ITSEW 2019
International Total Survey Error Workshop

UNIVERSITÀ DEGLI STUDI DI BERGAMO

Integrating administrative and survey agricultural data with Statistical Matching

D'Alberto Riccardo*, Raggi Meri

*riccardo.dalberto@unibo.it

Dept. of Statistical Sciences "P. Fortunati"

Alma Mater Studiorum
University of Bologna

Introduction - Why integrating data?

Growing amount of data sources, e.g. digitalized administrative data, big data (social media, smart-phones geodata, etc.), and ad hoc project surveys...

VS

Elementary data availability limited due to strict privacy claim constraints and the costs reduction rationale adopted by the Official Statistics (accessibility of only aggregate information, not unequivocal units' ids, etc.).

Several examples from the Official Statistics fostering the integration of different data sources to improve the comprehensiveness and timeliness of statistics...

An ongoing debate!

Introduction - Why integrating **farm** data?

Introduction - Why integrating **farm** data?

1. Relevance of ex ante and ex post impact assessment within the Common Agricultural Policy (CAP), but also relevance of the new, discussed, "result-based" approach to evaluation.

Introduction - Why integrating **farm** data?

1. Relevance of ex ante and ex post impact assessment within the Common Agricultural Policy (CAP), but also relevance of the new, discussed, "result-based" approach to evaluation.
2. Comprehensive and up-to-date data are more and more needed.

Introduction - Why integrating **farm** data?

1. Relevance of ex ante and ex post impact assessment within the Common Agricultural Policy (CAP), but also relevance of the new, discussed, "result-based" approach to evaluation.
2. Comprehensive and up-to-date data are more and more needed.
3. Relevance of behavioural information on farmers.

Introduction - Why integrating **farm** data?

1. Relevance of ex ante and ex post impact assessment within the Common Agricultural Policy (CAP), but also relevance of the new, discussed, "result-based" approach to evaluation.
2. Comprehensive and up-to-date data are more and more needed.
3. Relevance of behavioural information on farmers.
4. Relevance of sample representativeness issues of the most used EU farm data: the Farm **Accountancy** Data Network (FADN).

Introduction - Why integrating **farm** data?

1. Relevance of ex ante and ex post impact assessment within the Common Agricultural Policy (CAP), but also relevance of the new, discussed, "result-based" approach to evaluation.
2. Comprehensive and up-to-date data are more and more needed.
3. Relevance of behavioural information on farmers.
4. Relevance of sample representativeness issues of the most used EU farm data: the Farm **Accountancy** Data Network (FADN).

For example, considering Italian agricultural holdings:

- General Agricultural Census: too much dated.
- Farm Structure Survey: unavailable for research purposes.
- FADN: biggest commercial farms are over-represented.

Introduction - How to integrate data?

Several methods, e.g. Record Linkage, Statistical (Up)Downscaling, Statistical Matching (**SM**).

We focus on the **non-parametric micro SM**, the so-called "hot deck" methods/techniques

Because they offer some relevant pros:

- Units' ids are not needed, sets of units cannot be (neither at least) overlapping.
- Data integration based on observed "live" information.
- Model misspecification bias from parametric SM can be avoided.
- Computational advantages with respect to parametric SM.

Method - State of the art in Statistical Matching

Developed from the first 70s (Okner 1972) and progressively implemented out of a coherent framework up to late 90s (Singh et al. 1993) and, then, swiftly further developed by D'Orazio et al. (2006) and Rässler (2012).

Method - State of the art in Statistical Matching

Developed from the first 70s (Okner 1972) and progressively implemented out of a coherent framework up to late 90s (Singh et al. 1993) and, then, swiftly further developed by D'Orazio et al. (2006) and Rässler (2012).

Parametric SM has been widely investigated...

VS

Non-parametric SM "left" to practitioners and a learn-by-doing approach.

Method - State of the art in Statistical Matching

Developed from the first 70s (Okner 1972) and progressively implemented out of a coherent framework up to late 90s (Singh et al. 1993) and, then, swiftly further developed by D'Orazio et al. (2006) and Rässler (2012).

Parametric SM has been widely investigated...

VS

Non-parametric SM "left" to practitioners and a learn-by-doing approach.

Still, some pending challenges:

- Are there proofs of the commonly accepted "prescriptions" adopted by the hot deck applications?
- Is it possible to further develop the existing hot deck techniques?
- How to assess the integration goodness in the non-parametric framework?

