
7th Conference on the Use of R in Official Statistics (uRos2019)
Bucharest, 20 May 2019

Integration of Data Sources in R through Statistical Matching

Marcello D'Orazio*
marcello.dorazio@fao.org
(marcello.dorazio@istat.it)

**Senior Researcher in Statistical Methodologies*

Office of Chief Statistician, Food and Agriculture Organization (FAO) of the United Nations (UN), Rome Italy

(Italian National Institute of Statistics – Istat, Rome, Italy)

Techniques for integrating data sources:

- 1) ***Record linkage***
- 2) ***Statistical matching***

Record linkage (RL)

Find couples of records in different data sources referred to the same entity (e.g. person, household, farm, business ...)

The method relies on comparisons of set of variables available in both the data sources.

- **Exact record linkage**: couple of record sharing the same values of the (error-free) identifying variables (Personal Id. Code, VAT number, etc.).
- **Probabilistic record linkage**: identifiers are NOT error-free or there is a set of variables that can be used for identification purposes (name, surname, gender, birthday ...).
estimate the probability that a couple of records refers to the same entity.

Main uses of RL in Official Statistics:

- Integrate registers/archives to create a sampling frame
- Enrich survey data with data from admin register
- Estimate coverage of censuses (capture-recapture)
- Integrate registers/archives to create a Statistical Register (-> register based statistics)
- ...

Statistical Matching (aka *data fusion* or *synthetic matching*)

A wide set of methods to integrate two data sources, typically:

- 1) From two sample surveys (microdata or aggregated data) referred to the same target population

GOAL: investigate **relationship between variables never jointly observed in a single data source**

- 2) A sample survey and data from a register (also with data collected on different units, e.g. household vs. person-level data)

GOAL: make **inference on parameters (mean, total, ratio ...)** referred to variables available only in the register

SM 'basic' case:

source *A*

<i>Y</i>	X
----------	----------

source *B*

X	<i>Z</i>
----------	----------

1. **X** are shared by *A* and *B*
2. *Y* e *Z* NEVER jointly available
3. The probability of finding the same unit in both the sources is 0

Goal of SM consists in:

case 1) explore relationship between *Y* and *Z* or among **X**, *Y* and *Z*

case 2) estimate parameters related to *Z* (Rivers, 2007; Lavallée, 2007)

Goals of Statistical Matching - case (1)

Case 1) investigate relationship between Y and Z or among X , Y and Z

✓ **micro**: a “synthetic” data-set including X , Y and Z is created

Option i) fill-in A with values of Z (the missing variable):

A is the *recipient*
 B is the *donor*

Y	X	Z

Option ii) A and B are concatenated ($A \cup B$) then missing parts are imputed (*file concatenation*; Rubin, 1986):

	Y	X	Z
A			
B			

“synthetic”: imputed values for missing variables are NOT the values actually observed through data collection

- ✓ **macro**: estimation of parameters concerning relationship between variables never jointly observed:
 - correlation coefficient ρ_{YZ}
 - regression coefficient β_{YZ}
 - two-way contingency table $Y \times Z$
 - ...

Macro estimation does NOT necessarily require:

- integration of sources at micro-data level, and /or
- Availability of micro-data sources

Goals of Statistical Matching - case (1)

E.g. estimation of contingency table $Y \times Z$ can be achieved starting from:

- table $X \times Y$ estimated from survey A
- table $X \times Z$ estimated from survey B

Table 3. *Distribution of Professional Status vs Age in file A*

Age	Professional Status			Total
	M	E	W	
1	–	–	9	9
2	–	5	17	22
3	179	443	486	1108
4	6	1	2	9
Tot.	185	449	514	1148

Table 4. *Distribution of Educational Level vs Age in file B*

Age	Educational Level				Total
	C	V	S	D	
1	6	0	–	–	6
2	14	6	13	–	33
3	387	102	464	158	1111
4	10	0	3	2	15
Tot.	417	108	480	160	1165

Goal	Approach		
	Parametric	Nonparametric	Mixed
mAcro	Yes	Yes	No
mIcro	Yes	Yes	Yes

Example **parametric mAcro**

Use estimation methods designed to deal with missing values to estimate ρ_{YZ} or the contingency table $Y \times Z$

Example **parametric micro**: linear regression

- 1) Estimate parameters of $z_k = \beta_0 + \beta_1 x_{Bk}$ on survey B
- 2) Impute predicted values of Z in A by $\hat{z}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{Ak}$

Example of **nonparametric micro**: nearest-neighbor donor

- 1) For each record in A search the closest unit in B according to distance on X
- 2) Impute in A the Z value observed on its closest donor in B

mixed approach consists in 2 steps

- 1) Parametric: a model is fitted (e.g. linear regression). This model is used to draw values for the missing variables (**intermediate** values)
- 2) Nonparametric: intermediate values become the input for a nonparametric method (e.g. donor based) that determines the **final** imputed values

Statistical Methods used for statistical matching purposes:

- estimation of parameters in presence of missing values
(Little & Rubin, 2002; Rassler, 2002; D'Orazio et al 2005)
- model based imputation (regression, ...)
(Moriarity & Scheuren, 2001, 2003; Rassler, 2002; D'Orazio et al 2005)
- multiple imputation *(Rubin, 1986; Rassler, 2002)*
- nonparametric imputation (donor based methods)
(Singh, 1993; D'Orazio et al, 2006b; D'Orazio, 2015)
- calibration of survey weights *(Renssen, 1998)*
- estimation under partial identification (uncertainty investigation)
(Moriarity & Scheuren, 2001, 2003; D'Orazio et al, 2006a, 2016, 2017; Conti & Marella, 2012, 2013)
- machine learning *(D'Orazio, 2019a)*
- ...

