International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

The use of uncertainty to choose matching variables in statistical matching ☆

Marcello D'Orazio [a], Marco Di Zio [b,*], Mauro Scanu [b]

[a] *Food and Agriculture Organization of the United Nations, FAO, Rome, Italy*
[b] *Istituto Nazionale di Statistica ISTAT, Rome, Italy*

ARTICLE INFO

Article history:
Received 21 December 2016
Received in revised form 7 August 2017
Accepted 26 August 2017
Available online xxxx

Keywords:
Data fusion
Synthetical matching
Consistency
Partial identifiability

ABSTRACT

Statistical matching aims at combining information available in distinct sample surveys referred to the same target population. The matching is usually based on a set of common variables shared by the available data sources. For matching purposes just a subset of all the common variables should be chosen, the so called matching variables. The paper presents a novel method for selecting the matching variables based on the analysis of the uncertainty characterizing the matching framework. The uncertainty is caused by unavailability of data for estimating parameters describing the association between variables not jointly observed in a single data source. The paper focuses on the case of categorical variables and presents a sequential procedure for identifying the most effective subset of common variables in reducing the overall uncertainty.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Statistical matching (sometimes called *data fusion* or *synthetical matching*) aims at combining information available in distinct sample surveys referred to the same target population when the two samples are disjoint. Formally, let Y and Z be two random variables; statistical matching techniques have the objective to estimate the joint (Y, Z) probability distribution function or one of its parameters (e.g. a contingency table or a regression coefficient) when:

(i) Y and Z are not jointly observed in a survey, but Y is observed in a sample A, of size n_A, and Z is observed in a sample B, of size n_B;
(ii) A and B are independent and units in the two samples do not overlap (it is not possible to use record linkage);
(iii) A and B both observe a set of additional variables X.

This problem was first studied methodologically in [3] in the case of a trivariate (X, Y, Z) Gaussian distribution. Studies on the lack of identifiability of this problem dates back to [18]. Lack of identifiability has the effect that multiple equally plausible estimates of the joint (Y, Z) distribution are available. Traditional approaches either implicitly or explicitly introduce assumptions to make the model identifiable in order to obtain a unique estimate of the model parameters. The usual assumption is the conditional independence of Y and Z given X. However, this is a strong assumption that – given the

☆ This paper is part of the Virtual special issue on Soft methods in probability and statistics, Edited by Barbara Vantaggi, Maria Brigida Ferraro, Paolo Giordani.

* Corresponding author.
E-mail addresses: marcello.dorazio@fao.org (M. D'Orazio), dizio@istat.it (M. Di Zio), scanu@istat.it (M. Scanu).

http://dx.doi.org/10.1016/j.ijar.2017.08.015
0888-613X/© 2017 Elsevier Inc. All rights reserved.

UNCORRECTED PROOF

data at hand – cannot be tested. In order to avoid the introduction of assumptions with the consequence of making the inferences more credible, a number of papers have started studying how to make inference taking into account the non-uniqueness of estimates in statistical matching. These inferences are about finding out which values of the true parameter of interest are compatible with the observations we made. This set of values is either named "uncertainty region" [11] or "partial identification region" [23,13]. Inference based on partial identification regions is used in other contexts such as for instance econometrics [22] and social sciences [15]. It is worthwhile to remark that the "uncertainty regions" should not be confused with the regions determined by "confidence intervals". In the first case they arise from partial identifiability of the joint distribution, in the second case the intervals are determined by the sampling nature of the observations. Results on a joint analysis of these two types of uncertainty are described in [14] and [21].

