Statistical learning in official statistics: the case of statistical matching

Keywords: Data Integration, Machine Learning.

1. INTRODUCTION

National Statistical offices are facing the challenge of modernizing their statistical production processes, beyond traditional sample surveys and censuses, so as to exploit all available data provided by administrative registers and big data. Taking advantage of large data sources requires adoption of modern statistical methods, as those based on *machine learning*. In addition, availability of different data sources on the same phenomena poses the challenge of integrating them for producing a wider set of statistical outputs so as to satisfy users' request. This work will show how statistical learning methods can be beneficial in integrating data.

Statistical learning (SL) is an area of statistics relatively recent (see e.g. [1] and [2]) that includes a wide set of techniques that "learn from the data". They have become very popular in marketing, finance, and other domains, because allow analysis of large data sources, with many variables and observations. Under SL umbrella falls many recent methods related to classification, regression and clustering (generalized additive models, classification and regression trees, neural networks, etc.).

Integration is the core of new statistical production processes aimed at providing a richer set of statistical outputs by taking advantage of already existing data, avoiding setting up new surveys. Focus here is on *statistical matching* (SM, also known as *data fusion*) whose objective is integration of data sources (mainly from sample surveys), lacking of units' identifiers, to investigate relationship between variables not jointly observed in the same survey (see e.g. [3]). These methods are frequently applied to integrate the survey on household income with the one on expenditures to get a thorough picture of people well-being [4]. SM methods include a variety of well-known methods developed to impute missing values in a dataset (predictive mean matching, hotdeck imputation, etc.), but adapted to the specific SM setting.

2. SUPERVISED STATISTICAL LEARNING IN STATISTICAL MATCHING

The typical SM setting consists of two independent data sources (sample survey data), *A* and *B*, sharing a high number of variables, **X**; the *target* variables, *Y* and *Z*, are not jointly observed, i.e. *Y* is available only in *A* and *Z* only in *B*; investigating relationship between *Y* and *Z* is the final goal. SM ends up with an estimate of the interest parameters (correlation/regression coefficients, etc.) or with "fused" data source (*synthetic* source) including all the interest variables. Integration is based on the common information, typically a suitable subset of the shared **X**, called *matching variables*. An overview of main SM methods is in [3]; Authors warn about acritical application of SM, because integration based on *X*s implicitly assumes independence between *Y* and *Z* conditional on the *X*s themselves. This latter assumption is seldom valid, unless one of the *X*s is a proxy of the targets, i.e. highly associated/correlated with it [4].

Application of SL in SM seems limited to widespread hotdeck techniques that, in practice, can be viewed as implementations of the *k*-NN (*k* Nearest Neighbors) approach. This paper goes a step further by suggesting adoption of other modern classification or regression techniques. Two different ideas will be investigated: (i) creation of fused dataset (defined as "micro" in [3]); (ii) assessment of *uncertainty* due to the partial identification of interest parameters ("macro" in [3]).

2.1. Creating the synthetic data set

The procedure presented in this Section aims at creating the fused data set by setting one dataset as *recipient*, say *A*, and filling in it with values selected (donated) from *B* (the *donor* file). Generally speaking, the procedure consists in:

Step 1) using a SL supervised method to build a prediction model of *Z* on *B*;
Step 2) set *A* as recipient and predict *Z* in it by applying the model fitted in step (1).

Different SL procedures are considered as candidates for step (1), ranging from "simple" naïve Bayes classifiers to more complex techniques based on fitting classification/regression trees to end with boosting.

2.2. Investigating uncertainty in statistical matching

Absence of proxies or additional sources of data providing insights about relationship between *Y* and *Z* usually results in poor SM results, since rarely conditional independence holds true. In this case, an alternative approach consists in viewing the problem as one of *partial identification* of the target parameters (correlation coefficient ρ_{yz}, cell probabilities $p_{y=j,z=k}$, etc.), the goal is the estimation of the *partial identification regions*, i.e. all the set of equally possible estimates of the interest parameters given the available data [5].

In case of categorical variables, following Frechet property, the set of equally possible values of $p_{y=j,z=k}$ is:

$$\left[max\left(p_{y=j} + p_{z=k} - 1\right), min\left(p_{y=j}; p_{z=k}\right)\right]$$

being $\sum_{j,k} p_{y=j,z=k} = 1 \;\; (j = 1, \dots, J; k = 1, \dots, K)$.

[6] and [7] show that uncertainty regions can be reduced when conditioning on predictor *Xs* of both *Y* and *Z*. In particular, shorter intervals are achieved by considering expectations of conditional bounds (see [6] or [7] for details). In general, uncertainty decreases by increasing the number of "powerful" predictors of both *Y* and *Z*; unfortunately, by adding variables to the set of predictors, increases the number of conditional probabilities to estimate and, even in very large datasets, contingency tables may become rapidly sparse, thus making difficult estimation. [8] suggests using a sequential procedure for selecting best subset of predictors taking into account the sparseness problem. In this work, to overcome the problem of selecting the proper set of *X*s and estimating the conditional probabilities in large sparse tables, it is introduced a two-step procedure:

Step 1) use a SL supervised method to build:
  1.a) a prediction model of *Y* on *A*;
  1.b) a prediction model of *Z* on *B*

Step 2) estimate expected values of conditional bounds using predictions of both *Y* and *Z*.

