Received: 18 December 2019 | Accepted: 26 August 2020

DOI: 10.1111/rssb.12401

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

Small area estimation with linked data

N. Salvati[1] | **E. Fabrizi**[2] | **M. G. Ranalli**[3] | **R. L. Chambers**[4]

[1]Dipartimento di Economia e Management, Università di Pisa, Pisa, Italy

[2]Dipartimento di Scienze Economiche e Sociali, Università Cattolica del Sacro Cuore, Milan, Italy

[3]Dipartimento di Scienze Politiche, Università degli Studi di Perugia, Perugia, Italy

[4]National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia

Correspondence
N. Salvati, Dipartimento di Economia e Management, Università di Pisa, Pisa, Italy.
Email: nicola.salvati@unipi.it

Abstract

Data linkage can be used to combine values of the variable of interest from a national survey with values of auxiliary variables obtained from another source, such as a population register, for use in small area estimation. However, linkage errors can induce bias when fitting regression models; moreover, they can create non-representative outliers in the linked data in addition to the presence of potential representative outliers. In this paper, we adopt a secondary analyst's point of view, assuming that limited information is available on the linkage process, and develop small area estimators based on linear mixed models and M-quantile models to accommodate linked data containing a mix of both types of outliers. We illustrate the properties of these small area estimators, as well as estimators of their mean squared error, by means of model-based and design-based simulation experiments. We further illustrate the proposed methodology by applying it to linked data from the European Survey on Income and Living Conditions and the Italian integrated archive of economic and demographic micro data in order to obtain estimates of the average equivalised income for labour market areas in central Italy.

KEYWORDS

exchangeable linkage error, finite population inference, linear mixed models, mean squared error estimation, robust estimation

1 | INTRODUCTION

Estimates of finite population parameters are often needed for subsets (domains) of the population, defined either by geographical disaggregation (areas) or by other classification criteria (e.g. region by

gender by age class). When the domain-specific portion of the available data is so small that standard estimators are unacceptably imprecise for most of the domains, we have a small area estimation (SAE) problem. See Pfeffermann (2013) and Rao and Molina (2015) for general introductions to the topic. From now on we refer to the domains of interest as areas.

Small area estimation methods complement available data, typically from a large population survey, with area specific auxiliary information. A standard setting is where it is reasonable to assume that the value y_{ij} of the target variable for unit j in area i is related to a known vector of covariates $\mathbf{x}_{ij}$ by means of a regression model. These $\mathbf{x}_{ij}$ values, assumed to be known for both the survey sample and the rest of the population, are then used to predict the area parameter of interest.

Data integration is fast becoming an intrinsic part of Official Statistics, in large part due to the increasing availability of administrative registers and other population data sources. Here we focus on the situation where the y_{ij} values are measured in a sample survey while the $\mathbf{x}_{ij}$ are from a population register that can be linked to either the population frame from which the survey sample is drawn or directly to the selected sample. If an error-free unique identifier exists in both the survey record and the population register, this linkage is unremarkable. However, in many cases, such an identifier is not error free or does not exist, in which case we need to allow for record linkage errors.

Due to its growing importance, record linkage has attracted considerable scientific interest. Broadly speaking, we can identify two main literature streams: the first is concerned with how to link records when an error-free unique identifier is missing; the second is focused on how to adjust statistical methods so that they are appropriate for the analysis of linked data containing linkage errors. We place ourselves in the latter stream of research. For recent reviews of the first literature stream see Winkler (2009, 2014), and Han and Lahiri (2018).

It is widely recognised that overlooking linkage errors when analysing linked data can lead to biased estimates even if most records are correctly linked. Bias correction methods when fitting linear regression models to linked data are discussed in Scheuren and Winkler (1993, 1997), Lahiri and Larsen (2005), Chambers (2009), Kim and Chambers (2012), Han and Lahiri (2018), Zhang and Chambers (2019) and Chambers and Diniz da Silva (2020). The impact of linkage errors on linear mixed models, which are often used in small area estimation, has received comparatively less attention (Samart and Chambers, 2014). Lahiri and Han (2017) consider probabilistically linked data in the context of SAE but focus on simple linear regression models; to the best of our knowledge, Briscolini et al. (2018) and Han (2018) are the only attempts to use mixed models with linked data for SAE.

There is a further aspect to linkage errors that seems to have attracted much less attention: this is when linkage errors generate artificial outliers in the linked dataset. Let y^*_{ij} denote the linked value corresponding to y_{ij}. An outlier can then be generated when there is linkage error as the residual associated with the correctly linked pair $(y_{ij}, \mathbf{x}_{ij})$ is small, while the residual associated with the incorrectly linked pair $(y^*_{ij}, \mathbf{x}_{ij})$ is large. This can happen when the variables used in the matching process (such as name, address, identification code) are independent of those used as regressors. In Figure 1, we illustrate this phenomenon using a synthetic population, which is described in more detail in Section S.4 of the Supplementary Material. In panel (a), a scatterplot shows the strong linear relationship between y and x when there are no linkage errors; in panel (b) the same relationship is shown when linkage errors occur at an overall rate of 28.5 per cent. In panel (b), incorrectly matched pairs (y^*_{ij}, x_{ij}) are shown as open circles. Outlying residuals are evident.

Chambers (1986) first distinguishes between representative and non-representative outliers in a survey sample. Using this distinction, these artificial outliers are non-representative, and so are fundamentally different from outliers associated with the correctly linked population units, which are representative. Unfortunately it is not possible to tell *a priori* whether an outlier is induced by linkage error (and so is non-representative) or whether it is representative. In particular, outliers due to linkage errors can violate the assumptions underpinning non-robust estimation methods, as well as cause

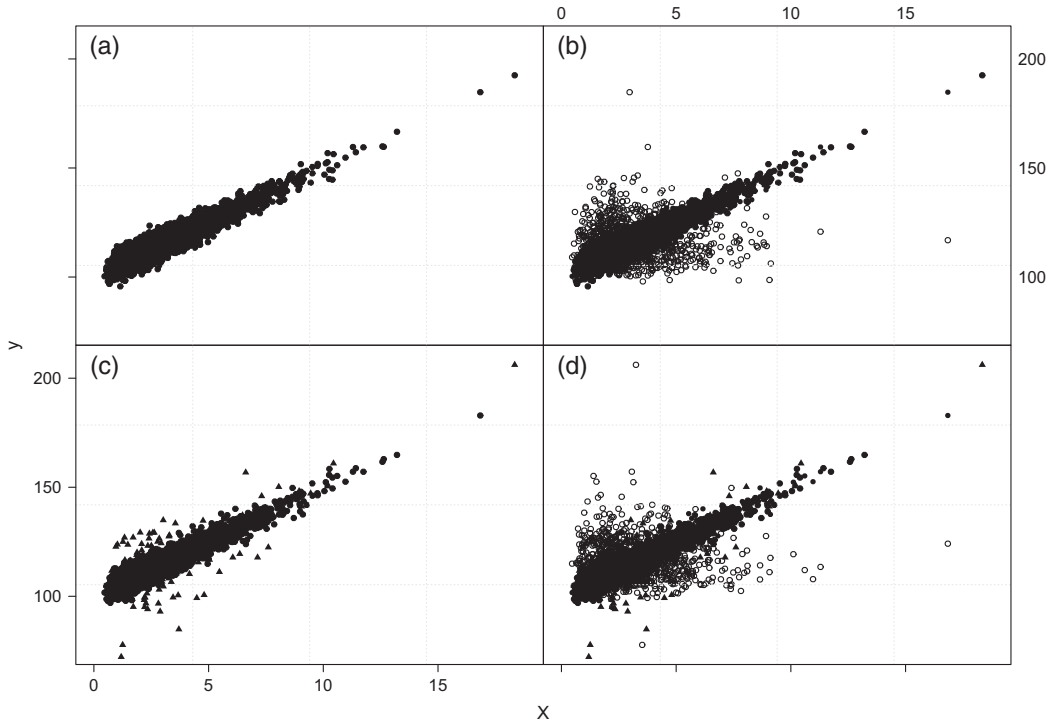

FIGURE 1 Scatterplot of four synthetic populations. Panel (a) shows the relationship between x and y when there are no linkage errors (filled circles). Panel (b) shows the same relationship when wrongly matched pairs (y^*_{ij}, x_{ij}) occur (circles). Panel (c) shows the relationship between x and y when there are representative outliers (filled triangles point-up). Panel (d) shows the same relationship when wrongly matched pairs (y^*_{ij}, x_{ij}) also occur (circles)

bias problems for robust projective estimation methods (Chambers et al., 2014) since down-weighting linkage error-induced outliers does not generally remove bias due to linkage errors. This problem is illustrated in the lower panels of Figure 1, where scatterplots depict an outlier-prone version of the same synthetic population underpinning in the upper panels. Here, panel (c) shows the relationship between x and y when there are representative outliers (denoted by filled triangles, point-up), whereas the scatterplot in panel (d) shows the same population when linkage errors lead to wrongly matched pairs (y^*_{ij}, x_{ij}) (denoted by open circles). It is difficult to distinguish the residuals due to representative outliers from those due to linkage errors in this panel.

Non-representative outliers inflate the variance of all small area predictors; moreover they can have an impact on bias of robust-projective predictors. Representative outliers, provided their distribution is approximately symmetric, have an impact on the variance of all predictors but not on their bias (Chambers et al., 2014; Jiongo et al., 2013; Sinha & Rao, 2009). In other words, such representative outlying values represent a stability problem for small area predictors, whereas linkage error-induced outliers can create a problem of both validity and stability for small area predictors.

Two questions immediately arise when looking at Figure 1. The first is whether well-known outlier robust methods for small area estimation can adequately deal with the mix of representative and non-representative outliers that can potentially occur in a linked data situation. The second concerns appropriate modifications to these methods to allow for linkage errors. We address both in this paper. In particular, we extend some popular small area estimators based on linear models, including the Empirical Best Linear Unbiased Predictor (EBLUP, Battese et al., 1988), its robust version (REBLUP,

Sinha & Rao, 2009), and the M-quantile-based predictor (Chambers & Tzavidis, 2006; Tzavidis et al., 2010) to non-deterministically linked data. We also propose analytical MSE estimators for the modified estimators that we introduce.

In our development, we adopt a secondary analyst viewpoint, that is, we assume that the researcher producing the small area estimates does not have access to all the information used in the linkage process. Instead, we assume that he/she has access to the linked dataset (including area indicators, which are assumed to be without error) and is also provided with minimal information regarding the linkage quality. We characterise this minimal information via a simple exchangeable linkage error structure within specified poststrata, referred to as blocks below. In contrast, Lahiri and Han (2017), Briscolini et al. (2018), and Han (2018) take a primary analyst viewpoint and so assume that a richer set of information on the linkage process is available; although not directly concerned with SAE, the same comment applies to Han and Lahiri (2018).

The paper is organised as follows. Section 2 is devoted to setting out the theoretical background and the assumptions on the linkage error model which is then used to extend the small area predictors. In Sections 3–5 we introduce the proposed extensions to the EBLUP, REBLUP and M-quantile-based predictors under linkage error and their corresponding MSE estimators respectively. In Section 6 the performance of these newly proposed predictors are empirically assessed, both in terms of point estimation performance as well as in terms of MSE estimation, by a design-based simulation. The performance of the predictors has also been evaluated by means of a model-based simulation study that considers a number of different scenarios. For reason of space these results are reported in Section S.5 of the Supplementary Material. In Section 7 we then apply these SAE methodologies to the estimation of equivalised income averages for 113 labour market areas (LMAs) in central Italy. LMAs are unplanned domains defined as clusters of municipalities, with many containing no sampled households. The estimates are obtained using 2016 EU-SILC survey data for Italy linked to the Italian integrated archive of economic and demographic microdata. Here, the only information about linkage accuracy provided to the user is the overall probability of correct linkage. Finally, in Section 8 we summarise our main findings and provide directions for future research.

2 | BACKGROUND AND ASSUMPTIONS

For simplicity of exposition, we restrict our development to the case of two registers. Extension to the case of multiple linked registers can be carried out along the same lines set out in Kim and Chambers (2015). In particular, we consider a situation where sample data on a target variable y are available from a survey, while auxiliary information on a vector of variables $\mathbf{x}$ comes from a register $\mathcal{X}$. Based on our secondary analyst viewpoint, we formalise this scenario by making the following assumptions:

(i) the linkage process is non-informative, that is, the probability of correct linkage does not depend on the values of y and on the realised sample given the auxiliary information;
(ii) the survey sample is randomly drawn from a population frame or register $\mathcal{Y}$, and sampling is non-informative, in the sense that it does not depend on the values of y given the auxiliary information;
(iii) the $\mathcal{Y}$ frame/register and the $\mathcal{X}$ register contain no duplicates, correspond to the same target population U of size N, and include a shared set of unit identifiers (linking variables) so that they can, in principle, be matched on a one to one basis. However, the linking variables are not unique identifiers, and linkage errors are possible.

Given these assumptions, the process of sampling from $\mathcal{Y}$ and then linking to register $\mathcal{X}$ is stochastically equivalent to the complete and one to one linkage of $\mathcal{X}$ and $\mathcal{Y}$, which essentially creates a unique linked register, followed by sampling from this linked register. In order to see why this equivalence holds, note that after collection of the sample information, the $\mathcal{Y}$ frame/register contains two types of variables: a set of variables, which we can label as attributes and include y, that are observed only on sampled units and the set of identifier variables to be used in linking the $\mathcal{Y}$ and $\mathcal{X}$ registers, that are known for all units in $\mathcal{Y}$. Attribute variables are not involved in the linking process. This does not necessarily mean that all units in the $\mathcal{Y}$ and $\mathcal{X}$ registers are linked, but register level information from both registers is available, so one to one and complete linkage of $\mathcal{X}$ and $\mathcal{Y}$ is in principle possible. Since sampling and linkage are two distinct processes, the equivalence noted previously in this paragraph follows.

The assumptions we make above can be viewed as strong, and we do not claim that they apply in all situations. However, we emphasise that they represent a reasonable formalisation of the problem from a secondary analyst perspective. Such an analyst is typically not from the organisation responsible for creating the linked dataset, and confidentiality protocols will usually preclude access to detailed information about the linkage process that could be included in the SAE model. These assumptions are also consistent with the common practice of outsourcing the linking process to another body, often referred to as a trusted third party, commissioned by the registers' owners to implement the linkage of the two registers using the linking information available on both. Trusted third parties are often used when identifiable linked data cannot be released to analysts, so there is a need to keep identifiers separate from attribute data (sometimes referred to as the separation principle, Kelman et al., 2002). A sample survey carried out on one of the registers can then be linked to the other register via the linking identifiers (match keys) created by the linking organisation, and the information from the non-sampled register combined with the sample information from the sampled register to create the linked sample data (Harron, 2016). Finally, these assumptions reflect how data linkage is carried out by National Statistical Offices in Europe and Australia (Abbott et al., 2016; Dygaszewicz, 2012; Garofalo, 2014; Schulte Nordholt, 2009; Swiss Federal Statistical Office, 2012). For example, they are consistent with the linking procedure adopted by the Italian National Statistics Office (ISTAT), and which underpins the data used in the application of Section 7. We also recognise that in some situations, linkage can depend on the sample design, for example, when only sampled identifiers are used in linkage, though it is unlikely that a secondary analyst would have access to such information. Nevertheless, in Section 6, we use simulations to evaluate the robustness of the methods described in this paper when linkage depends on the sampled units.

The data available to secondary analysts are somewhat limited, as they are not involved in the linking process, although some information about the accuracy of the linkage process may still be available. In order to allow for heterogeneity of linkage errors, we assume that the analyst has access to identifiers that allow each register to be partitioned into Q non-overlapping subsets or blocks such that linkage error probabilities are homogeneous within a block and possibly heterogeneous between blocks. The block identifiers themselves are assumed to depend on one or more linking variables measured without error and so are known to the analyst. We also assume that the target population U can be partitioned into D non-overlapping areas or domains, and that linkage is carried out within each area, so two population units from different areas cannot be erroneously matched. In effect, the known domain identifier is one of the linking variables. Cross-classifying U by area and block indicators, we then define U_{iq} to be the subset of N_{iq} population units that make up the segment of area i nested within block q, with $i = 1, \ldots, D$ and $q = 1, \ldots, Q$. We use $\mathbf{x}_{iqj}$ and y_{iqk} to denote individual population values from the two registers associated within the iq cell.

