How to integrate information from different data files?

An introduction to statistical matching

Eva Endres and Thomas Augustin

Department of Statistics, LMU Munich

April 4, 2018

EMOS Webinar

Table of contents

motivation & basic framework

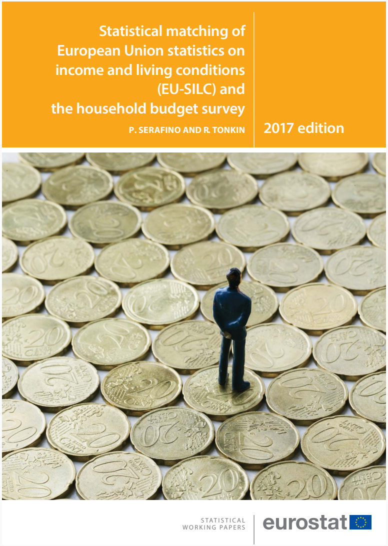
how to obtain data to statistically answer a research question?

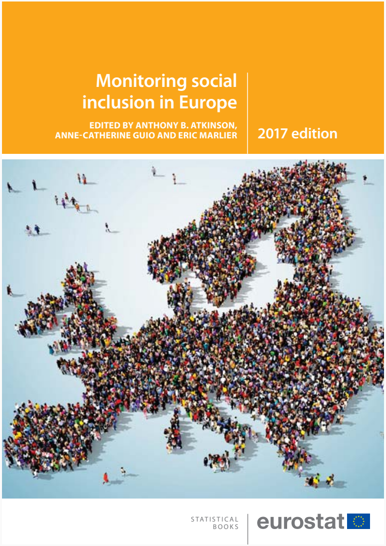
(D'Orazio et al. (2006), Chap. 1)

- carry out surveys or experiments but
    - time-consuming
    - hight cost
    - too long questionnaire might lead to nonresponse or low quality
- practical solution: exploit information from already available data sources (secondary data analysis)
- but: what can we do if we need joint information on features which are only available in different sources?

Eurostat example (simplified)

(Serafino and Tonkin (2017b), Serafino and Tonkin (2017a))





Eurostat example (simplified)

(Serafino and Tonkin (2017b))

- background: measure poverty and social exclusion to monitor the progress of the social inclusion target
- income is not adequate as sole measure of poverty (especially if poverty is interpreted in terms of achieved standards of living)
- the question arises whether expenditure or material deprivation provide more appropriate measures of standards of living than income
- compare people's exposure to poverty using three different measures: income, expenditure and material deprivation
- no single data source provides joint information on all these variables
- statistically match the Household Budget Survey with the EU-SILC for six EU countries

Eurostat example (simplified)
(Serafino and Tonkin (2017b))

material deprivation	income		EU-SILC
	income	expenditure	HBS

$\Downarrow$

material deprivation	income	expenditure	joint information

the statistical matching framework

(D'Orazio et al. (2006))

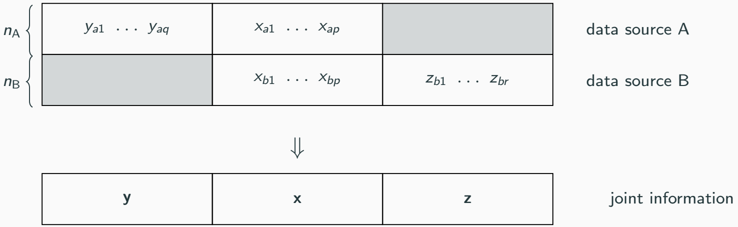
specific **y**	common **x**		data source A
	common **x**	specific **z**	data source B

$\Downarrow$

specific **y**	common **x**	specific **z**	joint information

the statistical matching framework

(D'Orazio et al. (2006))



how to achieve joint information?

