Frascati, 12 November 2019

Integration of Sample Survey Data

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

**Senior Researcher in Statistics*
Office of Chief Statistician, Food and Agriculture Organization (FAO) of the United Nations
(Italian National Institute of Statistics - Istat)

1. Reasons for data integration
2. Techniques for integration and purposes
3. Integration of sample survey data: statistical matching
4. Pros and cons in statistical matching
5. Integration in the survey design phase

Exploit data already available

- No need to design and carry out a new comprehensive survey (sometimes not feasible)
  - No response burden
  - No additional costs of a new survey
  - Statistics produced through integration more timely than those of a new survey

BUT

- integration is not always feasible (mostly when not planned in advance) or feasible but with poor results

- maybe a time-consuming task (computational and human efforts; no at zero cost…)

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 …)

Typically applied to:

- ✓ Integrate **survey** data with **administrative** data
- ✓ Integrate two (or more) admin sources

Example of uses of RL in Official Statistics:

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

RL techniques based on ID variables that should be available in both the data sources

- **Exact record linkage (merging)**: couple of records sharing the same value of the (error-free) identifying variable(s) (Personal Id. Code, VAT number, etc.).

- **Probabilistic record linkage**: identifier is NOT error-free or it is not available but 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

  Estimated probabilities are used to link the records

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

A wide set of methods to integrate two data sources, **lacking of ID variables**, typically:

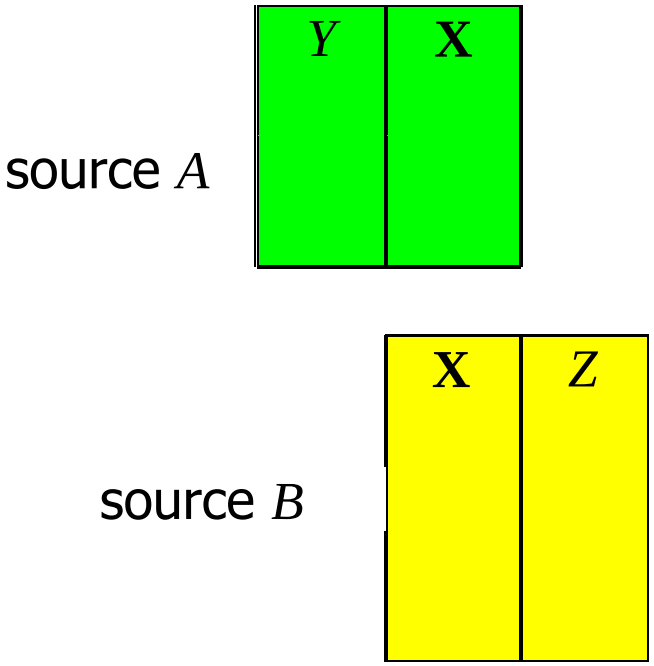
1) Two sample surveys referred to the same target population

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

2) A sample survey and data from another source (e.g. non-probabilistic sample)

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

SM 'basic' case:



1. $\mathbf{X}$ 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 (see e.g. D'Orazio et al., 2006)

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

Examples of SM at Istat with the goal of exploring the relationship between variables observed in separate sample surveys:

- Integrate Household Budget Survey (HBS) with Survey on Income and Living Conditions (SILC)
  To investigate the **relationship between income and expenditures** (Donatiello et al., 2014, 2016a, 2016b; ongoing project)

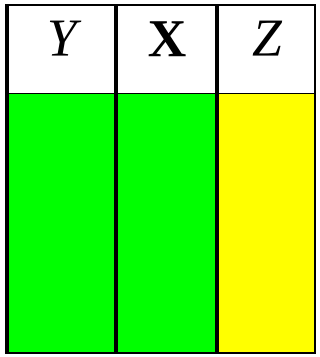
- Integrate Labor Force Survey (LFS) with Time Use Survey (TUS)
  To study the **relationship between time use and labor activity** (Gazzelloni et al, 2007)

- Integrate Farm Structure Survey (FSS) with Farm Accountancy Data network Survey (FADN)
  To investigate the **relationship between farm structure and farm economic performances** (Ballin et al., 2009)

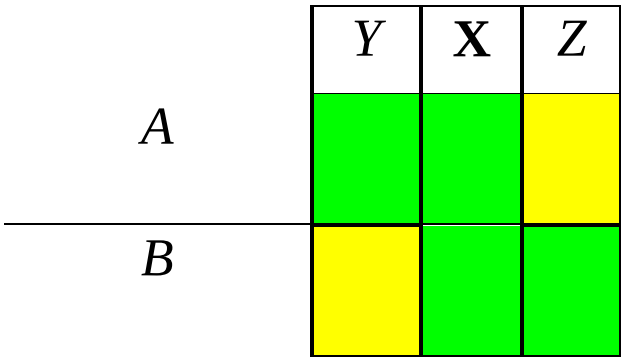
Case 1) investigate relationship between Y and Z

✓ **micro**: a "synthetic" data-set including $\mathbf{X}$, Y and Z is created

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



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



"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} (e.g. relationship between HH income and consumption)
  - 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

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 (selected) Xs
2) Impute in A the Z value observed on its closest donor in B

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)

In R some SM methods in the **StatMatch** package (D'Orazio, 2019)

