `data.y` will be computed. If not provided, by default it is assumed equal to `data.x` and only dissimilarities between rows of `data.x` will be computed.
- `rngs` A vector with the ranges to scale the variables. Its length must be equal to number of variables in `data.x`. In correspondence of nonnumeric variables, just put 1 or NA. When `rngs=NULL` (default) the range of a numeric variable is estimated by jointly considering the values for the variable in `data.x` and those in `data.y`. Therefore, assuming `rngs=NULL`, if a variable "X1" is considered:
- ```
rngs["X1"] <- max(data.x[, "X1"], data.y[, "X1"]) -
 min(data.x[, "X1"], data.y[, "X1"])
```
- .
- `KR.corr` When `TRUE` (default) the extension of the Gower's dissimilarity measure proposed by Kaufman and Rousseeuw (1990) is used. Otherwise, when `KR.corr=FALSE`, the original Gower's (1971) dissimilarity is considered.

## Details

This function computes distances among records when variables of different type (categorical and continuous) have been observed. In order to handle different types of variables, the Gower's dissimilarity coefficient (Gower, 1971) is used.

By default (`KR.corr=TRUE`) the Kaufman and Rousseeuw (1990) extension of the Gower's dissimilarity coefficient is used. The final dissimilarity between the  $i$ th and  $j$ th units is obtained as a weighted sum of dissimilarities for each variable:

$$d(i, j) = \frac{\sum_k \delta_{ijk} d_{ijk}}{\sum_k \delta_{ijk}}$$

In particular,  $d_{ijk}$  represents the distance between the  $i$ th and  $j$ th unit computed considering the  $k$ th variable. It depends on the nature of the variable:

- `logical` columns are considered as asymmetric binary variables, for such case  $d_{ijk} = 0$  if  $x_{ik} = x_{jk} = \text{TRUE}$ , 1 otherwise;
- `factor` or `character` columns are considered as categorical nominal variables and  $d_{ijk} = 0$  if  $x_{ik} = x_{jk}$ , 1 otherwise;
- `numeric` columns are considered as interval-scaled variables and

$$d_{ijk} = \frac{|x_{ik} - x_{jk}|}{R_k}$$

being  $R_k$  the range of the  $k$ th variable. The range is the one supplied with the argument `rngs` (`rngs[k]`) or the one computed on available data (when `rngs=NULL`);

- `ordered` columns are considered as categorical ordinal variables and the values are substituted with the corresponding position index,  $r_{ik}$  in the factor levels. These position indexes (that are different from the output of the R function `rank`) are transformed in the following manner

$$z_{ik} = \frac{(r_{ik} - 1)}{\max(r_{ik}) - 1}$$

These new values,  $z_{ik}$ , are treated as observations of an interval scaled variable.

As far as the weight  $\delta_{ijk}$  is concerned:

- $\delta_{ijk} = 0$  if  $x_{ik} = \text{NA}$  or  $x_{jk} = \text{NA}$ ;
- $\delta_{ijk} = 0$  if the variable is asymmetric binary and  $x_{ik} = x_{jk} = 0$  or  $x_{ik} = x_{jk} = \text{FALSE}$ ;
- $\delta_{ijk} = 1$  in all the other cases.

In practice, NAs and couple of cases with  $x_{ik} = x_{jk} = \text{FALSE}$  do not contribute to distance computation.

## Value

A `matrix` object with distances among rows of `data.x` and those of `data.y`.

## Author(s)

Marcello D'Orazio (`madorazi@istat.it`)

## References

Gower, J. C. (1971), "A general coefficient of similarity and some of its properties". *Biometrics*, **27**, 623–637.

Kaufman, L. and Rousseeuw, P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

## See Also

`daisy`, `dist`

## Examples

```
x1 <- as.logical(rbinom(10,1,0.5))
x2 <- sample(letters, 10, replace=TRUE)
x3 <- rnorm(10)
x4 <- ordered(cut(x3, -4:4, include.lowest=TRUE))
xx <- data.frame(x1, x2, x3, x4, stringsAsFactors = FALSE)

matrix of distances among observations in xx
gower.dist(xx)

matrix of distances among first obs. in xx
and the remaining ones
gower.dist(data.x=xx[1:3,], data.y=xx[4:10,])
```

---

mixed.mtc

*Statistical Matching via Mixed Methods*

---

## Description

This function implements some mixed methods to perform statistical matching between two data sources such that no units are in common and one or more continuous variables are shared.