Method - Non-parametric Statistical Matching framework

We have two datasets (a recipient and a donor); $\forall$ unit i and j, with $i = 1, \ldots n_R$, $j = 1, \ldots n_D$, we observe:

Recipient (R) dataset:

- $\mathbf{X}^R = \{X_1, \ldots, X_l, \ldots, X_L\}^R$
- and others, such as: $\mathbf{Z}^R = \{Z_1, \ldots, Z_p, \ldots, Z_P\}^R$

Donor (D) dataset:

$\mathbf{X}^D = \{X_1, \ldots, X_l, \ldots, X_L\}^D$

and others, such as: $\mathbf{K}^D = \{K_1, \ldots, K_m, \ldots, K_M\}^D$

We have four hot deck techniques:

I Nearest Neighbour Distance Hot Deck (**nnd**)
II Constrained Nearest Neighbour Hot Deck (**cnnd**)
III Rank Hot Deck (**rkhd**)
IV Random Hot Deck (**rhd**)

We can select three distance functions:

I Manhattan distance (**mn**)
II Mahalanobis distance (**ms**)
III Exact distance (**et**)

Method - Non-parametric Statistical Matching prescriptions

P.I Being equal the dimensionality ratio between the recipient (R) and the donor (D) datasets, the condition of the variability of the matching variables in R being minor than the variability of the matching variables in D, is always preferable.

P.II If P.I does not hold, the widest dimensionality ratio condition between R and D is always preferable.

P.III The so called "the biggest, the best" condition (i.e. the donor dataset has to be the biggest one) is always preferable.

P.IV The donation classes always benefit the integration goodness.

Method - Nearest Neighbour and Constrained (1/2)

Nearest Neighbour Distance Hot Deck (**nnd**):

$$\delta_{ij^*} = |x_i^R - x_{j^*}^D| = \min_{j=1,\ldots,n_D} |x_i^R - x_j^D|,$$

where δ_{ij} is the absolute minimum value of the difference between the i-th and j-th units (with the j^*-th unit being the donor unit chosen to be matched).

Method - Nearest Neighbour and Constrained (1/2)

Nearest Neighbour Distance Hot Deck (**nnd**):

$$\delta_{ij^*} = |x_i^R - x_{j^*}^D| = \min_{j=1,\dots,n_D} |x_i^R - x_j^D|,$$

where δ_{ij} is the absolute minimum value of the difference between the i-th and j-th units (with the j^*-th unit being the donor unit chosen to be matched).

If we want to exclude an already matched observation from the set of the possible donors, the Constrained Nearest Neighbour Hot deck (**cnnd**) defines the donor pattern as follows:

$$\sum_{i=1}^{n_R} \sum_{j=1}^{n_D} (\delta_{ij} \omega_{ij}),$$

with $\omega_{ij} = 1$ for a matched pair of units, $\omega_{ij} = 0$ otherwise.

Method - Nearest Neighbour and Constrained (2/2)

Approaching the goal of minimization of the donor pattern as if we are in a linear programming framework, if $n_R = n_D$ the following constraints do hold:

$$\sum_{j=1}^{n_D} \omega_{ij} = 1, \tag{1a}$$

$$\sum_{i=1}^{n_R} \omega_{ij} = 1. \tag{1b}$$

If $n_R < n_D$, Eq. 1b changes such that:

$$\sum_{i=1}^{n_R} \omega_{ij} \leq 1.$$

Method - Random Hot Deck

Random Hot Deck (**rhd**) picks at random the donor unit to be matched with the recipient. Considering the (initial) potential set of donor and recipient units' pairs:

$$n_{\mathrm{D}}{}^{n_{\mathrm{R}}}, \tag{2}$$

we can reduce it by means of donation classes; for example, being X_1 and X_2 two variables jointly observed both in R and in D, iff a donation class defined by these variables holds, the set in Eq. 2 is restricted such as:

$$(\mathrm{n}^{\mathrm{D}}_{X_1})^{\mathrm{n}^{\mathrm{R}}_{X_1}} + (\mathrm{n}^{\mathrm{D}}_{X_2})^{\mathrm{n}^{\mathrm{R}}_{X_2}}.$$

Method - Rank Hot Deck

Rank Hot Deck (**rkhd**) first:

$$f_{X^R}(x^R) = \frac{1}{n_R} \sum_{i=1}^{n_R} I\,(x_i \leq x),$$

$$f_{X^D}(x^D) = \frac{1}{n_D} \sum_{j=1}^{n_D} I\,(x_j \leq x),$$

considering R and D, respectively, with I that is the set of the indices of $x_i \leq x$ and of $x_j \leq x$.

