STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 103–122, DOI 10.21307/stattrans-2020-033
Received – 31.01.2020; accepted – 30.06.2020

A comparison of area level and unit level small area models in the presence of linkage errors

Loredana Di Consiglio[1], **Tiziana Tuoto**[2]

ABSTRACT

In Official Statistics, interest in data integration has grown enormously, but the effect of integration procedures on statistical analysis has not yet been sufficiently developed. Data integration is not an error-free procedure and linkage errors, as false links and missed links can invalidate standard estimates. Recently, increasing attention has been paid to the effect of linkage errors on the statistical analyses and on statistical predictions.

Recently, methods to adjust the unit level small area estimators for linkage errors have been proposed when the domains are correctly specified. In this paper we compare the naïve and the adjusted unit level estimators with the area level estimators that are not affected by the linkage errors. The comparison encourages the use of the adjusted unit level estimator.

Key words: linear mixed models, data integration, linkage errors.

1. Data integration and the impact of linkage errors

In Official Statistics, data integration has been acquiring more and more importance; the effect of this procedure on statistical analyses has long been disregarded for a long time but in recent years the impact of linkage errors, false links and missed links, on standard estimates has begun to be analysed. The effect of linkage errors on subsequent analyses has first been investigated by Neter et al. (1965) where first solutions can be found.

Scheuren and Winkler (1993, 1997) and Lahiri and Larsen (2005) analyse the problem from a primary user perspective; in this case the evaluation of the linkage errors is a by-product of the linkage procedure and they propose different methods to use this information to adjust for the linkage biases in subsequent analyses. Clearly, the resulting unbiased estimators depend on the parameters of the linkage model. Recently, Han and Lahiri (2018) propose a general framework for statistical analysis with linked data under general assumptions. A different perspective is in Chambers (2009); secondary data users generally do not have detailed information on linkage model and parameters, in this setting, Chambers (2009) suggests an approximated Best Linear Unbiased Estimator and its empirical version and proposes a maximum likelihood estimator with application to linear and logistic regression functions. An extension to sample-to-register linkage is also proposed.

[1] Istituto Nazionale di Statistica - Istat, Italy. E-mail: diconsig@istat.it
[2] Istituto Nazionale di Statistica - Istat, Italy. E-mail: tuoto@istat.it.
ORCID: https://orcid.org/0000-0003-3436-9474.

In the context of fitting mixed models with linked data, Samart and Chambers (2014) extend the settings in Chambers (2009) and suggest linkage error adjusted estimators of variance effects under alternative methods. In Official Statistics, mixed models are largely used for small area estimation to increase the detail of dissemination of statistical information at local level.

Administrative data can be used to augment the information collected by sample surveys. They can, therefore, increase the set of auxiliary variables and help to 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, see section 2 for the linkage model and errors we adopt in this paper.

Di Consiglio and Tuoto (2016) extend the analysis on the effects of linkage errors on the predictors based on unit level mixed models for small area estimation when auxiliary variables are obtained through a linkage procedure with an external register.

Under the assumption that false matches occur only within the same small area - i.e. in Chambers's terminology the block coincides with the small area-, the linkage errors affect small area predictors both through the impact on the estimation of the fixed and random components, and through the impact on the variance matrix of the linked values. Finally, linkage errors also result in an erroneous evaluation of the covariates means over the sampled units and consequently of the unobserved population units.

Following Chambers (2009) in the sample-to-register linkage setting, and in particular, assuming that the sampling mechanism does not affect the outcome of the linkage process (see Chambers 2009 for details), a pseudo-EBLUP estimator based on the derived distribution of the linked variable can be obtained. Section 3.4 illustrates the method in more detail.

Briscolini et al. (2018) introduce a Bayesian approach that jointly solves the record linkage problem and the small area predictions. They also compare the Bayesian approach with the frequentist estimator proposed in Di Consiglio and Tuoto (2016). In the context of secondary data analysis, Han (2018) put forward an approach to solve small area estimation in presence of linkage errors.

The cited studies focus on the evaluating and the adjustment of linkage errors when small area prediction is performed by a unit level model. However, one might question whether the complexity of adjusting for linkage errors at unit level is in fact overwhelmed by the simplicity of area level models, which do not require unit level linkage for the estimation.

This paper aims at comparing the unit level estimator with the area level estimator in the presence of linkage errors, illustrating advantages and drawbacks by means of the application to real case and the simulation of various scenarios.

2. Linkage model and linkage errors

The reference theory for record linkage dates back to Fellegi and Sunter (1969). They consider the linkage between two lists, L_1 and L_2, of size N_1 and N_2 respectively. Within this context, we can consider, for instance, the linkage between a register and a sample. From a statistical viewpoint, the linking process is a classification problem; it aims to classify all the pairs generated by the lists' comparison $\Omega = \{L_1 \times L_2\} = \{\omega = (i, j)\}$ where $i \in L_1$ and $j \in L_2$ into two independent and mutually exclusive subsets, M and U respectively;

- M is the set of links, grouping all the pairs composed by records belonging to the same unit $M = \{\omega = (i, j) \quad | \quad i = j\}$;
- U is the set of non-links $U = \{\omega = (i, j) \quad | \quad i \neq j\}$, where $i \in L_1$ and $j \in L_2$.

The classification decision is taken for each pair ω on the basis of the comparison on K linking variables, common to the two lists, e.g. name, surname, date of birth, address. The comparison on the linking variables results in a comparison vector γ_{ij}, e.g. $\gamma_{ij} = (1, 1, 0, 1)$ if unit $i \in L_1$ and unit $j \in L_2$ present the same (or similar) values for the first, the second, and the forth linking variables and different (or quite dissimilar) value for the third linking variable. From the observed probability distribution of γ over the pair space Ω, two probability distributions are estimated:

- $m(\gamma_{ij})$, i.e. the probability of γ given that the pair (i, j) belongs to set M;
- $u(\gamma_{ij})$, i.e. the probability of γ given that the pair (i, j) belongs to set U.