These assumptions now allow us to treat the sample as drawn from a linked register that has a one to one correspondence with the units making up U. To simplify notation (see Chambers, 2009; Chambers & Diniz da Silva, 2020; Kim & Chambers, 2012; Samart & Chambers, 2014), we shall assume that sampling is from the records in register $\mathcal{X}$ containing values $\mathbf{x}_{iqj}$ and that the sampled records are then combined with matched records from frame/register $\mathcal{Y}$ containing values y_{iqk}. Although this may seem the opposite of the situation that is described in the previous section, it is easy to see that since linkage is one to one, the only difference between sampling from $\mathcal{X}$ and augmenting the sampled records with linked data from $\mathcal{Y}$, and sampling from $\mathcal{Y}$ and augmenting the sampled records with linked data from $\mathcal{X}$ is the assumed 'correct' ordering of the units in U. What is important in both cases is that the frame/register $\mathcal{Y}$, with its area and block identifiers, is not available to the analyst. The linked sample data that are available, however, will be assumed to include area and block identifiers, together with any auxiliary information that can be extracted from $\mathcal{X}$, including, at a minimum, area by block-specific summary data for both the linkage and analysis model covariates. In what follows, we will always condition on the availability of this auxiliary information when evaluating moments of random variables.

Let s_{iq} denote the set corresponding to the n_{iq} population indexes of the sample units in small area i and block q, with $n = \sum_{i=1}^{D} \sum_{q=1}^{Q} n_{iq}$. The set containing the $N_{iq} - n_{iq}$ indices of the non-sampled units in small area i and block q is denoted by r_{iq}. For ease of notation, we assume that all areas are sampled, noting that non-sampled areas are easily accommodated. Let $\mathbf{y}_{iq}$ denote the N_{iq} vector of values for y_{iqk} in U_{iq}, with $\mathbf{X}_{iq}$ denoting the $N_{iq} \times p$ matrix with rows defined by the $\mathbf{x}_{iqj}$ values of the corresponding population units. The sample components of these quantities are then denoted by $\mathbf{y}_{siq}$ and $\mathbf{X}_{siq}$, respectively. As noted in the previous paragraph, assumptions (*i*)–(*iii*) introduced at the beginning of this section allow us to assume that the 'correct' ordering of the population is that of register $\mathcal{X}$. As a consequence, unless the linkage is perfect, $\mathbf{y}_{siq}$ is unknown. What we observe is a sample from the vector $\mathbf{y}_{iq}^{*}$ containing the values y_{iqj}^{*} generated by the linkage process. Given that both registers can be matched on a one to one basis, we characterise the relationship between $\mathbf{y}_{iq}$ and $\mathbf{y}_{iq}^{*}$ via a latent random permutation matrix $\mathbf{A}_{iq} = [a_{jk}^{iq}]$ of order N_{iq}, whose row sums and column sums are equal to 1. That is, we define

$$\mathbf{y}_{iq}^{*} = \mathbf{A}_{iq}\mathbf{y}_{iq}. \tag{1}$$

The distribution of linkage errors in U_{iq} is then determined by the distribution of $\mathbf{A}_{iq}$. In general, this distribution depends on all the information used in the linking process, which, as already noted, is typically unavailable. However, a minimal amount of information about the accuracy of the linkage may be available, in which case, a secondary analyst should be able to model the linkage errors within U_{iq} by combining an assumption of conditional independence of $\mathbf{A}_{iq}$ and $\mathbf{y}_{iq}$ given $\mathbf{X}_{iq}$ with the simple exchangeable linkage error (ELE) specification:

$$Pr(\text{correct linkage}) = Pr(a_{jj}^{iq} = 1) = \lambda_{iq} \tag{2}$$

$$Pr(\text{incorrect linkage}) = Pr(a_{jk}^{iq} = 1) = \gamma_{iq} = \frac{1 - \lambda_{iq}}{N_{iq} - 1}, \tag{3}$$

with $j, k = 1, \cdots, N_{iq}$. Given that we know which block is being referred to, the probability of correct linkage λ_{iq} is the same irrespective of the values of $\mathbf{X}_{iq}$ and $\mathbf{y}_{iq}$. As a consequence,

$$E_{\mathbf{A}}(\mathbf{A}_{iq}) = \mathbf{T}_{iq} = (\lambda_{iq} - \gamma_{iq})\mathbf{I}_{N_{iq}} + \gamma_{iq}\mathbf{1}_{N_{iq}}\mathbf{1}'_{N_{iq}},$$

where $\mathbf{I}_{N_{iq}}$ denotes the identity matrix of order N_{iq}, $\mathbf{1}_{N_{iq}}$ denotes a vector of ones of length N_{iq} and $E_{\mathbf{A}}(.)$ denotes expectation with respect to the linkage error model.

Let $E_M(.)$ denote expectation with respect to the model for $\mathbf{y}_{iq}$. The conditional independence assumption referred to just before (2) then allows us to write

$$E_{\mathbf{A},M}(\mathbf{A}_{iq}\mathbf{y}_{iq}) = E_{\mathbf{A}}(\mathbf{A}_{iq})E_M(\mathbf{y}_{iq}) = \mathbf{T}_{iq}E_M(\mathbf{y}_{iq}), \tag{4}$$

where $E_{\mathbf{A},M}$ denotes joint expectation with respect to the linkage error model and the model for the distribution of $\mathbf{y}_{iq}$. Without loss of generality, we partition the matrix $\mathbf{A}_{iq}$ as

$$\mathbf{A}_{iq} = \begin{bmatrix} \mathbf{A}_{siq} \\ \mathbf{A}_{riq} \end{bmatrix},$$

where $\mathbf{A}_{siq}$ is a $n_{siq} \times N_{iq}$ matrix and $\mathbf{A}_{riq}$ is a $(N_{iq} - n_{iq}) \times N_{iq}$ matrix. These matrices contain the rows of $\mathbf{A}_{iq}$ corresponding to sampled and non-sampled units, respectively. Then $\mathbf{y}^*_{siq}$ and $\mathbf{X}_{iq}$ are available to the analyst, while $\mathbf{y}_{siq}$ is not observed, where

$$\mathbf{y}^*_{siq} = \mathbf{A}_{siq}\mathbf{y}_{iq}. \tag{5}$$

As noted above, the matrix $\mathbf{A}_{siq}$ is not observable, but under the ELE assumptions (2) and (3) we have that

$$E_{\mathbf{A}}(\mathbf{A}_{siq}) = \mathbf{T}_{siq} = [(\lambda_{iq} - \gamma_{iq})\mathbf{I}_{n_{iq}} \mid \mathbf{0}_{r_{iq}}] + \gamma_{iq}\mathbf{1}_{n_{iq}}\mathbf{1}'_{N_{iq}}, \tag{6}$$

where $\mathbf{0}_{r_{iq}}$ is a $n_{iq} \times (N_{iq} - n_{iq})$ matrix of zeroes, $\mathbf{I}_{n_{iq}}$ denotes the identity matrix of order n_{iq}, and $\mathbf{1}_{n_{iq}}$ denotes a vector of ones of length n_{iq}.

Unless stated otherwise, we treat the parameter λ_{iq} as known (as is then the matrix $\mathbf{T}_{siq}$), with its value communicated to the secondary analyst by the linking agency as part of the linkage *paradata* (Gilbert et al., 2018). In practice, it is likely that linkage error quality will be assessed using a clerical audit of a sample of linked pairs or via some other measure of the accuracy of the achieved linkage (see e.g. Chambers & Diniz da Silva, 2020). It is unlikely that individual λ_{iq} values will be passed on to the secondary analyst: it is more likely that block-specific λ_q values will be made available, or even a single overall average value λ as in the application described in Section 7. In any case, when linked data are released to the user, the small areas of interest may not be known to the linkage agency and/or the secondary analyst may want to use the data to obtain estimates for different sets of small areas. We note that Briscolini et al. (2018) assume that the area indicator is the only blocking variable, so that the probabilities of correct linkage are area-specific (the authors use λ_i instead of λ_{iq}). All these settings fit into our framework. We address the issue of accounting for uncertainty due to estimation of λ_{iq} in Section 8. Finally, we shall assume that the sampled rows $\mathbf{X}_{siq}$ and the column means $\bar{\mathbf{x}}_{iq}$ of $\mathbf{X}_{iq}$ are known. It immediately follows that the matrix $\mathbf{X}^*_{siq}$ defined by

$$\mathbf{X}^*_{siq} = E_{\mathbf{A}}(\mathbf{A}_{siq}\mathbf{X}_{iq}) = \mathbf{T}_{siq}\mathbf{X}_{iq} = \{(\lambda_{iq} - \gamma_{iq})\mathbf{X}_{siq} + \gamma_{iq}N_{iq}\mathbf{1}_{n_{iq}}\bar{\mathbf{x}}'_{iq}\} \tag{7}$$

can be treated as known when λ_{iq} is known.

3 | LINEAR MIXED MODELS FOR SMALL AREA ESTIMATION WITH LINKED DATA

Linear mixed models for population unit data are widely used for SAE. These models include area-specific random effects that are used to characterise area-level heterogeneity in the model residuals. See Battese et al. (1988) for an early example of their application. A general specification for a unit level linear mixed model used in SAE is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{8}$$

where $\mathbf{y}$ is the vector of the N population values of the response variable, and $\mathbf{X}$,$\mathbf{Z}$ are known matrices, of dimension $N\times p$ and $N\times Dm$ respectively, containing the corresponding population level values of the p covariates used to model unit level heterogeneity and the m covariates used to model area-level heterogeneity (Section 4.5.3 in Haslett, 2016; Rao & Molina, 2015); $\mathbf{u} = (\mathbf{u}_1', \cdots, \mathbf{u}_D')$ is a vector of length Dm made up of D independent realisations $\{\mathbf{u}_i; i = 1, \cdots, D\}$ of an m-dimensional random area effect with $\mathbf{u} \sim N(\mathbf{0}, \Sigma_u)$ and $\mathbf{e} \sim N(\mathbf{0}, \Sigma_e)$ is the vector of individual errors. The random effects $\mathbf{u}$ and the individual errors $\mathbf{e}$ are assumed to be independent, so the covariance matrix of $\mathbf{y}$ is $\Sigma(\boldsymbol{\delta}) = \Sigma_e + \mathbf{Z}\Sigma_u\mathbf{Z}'$. Here D is the total number of small areas that make up the population, and we assume that the m covariates defining $\mathbf{Z}$ are contextual and so do not vary within an area. That is, $\mathbf{Z} = \text{diag}(\mathbf{Z}_i)$, where $\mathbf{Z}_i$ is the $N_i \times m$ matrix with rows equal to the area i values $\mathbf{z}_i$ of these covariates. Note that there is no restriction on the contextual variables defining $\mathbf{Z}$ also being included in the set of variables defining $\mathbf{X}$ (Section 4.5.3 in Rao & Molina, 2015).

We also assume that the covariance matrices Σ_u and Σ_e are defined in terms of a lower dimensional set of parameters $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_K)$, which are typically referred to as the variance components of model (8), whereas the vector $\boldsymbol{\beta}$ stands for the $p\times 1$ vector of regression coefficients. In particular, we shall assume $\Sigma_e = \sigma_e^2\text{diag}(\mathbf{w})$ where $\sigma_e > 0$ is an unknown scale parameter and $\mathbf{w}$ is a vector of order N of strictly positive values that are known functions of the values in $\mathbf{X}$. This allows for heteroskedasticity in individual residuals, an assumption often needed in small area estimation and especially in poverty mapping (Section 7.6.1 in Das et al., 2017; Molina & Rao, 2010; Rao & Molina, 2015).

Provided that it is reasonable to assume that the distribution of the unit level residuals in $\mathbf{e}$ remains the same from block to block (i.e. the blocking variables are independent of the response variable given auxiliary information), then at the U_{iq} level, (8) can be written as

$$\mathbf{y}_{iq} = \mathbf{X}_{iq}\boldsymbol{\beta} + \mathbf{Z}_{iq}\mathbf{u}_i + \mathbf{e}_{iq}, \tag{9}$$

where $\mathbf{Z}_{iq}$ is the $N_{iq} \times m$ incidence matrix defined by the rows of $\mathbf{Z}$ corresponding to area i units in block q.

Now suppose that linked data are used to define the sample values of the response variable within sub-population U_{iq}. In view of (1) and (5), we can consider the following model for the linked sample values $\mathbf{y}^*_{siq}$ that takes into account the linkage error process:

$$\mathbf{y}^*_{siq} = \mathbf{A}_{siq}\mathbf{y}_{iq} = \mathbf{A}_{siq}\mathbf{X}_{iq}\boldsymbol{\beta} + \mathbf{A}_{siq}\mathbf{Z}_{iq}\mathbf{u}_i + \mathbf{e}^*_{siq}, \tag{10}$$

where $E_{\mathbf{A},M}(\mathbf{e}^*_{siq}) = \mathbf{0}$ and $V_{\mathbf{A},M}(\mathbf{e}^*_{siq}) = \Sigma^*_{seiq} = \sigma_e^2\text{diag}(\mathbf{T}_{siq}\mathbf{w}_{iq})$. Here, $\mathbf{w}_{iq}$ denotes the component of vector $\mathbf{w}$ for area i and block q, $E_{\mathbf{A},M}$ and $V_{\mathbf{A},M}$ denote the joint expectation and variance with respect to the linkage error model and the linear mixed model. Note that the expression for Σ^*_{seiq} above follows from the fact that $E_{\mathbf{A},M}(\mathbf{e}^*_{siq}) = E_{\mathbf{A}}(\mathbf{A}_{siq})E_M(\mathbf{e}_{iq}) = \mathbf{T}_{siq}\mathbf{0} = \mathbf{0}$ so we can write $V_{\mathbf{A},M}(\mathbf{e}^*_{siq}) = E_{\mathbf{A}}\left(\mathbf{A}_{siq}E_M(\mathbf{e}_{iq}\mathbf{e}_{iq}'|\mathbf{A}_{siq})\mathbf{A}_{siq}'\right) = \sigma_e^2E_{\mathbf{A}}(\mathbf{A}_{siq}\text{diag}(\mathbf{w}_{iq})\mathbf{A}_{siq}') = \sigma_e^2\text{diag}\left\{E_{\mathbf{A}}(\mathbf{A}_{siq})\mathbf{w}_{iq}\right\} = \sigma_e^2\text{diag}(\mathbf{T}_{siq}\mathbf{w}_{iq})$.

In most of the SAE literature, $\Sigma_e = \sigma_e^2 \mathbf{I}_N$ so $\Sigma^*_{seiq} = \Sigma_{seiq}$, that is, the sampled rows and columns of the area i by block q component of Σ_{eiq}. Let $\mathbf{Z}_{siq}$ denote the sampled rows of $\mathbf{Z}_{iq}$. Then, since matching across areas is impossible, and since the values in each column of $\mathbf{Z}_{iq}$ are identical, $\mathbf{A}_{iq}\mathbf{Z}_{iq} = \mathbf{Z}_{iq}$ and so $\mathbf{A}_{siq}\mathbf{Z}_{iq} = \mathbf{Z}_{siq}$. As a consequence, we can write

$$E_{\mathbf{A},M}(\mathbf{y}^*_{siq}) = \mathbf{X}^*_{siq}\boldsymbol{\beta}, \tag{11}$$

and

$$V_{\mathbf{A},M}(\mathbf{y}^*_{siq}) = \Sigma_{siq} = \Sigma_{siq}(\boldsymbol{\delta}, \boldsymbol{\beta}) = \mathbf{Z}_{siq}\Sigma_{\mathbf{u}_i}\mathbf{Z}'_{siq} + \Sigma^*_{seiq} + \mathbf{V}_{siq}. \tag{12}$$

See Section S.1 in Supplementary Material for details concerning the derivation of Equation (12). Here $\Sigma_{siq}(\boldsymbol{\delta}, \boldsymbol{\beta})$ is the (iq) component of the block diagonal covariance matrix of $\mathbf{y}^*_s$. The matrices $\Sigma_{\mathbf{u}_i}$ and Σ^*_{seiq} depend only on the variance components $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_K)\, \forall i, q$, that are not area or block specific, that is, do not change across areas or across blocks. The last component of the covariance matrix, $\mathbf{V}_{siq}$, is defined in terms of the regression coefficients $\boldsymbol{\beta}$. An exact expression for $\mathbf{V}_{siq}$ is unavailable. However, using the arguments set out in Appendix 1 of Chambers (2009), we can write down the approximation

$$\mathbf{V}_{siq} \approx \text{diag}(v_{siq}) = \text{diag}\left[(1-\lambda_q)\left\{\lambda_q(f_{iqj} - \bar{f}_{siq})^2 + \bar{f}^{(2)}_{siq} - \bar{f}^2_{siq}\right\}\right], j = 1, \cdots, n_{iq}, \tag{13}$$

where $\mathbf{f}_{siq} = \{f_{iqj}\} = \mathbf{x}'_{iqj}\boldsymbol{\beta}$ and $\bar{f}_{siq}$, $\bar{f}^{(2)}_{siq}$ denote the block q averages of the components of $\mathbf{f}_{siq}$ and their squares, respectively. Note that in the case of the widely used random intercepts model, $\boldsymbol{\delta} = (\sigma_u^2, \sigma_e^2)$ and $\Sigma_{siq}(\sigma_u^2, \sigma_e^2, \boldsymbol{\beta}) = \sigma_u^2 \mathbf{1}_{n_{iq}}\mathbf{1}'_{n_{iq}} + \sigma_e^2 \mathbf{I}_{n_{iq}} + \mathbf{V}_{siq}$.