Second, each recipient unit is associated with a donor one and a matched units' pair is constituted as follows:

$$|f_{X^R}(x_i^R) - f_{X^D}(x_{j^*}^D)| = \min_{j=1,\ldots,n_D} |f_{X^R}(x_i^R) - f_{X^D}(x_j^D)|.$$

Method - Manhattan and Mahalanobis distance functions

The Manhattan (**mn**) distance computes the distance between the i-th and j-th units such that:

$$\Delta_{ij}^{mn} = \sum_{l=1}^{L} |x_{li} - x_{lj}|,$$

i.e. by means of the sum of the absolute value of the differences between the donor and the recipient units in terms of the values of the chosen matching variables.

The Mahalanobis (**ms**) distance computes the distance between the i-th and j-th units, taking into account the statistical relation among the observed covariates $\mathbf{X}$ such that:

$$\Delta_{ij}^{ms} = \left(\mathbf{X}_i^R - \mathbf{X}_j^D\right)' \Sigma_{\mathbf{X}^R\mathbf{X}^D}^{-1} \left(\mathbf{X}_i^R - \mathbf{X}_j^D\right),$$

where Σ is the covariance matrix of the matching variables $\mathbf{X}$.

Method - Exact distance function

The Exact (**et**) distance function has to be conceived more properly as a semi-metric since, for it, the assumption of the triangle inequality does not hold. It is defined as follows:

$$\Delta_{ij}^{et} = \frac{1}{L} \sum_{l=1}^{L} s_l \, |x_{li} - x_{lj}|,$$

where s_l is a scaling factor for the l-th variable that is equal to 1 for binary variables and equal to $\frac{1}{j_l}$ for continuous and categorical ones (with $j_l = \max_i \{x_{li}\} - \min_i \{x_{li}\}$).

Method - Validation strategy

Parametric SM assesses the integration goodness assuming a statistical relation between $\mathbf{Z}$ and $\mathbf{K}$ by means of the joint distribution function $f(\mathbf{X}, \mathbf{K}, \mathbf{Z})$, such that:

$$f(\mathbf{K}|\mathbf{Z}, \mathbf{X}) \propto f(\mathbf{X}|\mathbf{Z}, \mathbf{K}) f(\mathbf{Z}|\mathbf{K}) f(\mathbf{K}) \propto f(\mathbf{X}|\mathbf{K}) f(\mathbf{Z}|\mathbf{K}) f(\mathbf{K}),$$

where $\mathbf{Z}$ is assumed to be a surrogate for $\mathbf{K}$.

Method - Validation strategy

Parametric SM assesses the integration goodness assuming a statistical relation between $\mathbf{Z}$ and $\mathbf{K}$ by means of the joint distribution function $\mathrm{f}(\mathbf{X}, \mathbf{K}, \mathbf{Z})$, such that:

$$\mathrm{f}(\mathbf{K}|\mathbf{Z}, \mathbf{X}) \propto \mathrm{f}(\mathbf{X}|\mathbf{Z}, \mathbf{K})\mathrm{f}(\mathbf{Z}|\mathbf{K})\mathrm{f}(\mathbf{K}) \propto \mathrm{f}(\mathbf{X}|\mathbf{K})\mathrm{f}(\mathbf{Z}|\mathbf{K})\mathrm{f}(\mathbf{K}),$$

where $\mathbf{Z}$ is assumed to be a surrogate for $\mathbf{K}$.

For the non-parametric SM we propose:

1. The graphical analysis of the distribution of the variables pre-and-post the integration.
2. The graphical analysis of the distribution of the variable W.
3. The MSE of the variable W.
4. The Hellinger index.

Application 1 - Simulation study

In D'Alberto & Raggi (2017) we simulate different scenarios:

- two datasets $\rightarrow$ R (recipient), D (donor)
- two sets of common variables $\rightarrow$ $\mathbf{X} = \{X_1, X_2, X_3\}$ and $\mathbf{K} = \{K_1, K_2\}$, both in R and D
- two different dimensionality ratio conditions $\rightarrow$ 1 to 3 and 1 to 10
- two different conditions of matching variables variability

...And we integrate data both with and without donation classes.