SM underlying assumptions

- i) A e B are representative samples of the same target population
- ii) The X variables, shared by both the data sources, follow the same definitions and have the same distributions in both A and B .
- iii) In the “basic” SM setting ($A=\{X,Y\}$ e $B=\{X,Z\}$) when matching is uniquely based on a subset of common variables \mathbf{X}_M ($\mathbf{X}_M \subseteq \mathbf{X}$) (**matching variables**), it is implicitly assumed the independence between Y and Z once conditioning on \mathbf{X}_M

$$f(x_M, y, z) = f(y|x_M) f(z|x_M) f(x_M)$$

This assumption is NOT holding in most of real cases.

For instance:

X : household typology ($i = 1, \dots, I$)

Y : classes of household income ($j = 1, \dots, J$)

Z : classes of total household expenditures ($k = 1, \dots, K$)

Conditional Independence Assumption implies:

$$\begin{aligned} P(X = \text{single male age} > 24, Y = 1, Z = 1) = \\ P(Y = 1 | X = \text{single male age} > 24) \times \\ P(Z = 1 | X = \text{single male age} > 24) \times \\ P(X = \text{single male age} > 24) \end{aligned}$$

Estimation is straightforward:

$$\begin{aligned} \hat{P}(Y = 1, Z = 1) \\ = \sum_{i=1}^I [\hat{P}^{(A)}(Y = 1 | X = i) \times \hat{P}^{(B)}(Z = 1 | X = i) \times \hat{P}^{(A \cup B)}(X = i)] \end{aligned}$$

Underlying Assumptions in Statistical Matching: Conditional Independence

Implications of CIA: X =gender; Y =having a cat; Z =purchase cat foot

$X \times Y$ estimated from A

Gender	Cat	No cat	Tot.
M	10	38	48
F	32	20	52
Tot.	42	58	100

$X \times Z$ estimated from B

Gender	Buy	Not buy	Tot
M	4	44	48
F	16	36	52
Tot.	20	80	100

Under Conditional independence:

$$\begin{aligned}\Pr(Y = 'no cat', Z = 'buy') &= \Pr(Y = 'no cat' | X = 'M') \times \Pr(Z = 'buy' | X = 'M') \times \Pr(X = 'M') + \\ &\quad + \Pr(Y = 'no cat' | X = 'F') \times \Pr(Z = 'buy' | X = 'F') \times \Pr(X = 'F') \\ &= 38/48 \times 4/48 \times 48/100 + 20/52 \times 16/52 \times 52/100 \\ &= 0,0317 + 0,0615 = 0,0932\end{aligned}$$

is NOT 0 as one would expect!!!

In case of continuous variables following the multivariate Gaussian distribution, CIA implies:

$$\rho_{yz} = \rho_{xy} \times \rho_{xz}$$

i.e. a relationship completely explained by X

Most of SM methods proposed in literature rely on Conditional Independence (CI) Assumption

This is a strong assumption; it **rarely holds true** in real world applications.

When CI assumption is NOT valid, then results of SM based on it will NOT be reliable.

NB: unfortunately in the basic SM setting ($A=\{X,Y\}$ and $B=\{X,Z\}$) it is NOT possible to test whether CI holds or not

Testing CI requires the availability of a data sources containing all the interest variables X , Y and Z

(Rough assessment possible through investigation of uncertainty)

What if...

- (i) If CI is valid => apply SM methods based on CI taking into account the final goal (macro or micro)

CI holds true when one of the X variables is **strongly correlated/associated** with one of the target variables (X is said **proxy**)

What does it mean 'proxy'?

Example: X, Y e Z are continuous and follow the Multiv. Gaussian

ρ_{xy} can be estimated on A

ρ_{xz} can be estimated on B

there are NO data to estimate ρ_{yz}

By considering the properties of the correlation matrix (should be positive semi-definite) it is possible to show that:

$$\rho_{xy}\rho_{xz} - \sqrt{(1 - \rho_{xy}^2)(1 - \rho_{xz}^2)} \leq \rho_{yz} \leq \rho_{xy}\rho_{xz} + \sqrt{(1 - \rho_{xy}^2)(1 - \rho_{xz}^2)}$$

If ρ_{xy} is close to 1 then $\rho_{yz} \cong \rho_{xy} \times \rho_{xz}$ (CIA holds)

in such a case, X is said 'proxy' of Y

Example: matching of SILC and HBS Istat surveys

Goal: explore relationship between:

