idescat

12, 13 de Novembre del 2012

Seminar on data integration: Data Fusion

Tomàs Aluja-Banet

Josep Daunis i Estadella

LiAM
Laboratory of Information Analysis and Modeling

UPC
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Universitat de Girona

Program

November 12: Data Fusion

9:00-10:30	The problem of Data Fusion. Conditions for Data Fusion.
10:30-10:50	Coffee break
10:50-13:00	Example of application. Steps for Data Fusion
13:00-14:00	Lunch break
14:00-15:30	SPAD-Graft: a toolkit for statistical matching
15:30-15:50	Tea break
15:50-17:15	Computer session with health survey case study

November 13: Data Fusion (cont.)

9:00-10:30	Steps for Data Fusion (cont.)
10:30-10:50	Coffee break
10:50-13:00	Validation tools. Accuracy of the obtained results
13:00-14:00	Lunch break
14:00-15:30	Computer session with ECV_EPA case study
15:30-15:50	Tea break
15:50-17:15	Analysis of matched data files. Discussion on the obtained results

The context of Data Fusion

Context:

- Increasing need of statistical information (new areas of interest, new economy, statistics of consumption, intangibles, …).
- Increasing cost of survey information (survey burden and non response rate).
- Increasing number of (secondary) sources of information with even a faster rate of fragmented information (IT systems).

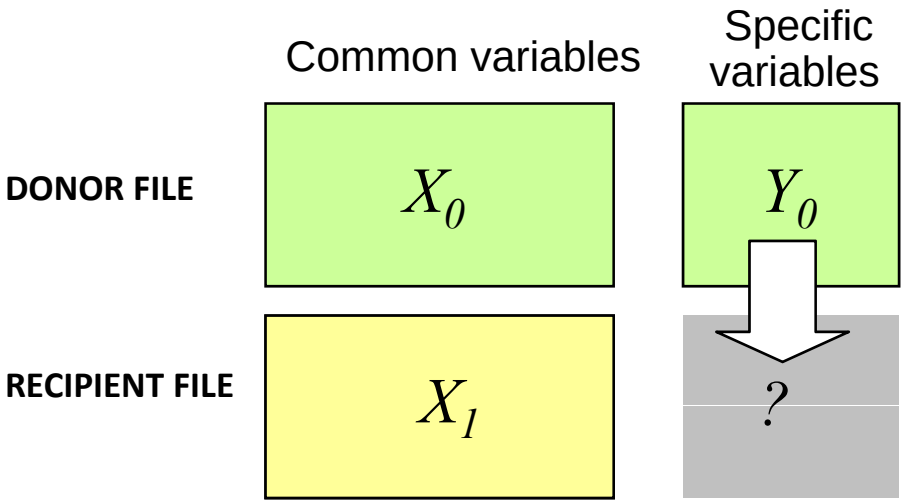
Jacques Antoine (Les Sondages: prospective des techniques et défis méthodologiques, 1999):

"développent de tout ce qui permet d'eviter d'interroger les gens":

- Utilisation des base de données existantes.
- Utilisation des techniques de scanning pour identifier les objets et leur caractéristiques.
- Utilisation d'information auxiliaires pour améliorer les estimations.
- Les techniques de fusion de fichiers permettant simuler les informations manquantes.

What Data Fusion is?

Data fusion is a technological operation whose aim is to integrate the information of two independent data sources.
Technically it involves the imputation of a complete block of missing variables.



Notation:

Donor file $(X_0,\ Y_0)$

Recipient file (X_1)

Common, hinge, bridge variables: x

Specific, fusing, imputing variables: y

Y_1 is a block of missing variables by design

The objective of the data fusion is to transfer the specific variables of the donor file to the recipient file at individual (micro) level

