Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 66(1) 2012 91-104

A Modeling Approach for Uncertainty Assessment of Register-based Small Area Statistics

L.-C. Zhang* **and J. Fosen**
Statistics Norway, Kongensgate 6, PB 8131 Dep, N-0033 Oslo, Norway
Received 30 April 2011; Revised 17 August 2011; Accepted 17 August 2011

SUMMARY

Statistical registers have great potentials when it comes to producing statistics at detailed spatial-demographic levels. However, population totals based on statistical registers are subjected to random variations that exist in the target population as well as errors that are associated with the registration (or measurement) process. While the former counts for heterogeneity across the areas (or domains), i.e. genuine 'signals' of interest, the latter ones are merely 'noises' in measurement. We propose a model-based sensitivity analysis approach, which allows us to distinguish between the different sources of randomness in the data, by which means the strength of the signals can be assessed against the noises. The data from the Norwegian Employer/Employee register are used to demonstrate the existence of measurement noises in administrative data sources, and to illustrate the proposed approach. We believe that both the conceptualization of the random nature of the register data and the sensitivity analysis approach can be useful for assessing detailed statistics produced from statistical registers on various subjects.

Keywords : Modeling, Register-based statistics, Measurement errors, Sensitivity analysis.

1. INTRODUCTION

Statistical registers have great potentials when it comes to producing statistics at detailed spatial-demographic levels. There are, however, several difficulties that need to be dealt with. As e.g. pointed out in Holt (2007), the two important issues to start with are the coverage of the target population and the relevance (or definition) with respect to the variables of interest. In this paper we assume that these matters have been successfully resolved. But we are concerned with the fact that the population totals based on any statistical register are subjected to different sources of randomness. While some of these represent genuine heterogeneity across the areas (or domains), others are merely undesirable measurement errors. Moreover, some of the random variations are subject to the law of Large Numbers and may be negligible at high aggregation levels, while others are not so and will always exist at the level of interest. It is thus important that one is able to conceptualize and assess the 'signals' of interest against the 'noises' that are ever present.

We propose a modeling approach by which one may distinguish between randomness of three different kinds in a statistical register. Firstly, the target small area population parameters are regarded as being stochastic both over time and space. Take for instance the employment rate at the Municipality level. This varies from one Municipality to another at any given time point, as well as over time for any given Municipality, in a non-deterministic fashion. Secondly, given the target population parameter, there is individual variation across the units. Thus, given the target employment rate for a certain Municipality, the employment status is still a random variable from one person to another within that Municipality. Thirdly, the registration, indeed the whole production, process of

* *Corresponding author* : L.-C. Zhang
E-mail address : lcz@ssb.no

any statistical register is associated with random variations. For instance, there may be delays or mistakes in reports to the administrative authority, but there may also be processing errors committed at the administration or the statistical agency. All such 'process errors' may affect the statistical register and the statistics produced from it.

The data we shall examine are taken from the Norwegian Employer/Employee register (NEER). The NEER consists initially of reports on various job events, such as hiring and dismissing. Retrospective inspection of historic data from the NEER reveals different types of process errors. Depending on the statistical use of the NEER, however, some of these errors may be 'ignored' because they do not affect the intended statistics, while others may have a large impact as we shall demonstrate. For this study we use the job-event reports in the NEER to derive a register-based employment status where, roughly speaking, someone is employed who holds at least one 'active' job according to the NEER, and the population (*i.e*., of persons with age between 16 and 74) is obtained from the Central Population Register. Misclassification of this register employment status may then be caused by certain process errors such as a delay in reporting.

Notice that for the purpose of employment statistics the information in the original NEER is reorganized using persons as the units. For convenience, we still refer to the latter register as the NEER, only bearing in mind the transformation that has taken place from the administrative register of job events to the statistical register of employment status. Such a transition is in fact characteristic of the statistical uses of many administrative registers, which initially consist of objects other than the statistical unit of interest. The conceptual distinction is important when setting up a framework for statistical analysis. For instance, a delay in the reporting of a job event causes initially under-coverage of the event-population by the administrative register, which constitutes an error in terms of representation of the initial NEER. For the statistical register, however, the same delay affects the classification of the employment status and, thus, causes a measurement error. We refer to Chapter 2 of Groves, Fowler Jr., Couper, Lepkowski, Singer, and Tourrangeau (2004) for the distinction between errors of measurement and representation. Zhang (2011) outlined an extension of the error framework to statistical micro data from multiple sources, including sample survey and register, either on their own or in combination.

It is also worth noting that, in reality, the production of a register-based employment status is likely to involve many additional relevant data sources (e.g. Aukrust *et al.* 2010), such as self-employment register, wage sum register, military services, Labor Force Survey, etc. We have deliberately simplified the data integration process here in order to focus on the central statistical methodological issues involved.

Having thus formulated the effects of the process errors in the NEER as a possible misclassification error of a binary variable, we find from the historic data that the error mechanism varies both over time and across the population domains. Direct estimation of the error mechanism at the production time point must therefore rely on model assumptions that are likely to result in bias for any particular domain at any particular reference time point. Moreover, whether or not to adjust the directly tabulated register-based statistics by means of such estimated error mechanisms is a broader issue in practice, which is beyond the scope of this paper. Here we propose a *sensitivity analysis* approach, where the results derived under alternative scenarios of the error mechanisms may lead to conclusions that are likely to withstand deviations from the default mechanism that corresponds to the unadjusted register-based statistics.

The rest of the paper is organized as follows. In Section 2 we outline a modeling approach for small area means of a binary variable, which accounts for random variations on three different levels as explained above. Both level and change statistics will be considered, and the associated model-fitting procedures described. In Section 3 we develop a conceptual framework for the categorization of the process errors that may affect the classification of a binary variable. We examine the historic data from the NEER, which provide empirical evidences for the nature of the error mechanism. We then outline and apply a sensitivity analysis approach to Municipality employment rates based on the NEER in 2005 and 2006. A summary is given in Section 4.

Again, we notice that, while the categorization of the process errors and the models that we use in this paper necessarily depend on the actual NEER data, it

is not our intention to adjust for these specific errors in order to arrive at better estimates. The NEER data serve primarily to illustrate both the conceptualization of the random nature of register-based statistics and the proposed sensitivity analysis approach, both of which we believe are applicable to detailed statistics produced from many statistical registers.