Y = HH income (observed in IT-SILC)

Z = HH overall consumption (observed in HBS)

results were acceptable if one of the X s was the income in classes (Y^*)

- in IT-SILC derived by categorizing Y
- roughly observed in HBS

(Cf. Donatiello et al., 2016a, 2016b)

(ii) If CI between Y and Z given X is **NOT holding** then:

ii.2) search for [auxiliary information](#):

- alternative data sources with **all** variables observed;
- estimates of the target parameters,
- etc.

and, if available, use them in the SM.

ii.1) adopt an alternative approach to SM based on [exploring uncertainty](#) (only with [macro goal](#); cf. D'Orazio et. al 2006a, 2006b)

Key steps in SM

Q1: Can we assume independence of Y and Z conditional on X ?

- **YES** -> apply SM methods based on CI
- **NO** -> go to Q2

Q2: Is auxiliary information available?

- **YES** -> apply SM methods exploiting auxiliary information
- **NO** -> assess **uncertainty** in your SM problem (only macro goal)

Usually, SM of surveys NOT designed and treated with integration purposes (matching ex-post) will be **unfeasible** or **feasible but with poor results** because of:

- Differences in the **definition of the target population**
- Differences in the **definitions of common variables** (non-reconcilable)
- Few common variables and not being good predictors of target ones
- **CI is NOT a valid assumption** (no proxies, nor auxiliary information)
- CI holds for (X, Y, Z_1) but not for (X, Y, Z_2)

Performing statistical matching in R

- use some packages developed to impute missing values
- use [StatMatch](#) (D'Orazio, 2019b, v. 1.3.0)

Example of a matching application in R using StatMatch

https://github.com/marcellodo/StatMatch/blob/master/2019-05_Tutorial_uRos2019/ExampleCode.R

Example data in StatMatch, `samp.A` and `samp.B`, i.e. artificial data set resembling EU–SILC survey

```
# X variables
> intersect(colnames(samp.A), colnames(samp.B))
[1] "HH.P.id" "area5" "urb" "hsize" "hsize5" "age"
[7] "c.age" "sex" "marital" "edu7" "ww"

> # variables only in samp.A (Y)
> setdiff(colnames(samp.A), colnames(samp.B))
[1] "n.income" "c.neti"

# variables only in samp.B (Z)
> setdiff(colnames(samp.B), colnames(samp.A))
[1] "labour5"
```

Step 1) Are the samples representing the same population?

- Check definition of target population
- Check definition of sampling unit (household in the example)
- Check sampling frames used to select the samples
- Collect information concerning nonsampling errors (nonresponse, measurement, etc.) and corrections of design weights
- Check definitions, estimates, distributions etc. for key variables (no. of households, no. of households by size, number of people, by gender, by age, etc.)

If samples reflect partially overlapping populations, matching can only be done for the overlapping part (discard non-overlapping from samples)

Example of Statistical Matching in R with StatMatch Package: Step (1)

```
> #####
> # check target population
>
> # estimated population size
> sum(samp.A$ww) # sum of survey weights
[1] 5094952
> sum(samp.B$ww) # sum of survey weights
[1] 5157582

> # distribution by regions
> ttA <- xtabs(ww~area5, data=samp.A)
> ttA
area5
      NE      NO      C      S      I
1215201.2 1050338.2 1089207.6 1204930.2 535274.6
> ttB <- xtabs(ww~area5, data=samp.B)
> ttB
area5
      NE      NO      C      S      I
1389409.6 998670.1 1071398.2 1174936.4 523167.8
```

Example of Statistical Matching in R with StatMatch Package: Step (1)

```
> cbind(A=prop.table(ttA) ,
+       B=prop.table(ttB))
      A      B
NE 0.2385108 0.2693917
NO 0.2061527 0.1936315
C  0.2137817 0.2077327
S  0.2364949 0.2278076
I  0.1050598 0.1014366

> # measure closeness between distributions
> comp.prop(p1 = ttA, p2 = ttB,
+          n1 = nrow(samp.A), n2 = nrow(samp.B), ref = F)
$meas
      tvd      overlap      Bhatt      Hell
0.03088081 0.96911919 0.99935387 0.02541911
$chi.sq
      Pearson      df      q0.05      delta.h0
10.571108      4.000000      9.487729      1.114187

$p.exp
area5
      NE      NO      C      S      I
0.2598073 0.1975176 0.2096101 0.2305038 0.1025611
```

Example of Statistical Matching in R with StatMatch Package: Step (1)

```
> # distribution by (region x gender x age-classes)
> ttA <- xtabs(ww~sex+c.age+area5, data=samp.A)
> ttB <- xtabs(ww~sex+c.age+area5, data=samp.B)
> comp.prop(p1 = ttA, p2 = ttB,
+           n1 = nrow(samp.A), n2 = nrow(samp.B), ref = F)
$meas
      tvd      overlap      Bhatt      Hell
0.07459763 0.92540237 0.99489746 0.07143203
```

Step 2) Identify common variables (Xs)


- Check definitions of variables; if different, is it possible to harmonize?
If harmonization is NOT possible --> discard
- Check marginal distributions of common variables

Example of Statistical Matching in R with StatMatch Package: Step (2)

```
> #####  
> # check marginal distr. of common variables  
> ttA <- xtabs(ww~edu7, data=samp.A)  
> ttB <- xtabs(ww~edu7, data=samp.B)  
> cc <- comp.prop(p1 = ttA, p2 = ttB,  
+                n1 = nrow(samp.A), n2 = nrow(samp.B),  
+                ref = F)  
> cc$meas  
      tvd      overlap      Bhatt      Hell  
0.01048456 0.98951544 0.99986854 0.01146559
```

Step 3) is CI assumption holding?