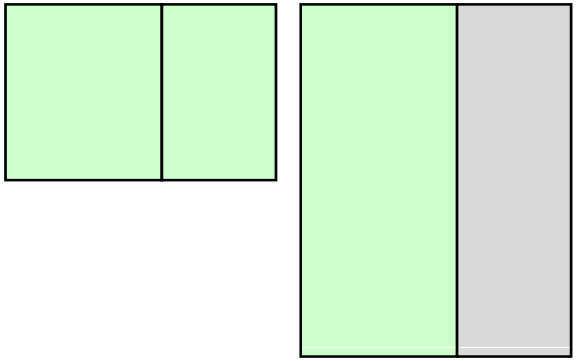
Principal Data Fusion applications

1. In official statistics
   - To integrate information coming from two sources:

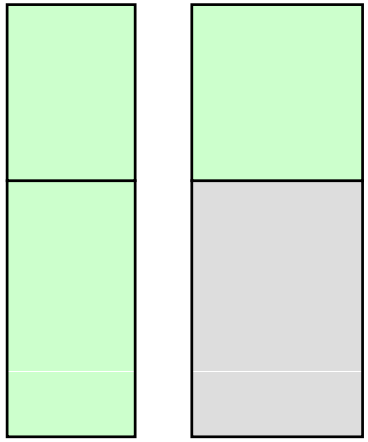
     Fusion of survey data with registries (to give more value to the existing registries).
2. In media surveys
   - To impute the TV audience in a consumption panel
   - To impute the web usage to a representative survey
3. In Data Mining to enrich existing Databases

Data Fusion is mainly an ad-hoc problem undertaken for specific purposes.

Enrichment of a data base from a puctual survey



Self-administered witth re-interview



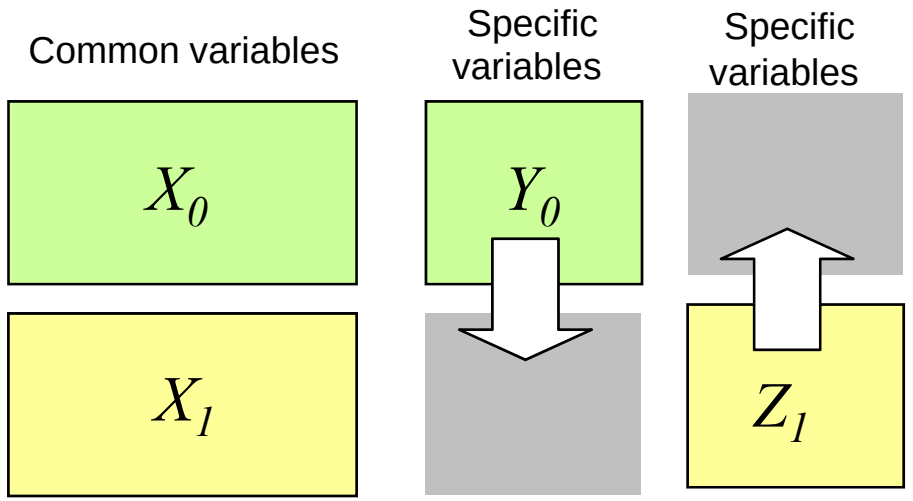
Data Fusion and Statistical Matching

Data Fusion is similar Statistical Matching, but …

Statistical Matching is mainly used by Academics. It refers to the bilateral fusion, where the objective is to impute jointly both missing blocks in both files: Z_0 and Y_1, assuming donors and recipients, representative samples of the same population and assuming some knowledge from the correlation between $(Y,Z|X)$ (default is CIA: conditional independence)

Observations in both files are iid. random draws of the same distribution

$$f(X_0, Y_0, Z_0) = g(X_1, Y_1, Z_1)$$



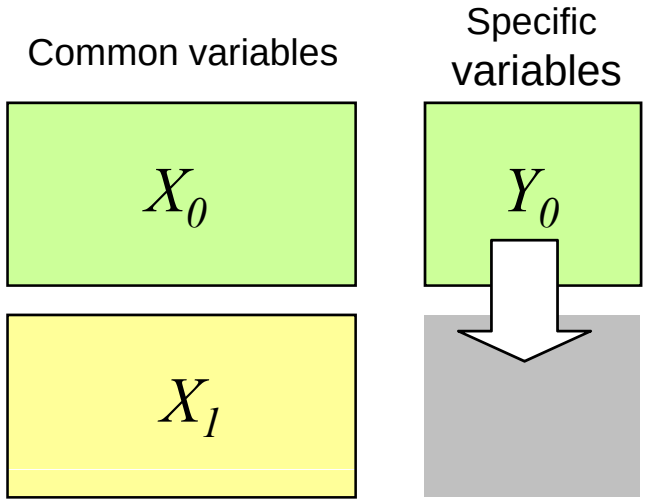
Data Fusion and Statistical Matching

Data Fusion is mainly used in entrepreneurial applications. It refers to the unilateral fusion, where the objective is to impute the block Y_1, assuming donors being a representative sample of the population and the same conditional distribution in both files.