(D'Orazio et al. (2006))

objectives of statistical matching:

- *micro approach*: create complete (synthetic) data file
- *macro approach*: estimate the joint distribution

solutions for the statistical matching task either

- are based on the conditional independence assumption (**CIA**),
- incorporate (sufficient) auxiliary information, or
- respect the uncertainty and yield set-valued results

the conditional independence assumption

(conditional) independence of random variables

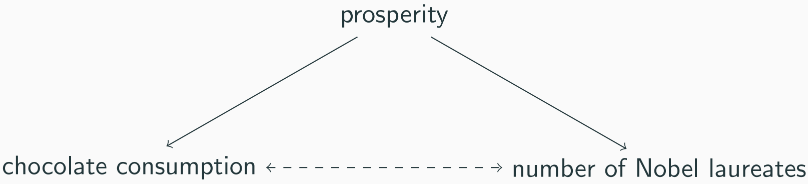
if X, Y and Z are (continuous) random variables

- Y and Z are stochastically independent iff
  $f_{Y,Z}(y,z) = f_Y(y) \cdot f_Z(z) \Leftrightarrow f_{Y|Z}(y|z) = f_Y(y)$
  i.e. knowing the value of Z, does not change my assessment of the distribution of Y
- Y and Z are conditionally independent given X iff
  $f_{Y,Z|X}(y,z|x) = f_{Y|X}(y|x) \cdot f_{Z|X}(z|x) \Leftrightarrow f_{Y|X,Z}(y|x,z) = f_{Y|X}(y|x)$
  i.e. knowing the value of Z, does not change my assessment of the distribution of Y given x is known

(conditional) independence of random variables

(Messerli (2012))

number of different colours in the national flag



statistical matching and the CIA

(D'Orazio et al. (2006))

- assume the conditional independence of Y and Z given X
- yields an identifiable model for (X, Y, Z) on the available data $A \cup B$, the joint density simplifies to

$$\begin{aligned} f_{Y,Z,X}(y,z,x) &= f_{Y|Z,X}(y|z,x) \quad \cdot f_{Z|X}(z|x) \quad \cdot f_X(x) \\ &\overset{CIA}{=} \underbrace{f_{Y|X}(y|x)}_{\text{A}} \quad \underbrace{\cdot f_{Z|X}(z|x)}_{\text{B}} \quad \underbrace{\cdot f_X(x)}_{\text{A and B}} \end{aligned}$$

selected macro & micro approaches

selected macro & micro approaches

a parametric macro approach

a parametric macro approach

(D'Orazio et al. (2006))

- $f(x, y, z; \theta) \in$ *parametric family of distributions* and $\boldsymbol{\theta} \in \Theta$
- aim of the macro approach is the estimation of $\boldsymbol{\theta}_{Y|X}$, $\boldsymbol{\theta}_{Z|X}$, $\boldsymbol{\theta}_X$
- likelihood approach:

$$L(\boldsymbol{\theta}|A \cup B) \overset{iid\&MCAR}{=} \prod_{a=1}^{n_A} f_{XY}(x_a, y_a; \boldsymbol{\theta}_{XY}) \cdot \prod_{b=1}^{n_B} f_{XZ}(x_b, z_b; \boldsymbol{\theta}_{XZ})$$

$$= \underbrace{\prod_{a=1}^{n_A} f_{Y|X}(y_a|x_a; \boldsymbol{\theta}_{Y|X})}_{A} \cdot \underbrace{\prod_{b=1}^{n_B} f_{Z|X}(z_b|x_b; \boldsymbol{\theta}_{Z|X})}_{B} \cdot \underbrace{\mathrm{C}(\mathrm{x})}_{A \text{ and } B}$$

with $\mathrm{C}(\mathrm{x}) = \prod_{a=1}^{n_A} f_X(x_a; \boldsymbol{\theta}_X) \cdot \prod_{b=1}^{n_B} f_X(x_b; \boldsymbol{\theta}_X)$

CIA $\Rightarrow$ sufficient to determine the joint distribution

Table of contents

a parametric macro approach
(D'Orazio et al. (2006))

If $(X, Y, Z) \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and

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

- marginal distribution of the common variable:

$$X \sim N(\mu_X, \sigma_X^2)$$

a parametric macro approach

(D'Orazio et al. (2006), Fahrmeir and Hamerle (1996), Wang (2018))

- conditional distribution of the specific variable(s) given the common variable:

$$Y|X \sim N(\mu_{Y|X}, \sigma^2_{Y|X})$$

  express unknown conditional parameters by known parameters which leads to a 'regression model form' ($Y = \alpha + \beta \cdot X + \epsilon$):

$$\mu_{Y|X} = \alpha + \beta \cdot X$$

$$\alpha = \mu_Y - \beta \cdot \mu_X$$

$$\beta = \frac{\sigma_{XY}}{\sigma^2_X}$$

$$\sigma^2_{Y|X} = \sigma^2_Y - \frac{\sigma^2_{XY}}{\sigma^2_X} = \sigma^2_Y - \beta^2\sigma^2_X$$