I.e., are $Y = \text{income}$ (`c.neti`) and $Z = \text{prof. status}$ (`labour5`) independent once conditioning on a subset of the available X s?

- Consult subject matter experts → NO in the case of example
 - Search X s for a proxy of Y or Z
 - Search for best predictors of Y
 - Search for best predictors of Z
- Select matching variables
- 

Example of Statistical Matching in R with StatMatch Package: Step (3)

```
> #####  
> # best predictors of n.income (Y, is continuous)  
> Hmisc::spearman2(n.income~area5+urb+hsize5+c.age+sex+marital+edu7,  
+                  data=samp.A)
```

```
Spearman rho^2      Response variable:n.income
```

	rho2	F	df1	df2	P	Adjusted rho2	n
area5	0.033	25.45	4	3004	0.0000	0.031	3009
urb	0.000	0.49	2	3006	0.6105	0.000	3009
hsize5	0.032	25.01	4	3004	0.0000	0.031	3009
c.age	0.100	83.33	4	3004	0.0000	0.099	3009
sex	0.120	410.25	1	3007	0.0000	0.120	3009
marital	0.034	53.02	2	3006	0.0000	0.033	3009
edu7	0.071	38.17	6	3002	0.0000	0.069	3009

no proxies of `n.income` !!!!

Example of Statistical Matching in R with StatMatch Package: Step (3)

```
> # best predictors of labour5 (Z, is categorical)
>
> pws <- pw.assoc(labour5~area5+urb+hsize5+c.age+sex+marital+edu7,
+               data=samp.B, out.df = TRUE)
Warning message:
In chisq.test(tab) : Chi-squared approximation may be incorrect
>
> pws[, c("norm.mi", "U", "AIC", "npar")]
```

	norm.mi	U	AIC	npar
labour5_area5	0.0163106664	0.0163106664	19232.25	20
labour5_urb	0.0009598301	0.0007064674	19520.70	12
labour5_hsize5	0.0174329815	0.0166305027	19226.01	20
labour5_c.age	0.2485754479	0.2485754479	14700.65	20
labour5_sex	0.0782041550	0.0371158413	18802.33	8
labour5_marital	0.0643267975	0.0420075216	18714.89	12
labour5_edu7	0.0803914798	0.0803914798	17998.00	28

No proxies of labour5!!!!

Which are the matching variables?

Are the X_M variables being good predictors of both Z and Y :

$$X_Y \cap X_Z \subseteq X_M \subseteq X_Y \cup X_Z$$

X_Y : subset of X variables ($X_Y \subseteq X$) being good predictors of Y

X_Z : subset of X variables ($X_Z \subseteq X$) being good predictors of Z

NB: Avoid choosing too many matching variables, they may add undesired additional noise affecting the results of SM (e.g. marginal distribution of the variable Z imputed in A may not be coherent with the one observed in B)

Example of Statistical Matching in R with StatMatch Package: Step (3)

```
> # matching variables

> x_y <- c("c.age", "sex", "edu7")

> x_z <- c("c.age", "sex", "marital", "edu7")

> intersect(x_y, x_z)
[1] "c.age" "sex"   "edu7"

> union(x_y, x_z)
[1] "c.age"   "sex"     "edu7"    "marital"
```

D'Orazio et al. (2017, 2019) introduced a strategy for selecting matching variables based on uncertainty reduction.

Requires all variables (Xs, Y and Z) to be categorical.

Background:

when estimating cells' probabilities in a two-way table $Y \times Z$ the **Frechét-Bonferroni** bounds apply, i.e.:

$$\max \left\{ 0; P_{Y=j} + P_{Z=k} - 1 \right\} \leq P_{Y=j, Z=k} \leq \min \left\{ P_{Y=j}; P_{Z=k} \right\}$$

$$j = 1, \dots, J, \quad k = 1, \dots, K$$

When conditioning on X_D (obtained by crossing a subset of the X_s):

$$P_{j,k}^{(low)} \leq P_{Y=j, Z=k} \leq P_{j,k}^{(up)}$$

With expected conditional bounds:

$$P_{j,k}^{(low)} = \sum_{i=1}^I P_{X_D=i} \max \left\{ 0; P_{Y=j|X_D=i} + P_{Z=k|X_D=i} - 1 \right\}$$

$$P_{j,k}^{(up)} = \sum_{i=1}^I P_{X_D=i} \min \left\{ P_{Y=j|X_D=i}; P_{Z=k|X_D=i} \right\}$$