To estimate the two distributions $m(\gamma_{ij})$ and $u(\gamma_{ij})$, and the prevalence of the links in the pairs $\pi = |M|/|\Omega|$ usually the EM algorithm is applied; details can be found in Jaro (1989), Herzog et al. (2007).

The classification procedure might produce two kinds of errors: the mismatch or false positive, when a pair (i, j) is classified as a link but in reality the two records i and j refer to different units, and the missing match or false negative, when the pair (i, j) is classified as a non-link but in reality the two records i and j belong to the same unit.

Linkage procedure aims at minimising both the probability of false match and the probability of missing match or, at least, to keep both below acceptable values. The classification procedure provides as a by-product the false positive rate and the false negative rate. For each pair, it also provides estimate of the probability of being a correct link given that the link is assigned:

$$\lambda_{ij} = \frac{m(\gamma_{ij})\pi}{m(\gamma_{ij})\pi + u(\gamma_{ij})(1 - \pi)}. \tag{1}$$

The quantities λ_{ij} will be exploited for adjusting the linkage errors in the small area estimation framework described in the next section. It is worthwhile noting that accurate estimation of these probabilities is not a trivial task, even when the probabilistic linkage strategies are very effective in identifying the correct links. We will go back to this point in section 4, however the estimation of λ_{ij} is not the focus of this paper.

3. Small area estimation

When the survey is not planned to provide estimates at a very fine disaggregation (e.g. by geography or by a cross-classification such as gender and age), the standard estimates are often too variable, because the sample size is too small or zero at the desired level. Small area estimation methods allow an improvement of the quality of the estimates exploiting relationships of the target variable with highly correlated auxiliary variables at unit level or area (domain) level. For an extensive review of small area methods, see Rao and Molina (2015).

In the following sub-sections we briefly overview the basic unit level (Battese-Harter-Fuller, 1988) and area level (Fay-Herriott, 1979) estimators. We describe how the former has to be modified to account for the linkage errors in the presence of auxiliary variables that are not recorded in the survey but obtained from an external source, such as administrative data.

3.1. The unit linear mixed model

Let the population units be partitioned into D different domains. Let Y be the target variable and X the auxiliary variables observed on the same units. Let us assume a linear mixed relationship between the target variable and the covariates

$$y_{id} = X_{id}^T\beta + u_d + e_{id}, \;\; i = 1,\ldots,N_d, \;\; d = 1,\ldots,D, \tag{2}$$

where β is a p-dimensional vector of fixed regression coefficients and u_d, $d = 1,\ldots,D$, are the i.i.d. random variables related to the specific or domain contributions, with $E(u_d) = 0$ and $V(u_d) = \sigma_u^2$, independently distributed to the random errors e_{id} i.i.d. with $E(e_{id}) = 0$ and $V(e_{id}) = \sigma_e^2$. In matrix notation

$$Y = X\beta + Zu + e$$

where Z is the design matrix denoting the belonging of units to the areas: $Z = Blockdiag(Z_d = 1_{N_d}; d = 1\cdots D)$.

The total variance is given by $V(Y) = V = \sigma_u^2 ZZ^T + \sigma_e^2 I$; equivalently, in matrix notation, $V = diag(V_d; d = 1\cdots D)$ with $V_d = \sigma_e^2 I_{N_d} + \sigma_u^2 Z_d Z_d^T$. When σ_u^2 and σ_e^2 are known, the BLUP estimator of a small area mean or totals $\bar{Y}_d$, is given by

$$\hat{\bar{Y}}_d^{BLUP} = \frac{1}{N_d}\left(\sum_{i\in s_d} y_{id} + \sum_{i\in s_d^c} \hat{y}_{id}^{BLUP}\right) \tag{3}$$

where s_d is the sample in area d, $\hat{y}_{id}^{BLUP} = X_{id}^T\hat{\beta} + \tilde{u}_d$ with

$$\hat{\beta} = (X_s^T V_{ss}^{-1} X_s)^{-1} X_s^T V_{ss}^{-1} y$$

and $\tilde{u} = \sigma_u Z_s^T V_{ss}^{-1}(y - X_s\hat{\beta})$, where y is the sample vector of Y and denoting with the subscript s the portion of vector and matrices related to the sample observations.

In real cases, the estimates are given by the EBLUP that is obtained by plugging the estimates $\hat{\sigma}_u$ and $\hat{\sigma}_e$ into V and then into the previous expressions of $\hat{\beta}$ and $\tilde{u}$. See the section (sec.3.5) for a brief overview of the variance components estimation.

3.2. Area level small area predictor

The basic area level model (Fay and Herriot, 1979) relies on a linear relationship between the direct estimates $\hat{\bar{Y}}_d$ and the true finite population values $\bar{Y}_d$ in each area d, and a linear relationship among the true values and known area totals X_d:

$$\hat{\bar{Y}}_d = \bar{Y}_d + \varepsilon_d \qquad d = 1, \ldots, D, \tag{4}$$

where ε_d is the sampling error in the estimation of $\bar{Y}_d$, with mean zero and assumed known variance σ^2_{ed}, and

$$\bar{Y}_d = X_d\beta + u_d \qquad d = 1, \ldots, D, \tag{5}$$

where β is the vector of regression coefficients and u_d is assumed to be normal with zero mean and variance σ^2_u. Combining (4) and (5) one gets:

$$\hat{\bar{Y}}_d = X_d\beta + \varepsilon_d + u_d \qquad d = 1, \ldots, D, \tag{6}$$

where ε and u are assumed to be independent.

The BLUP estimator based on the model in (6) is given by:

$$\tilde{\bar{Y}}_d^{FH} = \gamma_d\hat{\bar{Y}}_d + (1-\gamma_d)X_d\hat{\beta} \qquad d = 1, \ldots, D, \tag{7}$$

where $\gamma_d = \sigma^2_u/(\sigma^2_u + \sigma^2_{ed})$. The EBLUP is obtained by replacing an estimate (e.g ML or REML estimate) of σ^2_u in formula (7). See Molina and Rao (2015) for more details. The FH model assumes known σ^2_{ed}. In practice it has to be estimated. See section (4) for more details on how it is estimated in the present work.