In statistical matching, Rubin [19] defines a non-proper Bayesian approach for the exploration of uncertainty regions. This method was generalized in [17] by presenting proper Bayesian approaches. [16] explores the set of equally plausible solutions fixing all the estimable parameters equal to the estimates obtained by means of consistent estimators; this approach is not completely satisfactory because some of the results are not admissible (e.g. covariance matrices in case of Gaussian variables that are negative-definite). The use of maximum likelihood estimators for estimating the identifiable parameters and then finding the corresponding likelihood ridge (i.e. the set of equally maximum likelihood estimates) avoids the possibility to include non-admissible solutions in this set (see [12]). The use of the term *uncertainty* in order to describe the width of the likelihood ridge was first proposed in [12]. Its properties in different contexts and its estimation together with the study of the asymptotic properties of the estimators are in [8].

Although uncertainty is a notion useful to describe how far from the case of identifiability we are in the statistical matching context, the idea of this paper is that it can also be used for operational purposes. More precisely, assume that the two surveys A and B observe a sufficiently large number of common variables X, being all categorical. The case of a large number of common variables X is quite frequent: e.g. in case of social surveys many socio-economic variables (residence, age, gender, professional, educational and marital status, characteristics of the head of the household, characteristics of the households and so on) are available. Should all these variables be used for matching purposes? Problems may arise, for instance it is known that the larger the number of categorical variables, the higher is the risk of having sparse tables, i.e. tables with cells with few or zero observations per cell. More in general, a high number of variables and consequently of categories may have impact on the efficiency of estimates. In this case the main problem consists in estimating simultaneously the high number of identifiable parameters of (Y, Z) given X.

In this paper we claim that an appropriate selection of the matching variables can be found through the notion of uncertainty, that is to select those variables that minimize uncertainty. In order to avoid selecting all the common variables at hand, a method based on a measure of uncertainty penalized by an indicator of sparseness is introduced and discussed in Section 2. Section 3 shows two applications on simulated and real data. Final conclusions are reported in Section 4.

2. Choice of the matching variables

In statistical matching (SM) the data sources A and B may share many common variables X. In performing SM, not all the X variables will be used but just the most important ones. The selection of the most relevant X_M ($X_M \subseteq X$), called *matching variables*, is usually performed by consulting subject matter experts and through appropriate statistical methods (see [10]).

The choice of the matching variables should be made in a multivariate sense [5] to identify the subset X_M connected, at the same time, with Y and Z. This would require the availability of a data source in which (X, Y, Z) are observed. In the basic SM framework, A permits to investigate the relationship between Y and X, while the relationship between Z and X can be studied in B. The results of the two separate analyses are then joined and, in general, the following rule can be applied:

$$X_Y \cap X_Z \subseteq X_M \subseteq X_Y \cup X_Z$$

where X_Y ($X_Y \subseteq X$) and X_Z ($X_Z \subseteq X$) are the subsets of the common variables that better explain Y and Z, respectively. The intersection $X_Y \cap X_Z$ provides a smaller subset of matching variables if compared to $X_Y \cup X_Z$; this is an important feature in achieving parsimony. For instance, too many matching variables in a distance hot-deck SM micro application can introduce undesired additional noise in the final results. Unfortunately, the risk with $X_Y \cap X_Z$ is that the predictors of one target variable will be excluded if they are not in the subset of the predictors of the other target variable, more in general, the intersection may be empty. For this reason, the final subset of the matching variables X_M is usually a compromise and the contribution of subject matter experts and data analysts is important in order to identify the "best" subset. Our proposal is to perform a unique analysis for choosing the matching variables by searching the set of common variables that are the most effective in reducing the *uncertainty* between Y and Z, avoiding selecting too many X variables which determine a high increase of the parameters to estimate.

2.1. Uncertainty in statistical matching

In a SM problem (where Y and Z are never jointly observed) there is an intrinsic uncertainty due to the structure of the data sets at hand: estimators of the parameters describing the association/correlation between Y and Z report multiple

UNCORRECTED PROOF

Please cite this article in press as: M. D'Orazio et al., The use of uncertainty to choose matching variables in statistical matching, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.08.015

Download English Version:

https://daneshyari.com/en/article/4945201

Download Persian Version:

https://daneshyari.com/article/4945201

Daneshyari.com