# Assessing the uncertainty in $k$nn Data Fusion

Tomàs Aluja-Banet*, Josep Daunis-i-Estadella**, Enric Ripoll***

*Universitat Politècnica de Catalunya, Campus Nord, C5-204, E-08034 Barcelona
tomas.aluja@upc.edu
**Universitat de Girona, Campus de Montilivi, Edifici P4, E-17071 Girona
josep.daunis@udg.edu
***Institut d'Estadística de Catalunya, Via Laietana 58, E-08003 Barcelona
eripoll@idescat.net

Data fusion, also known as statistical matching, is a technological operation whose aim is to integrate the information of two independent data sources. Let $(X_0; Y_0)$ the *donor* file and $(X_1)$ the *recipient* file, where the $X$ are the *common* variables and the $Y$ are the *specific* ones. The goal is to complete the recipient file $(X_1, \hat{Y}_1)$ in such a way that it can a be a realization of the joint density function $f(X, Y)$.

There are three basic approaches for data fusion. The first one consists of embedding the common and specific variables within a *parametric* multivariate distribution $f(X, Y|\theta)$, assuming donors and receptors independently and randomly draw from this distribution. This distribution can be factored into $f(X, Y|\theta) = f(Y|X, \theta_{Y|X})f(X, \theta_X)$; hence, it is possible to estimate its parameters $\theta_X$ and $\theta_{Y|X}$ from the available information and use them to impute the missing block of data. The second approach consists of directly *modelling* the relationship between the $Y$ variables and the $X$ variables in the donor file by means a regression function: $E(Y|X) = r(X) + \varepsilon$ and applying this model in the *recipient* file (*explicit modelling*). The last approach consists of finding for each individual of the recipient file one or more donor individuals as similar as possible, and then in some way, transferring the values of $Y$ variables to the *recipient* individual (*implicit modelling*). This method is known as *hot deck*, a term borrowed from data editing in data bases.

**Validity of the imputation**

We will say that a data fusion is valid if the fused data set $(X_1, \hat{Y}_1)$ is an instance of the distribution function $f(X, Y)$. In general the distribution function $f$ is unknown, thus we are compelled to compare the empirical distribution functions $ef(X_1, \hat{Y}_1)$ with the $ef(X_1, Y_1)$.

We call the discrepancy between both distributions *matching noise*; following Paass (1985) the matching noise depends on the rightness of the imputation function $i(X)$ to approximate instances of the joint true distribution, which in the parametric case it depends on how well the assumed multivariate distribution represents the true data, and in the hot deck methodology depends on, as before, the assumed imputation model $i(X)$ and in addition to the existing discrepancies between the recipients and their corresponding donors.

However, whatever the imputation method chosen, imputed data is not like observed data, since it has inherent uncertainty, this is **the uncertainty problem**. Imputed values $\hat{Y}_1$ are estimates, thus, to be realistic, we need to take into account the variability of the imputed data when analyzing it. This variability comes from the random fluctuation of the distribution

$f(Y|X, \theta_{Y|X})$ and also from the fact that model parameters $\theta_{Y|X}$ are unknown and consequently they also convey random fluctuation. Multiple Imputation is the classical way to cope with this problem (Rubin (2004)). It consists of repeating several times the single imputation procedure, from the predictive distribution of $f(Y|X, \theta_{Y|X})$ under realistic conditions of the parameters $\theta_{Y|X}$ and then just concatenate the several single imputation files.

**Suite of validation statistics** (Aluja-Banet et al., 2007)

*ASLm*: comparison of marginal means in $\hat{Y}_1$ and $Y_1$.

*ASLs*: comparison of marginal variances in $\hat{Y}_1$ and $Y_1$.

*ACDi*: comparison of the pairwise correlations among the specific variables in $\hat{Y}_1$ and $Y_1$.

*ACDe*: comparison of the pairwise correlations between the specific variables and the common ones in $(X_1, \hat{Y}_1)$ and $(X_1, Y_1)$.

*wc*: reproduction of the eigenstructure of $Y_1$ in $\hat{Y}_1$.

*ASD*: computation of the Smirnov distances between the empirical distributions of the specific variables $Y_1$ and $\hat{Y}_1$.

$\tau$: Computation of the individual generalization error.

**Application to an official survey data on safety and victimization in Catalonia**

We have taken the data collected in 2006 Idescat survey to perform a data fusion operation of some selected variables on the 2007 survey and compare them with the actual values collected in 2007. We have proceed to extract 400 bootstrap resamples from each file to assess the validity of the results.

**Main results**

We present the mean value and the 95% interval of the different validation statistics of the nearest neighbor as baseline method, the usual DA-MI and the $k$nn-MI proposed method.

| | ASLm | ASLs | ACDi | ACDe | wc | ASD | $\tau$ |
|---|---|---|---|---|---|---|---|
| 1nn | 0.021 | 0.027 | 0.132 | 0.040 | 0.741 | 0.100 | 1.934 |
| | 0.000 0.100 | 0.000 0.112 | 0.085 0.209 | 0.035 0.046 | 0.190 0.956 | 0.065 0.137 | 1.751 2.195 |
| DA-MI | 0.055 | 0.065 | 0.051 | 0.031 | 0.950 | 0.267 | 1.894 |
| | 0.001 0.151 | 0.004 0.138 | 0.043 0.059 | 0.027 0.036 | 0.931 0.969 | 0.261 0.274 | 1.865 1.923 |
| $k$nn-MI | 0.031 | 0.054 | 0.068 | 0.038 | 0.915 | 0.065 | 1.935 |
| | 0.000 0.111 | 0.000 0.167 | 0.043 0.100 | 0.035 0.042 | 0.708 0.986 | 0.045 0.086 | 1.841 2.048 |

TAB. 1 – *Mean validation statistics*

The $k$nn multiple imputation clearly improves the results obtained by the single imputation, but it stands below the performances of the parametric multiple imputation, except for the matching noise, where the $k$nn method assures realistic imputations.

## References

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

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 2.*

Rubin, D. (2004). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley & Sons.