Donor file is (must be) a representative sample of the population respect to the recipients. Recipient sample can be taken with a different design

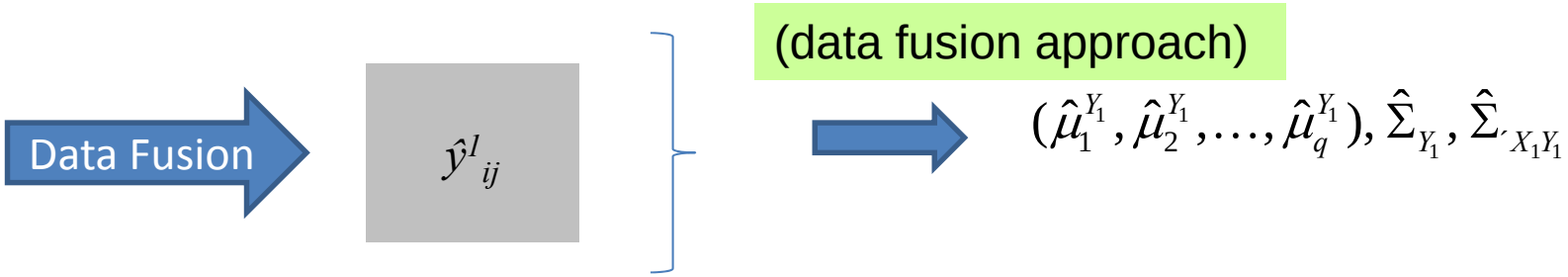
$$f(X_0, Y_0) \neq g(X_1, Y_1)$$

$$f(Y_0 / X_0) = g(Y_1 / X_1)$$

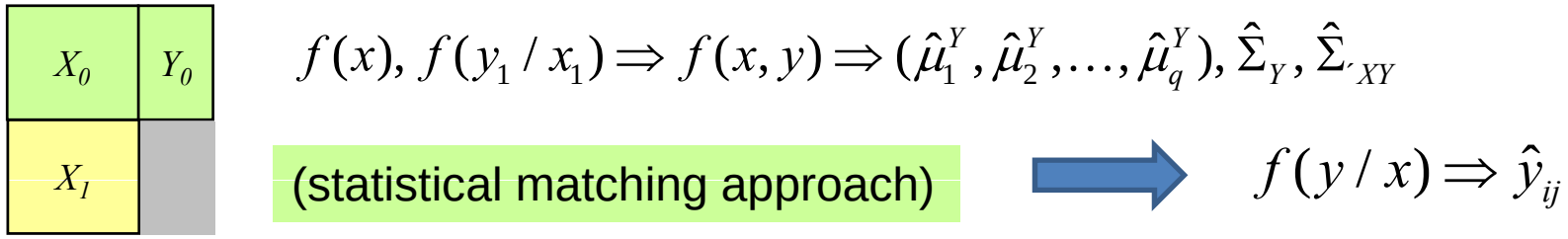


Micro and Macro approach

The **micro approach** means, first to impute (fill in, estimate) the specific variables at individual level, then to compute the aggregate statistics (means, proportions, covariances, …) of the considered populations (donors and recipients).



The **macro approach**, where the emphasis is in the aggregate statistics, means first to compute the parameters of the considered population (assuming a given multivariate distribution, i.e. multinormal, multinomial), and then estimate the individual data from the corresponding conditional distribution.



What we want in a Data Fusion operation?

Data Fusion objectives:

1. **Minimizing the prediction error** $\mathbf{Min}\ E[y_1-\hat{y}_1]^2$

   The imputed values should be as close as possible of the true unknown values.

2. **Reproduction of the conditional distribution of donors in the recipients:**

   $f(\hat{Y}_1/X_1) \approx f(Y_0/X_0)$

   Simulation of real data.

   Both objectives can't be jointly optimized.

If both samples are representative of the same population:

3. **Reproduction of the multivariate distribution of specific variables**

   $f(\hat{Y}_1) \approx f(Y_0)$.

   (but in practice we content to reproduce the values of the marginal statistics and the second order moments of the donor set in the recipient file)

More objectives of Data Fusion

3. Individual Coherence

To prevent nonsense values. The imputed values should be realistic for the whole set of Y_I variables.

4. Avoiding the imputation bias *(black hole* effect of ***knn*** imputation)

- $E[Y_I] = E[\hat{Y}_I]$
- Reproducing the donor variability

  Not repeating the same donor, taking more donors than recipients, introducing variability by sequential imputation, …

5. Measuring the uncertainty of fused data

- Imputed data is not observed data. It is predicted data, so it is uncertain.