Given (11) and (12), and assuming that $\Sigma_{\mathbf{u}_i}$ and Σ^*_{seiq} are known $\forall i \in D$ and $q \in Q$, with $\mathbf{V}_{siq}$ approximated by (13), the optimal unbiased estimating function for $\boldsymbol{\beta}$ based on the linked sample values $\mathbf{y}^*_{siq}$ is $\mathbf{H}(\boldsymbol{\beta}_0) = \sum_{i=1}^{D}\sum_{q\in i}\mathbf{X}^{*\prime}_{siq}\Sigma^{-1}_{siq}\left(\mathbf{y}^*_{siq} - \mathbf{X}^*_{siq}\boldsymbol{\beta}_0\right)$, where $\boldsymbol{\beta}_0$ is the 'true' value of $\boldsymbol{\beta}$ and the notation $q \in i$ is used to allow for the case where different numbers of blocks are represented in different areas. The solution to $\mathbf{H}(\boldsymbol{\beta}) = \mathbf{0}$ is then the feasible generalised least squares estimator of $\boldsymbol{\beta}$,

$$\tilde{\boldsymbol{\beta}}^* = \left(\sum_{i=1}^{D}\sum_{q\in i}\mathbf{X}^{*\prime}_{siq}\tilde{\Sigma}^{-1}_{siq}\mathbf{X}^*_{siq}\right)^{-1}\left(\sum_{i=1}^{D}\sum_{q\in i}\mathbf{X}^{*\prime}_{siq}\tilde{\Sigma}^{-1}_{siq}\mathbf{y}^*_{siq}\right), \tag{14}$$

where $\tilde{\Sigma}_{siq}$ denotes the value of Equations (12) at (14). The consistency and asymptotic efficiency properties of Equation (14) are well known, with Bera et al. (2006) providing an extensive discussion. Computation of Equation (14) is carried out iteratively, with an initial estimate of $\boldsymbol{\beta}$ used to compute an estimate of $\mathbf{V}_{siq}$ based on (13). This is substituted into (14) to obtain an updated estimate of $\boldsymbol{\beta}$, which is then used to update the estimate for $\mathbf{V}_{siq}$. This process continues until convergence.

We refer to (14) as the pseudo-Best Linear Unbiased Estimator (pseudo-BLUE) since it is the standard BLUE for $\boldsymbol{\beta}$ when $\tilde{\Sigma}_{siq}$ is replaced by Σ_{siq} and there are no linkage errors. Following this analogy, we define the pseudo-Best Linear Unbiased Predictor (pseudo-BLUP) of $\mathbf{u}_i$ as

$$\tilde{\mathbf{u}}^*_i = \sum_{q\in i}\Sigma_{\mathbf{u}_i}\mathbf{Z}'_{siq}\tilde{\Sigma}^{-1}_{siq}\left(\mathbf{y}^*_{siq} - \mathbf{X}^*_{siq}\tilde{\boldsymbol{\beta}}^*\right). \tag{15}$$

When there are no linkage errors, (15) is the usual BLUP of $\mathbf{u}_i$ where $\tilde{\Sigma}_{siq}$ is replaced by Σ_{siq}.

From (1) it is straightforward to see that the sum of the values in $\mathbf{y}_{iq}$ will be the same as the corresponding sum for $\mathbf{y}^*_{iq}$. Consequently, the small area means $\overline{Y}_i$ and $\overline{Y}^*_i$ will be identical. Given (11) and (12), and the estimators (14) and (15), the pseudo-BLUP of $\overline{Y}_i$ (referred to from now on as the *BLUP) can be written as

$$\tilde{\overline{Y}}_i^{*BLUP} = N_i^{-1}\left\{n_i\overline{y}^*_{si} + (N_i - n_i)\left(\overline{\mathbf{x}}^{*\prime}_{ri}\tilde{\boldsymbol{\beta}}^* + \mathbf{z}'_i\tilde{\mathbf{u}}^*\right)\right\}, \tag{16}$$

where $\overline{y}^*_{si} = n_i^{-1}\sum_{q\in i}\sum_{j\in s_{iq}} y^*_{iqj}$ and $\overline{\mathbf{x}}^*_{ri}$ is the vector of column averages for the non-sampled rows of $\mathbf{X}^*_i$.

The 'empirical' versions of Equations (14) and (15), denoted by $\hat{\boldsymbol{\beta}}^*$ and $\hat{\mathbf{u}}^*_i$ respectively, are obtained by substituting estimates $\hat{\boldsymbol{\delta}}^*$ for the unknown variance components $\boldsymbol{\delta}$ that define Σ_{siq}. These estimates can be obtained using either the pseudo maximum likelihood (pseudo-ML) or the pseudo restricted maximum likelihood (pseudo-REML) approach of Samart and Chambers (2014), and only depend on the conditional moments (11) and (12). Note that the resulting estimators are referred to as pseudo-ML (or pseudo-REML) because their estimating functions are based on the assumption that the matrices Σ_{siq} are known. The pseudo-EBLUP of the small area mean $\overline{Y}_i$ is then defined by substituting $\hat{\boldsymbol{\beta}}^*$ and $\hat{\mathbf{u}}^*$ for $\tilde{\boldsymbol{\beta}}^*$ and $\tilde{\mathbf{u}}^*$, respectively, in Equation (16). This is denoted $\hat{\overline{Y}}_i^{*EBLUP}$, and referred to as the *EBLUP, in what follows.

Note that (15) can be expressed as

$$\tilde{\mathbf{u}}^*_i = \sum_{q\in i}\Sigma_{\mathbf{u}_i}\mathbf{Z}'_{siq}\tilde{\Sigma}^{-1}_{siq}\{\mathbf{y}^*_{siq} - \mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^* + (1-\lambda_{iq})\frac{N_{iq}}{N_{iq}-1}\left(\mathbf{X}_{siq} - \mathbf{1}_{n_{iq}}\overline{\mathbf{x}}'_{iq}\right)\tilde{\boldsymbol{\beta}}^*\}, \tag{17}$$

under the assumed ELE model since $\mathbf{y}^*_{siq} - \mathbf{X}^*_{siq}\tilde{\boldsymbol{\beta}}^* = \mathbf{y}^*_{siq} - \{(\lambda_{iq} - \gamma_{iq})\mathbf{X}_{siq} + \gamma_q N_{iq}\mathbf{1}_{n_{iq}}\overline{\mathbf{x}}'_{iq}\}\tilde{\boldsymbol{\beta}}^*$ and $-\lambda_{iq}\mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^* = (1-\lambda_{iq})\mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^* - \mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^*$. This implies

$$\mathbf{y}^*_{siq} - \mathbf{X}^*_{siq}\tilde{\boldsymbol{\beta}}^* = \mathbf{y}^*_{siq} - \mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^* + (1-\lambda_{iq})\mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^* + \gamma_{iq}\mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^* - \gamma_q N_{iq}\mathbf{1}_{n_{iq}}\overline{\mathbf{x}}'_{iq}\tilde{\boldsymbol{\beta}}^*.$$

As $\gamma_q = (1-\lambda_{iq})/(N_{iq}-1)$ and $(1-\lambda_{iq}) = \gamma_{iq}(N_{iq}-1)$ it follows that $(1-\lambda_{iq})\mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^* + \gamma_{iq}\mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^* = , (1-\lambda_{iq})\frac{N_{iq}}{N_{iq}-1}\mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^*$ from which (17) easily follows.

Ignoring $N_{iq}/(N_{iq}-1)$ in Equation (17), the jth component of the residual vector $\mathbf{y}^*_{siq} - \mathbf{X}^*_{siq}\tilde{\boldsymbol{\beta}}^*$ becomes $y^*_{iqj} - \mathbf{x}'_{iqj}\tilde{\boldsymbol{\beta}}^* + \left(1-\lambda_{iq}\right)(\mathbf{x}_{iqj} - \overline{\mathbf{x}}_{iq})'\tilde{\boldsymbol{\beta}}^*$. That is, this modified residual can be interpreted as the naive residual that ignores linkage error plus a bias correction that increases in absolute value as the probability of an incorrect match increases and also as the leverage exerted by the individual fitted value $\mathbf{x}'_{iqj}\tilde{\boldsymbol{\beta}}^*$ increases. Alternatively, we can note the approximation

$$\mathbf{y}^*_{siq} - \mathbf{X}^*_{siq}\tilde{\boldsymbol{\beta}}^* \approx \lambda_{iq}\left(\mathbf{y}^*_{siq} - \mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^*\right) + (1-\lambda_{iq})\left(\mathbf{y}^*_{siq} - \mathbf{1}_{n_{iq}}\overline{\mathbf{x}}'_{iq}\tilde{\boldsymbol{\beta}}^*\right). \tag{18}$$

When the probability of an incorrect match is significantly greater than zero, the second term on the right hand side of Equation (18) can be unstable. Consequently, we consider an alternative expression for the predicted area effect that ignores this second term. This leads to a modified predictor of $\mathbf{u}_i$ of the form

$$\tilde{\mathbf{u}}^{**}_i = \sum_{q\in i}\Sigma_{\mathbf{u}_i}\mathbf{Z}'_{siq}\tilde{\Sigma}^{-1}_{siq}\lambda_{iq}\left(\mathbf{y}^*_{siq} - \mathbf{X}_{siq}\tilde{\boldsymbol{\beta}}^*\right). \tag{19}$$

The corresponding versions of the pseudo-BLUP and pseudo-EBLUP are referred to as **BLUP and **EBLUP, respectively, in what follows.

3.1 | MSE estimation for the *EBLUP and **EBLUP

Methods for estimating the unconditional MSE of small area EBLUPs are typically based on averaging over the distribution of the random area effects, with the standard estimator of the unconditional MSE of the EBLUP predictor being the one suggested by Prasad and Rao (1990). In what follows, we derive an estimator of the unconditional MSE of $\widehat{\overline{Y}}_i^{*EBLUP}$ along similar lines. In particular, we assume the regularity conditions (RCs) 1-6 set out in Section S.2 of Supplementary Material and use the decomposition

$$MSE_{\mathbf{A},M}\left(\widehat{\overline{Y}}_i^{*EBLUP}\right) = MSE_{\mathbf{A},M}\left(\tilde{\overline{Y}}_i^{*BLUP}\right) + E_{\mathbf{A},M}\left(\left(\widehat{\overline{Y}}_i^{*EBLUP} - \tilde{\overline{Y}}_i^{*BLUP}\right)^2\right). \tag{20}$$

Under normality of the random area and individual effects, the cross-product term missing from (20) is zero provided the vector of variance components is translation invariant. Furthermore, the last component on the right hand side of Equation (20) is generally intractable. We therefore approximate this term using a first-order Taylor expansion. Finally, given the values of the variance components, the estimators of $\boldsymbol{\beta}$ can be represented as the solution to estimating equations, as we have stated above, and consequently, we follow the approach described in Chambers (2009) in order to define large sample estimators of the variances of these regression parameter estimators. Following Henderson (1975), we first note that the *MSE* of $\tilde{\overline{Y}}_i^{*BLUP}$ is

$$MSE_{\mathbf{A},M}\left(\tilde{\overline{Y}}_i^{*BLUP}\right) = g_{1i}^*(\boldsymbol{\delta}) + g_{2i}^*(\boldsymbol{\delta}), \tag{21}$$

where

$$g_{1i}^*(\boldsymbol{\delta}) = \mathbf{z}_i'\{\boldsymbol{\Sigma}_{\mathbf{u}_i} - \boldsymbol{\Sigma}_{\mathbf{u}_i}\sum_{q\in i}(\mathbf{Z}_{siq}'\tilde{\boldsymbol{\Sigma}}_{siq}^{-1}\mathbf{Z}_{siq})\boldsymbol{\Sigma}_{\mathbf{u}_i}\}\mathbf{z}_i,$$

and $g_{2i}^*(\boldsymbol{\delta}) = \mathbf{C}_i\left(\sum_{i=1}^D\sum_{q\in i}\mathbf{X}_{siq}^{*\prime}\tilde{\boldsymbol{\Sigma}}_{siq}^{-1}\mathbf{X}_{siq}^*\right)^{-1}\mathbf{C}_i'$. Here $\mathbf{C}_i = \overline{\mathbf{x}}_i' - \mathbf{z}_i'\boldsymbol{\Sigma}_{\mathbf{u}_i}\sum_{q\in i}(\mathbf{Z}_{siq}'\tilde{\boldsymbol{\Sigma}}_{siq}^{-1}\mathbf{X}_{siq})$ and $\overline{\mathbf{x}}_i = N_i^{-1}\sum_{q\in i}N_{iq}\overline{\mathbf{x}}_{iq}$. Next, a first order approximation to the second term on the right hand side of Equation (20) follows from:

Proposition 1 *Suppose that population model (8) holds, the sample design is non-informative so that sample model (10) holds, and linkage error is non-informative, so that (4) holds. In addition, assume that the regularity conditions 1–6 specified in Section S.2 of Supplementary Material hold, and that the random errors in Equation (8) are normally distributed. Let* $\widehat{\boldsymbol{\delta}}$ *be the pseudo-REML estimator of* $\boldsymbol{\delta}$ *and put*

$$g_{3i}^*(\boldsymbol{\delta}) = tr\{(\nabla\mathbf{b}_i')(\sum_{q\in i}\tilde{\boldsymbol{\Sigma}}_{siq})(\nabla\mathbf{b}_i')'E_{\mathbf{A},M}(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta})(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta})'\}, \tag{22}$$

where $\nabla\mathbf{b}_i' = col_{1\leqslant l\leqslant K}(\partial\mathbf{b}_i'/\partial\delta_l)$ and $\mathbf{b}_i' = \sum_{q\in i}\boldsymbol{\Sigma}_{\mathbf{u}_i}\mathbf{Z}_{siq}'\tilde{\boldsymbol{\Sigma}}_{siq}^{-1}$. Then

$$E_{\mathbf{A},M}\left(\left(\widehat{\overline{Y}}_i^{*EBLUP} - \tilde{\overline{Y}}_i^{*BLUP}\right)^2\right) = g_{3i}^*(\boldsymbol{\delta}) + o(D^{-1}). \tag{23}$$

Proof The expansion (23) is a modified version of that obtained by Prasad and Rao (1990).