3. APPLICATION OF MATCHING VIA STATISTICAL LEARNING

3.1. The data

The SM procedures presented in Section 2. are investigated by applying them to:

(d1) subset from 7th round of European Social Survey (ESS, see [9]);
(d2) subset of Istat's 2011 Household Budget Survey (HBS) and Survey on Income and Living Conditions (IT-SILC).

The (d1) dataset consists of $n = 28{,}769$ individuals (after discarding missing values, "Don't know" and refusals) and a subset of 13 socio-demographic variables. For SM purposes, a series a simulations is run; in each iteration: (a) the whole data set is randomly split in two subsamples, 1/3 of observations in *A* and the remaining ones in *B*; (b) the variable related to income (deciles) plays the role of *Y* and is removed from *B*, while "living comfortably on present income" is the *Z* variable (dichotomized 1="Yes", 2="No") and is removed from *A*; (c) the procedures introduced in Section 2. are applied to *A* and *B*.

Data (d2) are those from surveys are referred to year 2011; in particular IT-SILC sample consists of 18,487 responding households (HHs) while HBS provides data on 22,933 responding HHs. The HH income (categorized in 7 classes) is *Y* target variable in IT-SILC; HH overall expenditures (categorized in 14 classes) plays the role of *Z* in HBS. The two surveys share a high number of *X* variables; the ones considered in the present application are just 8 referred to both the HH and its reference person. Procedure in 2.2. is applied considering IT-SILC as recipient (*A*) while HBS is the donor (*B*).

3.2. Main results

When the goal of the procedure is the creation of the fused dataset, the procedure introduced in Section 2.1. is applied and its results are compared with the corresponding "standard" hotdeck procedures (distance or random hotdeck). In simulations on (d1) basically the focus is on how imputed *Z* in *A*: (i) is accurate (prediction accuracy); (ii) preserves its "true" marginal distribution; and (iii) preserves of relationship with *X*s. Only (ii) and (iii) are considered when applying the procedure in case (d2).

First initial results show that traditional hotdeck procedures perform well in terms of all the criteria considered, they however require selection of few good matching variables and wrong choices can add noise. The SL classifications procedure tend to perform better in terms of prediction accuracy and worst in preserving marginal distribution of imputed variable, however this latter feature is improved by introducing a randomization step, i.e. drawing predicted category of *Z* from the estimated probabilities provided by the SL procedure. In almost all the cases an underestimation of true association between *Y* and the *X*s occurs when using imputed *Y* in the synthetic data set (*A*).

When investigating uncertainty, the procedure presented in 2.2 is compared with the "standard" procedure based on conditioning on a subset of the available *X*s, instead of predictions of both *Y* and *Z*. In particular, comparison is based on a rough summary measure of uncertainty, obtained as the average width of uncertainty bounds. The preliminary results show that in some cases application of SL methods helps in reducing uncertainty, however results are very close; in this case adding a randomization step in

prediction (predicted class obtained by draws based on estimated class probabilities) produces slight worse results if compared to direct predictions.

These preliminary results are quite encouraging and lead to believe that the SL methods can profitably be used in the integration of data sources via SM. Further investigation is deserved, also to cover the case of continuous target variables (*Y* and *Z*, or mixed case), where a wider set of prediction SL methods can be applied.

SL methods are usually considered as computational demanding, but power of today machines makes their application quite fast. The procedure presented in the Section 2. Are more computational demanding then the standard traditional ones, but this ratio can be reverted when considering the time-consuming tsk related to the selection of the matching variables required as input by traditional SM methods.

REFERENCES

[1] T. Hastie, R. Tibshirani and J. Friedman (2001), *The Elements of Statistical Learning* Springer, New York.

[2] T. Hastie, R. Tibshirani and J. Friedman (2009), *The Elements of Statistical Learning, 2nd Edition*, Springer, New York.

[3] M. D'Orazio, M. Di Zio and M. Scanu (2006), *Statistical Matching, Theory and Practice*, Wiley, Chichester.

[4] G. Donatiello, M. D'Orazio, D. Frattarola, A. Rizzi, M. Scanu and M. Spaziani (2016) "The role of the conditional independence assumption in statistically matching income and consumption", *International Journal of the IAOS*, 32, pp. 667-675

[5] C. Moriarity and F. Scheuren, (2001) "Statistical matching: a paradigm for assessing the uncertainty in the procedure", *Journal of Official Statistics*, 17, pp. 407-422.

[6] M. D'Orazio, M. Di Zio and M. Scanu (2006) "Statistical matching for categorical data: displaying uncertainty and using logical constraints", Journal of Official Statistics, 22, pp. 1-22

[7] P.L. Conti, D. Marella and M. Scanu (2012) "Uncertainty analysis in statistical matching", *Journal of Official Statistics*, 28, pp. 69-88.

[8] M. D'Orazio, M. Di Zio and M. Scanu (2006) "The use of uncertainty to choose matching variables in statistical matching", *International Journal of Approximate Reasoning* 90, pp. 433-440

[9] ESS (2014) European Social Survey Round 7 Data (2014). Data file edition 2.1. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.