## Usage

```
mixed.mtc(data.rec, data.don, match.vars, y.rec, z.don, method="ML",
 rho.yz=0, micro=FALSE, constr.alg="lpSolve")
```

## Arguments

`data.rec` A matrix or data frame that plays the role of *recipient* in the statistical matching application. This data set must contain all variables (columns) that should be used in statistical matching, i.e. the variables called by the arguments `match.vars` and `y.rec`. Note that, all variables must be continuous. Missing values (NA) are not allowed.

|                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|-------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>data.don</code>   | A matrix or data frame that plays the role of <i>donor</i> in the statistical matching application. This data set must contain all the numeric variables (columns) that should be used in statistical matching, i.e. the variables called by the arguments <code>match.vars</code> and <code>z.don</code> . All variables must be continuous. Missing values (NA) are not allowed.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <code>match.vars</code> | A character vector with the names of the common variables (the columns in both the data frames) to be used as matching variables ( <b>X</b> ).                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| <code>y.rec</code>      | A character vector with the name of the target variable Y that is observed only for units in <code>data.rec</code> . Only one continuous variable is allowed.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| <code>z.don</code>      | A character vector with the name of the target variable Z that is observed only for units in <code>data.don</code> . Only one continuous variable is allowed.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| <code>method</code>     | A character vector that identifies the method that should be used to estimate the parameters of the regression models: Y vs. <b>X</b> and Z vs. <b>X</b> . Maximum Likelihood method is used when <code>method="ML"</code> (default); on the contrary, when <code>method="MS"</code> the parameters are estimated according to Moriarity and Scheuren (2001 and 2003). See Details for further information.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| <code>rho.yz</code>     | <p>A numeric value representing the guess for the correlation among the Y (<code>y.rec</code>) and the Z variable (<code>z.don</code>) that are not jointly observed. Note that when <code>method="MS"</code>, <code>cor.yz</code> must specify the value of the correlation coefficient <math>\rho_{YZ}</math>; on the contrary, when <code>method="ML"</code>, it must specify the <i>partial correlation coefficient</i> among Y and Z given <b>X</b> (<math>\rho_{YZ \mathbf{X}}</math>).</p> <p>By default (<code>rho.yz=0</code>), in absence of auxiliary information concerning the correlation coefficient or the partial correlation coefficient, statistical matching is carried out under the assumption of independence among Y and Z given <b>X</b> (Conditional Independence Assumption, CIA), i.e. <math>\rho_{YZ \mathbf{X}} = 0</math>.</p>                                                       |
| <code>micro</code>      | Logical. When <code>micro=FALSE</code> (default) only the parameter estimates are returned. On the contrary, when <code>micro=TRUE</code> <code>data.rec</code> filled in with the values for the variable Z is returned too. The donors for filling in Z in <code>data.rec</code> are identified using a constrained distance hot deck method. In this case, the number of units (rows) in <code>data.don</code> must be greater or equal to the number of units (rows) in <code>data.rec</code> . See next argument and Details for further information.                                                                                                                                                                                                                                                                                                                                                          |
| <code>constr.alg</code> | <p>A string that has to be specified when <code>micro=TRUE</code>, in order to solve the transportation problem involved by the constrained distance hot deck method. Two choices are available: "lpSolve" and "relax". In the first case,</p> <p><code>constr.alg="lpSolve"</code>, the transportation problem is solved by means of the function <code>lp.transport</code> available in the package <b>lpSolve</b>. When</p> <p><code>constr.alg="relax"</code> the transportation problem is solved using RELAX-IV algorithm from Bertsekas and Tseng (1994), implemented in function <code>pairmatch</code> available in the package <b>optmatch</b>. Note that <code>constr.alg="relax"</code> is faster and requires less computational effort, but the usage of this algorithm is allowed only for research purposes (for details see function <code>relaxinfo()</code> in the package <b>optmatch</b>).</p> |

## Details

This function implements some mixed methods to perform statistical matching. A mixed method consists of two steps:

- (i) adoption of a parametric model for the joint distribution of  $(\mathbf{X}, Y, Z)$  and estimation of its parameters;
- (ii) derivation of a complete “synthetic” data set (recipient data set filled in with values for the  $Z$  variable) using a nonparametric approach.

In this case, as far as (i) is concerned, it is assumed that  $(\mathbf{X}, Y, Z)$  follows a multivariate normal distribution. In particular, dealing with continuous variables, a version of the imputation method known as *predictive mean matching* is used. This method consists of three steps:

step 1) – Regression step: The two linear regression models  $Y$  vs.  $\mathbf{X}$  and  $Z$  vs.  $\mathbf{X}$  are considered and their parameters are estimated.

step 2) – Computation of intermediate values. For the units in `data.rec` the following intermediate values are derived:

$$\tilde{z}_a = \hat{\alpha}_Z + \hat{\beta}_{Z\mathbf{X}}\mathbf{x}_a + e_a$$

for each  $a = 1, \dots, n_A$ , being  $n_A$  the number of units in `data.rec` (rows of `data.rec`). Note that,  $e_a$  is a random draw from the multivariate normal distribution with zero mean and estimated residual variance  $\hat{\sigma}_{Z|\mathbf{X}}$ .

