4

Small Area Methods and Administrative Data Integration

Li-Chun Zhang[1] and Caterina Giusti[2]
[1]*S3RI/University of Southampton, Southampton, UK and Statistics Norway, Oslo, Norway*
[2]*Department of Economics and Management, University of Pisa, Pisa, Italy*

4.1 Introduction

The literature of Small Area Estimation (SAE) is dominated by sample-survey-based applications, where administrative register data are used as covariates in the models. Register-based statistics, however, are becoming more and more common, and integration of survey and administrative data can raise many distinct issues. Compared with sample survey data, an important advantage of the register data is that statistics can be produced at much more detailed aggregation levels. For instance, register-based census have been carried out in a number of European countries. Statistical measures of uncertainty, however, are rarely produced for register-based statistics, partly due to a lack of theoretical developments, partly due to the complexity of the errors involved (Zhang 2012). These issues are particularly relevant also for the production of poverty and well-being indicators. In many countries data coming from large sample surveys, such as the Labour Force Survey (LFS) or the EU-SILC (EU Statistics on Income and Living Conditions), can be complemented or integrated with data coming from several population registers to obtain either more accurate estimates at the local level, or multidimensional indicators that could not have been produced using each source on its own.

We shall characterize the settings of SAE which involve administrative data according to (a) whether there are relevant additional sample survey data present and, if so, (b) how the target measure is related to the available data in the two sources.

The term register-based is often understood to refer to statistics that are produced by tabulation of statistical registers processed purely from administrative data. It may be that no

Analysis of Poverty Data by Small Area Estimation, First Edition. Edited by Monica Pratesi.

Companion Website: www.wiley.com/go/pratesi/poverty

additional survey data are available at all, such as when health statistics are exclusively compiled based on clinical or pharmaceutic records. Sometimes, relevant survey data are available, but are only used 'indirectly' to define the processing rules and/or to assess the accuracy of the register data but not to adjust them. For example, analysis of past census and sample survey household data may help to define the 'rules' by which administrative data are combined to construct the statistical register households in a register-based census (Zhang 2011). As another example, to size the inter-regional over-/under-coverage of the register-based population enumeration, additional questions were administered in the Norwegian LFS in the 4th quarter of 2011. Table 4.1 cross-tabulates the 1029 persons (out of over 20 000 LFS respondents) who have reported a different address in the LFS to the Population Register. Further analysis reveals two causes that dominate the discrepancy. For instance, out the 150 persons in column I, 100 have permanently moved to another address and 32 of the rest are students. The former demonstrates the time lag that exists in many registers, which will be discussed in Section 4.2.2. The latter reflects potential relevance error, as the population registration abides by different regulations compared with the traditional census residence definition. We discuss the relevance error in Section 4.3.2.

It is important to emphasize that combining data from multiple sources is generally necessary for the register-based statistics. In particular, integration with one or several base registers (Wallgren and Wallgren 2006), including Population Register, Business Register and Immobility Register of building, property and land, is almost always required to obtain the target population frame and to improve the data quality. For instance, part of the input to register-based education statistics are university exam results, which may be utilized for deriving, say, the variable highest level of education. Matching these data to the Population Register is necessary in order to verify the in-scope target population, and it helps to combine multiple exam results of the same person and check their plausibility, and so on, and it enables these data to be 'linked' with other relevant sources of educational data for the stated purpose. Chapter 3 of this book focuses on methods to integrate survey and register data with the aim of producing measures of well-being at the macro or micro level.

A different setting is when sample survey data are available and considered to provide the target measure. This is by and large the commonly treated scenario in the SAE literature, where

Table 4.1 Region of residence by register enumeration and LFS

Region	Region in register enumeration								
in LFS	Missing	I	II	III	IV	V	VI	VII	Total
Missing	1	31	10	29	22	40	20	19	172
I	0	116	15	24	14	8	5	8	190
II	0	5	53	1	1	3	1	1	65
III	3	11	1	90	4	2	2	4	117
IV	1	4	0	2	94	5	1	5	112
V	0	6	2	7	14	118	6	3	156
VI	0	5	2	7	2	11	58	10	95
VII	1	3	0	4	2	4	5	103	122
Total	5	150	73	135	131	151	78	134	1029

Source: Internal quality report, Norwegian register-based Census (2011).

the administrative data supply the covariates of the model. A special situation arises when the administrative source contains a proxy to the target survey measure. To distinguish it from the other auxiliary variables, we loosely define a proxy measure to be a variable that is similar in definition and has the same support compared with the target variable. For instance, while variables such as age, sex, education, income, and so on are auxiliary variables to the binary unemployment status, the binary register-based job-seeker status is a proxy measure. But the job-seeker status is not a proxy variable to the activity status defined (employed, unemployed, inactive) because the two have different support.

Statistical registers are rich sources of concurrent proxy measures that have complete (or virtually complete) coverage. A proxy variable is typically the most powerful among all the auxiliary variables for regression modelling of sample survey data. But it also enables another perspective, namely to adjust the relevance error inherent in the proxy variable that has complete coverage by the target survey variable that is only available for a sample of the population, where the sample size is typically small in the SAE context.

In short, the two settings to be considered in more detail are: Section 4.2, register-based SAE without any survey data; and Section 4.3, SAE based on integration of sample survey and register data. The topics will be organized by the nature of the predominant error of concern—see Zhang (2012) for a total-error framework for data integration. These include sampling error, measurement error arising from progressive administrative data, coverage error, relevance error and probability linkage error. The discussions will be focused on their relevance to the different settings of SAE, in such a way that the general concepts and issues involved are applicable to the production of poverty and well-being indicators.

The key message we wish to convey is that there are considerable challenges for SAE based on administrative data integration beyond conventional regression modelling of the sample survey data. With or without sample surveys, there are many important areas of application for administrative data, including the production of poverty and well-being indicators, but also plenty of problems that need to be solved.

4.2 Register-based Small Area Estimation

4.2.1 Sampling Error: A Study of Local Area Life Expectancy

When the target parameter is of a theoretical nature, such as life expectancy, disease prevalence or well-being, the actual finite-population can be conceived as a sample from an appropriately defined infinite theoretical or super-population, and the register-based direct domain (or sub-population) counts the direct sample estimates. Despite the total sample size being very large, the effects of sampling errors will be noticeable at very detailed levels.

For example, domain mortality rates are needed in order to calculate life expectancy (Chiang 1984) at a disaggregated level. Let the domains be classified according to: (1) area, denoted by $i = 1, \ldots, m$ for fixed m; (2) sex, denoted by $j = 1$ for male and $j = 2$ for female; and (3) age, denoted by $a = 0, \ldots, 99$. (The few people over 99 are grouped with those of the age 99.) Suppose that the domain-specific number of deaths over a given period is available in the relevant administrative register, denoted by y_{ija}. Since the domain population size is not constant over the period during which the death records are accumulated, some kind of a hypothetical *equivalent* population size is needed, denoted by n_{ija}.

Poisson distribution of y_{ija} seems natural, with parameter $\lambda_{ija} = n_{ija}\tau_{ija}$, where τ_{ija} is the theoretical domain mortality rate. A *direct* estimator of τ_{ija} is then given by

$$\hat{\tau}_{ija} = y_{ija}/n_{ija}$$

But $\hat{\tau}_{ija}$ can be highly unstable, yielding many extreme mortality rates in the smallest population domains. A simple alternative is the *synthetic* estimator given by

$$\hat{\tau}^S_{ija} = \left(\sum_{i=1}^{m} y_{ija}\right) / \left(\sum_{i=1}^{m} n_{ija}\right) = \xi_{ja}$$

yielding a single mortality rate for each sex-age group without any between-area variation.

SAE of the domain mortality rates is thus needed. On the one hand, this should reduce the large variance of the direct estimators in order to bring stability over time and to avoid an implausibly huge range of the local area life expectancy. On the other hand, this should avoid the evident over-smoothing of the synthetic estimator.