A rough estimate of the **overall uncertainty** is provided by the [Average width of intervals](#):

$$\bar{d} = \frac{1}{J \times K} \sum_{j,k} \left[\hat{P}_{j,k}^{(up)} - \hat{P}_{j,k}^{(low)} \right]$$

Example of Statistical Matching in R with StatMatch Package: Step (3)

Rationale: identify the subset of X s more effective in reducing uncertainty, avoiding selecting too many (avoid sparse contingency tables)

```
> # rescale weights to sum up to n
> wwA <- samp.A$ww / sum(samp.A$ww) * nrow(samp.A)
> wwB <- samp.B$ww / sum(samp.B$ww) * nrow(samp.B)

> #estimate joint ditribution of starting Xs
> txA <- xtabs(wwA~area5+urb+hsize5+c.age+sex+marital+edu7,
+             data=samp.A)

> txB <- xtabs(wwB~area5+urb+hsize5+c.age+sex+marital+edu7,
+             data=samp.B)
> txx <- txA+txB

> # estimate table Xs vs. Y
> txyA <- xtabs(wwA~area5+urb+hsize5+c.age+sex+marital+edu7+c.neti,
+             data=samp.A)
> # estimate table Xs vs. Z
> txzB <- xtabs(wwB~area5+urb+hsize5+c.age+sex+marital+edu7+labour5,
+             data=samp.B)
```

Example of Statistical Matching in R with StatMatch Package: Step (3)

```
> unc <- selMtc.by.unc(tab.x=txx, tab.xy=txyA, tab.xz=txzB,  
+                      corr.d=2)
```

```
> unc$av.df
```

	x.vars	nxv	nc.x	nc0.x	av.crf.x	veq.x	nc.xy	nc0.xy
1	<NA>	0	NA	NA	NA	NA	7	0
2	c.age	1	5	0	1939.0	0.1817893	35	0
3	c.age*sex	2	10	0	969.5	0.2001697	70	0
4	c.age*sex*edu7	3	70	2	138.5	1.1689166	490	168
5	c.age*sex*edu7*area5	4	350	41	27.7	1.2738578	2450	1485

	av.crf.xy	veq.xy	nc.xz	nc0.xz	av.crf.xz	veq.xz	min.av
1	429.857143	0.4087089	5	0	1337.200000	0.5600253	429.857143
2	85.971429	0.7108845	25	1	267.440000	1.0641151	85.971429
3	42.985714	0.8405931	50	3	133.720000	1.1520055	42.985714
4	6.140816	1.9258753	350	111	19.102857	2.1197475	6.140816
5	1.228163	2.2733175	1750	910	3.820571	2.3572067	1.228163

	avw	penalty	avw.pen
1	0.11380081	0.0000000000	0.11380081
2	0.08714187	0.0003362475	0.08747812
3	0.07735921	0.0003402518	0.07769946
4	0.06990577	0.0003969829	0.07030275
5	0.05772277	0.0017889088	0.05951168

CI assumption does not seem valid

Step 4) perform matching

in absence of proxies and additional data sources, where Y and Z are jointly observed, one should only perform **assessment of uncertainty**. When X, Y and Z are all categorical, `Frechet.bounds.cat` function can be used (summary info already provided by `selMtc.by.unc` function)

```
> # joint of matching variables Xs
> txA <- xtabs(wwA~area5+c.age+sex+edu7,
+             data=samp.A)
>
> txB <- xtabs(wwB~area5+c.age+sex+edu7,
+             data=samp.B)
> txx <- txA+txB
>
> # joint X vs. Y
> txyA <- xtabs(wwA~area5+c.age+sex+edu7+c.neti,
+             data=samp.A)
```

Example of Statistical Matching in R with StatMatch Package: Step (4)

```
> # joint X vs. Z
> txzB <- xtabs(wxB~area5+c.age+sex+edu7+labour5,
+              data=samp.B)

> # estimate frechet-bonferroni bounds for relative frequencies
> # in table c.neti vs. labour5

> fbw <- Frechet.bounds.cat(tab.x = txx,
+                           tab.xy = txyA, tab.xz = txzB,
+                           print.f = "data.frame",
+                           align.margins =TRUE)
> head(fbw$bounds, 4)
```

	c.neti	labour5	low.u	low.cx	CIA	up.cx	up.u
1	(-Inf,0]	1	0	0.006529978	0.06022077	0.1129133	0.1937787
2	(0,10]	1	0	0.005858603	0.05545869	0.1094139	0.2235263
3	(10,15]	1	0	0.006609817	0.05455212	0.1071773	0.1775354
4	(15,20]	1	0	0.015010174	0.06866731	0.1210169	0.1632569

Just for illustrative purposes let's assume that CI is valid.

Step 4) perform matching

A number of method can be applied.