- analogously for $Z|X$

selected macro & micro approaches

a nonparametric micro approach

a nonparametric micro approach

(D'Orazio et al. (2006))

hot deck imputation

- no assumption of any parametric family of distributions
- substitute missing entries with *live* values
- assign the roles of *recipient* file and *donor* file

$y_{a1} \ \dots \ y_{aq}$	$x_{a1} \ \dots \ x_{ap}$	$\tilde{z}_{a1} \ \dots \ \tilde{z}_{ar}$	recipient file
	$x_{b1} \ \dots \ x_{bp}$	$z_{b1} \ \dots \ z_{br}$	donor file

a nonparametric micro approach

(D'Orazio et al. (2006))

common hot deck methods in statistical matching:

- random hot deck
- rank hot deck
- distance hot deck
- **distance hot deck**
    - match each recipient record with the closest donor record in terms of a predefined (distance) metric
    - use, for example, the Manhattan distance for (standardised) continuous common variables:

$$\Delta(a, b) = \sum_{\ell=1}^{p} |x_{a\ell} - x_{b\ell}|$$

a nonparametric micro approach

$$\Delta(a, b) = \sum_{\ell=1}^{p} |x_{a\ell} - x_{b\ell}|$$

recipient

a	y	x_1	x_2	$\tilde{z}$
1	27	22	88	202
2	35	19	101	155
3	39	27	93	182

donor

b	y	x_1	x_2	z
1		18	96	155
2		30	92	182
3		22	89	202

$$\Delta(1,1) = |22 - 18| + |88 - 96| = 12$$
$$\Delta(1,2) = |22 - 30| + |88 - 92| = 12$$
$$\Delta(1,3) = |22 - 22| + |88 - 89| = 1$$

$$\Delta(2,1) = |19 - 18| + |101 - 96| = 6$$
$$\Delta(2,2) = |19 - 30| + |101 - 92| = 20$$
$$\Delta(2,3) = |19 - 22| + |101 - 89| = 15$$

$$\Delta(3,1) = |27 - 18| + |93 - 96| = 12$$
$$\Delta(3,2) = |27 - 30| + |93 - 92| = 4$$
$$\Delta(3,3) = |27 - 22| + |93 - 89| = 9$$

a nonparametric micro approach in R

(R Core Team (2017), D'Orazio (2017))

```
# install and load package
install.packages("StatMatch")
library(StatMatch)

# create data files
A <- data.frame(y = c(27,35,39), x1 = c(22,19,27),
      x2 = c(88,101,93))
> A
   y x1  x2
1 27 22  88
2 35 19 101
3 39 27  93

B <- data.frame(z = c(155,182,202), x1 = c(18,30,22),
      x2 = c(96,92,89))
> B
    z x1 x2
1 155 18 96
2 182 30 92
3 202 22 89
```

a nonparametric micro approach in R

(R Core Team (2017), D'Orazio (2017))

```
# detect specific and common variables
common.x <- intersect(names(A), names(B))
common.x
[1] "x1" "x2"

specific.y <- setdiff(names(A), names(B))
specific.y
[1] "y"

specific.z <- setdiff(names(B), names(A))
specific.z
[1] "z"
```

a nonparametric micro approach in R

(R Core Team (2017), D'Orazio (2017))

```
# nearest neighbour using Manhattan distance
matching.ids <- NND.hotdeck(data.rec=A, data.don=B,
                    match.vars=common.x, dist.fun="Manhattan")
> matching.ids
$mtc.ids
     rec.id don.id
[1,] "1"    "3"
[2,] "2"    "1"
[3,] "3"    "2"

$dist.rd
[1] 1 6 4

$noad
[1] 1 1 1

$call
NND.hotdeck(data.rec = A, data.don = B, match.vars = common.x,
    dist.fun = "Manhattan")
```

a nonparametric micro approach in R

(R Core Team (2017), D'Orazio (2017))

```
# create complete synthetic file
synthetic.file <- create.fused(data.rec=A, data.don=B,
                      mtc.ids=matching.ids$mtc.ids, z.vars=specific.z)
> synthetic.file
   y x1  x2   z
1 27 22  88 202
2 35 19 101 155
3 39 27  93 182
```

selected macro & micro approaches

outlook on mixed methods

basic idea of mixed methods

(D'Orazio et al. (2006))