Let each sex-age group form a cohort. Initially, within each cohort $h = (j, a)$, the observed area-specific mortality rates can be smoothed to yield estimates of $\{\tau_{hi}; i = 1, \ldots, m\}$. Repeating the same procedure separately in each cohort yields then the estimates of all $\{\tau_{hi}; i = 1, \ldots, m$ and $h = 1, \ldots, H\}$. This *basic* smoothing approach has a theoretical drawback because the estimates of $V(\tau_{hi})$ do not necessarily vary smoothly over the 'neighbouring' cohorts for fixed i. The neighbouring cohorts may be the neighbouring age groups, either for a given sex or when both sexes are considered pairwise. Yet there appears to be no reason *a priori* why one should expect such 'jerky' dispersions.

A variance-component model for domain relative risk was developed in ESSnet SAE (2011, pp. 86–109) to address the problem. As commonly found in the literature of disease mapping (e.g. Rao 2003, section 9.5), let $\theta_{hi} = \tau_{hi}/\xi_h$ be the *relative risk* (RR), or the standardized mortality rate (SMR), where ξ_h denotes the cohort mortality rate that is calculated using the data from all the areas and treated as fixed. The variance-component model is given as

$$\begin{aligned} &y_{hi}\,|\theta_{hi} \sim \text{Poisson}\,(\lambda_{hi}) \quad \text{where} \quad \lambda_{hi} = \mu_{hi}\theta_{hi} \quad \text{and} \quad \mu_{hi} = n_{hi}\xi_h \\ &\theta_{hi} = \psi_h\psi_{hi} \quad \text{where} \quad E(\psi_h) = E(\psi_{hi}) = 1 \quad \text{and} \quad V(\psi_h) = \sigma^2_\psi \quad \text{and} \quad V(\psi_{hi}) = \sigma^2_h \end{aligned} \tag{4.1}$$

and ψ_h and ψ_{hi} are independent of each other, and ψ_{hi} and ψ_{hj} are independent of each other for $i \neq j$. The domain RR θ_{hi} is then the product of a cohort-level random effect ψ_h and a cohort-domain-level random effect ψ_{hi}. The basic separate-cohort smoothing approach corresponds to the special case of $\sigma^2_\psi = 0$. In general $V(\psi_{hi}) = \sigma^2_h$ is allowed to vary across the cohorts, possibly with a functional expression. A special case is $\sigma^2_h = \sigma^2$, which is referred to as the *variance homogeneity* assumption. Notice that the data across all the domains will be used to estimate the model (4.1).

The variance-component model (4.1) was applied in a case study of the life expectancy across the Norwegian municipalities $i = 1, \ldots, m$. An approach based on moving-average with different choices of the window width was explored for the estimation of σ^2_h, in addition to the variance homogeneity assumption. Some results for the quantiles of the resulting estimated life expectancy across the over 400 municipalities are given in Table 4.2. For the present context, we notice only that the direct register-based estimates have an implausibly huge range across

Table 4.2 Quantiles of estimated life expectancies across the municipalities. Basic: Separate basic smoothing in each cohort. Neighbour: Neighbouring variance homogeneity assumption and moving average variance estimator of bandwidth 50. Global: Variance homogeneity assumption

Sex	Method	Minimum	25% Quantile	Median	75% Quantile	Maximum
Male	Direct Estimator	62.2	76.7	78.0	79.1	83.6
	Basic	77.6	77.9	78.0	78.0	78.8
	Neighbour	76.9	77.6	78.0	78.4	79.8
	Global	75.1	77.3	78.0	78.6	80.7
Female	Direct Estimator	75.2	81.8	82.8	83.9	88.3
	Basic	82.1	82.5	82.6	82.6	82.9
	Neighbour	81.1	82.4	82.6	82.9	84.1
	Global	79.9	82.1	82.6	83.3	85.3

Source: ESSnet SAE (2011, p. 101).

the municipalities. The basic cohort-specific smoothing more or less wipes out all the potential between-area variation. The estimated cohort variance components $\hat{\sigma}_h^2$ are especially jerky for the lower age groups, whilst being zero in many cohorts which causes over-shrinkage in those cohorts. The variance-component model provides an attractive alternative for the smoothing of the direct register-based estimates. In practice, the appropriate degree of smoothing may be decided together with the demographers.

4.2.2 Measurement Error due to Progressive Administrative Data

It is customary that sample survey or census data are collected over a specified period of time for the field operation, after which the observations become static and no longer change. This is not the case with administrative data. Most administrative data are event-triggered/-based, such as birth or death, taking an exam or graduation, paying tax, hiring an employee, and so on. The mandatory registration of the event is often self-administered. Delays and mistakes are not avoidable entirely, whether by allowance or negligence. Measures from the administrative sources are thus progressive in the sense that the 'observed' value for a given statistical reference time point t may differ depending on the measurement time point $t + s$, for $s \geq 0$, where s may be referred to as the measurement delay.

As explained in Zhang (2012), the progressiveness of the input data source can affect either the measurement (i.e. variables) of the secondary integrated data or the representation (i.e. units and population frame), depending on the process of data integration. For example, a delay in the reporting of a new employee may cause a measurement error of the register-based activity status, as along as the target population is defined by the Population Register which nevertheless contains the person. Whereas a delay in the registration of a newly finished building may cause an under-coverage error in the register-based Building Statistics.

Under-coverage caused by reporting delay has been studied for example for epidemiological, insurance, and product warranty applications. Hedlin *et al.* (2006) applied a log-linear type of model to estimate the reporting delays for the introduction of birth units to the Business Register. Linkletter and Sitter (2007) used a non-parametric method to estimate and adjust

for delays in Natural Gas Production reports in Texas. Similar issues can arise in SAE applications, such as utilizing the VAT register for short-term business statistics, where the numerous industrial groups (e.g. at the 5-digit NACE level) form the domains of interest.

We are currently unaware of any established register-based SAE application that deals with the coverage error arising from progressive administrative data. Coverage errors in the context of population size estimation will be discussed in Section 4.3.1, where the setting involves coverage or post-enumeration surveys. Here we focus on the measurement error.

Zhang and Fosen (2012) study the progressive measurement errors in the register-based small-area employment rate. Let $y_k(t; t+s)$ be the binary employment status of person k at statistical time point t, which is available at the measurement time point $t+s$, for $s \geq 0$. Take any two measurement time points $(t+r, t+s)$ where $0 \leq r < s$. Person k is said to have a *delayed entry* between $t+r$ and $t+s$ if $y_k(t; t+s) \neq y_k(t; t+r)$ due to updates between r and s. Person k is said to have a *recurred entry* between $t+r$ and $t+s$ if $y_k(t; t+r) = y_k(t; t+s)$ despite there being updates between $t+r$ and $t+s$ concerning person k. For the register-based production at time $t+s_0$, the recurred entries between t and $t+s_0$ are *ignorable* progressive data, but the delayed entries are *non-ignorable*.

Let $y_k(t)$ be the true register-status of person k for time t, based on ideal error-free input administrative data. To simplify the matter we shall assume that

$$y_k(t) = \lim_{s\to\infty} y_k(t; t+s)$$

Let $N_{ab}(t+r, t+s)$ be the number of persons with $y_k(t; t+r) = a$ and $y_k(t; t+s) = b$ for $a, b = 0, 1$. Let $t+s_0$ be the production time point. A simple selection model of the measurement mechanism (i.e. binary classification) at s_0 is given by

$$p_1 = P[y_k(t; t+s_0) = 1 | y_k(t) = 1] = \lim_{s\to\infty} \frac{N_{11}(t+s_0, t+s)}{N_{11}(t+s_0, t+s) + N_{01}(t+s_0, t+s)}$$

$$p_0 = P[y_k(t; t+s_0) = 1 | y_k(t) = 0] = \lim_{s\to\infty} \frac{N_{10}(t+s_0, t+s)}{N_{00}(t+s_0, t+s) + N_{10}(t+s_0, t+s)}$$

The classification is assumed to be independent and identically distributed across the units.