2. A MODELING APPROACH FOR BINARY REGISTER DATA

2.1 Level Estimation : Without Process Errors

Let y_{ij} be a binary register variable for unit j in area i, where $i = 1, ..., m$ and $j = 1, ..., N_i$. Assume that y_{ij} is free of measurement errors. Let θ_i be the theoretical small area mean. Put

$$\theta_i = \theta + u_i \tag{1}$$

$$y_{ij} \mid \theta_i \sim \text{Binomial}(1, \theta_i) \tag{2}$$

i.e. Bernoulli(θ_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

$$\bar{y}_i = \theta_i + e_i = \theta + u_i + e_i \tag{3}$$

where $E(e_i \mid u_i) = 0$, and $\psi_i = V(e_i \mid u_i) = \theta_i(1 - \theta_i)/N_i$, and $\text{Cov}(u_i, e_i) = 0$.

We notice that a generalized linear mixed model, such as a logistic regression model with a normally distributed random effect on the linear predictor scale, is a default model choice for binary data. However, the computation would have been more complicated. A linear mixed predictor for $\bar{y}_i$ has two main advantages. Firstly, it is easier to compute. Because we are dealing with population registers, the underlying denominator N_i is usually large enough to warrant a normal approximation to the distribution of e_i and, thereby, the computational simplicity of a linear model, as long as the θ_i's are not very close to either 0 or 1. Secondly, the two random components u_i and e_i are on the same scale, allowing for straightforward interpretation. The random effect u_i represents the heterogeneity across the areas. It is of the order $O_p(1)$ at whichever aggregation level that is being modeled. The random error e_i is due to within-area individual variations. It is of the order $O_p(1/\sqrt{N_i})$ and is subjected to the Law of Large Numbers. It may obscure the underlying signal of interest, i.e. u_i, for sufficiently small areas.

Now, model (3) looks like a so-called basic area-level model (Section 5.2, Rao 2003), except that the variance $V(e_i)$ actually depends on the mean parameter $\theta_i = \theta + u_i$. Still, a method of moment estimator of σ_u^2 given $\hat{\theta}$ can be obtained by straightforward algebra. To estimate all the parameters and the random effects, we iterate until convergence between the method of moment estimator of σ_u^2 given by (4), and a weighted least-square estimator of θ given by (5), and an empirical best linear predictor of u_i given by (6), i.e.

$$\hat{\sigma}_u^2 = \max\left\{0, \sum_{i=1}^{m}((\bar{y}_i - \hat{\theta})^2 - \hat{\theta}(1-\hat{\theta})/N_i) / \sum_{i=1}^{m}(1 - 1/N_i)\right\} \tag{4}$$

$$\hat{\theta} = \left(\sum_{i=1}^{m} 1/\hat{v}_i\right)^{-1}\left(\sum_{i=1}^{m} \bar{y}_i/\hat{v}_i\right) \tag{5}$$

$$\hat{u}_i = \hat{\gamma}_i(\bar{y}_i - \hat{\theta}) \tag{6}$$

where $\hat{v}_i = \hat{\sigma}_u^2 + \hat{\psi}_i$ and $\hat{\gamma}_i = \hat{\sigma}_u^2 / \hat{v}_i = \hat{\sigma}_u^2 / (\hat{\sigma}_u^2 + \hat{\psi}_i)$.

To evaluate the mean squared error (MSE) of $\hat{\theta}_1 = \hat{\theta} + \hat{u}_i$ we use a jackknife MSE estimator proposed by Lohr and Rao (2009). We assume that the MSE of the best predictor (BP) when all the parameters are known is given by

$$g_i(\theta_i; \xi) = \psi_i\gamma_i = \sigma_u^2\psi_i / (\sigma_u^2 + \psi_i)$$

where ξ denotes the model parameters. Denote by $\hat{\xi}$ the full-data parameter estimator and by $\hat{\xi}_{(-j)}$ the delete-j estimator at the jth jackknife iteration, i.e. after the jth area is removed from the data. Similarly, denote by $\hat{\theta}_i$ the full-data estimator of θ_i and by $\hat{\theta}_{i(-j)}$ the corresponding delete-j estimator. The MSE estimator is then given by

$$\text{mse}(\hat{\theta}_i) = \hat{M}_{1i} + \hat{M}_{2i} \tag{7}$$

where $\hat{M}_{1i} = g_i(\hat{\xi}) - \sum_{j \neq i}\left\{g_i\left(\hat{\xi}_{(-j)}\right) - g_i\left(\hat{\xi}\right)\right\}$ and $\hat{M}_{2i} = (1 - 1/m)\sum_{j=1}^{m}(\hat{\theta}_{i(-j)} - \hat{\theta}_i)^2$. We refer to Lohr and Rao (2009) for more detailed elaboration on the jackknife MSE estimation.

2.2 Estimation of Change: Without Process Errors

Consider now the change in small area means from period $t = 1$ to period $t = 2$. Obviously, one may apply level estimation to each time point, and use the difference between the two level estimates as a change estimate. But a more direct approach is possible. Denote by $y_{ij,1}$ and $y_{ij,2}$ the binary register variables in each period, respectively. Let $z_{ij} = y_{ij,\,2} - y_{ij,\,1}$. Then

$$\overline{z}_i = \overline{y}_{i,\,t=2} - \overline{y}_{i,\,t=1} = \delta + d_i + \varepsilon_i \qquad (8)$$

where $\delta = \theta_{i,\,t=2} - \theta_{i,\,t=1}$, and $d_i = u_{i,\,t=2} - u_{i,\,t=1}$ with $V(d_i) = \sigma_d^2$, and $\varepsilon_i = e_{i,\,t=2} - e_{i,\,t=1}$ with variance $\tau_i = V(\varepsilon_i \mid u_{i,\,t=1},\, u_{i,\,t=2}) = V(z_{ij} \mid \theta_{i,\,t=1},\, \theta_{i,\,t=2})/N_i$.