Following the same line of reasoning as in Prasad and Rao (1990), an approximately unbiased estimator of the MSE of $\widehat{\overline{Y}}_i^{*EBLUP}$ under (8) is then

$$\widehat{MSE}_{\mathbf{A},M}(\widehat{\overline{Y}}_i^{*EBLUP}) = \left(1-\frac{n_i}{N_i}\right)^2 \left\{ g_{1i}^*(\widehat{\boldsymbol{\delta}}^*) + g_{2i}^*(\widehat{\boldsymbol{\delta}}^*) + 2g_{3i}^*(\widehat{\boldsymbol{\delta}}^*) \right\}. \tag{24}$$

An estimator of the covariance matrix of the variance component estimator $\widehat{\boldsymbol{\delta}}^*$ is necessary in order to compute (22). This can be obtained as the inverse of the expected information matrix developed in Samart and Chambers (2014). If a pseudo-ML estimator is used instead, a bias correction to (24) is necessary, along the same lines in Rao and Molina (2015, Section 5.2.5). If the random area effects are estimated using (19), then the three components of the estimated MSE become:

$$g_{1i}^{**}(\widehat{\boldsymbol{\delta}}^*) = \mathbf{z}_i' \{ \widehat{\Sigma}_{\mathbf{u}_i} + \widehat{\Sigma}_{\mathbf{u}_i} \sum_{q\in i} \lambda_{iq}^2 (\mathbf{Z}_{siq}' \widehat{\Sigma}_{siq}^{-1} \mathbf{Z}_{siq}) \widehat{\Sigma}_{\mathbf{u}_i} - \widehat{\Sigma}_{\mathbf{u}_i} \sum_{q\in i} \lambda_{iq} (\mathbf{Z}_{siq}' \widehat{\Sigma}_{siq}^{-1} \mathbf{Z}_{siq}) \widehat{\Sigma}_{\mathbf{u}_i} \} \mathbf{z}_i,$$

$$g_{2i}^{**}(\widehat{\boldsymbol{\delta}}^*) = \widehat{\mathbf{C}}_i^{**\prime} \left(\sum_{i=1}^{D} \sum_{q\in i} \mathbf{X}_{siq}^{*\prime} \widehat{\Sigma}_{siq}^{-1} \mathbf{X}_{siq}^* \right)^{-1} \widehat{\mathbf{C}}_i^{**}, \text{ and } g_{3i}^{**}(\widehat{\boldsymbol{\delta}}^*) = tr\left\{ \nabla \mathbf{b}_i^{**\prime} \sum_{q\in i} \widehat{\Sigma}_{siq} (\nabla \mathbf{b}_i^{**\prime})' \widehat{V}(\widehat{\boldsymbol{\delta}}) \right\}.$$

Here $\widehat{\mathbf{C}}_i^{**} = \overline{\mathbf{x}}_i' - \mathbf{z}_i' \widehat{\Sigma}_{\mathbf{u}_i} \sum_{q\in i} \lambda_{iq} (\mathbf{Z}_{siq}' \widehat{\Sigma}_{siq}^{-1} \mathbf{X}_{siq})$, $\nabla \mathbf{b}_i^{**\prime} = col_{1\leqslant l\leqslant K}(\partial \mathbf{b}_i^{**\prime} / \partial \delta_l)$ and $\mathbf{b}_i^{**\prime} = \sum_{q\in i} \lambda_{iq} \widehat{\Sigma}_{\mathbf{u}_i} \mathbf{Z}_{siq}' \widehat{\Sigma}_{siq}^{-1}$.

4 | ROBUST LINEAR MIXED MODELS FOR SMALL AREA ESTIMATION WITH LINKED DATA

Outliers are a problem for any model-based survey estimation method, but particularly so for small area estimates. Sinha and Rao (2009) proposed an estimator of a small area mean based on outlier robust estimation of the parameters of the linear mixed model. This robust-projective approach (Chambers et al., 2014) uses plug-in robust prediction, that is, the authors substitute these outlier robust parameter estimates for the optimal, but outlier-sensitive, parameter estimates used in the EBLUP. In particular, they estimate fixed effects and variance components using a modified version of the estimating equations corresponding to the Robust ML Proposal II of Richardson and Welsh (1995) and compute outlier robust predictions for the random area effects using the robust estimating equations suggested by Fellner (1986). The solutions to these estimating equations depend on specification of a bounded influence function ψ, which we take to be the Huber (1981) influence function, defined as $\psi(u) = u\min(1, c/|c|)$ where c is a tuning constant. Quantities that depend on this influence function (and hence on the choice of the tuning constant) will be denoted by a superscript ψ in what follows. Under (8), the Sinha and Rao (2009) robust version of the EBLUP, denoted by REBLUP, of the small area mean $\overline{Y}_i$ is given by

$$\widehat{\overline{Y}}_i^{\psi REBLUP} = N_i^{-1} \left\{ n_i \overline{y}_{si} + (N_i - n_i) \left(\overline{\mathbf{x}}_{ri}' \widehat{\boldsymbol{\beta}}^{\psi} + \mathbf{z}_i' \widehat{\mathbf{u}}^{\psi} \right) \right\}, \tag{25}$$

where $\overline{y}_{si} = \sum_{j\in s_i} y_{ij}/n_i$, $\overline{\mathbf{x}}_{ri}$ denotes the vector of average values of $\mathbf{X}_i$ for non-sampled units in small area i, $\widehat{\boldsymbol{\beta}}^{\psi}$ is the robust estimated vector of regression coefficients and $\widehat{\mathbf{u}}^{\psi}$ is the vector of robust predicted values of the area effects.

Given linked data, we modify the estimating equations of both the Robust ML Proposal II of Richardson and Welsh (1995) and of Fellner (1986) to account for linkage errors, making the same

assumptions (ELE errors, one to one, complete and non-informative linkage) as in Section 2. The modified Fellner (1986) estimating equations assume the variance components $\boldsymbol{\delta}$ are known, and define robust estimates $\boldsymbol{\beta}^*$ and $\mathbf{u}^*$ of the fixed effects and the area random effects, respectively. These are:

$$\sum_{i=1}^{D}\sum_{q\in i}\mathbf{X}_{siq}^{*\prime}\Sigma_{siq}^{-1}\mathbf{U}_{siq}^{1/2}\psi\{\mathbf{r}_{siq}^{*}\}=\mathbf{0}, \tag{26}$$

and

$$\sum_{i=1}^{D}\sum_{q\in i}\left[\mathbf{Z}_{siq}^{\prime}\Sigma_{seAiq}^{-1/2}\psi\left\{\Sigma_{seAiq}^{-1/2}(\mathbf{y}_{siq}^{*}-\mathbf{X}_{siq}^{*}\boldsymbol{\beta}^{*}-\mathbf{Z}_{siq}\mathbf{u}_{i}^{*})\right\}-\Sigma_{\mathbf{u}_i}^{-1/2}\psi\{\Sigma_{\mathbf{u}_i}^{-1/2}\mathbf{u}_{i}^{*}\}\right]=\mathbf{0}. \tag{27}$$

Here $\mathbf{r}_{siq}^{*}=\mathbf{U}_{siq}^{-1/2}(\mathbf{y}_{siq}^{*}-\mathbf{X}_{siq}^{*}\boldsymbol{\beta}^{*})$, $\Sigma_{seAiq}=\Sigma_{seiq}^{*}+\mathbf{V}_{siq}$, Σ_{siq} is given in Equation (12), and $\mathbf{U}_{siq}$ is a diagonal matrix with the same diagonal entries as Σ_{siq}. As with the computation of Equation (14), the estimation of $\boldsymbol{\beta}^*$ is carried out iteratively, with an initial estimate of $\boldsymbol{\beta}^*$ first used to compute an estimate of $\mathbf{V}_{siq}$ based on (13). This is substituted into (26) to obtain an updated estimate of $\boldsymbol{\beta}^*$, which is then used to re-estimate $\mathbf{V}_{siq}$. At convergence, the estimates of $\boldsymbol{\beta}^*$ and $\mathbf{V}_{siq}$ are substituted into (27) in order to estimate $\mathbf{u}_i^*$. Robust estimates of the variance components $\boldsymbol{\delta}$ are obtained as solutions to the modified version of the Robust ML Proposal II estimating equations of Richardson and Welsh (1995), and are then substituted into (26) and (27). Let $\mathbf{K}_{siq}=E\{\psi^2(R)\}\mathbf{I}_{n_{iq}}$, where R is a standard normal random variable. These estimating equations can be written as

$$\sum_{i=1}^{D}\sum_{q\in i}\left\{\boldsymbol{\psi}^{\prime}\{\mathbf{r}_{siq}^{*}\}\mathbf{U}_{siq}^{1/2}\Sigma_{siq}^{-1}\frac{\partial\Sigma_{siq}}{\partial\delta_l}\Sigma_{siq}^{-1}\mathbf{U}_{siq}^{1/2}\boldsymbol{\psi}\{\mathbf{r}_{siq}^{*}\}-tr\left(\mathbf{K}_{siq}\Sigma_{siq}^{-1}\frac{\partial\Sigma_{siq}}{\partial\delta_l}\right)\right\}=\mathbf{0}, \tag{28}$$

and the corresponding linkage error-adjusted version of the REBLUP is

$$\hat{\bar{Y}}_{i}^{\psi*REBLUP}=N_i^{-1}\left\{n_i\bar{y}_{si}^{*}+(N_i-n_i)\left(\bar{\mathbf{x}}_{ri}^{*\prime}\hat{\boldsymbol{\beta}}^{\psi*}+\mathbf{z}_i^{\prime}\hat{\mathbf{u}}^{\psi*}\right)\right\}. \tag{29}$$

Here $\hat{\boldsymbol{\beta}}^{\psi*}$ and $\hat{\mathbf{u}}^{\psi*}$ depend on the influence function ψ, and are the solutions to the robust estimating equations (26) and (27) respectively, using values of $\boldsymbol{\delta}$ obtained by solving (28). Note that the value of Equation (29) when the true value of $\boldsymbol{\delta}$ is used in Equations (26) and (27) is referred to as the pseudo-RBLUP and denoted by $\tilde{\bar{Y}}_{i}^{\psi*RBLUP}$, with the corresponding solutions to (26) and (27) denoted by $\tilde{\boldsymbol{\beta}}^{\psi*}$ and $\tilde{\mathbf{u}}^{\psi*}$, respectively. We refer to these linkage-error adjusted versions of RBLUP and REBLUP as *RBLUP and *REBLUP respectively below.

When the probability of an incorrect match is significantly greater than zero, the solution to equation (27) can be unstable. Consequently, as with the pseudo-EBLUP, we propose an alternative version of the Fellner (1986) estimating equation for computing the robust random area effects:

$$\sum_{i=1}^{D}\sum_{q\in i}\left[\mathbf{Z}_{siq}^{\prime}\Sigma_{seAiq}^{-1/2}\psi\left\{\lambda_{iq}\Sigma_{seAiq}^{-1/2}(\mathbf{y}_{siq}^{*}-\mathbf{X}_{siq}^{*}\boldsymbol{\beta}^{*}-\mathbf{Z}_{siq}\mathbf{u}_{i}^{*})\right\}-\Sigma_{\mathbf{u}_i}^{-1/2}\psi\{\Sigma_{\mathbf{u}_i}^{-1/2}\mathbf{u}_{i}^{*}\}\right]=\mathbf{0}. \tag{30}$$

The solution to (30) is denoted by $\tilde{\mathbf{u}}^{\psi**}$, and the version of the pseudo-REBLUP obtained by substituting $\hat{\mathbf{u}}^{\psi**}$ for $\hat{\mathbf{u}}^{\psi*}$ in Equation (29) is referred to as the **REBLUP below.

4.1 | MSE estimation for the *REBLUP and **REBLUP

We develop an analytic estimator for the MSE of Equation (29) under a working mixed model that conditions on the realised values of the area effects, that is, the proposed MSE estimator is an estimator of the conditional MSE of $\widehat{\bar{Y}}_i^{\psi*REBLUP}$. It is based on the assumption that a consistent estimator of the MSE of a linear approximation to a non-linear small area estimator can be used as an estimator of the MSE of that small area estimator. See Booth and Hobert (1998) and Chambers et al. (2014). Our theoretical development is based on approximations that correspond to assuming that $max(n_{iq}) = O(1)$, so that, as $D \to +\infty$, the prediction variance and the squared bias are $O(1)$ and the correction term is $O(D^{-1})$ (see Chambers et al., 2014). The same arguments are valid for approximating the MSE of the M-quantile predictors described in Section S.4. Such linearisation-based MSE estimators are generally not consistent, and can be biased low, see Harville and Jeske (1992). However, in small sample problems, this is not an issue since it is the variability, rather than the bias, of the MSE estimator that is of concern. The development below omits some technical details, which are available from the authors upon request.

Proposition 2 *Put* $\tilde{\boldsymbol{\theta}}^{\psi*} = (\tilde{\boldsymbol{\beta}}^{\psi*\prime}, \tilde{\mathbf{u}}^{\psi*\prime})'$ *and assume that this random variable converges in probability to* $\boldsymbol{\theta}_0^{\psi*} = (\boldsymbol{\beta}_0^{\psi*\prime}, \mathbf{u}_0^{\psi*\prime})'$. *Also let* $V_{\mathbf{A},M|\mathbf{u}}$ *denote variance with respect to both the linkage error model and the linear mixed model for* $\mathbf{y}$ *in terms of* $\mathbf{X}$ *given the realised values of the area effects. Suppose the same assumptions are made as in Proposition 1, regularity conditions 1–5 and 7–9 in Section S.2 of Supplementary Material apply, and* ψ *corresponds to the Huber influence function. Then the conditional prediction variance of the* *RBLUP*RBLUP* *can be expressed as*

$$V_{\mathbf{A},M|\mathbf{u}}(\tilde{\bar{Y}}_i^{\psi*RBLUP} - \bar{Y}_i) = \left(1 - \frac{n_i}{N_i}\right)^2 \{[\bar{\mathbf{x}}_{ri}^{*\prime} \mid \mathbf{z}_i'] V_{\mathbf{A},M|\mathbf{u}}(\tilde{\boldsymbol{\theta}}^{\psi*})[\bar{\mathbf{x}}_{ri}^{*\prime} \mid \mathbf{z}_i']' + V_{\mathbf{A},M|\mathbf{u}}(\bar{\mathbf{e}}_{ri}^{\psi*})\} + o(D^{-1}) \tag{31}$$

where $\bar{\mathbf{e}}_{ri}^{\psi*} = (N_i - n_i)^{-1} \sum_{j \in r_i} (y_{ij}^* - \mathbf{x}_{ij}^{*\prime} \boldsymbol{\beta}_0^{\psi*} - \mathbf{z}_i' \mathbf{u}_0^{\psi*})$.

Proof Formula (31) is a modified version of a similar prediction variance formula developed in Chambers et al. (2014).

