7th Conference on the Use of R in official Statistics (uRos2019)
Bucharest, 12-14 March 2019

Integration of Survey Data in R Based on Machine Learning

Spaziani M., Frattarola D.*, **M. D'Orazio***[+]*
marcello.dorazio(at)fao.org
marcello.dorazio(at)istat.it

**Italian National Institute of Statistics – Istat, Rome, Italy*
[+]Office of Chief Statistician, Food and Agriculture Organization of the UN, Rome, Italy

Istat many years ago started a project for integration of data from social surveys with the objective of better measuring households' economic well-being at micro and macro level

Core of the project is the integration of:

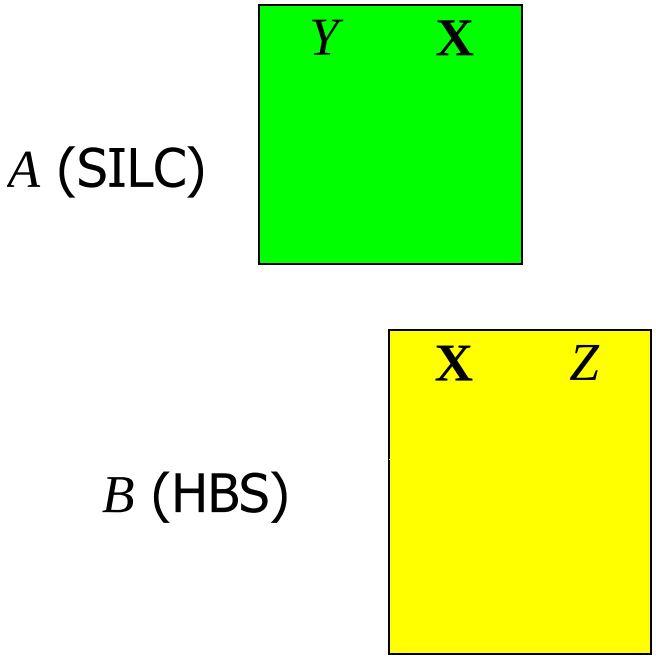
Survey on Income and Living Conditions (IT-SILC)
provides a picture of household income

Household Budget Survey (HBS).
investigates household expenditures

Objective: study relationship between income and consumption

Integration based on **statistical matching** (aka data fusion)

Statistical Matching (SM)
'basic' case:



1. A and B are representative samples of the same population (Italian HHs)
2. $\mathbf{X}$ are common variables (follow the same definitions)
3. Y (HH income) and Z (HH consumption) are NOT jointly observed
4. The probability of finding the same unit in both the sources is **0**

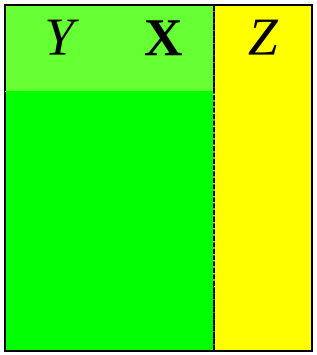
Goal of SM: exploring relationship between Y (income) and Z (consumption)

✓ **micro**: by creating a "synthetic" data-set that includes $\mathbf{X}$, Y and Z,

usually by imputing Z (HH consumption) in A (SILC)

A (SILC) is the *recipient*

B (HBS) is the *donor*



✓ **macro**: by estimating parameters, e.g.:

- correlation coefficient ρ_{YZ} (or β_{YZ})
- contingency table $Y \times Z$
- …

Basics of Statistical Matching:

- Integration is based on the common information, typically a suitable subset X_M of all the Xs is considered ($X_M \subseteq \boldsymbol{X}$)

  X_M are the **matching variables**

- Integration based on X_M implicitly assumes independence between Y and Z conditional on the X_M (**conditional independence**), i.e.

  HH income and HH consumption are independent, conditional on a series of characteristics of HHs, observed in both HBS and SILC

Conditional independence is seldom valid, unless one of the Xs in X_M is a **proxy** of one of the targets, i.e. highly associated/correlated with Y or Z.