Similarly, for the units in `data.don` the following intermediate values are derived:

$$\tilde{y}_b = \hat{\alpha}_Y + \hat{\beta}_{Y\mathbf{X}}\mathbf{x}_b + e_b$$

for each  $b = 1, \dots, n_B$ , being  $n_B$  the number of units in `data.don` (rows of `data.don`).  $e_b$  is a random draw from the multivariate normal distribution with zero mean and estimated residual variance  $\hat{\sigma}_{Y|\mathbf{X}}$ .

step 3) – Matching step. For each observation (row) in `data.rec` a donor is chosen in `data.don` through a nearest neighbor constrained distance hot deck procedure. The distances are computed between  $(y_a, \tilde{z}_a)$  and  $(\tilde{y}_b, z_b)$  using Mahalanobis distance.

For further details see Sections 2.5.1 and 3.6.1 in D’Orazio *et al.* (2006).

Note that in step 1) the parameters of the regression model can be estimated by means of the Maximum Likelihood method (`method="ML"`) (see D’Orazio *et al.*, 2006, pp. 19–23, 73–75) or, using the Moriarity and Scheuren (2001 and 2003) approach (`method="MS"`) (see also D’Orazio *et al.*, 2006, pp. 75–76). The two estimation methods are compared in D’Orazio *et al.* (2005).

When `method="MS"`, if the value specified for the argument `rho.yz` is not compatible with the other correlation coefficients estimated from the data, then it is substituted with the closest value compatible with the other estimated coefficients.

When `micro=FALSE` only the estimation of the parameters is performed (step 1). Otherwise, (`micro=TRUE`) the whole procedure is carried out.

**Value**

A list with a varying number of components depending on the values of the arguments `method` and `rho.yz`.

|                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|----------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>mu</code>            | The estimated mean vector.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| <code>vc</code>            | The estimated variance–covariance matrix.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| <code>cor</code>           | The estimated correlation matrix.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| <code>res.var</code>       | A vector with estimates of the residual variances $\sigma_{Y Z\mathbf{X}}$ and $\sigma_{Z Y\mathbf{X}}$ .                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| <code>start.prho.yz</code> | It is the initial guess for the partial correlation coefficient $\rho_{YZ \mathbf{X}}$ passed in input via the <code>rho.yz</code> argument when <code>method="ML"</code> .                                                                                                                                                                                                                                                                                                                                                                                               |
| <code>rho.yz</code>        | Returned in output only when <code>method="MS"</code> . It is a vector with four values: the initial guess for $\rho_{YZ}$ ; the lower and upper bounds for $\hat{\rho}_{YZ}$ in the statistical matching framework given the correlation coefficients among Y and Xs and the correlation coefficients among Z and Xs estimated from the available data; and, finally, the closest admissible value used in computations instead of the initial <code>rho.yz</code> that resulted not coherent with the other correlation coefficients estimated from the available data. |
| <code>phi</code>           | When <code>method="MS"</code> . Estimates of the $\phi$ terms introduced by Moriarity and Scheuren (2001 and 2003).                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| <code>filled.rec</code>    | The <code>data.rec</code> filled in with the values of Z. It is returned only when <code>micro=TRUE</code> .                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| <code>mtc.ids</code>       | when <code>micro=TRUE</code> . This is a matrix with the same number of rows of <code>data.rec</code> and two columns. The first column contains the row names of the <code>data.rec</code> and the second column contains the row names of the corresponding donors selected from the <code>data.don</code> . When the input matrices do not contain row names, a numeric matrix with the indexes of the rows is provided.                                                                                                                                               |
| <code>dist.rd</code>       | A vector with the distances among each recipient unit and the corresponding donor, returned only in case <code>micro=TRUE</code> .                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| <code>call</code>          | How the function has been called.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |

**Author(s)**

Marcello D’Orazio (madorazi@istat.it)

**References**

- Bertsekas, D.P. and Tseng, P. (1994). “RELAX–IV: A Faster Version of the RELAX Code for Solving Minimum Cost Flow Problems”. *Technical Report*, LIDS-P-2276, Massachusetts Institute of Technology, Cambridge. [http://web.mit.edu/dimitrib/www/RELAX4\\_doc.pdf](http://web.mit.edu/dimitrib/www/RELAX4_doc.pdf)
- D’Orazio, M., Di Zio, M. and Scanu, M. (2005). “A comparison among different estimators of regression parameters on statistically matched files through an extensive simulation study”, *Contributi*, **2005/10**, Istituto Nazionale di Statistica, Rome. [http://www.istat.it/dati/pubbsci/contributi/Contributi/contr\\_2005/2005\\_10.pdf](http://www.istat.it/dati/pubbsci/contributi/Contributi/contr_2005/2005_10.pdf)
- D’Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley, Chichester.