A first-order approximation to the prediction variance (31) is then:

$$\widehat{V}_{\mathbf{A},M|\mathbf{u}}(\tilde{\bar{Y}}_i^{\psi*RBLUP} - \bar{Y}_i) = h_{1i}(\tilde{\boldsymbol{\theta}}^{\psi*}) + h_{2i}(\tilde{\boldsymbol{\theta}}^{\psi*}), \tag{32}$$

where

1. $h_{1i}(\tilde{\boldsymbol{\theta}}^{\psi*}) = \left(1 - \frac{n_i}{N_i}\right)^2 [\bar{\mathbf{x}}_{ri}^{*\prime} \mid \mathbf{z}_i'] \widehat{V}_{\mathbf{A},M|\mathbf{u}}(\tilde{\boldsymbol{\theta}}^{\psi*})[\bar{\mathbf{x}}_{ri}^{*\prime} \mid \mathbf{z}_i']'$. Here $\widehat{V}_{\mathbf{A},M|\mathbf{u}}(\tilde{\boldsymbol{\theta}}^{\psi*})$ is the sandwich-type estimator of $V_{\mathbf{A}M|\mathbf{u}}(\tilde{\boldsymbol{\theta}}^{\psi*})$ set out in Section S.3 of the Supplementary Material.
2. $h_{2i}(\tilde{\boldsymbol{\theta}}^{\psi*}) = \left(1 - \frac{n_i}{N_i}\right)^2 \widehat{V}_{\mathbf{A},M|\mathbf{u}}(\bar{\mathbf{e}}_{ri}^{\psi*})$ where

$$\widehat{V}_{\mathbf{A},M|\mathbf{u}}(\bar{\mathbf{e}}_{ri}^{\psi*}) = (N_i - n_i)^{-1}(n_i - 1)^{-1} \sum_l \sum_{j \in s_l} (y_{lj}^* - \mathbf{x}_{lj}^{*\prime} \tilde{\boldsymbol{\beta}}^{\psi*} - \mathbf{z}_l' \tilde{\mathbf{u}}^{\psi*})^2.$$

Note that $\widehat{V}_{\mathbf{A},M|\mathbf{u}}(\bar{\mathbf{e}}_{ri}^{\psi*})$ above pools data from the entire sample. This leads to more stable MSE estimates when area sample sizes are very small. The estimator of the MSE of the *RBLUP is obtained by adding an estimator of the squared conditional bias to (32):

$$\widehat{MSE}_{\mathbf{A},M|\mathbf{u}}(\bar{\bar{Y}}_i^{\psi*RBLUP}) = h_{1i}(\tilde{\boldsymbol{\theta}}^{\psi*}) + h_{2i}(\tilde{\boldsymbol{\theta}}^{\psi*}) + \widehat{B}^2_{\mathbf{A},M|\mathbf{u}}(\bar{\bar{Y}}_i^{\psi*RBLUP}), \tag{33}$$

where

$$\widehat{B}_{\mathbf{A},M|\mathbf{u}}(\bar{\bar{Y}}_i^{\psi*RBLUP}) = \sum_{j\in s_i} w_{ij}^{\psi*RBLUP}\tilde{\mu}_{ij}^{*} - N_i^{-1}\sum_{j\in U_i}\tilde{\mu}_{ij}^{*}. \tag{34}$$

Here $\tilde{\mu}_{ij}^{*}$ is an unbiased linear estimator of the conditional expected value $\mu_{ij}^{*} = E_{\mathbf{A},M}(y_{ij}^{*}|\mathbf{x}_{ij}, \mathbf{u}^{\psi*})$ and $w_{ij}^{\psi*RBLUP}$ is the weight for unit j in area i based on the pseudo-linearisation approach to MSE estimation described in Chambers et al. (2011) and extended to RBLUP in Chambers et al. (2014). Note that the weights $w_{ij}^{\psi*RBLUP}$ can be obtained in a straightforward manner in the case of linkage data following Chambers et al. (2011).

Finally, we consider MSE estimation for the *REBLUP (29), noting that an estimator of its conditional MSE can based on a decomposition similar to that used in Prasad and Rao (1990):

$$\begin{aligned} MSE_{\mathbf{A},M|\mathbf{u}}(\hat{\bar{Y}}_i^{\psi*REBLUP}) &= MSE_{\mathbf{A},M|\mathbf{u}}(\bar{\bar{Y}}_i^{\psi*RBLUP}) + \\ &+ E_{\mathbf{A},M|\mathbf{u}}\left(\left(\hat{\bar{Y}}_i^{\psi*REBLUP} - \bar{\bar{Y}}_i^{\psi*RBLUP}\right)^2\right) + O(D^{-1}). \end{aligned} \tag{35}$$

An approximation to the second term on the right-hand side of equation (35) can be obtained using the following proposition.

Proposition 3 *We make the same assumptions as in Proposition 1. In addition, we assume that the regularity conditions 1–5 and 7–9 in Section S.2 of Supplementary Material hold. Let* $\hat{\boldsymbol{\delta}}^{\psi*}$ *be the vector of estimated variance components obtained by solving the robust estimating equations (26)–(28). Then*

$$E_{\mathbf{A},M|\mathbf{u}}\left(\left(\hat{\bar{Y}}_i^{\psi*REBLUP} - \bar{\bar{Y}}_i^{\psi*RBLUP}\right)^2\right) = h_{3i}(\hat{\boldsymbol{\delta}}^{\psi*}) + O(D^{-1}), \tag{36}$$

where $h_{3i}(\hat{\boldsymbol{\delta}}^{\psi*}) = \mathbf{z}_i'\sum_{q\in i}\Omega_{siq}V_{\mathbf{A},M|\mathbf{u}}(\hat{\boldsymbol{\delta}}^{\psi*})\mathbf{z}_i$. Here

$$\Omega_{siq} = \sum_{k=1}^{K}\sum_{g=1}^{K}\left[(\partial_{\delta_k}\mathbf{B}_{siq})\left\{(\mathbf{z}_i'\mathbf{u}_0^{\psi*})(\mathbf{z}_i'\mathbf{u}_0^{\psi*})' + \tilde{\Sigma}_{seAiq}\right\}(\partial_{\delta_g}\mathbf{B}_{siq})'\right],$$

$$\mathbf{B}_{siq} = \left(\mathbf{Z}_{siq}'\tilde{\Sigma}_{seAiq}^{-1/2}\mathbf{W}_{2siq}\tilde{\Sigma}_{seAiq}^{-1/2}\mathbf{Z}_{siq} + \Sigma_{\mathbf{u}_i}^{-1/2}\mathbf{W}_{3siq}\Sigma_{\mathbf{u}_i}^{-1/2}\right)^{-1}\left(\mathbf{Z}_{siq}'\tilde{\Sigma}_{seAiq}^{-1/2}\mathbf{W}_{2siq}\tilde{\Sigma}_{seAiq}^{-1/2}\right), \tag{37}$$

$$\mathbf{W}_{2siq} = \psi\left\{\tilde{\Sigma}_{seAiq}^{-1/2}(\mathbf{y}_{siq}^{*} - \mathbf{X}_{siq}^{*}\tilde{\boldsymbol{\beta}}^{\psi*} - \mathbf{Z}_{siq}\tilde{\mathbf{u}}_i^{\psi*})\right\}\left\{\tilde{\Sigma}_{seAiq}^{-1/2}(\mathbf{y}_{siq}^{*} - \mathbf{X}_{siq}^{*}\tilde{\boldsymbol{\beta}}^{\psi*} - \mathbf{Z}_{siq}\tilde{\mathbf{u}}_i^{\psi*})\right\}^{-1}$$

is a $n_{iq} \times n_{iq}$ diagonal matrix of weights for the individual effects in area i, and $\mathbf{W}_{3siq} = \psi\{\Sigma_{\mathbf{u}_i}^{-1/2}\tilde{\mathbf{u}}_i^{\psi*}\}\left(\Sigma_{\mathbf{u}_i}^{-1/2}\tilde{\mathbf{u}}_i^{\psi*}\right)^{-1}$ is a $m \times m$ diagonal matrix of weights for the area effect associated with area i and $\tilde{\Sigma}_{seAiq}$ denotes the value of Σ_{seAiq} at the solution $\boldsymbol{\beta}^*$ to (26).

Proof Formula (22) takes into account expectation with respect to the linkage error model and the model that relates $\mathbf{y}$ to $\mathbf{X}$ and $\mathbf{Z}$, and is a suitably modified version of a similar formula in Chambers et al. (2014).

We define an estimator $\widehat{V}_{\mathbf{A},M|\mathbf{u}}(\widehat{\boldsymbol{\delta}}^{\psi*})$ of the variance-covariance matrix of the estimated variance components in Section S.3 of the Supplementary Material, based on the approach taken by Sinha and Rao (2009). Let $\widehat{h}_{1i}(\widehat{\boldsymbol{\theta}}^{\psi*})$ and $\widehat{h}_{2i}(\widehat{\boldsymbol{\theta}}^{\psi*})$ denote the values of $h_{1i}(\tilde{\boldsymbol{\theta}}^{\psi*})$ and $h_{2i}(\tilde{\boldsymbol{\theta}}^{\psi*})$ respectively when all unknown parameters are replaced by robust estimates, and put $\widehat{h}_{3i}(\widehat{\boldsymbol{\delta}}^{\psi*})$ equal to $h_{3i}(\widehat{\boldsymbol{\delta}}^{\psi*})$ in which the estimator $\widehat{V}_{\mathbf{A},M|\mathbf{u}}(\widehat{\boldsymbol{\delta}}^{\psi*})$ is substituted. An estimator of the conditional MSE of the *REBLUP for area i is then:

$$\widehat{MSE}_{\mathbf{A},M|\mathbf{u}}(\widehat{\bar{Y}}_i^{\psi*REBLUP}) = \widehat{h}_{1i}(\widehat{\boldsymbol{\theta}}^{\psi*}) + \widehat{h}_{2i}(\widehat{\boldsymbol{\theta}}^{\psi*}) + \widehat{h}_{3i}(\widehat{\boldsymbol{\theta}}^{\psi*}) + \widehat{B}^2_{\mathbf{A},M|\mathbf{u}}(\widehat{\bar{Y}}_i^{\psi*REBLUP}). \tag{38}$$

Note that if the random area effects are estimated using (30), then the components of the estimated MSE (38) can be derived in straightforwardly. Details for this case are not reported here for reasons of space, but they can be made available to the interested reader upon request.

5 | M-QUANTILE MODELS FOR SMALL AREA ESTIMATION WITH LINKED DATA

M-quantile regression models were first suggested for small area estimation by Chambers and Tzavidis (2006). See Bianchi et al. (2018) for a recent review of further applications and theoretical extensions. Here, we briefly discuss basic ideas and develop appropriate notation.

Breckling and Chambers (1988) introduced M-quantile regression as a 'quantile-like' generalisation of regression based on a bounded influence function ψ (Huber, 1981), with associated loss function ρ such that $\psi = d\rho(u)/du$. For a given $\tau \in (0,1)$, the M-quantile of order τ of a random variable is defined as the value minimising the expectation of the tilted loss function $\rho_\tau = |\tau - I(u < 0)|\rho(u)$, where $\rho(u)$, $u \in \Re$, is continuously differentiable with $\rho(0) = 0$. M-quantiles aim at combining the robustness properties of quantiles ($\rho(u) = |u|$) with the efficiency properties of expectiles ($\rho(u) = u^2$). By definition, any M-quantile depends on the specification of the influence function ψ, and so we will not explicitly refer to ψ in our notation below for quantities that depend on values of M-quantiles that are all defined using the same ψ.

In the linear case, M-quantile regression leads to a family of hyperplanes indexed by the order τ of the corresponding M-quantile, that is,

$$MQ_\tau(y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}_\tau.$$

For specified τ and influence function ψ (with $\psi_\tau = d\rho_\tau(u)/du$), an estimate $\widehat{\boldsymbol{\beta}}_\tau$ of the vector of regression parameters $\boldsymbol{\beta}_\tau$ may be obtained as the solution to the normal equations,

$$\sum_{i=1}^{D}\sum_{j\in s_i}\psi_\tau\left\{\frac{y_{ij} - \mathbf{x}'_{ij}\widehat{\boldsymbol{\beta}}_\tau}{\sigma_\tau}\right\}\mathbf{x}_{ij} = \mathbf{0}, \tag{39}$$

where σ_τ is a scale parameter that characterises the spread of the distribution of the residuals $y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta}_\tau$. Following standard practice in robust M-regression, this scale parameter can be estimated by $\hat{\sigma}_\tau = median(|y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}_\tau|)/0.6745$. When ψ is a continuous function, Breckling and Chambers (1988) adapt the iteratively reweighted least squares (IRWLS) approach to M-regression and show that a linear regression M-quantile of specified order τ can be estimated by weighting positive residuals from the M-quantile line by τ and negative residuals by $1 - \tau$. Note that in what follows we will assume that ψ is the Huber influence function, with tuning constant $c > 0$, and so ψ is continuous.

Following Chambers and Tzavidis (2006), we characterise conditional variability given $\mathbf{x}_{ij}$ across the population of interest by the M-quantile coefficients of the population units. For unit j in area i this coefficient is the value τ_{ij} such that $MQ_{\tau_{ij}}(y_{ij}|\mathbf{x}_{ij}) = y_{ij}$. If a hierarchical structure does explain part of the variability in the population data, units within areas defined by this hierarchy are expected to have similar M-quantile coefficients. When the conditional M-quantiles are assumed to follow a linear model, with $\boldsymbol{\beta}_\tau$ a sufficiently smooth function of τ, Chambers and Tzavidis (2006) suggest a predictor of $\overline{Y}_i$ of the form

$$\widehat{\overline{Y}}_i^{MQ} = N_i^{-1}\left\{n_i\overline{y}_{si} + (N_i - n_i)\overline{\mathbf{x}}'_{ri}\hat{\boldsymbol{\beta}}_{\hat{\tau}_i}\right\}, \tag{40}$$

where $\hat{\tau}_i = n_i^{-1}\sum_{j\in s_i}\hat{\tau}_{ij}$ is an estimate of the average value of the M-quantile coefficients τ_{ij} for population units in area i, and $\hat{\tau}_{ij}$ is defined by the estimating equation $y_{ij} = \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}_{\hat{\tau}_{ij}}$.

Naive use of M-quantile regression modelling when data contain linkage errors leads to biased estimates of the true M-quantile fits. This is intuitively clear from the fact that the combined impact of natural variability as well as linkage error variability leads to conditional distributions at the different $\mathbf{x}_{ij}$ that are not those of interest. This is also clear from a cursory inspection of (11) and (12). Following the approach of Chambers (2009), we therefore modify the M-quantile normal Equations (39) to take account of the linkage error structure, using the notation introduced in Section 2. This leads to the modified M-quantile normal equations,

$$\sum_{i=1}^{D}\sum_{q\in i}\mathbf{X}^{*\prime}_{siq}\Upsilon_{siq\tau}^{-1/2}\psi_\tau\left\{\Upsilon_{siq\tau}^{-1/2}(\mathbf{y}^*_{siq} - \mathbf{X}^{*\prime}_{siq}\boldsymbol{\beta}^*_\tau)\right\} = \mathbf{0}, \tag{41}$$

where $\Upsilon_{siq\tau} = \text{diag}\left[\sigma^{*2}_\tau + (1-\lambda_{iq})\left\{\lambda_{iq}(f_{iqj\tau} - \bar{f}_{siq\tau})^2 + \bar{f}^{(2)}_{siq\tau} - \bar{f}^2_{siq\tau}\right\}\right]$, $\mathbf{f}_{siq\tau} = \{f_{iqj\tau}\} = \mathbf{x}'_{iqj}\boldsymbol{\beta}^*_\tau$, and $\bar{f}_{siq\tau}$, $\bar{f}^2_{siq\tau}$ denote the block q averages of the components of $\mathbf{f}_{siq\tau}$ and their squares respectively. Note that σ^*_τ here is the scale coefficient associated with the skewed residuals from the M-quantile regression line of order τ. Given $\Upsilon_{siq\tau}$, the solution to (41) can be obtained via IRWLS and is of the form

$$\tilde{\boldsymbol{\beta}}^*_\tau = \left(\sum_{i=1}^{D}\sum_{q\in i}\mathbf{X}^{*\prime}_{siq}\Upsilon_{siq\tau}^{-1/2}\mathbf{W}^*_{siq\tau}\Upsilon_{siq\tau}^{-1/2}\mathbf{X}^*_{siq}\right)^{-1}\left(\sum_{i=1}^{D}\sum_{q\in i}\mathbf{X}^{*\prime}_{siq}\Upsilon_{siq\tau}^{-1/2}\mathbf{W}^*_{siq\tau}\Upsilon_{siq\tau}^{-1/2}\mathbf{y}^*_{siq}\right), \tag{42}$$

where $\mathbf{W}^*_{siq\tau}$ is a diagonal matrix of weights defined by component-wise division of the vector $\psi_\tau\left\{\Upsilon_{siq\tau}^{-1/2}(\mathbf{y}^*_{siq} - \mathbf{X}^*_{siq}\tilde{\boldsymbol{\beta}}^*_\tau)\right\}$ by the vector $\Upsilon_{siq\tau}^{-1/2}(\mathbf{y}^*_{siq} - \mathbf{X}^*_{siq}\tilde{\boldsymbol{\beta}}^*_\tau)$. Similarly, given $\boldsymbol{\beta}^*_\tau$ and $\Upsilon_{siq\tau}$, a robust estimator of σ^{*2}_τ is

$$\tilde{\sigma}^{*2}_\tau = \left\{\sum_{i=1}^{D}\sum_{q\in i}tr(\mathbf{W}^*_{siq\tau})\right\}^{-1}\sum_{i=1}^{D}\sum_{q\in i}\left\{(\mathbf{y}^*_{siq} - \mathbf{x}^{*\prime}_{iqj}\tilde{\boldsymbol{\beta}}^*_\tau)'\mathbf{W}^*_{siq\tau}(\mathbf{y}^*_{siq} - \mathbf{x}^{*\prime}_{iqj}\tilde{\boldsymbol{\beta}}^*_\tau)\right\}. \tag{43}$$

The 'empirical' versions of $\tilde{\boldsymbol{\beta}}_{\tau}^{*}$ and $\tilde{\sigma}_{\tau}^{*2}$, which we denote by $\widehat{\boldsymbol{\beta}}_{\tau}^{*}$ and $\widehat{\sigma}_{\tau}^{*2}$ respectively, are then defined by iterating between (42) and (43).