D'Orazio (2019) two proposals for using **Statistical Learning*** in data fusion

1) predict Z (HH consumption) in A (SILC) using SL
   ↳ unsatisfactory results

2) assess uncertainty conditional on SL predictions of both Y and Z
   ↳ Promising results, i.e. reduction of uncertainty if compared to results obtained conditional on X_M (best predictors of Y and/or Z)

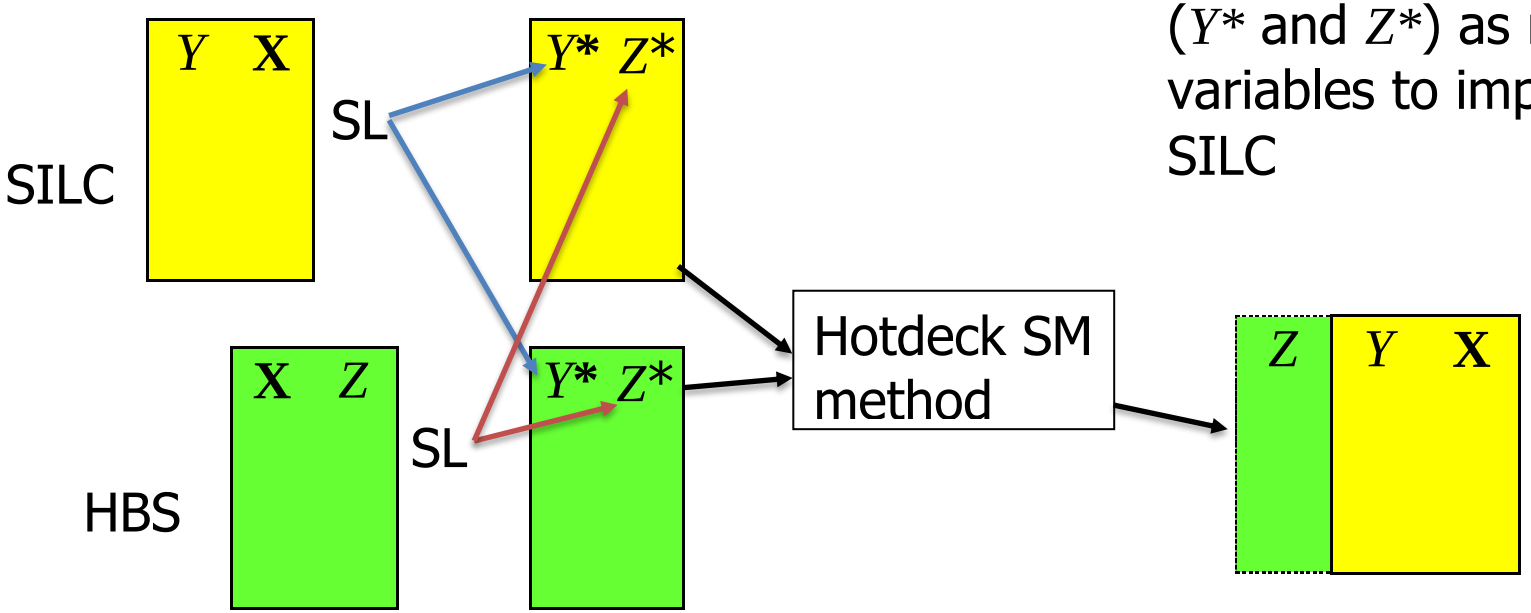
*Statistical (Machine) Learning (SL):

Wide set of techniques that "learn from the data" (Hastie et al., 2009)
Methods related to classification, regression and clustering (generalized additive models, classification and regression trees, neural networks, etc.). Very popular in marketing, finance, etc.
Allow analysis of large data sources.

