3rd International Conference
New Challenges for Statistical Software - The Use of R in Official Statistics
Bucharest, 23-24 April 2015

INTEGRATION AND IMPUTATION OF SURVEY DATA IN R: THE STATMATCH PACKAGE

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

**Italian National Institute of Statistics,*

Introduction

Increasing demand for new statistical outputs represents a big challenge for National Statistical Institutes

Traditional solution:

set up new surveys or modify existing ones

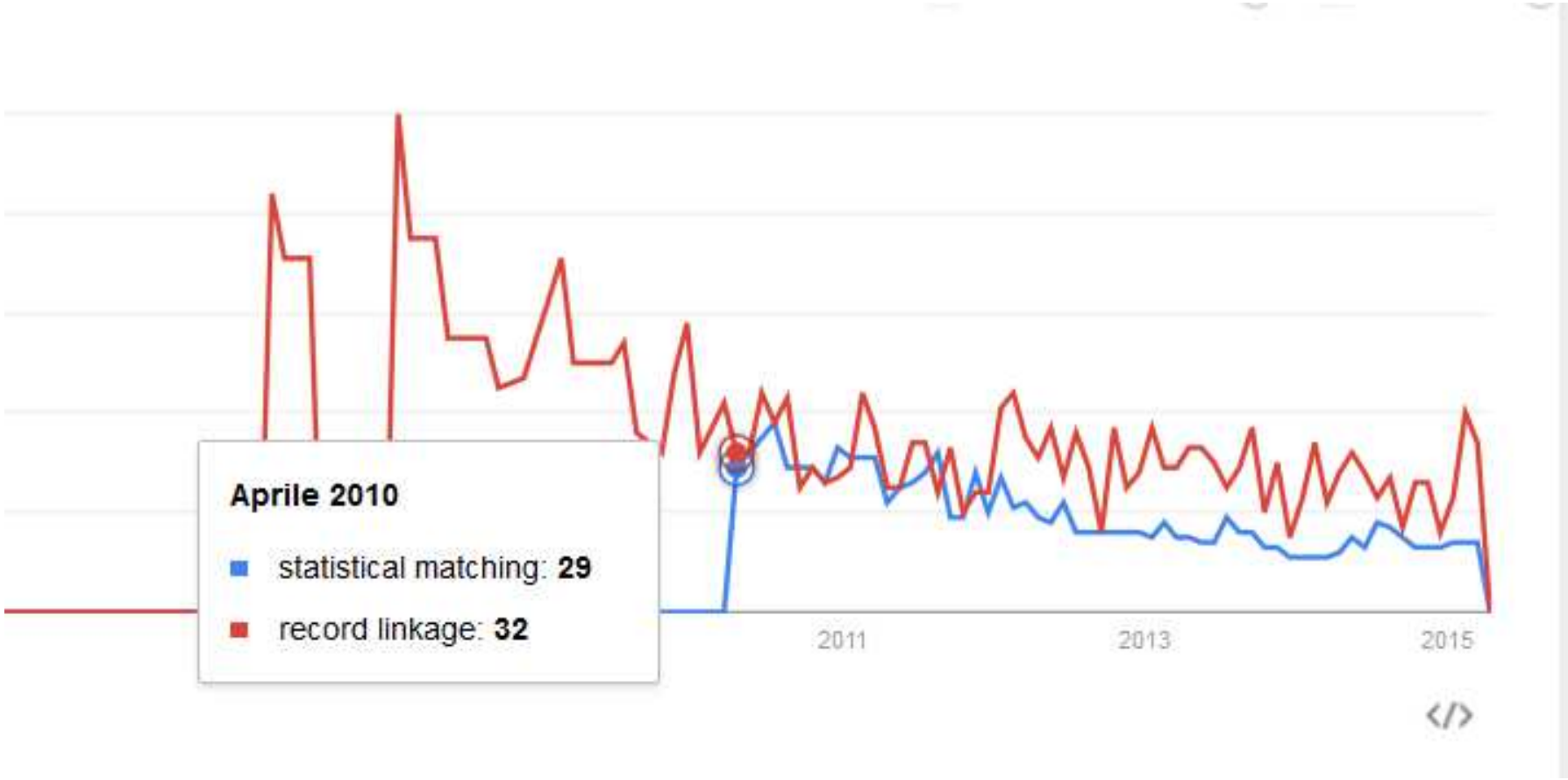
- ↳ often difficult because of budget constraints and risk of increasing the respondents' burden.