Moriarity, C., and Scheuren, F. (2001). “Statistical matching: a paradigm for assessing the uncertainty in the procedure”. *Journal of Official Statistics*, **17**, 407–422. <http://www.jos.nu/Articles/abstract.asp?article=173407>

Moriarity, C., and Scheuren, F. (2003). “A note on Rubin’s statistical matching using file concatenation with adjusted weights and multiple imputation”, *Journal of Business and Economic Statistics*, **21**, 65–73.

## See Also

[NND.hotdeck](#)

## Examples

```
Example with fictitious data
Set the correlation matrix
mat.cor <- matrix(0, 4, 4)
mat.cor[lower.tri(mat.cor)] <- c(0.3, 0.5, 0.7, 0.8, 0.4, 0.8)
mat.cor <- mat.cor+t(mat.cor)
diag(mat.cor) <- 1
dimnames(mat.cor) <- list(c("x1", "x2", "y", "z"), c("x1", "x2", "y", "z"))

generate data from multivariate normal distribution
library(mvtnorm)
data.all <- rmvnorm(n=100, mean=rep(0,4), sigma=mat.cor)
dimnames(data.all) <- list(1:100, c("x1", "x2", "y", "z"))

reproduce statistical matching framework
data.A <- data.all[1:50, 1:3] #z deleted
data.B <- data.all[51:100, c(1:2,4)] #y deleted

ML estimation method under CIA ((rho_YZ|X=0));
only parameter estimates (micro=FALSE)
mtc.1 <- mixed.mtc(data.rec=data.A, data.don=data.B,
 match.vars=c("x1", "x2"), y.rec="y", z.don="z")

estimated vs. true correlation matrix
mtc.1$cor - mat.cor

ML estimation method with partial correlation coefficient
set equal to 0.5 (rho_YZ|X=0.5)
only parameter estimates (micro=FALSE)

mtc.2 <- mixed.mtc(data.rec=data.A, data.don=data.B,
 match.vars=c("x1", "x2"), y.rec="y", z.don="z", rho.yz=0.5)

estimated vs. true correlation matrix
mtc.2$cor - mat.cor

ML estimation method with partial correlation coefficient
set equal to 0.5 (rho_YZ|X=0.5)
with imputation step (micro=TRUE)
```

```

mtc.3 <- mixed.mtc(data.rec=data.A, data.don=data.B,
 match.vars=c("x1","x2"), y.rec="y", z.don="z", rho.yz=0.5,
 micro=TRUE, constr.alg="lpSolve")

estimated vs. true correlation matrix
mtc.3$cor - mat.cor

first rows of data.rec filled in with z
head(mtc.3$filled.rec)

Moriarity and Scheuren estimation method under CIA;
only with parameter estimates (micro=FALSE)
mtc.4 <- mixed.mtc(data.rec=data.A, data.don=data.B,
 match.vars=c("x1","x2"), y.rec="y", z.don="z", method="MS")

estimated vs. true correlation matrix
mtc.4$cor - mat.cor

Moriarity and Scheuren estimation method
with correlation coefficient set equal to 0.2 (rho_YZ=0.2)
only parameter estimates (micro=FALSE)

mtc.5 <- mixed.mtc(data.rec=data.A, data.don=data.B,
 match.vars=c("x1","x2"), y.rec="y", z.don="z",
 method="MS", rho.yz=0.2)

the starting value of rho.yz and the value used
in computations
mtc.5$rho.yz

estimated vs. true correlation matrix
mtc.5$cor - mat.cor

Moriarity and Scheuren estimation method
with correlation coefficient set equal to 0.6 (rho_YZ=0.6)
with imputation step (micro=TRUE)

mtc.6 <- mixed.mtc(data.rec=data.A, data.don=data.B,
 match.vars=c("x1","x2"), y.rec="y", z.don="z", rho.yz=0.6,
 method="MS", micro=TRUE, constr.alg="lpSolve")

estimated vs. true correlation matrix
mtc.6$cor - mat.cor

first rows of data.rec filled in with z imputed values
head(mtc.6$filled.rec)

```

## Description

This function implements the distance hot deck method to match the records of two data sources such that no units are in common and one or more variables are shared.