3.3. Linear mixed model under Record Linkage

When the auxiliary variables X and target variable Y are not jointly observed on the same data set but are obtained, for instance, by linking a sample with a register, the use of the relationship (2) and the corresponding estimator can produce biased estimates, if naively applied on linked data. Di Consiglio and Tuoto (2016) analyse the effect of linkage errors on unit level small area estimators and propose an adjustment to account for linkage errors, following the setting in Chambers (2009) and Samart and Chambers (2014).

The proposed adjustment, however, requires that no linkage errors occur between blocks/small areas. Under this assumption, the area level estimator is not affected by linkage errors and therefore linkage bias, since it only needs the mean value of X for each of the target domains. Hence, under the assumption of no linkage errors between areas, the standard Fay-Herriot estimator can be applied even in the presence of linkage errors within the small areas.

Let us first consider a register-register linkage and describe the linear mixed model and the proposed adjustment in this linkage setting.

Let us denote with y^*_{id} the value of the variable Y from one register that is matched with the value X_{id} in the other register, for unit i in domain d.

Let us assume that the blocking variable Z is measured without error on both the Y-register and the X-register, and that the partition of the registers introduced by Z is such that linkage errors only occur within this blocking variable.

Finally, let us assume an exchangeable linkage error model (see Chambers, 2009), i.e. the probability of correct linkage is the same for all records in block q, $q = 1, \cdots, Q$.

Under the following standard assumptions, as in Chambers (2009) and in Samart and Chamber (2010):

1. the linkage is complete, i.e. the X-register and Y-register refer to the same population and have no duplicates;

2. the linkage is one to one between the Y- and X-registers;

3. exchangeable linkage error model;

the observed linked variable Y^* is a permutation of the true one Y: $Y^* = AY$, where A is a random permutation matrix such that $E(A|X) = E$. The blocking index q is omitted in previous equations for simplicity of notation.
Being $Pr(a_{ii} = 1|X) = Pr(correct\ \ linkage) = \lambda$ and $Pr(a_{ij} = 1|X) = Pr(incorrect\ \ linkage) = \psi$, the expected value $E(A|X) = E$ can be written as:

$$E = (\lambda - \psi)I + \psi 1 1^T. \tag{8}$$

In this setting, Samart and Chambers (2014) proposed a ratio type corrected estimator for the regression coefficients β:

$$\tilde{\beta}_R = (X^T V^{-1} E X)^{-1} X^T V^{-1} y^* \tag{9}$$

following the same rationale of the bias correction estimator in the linear model (Chambers, 2009). They also proposed an approximation of the BLUE estimator by exploiting the new relationship between Y^* and X:

$$\tilde{\beta}_C = (X^T E^T \Sigma^{-1} E X)^{-1} X^T E^T \Sigma^{-1} y^* \tag{10}$$

where the derived variance $V(Y^*)$ of the observed y^* is considered:

$$V(Y^*) = \Sigma = \sigma^2_u K + \sigma^2_e I + W \tag{11}$$

with

$$W \approx diag((1-\lambda)(\lambda(f_i - \bar{f}) + \bar{f}^{(2)} - \bar{f}^2)) \tag{12}$$

being $f_i = X_i \beta$ and K a function of the number of areas within a block, block-group sizes and λs; see Samart and Chambers (2014) for more details. Clearly, the estimation of β requires an iterative process as Σ depends on β via the f. Moreover, the variance components are unknown and have to be estimates. The linkage errors can affect also their estimation, see section 3.5 for a short description of how Samart and Chambers (2014) propose to deal with this issue.

3.4. Unit level small area predictor under linkage errors

Let us now consider the more realistic situation when the linkage is between a sample, where the variable Y is observed, and a register where X is recorded; this is the case where mixed models are useful for small area estimation.

In the sample-to-register setting, Chambers (2009) adds the assumption that the sampling does not change the outcome of the linkage process, i.e. selecting a record to be in sample does not change the register record to which it would be linked if all records were linked. Hence, the same permutation of the y described above would apply. This scheme works as if a hypothetical linkage can be performed before the sampling process and then we observe the sampled sub-set.

This assumption, as already pointed out by Chambers (2009), can be easily challenged as the sampling process may indeed affect the linkage process, but it is very useful in extending the register-register estimation setting to the survey-register situation.

Under the given conditions, the matrices E, V and Σ depend only on blocking variables and linkage errors, so there is no need to use sampling weights.

If the exchangeable linkage error model is assumed, as in section 3.3, the linkage errors occur only within the same block where records have the same probability of being correctly linked, then the mixed model can be fitted with the observed sample quantities applying the same argument as in the register-to-register case. See Chambers (2009) for more details.

Finally, for the small area estimation, we assume that small areas coincide with blocks. Note that with the latter assumption, the target mean of y is the same as the mean of the linked Y^*:

$$\hat{\bar{Y}}^* = \hat{\bar{Y}}.$$

Di Consiglio and Tuoto (2016) propose to exploit the distribution of Y^* to obtain the pseudo-BLUP estimator of $\bar{y}^*$ and then an estimation of $\bar{y}$:

$$\hat{\bar{Y}}_d^{*BLUP} = \frac{1}{N_d}\left(\sum_{i\in s_d} y_{id}^* + \sum_{i\in s_d^c} \hat{y}_{id}^{*BLUP}\right) \tag{13}$$

where $\hat{y}_{id}^{*BLUP} = EX\tilde{\beta}_C + \tilde{u}_d$, $\tilde{u} = \sigma_u Z^T \Sigma^{-1}(y^* - EX\tilde{\beta}_C)$ and $\tilde{\beta}_C$ is given in formula (10).

The pseudo-EBLUP estimator is given by replacing the estimates of the variance components (as in section 3.5) into the estimates of $\tilde{\beta}_C$ and $\tilde{u}$ and then in (13).