Recent strategy:

Exploitation of all the available data sources (surveys, admin. data, etc.)

- ↳ new statistical processes
  - ↳ need of integrating data from different sources
    - **Record linkage**
    - **Statistical matching**

Integration trends



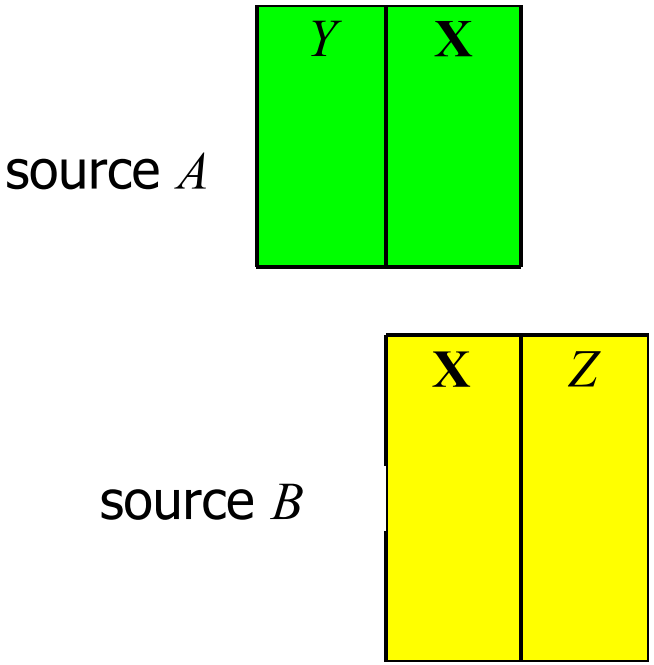
Relative number of searches (max is 100)
Source: Google trends (region USA)

Objectives of Statistical Matching

Statistical Matching (data fusion or synthetic matching):

Series of statistical methods for integrating two data sources (usually samples) referred to the same target population.

Objective: study the relationship among variable not jointly observed in a single data source



X variables in common

Y and Z are NOT jointly observed

Objectives of Statistical Matching

- ✓ **micro**: derive a "synthetic" data set with X, Y and Z; for instance:
  - A filled-in with Z (imputation of Z in A)
  - file concatenation ($A \cup B$) with Z filled in A and Y filled in B

- ✓ **macro**: estimation of parameters; for instance:
  - correlation coefficient (ρ_{YZ})
  - regression coefficient (β_{YZ})
  - a contingency table $Y \times Z$

Statistical Matching at Istat

1) Household Budget Survey (Istat) &
Survey on Household Income and Wealth (Italian Central Bank)

   to build the Social Accounting Matrix (Coli et al., 2006)

2) Labour Force Survey (Istat) & Time Use Survey (Istat)

   to investigate relationship between time spent at work and the type of work (Gazzelloni et al, 2007)

3) Farm Structure Survey (Istat) &
Farm Accountancy Data Network Survey (Istat+Inea)

   To explore relationship between structure of the farms and their economic performances (Ballin et al., 2009)

4) HBS (Istat) & EU-SILC (Istat)

   To investigate relationship between household income and expenditures (Donatiello et al., 2014)

Methods for Statistical Matching

Various methods available, depending on the objective (micro or macro) and on the framework (parametric, nonparametric or mixed).

	Parametric	Nonparametric	Mixed
Macro	- estimation of parameters in the presence of missing values	- estimation of the empirical cumulative distribution - kernel density estimators	
Micro	- conditional mean matching - stochastic regression imputation	- hot deck imputation	- combination of predictive mean matching and hot deck imputation

The R package StatMatch

- appears on CRAN in 2008 (based on codes in D'Orazio *et al.*, 2006)
- major updates in 2011 in the framework of ESSnet "Data Integration"
- latest version 1.2.3 (January 2015)

5 groups of functions:

- **hot deck imputation**: `NND.hotdeck`, `RANDwNND.hotdeck`, `rankNND.hotdeck`
- **mixed SM** at macro or micro level (continuous variables with MVN distr.): `mixed.mtc`
- **SM of data from complex sample surveys** via weights calibrations (Renssen, 1998): `harmonize.x` and `comb.samples`
- **exploration of uncertainty** in estimating the contingency table $Y \times Z$: `Frechet.bounds.cat` and `Fbwidhts.by.x`
- miscellanea: compare marg. distributions; compute distances, etc.