## Usage

```
NND.hotdeck(data.rec, data.don, match.vars, don.class=NULL,
 dist.fun="Euclidean", constrained=FALSE, constr.alg=NULL)
```

## Arguments

- |                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|-------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>data.rec</code>   | A matrix or data frame that plays the role of <i>recipient</i> in the statistical matching application. This data frame must contain the variables (columns) that should be used, directly or indirectly, in the computation of distances between its observations (rows) and those of <code>data.y</code> .<br>Missing values (NA) are allowed.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| <code>data.don</code>   | A matrix or data frame that plays the role of <i>donor</i> in the statistical matching application. The variables (columns) involved, directly or indirectly, in the computation of distance must be the same and of the same type as those in <code>data.rec</code> .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| <code>match.vars</code> | A character vector with the names of the variables (the columns in both the data frames) that have to be used to compute distances among records (rows) in <code>data.rec</code> and those in <code>data.don</code> .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <code>don.class</code>  | A character vector with the names of the variables (columns in both the data frames) that have to be used to identify donation classes. In this case the computation of distances is limited to those units of <code>data.rec</code> and <code>data.don</code> that belong to the same donation classes. The case of empty donation classes should be avoided. To avoid confusion, it would be preferable that variables used to form donation classes are defined as <code>factor</code> .<br>When not specified (default) no donation classes are used. This may result in a heavy computational effort.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <code>dist.fun</code>   | A string with the name of the distance function that has to be used. The following distances are allowed: “Euclidean” (default), “Manhattan” (aka “City block”), “exact” or “exact matching”, “Gower” or one of the distance functions available in the package <b>proxy</b> . Note that the distance is computed using the function <code>dist</code> of the package <b>proxy</b> with the exception of the “Gower” case.<br>When <code>dist.fun= "Euclidean"</code> or “Manhattan” all the non numeric variables in <code>data.rec</code> and <code>data.don</code> will be converted to numeric. On the contrary, when <code>dist.fun="exact"</code> or “exact matching”, all the variables in <code>data.rec</code> and <code>data.don</code> will be converted to character and, as far as the distance computation is concerned, they will be treated as categorical nominal variables, i.e. distance is 0 if a couple of units shows the same response category and 1 otherwise.<br>When <code>dist.fun="Gower"</code> the Gower dissimilarity is considered. See function <code>gower.dist</code> for details. |

|                          |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|--------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>constrained</code> | Logical. When <code>constrained=FALSE</code> (default) each record in <code>data.don</code> can be used as a donor more than once. On the contrary, when <code>constrained=TRUE</code> each record in <code>data.don</code> can be used as a donor only once. In this case, the set of donors is selected by solving a transportation problem, so to minimize the overall matching distance. See description of the argument <code>constr.alg</code> for details.                                                                                                                                                                                                                                                                                                                                      |
| <code>constr.alg</code>  | A string that has to be specified when <code>constrained=TRUE</code> . Two choices are available: “lpSolve” and “relax”. In the first case, <code>constr.alg="lpSolve"</code> , the transportation problem is solved by means of the function <code>lp.transport</code> available in the package <b>lpSolve</b> . When <code>constr.alg="relax"</code> the transportation problem is solved using RELAX-IV algorithm from Bertsekas and Tseng (1994), implemented in function <code>pairmatch</code> available in the package <b>optmatch</b> . Note that <code>constr.alg="relax"</code> is faster and requires less computational effort, but the usage of this algorithm is allowed only for research purposes (for details see function <code>relaxinfo()</code> in the package <b>optmatch</b> ). |

### Details

This function finds a donor record in `data.don` for each record in the recipient data frame `data.rec`. In the unconstrained case, the closest donor record is found according to the chosen distance function in correspondence of each record in the recipient data set. When more donor records are at the minimum distance from the given recipient record, one of them is picked at random.

In the constrained case the set of donors is chosen in order to minimize the overall matching distance. In this case the number of units (rows) in the donor data set has to be larger or equal to the number of units of the recipient data set. In case donation classes are used, this condition must be satisfied in each donation classes. For further details on nearest neighbor distance hot deck refer to Chapter 2 in D’Orazio *et al.* (2006).

Note that this function can also be used to impute missing values in a data set using the nearest neighbor distance hot deck. In this case `data.rec` is the part of the initial data set that contains missing values; on the contrary, `data.don` is the part of the data set without missing values. See R code in the Examples for details.