In order to use the M-quantile approach for small area estimation, it is first necessary to estimate the M-quantile coefficients defined by the correctly linked sample values $\mathbf{y}_{siq}$, that is, the values $\widehat{\tau}_{iqj}$ such that $y_{iqj} = \mathbf{x}'_{iqj}\widehat{\boldsymbol{\beta}}_{\widehat{\tau}_{iqj}}$ for $j \in s_{iq}$. Unfortunately, replacing y_{iqj} by its linked value y^{*}_{iqj} leads to biased estimates of these coefficients. We therefore propose to use an approximation to $\widehat{\tau}_{iqj}$ based on linked data that are corrected for linkage error induced bias. Let $\widehat{\tau}^{**}_{iqj}$ satisfy $y^{*}_{iqj} = \mathbf{x}'_{iqj}\widehat{\boldsymbol{\beta}}_{\widehat{\tau}^{**}_{iqj}}$. We then define our linked data-based estimate of the M-quantile coefficient for $j \in s_{iq}$ as $\widehat{\widehat{\tau}}^{*}_{iqj} = (\lambda_{iq} - \gamma_{iq})\widehat{\tau}^{**}_{iqj} + \gamma_{iq}N_{iq}\widehat{\overline{\tau}}^{**}_{iq}$ where $\widehat{\overline{\tau}}^{**}_{iq} = n^{-1}_{iq}\sum_{k\in,s_{iq}}\widehat{\tau}^{**}_{iqk}$. If the values $\widehat{\tau}^{**}_{iqk}$ vary widely over (0, 1), a more stable approximation to the M-quantile coefficient for unit $j \in s_{iq}$ is $\widehat{\tau}^{*}_{iqj} \approx (\lambda_{iq} - \gamma_{iq})\widehat{\tau}^{**}_{iqj} + (\gamma_{iq}N_{iq})/2$. In any case, the estimated area-specific M-quantile coefficient is then computed as $\widehat{\tau}^{*}_{i} = n^{-1}_{i}\sum_{q\in i}\sum_{j\in s_{iq}}\widehat{\tau}^{*}_{iqj}$, and the M-quantile predictor of $\overline{Y}_i$ using linked data (the *M-quantile predictor) becomes

$$\widehat{\overline{Y}}^{*MQ}_{i} = N^{-1}_{i}\left\{n_i\overline{y}^{*}_{si} + (N_i - n_i)\overline{\mathbf{x}}^{*\prime}_{ri}\widehat{\boldsymbol{\beta}}^{*}_{\widehat{\tau}^{*}_{i}}\right\}. \tag{44}$$

An estimator of the MSE for $\widehat{\overline{Y}}^{*MQ}_{i}$ based on the linearisation approach of Chambers et al. (2014) can be written as

$$\widehat{MSE}_{\mathbf{A},M}(\widehat{\overline{Y}}^{*MQ}_{i}) = \widehat{V}_{\mathbf{A},M}(\widehat{\overline{Y}}^{*MQ}_{i}) + \widehat{B}^{2}_{\mathbf{A},M}(\widehat{\overline{Y}}^{*MQ}_{i}) + \widehat{V}_{\mathbf{A},M}(\widehat{\tau}^{*}_{i}). \tag{45}$$

Here, $\widehat{V}_{\mathbf{A},M}(\widehat{\overline{Y}}^{*MQ}_{i})$ is an estimator of the prediction variance of Equation (44), $\widehat{B}^{2}_{\mathbf{A},M}(\widehat{\overline{Y}}^{*MQ}_{i})$ is an estimator of its area-specific bias, and the last term is an approximation to the contribution to the MSE due to estimation of the area M-quantile coefficient τ^{*}_{i}. The derivation and the details on the components of this MSE estimator are provided in Section S.4 of the Supplementary Material.

6 | SIMULATION STUDIES

Design-based simulations allow us to evaluate the performance of SAE methods in the context of a real population and realistic sampling methods where we do not know the precise source of outlier contamination. From a finite population perspective, we believe that this type of simulation constitutes a more practical and appropriate representation of SAE performance.

The synthetic population underpinning the design-based simulation is based on the simulation experiment reported in Briscolini et al. (2018); it comes from the European Statistical System Data Integration project (ESSnet, McLeod et al., 2011) and from the Survey on Household Income and Wealth, Bank of Italy (SHIW), whose data are freely available in anonymous form. Specifically, the synthetic ESSnet population contains information on over 26,000 individuals including *name*, *surname*, *gender* and *date of birth*. Two new variables have been added to the original dataset: the *annual income* and the *domain* indicator. The latter comprises 18 areas resulting from aggregation of Italian administrative regions. The area population sizes range from 102 to 3262 with an average of 1407.

Following Briscolini et al. (2018), we carry out a realistic record linkage and SAE simulation experiment by perturbing the ESSnet dataset via the introduction of missing values and typos in some potential linking variables (name, surname, gender and date of birth). Moreover, for the purposes of the simulation study, annual income has been removed from the perturbed dataset and the

corresponding value of *annual consumption* obtained from the SHIW survey has been added. As far as record linkage is concerned we consider two scenarios. Scenario (*i*) is in agreement with the theoretical assumptions of Section 2, with linking variables available at both register level. Note that this is the level of information available for the application described in Section 7. In Scenario (*ii*), linking variables for register $\mathcal{Y}$ are available only for sampled units.

In both scenarios, the classical version of the probabilistic record linkage procedure described in Fellegi and Sunter (1969) and Jaro (1989) has been implemented, using the function `compare.linkage` of the package `RecordLinkage` in `R` (Sariyar & Borg, 2020) to link the perturbed dataset with the original register population using *surname* as key-variable and *age*, grouped in four categories, and *domain* as blocking variables. The domain indicator has been used as block in the linkage process to guarantee the assumption that both registers include an area identifier measured without error (see Section 2). Empirical estimates of λ_{iq} are then calculated following this linkage process. Finally, in order to guarantee stable probabilities of correct linkage we used the average of these λ_{iq} estimates from all areas represented in the same block as our value of λ_q.

Note that in Scenario (*i*), after the sample is drawn from register $\mathcal{Y}$, the linkage process is carried out by matching information at the entire $\mathcal{Y}$ and $\mathcal{X}$ registers level. This linking is complete, one to one and independent of the sample outcome. The proportion of correct links for the four categories of age are $\lambda_q = (0.86, 0.93, 0.88, 0.91)$. In Scenario (*ii*), again after the sample is drawn, the linkage process only uses matching information for the sampled part of register $\mathcal{Y}$ (so the linking process is restricted to the sample values of the matching variables). Linking remains one to one, but it is clearly not complete. This corresponds to a violation of the first assumption of Section 2, so that linkage and sampling cannot be exchanged, and is used to assess the robustness of the properties of the proposed small area predictors. Since the linkage in this second scenario depends on the realised sample, the proportion of correct links for the four categories of age varies from sample to sample.

The aim of the design-based simulation is to compare the performance of different estimators, and their MSE estimators, for mean consumption in each domain under repeated sampling from a fixed population using income as the auxiliary variable. A total of 1,000 independent random samples of size $n = 268$ are taken from the synthetic fixed population described above, by randomly selecting units in the 18 domains, with domain sample sizes set proportional to domain population sizes unless the resulting sample size is less than 5, in which case the domain sample size is set to 5.

Six different estimators are used for this purpose: the standard EBLUP, $\widehat{\overline{Y}}_i^{EBLUP}$ (Rao & Molina, 2015), which serves as a reference, its corrected version in case of linkage error $\widehat{\overline{Y}}_i^{*EBLUP}$, the REBLUP estimator of Sinha and Rao (2009), $\widehat{\overline{Y}}_i^{REBLUP}$, equation (25), and its corrected version $\widehat{\overline{Y}}_i^{*REBLUP}$, expression (29), the estimator based on M-quantile regression model $\widehat{\overline{Y}}_i^{MQ}$ (40) and its corrected version with linked data $\widehat{\overline{Y}}_i^{*MQ}$ (44). In all cases, the influence function ψ is a Huber-type function with tuning constant $c = 1.345$. For $\widehat{\overline{Y}}_i^{*EBLUP}$ we consider two different methods for estimating the area-specific random effects: prediction of the random effects as in expression (17) and prediction of the random effects neglecting the second addend on the right hand side of Equation (18). We denote this alternative estimator by $\widehat{\overline{Y}}_i^{**EBLUP}$. Also for $\widehat{\overline{Y}}_i^{*REBLUP}$ we assess the beahviour of two different methods for estimating the area-specific random effect: prediction of the random effects as in expression (27) and prediction of the random effects as in expression (30). The estimator that uses the second approach will be denoted by $\widehat{\overline{Y}}_i^{**REBLUP}$. For each estimator and for each small area, we computed the Monte Carlo estimate of the percentage of relative bias in absolute value (ARB) and the percentage of Relative Root MSE (RRMSE) and the corresponding efficiency. The relative bias in absolute value of an estimator $\widehat{\overline{Y}}_i$ of the actual mean $\overline{Y}_i$ is the absolute value of the average across simulations of the errors $\widehat{\overline{Y}}_i - \overline{Y}_i$ divided by the corresponding value of $\overline{Y}_i$, its RRMSE is the square root of the average across simulations of

TABLE 1 Across areas distributions of the percentage of relative bias in absolute value (ARB), RRMSE and efficiency (EFF) of predictors of small area means in design-based simulations with $n = 262$

		Summary of across areas distribution									
Predictor	**Indicator**	**Min**	**Q1**	**Median**	**Q3**	**Max**	**Min**	**Q1**	**Median**	**Q3**	**Max**
		Scenario (i)					Scenario (ii)				
$\hat{\bar{Y}}_i^{EBLUP}$	ARB	0.17	2.58	3.15	6.78	12.05	0.05	1.94	2.80	5.69	11.46
	RRMSE	5.08	6.16	7.40	8.17	12.91	5.59	6.43	8.08	8.48	14.35
	EFF	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
$\hat{\bar{Y}}_i^{*EBLUP}$	ARB	0.02	1.31	2.34	4.29	9.62	0.02	1.15	2.77	4.43	9.37
	RRMSE	4.00	5.03	6.45	7.56	11.26	4.47	5.67	6.62	7.68	10.79
	EFF	72.3	78.7	86.8	95.3	112.6	62.1	79.7	84.6	93.3	126.4
$\hat{\bar{Y}}_i^{**EBLUP}$	ARB	0.11	1.28	2.60	4.36	9.69	0.11	1.13	3.20	4.57	9.52
	RRMSE	3.91	4.89	6.40	7.48	11.28	4.24	5.54	6.42	7.54	10.82
	EFF	69.6	77.2	86.6	95.1	112.8	62.5	76.0	82.1	91.1	126.8
$\hat{\bar{Y}}_i^{REBLUP}$	ARB	0.36	1.83	2.62	3.68	10.55	0.77	1.75	3.42	4.18	9.72
	RRMSE	4.42	5.01	5.35	6.37	11.08	4.18	5.54	5.97	6.74	10.29
	EFF	43.1	68.6	81.3	101.8	112.6	47.0	66.5	85.7	97.1	120.6
$\hat{\bar{Y}}_i^{*REBLUP}$	ARB	0.29	1.35	2.32	2.90	11.74	0.28	1.16	2.55	3.33	12.17
	RRMSE	3.86	4.83	5.01	5.65	12.33	3.56	4.97	5.59	6.32	12.9
	EFF	30.4	63.9	79.7	91.3	123.3	27.8	71.0	79.3	91.5	151.4
$\hat{\bar{Y}}_i^{**REBLUP}$	ARB	0.12	1.08	2.10	2.96	11.79	0.13	1.46	2.17	3.18	12.39
	RRMSE	3.62	4.66	4.96	5.51	12.31	3.30	4.71	5.33	5.89	12.93
	EFF	28.0	61.8	78.7	91.9	123.2	25.7	62.5	78.4	90.5	151.5
$\hat{\bar{Y}}_i^{MQ}$	ARB	0.64	3.26	3.90	5.15	9.17	1.35	3.97	4.24	5.83	8.52
	RRMSE	5.28	6.12	7.47	10.23	11.71	5.70	7.04	8.32	11.04	11.91
	EFF	60.4	88.4	101.4	126.2	197.0	65.2	90.8	101.9	137.7	213.2
$\hat{\bar{Y}}_i^{*MQ}$	ARB	0.52	1.49	2.65	3.76	8.22	0.06	1.45	2.63	3.68	8.14
	RRMSE	4.83	5.27	6.72	9.25	10.75	4.93	5.80	7.02	9.64	10.37
	EFF	55.5	76.2	86.7	113.1	177.1	59.2	75.9	84.1	120.4	183.9

TABLE 2 Across areas distributions of the percentage of relative bias in absolute value (ARB) and RRMSE of MSE estimators in design-based simulations with $n = 262$

		Summary of across areas distribution									
MSE Estimator	**Indicator**	**Min**	**Q1**	**Median**	**Q3**	**Max**	**Min**	**Q1**	**Median**	**Q3**	**Max**
		Scenario (i)					Scenario (ii)				
$\widehat{MSE}(\hat{\bar{Y}}_i^{EBLUP})$	ARB	11.6	28.5	49.9	69.2	97.6	9.9	25.0	38.7	55.4	89.8
	RRMSE	57.5	83.4	95.6	109.4	124.0	46.7	68.0	76.3	84.9	104.5
$\widehat{MSE}_{\mathbf{A},M}(\hat{\bar{Y}}_i^{*EBLUP})$	ARB	0.6	5.2	8.4	29.7	52.1	1.5	7.6	18.7	30.4	54.9
	RRMSE	29.9	43.3	47.9	55.1	61.6	25.4	39.4	40.7	45.4	53.8
$\widehat{MSE}_{\mathbf{A},M}(\hat{\bar{Y}}_i^{**EBLUP})$	ARB	1.5	4.9	9.3	29.1	54.3	1.1	7.6	19.7	30.9	62.0
	RRMSE	29.9	43.2	48.0	55.1	61.6	25.4	39.3	40.7	45.4	54.3
$\widehat{MSE}(\hat{\bar{Y}}_i^{REBLUP})$	ARB	3.6	9.8	20.6	40.3	100.9	1.1	4.3	7.8	22.7	84.7
	RRMSE	27.0	41.5	50.1	59.9	77.9	21.8	38.0	45.0	52.3	71.4
$\widehat{MSE}_{\mathbf{A},M\|\mathbf{u}}(\hat{\bar{Y}}_i^{*REBLUP})$	ARB	7.4	18.2	23.9	47.3	108.2	0.5	17.8	32.0	42.9	135.4
	RRMSE	31.1	53.1	57.4	64.6	82.8	21.6	45.2	52.9	60.2	90.3
$\widehat{MSE}_{\mathbf{A},M\|\mathbf{u}}(\hat{\bar{Y}}_i^{**REBLUP})$	ARB	1.7	11.9	19.7	31.7	124.8	3.6	12.8	24.8	32.4	146.7
	RRMSE	21.1	39.7	50.4	54.4	90.5	23.2	45.9	52.0	59.9	95.0
$\widehat{MSE}(\hat{\bar{Y}}_i^{MQ})$	ARB	0.2	14.5	22.8	32.4	46.9	1.8	17.1	25.7	30.0	41.5
	RRMSE	36.1	40.3	51.3	64.5	93.3	34.8	39.1	48.7	59.9	92.7
$\widehat{MSE}_{\mathbf{A},M}(\hat{\bar{Y}}_i^{*MQ})$	ARB	2.1	12.6	28.6	60.9	114.7	0.7	3.9	16.1	20.2	26.4
	RRMSE	32.2	46.8	59.9	74.1	113.8	36.3	39.9	43.5	53.6	70.4

the squares of these errors, again divided by the value of $\overline{Y}_i$, and its efficiency (EFF) is the value of the ratio of the actual MSE of the predictor to the actual MSE of the corresponding EBLUP.