Zhang and Fosen (2012) explore the measurement errors using historic data from the Norwegian Employer Employee Register (NEER). Put

$$a_s = N_{01}(t, t+s)/N_{11}(t, t) \qquad \text{and} \qquad b_s = N_{10}(t, t+s)/N_{11}(t, t)$$

which gives, respectively, the relative increase and decrease of the register-based employment rate due to the delayed entries between measurement time points t and $t+s$.

Table 4.3 shows the historic values of a_s and b_s in the NEER for reference time points in years 2002, 2004 and 2006, respectively. The first measurement delay $s = 140$ corresponds roughly to the actual production time point $t+s_0$ of the register-based employment rate. It can be seen that delayed entries may keep arriving a long time after that. Only b_s seems to have converged after about 6 years (say, for $t > 2190$) for the reference year 2002. Convergence does not seem to be the case for the other series. Nevertheless, one can get a feeling of the level of the classification probabilities (p_1, p_0) on substituting $s = 2555$ for $s = \infty$ for reference year 2002, which gives $\hat{p}_1 = 1 - 0.053/1.052 = 0.950$ and $\hat{p}_0 = 0.576 \cdot 0.030/$

Table 4.3 Historic data in the NEER. Statistical reference time point (t) in week 45 of 2002, 2004 and 2006. Measurement delay (s) in days after the reference time point (t). Increase (a_s) and decrease (b_s) due to delayed entries

Measurement delay (s)	Reference time point (t)					
	Year 2002		Year 2004		Year 2006	
	a_s	b_s	a_s	b_s	a_s	b_s
140	.043	.014	.031	.025	.041	.027
365	.070	.036	.044	.036	.056	.037
548	.080	.040	.051	.041	.064	.041
730	.084	.041	.055	.043	.068	.042
1095	.089	.042	.060	.045	.070	.044
1460	.091	.043	.062	.046		
1825	.094	.043	.063	.047		
2190	.095	.044				
2555	.096	.044				

Source: Zhang and Fosen (2012, Table 2, p. 99). Reproduced with permission of Indian Agricultural Statistics Research Institute.

$(1 - 0.576 \cdot 1.052) = 0.044$, that is 5% misclassification for those with $y_k(t) = 1$ and 4.4% for those with $y_k(t) = 0$.

To see the relevance for SAE, consider the following simple model. Let θ_i be the theoretical area mean, and let y_{ij} be the error-free binary register variable for unit j in area i, where $i = 1, \ldots, m$ and $j = 1, \ldots, N_i$. Let x_{ij} be the observed binary register variable. The fixed reference and production time points $(t, t + s_0)$ are dropped to simplify the notation. Put

$$P(x_{ij} = 1|y_{ij}) = \begin{cases} p_{i1} & \text{if } y_{ij} = 1 \\ p_{i0} & \text{if } y_{ij} = 0 \end{cases}$$

$$y_{ij}|\theta_i \sim \text{ Bernoulli } (\theta_i)$$

$$\theta_i = \theta + u_i$$

where $E(u_i) = 0$ and $V(u_i) = \sigma_u^2$. Let $\bar{y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i$. We have

$$\bar{y}_i = \theta_i + e_i = \theta + u_i + e_i$$

where $E(e_i|u_i) = 0$, and $V(e_i|u_i) = \theta_i(1 - \theta_i)/N_i$, and $Cov(u_i, e_i) = 0$. Let $\bar{x}_i = \sum_{j=1}^{N_i} x_{ij}/N_i$. Let $\lambda_i = p_{i1} - p_{i0}$. We have

$$\bar{x}_i = p_{i0} + \lambda_i\theta + \lambda_i u_i + b_i \tag{4.2}$$

where $E(b_i|u_i) = 0$ and $V(b_i|u_i) = V(x_{ij}|\theta_i)/N_i = \theta_i(1 - \theta_i)\lambda_i^2 + \theta_i p_{i1}(1 - p_{i1}) + (1 - \theta_i)$ $p_{i0}(1 - p_{i0})$. The expected true area mean $\bar{y}_i$ conditional on the observed $\bar{x}_i$ is

$$E(\bar{y}_i|\bar{x}_i) = \frac{\bar{x}_i\theta_i p_{i1}}{\theta_i p_{i1} + (1 - \theta_i)p_{i0}} + \frac{(1 - \bar{x}_i)\theta_i(1 - p_{i1})}{\theta_i(1 - p_{i1}) + (1 - \theta_i)(1 - p_{i0})}$$

which is not equal to $\bar{x}_i$ unless $(p_{i1}, p_{i0}) = (1, 0)$, that is no measurement error.

The random effect u_i represents the heterogeneity across the areas. It is of the order $O_p(1)$. The random error b_i arises from the within-area individual variations and the measurement errors. It is of the order $O_p(1/\sqrt{N_i})$. The model (4.2) differs from the model of Fay and Herriot (1979) in that the sampling variance $V(b_i|\theta_i)$ depends on the mean parameter θ_i. It is nevertheless easier to handle compared with an alternative generalized linear mixed model. Because we are dealing with population registers, N_i is usually large enough to warrant a normal approximation to the distribution of b_i, as long as θ_i is not very close to either 0 or 1.

Zhang and Fosen (2012) explore in addition a model for the register-based change. They carry out sensitivity analysis of the level and change estimates of small-area employment rates, using alternative values of (p_{i1}, p_{i0}) under the simplification that $(p_{i1}, p_{i0}) = (p_1, p_0)$ does not vary over time. However, as the historic data in Table 4.3 suggest, the measurement error mechanism does vary over time, and may well be expected to vary across the areas. For future research it will be interesting to develop more elaborate models, which allow for differential error mechanisms both over time and across the population domains.

Finally, we notice that provided relevant sample survey data are available, such as the LFS in the context of Employment Statistics, alternative approaches are possible based on administrative and survey data integration, as we shall discuss later in Section 4.3.2.

4.3 Administrative and Survey Data Integration

4.3.1 Coverage Error and Finite-population Bias

A register has under-coverage (or missing enumeration) of the target population if there exist population units that are not enlisted in the register; it has over-coverage (or erroneous enumeration) if there are units in the register that do not belong to the target population.

An important SAE application that involves the coverage error is list enumeration adjustment using separate coverage surveys. The list may be the census or Population Register enumeration. See Hogan (1993) for an early account of the methodology in the US. See Nirel and Clickman (2009) for a recent and more comprehensive review of the uses of sample surveys in censuses.

A potential adaption to the production of poverty indicators can be as follows. On the one hand, one has from the tax authority some past, say previous-year, income, based on which it is possible to classify the 'poor' persons. Such a 'poor' person, however, may be erroneous in the current situation, if his/her income has changed to above the poverty threshold in the meantime; whereas, a 'not-poor' person may be a missing enumeration in the reverse case. On the other hand, one has a sample of the current population, where the poverty status can be correctly classified, but for the nonrespondents. In other words, one recognizes that the survey enumeration has under-coverage.

Disregarding several additional issues including sampling design, weighting and imputation for missing data, matching and processing of datasets, a stylized population size estimator can be given as

$$\hat{N} = n(1-\theta)x/m$$

where x is the list population count, and n an independent under-coverage survey (U-sample) count, and m the number of enumerations in both the list and the U-sample. The parameter

θ is the proportion (or probability) of erroneous list enumeration, and is traditionally estimated using a separate O-sample drawn from the list enumeration. Wolter (1986) details the underlying assumptions without explicit reference to $1-\theta$, but the account can be rephrased to acknowledge the additional O-sample. See Zhang (2015) for modelling approaches to the over- and under-coverage errors, where the adjustment requires only the U-sample.