3.5. Estimation of variance components

The BLUE and the approximate BLUE estimators considered in the previous sections are based on known variance components. However, the variance components σ_u and σ_e are usually unknown, they are commonly estimated by methods of moments, ML or REML (Harville, 1977, Searle et al 2006). In Samart and Chambers (2014), a Pseudo-ML and

Pseudo-REML are proposed for adjusting variance component estimation for linkage errors. In the application and simulation study reported in section 4, we consider only ML approach, and pseudo-ML for the linkage error framework, assuming multivariate normal distribution.

In general, there is no analytical expression for the ML variance component estimator and the method of scoring is applied. When variables X and Y are both recorded on the sample, hence no linkage errors, the target variable is $y \sim N(X\beta;V)$. On the other hand, in the presence of linkage errors, we should use the modified distribution $y^* \sim N(Ef;\Sigma)$. The scoring algorithm can be applied on the derivatives of this likelihood rather than of the likelihood of the un-observed target variable y.

In the presence of linkage errors, estimates of β can be obtained from formulas (9) or (10) by replacing the variance components with their estimates. An iterative process is needed between the pseudo-ML estimates of the variance components and the estimate of β. See Samart and Chambers (2014) for more details.

4. Results on real and simulated data

Previous estimators are applied to a realistic case for estimating small areas in the presence of linkage errors. In addition, several synthetic populations have been generated based on two different mixed linear models to test the performance of estimators in a controlled environment. This section illustrates the real case and the data generation for the controlled experiment and describes the result.

4.1. The real case data

Microdata from the Survey on Household Income and Wealth, Bank of Italy, (SHIW), can be used to study the relationship between the consumption (the variable Y observed throughout the survey) and the net disposable income (the variable X available for the whole population). The survey sample is designed to produce reliable estimates at NUTS1 level, but the relevance of the topic prompts analysis of the results at the finer level, i.e. the NUTS2 administrative regions, which therefore represent a small area of estimation. In fact, variables Y and X are both observed by the survey: this allows us to compare different settings for linkage and mixed model estimation, knowing the true value of the regression model parameters. However, in principle one can imagine to study the relationship between the consumption recorded via the survey and the income from the tax register, available to the entire Italian population, thus overcoming the households' reluctance to provide information on income via a survey.

To overcome privacy issue and guarantee the reproducibility of the experiment, the record linkage procedure is applied to the fictitious population census data (McLeod et al. 2011) created for the ESSnet DI, an European project on data integration that run from 2009 to 2011. The population size is over 20000 records; data contain linking variables (names, dates of birth, addresses) for individual identification with missing values and typos, mimicking a real situation. The small domains are defined as aggregation of postal codes, assigning 18 areas. From this population, 100 replicated samples of size 1000 were

Table 1: True values of the correct linkage rates

Scenario	Min(λ)	Mean(λ)	Max(λ)	MMR
A	0.9525	0.9730	0.9834	0.0629
B	0.8430	0.8757	0.9043	0.0424

average values in 100 replications, over the 18 areas

independently randomly selected without replacement. Finally on each replication, the sample containing the Y variable was linked with the register reporting the X variables. The linkage was performed by means of the batch version of the software RELAIS (2015) that implements the probabilistic record linkage model (Fellegi and Sunter, 1969; Jaro, 1989).

We considered two linkage scenarios, characterized by two different sets of linking variables: in Scenario A we used "Day, Month, and Year of Birth"; in Scenario B we adopted "Day and Year of Birth", and "Gender". The first scenario uses linking variables with higher identifying power than the second scenario, producing fewer linkage errors in the results (both in terms of missing and false links). In both scenarios we assume that false linkage errors between different areas do not occur, in other words the administrative areas, i.e. the small domains are the blocking variable for the linkage procedures. Both scenarios also contain missing matches, mimicking the real outcomes of linkage procedures. Missing matches are mainly due to typos in the linking variables and hence they are independent from the target variable Y and the auxiliary variables X. In few words, they can be considered missing at random. However, they have the effect of reducing the sample size. Therefore, in the presence of linkage procedure, the estimators rely on the linked subset s_{Ld} of the sample s_d for the domain d.

True matches are known for the ESSnet DI data, so one can calculate the true value of the linkage errors for the proposed scenarios by comparing the obtained links with the true matches. Therefore, the value of the probability of correct link, λ, is calculated for each block (small area), as the ratio between the true matches in the linked set and the links within each area. Table 1 summarizes the results of the linkage procedures for the 100 replicas, showing the statistics for the probability of correct link λ, on average in the 18 areas. Moreover, Table 1 reports the average of the missing match rate, MMR, in the 18 areas for the 100 replicas, calculated as one minus the ratio between the numbers of identified links and the true matches. As expected, in the two scenarios, there is a trade-off between false matches and missing matches: scenario A has a lower false match rate but a higher missing match rate and vice-versa for scenario B.

For the adjusted estimator introduced in section 3.4, we use the true false linkage rate, $1-\lambda$, in each area. We do not simulate additional evaluation of λs, as the accurate estimation of λ is still an open research question in record linkage and it is not in the focus of this paper. However, at the end of the simulation study, we propose an insight into the behavior of the estimators when the linkage errors are overestimated.

The experiment considers five estimators for comparison:

1. BHF : is the EBLUP based on the Battese-Harter-Fuller model with X and Y observed on the same dataset, i.e. no linkage is assumed in this setting:

$$\hat{\bar{Y}}_d^{BHF} = \frac{1}{N_d}\left(\sum_{i \in s_d} y_{id} + \sum_{i \in s_d^c} \hat{y}_{id}^{EBLUP}\right),$$

   where s_d is the sample in area d, $\hat{y}_{id}^{EBLUP} = X_{id}^T\hat{\beta} + \hat{u}_d$ with

$$\hat{\beta} = (X_s^T \hat{V}_{ss}^{-1} X_s)^{-1} X_s^T \hat{V}_{ss}^{-1} y$$

   and $\hat{u} = \hat{\sigma}_u Z_s^T \hat{V}_{ss}^{-1}(y - X\hat{\beta})$.