Notice that model (8) for the mean of changes has the same structure as model (3) for the mean at a given time point, except that the variance of the random error, i.e. ε_i, is not directly a function of the mean change parameter δ_i. Indeed, we have

$$\begin{aligned} &V(z_{ij} \mid \theta_{i,\,t=1},\, \theta_{i,\,t=2}) \\ &\quad = V(y_{ij,1} \mid \theta_{i,\,t=1}) + V(y_{ij,2} \mid \theta_{i,\,t=2}) \\ &\qquad - 2\mathrm{Cov}(y_{ij,1}, y_{ij,2} \mid \theta_{i,\,t=1}, \theta_{i,\,t=2}) \\ &\quad = \theta_{i,\,t=1}(1 - \theta_{i,\,t=1}) + \theta_{i,\,t=2}(1 - \theta_{i,\,t=2}) \\ &\qquad - 2\theta_{i,\,t=1}(\alpha_i - \theta_{i,\,t=2}) \end{aligned}$$

where $\theta_{i,t=1}$ and $\theta_{i,t=2}$ are the mean parameters in each period, and $\alpha_i = P(y_{ij,2} = 1 \mid y_{ij,1} = 1)$ is conditional mean of $y_{ij,2}$ given $y_{ij,1} = 1$. The α_i can be estimated using the same approach as for θ_i in (3), if we introduce a random effect model

$$\alpha_i = \alpha + a_i \quad \text{where} \quad E(a_i) = 0 \quad \text{and} \quad V(a_i) = \sigma_a^2$$

Estimates of the parameters δ and σ_d^2 as well as the d_i's can be obtain in two stages. At the first stage, we estimate the parameters ($\theta_{i,t=1}$, $\theta_{i,t=2}$, α_i) that are needed to compute the variance component τ_i. At the second stage, we estimate the model (8) as if it were a standard area-level linear mixed model with known variances of the random error ε_i's, i.e. treating $\hat{\tau}_i$ from the first stage as fixed. The algorithm is given as follows.

- Stage One:
  - Apply the estimation approach for (3), separately for each period, to obtain $\hat{\theta}_{t=1}$ and $\hat{\theta}_{t=2}$, as well as $\hat{\theta}_{i,\,t=1}$ and $\hat{\theta}_{i,\,t=2}$ for $i = 1, \ldots, m$.
  - Apply the estimation approach for (3) to the subpopulation of $y_{ij,1} = 1$ to obtain $\hat{\alpha}_i$.
- Stage Two, i.e. for fixed $\hat{\tau}_i$ from Stage One:
  - Obtain $\hat{\delta}$ and $\hat{\sigma}_d^2$ by iteration till convergence, where

$$\hat{\delta} = \left(\sum_{i=1}^{m} 1/(\hat{\sigma}_d^2 + \hat{\tau}_i) \right)^{-1} \left(\sum_{i=1}^{m} \overline{z}_i /(\hat{\sigma}_d^2 + \hat{\tau}_i) \right)$$

$$\hat{\sigma}_d^2 = \max\left\{ 0, \sum_{i=1}^{m} \left((\overline{z}_i - \hat{\delta})^2 - \hat{\tau}_i \right) \Big/ m \right\}$$

  - Obtain $\hat{d}_i = \hat{\gamma}_i (\overline{z}_i - \hat{\delta})$ where

$$\hat{\gamma}_i = \hat{\sigma}_d^2 \Big/ (\hat{\sigma}_d^2 + \hat{\tau}_i), \text{ for } i = 1, \ldots, m.$$

We notice that alternative method-of-moment estimators of σ_d^2 can be found in Section 7.1.2 of Rao (2003), all of which are consistent as $m \to \infty$ without requiring normality. As before we use the jackknife to estimate the MSE where g_i is now given as $\hat{\sigma}_d^2 \tau_i \big/ (\sigma_d^2 + \tau_i)$. All the estimation steps are replicated at each jackknife iteration, including the estimation of τ_i.

2.3 Level Estimation: With Process Errors

Let us now consider the situation with possible misclassification of the register binary variable due to the underlying process errors. Let x_{ij} be a register binary variable that may be subject to process-generated measurement errors. In addition to (1) and (2), we assume the following *misclassification mechanism* where, given the underlying correct variable y_{ij},

$$P(x_{ij} = 1 \mid y_{ij}) = \begin{cases} p_{i1} & \text{if } y_{ij} = 1 \\ p_{i0} & \text{if } y_{ij} = 0 \end{cases} \qquad (9)$$

Let $\lambda_i = p_{i1} - p_{i0}$. Let $v_{i1} = p_{i1}(1 - p_{i1})$ and $v_{i0} = p_{i0}(1 - p_{i0})$. We have

$$\begin{aligned} E(x_{ij} \mid \theta_i) &= E(E(x_{ij} \mid y_{ij}) \mid \theta_i) = \theta_i p_{i1} + (1 - \theta_i)\, p_{i0} \\ &= p_{i0} + \lambda_i \theta_i \end{aligned}$$

$$V(x_{ij} \mid \theta_i) = \theta_i(1-\theta_i)\lambda_i^2 + \theta_i v_{i1} + (1-\theta_i)v_{i0} \tag{10}$$

since $E(V(x_{ij} \mid y_{ij})|\theta_i) = \theta_i v_{i1} + (1-\theta_i)v_{i0}$ and $V(E(x_{ij} \mid y_{ij})|\theta_i) = \theta_i(1-\theta_i)\lambda_i^2$. Let $\bar{x}_i = \sum_{j=1}^{N_i} x_{ij}/N_i$. A model derived from assumptions (1), (2) and (9) is then given by

$$\bar{x}_i = p_{i0} + \lambda_i\theta + \lambda_i u_i + b_i \tag{11}$$

where $E(b_i \mid u_i) = 0$ and $\phi_i = V(b_i|u_i) = V(x_{ij})/N_i$ for $V(x_{ij})$ given by (10), and $Cov(u_i, b_i) = 0$. Notice that u_i is given in (1) and is the same as in (3).

Model (11) can be fitted in two stages. At the first stage, we estimate the error mechanism (9) in one way or another. More discussions of the error mechanism will be given in Section 3. In fact, for the proposed sensitivity analysis approach, the misclassification probabilities will be set at values according to the alternative scenarios. In any case, at the second stage, we fix the parameters (p_{i1}, p_{i0}) at these given values, by which the model attains a form similar to model (3), i.e. an area-level linear mixed model but with the variance of the error component b_i depending on the mean parameter θ_i. By a similar method-of-moment derivation for $\hat{\sigma}_u^2$, then, we obtain $\hat{\theta}$, $\hat{\sigma}_u^2$ and $\hat{u}_i$, for $i = 1, ..., m$, by iterations till convergence, where

$$\hat{\theta} = \left(\sum_{i=1}^{m} \frac{\lambda_i^2}{(\lambda_i^2\hat{\sigma}_u^2 + \hat{\phi}_i)}\right)^{-1} \left(\sum_{i=1}^{m} \lambda_i \frac{(\bar{x}_i - p_{i0})}{(\lambda_i^2\hat{\sigma}_u^2 + \hat{\phi}_i)}\right)$$