A key difficulty from the SAE perspective arises when the coverage samples are not large enough to support the direct estimator $\hat{N}_i$ in all the local areas $i = 1, \ldots, m$. In particular, when a multi-stage sampling design is used, there may be many areas that are not represented in the sample. To focus on the key issue at hand, put, to start with,

$$\tilde{N}_i = x_i\xi \qquad \text{and} \qquad N_i = x_i\xi + v_i$$

where N_i is the local area population size, and x_i the known list enumeration, and ξ the known global adjustment factor $\xi = N/x$ for $N = \sum_{i=1}^{m} N_i$ and $x = \sum_{i=1}^{m} x_i$. Then, $\tilde{N}_i$ is the best estimator for an out-of-sample area, and $E(\tilde{N}_i - N_i|N_i) = -v_i$ is its finite-population (FP) bias for fixed $U_N = \{N_1, \ldots, N_m\}$, or $U_v = \{v_1, \ldots, v_m\}$.

Zhang (2007) uses an area-level mixed model to assess the squared FP-bias of a synthetic estimator. An application to the register-based census employment rate is given by Fosen and Zhang (2011). For population size estimation, consider the following adaption. Let $\hat{N}_i = N_i + e_i = x_i\xi + v_i + e_i$ be the direct estimator for an in-sample area, where e_i is its FP sampling error such that $E(\hat{N}_i|v_i) = N_i$ and $V(\hat{N}_i|v_i) = \psi_i$. Put

$$z_i = \hat{N}_i - \tilde{N}_i = v_i + e_i \tag{4.3}$$

and assume independence between v_i and e_i. Next, since $N = x\xi$, assume $E(v_i) = 0$. Finally, for the variance, assume $V(v_i) = x_i^{\alpha}\sigma_v^2$ for some fixed constant α. The mixed model (4.3) is then a special case of the area-level models (Rao 2003). Having fitted the model to the in-sample areas, we obtain $\hat{\sigma}_v^2$ as the estimate of σ_v^2 and, for an out-of-sample area,

$$\hat{E}(v_i^2) = x_i^{\alpha}\hat{\sigma}_v^2$$

One may interpret $E(v_i^2)$ as the anticipated squared FP-bias under model (4.3). Other measures are possible, for example $E(|v_i|)$ under an additional normality assumption of v_i. Or, it may be more suitable to model v_i on a different scale.

In practice, the global factor ξ needs to be estimated. Let $\hat{\xi} = \xi + \epsilon$ with sampling error ϵ. Then, instead of model (4.3), put

$$z_i = \hat{N}_i - x_i\hat{\xi} = v_i + u_i \qquad \text{and} \qquad u_i = e_i - x_i\epsilon$$

where e_i and ϵ are correlated. Re-sampling methods may be used to estimate the variance of u_i directly, instead of those of e_i and ϵ and their covariance separately.

More importantly, the FP-bias v_i is uncontrollable for an out-of-sample area. This suggests that, subjected to the total sample size affordable, one may consider adopting a sampling design which ensures that all the local areas are represented in the sample. A necessary consequence is the reduction of the sample size in certain (if not all) local areas, and the direct estimation may no longer have an acceptable accuracy in a number of areas.

Notice that indirect SAE via the random effects is also FP-biased. To fix the idea, assume known ξ, σ_v^2 and $\psi_i(>0)$ for $i = 1, \ldots, m$. The best predictor (BP) of v_i is then

$$\tilde{v}_i = \gamma_i(\hat{N}_i - x_i\xi) \qquad \text{and} \qquad \gamma_i = x_i^{\alpha}\sigma_v^2/(x_i^{\alpha}\sigma_v^2 + \psi_i)$$

For fixed U_N, or U_v, the FP-bias of the BP is then $E(\tilde{v}_i - v_i|v_i) = -(1-\gamma_i)v_i$. Thus, the unavailability of an acceptable direct estimator for the in-sample areas calls for a careful evaluation of the FP-bias under alternative sampling designs.

Future research in this area is important because a viable methodology for population size estimation that is able to address non-negligible over- and under-coverage errors in the input registers can potentially be useful in many situations beyond population census, including the production of poverty indicators as indicated above.

4.3.2 Relevance Error and Benchmarked Synthetic Small Area Estimation

Benchmarking the aggregates of mixed-effects model-based SAEs to accepted estimates or known totals can yield some protection against model misspecification and achieve output consistency that is important in Official Statistics. See Pfeffermann (2013, section 6.3) for a review. But there is a limit to which sample survey data can support the various mixed-effects models. As one descends the hierarchy of aggregation, sooner or later, a level will be reached where many (or most) areas are not represented in the sample, and many areas will have only very few sample observations. Synthetic methods are necessary from then on. However, the stringent assumptions required for the synthetic estimates to be unbiased are often plainly unattainable. So the issue becomes more than protection against the misspecification of certain aspects of the model. It is about actively reducing the model bias.

Proxy measures from census or administrative sources can provide substitutions to the 'random effects' whose estimation is poorly supported by the sample data available. A well known technique in the SAE literature is the Structure PREserving Estimation (SPREE) following Purcell and Kish (1980). See also Noble *et al*. (2002) for a generalized linear model (GLM) framework which includes the log-linear model underpinning SPREE. Identical or similar applications exist for example in demography and population geography, where the approach is known under different names such as iterative proportional fitting (IPF) or models with spatial-interaction offsets, and so on. See for example Simpson and Tranmer (2005) and Raymer *et al*. (2011). Benchmarking by IPF in all these applications of proxy data can as well be motivated as a means for reducing the definition bias (or relevance error) of the proxy measure, so as to achieve statistical relevance at which level the estimates are benchmarked.

The SPREE and related methods mentioned above can easily become relevant to the production of multidimensional poverty indicators, where the dimensions of poverty classification may involve economic, social and other factors. For instance, suppose that a complete cross-classification is feasible in the previous census. However, based on the current separate surveys and/or registers, only the updated marginal distributions are available, but not the joint distribution of the cross-classifications.

Below we outline some techniques for reducing the model bias, or the relevance error, of the synthetic estimation methods for SAE, from one-way benchmarking to multivariate generalized structure preserving estimation. A unifying formulation is optimal adjustment of the initial synthetic estimates subjected to the benchmarking constraints.

4.3.2.1 One-way Optimal Benchmarking

Take first the one-way case. Let μ_i be the best synthetic estimate of area (or domain) mean θ_i, which is given when all the involved parameters are known. Let μ_i^B be the *one-way* benchmarked best synthetic estimates that satisfy the constraint

$$\sum_{i=1}^{m} W_i \mu_i^B - \sum_{i=1}^{m} W_i \theta_i = 0 \tag{4.4}$$

where $\boldsymbol{\mu}^B = (\mu_1^B, \ldots, \mu_m^B)^T$ and $W_i = N_i/N$, and N_i is the area population size and $N = \sum_{i=1}^{m} N_i$. To obtain $\boldsymbol{\mu}^B$ as the solutions of optimal adjustment of $\boldsymbol{\mu}$ subjected to the benchmark constraint (4.4), one needs to specify: (i) the form of adjustment; and (ii) the loss function. Consider the following.