2. BHF_L : is the EBLUP based on the Battese-Harter-Fuller model on the subset of linked records. In this estimator we reduce the sample size to the linked records but we do not introduce linkage errors; this is our benchmark:

$$\hat{\bar{Y}}_d^{BHF\_L} = \frac{1}{N_d}\left(\sum_{i \in s_{Ld}} y_{id} + \sum_{i \in s_{Ld}^c} \hat{y}_{id}^{EBLUP}\right),$$

   where s_{Ld} is the sub-set of linked sample units in area d.

3. BHF_naive : is the naïve EBLUP based on the Battese-Harter-Fuller model on the subset of linked records, considering X and Y observed on two different datasets, without adjustment for linkage error:

$$\hat{\bar{Y}}_d^{BHF\_naive} = \frac{1}{N_d}\left(\sum_{i \in s_{Ld}} y_{id}^* + \sum_{i \in s_{Ld}^c} \hat{y}_{id}^{*EBLUP\_naive}\right),$$

   where s_{Ld} is the sub-set of linked sample units in area d, $\hat{y}_{id}^{*EBLUP\_naive} = X_{id}^T\hat{\beta}^* + \hat{u}_d$ with

$$\hat{\beta}^* = (X_{s_L}^T \hat{V}_{s_L s_L}^{-1} X_{s_L})^{-1} X_{s_L}^T V_{s_L s_L}\hat{}^{\,-1} y^*$$

   and $\hat{u} = \hat{\sigma}_u Z^T \hat{V}_{s_L s_L}^{-1}(y^* - X_{s_L}\hat{\beta}^*)$.

4. BHF_adj: is the adjusted EBLUP based on the Battese-Harter-Fuller model:

$$\hat{\bar{Y}}_d^{BHF\_adj} = \frac{1}{N_d}\left(\sum_{i \in s_{Ld}} y_{id}^* + \sum_{i \in s_{Ld}^c} \hat{y}_{id}^{*EBLUP}\right),$$

   where $\hat{y}_{id}^{*EBLUP} = EX\hat{\beta}_C + \hat{u}_d$ and $\hat{u} = \hat{\sigma}_u Z^T \hat{\Sigma}^{-1}(y^* - EX\hat{\beta}_C)$, and $\hat{\beta}_C$ is given by

$$\hat{\beta}_C = (X_{s_L}^T E_{s_L}^T \hat{\Sigma}_{s_L s_L}^{-1} E_{s_L} X_{s_L})^{-1} X_{s_L}^T E_{s_L}^T \hat{\Sigma}_{s_L s_L}^{-1} y^*.$$

5. FH : is the EBLUP based on the Fay-Herriot model:

$$\tilde{\hat{Y}}^{FH} = \hat{\gamma}_d \hat{\hat{Y}} + (1-\hat{\gamma}_d) X_d \hat{\beta},$$

where $\hat{\gamma}_d = \hat{\sigma}_u/(\hat{\sigma}_u + \hat{\sigma}_{ed})$. The FH model assumes known sampling variance σ^2_{ed}, however it needs to be estimated in practice. In this simulation, we used a simple minded smoothing method, which assumes that the population variances of all the domains are identical, σ^2_e. The variances of the direct estimators are then evaluated as $\hat{\sigma}^2_e/n_d$ where $\hat{\sigma}^2_e$ is estimated from the unit linear model.

It is worth noting that the five estimators are evaluated on different sub-sets; the BHF estimator and the FH estimator are evaluated on the sample s_d, the BHF_naive and the BHF_adj estimators are evaluated on the linked sub-set s_{Ld} that might include linkage errors; the estimator BHF_L is evaluated on the sub-sample s_{Ld} but the correct values of X in the register have been used.

Table 2 reports the average of the Absolute Relative Error (ARE) over the 18 areas, the average of the Standard Deviation (SD), and the average of the Mean Square Error (MSE). Results in table 2 show that in terms of bias the area level estimator outperforms the unit level estimators, even when linkage error correction is applied. However, in terms of variability, the area level estimator shows values considerably higher compared to the other estimators. We assumed equal population variances in all domains in the implementation of the Fay-Herriot model. This assumption may be not appropriate in our context, highlighting that sampling variance smoothing deserves great attention in the application of the FH estimator. We will return to this point in the concluding remarks, though the variance estimation is not the focus of this paper, see Hawala and Lahiri (2018) for some ideas on variance modeling.

Table 2 shows that the adjusted unit level EBLUP (BHF_adj) reduces the bias with respect to the naïve estimator (BHF_naive), at the price of an increase in variance that is, however, compensated at MSE level. In fact, the MSE of the adjusted unit level EBLUP (BHF_adj) is similar to that of the benchmark estimator (BHF_L), based on the linked sample without errors. Similar results are also in Di Consiglio and Tuoto (2016), and in Briscolini et al. (2018). It is worth noting that the adjustment for linkage errors does not completely eliminate the bias. We will return to this point in our concluding remarks.

4.2. Simulated data

In the previous subsection, the comparison of the unit level and area level estimators in the presence of linkage errors can be affected by the actual relationship between the variables, which are observed in the field and interpreted with linear mixed models to pursue our purposes.