For Statistical matching at micro level, **StatMatch** offers possibility of:

- Stochastic regression imputation (Y and Z continuous, Xs also)
- Hotdeck (random, Nearest Neighbor Distance, rank)
- Mixed (regression + NND hot deck)

Since the matching variables

`c.age (or age), sex, edu7, area5`

are all categorical with exception of age:

- random hotdeck within fixed classes formed by crossing some variables
- random hotdeck within non-fixed classes
- Nearest Neighbour Distance hotdeck within classes, i.e. distance between units belonging to the same classes

Pick at random a donor having same characteristic of recipient, i.e. divide units in groups (donation classes) and select donors at random within them

```
> # check for empty classes in donor
> dcA <- xtabs(~c.age+sex+edu7+area5, data=samp.A)
> dcB <- xtabs(~c.age+sex+edu7+area5, data=samp.B)

> tst <- dcA>0 & dcB==0
> sum(tst)
[1] 10
```

There 10 classes with 0 donors in B, while corresponding classes in recipient (A) are non-empty → drop one X variable (*area5*)

```
> # discard area5
> dcA <- xtabs(~c.age+sex+edu7, data=samp.A)
> dcB <- xtabs(~c.age+sex+edu7, data=samp.B)

> tst <- dcA>0 & dcB==0
> sum(tst)
[1] 1
```

Example of Statistical Matching in R with StatMatch Package: Step (4)

```
> # discard edu7
>
> dcA <- xtabs(~c.age+sex, data=samp.A)
> dcB <- xtabs(~c.age+sex, data=samp.B)

> tst <- dcA>0 & dcB==0
> sum(tst)
[1] 0
```

NO empty donor classes --> run random hotdeck with fixed classes formed crossing c.age and sex

```
> out.rnd <- RANDwNND.hotdeck(data.rec = samp.A, data.don = samp.B,
                             don.class = c("c.age", "sex"))
```

Create the synthetic data set (samp.A is the recipient)

```
> fillA.rnd.1 <- create.fused(data.rec = samp.A, data.don = samp.B,
                             mtc.ids = out.rnd$mtc.ids, dup.x = T,
                             match.vars = c("c.age", "sex"),
                             z.vars = "labour5")
```

NON-fixed classes: Random selection of one of k=5 closest donors in terms of age, having the same gender and education level

```
> #####  
> ## Random Hot-deck within NON-fixed classes  
> out.rnd2 <- RANDwNND.hotdeck(data.rec = samp.A, data.don = samp.B,  
+                             don.class = c("edu7", "sex"),  
+                             match.vars = "age", cut.don = "exact",  
+                             k = 5)  
  
> head(out.rnd2$sum.dist, 4)  
      min max      sd cut dist.rd  
[1,]   0  49 11.97574   4       4  
[2,]   0  58 17.13920   1       1  
[3,]   1  46 10.61782   3       3  
[4,]   0  42 11.17538   1       1  
  
> # create synthetic data set, samp.A is the recipient  
> fillA.rnd.2 <- create.fused(data.rec = samp.A, data.don = samp.B,  
+                             mtc.ids = out.rnd2$mtc.ids, dup.x = T,  
+                             match.vars = c("c.age", "sex"),  
+                             z.vars = "labour5")
```

Nearest Neighbour distance hotdeck: `NND.hotdeck()` function

- within classes formed on 'sex'
- distance calculated on 'age'
- **unconstrained**: a donor can be used more than once

```
> #####  
> ## Nearest neighbour distance hot-deck  
  
> out.nnd1 <- NND.hotdeck(data.rec = samp.A, data.don = samp.B,  
+                          don.class = "sex",  
+                          match.vars = "age")  
Warning: The Manhattan distance is being used  
All the categorical matching variables in rec and don  
data.frames, if present are recoded into dummies  
  
> summary(out.nnd1$dist.rd)  
      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.   
0.0000000 0.0000000 0.0000000 0.0006647 0.0000000 1.0000000
```


Example of Statistical Matching in R with StatMatch Package: Step (4)

```
> # create synthetic data set, samp.A is the recipient
> fillA.nnd.1 <- create.fused(data.rec = samp.A, data.don = samp.B,
+                             mtc.ids = out.nnd1$mtc.ids, dup.x = T,
+                             match.vars = c("age", "sex"),
+                             z.vars = "labour5")
```

- **constrained**: a donor can be used just once

```
> out.nnd2 <- NND.hotdeck(data.rec = samp.A, data.don = samp.B,
+                          don.class = "sex",
+                          match.vars = "age",
+                          constrained = T, k = 1,
+                          constr.alg = "hungarian")
>
> summary(out.nnd2$dist.rd)
      Min.  1st Qu.  Median    Mean  3rd Qu.   Max.
0.000000 0.000000 0.000000 0.001329 0.000000 1.000000
> summary(out.nnd1$dist.rd)
      Min.  1st Qu.  Median    Mean  3rd Qu.   Max.
0.0000000 0.0000000 0.0000000 0.0006647 0.0000000 1.0000000
```

Example of Statistical Matching in R with StatMatch Package: Step (4)

```
> # create synthetic data set, samp.A is the recipient
> fillA.nnd.2 <- create.fused(data.rec = samp.A, data.don = samp.B,
+                             mtc.ids = out.nnd2$mtc.ids, dup.x = T,
+                             match.vars = c("age", "sex"),
+                             z.vars = "labour5")
```