### Value

A R list with the following components:

|                      |                                                                                                                                                                                                                                                                                                                                                                                      |
|----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>mtc.ids</code> | A matrix with the same number of rows of <code>data.rec</code> and two columns. The first column contains the row names of the <code>data.rec</code> and the second column contains the row names of the corresponding donors selected from the <code>data.don</code> . When the input matrices do not contain row names, a numeric matrix with the indexes of the rows is provided. |
| <code>dist.rd</code> | A vector with the distances among each recipient unit and the corresponding donor.                                                                                                                                                                                                                                                                                                   |
| <code>noad</code>    | When <code>constrained=FALSE</code> , it reports the number of available donors at the minimum distance for each recipient unit.                                                                                                                                                                                                                                                     |
| <code>call</code>    | How the function has been called.                                                                                                                                                                                                                                                                                                                                                    |

**Author(s)**

Marcello D’Orazio (madorazi@istat.it)

**References**

Bertsekas, D.P. and Tseng, P. (1994). “RELAX–IV: A Faster Version of the RELAX Code for Solving Minimum Cost Flow Problems”. *Technical Report*, LIDS-P-2276, Massachusetts Institute of Technology, Cambridge. [http://web.mit.edu/dimitrib/www/RELAX4\\_doc.pdf](http://web.mit.edu/dimitrib/www/RELAX4_doc.pdf)

D’Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley, Chichester.

Rodgers, W.L. (1984). “An evaluation of statistical matching”. *Journal of Business and Economic Statistics*, **2**, 91–102.

Singh, A.C., Mantel, H., Kinack, M. and Rowe, G. (1993). “Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption”. *Survey Methodology*, **19**, 59–79.

**See Also**

[RANDwNND.hotdeck](#)

**Examples**

```
reproduce the classical matching framework
lab <- c(1:15, 51:65, 101:115)
iris.rec <- iris[lab, c(1:3,5)] # recipient data.frame
iris.don <- iris[-lab, c(1:2,4:5)] #donor data.frame

Now iris.rec and iris.don have the variables
"Sepal.Length", "Sepal.Width" and "Species"
in common.
"Petal.Length" is available only in iris.rec
"Petal.Width" is available only in iris.don

Find the closest donors computing distance
on "Sepal.Length" and "Sepal.Width"
unconstrained case, Euclidean distance

out.NND.1 <- NND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 match.vars=c("Sepal.Length", "Sepal.Width"))

create the synthetic data.set:
fill in "Petal.Width" in iris.rec

fused.1 <- create.fused(data.rec=iris.rec, data.don=iris.don,
 mtc.ids=out.NND.1$mtc.ids, z.vars="Petal.Width")

Find the closest donors computing distance
on "Sepal.Length", "Sepal.Width" and Species;
unconstrained case, Gower's distance
```

```

out.NND.2 <- NND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 match.vars=c("Sepal.Length", "Sepal.Width", "Species"),
 dist.fun="Gower")

find the closest donors using "Species" to form donation classes
and "Sepal.Length" and "Sepal.Width" to compute distance;
unconstrained case.

out.NND.3 <- NND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 match.vars=c("Sepal.Length", "Sepal.Width"),
 don.class="Species")

find the donors using "Species" to form donation classes
and "Sepal.Length" and "Sepal.Width" to compute distance;
constrained case, "RELAX" algorithm

library(optmatch)
out.NND.4 <- NND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 match.vars=c("Sepal.Length", "Sepal.Width"),
 don.class="Species", constr=TRUE, constr.alg="relax")

find the donors using "Species" to form donation classes
and "Sepal.Length" and "Sepal.Width" to compute distance;
constrained case, transportation problem solved by functions
in package "lpSolve"

library(lpSolve)
out.NND.5 <- NND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 match.vars=c("Sepal.Length", "Sepal.Width"),
 don.class="Species", constr=TRUE, constr.alg="lpSolve")

Example of Imputation of missing values.
Introducing missing values in iris
ir.mat <- iris
miss <- rbinom(nrow(iris), 1, 0.3)
ir.mat[miss==1,"Sepal.Length"] <- NA
iris.rec <- ir.mat[miss==1,-1]
iris.don <- ir.mat[miss==0,]

#search for NND donors
imp.NND <- NND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 match.vars=c("Sepal.Width", "Petal.Length", "Petal.Width"),
 don.class="Species")

imputing missing values
iris.rec.imp <- create.fused(data.rec=iris.rec, data.don=iris.don,
 mtc.ids=imp.NND$mtc.ids, z.vars="Sepal.Length")

rebuild the imputed data.frame
final <- rbind(iris.rec.imp, iris.don)

```

---

RANDwNND.hotdeck *Random Distance hot deck method for Statistical Matching.*

---

## Description

This function implements a variant of the distance hot deck method. For each recipient record a subset of units consisting of the closest donors is retained and then a donor is selected at random