$$\hat{\sigma}_u^2 = \max\left\{0, \frac{\sum_{i=1}^{m}((\bar{x}_i - p_{i0} - \lambda_i\hat{\theta})^2 - \hat{s}_i^2)}{\sum_{i=1}^{m}((1-1/N_i)\lambda_i^2)}\right\}$$

$$\hat{u}_i = \hat{\gamma}_i\,(\bar{x}_i - p_{i0} - \hat{\theta}\lambda_i)/\lambda_i$$

and $\hat{\theta}_i = \hat{\theta} + \hat{u}_i$, and $\hat{s}_i^2 = (\hat{\theta}(1-\hat{\theta})\lambda_i^2 + \hat{\theta}v_{i1} + (1+\hat{\theta})v_{i0})/N_i$, and $\hat{\gamma}_i = \lambda_i^2\hat{\sigma}_u^2/(\lambda_i^2\hat{\sigma}_u^2 + \hat{\phi}_i)$. Finally, for MSE estimation we use the jackknife estimator with $g_i = \sigma_u^2\phi_i/(\lambda_i^2\sigma_u^2 + \phi_i)$.

Now, having fitted the model and obtained $\hat{\theta}_i$, we may derive the expected correct register status of y_{ij} conditional on the observed x_{ij}. We have

$$P(y_{ij} = 1 \mid x_{ij}, \theta_i, p_{i1}, p_{i0}) = \begin{cases} \dfrac{\theta_i p_{i1}}{(\theta_i p_{i1} + (1-\theta_i)p_{i0})} & \text{if } x_{ij} = 1 \\ \dfrac{\theta_i(1-p_{i1})}{(\theta_i(1-p_{i1}) + (1-\theta_i)(1-p_{i0}))} & \text{if } x_{ij} = 0 \end{cases}$$

It follows that an estimate of the expected corrected $\bar{y}_i$ given the observed $\bar{x}_i$ is given by

$$\hat{\bar{y}}_i = \frac{\bar{x}_i\hat{\theta}_i p_{i1}}{\hat{\theta}_i p_{i1} + (1-\hat{\theta}_i)p_{i0}} + \frac{(1-\bar{x}_i)\hat{\theta}_i(1-p_{i1})}{\hat{\theta}_i(1-p_{i1}) + (1-\hat{\theta}_i)(1-p_{i0})} \tag{12}$$

The MSE of $\hat{\bar{y}}_i$ can be derived from that of $\hat{\theta}_i$ using the linearization technique, where

$$\frac{\partial\hat{\bar{y}}_i}{\partial\theta_i} = \frac{\bar{x}_i p_{i1}p_{i0}}{(\hat{\theta}_i p_{i1} + (1-\hat{\theta}_i)p_{i0})^2} + \frac{(1-\bar{x}_i)(1-p_{i1})(1-p_{i0})}{(\hat{\theta}_i(1-p_{i1}) + (1-\hat{\theta}_i)(1-p_{i0}))^2}$$

2.4 Change Estimation: With Process Errors

When it comes the estimation of change, again, one may apply level estimation to each time point in the presence of process errors, and use the difference between the two level estimates as a change estimate. A more direct approach is possible, but extra assumptions about the error mechanism are necessary. First, the misclassification probabilities (p_{i1}, p_{i0}) must remain the same for both time period. Otherwise, the observed change in means can not be expected to depend *only* on the underlying change parameter and, conditionally, the difference of the random effects over time. Next, the joint misclassification probabilities are needed in order to calculate the covariance between the pair of observed register variables. Independent classification conditional on the underlying correct register variable, i.e. the so-called conditional independence assumption (CIA), is perhaps the most common assumption in the literature concerning measurement errors. However, the situation here is somewhat different. As explained earlier, the register-based employment status is derived from the reported job events, not repeated measurements (or observations) of the statistical units (i.e. persons in this case). For instance, the chance that a delay should cause a

misclassification at both $t = 1$ and $t = 2$ probably depends on how long it has been since the event occurred, in addition to the true $y_{ij,t}$ at the two time points. Yet, despite the restrictions, the direct approach described below can still be a useful tool for sensitivity analysis.

Let $\tilde{z}_{ij} = x_{ij,2} - x_{ij,1}$, and $\bar{\tilde{z}}_i = \sum_{j=1}^{N_i} \tilde{z}_{ij} / N_i$

Provided constant (p_{i1}, p_{i0}), we have

$$\bar{\tilde{z}}_i = \lambda_i \delta + \lambda_i d_i + \tilde{b}_i \tag{13}$$

where $\delta = \theta_{t=2} - \theta_{t=1}$, and $d_i = u_{i,\,t=2} - u_{i,\,t=1}$ with $V(d_i) = \sigma_d^2$ as in (8), and $E(\tilde{b}_i \mid \theta_{i,\,1},\ \theta_{i,\,2}) = 0$ and $\tilde{\tau}_i = V(\tilde{b}_i \mid \theta_{i,1}, \theta_{i,2}) = V(\tilde{z}_{ij} \mid \theta_{i,1}, \theta_{i,2}) / N_i$ for shorthanded $\theta_{i,\,1} = \theta_{i,\,t=1}$ and $\theta_{i,\,2} = \theta_{i,\,t=2}$.

The marginal variance of $x_{ij,t}$ for given t is given by (10). To calculate the variance component $\tilde{\tau}_i$, we need to find the covariance between $x_{ij,1}$ and $x_{ij,2}$. Under the CIA, we have

$$E(x_{ij,\,1}x_{ij,\,2} \mid \theta_{i,\,1},\ \theta_{i,\,2}) = \theta_{i,1}\,\alpha_i\,p_{i1}^2 + \theta_{i,\,1}(1-\alpha_i)p_{i1}\,p_{i0} + (1-\theta_{i,\,1})\beta_i p_{i0}p_{i1} + (1-\theta_{i,\,1})(1-\beta_i)\,p_{i0}^2$$

where $\alpha_i = P(y_{ij,\,2} = 1 \mid y_{ij,\,1} = 1, \theta_{i,\,1},\ \theta_{i,\,2})$ and $\beta_i = P(y_{ij,\,2} = 1 \mid y_{ij,\,1} = 0, \theta_{i,\,1},\ \theta_{i,\,2})$. Moreover, $E(x_{ij,\,1} \mid \theta_{i,\,1}) = \theta_{i,1}p_{i1} + (1-\theta_{i,\,1})p_{i0}$ and $E(x_{ij,\,2} \mid \theta_{i,\,2}) = \theta_{i,\,2}p_{i1} + (1-\theta_{i,\,2})p_{i0}$, such that