2-step procedure combining parametric and nonparametric methods

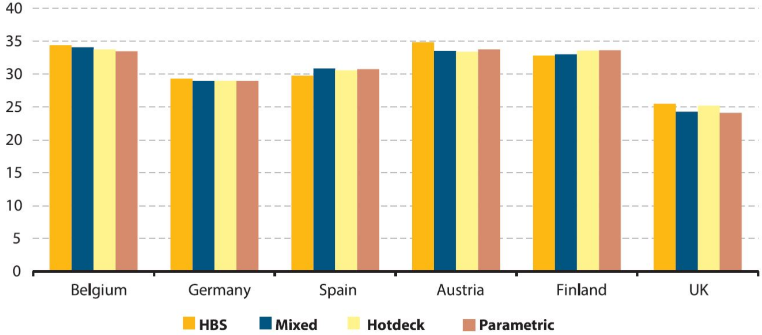
1. estimate parameters for the parametric model and create *intermediate* values
2. apply a hot deck method: choose donor records based on intermediate values and impute correpsonding *live* values

selected results of the Eurostat application

selected results of the Eurostat application i

(figure taken from Serafino and Tonkin (2017b))

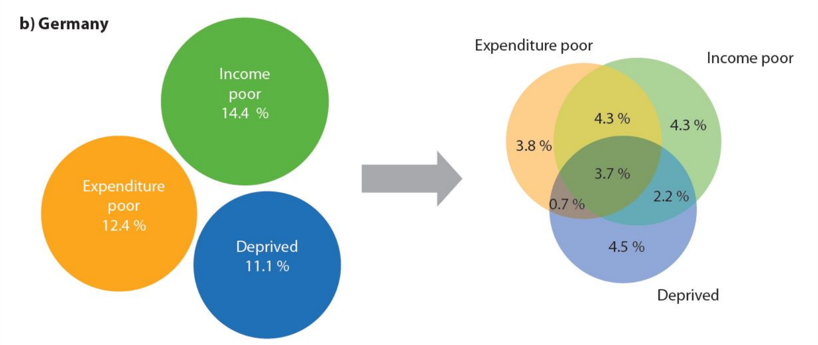
Figure 1: Mean expenditure for HBS and each of the matching methods
(thousands € per annum)



Source: EU-SILC 2009 (Austria), 2010 and 2012 (Finland): EU-SILC Users' database; HBS 2010: Eurostat/ONS.

selected results of the Eurostat application ii

(figure taken from Serafino and Tonkin (2017a))



summary

summary

- fusion of data files with
  - a partially overlapping set of variables
  - disjoint observation units
- parametric and nonparametric approaches to estimate the joint distribution or to create a complete synthetic file
- common assumption: conditional independence of the specific variables given the common variables
- carefully assess whether the assumptions are justified to produce credible results from the matched data files

References

D'Orazio, M. (2017). *StatMatch: Statistical Matching*. R package version 1.2.5.
URL: *https://CRAN.R-project.org/package=StatMatch*

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*, Wiley, Chichester, United Kingdom.

Fahrmeir, L. and Hamerle, A. (1996). Mehrdimensionale Zufallsvariablen und Verteilungen, *in* L. Fahrmeir, A. Hamerle and G. Tutz (eds), *Multivariate statistische Verfahren*, 2 edn, Walter de Gryter, Berlin, pp. 18–48.

Messerli, F. (2012). Chocolate consumption, cognitive function, and nobel laureates, *New England Journal of Medicine* **367**(16): 1562–1564.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Serafino, P. and Tonkin, R. (2017a). Comparing poverty estimates using income, expenditure and material deprivation, *in* A. B. Atkinson, A.-C. Guio and E. Marlier (eds), *Monitoring social inclusion in Europe*, Publications Office of the European Union, Luxembourg, pp. 241–258.
URL: *http://ec.europa.eu/eurostat/documents/3217494/8031566/KS-05-14-075-EN-N.pdf/c3a33007-6cf2-4d86-9b9e-d39fd3e5420c*

Serafino, P. and Tonkin, R. (2017b). Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey. Theme: Population and social conditions, Collection: Statistical working papers.
URL: *http://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/KS-TC-16-026*

Wang, R. (2018). Marginal and conditional distributions of multivariate normal distribution. Accessed on March 19, 2018.
URL: *http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html*