## Usage

```
RANDwNND.hotdeck(data.rec, data.don, match.vars, don.class=NULL,
 dist.fun="Euclidean", cut.don="rot", k=NULL)
```

## Arguments

|                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>data.rec</code>   | A numeric matrix or data frame that plays the role of <i>recipient</i> in the statistical matching application. This data frame must contain the variables (columns) that should be used, directly or indirectly, in the computation of distances between its observations (rows) and those in <code>data.y</code> .<br>Missing values (NA) are allowed.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| <code>data.don</code>   | A matrix or data frame that plays the role of <i>donor</i> in the statistical matching application. The variables (columns) involved, directly or indirectly, in the computation of distances be the same and of the same type as those in <code>data.rec</code> .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <code>match.vars</code> | A character vector with the names of the variables (the columns in both the data frames) that have to be used to compute distances among records (rows) in <code>data.rec</code> and those in <code>data.don</code> .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <code>don.class</code>  | A character vector with the names of the variables (columns in both the data frames) that have to be used to identify donation classes. In this case the computation of distances is limited to those units in <code>data.rec</code> and <code>data.doc</code> that belong to the same donation classes. The case of empty donation classes should be avoided. To avoid confusion, it is preferable the variables used to form donation classes are defined as <code>factor</code> .<br>When not specified (default), no donation classes are used. This may result in a heavy computational effort.                                                                                                                                                                                                                                                                                                                                                                           |
| <code>dist.fun</code>   | A string with the name of the distance function that has to be used. The following distances are allowed: “Euclidean” (default), “Manhattan” (aka “City block”), “exact” or “exact matching”, “Gower” or one of the distance functions available in the package <b>proxy</b> . Note that the distance is computed using the function <code>dist</code> of the package <b>proxy</b> with the exception of the “Gower” case.<br>When <code>dist.fun= "Euclidean"</code> or “Manhattan” all the non numeric variables in <code>data.rec</code> and <code>data.don</code> will be converted to numeric. On the contrary, when <code>dist.fun="exact"</code> or “exact matching”, all the variables in <code>data.rec</code> and <code>data.don</code> will be converted to character and, as far as distance computation is concerned, they will be treated as categorical nominal variables, i.e. the distance is 0 if two units show the same response category and 1 otherwise. |

|                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                      | When <code>dist.fun="Gower"</code> the Gower dissimilarity is considered. See function <code>gower.dist</code> for details.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| <code>cut.don</code> | A character string that, jointly with the argument <code>k</code> , identifies the rule to be used to form the subset of the closest donor records. <ul style="list-style-type: none"> <li>• <code>cut.don="rot"</code>: (default) then the number of the closest donors to retain is given by <math>\lceil \sqrt{n_D} \rceil + 1</math>; being <math>n_D</math> the total number of available donors. In this case <code>k</code> must not to be specified.</li> <li>• <code>cut.don="span"</code>: the number of closest donors is determined as the proportion <code>k</code> of all the available donors, i.e. <math>\lceil n_D \times k \rceil</math>. Note that, in this case, <math>0 &lt; k \leq 1</math>.</li> <li>• <code>cut.don="exact"</code>: the <code>k</code>th closest donors out of the <math>n_D</math> are retained. In this case, <math>0 &lt; k \leq n_D</math>.</li> <li>• <code>cut.don="min"</code>: the donors at the minimum distance from the recipient are retained.</li> <li>• <code>cut.don="k.dist"</code>: only the donors whose distance from the recipient is less or equal to the value specified with the argument <code>k</code>.</li> </ul> |
| <code>k</code>       | Depends on the <code>cut.don</code> argument.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |

### Details

This function finds a donor record for each record in the recipient data set. This donor is chosen at random in the subset of available donors. The number of closest donors retained to form the subset is determined according to criterion specified with the argument `cut.don`.

Note that the same donor can be used more than once.

Note that this function can also be used to impute missing values in a data set. In this case `data.rec` is the part of the initial data set that contains missing values; on the contrary, `data.don` is the part of the data set without missing values. See R code in the Examples for details.