Step 5) Assess matching results

- check marginal distribution of imputed variable

```
> # compare marginal distribution of imputed income vs. reference one
> # unweighted
> t.imp <- table(fillA.rnd.1$labour5)
> t.don <- table(samp.B$labour5)

> ck1 <- comp.prop(p1 = t.imp, p2 = t.don,
+                 n1 = nrow(fillA.rnd.1),
+                 n2 = nrow(samp.B),
+                 ref = T)

> ck1$meas
      tvd      overlap      Bhatt      Hell
0.008380383 0.991619617 0.999940477 0.007715100
```

Example of Statistical Matching in R with StatMatch Package: Step (5)

```
> # weighted
> t.imp <- xtabs(ww~labour5, data = fillA.rnd.1)
> t.don <- xtabs(ww~labour5, data = samp.B)

> ck1.w <- comp.prop(p1 = t.imp, p2 = t.don,
+                   n1 = nrow(fillA.rnd.1),
+                   n2 = nrow(samp.B),
+                   ref = T)

> ck1.w$meas
      tvd      overlap      Bhatt      Hell
0.01488083 0.98511917 0.99979976 0.01415066
```

Example of Statistical Matching in R with StatMatch Package: Step (5)

- check joint distribution of imputed variable with matching variables

```
> # check joint distribution of imputed income with matching variables
> # unweighted
> t.imp <- xtabs(~labour5+c.age+sex, data = fillA.rnd.1)
> t.don <- xtabs(~labour5+c.age+sex, data = samp.B)
>
> cc <- comp.prop(p1 = t.imp, p2 = t.don,
+               n1 = nrow(fillA.rnd.1),
+               n2 = nrow(samp.B),
+               ref = T)
> cc$meas
      tvd      overlap      Bhatt      Hell
0.03867891 0.96132109 0.99734804 0.05149721

> # weighted
> t.imp <- xtabs(ww~labour5+c.age+sex, data = fillA.rnd.1)
> t.don <- xtabs(ww~labour5+c.age+sex, data = samp.B)
>
> cc <- comp.prop(p1 = t.imp, p2 = t.don,
+               n1 = nrow(fillA.rnd.1),
+               n2 = nrow(samp.B),
+               ref = T)
> cc$meas
      tvd      overlap      Bhatt      Hell
0.06122068 0.93877932 0.99553237 0.06684035
```

A comparison of different methods

```
> # marginal of imputed (weighted)
> t.don <- xtabs(wv~labour5, data = samp.B)
> t.imp.rnd1 <- xtabs(wv~labour5, data = fillA.rnd.1)
> t.imp.rnd2 <- xtabs(wv~labour5, data = fillA.rnd.2)
> t.imp.nndu <- xtabs(wv~labour5, data = fillA.nnd.1)
> t.imp.nndc <- xtabs(wv~labour5, data = fillA.nnd.2)
>
> a <- comp.prop(p1 = t.imp.rnd1, p2 = t.don,
+               n1 = nrow(fillA.rnd.1),
+               n2 = nrow(samp.B),
+               ref = T)
...

> rbind(rnd.1=a$meas, rnd.2=b$meas,
+       nnd.unc=c$meas, nnd.c=d$meas)
              tvd    overlap    Bhatt      Hell
rnd.1    0.02369220 0.9763078 0.9996303 0.01922830
rnd.2    0.03958675 0.9604133 0.9987532 0.03530970
nnd.unc  0.02251686 0.9774831 0.9994403 0.02365720
nnd.c    0.01731528 0.9826847 0.9997257 0.01656167
```

Example of Statistical Matching in R with StatMatch Package: Step (5)

```
> # Joint of imputed vs. matching (weighted)
> t.don <- xtabs(wv~labour5+c.age+sex, data = samp.B)
> t.imp.rnd1 <- xtabs(wv~labour5+c.age+sex, data = fillA.rnd.1)
> t.imp.rnd2 <- xtabs(wv~labour5+c.age+sex, data = fillA.rnd.2)
> t.imp.nndu <- xtabs(wv~labour5+c.age+sex, data = fillA.nnd.1)
> t.imp.nndc <- xtabs(wv~labour5+c.age+sex, data = fillA.nnd.2)

> a <- comp.prop(p1 = t.imp.rnd1, p2 = t.don,
+               n1 = nrow(fillA.rnd.1),
+               n2 = nrow(samp.B),
+               ref = T)

...

> rbind(rnd.1=a$meas, rnd.2=b$meas,
+       nnd.unc=c$meas, nnd.c=d$meas)
      tvd   overlap   Bhatt   Hell
rnd.1  0.06122068 0.9387793 0.9955324 0.06684035
rnd.2  0.06522719 0.9347728 0.9928784 0.08438958
nnd.unc 0.05517412 0.9448259 0.9959244 0.06384012
nnd.c   0.06895387 0.9310461 0.9955954 0.06636696
```