- Global additive adjustment δ and loss function Δ given by, respectively,

$$\begin{aligned} & \mu_i^B = \mu_i + \delta \\ & \Delta = \frac{1}{2} \sum_{i=1}^{m} W_i (\mu_i^B - \theta_i)^2 = \frac{1}{2} \sum_{i=1}^{m} W_i (\mu_i + \delta - \theta_i)^2 \\ \Rightarrow \quad & \mu_i^B = \mu_i + \delta \qquad \text{and} \qquad \delta = \bar{\theta}_w - \bar{\mu}_w \end{aligned}$$

  where $\bar{\theta}_w = \sum_{i=1}^{m} W_i \theta_i$ and $\bar{\mu}_w = \sum_{i=1}^{m} W_i \mu_i$.
- Global multiplicative adjustment δ and loss function Δ given by

$$\begin{aligned} & \mu_i^B = \delta \mu_i \\ & \Delta = \frac{1}{2} \sum_{i=1}^{m} W_i (\mu_i^B - \theta_i)^2 / \mu_i = \frac{1}{2} \sum_{i=1}^{m} W_i (\delta \mu_i - \theta_i)^2 / \mu_i \\ \Rightarrow \quad & \mu_i^B = \delta \mu_i \qquad \text{and} \qquad \delta = \bar{\theta}_w / \bar{\mu}_w \end{aligned}$$

- Area-specific additive adjustment δ and loss function Δ given by

$$\begin{aligned} & \mu_i^B = \mu_i + \delta_i \\ & \Delta = \frac{1}{2} \sum_{i=1}^{m} W_i (\mu_i^B - \theta_i)^2 = \frac{1}{2} \sum_{i=1}^{m} W_i (\mu_i + \delta_i - \theta_i)^2 \\ \Rightarrow \quad & \mu_i^B = \mu_i + \delta_i \qquad \text{and} \qquad \delta_i = \theta_i - \mu_i \end{aligned}$$

In practice, θ_i is replaced by an unbiased direct estimate $\hat{\theta}_i$, and μ_i by a synthetic estimate $\hat{\mu}_i$ depending on some global parameter estimates. The benchmarked synthetic estimate $\hat{\bar{\mu}}_w^B = \sum_{i=1}^{m} W_i \hat{\mu}_i^B$ is unbiased for $\bar{\theta}_w$, because $\hat{\bar{\mu}}_w^B = \hat{\bar{\theta}}_w$. Moreover, it is clear that area-specific optimal adjustment is difficult because it essentially recast the SAE problem one had from the beginning, that is to obtain a good estimate of θ_i.

Wang *et al*. (2008) derive benchmarked best linear unbiased predictor (BBLUP) under the Fay–Herriot model. Replacing the empirical best linear unbiased predictor (EBLUP) $\hat{\theta}_i^{EBLUP}$

by $\hat{\mu}_i$, one obtains a benchmarked synthetic estimate in analogy to the BBLUP, which is given by

$$\hat{\theta}_i^B = \hat{\mu}_i + (a_i/N_i)N(\hat{\bar{\theta}}_w - \hat{\bar{\mu}}_w)$$

where $\sum_{i=1}^m a_i = 1$. First, under the additive adjustment, apportioning the overall difference $N\hat{\bar{\theta}}_w - N\hat{\bar{\mu}}_w$ to all the areas according to the area population sizes leads to a difference estimator for θ_i, that is

$$a_i = N_i/N \qquad \Rightarrow \qquad \hat{\theta}_i^B = \hat{\mu}_i + (\hat{\bar{\theta}}_w - \hat{\bar{\mu}}_w) = \hat{\bar{\theta}}_w + (\hat{\mu}_i - \hat{\bar{\mu}}_w)$$

which is identical to the optimal additive adjustment above. In particular, $\approxeq_i = \hat{\mu}_i - \hat{\bar{\mu}}_w$ can be considered as a substitution estimate of the random effect $u_i = \theta_i - \bar{\theta}_w$. Secondly, the commonly used proportional or pro-rata adjustment can be given by

$$a_i = (N_i\hat{\mu}_i)/(N\hat{\bar{\mu}}_w) \qquad \Rightarrow \qquad \hat{\theta}_i^B = (\hat{\mu}_i/\hat{\bar{\mu}}_w)\hat{\bar{\theta}}_w$$

which is identical to the optimal multiplicative adjustment above. Moreover, $\hat{\mu}_i = \hat{\mu}_i/\hat{\bar{\mu}}_w$ can be regarded as a substitution estimate of the multiplicative random effect $u_i = \theta_i/\bar{\theta}_w$.

As a property of the additive adjustment, we notice that it leads to the familiar post-stratification estimator as follows. Let the categorical target (Y) and proxy (X) variables take value $j = 1, \ldots, J$. Let Y_{ij} and $Y_{.j}$ be the area and overall count of units with $y = j$, and θ_{ij} and $\theta_{.j}$ the corresponding area and overall proportions. Similarly for X_{ij} and $X_{.j}$. Let $Y_{ij;k} = X_{ik}p_{i,kj}$ be the area count of the units with $(x, y) = (k, j)$. Let the initial proxy substitution estimate be $\hat{Y}_{ij;k}^S = X_{ij} = X_{ik}p_{i,kj}^S$ where $p_{i,kj}^S = 1$ if $j = k$ and $p_{i,kj}^S = 0$ if $j \neq k$. Overall we have $\hat{Y}_{j;k}^S = X_k p_{kj}^S$, where $p_{kj}^S = 1$ if $j = k$, and $p_{kj}^S = 0$ if $j \neq k$. It is straightforward to verify that, whether $j = k$ or not, additive adjustment of $p_{i,kj}^S$ yields $\hat{p}_{i,kj} = \hat{p}_{kj} = \hat{Y}_{j;k}/X_{.k}$, where $\hat{Y}_{j;k}$ is an unbiased estimate of the population count of the units with $(x, y) = (k, j)$. Thus, the additive adjustment yields

$$\hat{Y}_{ij}^B = \sum_{k=1}^{J} X_{ik}(\hat{Y}_{j;k}/X_{.k})$$

which is the synthetic post-stratification estimate of Y_{ij} for $i = 1, \ldots, m$ and $j = 1, \ldots, J$.

4.3.2.2 Two- or Multi-way Optimal Benchmarking

Consider now multi-way benchmarking constraints. We explain the approaches for the two-way case; the multi-way cases follow similarly. Let Y_{ij} denote the target two-way totals of interest, where $i = 1, \ldots, m$ denotes the areas and $j = 1, \ldots, J$ denotes the categories. The underlying variable can either be categorical such as the number of people by activity status, or continuous such as the VAT turnover by product type. Let $Y_{i.} = \sum_{j=1}^J Y_{ij}$ and $Y_{.j} = \sum_{i=1}^m Y_{ij}$. Let $(\hat{Y}_{i.})_{i=1}^m$ and $(\hat{Y}_{.j})_{j=1}^J$ be the benchmark totals for the two-way table $\{Y_{ij}\}$. In particular, due to the constraints $(\hat{Y}_{i.})_{i=1}^m$, the data are referred to as compositions, and it is equivalent to estimate the totals or proportions of the categories.

Let X_{ij} be a substitution estimate of Y_{ij} based on a proxy measure. Consider the ANOVA decomposition

$$X_{ij} = \mu_0^X + \mu_i^X + \mu_j^X + \mu_{ij}^X$$

where $\mu_0^X = \bar{X} = \sum_{i,j} X_{ij}/(mJ)$, and $\mu_i^X = \bar{X}_i - \bar{X} = \sum_j X_{ij}/J - \bar{X}$, and $\mu_j^X = \bar{X}_j - \bar{X} = \sum_i X_{ij}/m - \bar{X}$, and $\mu_{ij}^X = X_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}$. Notice that we have $\sum_i \mu_i^X = \sum_j \mu_j^X = \sum_i \mu_{ij}^X = \sum_j \mu_{ij}^X = 0$ by construction. Put

$$\hat{Y}_{ij}^B = \hat{\mu}_0^Y + \hat{\mu}_i^Y + \hat{\mu}_j^Y + \mu_{ij}^X \tag{4.5}$$

where $\hat{\mu}_0^Y = \hat{\bar{Y}} = \sum_i \hat{Y}_{i.}/m = \sum_j \hat{Y}_{.j}/J$, and $\hat{\mu}_i^Y = \hat{\bar{Y}}_{i.} - \hat{\mu}_0^Y = \hat{Y}_{i.}/J - \hat{\mu}_0^Y$, and $\hat{\mu}_j^Y = \hat{\bar{Y}}_{.j} - \hat{\mu}_0^Y = \hat{Y}_{.j}/m - \hat{\mu}_0^Y$. Notice that $\sum_i \hat{\mu}_i^Y = \sum_j \hat{\mu}_j^Y = 0$. We observe the following.