### Value

A R list with the following components:

|                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|-----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>mtc.ids</code>  | A matrix with the same number of rows of <code>data.rec</code> and two columns. The first column contains the row names of the <code>data.rec</code> and the second column contains the row names of the corresponding donors selected from the <code>data.don</code> . When the input matrices do not contain row names, then a numeric matrix with the indexes of the rows is provided.                                                                                                                                                                                                |
| <code>sum.dist</code> | A matrix with summary statistics concerning the subset of the closest donors from which a donor for a given recipient is chosen. The first three columns report the minimum, the maximum and the standard deviation of the distances among the recipient record and the donors in the subset of the closest donors, respectively. The 4th column reports the cutting distance, i.e. the value of the distance such that donors at a higher distance are discarded. The 5th column reports the distance between the recipient and the donor chosen at random in the subset of the donors. |
| <code>noad</code>     | For each recipient unit, reports the number of donor records in the subset of closest donors.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| <code>call</code>     | How the function has been called.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |

**Author(s)**

Marcello D’Orazio (madorazi@istat.it)

**References**

- D’Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley, Chichester.
- Rodgers, W.L. (1984). “An evaluation of statistical matching”. *Journal of Business and Economic Statistics*, **2**, 91–102.
- Singh, A.C., Mantel, H., Kinack, M. and Rowe, G. (1993). “Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption”. *Survey Methodology*, **19**, 59–79.

**See Also**

[NND.hotdeck](#)

**Examples**

```
reproduce the classical matching framework
lab <- c(1:10, 51:60, 101:110)
iris.rec <- iris[lab, c(1:3,5)] # recipient data.frame
iris.don <- iris[-lab, c(1:2,4:5)] # recipient data.frame

Now iris.rec and iris.don have the variables
"Sepal.Length", "Sepal.Width" and "Species"
in common.
"Petal.Length" is available only in iris.rec
"Petal.Width" is available only in iris.don

find a donor in the subset of closest donors using cut.don="rot";
the distance is computed using "Sepal.Length" and "Sepal.Width"

out.NND.1 <- RANDwNND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 match.vars=c("Sepal.Length", "Sepal.Width"))

create the synthetic (or fused) data.frame:
fill in "Petal.Width" in iris.rec
fused.1 <- create.fused(data.rec=iris.rec, data.don=iris.don,
 mtc.ids=out.NND.1$mtc.ids, z.vars="Petal.Width")

find a donor in the subset of closest donors using cut.don="rot";
the distance is computed using "Sepal.Length" and "Sepal.Width"
"Species" is used to form donation classes

out.NND.2 <- RANDwNND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 match.vars=c("Sepal.Length", "Sepal.Width") , don.class="Species")

as before, but with a different criteria to reduce the no. of donors:
the first half (k=0.5) of the closest available donors is retained,
```

```
then a donor is chosen at random among them

out.NND.3 <- RANDwNND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 don.class="Species", match.vars=c("Sepal.Length", "Sepal.Width"),
 cut.don="span", k=0.5)

as before, but the subset of closest donors is formed by considering
only the first k=5 closest donors

out.NND.4 <- RANDwNND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 don.class="Species", match.vars=c("Sepal.Length", "Sepal.Width"),
 cut.don="exact", k=5)

as before, but the subset of closest donors is formed by considering
the donors whose distance (Gower) is less or equal to k=0.33

out.NND.5 <- RANDwNND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 don.class="Species", match.vars=c("Sepal.Length", "Sepal.Width"),
 dist.fun="Gower", cut.don="k.dist", k=0.33)

Example of Imputation of missing values
introducing missing vales in iris
ir.mat <- iris
miss <- rbinom(nrow(iris), 1, 0.3)
ir.mat[miss==1,"Sepal.Length"] <- NA
iris.rec <- ir.mat[miss==1,-1]
iris.don <- ir.mat[miss==0,]

#search for NND donors
imp.NND <- RANDwNND.hotdeck(data.rec=iris.rec, data.don=iris.don,
 match.vars=c("Sepal.Width", "Petal.Length", "Petal.Width"),
 don.class="Species")

imputing missing values
iris.rec.imp <- create.fused(data.rec=iris.rec, data.don=iris.don,
 mtc.ids=imp.NND$mtc.ids, z.vars="Sepal.Length")

rebuild the imputed data.frame
final <- rbind(iris.rec.imp, iris.don)
```

# Index

\*Topic **cluster**

fact2dummy, 3

gower.dist, 5

\*Topic **design**

fact2dummy, 3

\*Topic **multivariate**

gower.dist, 5

\*Topic **nonparametric**

create.fused, 2

mixed.mtc, 7

NND.hotdeck, 13

RANDwNND.hotdeck, 17

\*Topic **regression**

mixed.mtc, 7

create.fused, 2

daisy, 7

dist, 7, 13, 17

fact2dummy, 3

gower.dist, 4, 5, 13, 18

lp.transport, 8, 14

mixed.mtc, 7

NND.hotdeck, 2, 3, 11, 12, 19

pairmatch, 8, 14

RANDwNND.hotdeck, 2, 3, 15, 17

rank, 6