$$\begin{aligned}
&\mathrm{Cov}(x_{ij,\,1},\ x_{ij,\,2} \mid \theta_{i,\,1},\ \theta_{i,\,2}) \\
&= \theta_{i,\,1}(\alpha_i - \theta_{i,\,2})\,p_{i1}^2 - \theta_{i,\,1}(\alpha_i - \theta_{i,\,2})p_{i1}p_{i0} \\
&\quad + (1-\theta_{i,\,1})(\beta_i - \theta_{i,\,2})p_{i0}p_{i1} - (1-\theta_{i,\,1})(\beta_i - \theta_{i,\,2})\,p_{i0}^2 \\
&= \theta_{i,\,1}(\alpha_i - \theta_{i,\,2})p_{i1}(p_{i1} - p_{i0}) \\
&\quad + (1-\theta_{i,\,1})(\beta_i - \theta_{i,\,2})p_{i0}(p_{i1} - p_{i0})
\end{aligned}$$

Notice that $\theta_{i,\,1}$ and $\theta_{i,\,2}$ can be estimated separately for each period. For estimation of α_i and β_i, we introduce the corresponding conditional means of the observed values, i.e.

$\tilde{\alpha}_i = P(x_{ij,\,2} = 1 \mid x_{ij,\,1} = 1, \theta_{i,\,1},\ \theta_{i,\,2})$ and $\tilde{\beta}_i = P(x_{ij,2} = 1 \mid x_{ij,\,1} = 0, \theta_{i,\,1},\ \theta_{i,\,2})$. We now notice that

$$\begin{cases}
\tilde{\alpha}_i = P\{(x_{ij,1}, x_{ij,2}) = (1,1) \mid \theta_{i,\,1},\ \alpha_i, \beta_i\} / P(x_{ij,1} = 1 \mid \theta_{i,1}) \\
\tilde{\beta}_i = P\{(x_{ij,1}, x_{ij,2}) = (0,1) \mid \theta_{i,1}, \alpha_i, \beta_i\} / P(x_{ij,1} = 0 \mid \theta_{i,1})
\end{cases}$$

Writing $q_i = P(x_{ij,\,1} = 1 \mid \theta_{i,\,1})$ and re-arranging the terms, we obtain the following linear equation system for $(\alpha_i,\ \beta_i)$, at each $i = 1, ..., m$,

$$\begin{cases}
\theta_{i,1}p_{i1}\lambda_i\alpha_i + (1-\theta_{i,1})p_{i0}\lambda_i\beta_i = \tilde{\alpha}_i q_i - \kappa_{i;\,\alpha} \\
\theta_{i,1}(1-p_{i1})\lambda_i\alpha_i + (1-\theta_{i,1})(1-p_{i0})\lambda_i\beta_i \\
\qquad = \tilde{\beta}_i(1-q_i) - \kappa_{i;\,\beta}
\end{cases} \tag{14}$$

where $\kappa_{i,\,\alpha} = \theta_{i,\,1}p_{i1}p_{i0} + (1-\theta_{i,\,1})\,p_{i0}^2$ and $\kappa_{i,\,\beta} = \theta_{i,\,1}(1-p_{i1})p_{i0} + (1-\theta_{i,\,1})(1-p_{i0})p_{i0}$.

We can now obtain $\hat{\delta}, \hat{\sigma}_d^2$ and $\hat{d}_i$ by the following two-stage algorithm:

- Stage One :
  - Apply model (11) separately to each period, and obtain $\hat{\theta}_{i,t=1}$ and $\hat{\theta}_{i,t=2}$ for $i = 1, ..., m$.
  - Apply the estimation approach for (3) separately to the subpopulation of $x_{ij,\,1} = 1$ and $x_{ij,\,1} = 0$ to obtain $\hat{\tilde{\alpha}}_i$ and $\hat{\tilde{\beta}}_i$, respectively. Obtain $\hat{\alpha}_i$ and $\hat{\beta}_i$ by (14), for $i = 1, ..., m$.
- Stage Two, i.e. for fixed $\hat{\tilde{\tau}}_i$ from Stage One :
  - Obtain $\hat{\delta}$ and $\hat{\sigma}_d^2$ by iteration till convergence, where

$$\hat{\delta} = \left(\sum_{i=1}^{m} \lambda_i^2 \big/ (\lambda_i^2\hat{\sigma}_d^2 + \hat{\tilde{\tau}}_i)\right)^{-1} \left(\sum_{i=1}^{m} \lambda_i \bar{\tilde{z}}_i \big/ (\lambda_i^2\hat{\sigma}_d^2 + \hat{\tilde{\tau}}_i)\right)$$

$$\hat{\sigma}_d^2 = \max\left\{0,\ \sum_{i=1}^{m}((\bar{\tilde{z}}_i/\lambda_i - \hat{\delta})^2 - \hat{\tilde{\tau}}_i/\lambda_i^2)/m\right\}$$

  - Obtain $\hat{d}_i = \hat{\gamma}_i(\bar{\tilde{z}}_i - \lambda_i\hat{\delta})/\lambda_i$ where $\hat{\gamma}_i = \lambda_i^2\hat{\sigma}_d^2/(\lambda_i^2\hat{\sigma}_d^2 + \hat{\tilde{\tau}}_i)$, for $i = 1, ..., m$.

For MSE estimation we use the jackknife estimator with $g_i = \sigma_d^2\tilde{\tau}_i/(\lambda_i^2\sigma_d^2 + \tilde{\tau}_i)$.

3. A SENSITIVITY ANALYSIS APPROACH TO REGISTER-BASED STATISTICS

3.1 A Reference Framework for Process Errors in NEER

The Norwegian Employer/Employee register (NEER) is maintained by the National Insurance Administration (NAV) for administrative purposes. A job that is eligible to the NEER is referred to as a *work*

relation. The beginning of a work relation is recorded through a *message* from the employer to the NAV. Both the *starting* date of a work relation and its *registration* date are recorded in the NEER. Likewise the *finishing* date of a work relation is recorded through another message, with its own registration date, if the event occurs. A work relation is said to be *active* at any time point between the starting and finishing dates.