To compare the unit level and the area level estimators in the presence of linkage errors in a fully controlled setting, we create two different models, Model1 and Model2,

Table 2: Average of the absolute relative error (ARE), standard deviation (SD), and Mean Square Error (MSE) for estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH

ARE

Scenario	BHF	BHF_L	BHF_naive	BHF_adj	FH
A	0.0330	0.0333	0.0350	0.0335	0.0231
B	0.0330	0.0334	0.0430	0.0347	0.0231

SD

Scenario	BHF	BHF_L	BHF_naive	BHF_adj	FH
A	0.4659	0.4820	0.4729	0.4762	2.3188
B	0.4659	0.5426	0.5107	0.5262	2.3188

MSE

Scenario	BHF	BHF_L	BHF_naive	BHF_adj	FH
A	0.6753	0.6906	0.6981	0.6913	2.3336
B	0.6753	0.7358	0.7938	0.7383	2.3336

based on the following linear mixed models:

$$\begin{aligned}
Model1 :& X \sim [1, Uniform(0,1)], \qquad \beta = [2,4], \qquad u \sim N(0,1), \qquad e \sim N(0,3),\\
& RealizedVar(u) = 1.5728\\
Model2 :& X \sim [1, Uniform(0,1)], \qquad \beta = [2,4], \qquad u \sim N(0,3), \qquad e \sim N(0,1),\\
& RealizedVar(u) = 4.7186.
\end{aligned}$$

The variables from the two models have been attached to the ESSnet DI data, containing the linking variables. The previous linking scenarios, A and B, have been considered for each model. Then, for each model, 100 replicated samples of size 1000 were independently and randomly selected without replacement; finally, for each replication, the sample containing the variable Y was linked to the register that reported the variables X.

As in the previous section, five estimators are compared: BHF, BHF_L, BHF_naive, BHF_adj and FH. Table 3 reports the Absolute Relative Error (ARE), the Standard Deviation (SD), and the Mean Square Error (MSE), averaged over the 18 areas, for linkage scenario B. The results for linkage scenario A are substantially similar and are not presented here for the sake of brevity.

For BF estimators, bias and variance are smaller in Model 2 than in Model 1. This is not the case for the FH estimator. As already observed with real data, the bias reduction of the adjusted estimator BHF_adj more than offsets the increase in variance, so the

Table 3: Average of the absolute relative error (ARE), standard deviation (SD), and Mean Square Error (MSE) for estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH

	ARE				
	BHF	BHF_L	BHF_naive	BHF_adj	FH
Model1	0.0412	0.0423	0.0476	0.0435	0.0401
Model2	0.0135	0.0137	0.0199	0.0161	0.0266

	SD				
	BHF	BHF_L	BHF_naive	BHF_adj	FH
Model1	0.3265	0.3447	0.3349	0.3424	0.9108
Model2	0.2263	0.2412	0.2522	0.2519	1.0040

	MSE				
	BHF	BHF_L	BHF_naive	BHF_adj	FH
Model1	0.3837	0.4018	0.4060	0.4013	0.9240
Model2	0.2333	0.2476	0.2652	0.2595	1.0064

MSE of estimator BHF_adj is always smaller than the MSE of estimator BHF_naive. The improvement is quite small when the linkage errors are small. As far as the area level estimator is concerned, it performs better than the BHF estimators in terms of bias in Model 1, whilst the FH performs worse than the unit level estimators, including the not-adjusted estimator BHF_naive in Model 2. In terms of variability, as anticipated in the previous section on real data, the area level estimator FH performs worse than the others, in both scenarios and in both models. The boxplot in figure 1 shows the relative errors for the estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH, in the 18 areas.

Figure 2 shows the standard deviations for the estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH in the 18 areas. The distribution over the areas basically confirms the behavior of the estimators highlighted in table 3.

These evidences do not allow us to answer in a definitive way to the initial question of the possible advantage of the FH which, unlike the unit level estimator in the presence of linkage errors, does not require unit linkage. This simulation study seems to suggest that there are situations (Model 1, real data of previous section) where the area level estimator can perform well enough and one can avoid to complicate the analysis introducing record linkage to apply an adjusted unit level estimator, if the FH guarantees enough accuracy. However, there are also contexts (e.g. Model 2) that show the advantages of considering auxiliary information at record level, even in the presence of uncertainty introduced by record linkage. As an aside, one should be careful on using an appropriate smoothing method for the variance of the direct for the FH estimator.

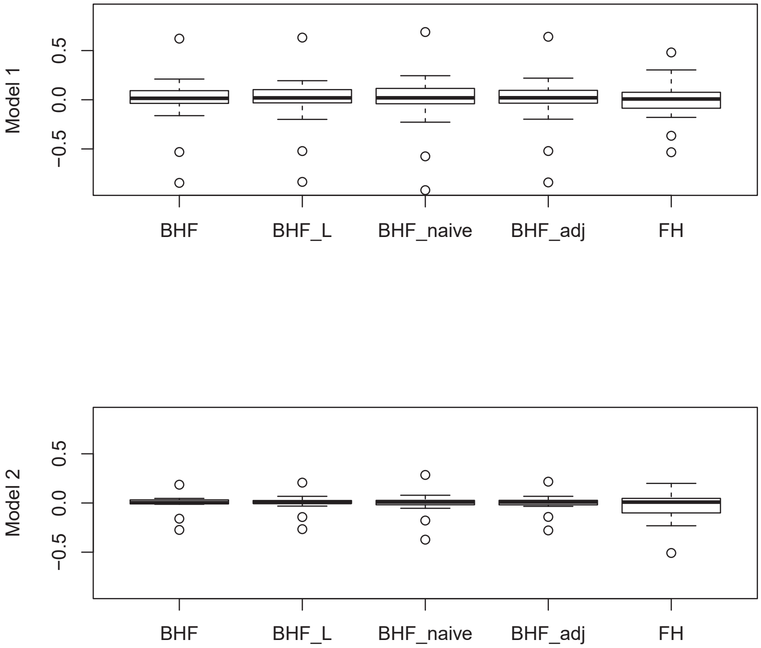

Figure 1: Boxplot of the relative errors for the estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH in the 18 areas

A comparison can be made between unit level and area level estimators when linkage errors are not accurately evaluated. As already discussed in the previous section, in this analysis we know the true value of the linkage errors and use them for the adjustments. However, generally in real cases, assessing linkage errors is not an easy task, the research on the topic is still active, some proposals include Belin and Rubin (1995), Tuoto (2016), and Chipperfield and Chambers (2015). To account for difficulties in assessing linkage errors, we propose a sketch on the behavior of the small area estimators when linkage errors are not accurately evaluated. When linkage errors are underestimated, we tend to make estimates such as the naïve. So, let's focus on the behavior of unit level and area level estimators when linkage errors are overestimated. To overestimate the linkage errors, within each small domain we treat the observed range of false linkage rate as if it were normally distributed, then we evaluate a 95% normality-based confidence interval for $1-\lambda$, and we consider the superior extreme of the confidence intervals as values of $1-\lambda$ in the estimator BHF_adj for the 100 replications.