1. Now that $\sum_j \mu_{ij}^X = \sum_i \mu_{ij}^X = 0$, $\sum_i \hat{Y}_{ij}^B = \hat{Y}_{.j}$ and $\sum_j \hat{Y}_{ij}^B = \hat{Y}_{i.}$ by construction.
2. The estimate (4.5) can be obtained from X_{ij} via the following additive adjustments

$$\hat{\mu}_0^Y - \mu_0^X \qquad \text{and} \qquad (\hat{\mu}_i^Y - \mu_i^X)_{i=1}^m \qquad \text{and} \qquad (\hat{\mu}_j^Y - \mu_j^X)_{j=1}^J$$

   The adjustments are uniquely determined by the benchmark constraints $(\hat{Y}_{i.})_{i=1}^m$ and $(\hat{Y}_{.j})_{j=1}^J$, together with the definitional constraints $\sum_i \hat{\mu}_i^Y = \sum_j \hat{\mu}_j^Y = 0$. Optimal adjustment is trivial in this case, irrespective of the choice of the loss function.
3. The interaction offsets μ_{ij}^X can be regarded as a substitution estimate for the ANOVA random effect μ_{ij}^Y. The approach is straightforward given a proxy measure. It is not so effective based only on auxiliaries. The key challenge then is how to construct a synthetic estimate of the interaction μ_{ij}^Y that is non-trivial.

When the underlying y-variable is non-negative, or when the Y_{ij}'s are the counts of a categorical variable, it is more common to apply the two-way proportional adjustment to the proxy table via IPF, or raking. At each iteration, the rows and the columns are adjusted successively by a multiplicative factor to satisfy the corresponding benchmark totals one at a time. One obtains the benchmarked estimates $\hat{Y}_{ij}^B$ on convergence of the IPF. In SAE this is known as the SPREE (Purcell and Kish 1980), where the log-linear interactions of the proxy table are preserved by the IPF.

The SPREE is an approach based on the log-linear ANOVA decomposition. Put

$$\log\,(X_{ij}) = \alpha_0^X + \alpha_i^X + \alpha_j^X + \alpha_{ij}^X$$

where

$$\alpha_0^X = \sum_{i,j} \log\,(X_{ij})/(mJ) \qquad \text{and} \qquad \alpha_i^X = \sum_j \log\,(X_{ij})/J - \alpha_0^X$$

$$\alpha_j^X = \sum_i \log\,(X_{ij})/m - \alpha_0^X \qquad \text{and} \qquad \alpha_{ij}^X = \log\,(X_{ij}) - \alpha_i^X - \alpha_j^X - \alpha_0^X$$

The IPF updates all the other α-terms except the α_{ij}^X's.

Let $\log\ (Y_{ij}) = \alpha_0^Y + \alpha_i^Y + \alpha_j^Y + \alpha_{ij}^Y$ be the log-linear ANOVA decomposition of Y_{ij}, defined in the same way as for X_{ij}. Under the SPREE, α_{ij}^X can be considered a proxy substitution estimate of the random effect α_{ij}^Y. Moreover, the IPF yields the maximum likelihood estimate of the log-linear model parameters, given the sufficient marginals $(\hat{Y}_{i.})_{i=1}^m$ and $(\hat{Y}_{.j})_{j=1}^J$ and with the interactions constrained at the α_{ij}^X's. One may therefore interpret the multiplicative adjustments from X_{ij} to Y_{ij} as the result of optimal adjustment subjected to these constraints, where the loss function is the Kullback–Leibler divergence between the constrained log-linear model and the saturated model for the two-way table $\{Y_{ij}\}$.

4.3.2.3 Modelling under Benchmark Constraints

The two-way benchmarking adjustments of the proxy table above amounts to substitution estimation of the random effects. By definition, however, the proxy measure suffers relevance error and entails definition bias. Flexible fixed-effects modelling of the relationship between the target interactions and the proxies can help to further reduce the bias. Conceptually, this points to a modelling approach that respects the benchmark constraints inherent to the problem at hand, rather than treating the necessary benchmark adjustments as an alien element to be administered ad hoc to the model-based estimates.

For instance, take the target two-way table $\{Y_{ij}\}$. Given the two sets of marginal benchmark totals $(\hat{Y}_{i.})_{i=1}^m$ and $(\hat{Y}_{.j})_{j=1}^J$, the only unknowns that remain are the linear or log-linear interactions. it is therefore natural and appealing that a modelling approach under the benchmark constraints should be developed in terms of these interactions. Zhang and Chambers (2004) propose a generalized SPREE (GPREE) model:

$$\alpha_{ij}^Y = \beta \alpha_{ij}^X \tag{4.6}$$

It is shown that the model (4.6) of the log-linear interactions can be fitted using the associated GLM of the within-area proportions $\theta_{ij}^Y = Y_{ij}/Y_{i.}$, that is

$$\eta_{ij}^Y = \lambda_j + \beta \eta_{ij}^X$$

where $\eta_{ij}^Y = \log\ \theta_{ij}^Y - \sum_{j=1}^J \log\ \theta_{ij}^Y$, and likewise for η_{ij}^X. The λ_j's are nuisance parameters subject to the constraint $\sum_{j=1}^J \lambda_j = 0$. On fitting the GLM, one obtains $\{\exp\ (\hat{\beta}\alpha_{ij}^X)\}$ as the starting table for the IPF, which then yields the final benchmarked synthetic estimates of $\hat{Y}_{ij}$. The SPREE is the special case of constraining β at 1.

More recently, Luna-Hernandez and Zhang (2013) have proposed an extension given by

$$\alpha_{ij}^Y = \sum_{k=1}^J \beta_{jk} \alpha_{ik}^X \tag{4.7}$$

where $\sum_{k=1}^J \beta_{jk} = \sum_{j=1}^J \beta_{jk} = 0$. This may be referred to as the multivariate GSPREE, in analogy to multivariate linear regression. Again, the model (4.7) can be fitted using the associated GLM

$$\eta_{ij}^Y = \lambda_j + \sum_{k=1}^J \beta_{jk} \eta_{ik}^X$$

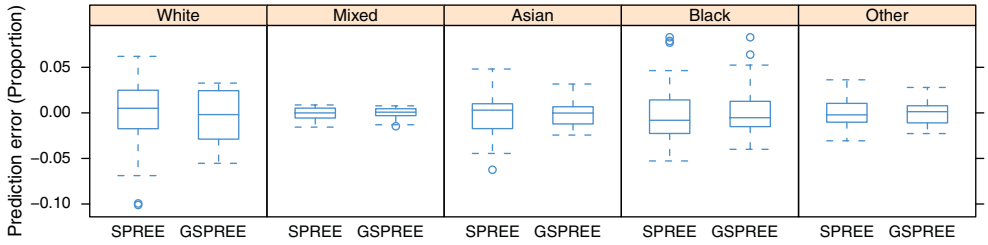

Figure 4.1 Boxplot of prediction error for ethnicity composition, Hackney Borough of London

with nuisance parameters λ_j's. On fitting the model, one obtains $\{\exp\ (\sum_{k=1}^{J} \hat{\beta}_{jk}\alpha_{ik}^{X})\}$ as the starting table of the IPF, and so on. Notice that the multivariate GSPREE model (4.7) includes as a special case the logistic multinomial model

$$\log\ (\theta_{ij}^{Y}/\theta_{iJ}^{Y}) = \gamma_j + \phi_j \log\ (\theta_{ij}^{X}/\theta_{iJ}^{X})$$

where category J is the arbitrarily chosen reference category, so that the model has $(J-1)$ free intercepts γ_j and slopes ϕ_j for $j \neq J$.