A particular time point at which statistics are of interest is called the *reference* time point, denoted by $t = 0$. The true status (or value) at the reference time point will be referred to as the *reference status (or value)*. An assessment of the reference status can be made at any *measurement* time point, at or after the reference time point, denoted by t for $t \geq 0$. The assessed reference status obtained on that occasion will be referred to as the *measurement status (or value)* at t. As long as the data are not ideal, the measurement status at any $t \geq 0$ can differ from the reference status. Finally, the time point at which the (register-based) statistics are produced will be referred to as the *production* time, denoted by $t = t'$ for a particular $t' > 0$. The measurement status at $t = t'$ may be referred to as the *production status (or value)*.

In this study of the NEER, we shall consider a person to be *employed* at $t = 0$, i.e. with reference value 1, if she or he has at least one active work relation at the corresponding reference time point. The reference status is 0 otherwise. The measurement status, on the other hand, is classified according to the reports of job events that have actually been recorded in the NEER up to the corresponding measurement time point. Explicitly, we fix the reference time point to be the first day of week 45 in a calendar year, say, 2002. The production time is then set to be 140 days after that, i.e. $t' = 140$, which is about the same as the actual time lag in annual register-based employment statistics. A measurement time can be any time point at or after the reference time point. For instance, it can be some time in week 50 in 2002, or week 5 in 2003, or week 13 in 2003 (i.e. around the production time), or week 13 in 2004, and so on.

Given the reference framework outlined above, one may distinguish between two types of *register process errors*. Fix the reference time point $t = 0$. Take any two measurement time points (t_1, t_2) where $0 \leq t_1 < t_2$. A person is said to have a *delayed entry* between t_1 and t_2 if there are messages, arriving between t_1 and t_2 and causing changes, such that the measurement status at $t = t_2$ is different than that at $t = t_1$. Whereas a person is said to have a *recurred entry* between t_1 and t_2 if there are messages, arriving between t_1 and t_2 and causing changes, yet the measurement status at $t = t_2$ remains the same as that at $t = t_1$.

For instance, take as t_1 and t_2 week 13 of 2003 and 2004, respectively. Suppose a person had measurement status "employed" at t_1. If the measurement status turned to "not employed" at t_2 due to, say, a message arriving in week 25 of 2003, then we say that this person had a delayed entry between week 13 of 2003 and of 2004. However, if there was a second message arriving, say, in week 40 of 2003 which changed the measurement status back to "employed" by week 13 of 2004, then we say that this person had a recurred entry between t_1 and t_2.

Table 1. Measured employment status in the NEER, including delayed and recurred entries. Reference time point: Week 45 of 2008. First measurement time point: Week 47 of 2008. Second measurement time point: Week 13 of 2009. E: "employed". N: "not employed".

	Measurement Status (First, Second)			
Entry	(E, E)	(E, N)	(N, E)	(N, N)
Delayed	–	70775	103211	–
Recurred	5259	–	–	2896
No Changes	2062976	–	–	1288140
Total	2068235	70775	103211	1291036

Evidences for both types of process errors in the NEER are given in Table 1. The reference time point is week 45 of 2008. The first measurement time is week 47 of 2008, i.e. two weeks after the reference time point; and the second measurement time is week 13 of 2009, i.e. around the production time point. Between these two time points, there were 70775 delayed entries which changed the measurement status from "employed" to "not employed", and 103211 delayed entries which changed the measurement status in the opposite direction from "not employed" to "employed". In addition, there were 5259 recurred entries of "employed" and 2896 recurred entries of "not employed". It appears that there are considerable amount of delayed entries in the NEER, whereas the amount of recurred entries is much smaller in comparison.

For the production of register-based employment statistics, the recurred entries are *ignorable* process errors because they do not affect the statistics. In contrast, the delayed entries are *non-ignorable* because they do affect the statistics, and the magnitude of this effect depends on the time of measurement in general, and the choice of the production time in particular.

The issue of reporting delay has been studied in the past, notably for epidemiological, insurance, and product warranty applications. More recently, Hedlin *et al.* (2006) applied a log-linear type of models to estimate the reporting delays for the introduction of birth units to the business register. Linkletter and Sitter (2007) applied a non-parametric method to estimate and adjust for delays in Natural Gas Production reports in Texas. In both cases, the reporting delays are postulated to cause under-coverage, and dealt with as an error of representation. There is a difference to the present framework, where delayed entries lead to misclassification errors in terms of measurement. The conceptual difference could easily have been resolved if misclassification had occurred only in one direction. For instance, had the NEER received only messages of hiring, under-coverage of the hiring messages would have been equivalent to misclassification of "employed" as "not employed". In general, however, these methods for dealing with under-coverage due to reporting delays are not adequate for handling delayed entries that may cause changes of the measurement status (or value) in multiple directions.

Perhaps even more important is the concern that the underlying error mechanism for delayed entries is hardly constant over time or across the population domains. The matter is certainly relevant to the production of detailed statistics, because it considerably raises the stake if the goal is to adjust the register-based statistics to obtain better estimates. We do not attempt at such adjustments for the NEER. Instead we outline below a sensitivity analysis approach as a means for assessing the uncertainty in the unadjusted register-based statistics.

3.2 Assessing Misclassification Probabilities Due to Delayed Entries

Formally, let x_t be a timely indexed binary measurement status. Let $N_{11}(t_1, t_2)$ be the number of units with measurement values $x_{t_1} = 1$ and $x_{t_2} = 1$ for $t_1 < t_2$. Similarly for $N_{10}(t_1, t_2)$, $N_{01}(t_1, t_2)$ and $N_{00}(t_1, t_2)$. For the NEER, let $x_t = 1$ denote "employed" and $x_t = 0$ denote "not employed". We then have $N_{11}(t_1, t_2) = 2068235$, $N_{10}(t_1, t_2) = 70775$, $N_{01}(t_1, t_2) = 103211$ and $N_{00}(t_1, t_2) = 1291036$, for t_1 = "week 47 of 2008" and t_2 = "week 13 of 2009" in Table 1.

Let $t = t'$ be the chosen production point time. Following the deffnition given by (9), the misclassification mechanism at $t = t'$, i.e. due to delayed entries, is given by

$$p_1 = \lim_{t\to\infty} \frac{N_{11}(t', t)}{N_{11}(t', t) + N_{01}(t', t)} \text{ and}$$

$$p_0 = \lim_{t\to\infty} \frac{N_{10}(t', t)}{N_{00}(t', t) + N_{10}(t', t)} \qquad (15)$$

Notice that for simplicity we have omitted the domain index i in the above. Domain-specific $N_{11}(t_1, t_2)$ to $N_{00}(t_1, t_2)$ and (p_1, p_0) can easily be specified in a similar fashion. Notice also that we assume in (15) that all the delays will eventually be updated. In reality, this may not be the case in a particular register e.g. due to coverage problems of certain sub-populations, or simply the practicality of the maintenance routines. We shall not pursue such eventualities here.