SM assumptions (case 1 from now on)

i) A e B are representative samples of the same target population

ii) The $\boldsymbol{X}$ variables, shared by both the data sources, follow the same definitions and have the same distributions in both A and B.

iii) when the SM 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 conditional on $\mathbf{X}_M$

$$f\left(x_M, y, z\right) = f\left(y|x_M\right) f\left(z|x_M\right) f\left(x_M\right)$$

I.e.: the relationship between Y and Z is fully explained by $\mathbf{X}_M$

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:

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

Estimation is straightforward:

$$\widehat{\mathrm{P}}(Y = 1, Z = 1) = \sum_{i=1}^{I} \left[\widehat{P}^{(A)}(Y = 1|X = i) \times \widehat{P}^{(B)}(Z = 1|X = i) \times \widehat{P}^{(A \cup B)}(X = i) \right]$$

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:

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

is NOT 0 as one would expect!!!

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.

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 if **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)

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

Example: matching of SILC and HBS Istat surveys

Goal: to explore relationship between:

Y = HH income (observed in IT-SILC)

Z = HH overall consumption (observed in HBS)

Results are acceptable if one of the Xs is a proxy of income (Y^*)

Donatiello et al. (2016a, 2016b) used Y^* =income in classes

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

Currently Y^* =income transform (percentiles, …; *test ongoing*):

- in IT-SILC derived from Y
- predicted in HBS using variables observed in a new ad-hoc module of the HBS survey questionnaire

(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). I.e. how large is interval of admissible values for ρ_{yz}?

$$\left[\hat{\rho}_{xy}\hat{\rho}_{xz} - \sqrt{(1-\hat{\rho}_{xy}^2)(1-\hat{\rho}_{xz}^2)};\quad \hat{\rho}_{xy}\hat{\rho}_{xz} + \sqrt{(1-\hat{\rho}_{xy}^2)(1-\hat{\rho}_{xz}^2)}\right]$$

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)

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)

Thinking at integration in the survey design phase, 2 options:

a) **Integrated surveys framework** (best; many options)
   - A kind of 'unique' sample
   - 3 modules: **module 0** is common part (Xs), **module 1** (Ys) and **module 2** (Zs)
   - 3 subsamples:
     - A Observing Xs and Ys
     - B Observing Xs and Zs
     - C Observing Xs, some Ys and some Zs

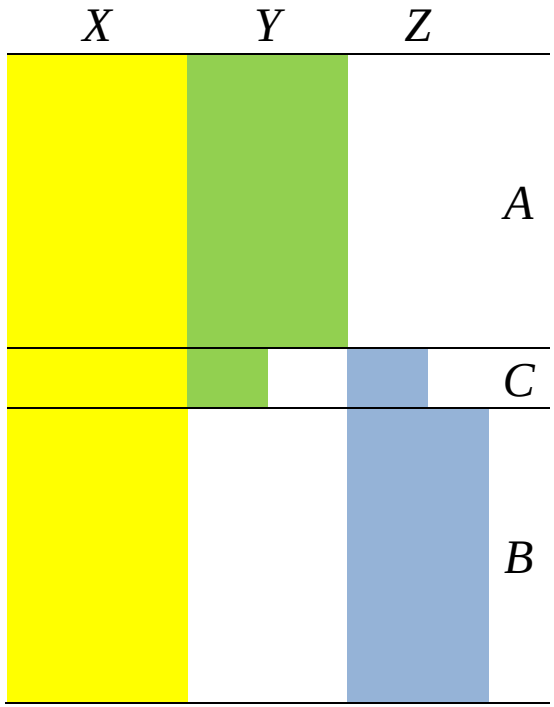
b) **Statistical matching** (ex-ante) (second best)
   - Two independent surveys
   - Same target population
   - Common definitions of the Xs (or of most of them)
   - Introduction of proxies to overcome CI, e.g.
     - Survey A observes Xs, Y and Z^* (a proxy of Z)
     - Survey B observes Xs, Z and Z^* (or Z^* is not observed but derived from Z)

Main issues in **Integrated surveys framework**

a) Major complexity in the design phase (size of subsamples? …)

b) Higher response burden for units in C

c) Estimation, various options, not straightforward



Thinking at integration in the design phase, Statistical Matching:

- SM can be performed to study relationship only between variables for which good proxies are observed/estimated with ad-hoc added questionnaire's modules (not all Zs or Ys)

Main References

Ballin M., D'Orazio M., Di Zio M., Scanu M., Torelli N. (2009) "Statistical Matching of Two Surveys with a Common Subset". *Working Paper*, N. 124, Università di Trieste

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 (2019) "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 (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. (2017) "The use of uncertainty to choose matching variables in statistical matching". *International Journal of Approximate* Reasoning, **90** (2017), 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.

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, At Vienna.

Gazzelloni S., Romano M.C., Corsetti G., Di Zio M., D'Orazio M., Pintaldi F., Scanu M., Torelli N. (2007) "Uso del tempo e Forze di lavoro: una proposta di integrazione dei dati mediante abbinamento statistico". In: (Romano, M. C. ed.) *I tempi della vita quotidiana: un approccio multidisciplinare all'analisi dell'uso del tempo*, *Argomenti* N. 32, Istat, pp. 375-403

Main References

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.

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**