Scenario Nr.	1		2		3		4	
Ratio	1 to 10		1 to 10		1 to 3		1 to 3	
Variability	var(R) > var(D)		var(R) < var(D)		var(R) > var(D)		var(R) < var(D)	
Integration Nr.	1	2	3	4	5	6	7	8
Donation classes	with	without	with	without	with	without	with	without

Table: Simulated scenarios

Application 2 - Real data (1/2)

CAP-IRE 2009

- CAP-IRE (**EU FP7**) project survey
- 300 Emilia-Romagna farms
- stratified by territory, specialization, Single Farm Payment

FADN 2009

- EUROSTAT data source
- 1055 Emilia-Romagna farms
- stratified by territory, specialization, economic size

Application 2 - Real data (2/2)

Main purpose: to have behavioural information on the farmers, socio-demographic characteristics on the farmers/farm households, structural characteristics of the farm, inputs, outputs, etc.

We select:

- the Total Agricultural Area (TAA) of the farm (in hectares) $\rightarrow$ matching variable
- the specialisation (categorical), the altitude (ordinal) and the legal status (ordinal) of the farm $\rightarrow$ variables used to build donation classes
- the Utilised Agricultural Area (UAA) of the farm (in hectares) of the single crops (e.g. cereals, mais, vegetables, permanent, fruit, etc.) $\rightarrow$ variables to be imputed

Results 1 - Simulation study

	Combinations of methods and distance functions									
	nnd.mn	nnd.ms	nnd.et	cnnd.mn	cnnd.ms	cnnd.et	rhd.mn	rhd.ms	rhd.et	rkhd
Overlap	●●●	●●●	●○○	●●○	●●○	●○○	●●○	●●○	●○○	○○○
Outliers control	●●●	●●●	●○○	●●○	●●○	●○○	●○○	●○○	●○○	○○○
Improvement with donation classes	↑	↑	↓	↓	↓	=	=	=	=	=
Legend: ●●● best performance, ○○○ worst performance, ↑ increase in goodness, ↓ decrease in goodness										

Table: Performances of methods and distance functions combinations

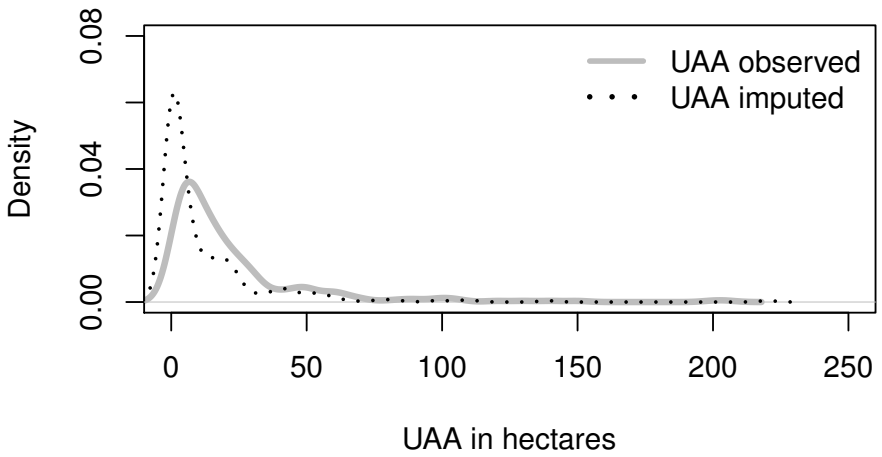
Main findings: all literature "prescriptions" are verified for the non-parametric framework **BUT** "the biggest, the best" one.

Indeed, P.II is not always preferable, i.e. **NOT MANDATORY!**

Results 2 - Real data

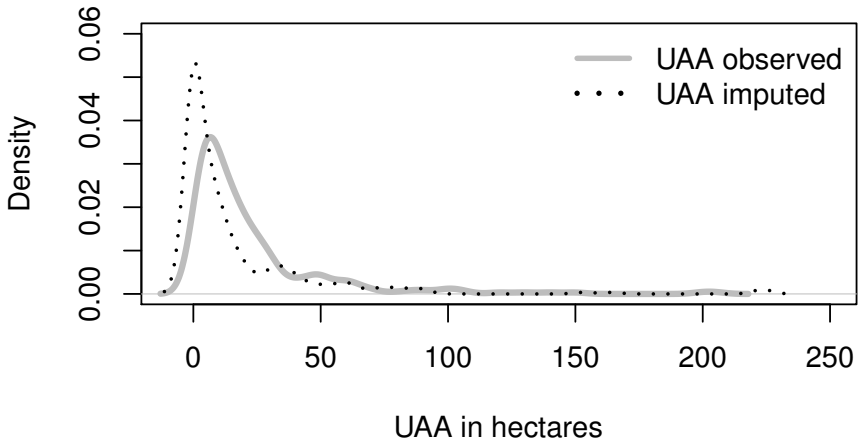
Good and bad combinations:

Figure: Combination nnd.ms



Hellinger index = 0.03728

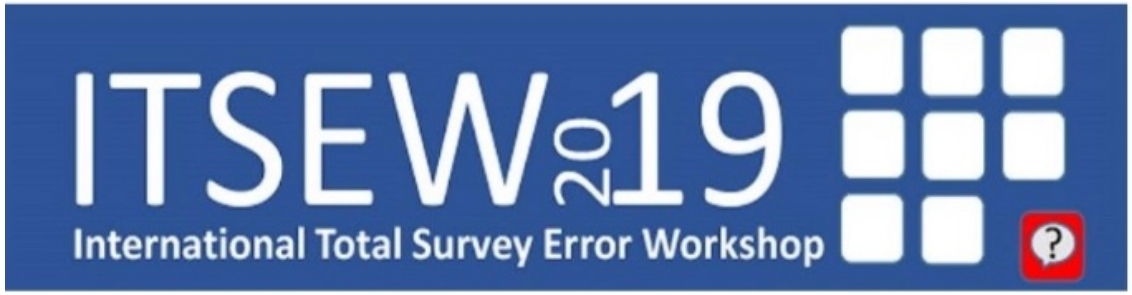
Figure: Combination nnd.et



Hellinger index = 0.05106

Results - Discussion

- Guidelines are useful but dealing with real data is struggling.
- Donation classes help the integration goodness but categorical variables have to be exhaustive.
- Exact distance needs full strata.
- Ex ante (logical) constrains can help the integration?
- What about correlation among matching and imputed variables?
- What about using a project survey (smaller dataset) to integrate official data (bigger dataset)?
- What if samples are not representative of the same target population?



Integrating administrative and survey agricultural data with Statistical Matching

riccardo.dalberto@unibo.it

Appendix 1 - Distance functions

Let be δ a generic distance function iff three properties hold (Mardia et al. 1980):

1. $\delta_{ij} = \delta_{ji}$
2. $\delta_{ij} \geq 0$
3. $\delta_{ii} = 0$

Implying symmetry, non-negativity and identity property.

Let be the h-th observation (with $h = 1, \ldots, n_D$) observed in D, therefore, given the generic distance function δ, we define Δ as a metric iff two assumptions hold:

1. $\Delta_{ij} = 0$, iff $i = j$
2. $\Delta_{ij} \leq \Delta_{ih} + \Delta_{hj}$

Implying the identity of the equals and the triangle inequality.

Appendix 2 - Variable W

The variable W is defined such as:

$$\underset{n_R \times P}{\mathbf{W}} = \underset{n_R \times P}{\mathbf{Z}} - \underset{n_R \times P}{\mathbf{K}^R},$$

for at least the p-th variable, where $\underset{n_R \times P}{\mathbf{Z}}$ and $\underset{n_R \times P}{\mathbf{K}^R}$ are, basically, the "same" variables.

Appendix 3 - Simulation study

- Two datasets $\rightarrow$ R (recipient), D (donor).
- Matching variables $\rightarrow$ $\mathbf{X} = \{\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3\}$ both in R and D, $\mathrm{X}_1^{\mathrm{R}}$ and $\mathrm{X}_1^{\mathrm{D}}$ are simulated as a Bernoulli(θ) with $\theta = 1/2$; $\mathrm{X}_2^{\mathrm{D}}$ is a categorical variable indicating the main variable's value between K_1 and K_2 while $\mathrm{X}_2^{\mathrm{R}}$ is a categorical variable indicating the main variable's value between Z_1 and Z_2; $\mathrm{X}_3^{\mathrm{R}}$ and $\mathrm{X}_3^{\mathrm{D}}$ are simulated as the sum of two log-Normals(μ, σ^2).
- Imputed variables $\rightarrow$ $\mathbf{K} = \{\mathrm{K}_1, \mathrm{K}_2\}$, both in R and D, where they are named $\{\mathrm{Z}_1, \mathrm{Z}_2\}$ are simulated as a log-Normal(μ, σ^2) multiplied by a Bernoulli(θ) with $\theta = 1/2$).
- Two dimensionality ratio conditions $\rightarrow$ 1 to 3 and 1 to 10.
- Two conditions of matching variables variability.