For an illustration, consider the proportions of ethnicity groups (White, Mixed, Asian, Black, Other) in the Hackney Borough of London. The UK Census 2001 and 2011 compositions are treated as X_{ij} and Y_{ij}, respectively. A boxplot of the prediction error of the SPREE and the multivariate GSPREE model (4.7) is given in Figure 4.1. It can be seen that the extra flexibility of the model (4.7) leads to further reduction of the bias of the SPREE. Notice that in practice the multivariate GSPREE modelling approach does not require any additional data compared with the SPREE. Notice also that the ability to produce less biased synthetic estimates is critical at a low aggregation level, where the random-effects modelling approach is not supported due to the sparseness of the sample survey data.

4.3.3 Probability Linkage Error

The EU-SILC survey is a national survey. It is a major source for poverty and living conditions estimates in Italy. Data on income and on working and living conditions are collected at both the household and individual levels. The target population of the Italian SILC are all the Italian households (and the individuals living in these households). However, the sample is not large enough to allow reliable estimates to be calculated at the local level, that is below the regional level (NUTS 2 level in the European Nomenclature for Territorial Units).

Probability linkage of EU-SILC data with several administrative data from the Province of Pisa (PI-SILC) was carried out under the FP7 project SAMPLE–Small Area Methods for Poverty and Living condition Estimates (www.sample-project.eu). The sample size for Pisa was boosted as part of the SAMPLE project. Indicators relevant to the measurement of poverty and living conditions were transferred from the administrative sources to the linked dataset. One of the primary aims was to explore the use of the EU-SILC sampling weights for the administrative indicators and to enhance the administrative data with the household characteristics collected in the survey. We explore here the linked data of the PI-SILC and the Job Centre

(JC) database. In Chapter 3 of this book a brief description is given of the linkage between the PI-SILC and Revenue Agency data under the SAMPLE project.

The Pisa JC register contains information on people looking for a job, being hired or being fired in the Province of Pisa, irrespective of their Province of residence. Importantly, the population accessible to the JC register consists only of those who have had dealings with the JC—people who do not contact the JC during their working life cannot be found in the JC database. In 2008 the total number of individuals in the Pisa JC register is 216 048. Meanwhile, the PI-SILC sample has 818 households, boosted from the original EU-SILC design of 162 households. The number of responding households and persons are 675 and 1685, respectively. There are 1476 persons with age above 15, which is the maximum number of persons that can be linked to the JC register.

Due to privacy restrictions, anonymized probability linkage between PI-SILC and JC data is based on the following key variables: municipality of residence, gender, birth month and birth year. A multiple pass procedure was applied. In the first pass, a pair of records is identified as a match provided exact match on all the key variables, referred to as level-1 linkage. In the next two passes, an increasing degree of mismatch between the key variables is allowed, resulting in level-2 and level-3 linkage.

The results, obtained using the package 'RecordLinkage' of the R software, are summarized in Table 4.4. It can be seen that multiple matches are found for the same PI-SILC unit at all three levels of linkage. The number of PI-SILC units with at least one matched JC record is 1113 out of the 1476 available units. The average number of matches is 7.9, irrespective of the level of linkage. The epiWeight shown in the table is a weight calculated for each matched pair of records based on the approach used by Contiero *et al*. (2005) in the EpiLink record linkage software: it is equal to 1 only in the case of exact match on all the key variables. Generally speaking, the weight reflects how likely a match identifies a true pair of records. For each of the 1113 PI-SILC units with at least one match, the match with the highest epiWeight is accepted as the linkage for that unit. In the case of ties, one of them is chosen at random. A linked dataset of 1113 units is obtained in this way.

Clearly, linkage errors exist in the linked dataset, which is the case if the accepted pair of records does not correspond to a true pair of records. From the SAE perspective, it is of interest to examine the heterogeneity of this linkage error. Of the 39 Municipalities in Pisa, 25 are represented in the linked dataset. The Municipalities are a partition of the provincial administrative area. These are treated as the small areas here.

Table 4.4 Results of the record linkage procedure between PI-SILC and JC data. Total number of matched pairs, number of unique PI-SILC units and average value of the epiWeights, by level of linkage

Level of linkage	Matched pairs	PI-SILC records	Average epiWeights
1	353	321	1
2	3097	715	0.76
3	5348	910	0.54
Total	8798	1946	0.64

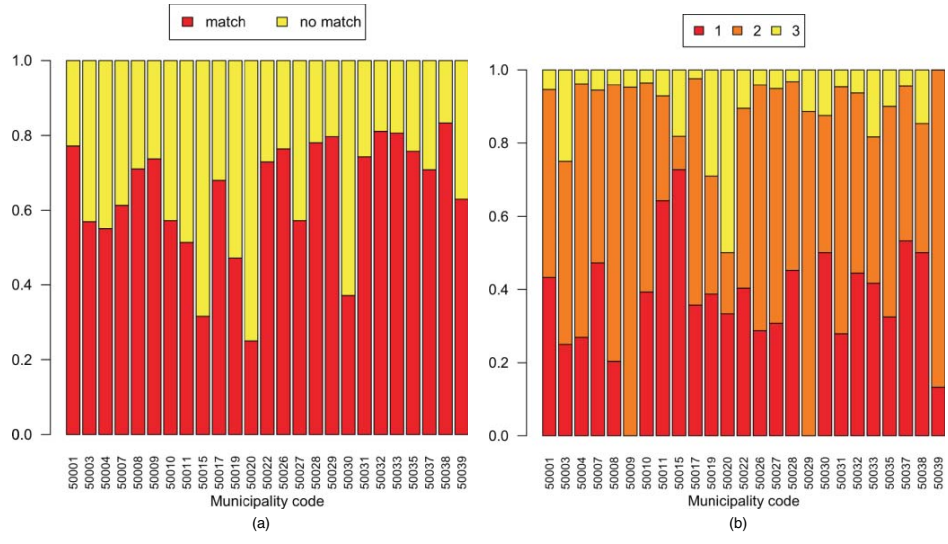

Figure 4.2 Percentage of linked units (a) and composition of linkage level (b) by Municipality

Simple exploratory analysis provides clear evidence for heterogeneous linkage error across the areas. For instance, it can be seen in Figure 4.2a that the overall percentage of linked PI-SILC units differs across the 25 Municipalities, with values between 25% and 83%. The Pearson χ^2-test rejects the null hypothesis of independence between Municipality and linkage percentage (p-value < 0.0001). Taking into consideration that the variation may have been caused, at least partly, by the varying sampling fraction across the Municipalities, we examine further the composition of the levels of linkage. Figure 4.2b shows that this also differs considerably across the Municipalities. For example, the percentage of near-deterministic linkages (level 1) varies greatly between 0% and 67%. Again, the Pearson χ^2-test rejects the null hypothesis of independence between Municipality and linkage level.