6. Obtaining valid inferences

- We want to use the imputed data to infer results about the population

Steps in Data Fusion

1. Preprocess of data:
   Stacking donors and recipients in one data file, cleaning of data, representativeness of donors, significant differences between donors and recipients, ….
2. Selection of the hinge:
   Predictive power of common variables, selection of the best subset of common variables (hinge variables) to predict the specific ones.
3. Visual display step of donors and recipients
4. Imputation step:
   Selection of the imputation model (model?, stochastic/deterministic, single/multiple).
5. Validation step:
   Selection of the validation procedure of the imputation

Technical aspects of Data Fusion

- Selection of the hinge variables among the common ones
- Selection of the imputation methodology
- Validation of the imputed data

1. *Which and how many common variables we need to select?*
2. *What assumptions can reasonably be made about such issues such as the "conditional independence" or the "prediction relevance" of the variables being used?*
3. *What algorithm (technique) should be used?*
4. *Which tests of validation are appropriate?*

Data Fusion assumptions

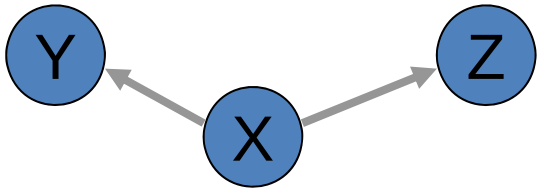
1.- The Predictive Relevance Assumption (PRA):

The Y variables are fully explained by the X variables.

$$Y = i(X) + \varepsilon$$

where ***i(X)*** *represents the* ***chosen*** *imputation model and* ε ***represents white noise.***

The PRA assumption implies the Conditional Independence Assumption:
The Y *variables are independent of any other set of variables* Z *given the* X*.*

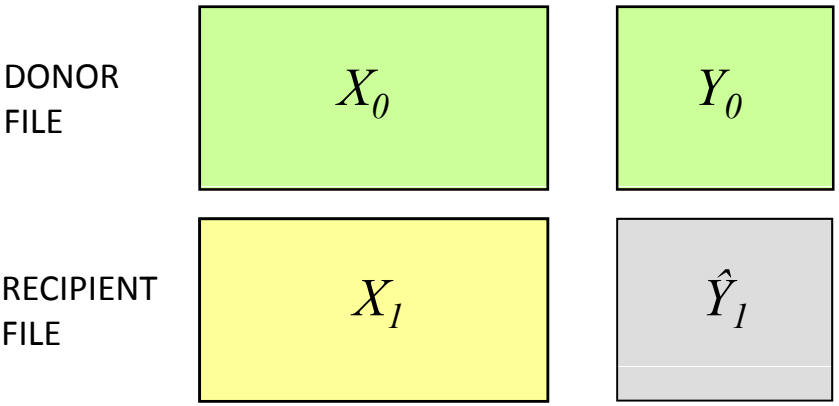


$$f(Y, Z / X) = f_{Y/X}(Y / X) f_{Z/X}(Z / X)$$

Data Fusion assumptions

2.- Preservation of the conditional distribution f(Y/X):

We assume to have the same conditional distribution in both files $f(Y_1/X_1)=f(Y_0/X_0)$, even though the joint distributions may differ (Lebart, Lejeune, 1995). The objective of data fusion is to transfer the specific variables of the donor file to the recipient file at individual level based on the $f(Y_0/X_0)$.



$$f(X_0, Y_0 / \theta_{f(X,Y)}) = f(Y / X, \theta_{Y/X}) f(X_0 / \theta_{f(X)})$$

$$g(X_1, Y_1 / \theta_{g(X,Y)}) = f(Y / X, \theta_{Y/X}) g(X_1 / \theta_{g(X)})$$

Data Fusion assumptions

3.- Representativeness of the donor file:

We don't assume that both files are random samples of the same parent population $f(X_0, Y_0) \neq g(X_1, Y_1)$. *We just assume that the donor file is a representative sample of the population (Santini 1988, van der Putten et al, 2002).*

This assumption need to be tested and assured by the usual way of weighing.

Additionally it is necessary to test whether both files stem from the same population respect the common information $f(X_0) = g(X_1)$.

Imputation model types

- ***Conditional or unconditional***
- ***Deterministic or stochastic***
- ***Parametric or non parametric***
- ***Single or Multiple***

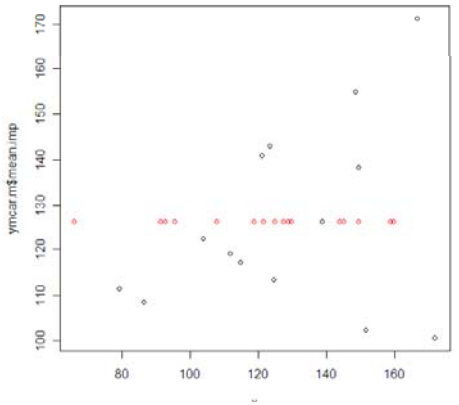
Conditional versus Unconditional imputation