It is possible to explore the misclassification mechanism using historic data from the NEER. Let $N_1(t)$ be the measurement total of "employed" at t for $t \geq 0$, and let $N_0(t)$ be the total of "not employed". The sum of the two is the population total of persons with age between 16 and 74 at the reference time point $t = 0$, denoted by $N \equiv N_1(t) + N_0(t)$. Thus, for any $0 \leq t_1 < t_2$, we have the following identities

$$N_1(t_2) = N_{11}(t_1, t_2) + N_{01}(t_1, t_2) \text{ and}$$
$$N_0(t_2) = N_{00}(t_1, t_2) + N_{10}(t_1, t_2)$$

In particular, at the reference time point $t = 0$, we have

$$N_{10}(0, 0) = N_{01}(0, 0) = 0 \text{ and}$$
$$N_1(0) = N_{11}(0, 0) \text{ and } N_0(0) = N_{00}(0)$$

Whereas, at any measurement time point t for $t \geq 0$, we have

$$\begin{aligned}\bar{x}_t &= N_1(t)/N = (N_{11}(0, t) + N_{01}(0, t))/N \\ &= (N_1(0) - N_{10}(0, t) + N_{01}(0, t))/N \\ &= (N_1(0) - N_1(0)b_t + N_1(0)a_t)/N = \bar{x}_0(1 + a_t - b_t)\end{aligned}$$

where $\bar{x}_0 = N_1(0)/N$, and

$a_t = N_{01}(0,t)/N_1(0)$ and

$b_t = N_{10}(0,t)/N_1(0)$

Provided there exist only delayed entries, or disregarding recurred entries, we have, for any $t_1 < t_2$,

$$N_{01}(t_1, t_2) = N_{01}(0, t_2) - N_{01}(0, t_1) = N_1(0)(a_{t_2} - a_{t_1})$$

and

$$N_{10}(t_1, t_2) = N_1(0)(b_{t_2} - b_{t_1})$$

It follows that the classification probabilities of (15) at the production time $t = t'$ are given by

$$p_1 = (N_1(\infty) - N_{01}(t', \infty))/N_1(\infty) = 1 - (a_\infty - a_{t'})/(1 + a_\infty - b_\infty) \quad (16)$$

$$p_0 = N_{10}(t', \infty)/N_0(\infty) = \bar{x}_0 (b_\infty - b_{t'})/(1 - \bar{x}_0 (1 + a_\infty - b_\infty)) \quad (17)$$

Table 2 contains the historic values of a_t and b_t in the NEER for reference year 2002, 2004 and 2006, respectively. The first measurement time $t = 140$ is the production time of this study. It can be seen that delayed entries may keep arriving a long time after that. Only b_t seems to have converged after about 6 years (say, for $t > 2190$) for the reference year 2002, i.e. the delayed entries that change employment status from employed to not employed in 2002. Convergence does not seem to be the case for the other series. Nevertheless, it is possible to assess the level of the classification probabilities given by (16) and (17) on substituting sufficiently late measurement time point for $t = \infty$. Thus, for the reference year 2002, we have $\bar{x}_0 = 0.576$. On replacing $t = 2555$ for $t = \infty$, we obtain

$\hat{p}_1 = 1 - 0.053/1.052 = 0.950$ and

$\hat{p}_0 = 0.576 \times 0.030/(1 - 0.576 \times 1.052) = 0.044$

Similarly, on substituting the latest measurement time point available in Table 2 for $t = \infty$, we obtain $(\hat{p}_1, \hat{p}_0) = (0.969, 0.032)$ for the reference year 2004 and $(\hat{p}_1, \hat{p}_0) = (0.972, 0.027)$ for year 2006. Since neither a_t nor b_t has converged for the reference years 2004 and 2006, we expect under-estimation of the misclassification probabilities $1 - p_1$ and p_0 here.

The examination above of the historic delayed entries in the NEER seems to suggest the following.

(I) When sufficiently late measurement point time is available, plausible estimates of the misclassification mechanism can be derived retrospectively. Such estimates provide valuable information on the magnitude of the classification probabilities (15).

(II) Clearly, there exists yearly variation in the process of delayed entries, such that it will be necessary to introduce model assumptions if 'real-time' estimation of the classification probabilities is to be carried out at the production time point, which are likely to be biased when compared to the true probabilities that are available retrospectively.

(III) For simplicity, we have disregarded the domain variation in the classification mechanism. One may

Table 2. Historic data in the NEER. Reference time point in week 45 of 2002, 2004 and 2006. Measurement time point (t) in days after the reference time point.

	Reference Time Point								
	Year 2002			Year 2004			Year 2006		
t	a_t	b_t	$a_t - b_t$	a_t	b_t	$a_t - b_t$	a_t	b_t	$a_t - b_t$
140	0.043	0.014	0.026	0.031	0.025	0.006	0.041	0.027	0.013
365	0.070	0.036	0.035	0.044	0.036	0.008	0.056	0.037	0.019
548	0.080	0.040	0.040	0.051	0.041	0.010	0.064	0.041	0.024
730	0.084	0.041	0.043	0.055	0.043	0.012	0.068	0.042	0.025
1095	0.089	0.042	0.047	0.060	0.045	0.014	0.070	0.044	0.026
1460	0.091	0.043	0.049	0.062	0.046	0.016			
1825	0.094	0.043	0.050	0.063	0.047	0.016			
2190	0.095	0.044	0.051						
2555	0.096	0.044	0.052						

apply a chosen estimation procedure independently within each domain. This will increase the estimation uncertainty. Additional model assumptions regarding the domain variation may be introduced to improve the efficiency of estimation. For any given domain of interest, however, this is likely to bring in an extra bias term to the estimates of the classification probabilities. In light of these considerations, we shall propose in the followings a sensitivity analysis approach to the assessment of register-based small area estimates.

3.3 Sensitivity Analysis of Register-based Municipality Employment Rate

Instead of directly estimating the classification probabilities due to the delayed entries at the production time point, we conduct a sensitivity analysis based on alternative scenarios of these probabilities. At least four scenarios are of interest.