In this analysis, we only consider the Scenario B, which shows the highest linkage error levels. The boxplot of the values of λ within the 18 areas in the 100 replications is

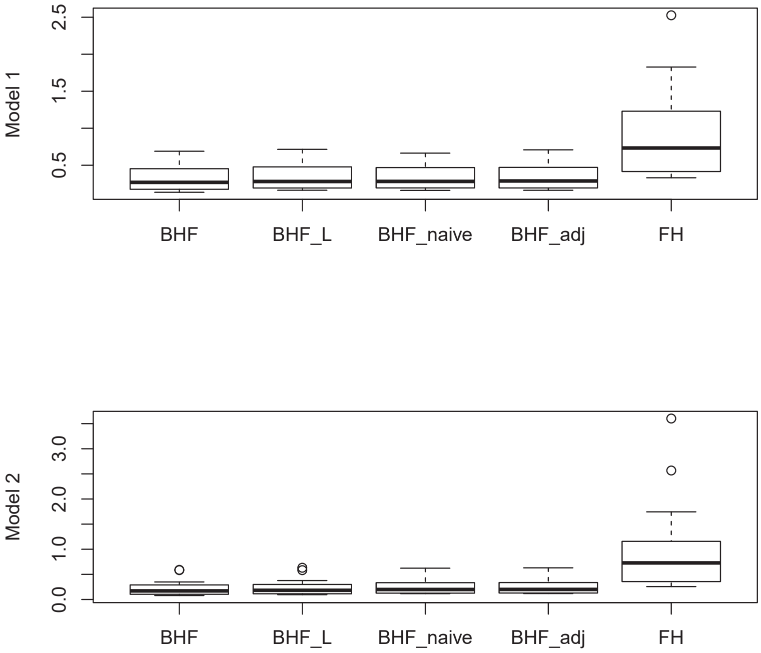

Figure 2: Boxplot of the standard deviations for the estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH in the 18 areas

shown in Figure 3. It is worth noting that the areas with the lowest linkage errors (i.e. area M3 and area M7) are the smallest ones, both in terms of population and sample. No linkage errors in these areas is a realistic assumption, since the small size of the areas avoids false matches.

Table 4 shows the average of the Absolute Relative Error (ARE), the Standard Deviation (SD), and the Mean Square Error (MSE), in the 18 areas, for estimators FH and BHF_adj.

Table 4 confirms the observed behavior and the relationship between area level and unit level estimator, even when the linkage errors are not accurately measured. Still in terms of bias, the FH estimator is preferable to the adjusted BHF estimator in Model 1, whilst the vice-versa in observed for Model 2. In terms of variability, the BHF estimator outperforms the FH estimator in both models.

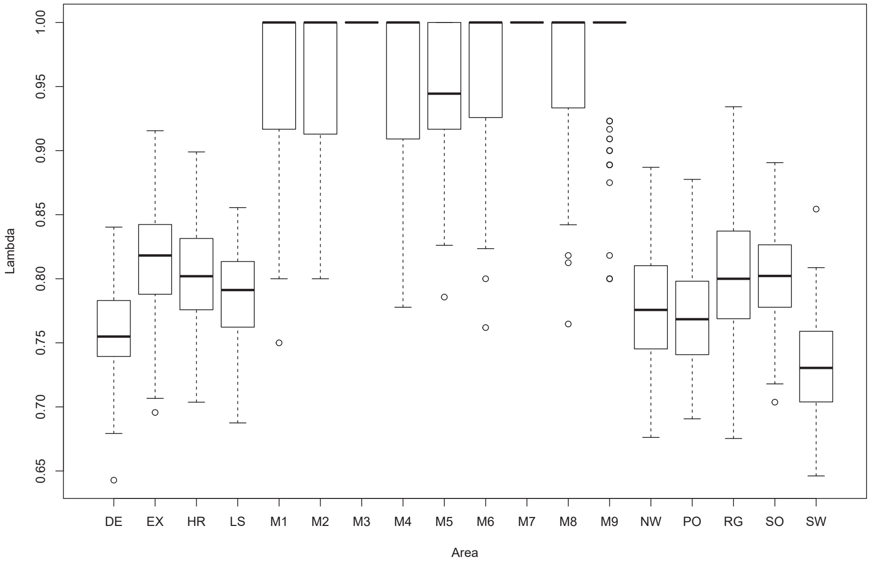

Figure 3: Boxplot of the values of λ in the 100 replications within the 18 small areas

5. Concluding remarks and future works

We explored the behavior of unit level and area level estimators in the presence of linkage errors. The area level is, in principle, quite attractive as it does not require record linkage at all. However, with both realistic and simulated data, we find that the use of auxiliary information at unit level is still useful, even if it exposes to the risk of unit identification errors.

As already noted, the implementation of the area level estimator under the Fay-Herriot model needs reliable smoothed estimates of the sampling variability. We used a simple minded smoothing method, which assumes that the population variances of all the domains are identical. This might be a strong assumption and it might have an

Table 4: Average of the Absolute Relative Error (ARE), Standard Deviation (SD), and Mean Square Error (MSE) for estimators BHF_adj and FH when linkage errors are over-estimated

ARE		SD		MSE		
	BHF_adj	FH	BHF_adj	FH	BHF_adj	FH
Model1	0.0422	0.0401	0.3470	0.9108	0.4009	0.9240
Model2	0.0142	0.0266	0.2554	1.0040	0.2608	1.0064

impact on our results. Further work is needed to improve the variance smoothing for the FH estimator.