Table 1 shows the key percentiles of the across areas distribution of the percentage ARB, the percentage RRMSE and the efficiency, while Table 2 reports the key percentiles of the across areas distribution of the percentage ARB and the percentage RRMSE of the corresponding estimator of the MSE for these estimators. In Scenario (*i*) we see that estimators that allow for linkage error work well in terms of both bias and RRMSE compared with the unmodified EBLUP, REBLUP and M-quantile-based predictors that ignore linkage error. The $\widehat{\overline{Y}}_i^{*EBLUP}$, $\widehat{\overline{Y}}_i^{**EBLUP}$, $\widehat{\overline{Y}}_i^{*REBLUP}$ and $\widehat{\overline{Y}}_i^{**REBLUP}$ estimators perform best in terms of bias, whereas $\widehat{\overline{Y}}_i^{*REBLUP}$ and $\widehat{\overline{Y}}_i^{**REBLUP}$ are best in terms of RRMSE. From these results we conclude that estimators that correct for linkage error seem to offer the most balanced performance in terms of both bias and MSE for this population.

With reference to MSE estimation, Table 2 shows that the MSE estimator for $\widehat{\overline{Y}}_i^{*EBLUP}$ and $\widehat{\overline{Y}}_i^{**EBLUP}$ performs better than the MSE estimators of $\widehat{\overline{Y}}_i^{*REBLUP}$, $\widehat{\overline{Y}}_i^{**REBLUP}$ and $\widehat{\overline{Y}}_i^{*MQ}$ that are based on the linearisation method. Furthermore, the MSE estimator for $\widehat{\overline{Y}}_i^{**REBLUP}$ improves on the MSE estimator for $\widehat{\overline{Y}}_i^{*MQ}$ in terms of efficiency. The results from Scenario (*ii*) are in line with those in Scenario (*i*). The only relevant difference is that the levels of ARB and RRMSE increase by around 5–10%. This is due to the use of sample, rather than register, information in the linking process. To complete the analysis in Table 2, the relationship between the 'true' (empirical) RMSE of each estimator and its estimator for each area is shown in Figures S.1 and S.2 in the Supplementary Material, where boxplots are used to illustrate the variability in the RMSE ratio, defined as the ratio of the average estimated RMSE for each area to the true RMSE. We can see that the MSE estimators proposed for linkage error corrected estimators perform better than MSE estimators for the EBLUP, REBLUP and M-quantile-based predictors, and especially so in Scenario (*i*).

The performances of the various linked data-based small area predictors that we have described in this paper, as well as those of their corresponding MSE estimators, has also been assessed via a series of model-based simulations. In particular, these simulations are based on data linkage scenarios in accord with the assumptions of Section 2 considering both linkage errors and actual population outliers. The simulation results show higher efficiency and lower bias for the linkage error corrected predictors compared with traditional small area estimators. Moreover, we note that the superior outlier robustness of REBLUP and M-quantile-based estimators with respect to the EBLUP still holds true when there are only artificial outliers due to linkage errors and when both artificial and real outliers are present. The proposed MSE estimators work very well under linkage errors with artificial outliers only. When both artificial outliers and real outliers are present, the MSE estimators are moderately biased. These additional results are not presented here for the sake of brevity, but they are available in Section S.5 of the Supplementary Material.

7 | ESTIMATING AVERAGE EQUIVALISED INCOME FOR LABOUR MARKET AREAS IN CENTRAL ITALY

The Italian component of the European Survey on Income and Living Conditions (EU-SILC) is conducted by ISTAT, the National Statistical Office, with the aim of producing estimates for indicators of poverty and social exclusion and, more generally, of the living conditions of the population at national and regional (NUTS-2) levels. Estimates for smaller areas are not released as the sample size of such areas are often too small to allow for direct estimates of adequate precision. In this paper, we focus on using 2016 EU-SILC data to produce estimates of the average equivalised net income for Labour

Market Areas (LMAs) in central Italy. Equivalised net income is the income measure used by Eurostat in its suite of poverty indicators. LMAs are unplanned domains obtained as clusters of municipalities and defined after the 2011 Census on the basis of daily working commuting flows. Central Italy is made of four (NUTS-2) regions and 113 LMAs; data from the 2016 EU-SILC cover only 71 LMAs so that for 42 LMAs it is not possible to compute a direct estimate. For in-sample LMAs, sample sizes range from 22 to a maximum of 2351 for Rome, with a mean of 166.3 and a median of 103. The overall sample size is 11,808 individuals making up 5,126 households.

In order to produce unit level-based small area estimates, EU-SILC data have been linked by ISTAT to data from the Italian integrated archive of economic and demographic micro data (Garofalo, 2014). Specifically, the linking procedure can be summarised as follows: (*i*) the sample is drawn from the EU-SILC frame, which is made of municipality-based population registers (which we label as $\mathcal{Y}$); (*ii*) information obtained from the survey is then appended to the sampled units on $\mathcal{Y}$; (*iii*) the identifier variables from $\mathcal{Y}$ (known before sampling for all units in this register) are used to link the sample records in $\mathcal{Y}$ to records in the Italian integrated archive of economic and demographic micro data (register $\mathcal{X}$); (*iv*) the sampled records from $\mathcal{Y}$ are then released, with each sample record containing EU-SILC data as well as linked data from $\mathcal{X}$, together with an estimate of the probability of correct linkage.

The integrated archive (register $\mathcal{X}$) itself is defined by the integration of data from the 2011 Census with administrative sources such as the municipal population registers, the tax returns register, the central register of pensioners, social security and fiscal sources for workers, and the social security benefit register. Among the variables available from the integrated archive, we select the following as having good predictive power for equivalised net income: a proxy for equivalised income; dummy for nationality (Italian citizen); dummy for income earner; work intensity, measured as the number of months for which there is an indication of employment (in at least one of the registers making up the archive) divided by twelve; indicator variables for gender by five age classes, 0–14, 15–29, 30–49, 50–64, 65 or more. Our data then consist of the population means of these variables for the 113 small areas, as well as individual values linked to the EU-SILC sample data. An overall probability of correct linkage is available, and is specified as 0.97. As a consequence, we have only one block, since there is no information provided regarding different values of the probability of correct linkage for the blocking variables used.

The assumptions detailed at the beginning of Section 2 seem reasonable for this application. In particular, the linkage process can be considered to be non-informative—assumption i. –, as it is based on individual identifiers which we can assume to be independent of the target variable given the model covariates. In addition, the linkage is conducted within municipalities, so there is perfect identification of each small area in the linked data. The sampling design is non-informative—assumption ii. –, as EU-SILC is based on a classical two-stage clustered design, where municipalities are PSUs selected with probability proportional to size, and households are SSUs selected by simple random sampling within PSUs. As for assumption iii., the reference population of EU-SILC is all private households and their current members residing in the area at the time of data collection. Persons living in collective households and in institutions are excluded from the target population. The Italian integrated archive of economic and demographic micro data is also based on households; in addition, for this application, we have used the 2015 edition of the archive, so that all our analysis variables refer to the same time interval. We note in passing that the archive does have a small undercoverage, since it can only include individuals who appear in at least one of its constituent registers; such undercoverage is estimated to be approximately 0.13% and it is likely to affect the EU-SILC survey frame (municipality registers) in the same way.

We start by fitting a linear mixed model with normally distributed random effects to the EU-SILC linked data. In particular, we use a two-level random-intercepts model with random effects specified at the level of the LMAs. Figure S.3 in the Supplementary Material shows the normal probability plot of individual residuals (Battese et al., 1988) and estimated LMA random effects (Lange & Ryan, 1989) obtained from fitting this model. These plots indicate some departures from normality particularly in the tails of the distribution. This is confirmed by a Shapiro-Wilk normality test, which rejects the null hypothesis that the random effects follow a normal distribution ($p = 0.0003$) and by the Kolmogorov-Smirnov test, which rejects the null hypothesis that the individual residuals follow a normal distribution (p-value < 0.00001).

The distribution of the standardised residuals indicates the presence of potentially influential observations defined by large individual residuals ($|r| > 2$). These can be due to the presence of real outliers and/or of artificial outliers generated by linkage errors. As discussed in previous sections, the two types of outliers are difficult to distinguish. In this dataset, in view of the relatively large value of λ and the marked deviations from normality shown in Figure S.3, it is likely that the observed outliers are a mixture of both representative and artificial outliers. Therefore, adopting a prediction approach that takes into account linkage errors and bounds the influence of outlying observations seems worthy of investigation for the EU-SILC linked data.

Motivated by this preliminary analysis, we now use the *EBLUP, **EBLUP, *REBLUP, **REBLUP, *MQ-based predictors to estimate the average equivalised net income for LMAs in central Italy. In order to provide some context for these estimates, we also implement the classical EBLUP, REBLUP and MQ estimation approaches and use a set of diagnostics based on the requirement that model-based small area estimates should be both consistent with corresponding unbiased direct estimates, as well as more precise.

We assess whether the model-based estimates are 'close' to the direct estimates by computing the correlations between the direct estimates and the model-based estimates. We note that the estimates obtained by the different approaches appear to be generally consistent with the direct estimates, with correlations varying between 0.75 and 0.80. The corresponding correlations between the traditional predictors (EBLUP, REBLUP and MQ) and those proposed in the paper which take into account the linkage error (*EBLUP, **EBLUP, *REBLUP, **REBLUP and *MQ) are very high (around 0.99). This implies, as expected, that given the high probability of correct linkage—0.97—there are only small differences between the estimates obtained using traditional predictors and the new predictors.

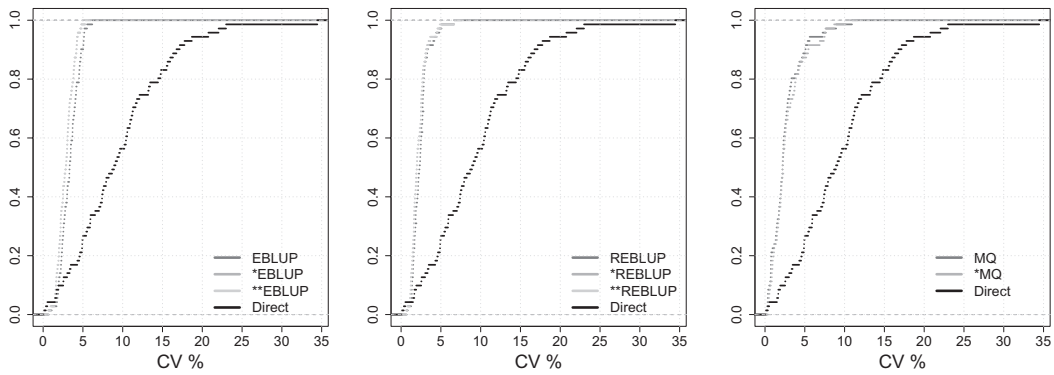

FIGURE 2 Empirical cumulative density functions for the estimated coefficient of variations of the small area estimators

However, the *EBLUP, **EBLUP, *REBLUP, **REBLUP and *MQ-based estimators all show potential gains in precision with respect to the EBLUP, REBLUP, MQ-based and the direct estimators.

In Figure 2 we compare the empirical cumulative density functions (ECDFs) of the coefficient of variation (CV) of all estimators. The first panel compares ECDFs of the estimated CVs for EBLUP, *EBLUP, **EBLUP with the ECDF of the CV of the direct estimator; the second and the third panels then focus on the CVs of the REBLUP and the MQ approaches. The CV of the direct estimator is computed accounting for the complex two-stage clustered sampling design employed for EU-SILC and using the `ReGenesees` `R` library described in Zardetto (2015). In the first panel, the ECDFs corresponding to *EBLUP and **EBLUP almost always dominate those for EBLUP and direct estimates, highlighting that CV values for the former approaches are lower than those estimated for the latter. It is only when the area sample size is large that this relation is inverted: CVs for the direct estimates are smaller than those for *EBLUP, **EBLUP, and EBLUP. This behaviour holds also in the second panel for the REBLUP-based estimates, while in the third panel, the MQ and *MQ-based predictors show an inverse behaviour in terms of ECDFs of CV values: in particular, MQ seems to be more precise than *MQ. This behaviour can be explained by the fact that the MSE estimator of the MQ-based predictor underestimates the real MSE in case of linkage error as shown by our model-based and the design-based simulations results.

Estimating the average equivalised net income at the LMA level enables us to investigate the gap in living conditions in the region of interest. Maps showing the estimates for the 113 LMAs are not displayed for reasons of space, but are available from the authors upon request. These show that income level varies considerably across the LMAs. There is also a clear North-South gradient: the LMAs in the southern part of the region are characterised by the lowest estimates of average equivalised income (between 10,000 and 15,000 Euros) and can be considered the most critical. Besides this North-South pattern, we also observe average equivalised net income values for larger cities and metropolitan areas (approx 20,000 Euros for Rome and 22,000 Euros for Florence) that are larger than those for neighboring LMAs.

8 | FINAL REMARKS

In this paper, we propose a number of small area estimation methods that allow for linkage error in the data. These proposed *EBLUP, **EBLUP, *REBLUP, **REBLUP and *MQ-based estimators have the potential to lead to significantly better small area estimates in important applications where linked data are available, such as in financial, economic, environmental and public health applications, and where outliers may also be present.

The properties of the proposed estimators have been studied through model-based and design-based simulation studies. The results from these studies suggest that these estimators represent a set of useful tools to allow for linkage error in SAE. In particular, the empirical results reported in Sections S.5 and 6 show that the proposed small area estimators are less biased and more efficient than the traditional predictors in the presence of artificial (i.e. linkage error-induced) and real outliers. In addition, the application reported in Section 7 provides evidence that the proposed approaches are viable options to deal with linkage error when small area estimation is conducted using survey data imprecisely matched with administrative data that provide useful auxiliary and, more importantly, proxy information. In addition, the performance of the proposed MSE estimators for these small area estimators seems promising, but we are aware that further research in this area is necessary. `R` code for calculating the *EBLUP, **EBLUP, *REBLUP, **REBLUP and *MQ-based estimators proposed in this paper and their corresponding MSE estimators can be obtained from: http://wileyonlinelibrary.com/journal/rss-datasets.

The approach to small area estimation using probability-linked data described in this paper is in the spirit of Scheuren and Winkler (1993), where it is suggested that one corrects the naive estimator using an estimate of its bias under an appropriate model for the linkage error process. In this paper, the adjustment we use for this purpose depends on assuming that linkage errors are generated via an ELE process and knowing the parameters (i.e. the λ_{iq}) that characterise this process. This is highly unlikely to be the case, and the probabilities λ_{iq} will usually be estimated. One way to estimate these parameters, suggested in Chambers (2009), is via access to a random 'audit' sample of the linked records in each block, where the only information required is whether a sampled link is correct or not. This could also be accomplished by calculating the achieved linkage error rate in a training set of 'gold standard' links, as would be possible if a classification-based approach to linkage is used (Chambers & Diniz da Silva, 2020).

Lahiri and Larsen (2005) consider the probabilities of correct linkage as parameters of a mixture model and estimate them using the expectation maximisation (EM) algorithm. In general, we can think of these estimated probabilities as part of the paradata for the linkage process, which should be made available to secondary users of the linked data (Gilbert et al., 2018). However, as discussed in Section 2, it is unrealistic to expect reliable estimates of λ_{iq} in the context of SAE where the sample size in area i and block q can be very small, or even 0. For this reason, we have assumed that accurate estimates of the probabilities of correct linkage are available at block level, perhaps by averaging over areas ($\widehat{\lambda}_q = n_q^{-1} \sum_{i=1}^{D} \widehat{\lambda}_{iq} n_{iq}$), or even as a single (overall average) $\widehat{\lambda}$, as in the application presented in Section 7. Such estimates can be substituted into the expression for the proposed small area estimators in Sections 3, 4 and 5. In order to assess the performance of the proposed small area estimators in this case (i.e. when linkage error rates are estimated), we have replicated the model-based simulations of Section S.5 with λ_q estimated by independently selecting a random 'audit' sample of linked records of 25 units in each block. The results in this case show a very small increase in the empirical variability of the proposed *EBLUP, **EBLUP, *REBLUP, **REBLUP and *MQ-based estimators. Interested readers can contact the authors to access these more detailed results.