StatMatch: hot deck methods

`NND.hotdeck`

- large set of **distance functions** (Gower's, maximum distance, ...)
- **constrained** selection of donors (donors used just once)

`RANDwNND.hotdeck`

- "**moving" donation classes** created according to different criteria (*k*NN; donors with dist<*k*; *k*% of closest donors)
- **Fast search** of the *k* Approximate Nearest Neighbors (ANN)
- Subset of **donors satisfying user defined equality/inequality constraints**
- **Weighted selection** of donors (weighted random hot deck)

These function can be used to impute missing values in a single survey!!

StatMatch: complex sample survey data

Need to account for:

- sampling design (stratification, clustering…)
- units' weights: design weights corrected to compensate non-observation errors

StatMatch implements some **naïve approaches** (`RADNwNND.hotdeck` and `rank.horteck`) and the **method suggested by Renssen** (1998):

Step 1) **harmonization** of marginal/joint distribution of matching variables
↳ function `harmonize.x`

Step 2) **estimation** of the contingency table $Y \times Z$
↳ function `comb.samples`

Both the functions perform **calibration of the survey weights** in A and B (functions in R package **survey**)

StatMatch: complex sample survey data

The estimated $Y \times Z$ table provides marginal distributions of Y and Z which are coherent with the ones estimates on respectively A and B, after step 1)

`comb.samples`

- **micro imputation** allowed:
  - predicted values of Y in B (pred. prob. with categorical var.)
  - predicted values of Z in A (pred. prob. with categorical var.)

  predictions obtained by fitting **linear probability models**:
  - risk of probs<0 or probs>1
  - marginal distributions of predicted values coherent with the ones in the origin data sources

Uncertainty in Statistical Matching

Uncertainty is due to lack of information: Y and Z which are NOT jointly observed in the basic SM setting

Focus on categorical X, Y and Z variables

objective of SM: estimation of the probabilities

$$p_{\cdot jk} = \Pr\left(Y = j, Z = k\right), \qquad j = 1,\ldots,J,\ k = 1,\ldots,K$$

In this case the uncertainty set can be computed by resorting to the **Fréchet bounds**:

$$\max\left(0, p_{j\cdot} + p_{\cdot k} - 1\right) \leq\ p_{jk} \leq \min\left(\ p_{j\cdot}, p_{\cdot k}\ \right)$$

By conditioning on X:

$$\sum_h p_{h\cdot\cdot} \max\left(0, p_{j|h} + p_{k|h} - 1\right) \leq p_{\cdot jk} \leq \sum_h p_{h\cdot\cdot} \min\left(\ p_{j|h}, p_{k|h}\ \right)$$