In this work, the linkage error adjusted unit level estimator is the one suggested in Di Consiglio and Tuoto (2016) and Briscolini et al. (2018). In the adjustment, we assumed block specific probabilities of correct link are known and this is indeed a strong assumption (see remark 2 (3) of Han and Lahiri, 2018). Moreover, the proposed adjustment assumes the exchangeability of linkage errors, and the small areas coinciding with the blocks of the linkage process. As already noted in Di Consiglio and Tuoto (2016) and in Section 4, the adjustment at unit level does not completely remove the bias introduced by linkage errors. This can be the result of the fact that the exchangeability assumption is not perfectly met.

While our evaluation does not provide a definite answer, we hope our paper encourages others to design an extensive evaluation experiment in order to compare BHF estimator corrected for linkage error with the EBLUP under the Fay-Herriot model that does not require any correction for linkage errors.

In the future, we propose to expand our simulation experiment to include the framework proposed by Han and Lahiri (2018) to correct the unit level small area estimation and to benefit from the use of unit level information to improve estimators, even in the presence of linkage errors. One of the promising advantages of the Han and Lahiri's setting is that it does not require any exchangeability assumption. In Han's dissertation thesis (Han, 2018), she suggests an integrated model where the information about the linkage is carried by all record pairs (links and non-links). In this way all record pairs contribute to the estimation process and to correct for linkage bias. This model is different from the secondary data analysis, adopted in this paper, where only the designated links are considered. More in details, the linkage process is viewed as a permutation of the true covariates associated with the observed target variables within a block/small area. Under the assumption that the random errors and random effects are independent from the observed linked covariates and the comparison matrix of the linkage, given the true covariates values, an Empirical Best Predictor is derived.

Acknowledgments

We wish to thank the Guest Editor-in-Chief Prof. Partha Lahiri for suggesting this research problem and encouraging us to follow up this work.

REFERENCES

BELIN, T., RUBIN, D. B., (1995). A method for calibrating false - match rates in record linkage, *Journal of the American Statistical Association*, 90, pp. 694–707.

BATTESE, G. E., HARTER, R.M., FULLER, W. A., (1988). An Error-Components Model for Prediction of Crop Areas Using Survey and Satellite Data, Journal of the American Statistical Association, 83, pp. 28–36.

BRISCOLINI, D., DI CONSIGLIO, L., LISEO, B., TANCREDI, A., TUOTO, T., (2018). New methods for small area estimation with linkage uncertainty. International Journal of Approximate Reasoning, 94, pp. 30–42.

CHAMBERS, R., (2009). Regression analysis of probability-linked data, Official Statistics Research Series, Vol. 4.

CHIPPERFIELD, J. O., CHAMBERS, R. L., (2015). Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data, Journal of Official Statistics, Vol. 31, No. 3,

DI CONSIGLIO, L., TUOTO, T., (2016). Small Area Estimation in the Presence of Linkage Errors. In International Conference on Soft Methods in Probability and Statistics, pp. 165–172. Springer, Cham.

DI CONSIGLIO, L., TUOTO, T., (2018). When adjusting for the bias due to linkage errors: A sensitivity analysis. Statistical Journal of the IAOS, 34(4), pp. 589–597.

FELLEGI, I. P., SUNTER, A. B., (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64, pp. 1183–1210.

FAY, HERRIOTT, (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association 74, pp. 269–277.

HAN, Y., (2018). Statistical Inference Using Data From Multiple Files Combined Through Record Linkage, PhD Dissertation thesis, downloadable at https://drum.lib.umd.edu/bitstream/handle/1903/21155/HAN_umd_0117E_19360.pdf

HAN, Y., LAHIRI, P., (2018). Statistical analysis with linked data. International Statistical Review, 87, S139–S157.

HARVILLE, D. A., (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of American Statistical Association, 72, pp. 320– 338.

HAWALA, S., LAHIRI, P., (2018). Variance Modeling for Domains. Statistics and Applications, 16, pp. 399– 409.

HERZOG, T. N., SCHEUREN F.J., WINKLER, W. E., (2007). Data Quality and Record Linkage Techniques, Springer Science & Business Media.

JARO, M., (1989). Advances in record linkage methodology as applied to matching the 1985 test census of Tampa, Florida. Journal of American Statistical Association, 84, pp. 414–420.

LAHIRI, P., LARSEN, M. D., (2005). Regression Analysis With Linked Data. Journal of the American Statistical Association, 100, pp. 222–230.

MCLEOD, P., HEASMAN, D. and FORBES, I., (2011). Simulated data for the on the job training, http://www.cros-portal.eu/content/job-training.

NETER, J., MAYNES, E. S, RAMANATHAN, R., (1965). The effect of mismatching on the measurement of response errors. Journal of the American Statistical Association, 60, pp. 1005–1027.

RAO, J. N. K., MOLINA, (2015). Small Area Estimation, Second Edition, Wiley, New York.

RELAIS 3.0 User's Guide, (2015). available at http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-elaborazione/relais.

SEARLE, S. R., CASELLA, G., MCCULLOCH, C. E., (2006). Variance Components, Wiley, New York.

SAMART, K., (2011). Analysis of probabilistically linked data, PhD thesis, School of Mathematics and Applied Statistics, University of Wollongong.

SAMART, K., CHAMBERS, R., (2010). Fitting Linear Mixed Models Using Linked Data, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper pp. 18–10.

SAMART, K., CHAMBERS, R., (2014). Linear regression with nested errors using probability-linked data, Australian and New Zealand Journal of Statistics 56.

SCHEUREN, F., WINKLER, W. E., (1993). Regression analysis of data files that are computer matched – Part I. Survey Methodology, Volume 19, pp. 39–58.

SCHEUREN F., WINKLER W. E., (1997). Regression analysis of data files that are computer matched- part II, Survey Methodology, 23, pp. 157–165.

TANCREDI, A., LISEO, B., (2011) A hierachical Bayesian approach to record linkage and population size problems. Annals of Applied Statistics, 5, pp. 1553–1585.

TUOTO, T., (2016). New proposal for linkage error estimation" Statistical Journal of the IAOS, Vol 32, no. 2, pp. 1–8.