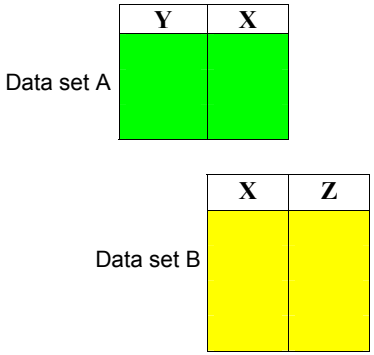
SOME EXPERIMENTS ON STATISTICAL MATCHING IN R

Marcello D'Orazio, Marco Di Zio, Mauro Scanu*
[madorazi, dizio, scanu](at)istat.it

*Italian National Statistical Institute, Istat

Statistical Matching (SM) (aka data fusion, or synthetical matching) aims to integrate two or more data sets related to the same target population.

Classical SM framework:



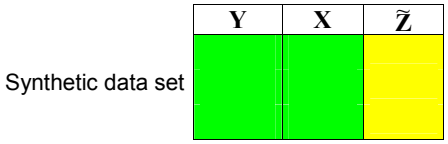
i) **X** is the set of variables common to both the data sets.

ii) **Y** and **Z** are not jointly observed

iii) The data sets refer to different units (no units in common)

SM objectives:

✓ Micro: construction of a complete synthetic data set. All the variables of interest, although collected in different sources, are contained in it.



✓ Macro: direct estimation of the joint distribution function, $f(\mathbf{x},\mathbf{y},\mathbf{z})$, (or of some of its key characteristics, e.g. ρ_{XY}) of the variables of interest which have not been jointly observed.

	Approaches to SM		
SM objectives	Parametric	Nonparametric	Mixed
Macro	✓	✓	
Micro	✓	✓	✓

SM mixed methods

Step 1) a parametric model is assumed and its parameters are estimated.

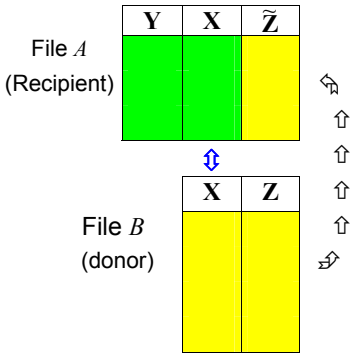
Step 2) a completed synthetic data set is derived through nonparametric micro approach.

Main advantage: nonparametric procedures are more robust to model misspecification.

Most of the first SM applications had a micro objective and used nonparametric methods.

In particular, imputation procedures have been used to fill in the missing variables in the recipient data set (or host file).

Each record of the recipient file is imputed using records (chosen suitably) from the other sample, the donor file.



Widely used hot deck imputation procedures:

- random hot deck (R function `RANDhd.mtc`)
- rank hot deck
- distance hot deck (R function `NNDhd.mtc` and `strNNDhd.mtc`)

Our R functions for SM can be downloaded at:

http://www.wiley.com/go/matching

The hot deck imputation procedures based on the values of the common variables $\mathbf{X}$ (the matching variables) implicitly assume the independence of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$:

$$f(\mathbf{x},\mathbf{y},\mathbf{z}) = f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \cdot f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}) \cdot f_{\mathbf{X}}(\mathbf{x}) \qquad [1]$$

usually known as Conditional Independence Assumption (CIA).

Under the CIA, if a parametric model is considered, this model is identifiable for $A \cup B$ and therefore its parameters can be directly estimated.

Eq. [1] can be estimated with the information in the two data-set.

Example: (X,Y,Z) follow a trivariate normal distribution

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{pmatrix}$$

All the parameters can be estimated directly on A, B and $A \cup B$ with the exception of σ_{YZ}.

If the CIA holds, $\sigma_{YZ} = (\sigma_{YX}\sigma_{ZX})/\sigma_X^2$ ($\rho_{YZ} = \rho_{XY} \cdot \rho_{ZX}$) and therefore

$$\hat{\sigma}_{YZ} = \hat{\sigma}_{YX}\hat{\sigma}_{ZX} / \hat{\sigma}_X^2$$

R function `MLmixed.mtc` provides ML estimates of all the parameters (D'Orazio *et al.*, 2006)

The CIA can NOT be tested from the available data sets A and B.

Very often the CIA does not hold.

SM techniques based on the CIA when it is not valid provide misleading results.