Unconditional Mean imputation

Substitute every missing value for the mean of the corresponding variable (i.e. the mode for the categorical case)

- Preserves the mean
- Reduces the variances and covariances
- Destroys the distribution
- Distorts the correlation with other variables

$$y_{ij,miss} \leftarrow \bar{y}_j$$

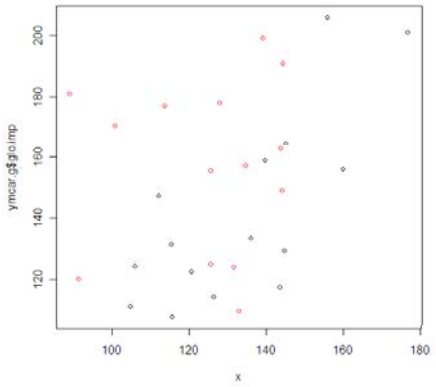


Random draw from the unconditional (marginal) distribution

Substitute every missing value for a random draw of the global distribution

- Preserves the mean,
- Preserves the variance
- Preserves the distribution of the variable
- Destroys the correlation with other variables

$$y_{ij,miss} \leftarrow f\left(y|\theta_Y\right)$$



Imputation should always be *conditional*

Deterministic versus Stochastic imputation

Deterministic or Stochastic

Imputed values $\hat{y}^1_{ij}$ can be obtained by a deterministic or stochastic mecahnism

We can impute a deterministic value (i.e. a conditional mean) or
We can impute a random draw from a distribution

Deterministic $i(X) \leftarrow E(Y / X)$

$$\hat{y}^1_{ij} \leftarrow E(y^0_{ij} \left| x^0_i, \theta_{Y/X} \right.)$$

Minimizes the prediction error $Min\ E\left[Y_1 - \hat{Y}_1 \right]^2$

Stochastic $i(X) \leftarrow f(Y / X)$

$$\hat{y}^1_{ij} \leftarrow f(y^0_{ij} \left| x^0_i, \theta_{Y/X} \right.)$$

Simulates a real instance of $f(x_1,y_1)$

$$f(\hat{Y}_1 / X_1) \approx f(Y_0 / X_0)$$

References

Aluja-Banet, T., Daunis-i-Estadella, J., Pellicer, D., 2007. GRAFT a complete system for data fusion. *Journal of Computational Statistics and Data Analysis*, 52, 2, 635-649.

Aluja-Banet, T., Daunis-i-Estadella, J., Chen Y. H., 2012. Enriching a Large Scale Survey from a Representative Sample by Data Fusion: Models and Validation. In *Survey Data Collection and Integration*. Cristina Davino, Luigi Fabbris (eds). Springer.

Antoine, J., 1999. Les sondages: prospective des techniques et défis méthodologiques, in *Enquêtes et sond*ages, Gildas Brossier, Anne-Marie Dussaix (eds). Dunod.

Conti, P.L. and D. Marella and M. Scanu, Nonparametric evaluation of matching noise., Compstat 2006. Proceedings in *Computational Statistics*, Physica-Verlag HD., 2006, 453-460

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

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*. 39, 1-38.

Lebart, L., Lejeune, M.,1995. Assessment of data fusions and injections., *Encuentro Internacional AIMC sobre Investigación de Medios*. 208-225

Little, R.J.A., Rubin, D.B., 2002. *Statistical Analysis With Missing Data*, John Wiley & Sons,New York.

Paass, G., 1985, Statistical record linkage methodology, state of the art and future prospects. *Bulletin of the International Statistical Institute, Proceedings of 45th Session*, LI, Vol. 2.

Rassler, S., 2004, Data Fusion: Identification problems, Validity and Multiple Imputation., *Austrian Journal of Statistics*., 33,1&2,153-171.

Rius, R., Aluja, T., Nonell, R.,1999, File grafting in market research., *Applied stochastic models in business and industry*, 15, 4, 451-460, 1524-1904

Rubin, D.B., 1976, Inference and missing data*. Biometrika*, 63, 467-474.

Rubin, D.B., 1987, *Multiple Imputation for Nonresponse in Surveys*, Wiley & Sons, New York

Scanu, M. and P.L. Conti, A comparison among different estimators of regression parameters on statistically matched files through an extensive simulation study., *Rivista di Statistica Ufficiale*, 2006, 1, 43-55, Istituto Nazionale di Statistica

Schafer, J.L., *Analysis of Incomplete Multivariate Data*, Chapman & Hall, 1997, London.

Tanner, M.A. and W.H. Wong, 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 398, 528-550