The extra uncertainty arising from the estimation of the probabilities λ_{iq} needs to be accounted for when carrying out MSE estimation for the small area estimators that use $\mathbf{A}_{iq}$ to correct for bias induced by linkage errors. This extra uncertainty can be taken into account in the estimated MSE of $\widehat{\overline{Y}}_i^{*EBLUP}$ by adding a term $g_{4i}(\widehat{\boldsymbol{\delta}}^*, \widehat{\lambda}_{iq})$ to expression (24):

$$g_{4i}(\widehat{\boldsymbol{\delta}}^*, \widehat{\lambda}_{iq}) = tr\left\{ \frac{\partial \mathbf{b}_i'}{\partial \lambda_{iq}} \sum_{q \in i} \widehat{\Sigma}_{siq} \widehat{V}(\widehat{\lambda}_{iq}) \left(\frac{\partial \mathbf{b}_i'}{\partial \lambda_{iq}} \right)' \right\},$$

where $\widehat{V}(\widehat{\lambda}_{iq})$ is an estimator of the variance of the estimators of the probabilities of correct linkage. If the estimates of the linkage probabilities are obtained via an 'audit' sample, $\widehat{V}(\widehat{\lambda}_{iq}) = n_{iq}^{-1} \widehat{\lambda}_{iq}(1 - \widehat{\lambda}_{iq})$. In this case, the estimator of the MSE of $\widehat{\overline{Y}}_i^{*EBLUP}$ becomes

$$\widehat{MSE}_{\mathbf{A},M}(\widehat{\overline{Y}}_i^{*EBLUP}) = (1 - n_i/N_i)^2 \left\{ g_{1i}(\widehat{\boldsymbol{\delta}}^*, \widehat{\lambda}_{iq}) + g_{2i}(\widehat{\boldsymbol{\delta}}^*, \widehat{\lambda}_{iq}) + 2g_{3i}(\widehat{\boldsymbol{\delta}}^*, \widehat{\lambda}_{iq}) + g_{4i}(\widehat{\boldsymbol{\delta}}^*, \widehat{\lambda}_{iq}) \right\} + o(D^{-1}).$$

As far as estimation of the MSE of $\widehat{\overline{Y}}_i^{*REBLUP}$ defined by (29) is concerned, we note that the component $h_{1i}(\widehat{\boldsymbol{\theta}})$ of equation (38) now becomes

$$h_{1i}(\widehat{\boldsymbol{\theta}}) = \left(1 - \frac{n_i}{N_i}\right)^2 [\overline{\mathbf{x}}_{ri}^{*\prime} | \mathbf{z}_i' | \mathbf{1}_{iq}'] \widehat{V}_{\mathbf{A},M|\mathbf{u}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}_i) [\overline{\mathbf{x}}_{ri}^{*\prime} | \mathbf{z}_i' | \mathbf{1}_{iq}']',$$

where $\widehat{\Lambda}_i$ denotes the vector defined by the area-block-specific values of λ_{iq} and $\widehat{V}_{\mathbf{A},\,M|\mathbf{u}}(\widehat{\boldsymbol{\theta}},\,\widehat{\Lambda}_i)$ is the estimated joint variance of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\Lambda}_i$ obtained by computing the asymptotic variance of solutions to the estimating equations. Using the same approach, we note that the first component of the MSE estimator (S.5) of the M-quantile-based estimator (44) also depends on the extra uncertainty arising from estimation of the probabilities of correct linkage and so needs to be written as

$$(1-n_i/N_i)^2\left\{[\overline{\mathbf{x}}^{*\prime}_{ri}|\mathbf{1}'_{iq}]\widehat{V}_{\mathbf{A},M}(\widehat{\boldsymbol{\beta}}^{*}_{\tau^{*}_i},\widehat{\Lambda}_i)[\overline{\mathbf{x}}^{*\prime}_{ri}|\mathbf{1}'_{iq}]'\right\}.$$

The development of corresponding MSE estimators for **EBLUP, **REBLUP is straightforward. The performance of the MSE estimators for *EBLUP, **EBLUP, *REBLUP, **REBLUP and the *MQ-based estimator when there is extra uncertainty arising from the estimation of probabilities of correct linkage is an area of current research.

Despite the fact that the linkage error corrected SAE estimation methods proposed in this paper provide encouraging results, further research remains to be done. In particular, we have assumed that both registers include an area identifier measured without error that is used in the linkage process. Consequently, we do not allow units from different areas to be erroneously linked. When this assumption is relaxed (i.e. linkage errors across areas are allowed), then within area heterogeneity is increased, with between area heterogeneity consequently decreased, leading to biases in estimation of model variance components and correlations between different units in different areas. This correlation needs to be taken into account when calculating *EBLUP, *REBLUP and *MQ. We are currently working on this scenario, with preliminary results showing that *MQ performs better than *EBLUP and *REBLUP. A possible explanation is that the estimation of model parameters in the M-quantile approach is independent of the area indicator, which only becomes relevant when computing area-specific M-quantiles coefficients.

Related to the previous topic, note that if a random slopes specification is of interest, so that values in $\mathbf{Z}$ vary within areas, then we need to replace $\mathbf{Z}_{siq}$ in Equation (12), and in all subsequent expressions based on this identity, by $\mathbf{Z}^{*}_{siq}=E_{\mathbf{A}}(\mathbf{A}_{siq}\mathbf{Z}_{iq})=\mathbf{T}_{siq}\mathbf{Z}_{iq}=\{(\lambda_{iq}-\gamma_{iq})\mathbf{Z}_{siq}+\gamma_{iq}N_{iq}\mathbf{1}_{n_{iq}}\overline{\mathbf{z}}'_{iq}\}$ where $\overline{\mathbf{z}}_{iq}$ is the vector of column means of $\mathbf{Z}_{iq}$. In this case also $V_{\mathbf{A},M}(\mathbf{y}^{*}_{siq})$ requires modifications that will be discussed in future research.

A key assumption of the paper is that linking variables are available at registers $(\mathcal{X},\,\mathcal{Y})$ level, as it is the case for the ISTAT linking procedure underpinning the application in Section 7. This may not be always the case, and it may be possible that linking variables from register $\mathcal{Y}$ are available only for sampled units. This violates our key assumption that sampling and linkage are independent, and the theory outlined in this paper then requires modification to allow for such an informative linkage situation. Though interesting, we believe that this case is not in line with our secondary user perspective since it implicitly assumes that the analyst and the linker are the same. However, it could be an interesting direction for future research.

Further research is also needed to investigate the performance of the proposed robust predictors when the errors due to representative outlying values are highly skewed (i.e. due to mean-shifted representative outliers). In this case, the robust-projective estimators described in this paper will typically have a low prediction variance, but may exhibit a large prediction bias. The behaviour of robust-projective estimators under skewed error distributions with outliers has been investigated by Chambers et al. (2014) in simulation scenarios without linkage errors. To reduce the bias impact when there are mean shift representative outliers, these authors propose a robust-predictive approach to SAE that attempts to partially adjust for the contribution of the population outliers to the population quantity of interest. A

full exploration of the behaviour of SAE estimators when there are mean-shifted representative outliers, as well as linkage errors, is beyond the scope of this paper, but is another interesting direction for future research.

Finally, we note once more that we have assumed a simple exchangeable linkage error model because we focus on SAE carried out by a secondary data analyst. It would be interesting to extend our estimators to situations where the information about the linkage process is richer and a different viewpoint can be adopted, as in Briscolini et al. (2018) and Han and Lahiri (2018).

ACKNOWLEDGEMENTS

The work of Nicola Salvati has been carried out with the support of the project InGRID 2 (Grant Agreement N. 730998, EU) and of project PRA2018-9 (From survey-based to register-based statistics: a paradigm shift using latent variable models). The work of Maria Giovanna Ranalli is conducted within the ISTAT project LABINN 'Integrazione di dati provenienti da più fonti per il calcolo di indicatori socio-economici a livello comunale' and has been supported by the project 'Uso di Metodi di Stima per Piccole Aree per la produzione corrente di indicatori del BES—Benessere Equo e Sostenibile—a livello locale da indagini ISTAT e informazioni da registro.'—Ricerca di Base 2019, University of Perugia. The authors are also grateful to the Editor, Associate Editor and referees for their comments and suggestions.

REFERENCES

Abbott, O., Jones, P. & Ralphs, M. (2016) *Methodological developments in data linkage*. Large scale linkage for total populations in official statistics, UK: John Wiley & Sons, Ltd, pp. 170–200.

Battese, G.E., Harter, R.M. & Fuller, W.A. (1988) An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28–36.

Bera, A., Bilias, Y. & Simlai, P. (2006) Estimating functions and equations: An essay on historical developments with applications to econometrics. *Palgrave handbook of econometrics*, vol. 1. Estimating Functions and Equations: An Essay on Historical Developments with Applications to Econometrics, New York: Palgrave MacMillan, pp. 427–476.

Bianchi, A., Fabrizi, E., Salvati, N. & Tzavidis, N. (2018) Estimation and testing in m-quantile regression with applications to small area estimation. *International Statistical Review*, 86, 541–570.

Booth, J. & Hobert, J. (1998) Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 262–272.

Breckling, J. & Chambers, R. (1988) M-quantiles. *Biometrika*, 75 (4), 761–771.

Briscolini, D., Consiglio, L.D., Liseo, B., Tancredi, A. & Tuoto, T. (2018) New methods for small area estimation with linkage uncertainty. *International Journal of Approximate Reasoning*, 94, 30–42.

Chambers, R. (1986) Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063–1069.

Chambers, R. (2009) Regression analysis of probability-linked data. *Technical report*, Official Statistics Research, Statistics New Zealand. Retrieved from: http://archive.stats.govt.nz/~/media/Statistics/about-us/statisphere/Files/official-statistics-research-series/osr-series-v4-2009-regression-analysis-probability-linked-data.pdf.

Chambers, R. & Diniz da Silva, A. (2020) Improved secondary analysis of linked data: A framework and an illustration. *Journal of Royal Statistical Society, Series A*, 183 (1), 37–59.

Chambers, R. & Tzavidis, N. (2006) M-quantile models for small area estimation. *Biometrika*, 93 (2), 255–268.

Chambers, R., Chandra, J. & Tzavidis, N. (2011) On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, 37 (2), 153–170.

Chambers, R., Chandra, H., Salvati, N. & Tzavidis, N. (2014) Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B*, 76 (1), 47–69.

Das, S., Chandra, H. & Chambers, R. (2017) Robust mean squared error estimation for ELL based poverty estimates under heteroskedasticity—an application to poverty estimation in Bangladesh. *Statistics and Applications*, 16, 375–397.

Dygaszewicz, J. (2012) Modern census in Poland. In *United Nations International Seminar on Population and Housing Censuses: Beyond the 2010 Round, November 2012*, Seoul, Republic of Korea.

Fellegi, I. & Sunter, A. (1969) A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183–1210.

Fellner, W.H. (1986) Robust estimation of variance components. *Technometrics*, 28 (1), 51–60.

Garofalo, G. (2014) Il progetto archimede obiettivi e risultati sperimentali (in italian), Istat. Working papers, 9

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L., Smith, P. et al. (2018) Guild: Guidance for information about linking data sets. *Journal of Public Health*, 40(1), 191–198.

Han, Y. (2018) *Statistical inference using data from multiple files combined through record linkages*. Ph.D. Dissertation, College Park: University of Maryland.

Han, Y. & Lahiri, P. (2018) Statistical analysis with linked data. *International Statistical Review*, 87, S139–S157.

Harron, K. (2016) Introduction to data linkage. Technical report, Administrative Data Research Network Publication. Retrieved from: https://adrn.ac.uk/media/1324/datalinkage.pdf

Harville, D. A. & Jeske, D. R. (1992) Mean square error of estimation or prediction under a general linear model. *Journal of the American Statistical Association*, 87, 724–731.

Haslett, S. (2016) Small area estimation using both survey and census unit record data. *Analysis of poverty data by small area estimation*. Small Area Estimation Using Both Survey and Census Unit Record Data. UK: John Wiley & Sons, Ltd, pp. 327–348.

Henderson, C. R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31 (2), 423–447.

Huber, P. (1981) *Robust statistics*. New York: Wiley.

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

Jiongo, V.D., Haziza, D. & Duchesne, P. (2013) Controlling the bias of robust small-area estimators. *Biometrika*, 100, 843–858.

Kelman, C., Bass, A. & Holman, C. (2002) Research use of linked health data—a best practice protocol. *Australian and New Zealand Journal of Public Health*, 26, 251–255.

Kim, G. & Chambers, R. (2012) Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, 56, 2756–2770.

Kim, G. & Chambers, R. (2015) Unbiased estimation in the presence of correlated linkage error. *Statistics*, 4, 32–45.

Lahiri, P. & Han, Y. (2017) Small area estimation with linked data. *Presented at the SAE 2017 ISI Satellite Meeting*, Paris, July 10-12.

Lahiri, P. & Larsen, M. (2005) Regression analysis with linked data. *Journal of the American Statistical Association*, 100 (469), 222–230.

Lange, N. & Ryan, L. (1989) Assessing normality in random effects models. *The Annals of Statistics*, 17 (2), 624–642.

McLeod, P., Heasman, D. & Forbes, I. (2011) Simulated data for the on the job training. ESSnet DI. Retrieved from: https://ec.europa.eu/eurostat/cros/content/job-training_en [Accessed 9 November 2020].

Molina, I. & Rao, J.N.K. (2010) Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38 (3), 369–385.

Pfeffermann, D. (2013) New important developments in small area estimation. *Statistical Science*, 28 (1), 40–68.

Prasad, N.G.N. & Rao, J.N.K. (1990) The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85 (409), 163–171.

Rao, J.N.K. (2003) *Small area estimation*. New York: Wiley.

Rao, J.N.K. & Molina, I. (2015) *Small area estimation*, 2nd edn. New York: Wiley.

Richardson, A.M. & Welsh, A.H. (1995) Robust restricted maximum likelihood in mixed linear models. *Biometrics*, 51 (4), 1429–1439.

Samart, K. & Chambers, R. (2014) Linear regression with nested errors using probabilitylinked data. *Australian and New Zealand Journal of Statistics*, 56 (1), 27–46.

Sariyar, M. & Borg, A. (2020) RecordLinkage: Record linkage functions for linking and deduplicating data sets, R package version 0.4-12.1. Retrieved from: https://CRAN.R-project.org/package=RecordLinkage.

Scheuren, F. & Winkler, W. (1993) Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39–58.

Scheuren, F. & Winkler, W. (1997) Regression analysis of data files that are computer matched—part ii. *Survey Methodology*, 23, 157–165.

Schulte Nordholt, E. (2009) Data integration activities on the way to the Dutch virtual census of 2011. Proceedings of MSP2009 Modernisation of Statistics Production. Retrived from: http://www.scb.se/Grupp/ProdukteTjanster/Kurser/ModernisationWorkshop/finalpapers/G

Sinha, S.K. & Rao, J.N.K. (2009) Robust small area estimation. *The Canadian Journal of Statistics*, 37 (3), 381–399.

Swiss Federal Statistical Office (2012) The Swiss census system: A comprehensive system of household & person statistics. Technical report, Paper presented at UNECE Conference of European Statisticians, Paris, June 2012.

Tzavidis, N., Marchetti, S. & Chambers, R. (2010) Robust estimation of small area means and quantiles. *Australian and New Zealand Journal of Statistics*, 52 (2), 167–186.

Winkler, W.E. (2009) Record Linkage. *Sample surveys: design, methods and applications*. Handbooks of Statistics, vol. 29A. New York: Elsevier B.V, pp. 351–380.

Winkler, W.E. (2014) Matching and record linkage. *WIREs Computational Statistics*, 6, 313–325.

Zardetto, D. (2015) Regenesees: An advanced R system for calibration, estimation and sampling error assessment in complex sample surveys. *Journal of Official Statistics*, 31, 177–203.

Zhang, L.C. & Chambers, R.L. (2019) *Analysis of integrated data*. Boca Raton: CRC Press.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Salvati N, Fabrizi E, Ranalli MG, Chambers RL. Small area estimation with linked data. *J R Stat Soc Series B*. 2021;83:78–107. https://doi.org/10.1111/rssb.12401