StatMatch: uncertainty

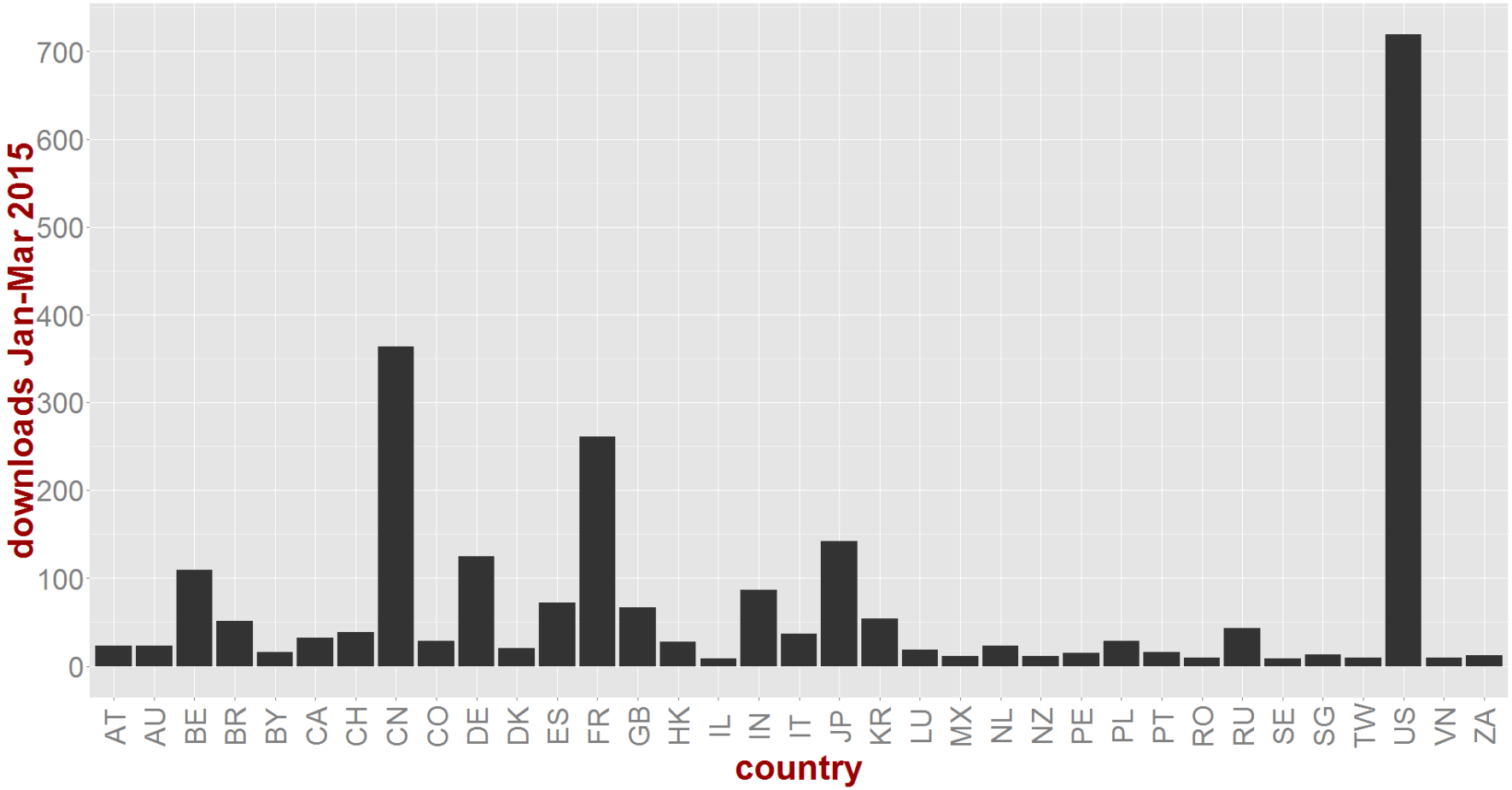
`Frechet.bounds.cat`

- **Frechét bounds** conditioned and unconditioned
- Estimates p_{jk} under the **conditional independence assumption**
- rough measure of uncertainty by considering average width of the cell bounds

`Fbwidths.by.x`

- explores uncertainty by considering **each possible combination of the X variables**
- useful for identifying the X variables more effective in reducing the uncertainty, that therefore can be used as matching variables

StatMatch: downloads



Note: downloads from RStudio server

StatMatch: to do list

- improvements in random hot deck procedure
- improvements of the techniques to match data from complex sample surveys
- automatic procedure for selecting the matching variables based on exploration of uncertainty (D'Orazio *et al*, 2015)
- improvements in the procedure for exploring uncertainty with categorical variables with sparse tables

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

Coli A., Tartamella F., Sacco G., Faiella I., D'Orazio M., Di Zio M., Scanu M., Siciliani I., Colombini S., Masi A. "La costruzione di un Archivio di microdati sulle famiglie italiane ottenuto integrando l'indagine ISTAT sui consumi delle famiglie italiane e l'Indagine Banca d'Italia sui bilanci delle famiglie italiane", *Documenti*, N. 12/2006, Istat.

D'Orazio, M. (2015) "StatMatch: Statistical Matching", R package version 1.2.3 http://CRAN.R-project.org/package=StatMatch

D'Orazio, M. (2015) "Statistical Matching and Imputation of Survey Data with StatMatch", R package vignette, http://cran.rstudio.com/web/packages/StatMatch/vignettes/Statistical_Matching_with_StatMatch.pdf

D'Orazio M., Di Zio M., and Scanu M. (2006) *Statistical Matching, Theory and Practice*. Wiley, New York.

D'Orazio M., Di Zio M., and Scanu M. (2015) "The use of uncertainty to choose the matching variables in statistical matching" NTTS 2015 Conference, Brussels, 10-12 March 2015.

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.

Gazzelloni S., Romano M.C., Corsetti G., Di Zio M., D'Orazio M., Pintaldi F., Scanu M., Torelli N. (2007) "Time Use and Labour Force: a proposal to integrate the data through statistical matching". In: (Romano, M. C. ed.) *Time Use in Daily Life: A Multidisciplinary Approach to the Time Use's Analysis*, National Institute of Statistics (Istat), pp.297-320

Renssen, R.H. (1998) "Use of Statistical Matching Techniques in Calibration Estimation". *Survey Methodology*, **24**, pp. 171–183