Good results in case (2) represent basis for this work, i.e. **test a two-steps approach (micro) to SM**:

Step 1) use SL to predict Y and Z in both A and B

Step 2) Apply SM hotdeck method using predictions (Y^* and Z^*) as matching variables to impute Z in SILC



- avoids time consuming procedure for selecting the matching variables X_M (how?, how many?, etc.)
- does not directly impute predictions

SL methods being considered:

i.) **Naïve Bayes classifier** (classification based on Bayes theorem)

ii.) **Random Forest** (sequence of uncorrelated classification trees)

iii.) **C5.0** (classification tree aimed at reducing entropy)

iv.) **Adaptive Boosting** (combination of "weak" classifiers; Breiman's extension of multi-class AdaBoost)

v.) **eXtreme Gradient Boosting** (combination of "weak" classifiers with "a more regularized model formalization to control over-fitting", trees considered as weak classifiers)

The choice is related to:

- the data used for simulation purposes, where both the target variables, Y and Z, are categorical
- availability of SL method in R

1) Train SL methods --> function `train()` in R package **caret**

   naive Bayes --> **klaR** (or **e1071**)

   random Forest --> **randomForest**

   C5.0 --> **C50**

   Adaptive Boosting --> **adaboost**

   eXtreme Gradient Boosting --> **xgboost**

   Training possible also using ad hoc functions in packages

2) Get predictions --> function `predict()` in **caret** or ad-hoc functions in the other packages

Istat's 2011 surveys on households (HHs):

Survey on Income and Living Conditions (IT-SILC, 18487 HHs)

Y = HH income in IT-SILC (7 classes)

Household Budget Survey (HBS, 22933 HHs).

Z = HH overall expenditures in HBS (11 classes)

$\boldsymbol{X}$ is a subset of available common variables, namely 8 variables related to the HH or to the reference person

A rough estimate of the **overall uncertainty** is provided by the Average width of intervals (`Frechet.bounds.cat()` in R package **StatMatch)**:

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

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

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

Results:

Conditional on	Prediction error Y	Prediction error Z	Uncertainty ($\bar{d}$)
Best 2 Xs			0.0612
Best 3 Xs			0.0582
Pred. naïve Bayes	68.7%	85.3%	0.0591
Pred. C5.0	43.0%	56.4%	0.0455
Pred. randomForest	38.6%	51.7%	0.0416
Pred. AdaBoost	35.9%	49.3%	0.0391
Pred. XGBoost	36.3%	49.7%	0.0391

Impute the HH consumption (Z) in SILC (A) through random hotdeck (donor chosen at random in donation class)

- donation classes obtained crossing SL predictions of Y and Z
- implemented in `RANDwNND.hotdeck()` function from **StatMatch**

Donation classes formed by	Prediction error		Preservation of distributions in synthetic dataset[1]		Estimated association $Y \times Z$ (Cramer's V)
	Y	Z	Z	$X \times Z$	
best 2 Xs			0.0359	0.1765	0.1666
best 3 Xs			0.0263	0.1785	0.1684
Pred. naïve Bayes	68.7%	85.3%	0.0267	0.2028	0.1660
Pred. C5.0	43.0%	56.4%	0.0278	0.1907	0.1823
Pred. randomForest	38.6%	51.7%	0.0256	0.1891	0.1830
Pred. AdaBoost	35.9%	49.3%	0.0240	0.1910	0.1882
Pred. XGBoost	36.3%	49.7%	0.0270	0.1898	0.1926

[1]Hellinger distance between distributions: imputed vs. observed in donor

Y and Z predictions provided by Statistical Learning methods

When exploring uncertainty:

- allow to reduce uncertainty (conditional on them), avoiding problem of sparse tables (due to crossing too many Xs)

When used to form donation classes in random hotdeck (two-steps procedure)

- avoid selection of matching variables
- avoid empty donation classes (when crossing too many Xs)
- permit to impute a Z variable with reliable marginal distribution
- provide a slightly better estimation of association between Y and Z

Major drawback in applying Statistical Learning:

- effort required by tuning SL is generally higher than that required by the selection of matching variables

Thank you for your attention