Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

New methods for small area estimation with linkage uncertainty [☆]

Dario Briscolini [a], Loredana Di Consiglio [b], Brunero Liseo [a,*], Andrea Tancredi [a], Tiziana Tuoto [b]

[a] *Sapienza Università di Roma, Via del Castro Laurenziano 9, Roma 00161, Italy*
[b] *Istat, Via Cesare Balbo, 16, 00184 Roma, Italy*

ARTICLE INFO

Article history:
Received 11 January 2017
Received in revised form 23 December 2017
Accepted 30 December 2017
Available online xxxx

Keywords:
Markov chain Monte Carlo
Measurement error
Nested error model
Record linkage
Uncertainty

ABSTRACT

In official statistics, interest for data integration has been increasingly growing, due to the need of extracting information from different sources. However, the effects of these procedures on the validity of the resulting statistical analyses has been disregarded for a long time. In recent years, it has been largely recognized that linkage is not an error-free procedure and linkage errors, as false links and/or missed links, can invalidate the reliability of estimates in standard statistical models. In this paper we consider the general problem of making inference using data that have been probabilistically linked and we explore the effect of potential linkage errors on the production of small area estimates. We describe the existing methods and propose and compare new approaches both from a classical and from a Bayesian perspective. We perform a simulation study to assess pros and cons of each proposed method; our simulation scheme aims at reproducing a realistic context both for small area estimation and record linkage procedures.

1. Data integration and impact of linkage errors

In official statistics, interest for data integration has been increasingly growing, due to the need of extracting information from different sources. However, the effects of these procedures on the validity of the resulting statistical analyses has been disregarded for a long time. In recent years, it has been largely recognized that linkage is not an error-free procedure and linkage errors, as false links and/or missed links can invalidate the reliability of estimates in standard statistical models. The effect of linkage errors on the calibration of linear regression models with variables observed in different sources was firstly illustrated by Neter et al. [18]. Major contributions to the development of this study can be found in Scheuren and Winkler ([24], [25]) and Lahiri and Larsen [15]. Chambers [3] also considers the construction of a Best Linear Unbiased Estimator and its empirical version. He also proposes a maximum likelihood estimator, providing examples with application in linear regression models, with a generalization to the logistic case. A possible extension to sample-to-register linkage is also suggested. On the Bayesian side, Tancredi and Liseo [27] and Tancredi et al. [28] have proposed an integrated model

[☆] 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: dario.briscolini@uniroma1.it (D. Briscolini), diconsig@istat.it (L. Di Consiglio), brunero.liseo@uniroma1.it (B. Liseo), andrea.tancredi@uniroma1.it (A. Tancredi), tuoto@istat.it (T. Tuoto).

https://doi.org/10.1016/j.ijar.2017.12.005
0888-613X/© 2018 Elsevier Inc. All rights reserved.

with a feed-back effect in which inferential procedures for the regression are able to borrow strength from the linkage process and vice versa.

This article focuses on the effects of linking errors on the production of small area estimates. In particular we consider the case of unit-level small area methods. They apply when some auxiliary variables X, whose totals are known for each small area, are available for each sampled unit. Small area predictions are usually constructed using (possibly generalized) linear mixed models expressing the survey variable Y in terms of X.

Samart and Chambers [23] consider the effect of linkage errors on mixed effect models, extending the settings in Chambers [3] and suggesting estimators of the variance effects which are adjusted for linkage errors. In official statistics, mixed models are largely used in small area estimation in order to increase the detail of information at local level. Administrative data can also be used to increase information collected in sample surveys, in order to expand auxiliary information and improve the model fitting for small area estimation. Linkage of external sources with basic statistical registers as well as with sample surveys can be carried out on different linkage scenarios. Di Consiglio and Tuoto [7] performed a sensitivity analysis for different alternative linkage error scenarios in linear and logistic regression settings.

In this paper, we present a comparative analysis of several different estimators of the parameters of a unit-level small area model both from a classical and a Bayesian perspective. We compare the results on a pseudo population, where the values of the survey variable Y and those of covariates X are obtained from the survey on Household Income and Wealth, Bank of Italy and the person identifiers come from the fictitious population census data [8] created for the ESSnet DI, an European project on data integration run from 2009 to 2011. The data set contains 26,625 observations and consists of 25 variables.

In a classical framework, under the assumption that false matches may occur only within the same small area, linkage errors affect small area predictors via a bias on the estimation of fixed components and random effects. Following Chambers [3], we assume that sampling does not change the outcome of the linkage process and we derive an adjusted EBLUP estimator. We also propose a Bayesian strategy where we jointly model the record linkage and the small area model using response variable and covariates available in different data sets. We believe that the latter approach is able – in a very natural way – to

- improve the performance of the linkage step through the use of the extra information contained in the Y's (the response variable values) and the covariates X's. This occurs because pairs of records which do not adequately fit the small area model, say $\mathcal{M}$, will be automatically down-weighted in the matching process;
- allow us to account for matching uncertainty in the estimation procedure related to model $\mathcal{M}$ involving Y's and X's;
- improve the accuracy of estimators of model parameters $\mathcal{M}$ in terms of bias.

Although we present several different strategies for estimating the parameters of the small area model, we stress the fact that a fair comparison among the different methods is not possible, since they consider different sets of assumptions. In the simulation study section we will discuss these issues in detail.

The linkage methods used in this paper refer to those implemented in *RELAIS* [21] on the frequentist side; for the Bayesian approaches we have used methods described in Tancredi and Liseo [27] and Tancredi et al. [28], where categorical variables are used for the linkage procedure, while either continuous or categorical variable can be considered in the inferential post-linkage step, as it might be the case in small area models.

The rest of the paper is organized as follows: Section 2 describes the statistical problem of linking data both from a classical and from a Bayesian perspective. Section 3 illustrates the different strategies of estimation in small area models. Section 4 compares the different methods using a simulation setting and a realistic pseudo-population, as described above, which mimics typical data sets to be used in record linkage problems and in small area estimation as well.

2. Linkage model and linkage errors

From a statistical perspective, the operation of merging two (or more) data sets can be important for two different and complementary reasons:

(i) to obtain a larger reference data set or frame, suitable to perform more accurate statistical analyses;
(ii) to make inference on suitable statistical models via the additional information which could not be extracted from either one of the two single data sets.

If the merging step can be accomplished without errors (maybe because an error-free identification key is available and it can be used to match units in different data sets), there are no specific consequences on the statistical procedures undertaken in both the situations. In practice, however, identification keys are rarely available and linkage between records is usually performed under uncertainty. This issue has caused a very active line of research among the statistical and the Information Technology communities, named “record linkage”, where the possibility to make wrong matching decisions must be accounted for, especially when the result of the linking operation, namely the merged data set, must be used for further statistical analyses.

Please cite this article in press as: D. Briscolini et al., New methods for small area estimation with linkage uncertainty, Int. J. Approx. Reason. (2018), https://doi.org/10.1016/j.ijar.2017.12.005