Heterogeneity is also found in the number of matches for each PI-SILC unit, regardless of whether these are accepted as linkage or not in the end. As shown in Figure 4.3, the average

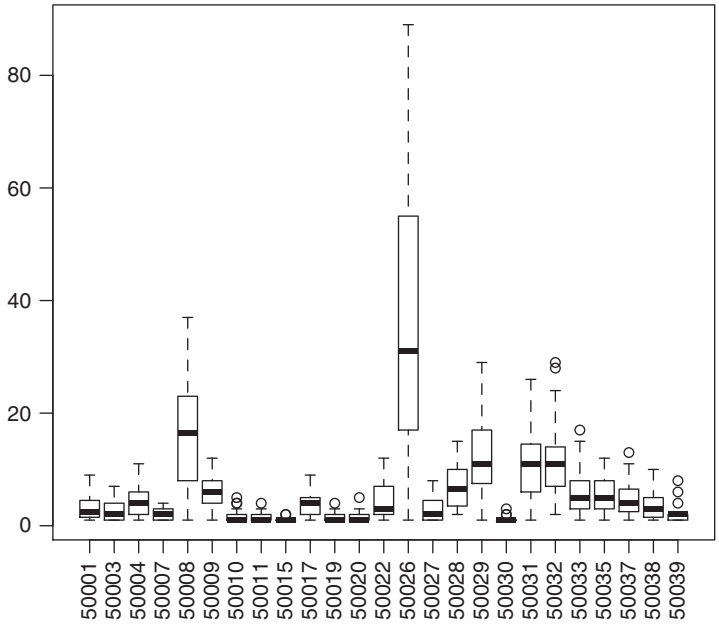

Figure 4.3 Boxplot: Number of matches for each PI-SILC unit with at least one match, by Municipality

Table 4.5 Cross-classification of JC job-seeker and PI-SILC labour status

Looking for job in JC	Labour status in PI-SILC			
	Employed	Looking for job	Not looking for job	Missing
Yes	104	26	55	5
No	541	28	336	16
Missing	1	0	1	0

Figure 4.4 Distribution of PI-SILC status by Municipality conditional on JC status: looking for a job (a) and not looking for a job (b)

number of matches varies between 1.15 and 35.43. Fitting an 'empty' mixed model with only random intercept yields an estimated intraclass correlation equal to 41.77%. Moreover, the percentage of household members with at least one match varies between 18.96% and 87.70%.

As mentioned above, one of the aims of the research within the SAMPLE project was to investigate the potential use of the JC register to produce small area statistics. A natural approach is to estimate the conditional distribution of the SILC labour-status given the JC status based on the linked dataset. Table 4.5 shows the cross-classified counts of JC and SILC status. Out of the 190 job-seekers in the JC, there are 104 (or 54.7%) who are employed according to the PI-SILC; the number is 541 (or 58.7%) out of the 921 non JC-job-seekers. This suggests that the binary JC job-seeker status will not be very effective for the estimation of SILC-employment total or rate, since the odds ratio 0.849 is quite close to 1. However, the association between the JC and PI-SILC is much higher among the persons who are not employed in the PI-SILC, that is columns 3 and 4 in Table 4.5, where the odds ratio is now 5.673. This suggests that a potential approach is to estimate, first, the SILC-employed total without necessarily using the JC register and, then, the SILC job-seekers total among the non-employed using the JC register. However, as Figure 4.4 shows, heterogeneity exists in the conditional distribution across the Municipalities, such that direct within-Municipality adjustment of the JC counts would have been too unstable, while synthetic estimation using some fixed effects model across the Municipalities may be too biased.

Adjustment of regression analysis for probability linkage errors arising between survey and register data has been discussed by Kim and Chambers (2012). Samart and Chambers (2010) consider fitting linear mixed models in the presence of linkage errors. In both cases, however, a homogeneous exchangeable linkage error (ELE) model is assumed. It will be interesting in future to develop methods that allow the application of SAE models, with or without random effects, to accommodate the heterogeneous linkage errors documented above.

4.4 Concluding Remarks

In this chapter we discuss SAE methods based on administrative data integration, where the settings may or may not involve survey data in addition. The most common sources of error and their implications for the SAE are described from a theoretical perspective and several real-life datasets are used for illustration. These data are directly relevant to the construction of multidimensional poverty and well-being indicators, including life expectancy, employment status, and so on. Potential applications of the methods for the production of local poverty and well-being statistics are indicated.

From the discussions it becomes clear that SAE based on administrative data integration can potentially involve many other types of errors than the sampling error alone. This presents some different and, often, rather difficult challenges, that have not received sufficient attention in the current literature of SAE. In particular, there arises generally the issue of how potentially to account for the cross-area heterogeneity in the predominant non-sampling error, be it progressive measurement error, coverage error or record-linkage error, in addition to the heterogeneity of the small area parameters themselves. One only needs to reflect on how scant the literature is, on SAE that involves heterogeneous nonresponse errors, in order to appreciate the challenges and open issues for future research.

References

Chiang CL 1984 *The Life Table and its Applications.* Malabar, FL: Robert E. Krieger Publishing Co.

Contiero P, Tittarelli A, Tagliabue G, Maghini A, Fabiano S, Crosignani P, Tessandori R 2005 The EpiLink Record Linkage Software Presentation and Results of Linkage Test on Cancer Registry Files. *Methods of Information in Medicine*, **4**, 66–71.

ESSnet SAE 2011 *WP5 Final Report on the Case Studies*. September 1, 2015. Available at http://www.cros-portal .eu/sites/default/files//ESSnet%20SAE%20WP5%20Report-final-rev2.pdf.

Fay R and Herriot R 1979 Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.

Fosen J and Zhang L-C 2011 Quality evaluation of employment status in register-based census. *Bulletin of the ISI 58th World Statistics Congress of the International Statistical Institute, Dublin*.

Hedlin D, Fenton T, McDonald JW, Pont M and Wang S 2006 Estimating the undercoverage of a sampling frame due to reporting delays. *Journal of Officia Statistics*, **22**, 53–70.

Hogan H 1993 The Post-Enumeration Survey: Operations and results. *Journal of the American Statistical Association*, **88**, 1047–1060.

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

Linkletter CD and Sitter RR 2007 Predicting natural gas production in Texas: An application of nonparametric reporting lad distribution estimation. *Journal of Officia Statistics*, **23**, 239–251.

Luna-Hernandez A and Zhang L-C 2013 On models for small area compositions. *Proceedings of Small Area Conference 2013,* Bangkok.

Nirel R and Clickman H 2009 Sample surveys and censuses. In *Sample Surveys: Design, Methods and Applications*, Vol 29A (eds. D. Pfeffermann and C.R. Rao), North-Holland, Elsevier, Amsterdam, The Netherland. Chapter 21, pp. 539–565.

Noble A, Haslett S and Arnold G 2002 Small area estimation via generalized linear models. *Journal of Officia Statistics*, **18**, 45–60.

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

Purcell NJ and Kish L 1980 Postcensal estimates for local areas (or domains). *International Statistical Review*, **48**, 3–18.

Rao JNK 2003 *Small Area Estimation*. Hoboken: John Wiley & Sons, Inc, Hoboken, NJ.

Raymer J, Smith PWF and Giuletti C 2011 Combining census and registration data to analyse ethnic migration patterns in England from 1991 to 2007. *Population, Space and Place*, **17**, 73–88.

Samart K and Chambers RL 2010. *Fitting linear mixed models using linked data*. Centre for Statistical and Survey Methodology, University of Wollongong, September 1, 2015. Working Paper 18-10, 25 pp. Available at http://ro.uow.edu.au/cssmwp/68.

Simpson L and Tranmer M 2005 Combining sample and census data in small area estimates: Iterative Proportional Fitting with standard software. *The Professional Geographer*, **57**, 222–234.

Wallgren A and Wallgren B 2006 *Register-based Statistics - Administrative Data for Statistical Purposes*. Chichester: John Wiley & Sons, Ltd.

Wang J, Fuller WA and Qu Y 2008 Small area estimation under a restriction. *Survey Methodology*, **34**, 29–36.

Wolter K 1986 Some coverage error models for census data. *Journal of the American Statistical Association*, **81**, 338–346.

Zhang L-C 2007 Finite population small area interval estimation. *Journal of Officia Statistics*, **23**, 223–237.

Zhang L-C 2011 A unit-error theory for register-based household statistics. *Journal of Officia Statistics*, **27**, 415–432.

Zhang L-C 2012 Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, **66**, 41–63.

Zhang L-C 2015 On modelling register coverage errors. *Journal of Officia Statistics*, **31**, 381–396.

Zhang L-C and Chambers RL 2004 Small area estimates for cross-classifications. *Journal of the Royal Statistical Society: Series B*, **66**, 479–496.

Zhang L-C and Fosen J 2012 A modelling approach for uncertainty assessment of register-based small area statistics. *Journal of the Indian Society of Agricultural Statistics*, **66**, 91–104.