Step 6) Estimate target parameters from the synthetic dataset

```
> # estimate table Y vs. Z in filled A
> # based on the chosen method: NND unconstrained

> t.yz <- xtabs(ww~c.neti+labour5, data = fillA.nnd.1)

> round(addmargins( prop.table(t.yz))*100,2)
      labour5
c.neti      1      2      3      4      5      Sum
(-Inf,0]    5.59    1.41    1.79    1.20    9.38    19.38
(0,10]      6.11    1.66    1.18    6.77    6.63    22.35
(10,15]     5.34    1.41    1.05    4.92    5.03    17.75
(15,20]     6.56    2.49    0.82    3.47    2.98    16.33
(20,25]     4.65    1.58    0.58    1.62    1.33     9.75
(25,35]     4.54    1.56    0.57    1.58    1.31     9.57
(35, Inf]   2.05    0.91    0.22    1.10    0.58     4.87
Sum        34.85   11.01    6.21   20.67   27.25   100.00

> # estimate association
> assoc <- pw.assoc(c.neti~labour5, data = fillA.nnd.1,
+                  out.df = T, weights = "ww")
> assoc$V
[1] 0.1791729
> assoc$norm.mi
[1] 0.04498098
```


Main References

- Conti, PL and Marella, D and Scanu, M (2012) “Uncertainty Analysis in Statistical Matching”, *Journal of Official Statistics*, **28**, pp. 69-88.
- Conti, P.L., D. Marella, and M. Scanu. (2013) “Uncertainty Analysis for Statistical Matching of Ordered Categorical Variables.” *Computational Statistics & Data Analysis*, **68**, pp. 311–325.
- D’Orazio, M. (2015) “Integration and imputation of survey data in R: the StatMatch package”. *Romanian Statistical Review*, 2/2015, pp. 57-68.
- D’Orazio, M. (2019a) “Statistical Learning in Official Statistics: the case of Statistical Matching”. Presentation at NTTS 2019 Conference, Bruxelles, 12-14 March 2019 (*paper in preparation*).
- D’Orazio, M (2019b) “StatMatch: Statistical Matching”, R package version 1.3.0
<http://CRAN.R-project.org/package=StatMatch>
- D’Orazio, M and 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 Istat*, 10 (2005)
- D’Orazio, M and Di Zio, M and Scanu, M (2006a), “Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints”, *Journal of Official Statistics*, **22**, pp. 137-157.
- D’Orazio, M and Di Zio, M and Scanu, M (2006b) *Statistical Matching: Theory and Practice*. Wiley, Chichester
- D’Orazio M., Di Zio M., Scanu M. (2016) “The Use of Uncertainty to Choose Matching Variables in Statistical Matching”, in: Ferraro et al. (eds.) *Soft Methods for Data Science*, Springer (ISBN 978-3-319-42971-7)
- D’Orazio M., Di Zio M., Scanu M. (2017) “The use of uncertainty to choose matching variables in statistical matching”. *International Journal of Approximate Reasoning*, **90**, pp. 433–440
- D’Orazio M., Di Zio M., Scanu M. (2019) “Auxiliary variable selection in a statistical matching problem”, in Zhang L.C. and Chambers R.L (eds.) *Analysis of Integrated Data*. Chapman and Hall/CRC, pp. 101–120 (*forthcoming*)
- Donatiello G., D’Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2014) “Statistical Matching of Income and Consumption expenditures”. *International Journal of Economic Science*, Vol. **III** (No. 3), pp. 50-65.

Main References

- Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M, Spaziani M. (2016a) "The role of the conditional independence assumption in statistically matching income and consumption", *International Journal of the IAOS*
- Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2016b) "The statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics". DGINS - Conference of the Directors General of the National Statistical Institutes, 26-27 September 2016, Vienna.
- Lavallée, P. (2007). *Indirect Sampling*. Springer, New York
- Little R.J.A., Rubin D.B. (2002) *Statistical Analysis with Missing Data, 2nd Edition*. Wiley, New York.
- Moriarity, C and Scheuren, F (2001) "Statistical Matching: a Paradigm for Assessing the Uncertainty in the Procedure", *Journal of Official Statistics*, **17**, pp. 407-422
- 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**, pp. 65-73
- Rässler, S, (2002) *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer-Verlag, New York.
- Rässler, S, (2003) "A Non-iterative Bayesian Approach to Statistical Matching". *Statistica Neerlandica*, **57**, pp. 58-74.
- Renssen RH (1998) "Use of statistical matching techniques in calibration estimation". *Survey Methodology*, **24**, pp. 171-183
- Rivers, D. (2007) "Sampling for web surveys", *Proceedings Joint Statistical Meeting*
- Rubin, DB (1986) "Statistical matching using file concatenation with adjusted weights and multiple imputations", *Journal of Business and Economic Statistics*, **4**, pp. 87-94
- Singh, AC and Mantel, H and Kinack, M and Rowe, G (1993) "Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption", *Survey Methodology*, **19**, pp. 59-79.
- Zhang, L-C. (2015) "On Proxy Variables and Categorical Data Fusion", *Journal of Official Statistics*, **31**