To obtain reliable results external auxiliary information about the joint relationship of $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ or between $(\mathbf{Y},\mathbf{Z})$ must be used in the SM procedures.

Examples of external auxiliary information:

a) a third file C where either $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ or $(\mathbf{Y},\mathbf{Z})$ are jointly observed;
b) plausible values of the inestimable parameters of $(\mathbf{Y},\mathbf{Z}|\mathbf{X})$ or $(\mathbf{Y},\mathbf{Z})$

In the case trivariate normal distribution we carried out simulation studies in order to:

i) compare mixed procedure vs. parametric procedures, under the CIA.

ii) evaluate some SM parametric procedures that make use of external estimates for ρ_{YZ} or $\rho_{YZ|X}$.

iii) compare some SM mixed procedures that make use of external estimates for ρ_{YZ} or $\rho_{YZ|X}$:
the mixed method by Moriarity and Scheuren (2003) (implemented in the R function `MoriSche.mtc`) vs.
a similar procedure based on ML estimation of the parameters (R function `MLmixed.mtc`)

A more general approach to SM

Consist in evaluating the uncertainty associated with the estimates provided by SM macro methods when the CIA does not hold.

Uncertainty refers to the fact that, due to the classical SM framework, no unique estimate can be provided for the parameter of interest.

For instance, in the case of the trivariate normal distribution it can be shown that:

$$\rho_{XY}\rho_{XZ} - \left[(1-\rho_{XY}^2)(1-\rho_{XZ}^2)\right]^{1/2} \leq \rho_{YZ} \leq \rho_{XY}\rho_{XZ} + \left[(1-\rho_{XY}^2)(1-\rho_{XZ}^2)\right]^{1/2}$$

In case of categorical variables we:

a) explored the uncertainty in estimating the parameters of a multinomial model (cell probabilities) by using the EM algorithm (Schafer, function `em.cat` in library `cat`)

b) showed how to reduce the uncertainty in parameters estimation via parameter constraints:
i) existence of some quantities (some cell probabilities must be 0; structural zeros);
ii) inequality constraints: $\theta_{ijk} \leq \theta_{i'j'k'}$

(i) can be easily introduced in `em.cat`.

(ii) achieved by using the EMH algorithm (R function `emcat.c` obtained by slightly modifying `em.cat`).

A SM application with sample survey data on Italian Households

The objective was the estimation of the module of the Social Accounting Matrix (SAM) related to the households:

Households Categories	Expenditure categories			Income categories		
	C_1	...	C_U	M_1	...	M_V
T_1	c_{11}	...	c_{1U}	m_{11}	...	m_{1V}
...	...	...	...	...	...	...
T_w	c_{w1}	...	c_{wU}	m_{w1}	...	m_{wV}
...	...	...	...	...	...	...
T_W	c_{W1}	...	c_{WU}	m_{W1}	...	m_{WV}

Attempt to estimate this SAM module through Statistical Matching of Istat Household Budget Survey (HBS) and Survey of Household Income and Wealth (SHIW) carried out by the Bank of Italy (data for 2000)

The micro approach to SM was considered.

We studied the possibility of using the hot deck imputation procedures; namely, random hot deck and distance hot deck.

SHIW as the recipient (file A); HBS as the donor (file B)

The CIA did not seem a valid assumption.

We introduced auxiliary information in the matching step (ranking of each household with respect to the total monthly household income).

Future directions of work

- ✓ A comprehensive R library on Statistical Matching
  - new functions to deal with parametric models
  - improvements of some computationally intensive functions (constrained distance hot deck)
- ✓ Further research work in the direction of exploring uncertainty due to SM (extension to continuous variables)
- ✓ Application of SM techniques to integrate some other sample surveys data.

References

D'Orazio, M., Di Zio, M., and Scanu, M. (2006) *Statistical Matching: Theory and Practice*. Wiley and Sons, Chichester. http://www.wiley.com/go/matching

Harding, T., and Tusell, F. (2004) *cat: Analysis of categorical-variable datasets with missing values* (original by Joseph L. Schafer). http://www.stat.psu.edu/~jls/misoftwa.html#aut

Moriarity, C., and Scheuren, F. (2003) "A Note on Rubin's Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations", *Journal of Business & Economic Statistics*, **21**, 65-73.

Räessler, S. (2002) *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer-Verlag, New York.