- In the *baseline* scenario the classification probabilities are fixed at some plausible values based on historic data. These can be a set of convergent (or almost convergent) estimates available, such as the estimates $(\hat{p}_1, \hat{p}_0) = (0.950, 0.044)$ for the reference year 2002 in Table 2, or an average of several sets of estimates of this kind.
- Two additional scenarios may be evaluated, one being likely an *over-statement* of the baseline values and the other being a likely *under-statement*. For the data in Table 2, a possible choice of the under-stated estimates are $(\hat{p}_1, \hat{p}_0) = (0.972, 0.027)$ for the reference year 2006, because these clearly have not converged and may be expected to bound the corresponding estimates towards the baseline values. A possible choice of over-stated estimates can be the equal-distance values on the opposite side of the baseline scenario, i.e. $\hat{p}_1 = 0.950 - (0.972 - 0.950) = 0.928$ and $\hat{p}_0 = 0.044 + (0.044 - 0.027) = 0.061$. The idea is that these over- and under-stated values may provide a range for the classification

Table 3. Municipality employment rates in 2005 (t = 1) and 2006 (t = 2) in Østfold, without process errors. Register rates ($\bar{y}_{i,t}$), Estimated theoretical rates ($\hat{\theta}_{i,t}$), Observed changes in the register ($\bar{z}_i$), Directly estimated theoretical changes ($\hat{\delta}_i$), Estimated root mean squared errors (rmse).

No.	Ni	Level 2005			Level 2006			Change 2005 - 2006			
		$\bar{y}_{i,1}$	$\hat{\theta}_{i,1}$	rmse	$\bar{y}_{i,2}$	$\hat{\theta}_{i,2}$	rmse	$\bar{z}_i$	$\hat{\theta}_{i,1} - \hat{\theta}_{i,2}$	$\hat{\delta}_i$	rmse
1	426	0.648	0.632	0.014	0.646	0.639	0.014	–0.001	0.007	0.012	0.005
2	961	0.602	0.611	0.012	0.637	0.636	0.011	0.034	0.025	0.018	0.006
3	2320	0.634	0.632	0.009	0.639	0.638	0.009	0.005	0.007	0.010	0.005
4	2427	0.587	0.595	0.009	0.596	0.604	0.009	0.009	0.010	0.012	0.004
5	2850	0.598	0.603	0.008	0.594	0.601	0.009	–0.004	–0.001	0.005	0.007
6	2899	0.664	0.656	0.009	0.679	0.670	0.009	0.015	0.015	0.014	0.004
7	3244	0.645	0.641	0.008	0.658	0.654	0.008	0.013	0.013	0.013	0.004
8	3453	0.631	0.629	0.007	0.648	0.646	0.007	0.017	0.016	0.015	0.003
9	3514	0.607	0.609	0.007	0.616	0.619	0.008	0.009	0.009	0.011	0.004
10	4664	0.645	0.642	0.007	0.654	0.652	0.007	0.010	0.010	0.011	0.004
11	5134	0.595	0.598	0.006	0.609	0.612	0.007	0.014	0.013	0.014	0.003
12	7184	0.614	0.615	0.005	0.629	0.629	0.005	0.015	0.015	0.015	0.003
13	9571	0.644	0.642	0.005	0.658	0.656	0.005	0.014	0.014	0.014	0.003
14	10049	0.620	0.620	0.005	0.635	0.635	0.005	0.016	0.015	0.015	0.003
15	19316	0.610	0.611	0.003	0.630	0.630	0.003	0.020	0.019	0.019	0.002
16	20315	0.623	0.623	0.003	0.636	0.636	0.003	0.013	0.013	0.014	0.002
17	35477	0.615	0.615	0.003	0.634	0.634	0.003	0.019	0.019	0.019	0.002
18	49975	0.617	0.617	0.002	0.634	0.634	0.002	0.017	0.017	0.017	0.001

probabilities that is likely to withstand both yearly and domain-wise deviations from the baseline values in most cases.

- Finally, there is the unrealistic but important *reference* scenario without misclassification, i.e. $(p_1, p_0) = (1, 0)$, because these correspond to the unadjusted register-based statistics.

The modeling approach described in Section 2 can now be applied under each alternative scenario, i.e. on plugging in the corresponding fixed and postulated values of the classification probabilities $(p_{i1}, p_{i0}) = (p_1, p_0)$. For an illustration we show the results for County Østfold in 2005 and 2006. Let the Municipality employment rate be the statistics of interest. There are 18 Municipalities in the population, denoted by $i = 1,..., 18$. The Municipality population size ranges from 426 to 49975, denoted by N_i. Given i, the theoretical rates are denoted by $\theta_{i,t}$ where $t = 1$ for year 2005 and $t = 2$ for year 2006. The corresponding observed register employment rates are denoted by $\bar{x}_{i,t}$. The underlying correct register employment rates are $\bar{y}_{i,t}$. Provided the process errors are disregarded, or assuming $(p_1, p_0) = (1, 0)$, we shall have $\bar{y}_{i,t} = \bar{x}_{i,t}$.

First of all we examine the reference scenario where we disregard the process errors in the register. In Table 3 we provide the register-based rates $\bar{y}_{i,t} = \bar{x}_{i,t}$, the estimated theoretical rates $\hat{\theta}_{i,t}$ and the associated estimated root mean squared errors (rmse), respectively, for levels in 2005 and 2006. Similarly for changes from 2005 to 2006. The theoretical rates are seen to agree with the register rates for the large Municipalities, both for level and change. However, the two rates differ clearly for the small Municipalities, indicating that the individual random variations may not be negligible for domain populations of such sizes. On average the rmse increases as the domain population size decreases. The effect is most evident for the change estimates. According to the register rates directly, there is a negative development in Municipality No. 1 and 5. However, for Municipality No. 1, the estimated change in the theoretical rates is 0.07 indirectly given by $\hat{\theta}_{i,t=2} - \hat{\theta}_{i,t=1}$ and 0.12 directly given by $\hat{\delta}_i$. For Municipality No. 5, the corresponding estimates are –0.01 indirectly and 0.05 directly. A possible interpretation is that the negative changes in register employment rates may be attributed to random individual variations, rather than reflections of a genuine negative trend in the labor marked of these two Municipalities, which would have been contrary to the general economic situation in the country at the time.

Next, we explore the three alternative scenarios of misclassification mechanism, using the baseline values of the classification probabilities and the corresponding over- and under-stated values. The alternative estimated underlying register employment rates $\hat{\bar{y}}_i$ given by (12), as well as the corresponding estimated theoretical rates $\hat{\theta}_i$, are given in Table 4 for the year 2005 and 2006, together with the associated rmse. The estimated underlying register rate $\hat{\bar{y}}_i$ has a much smaller rmse compared to the corresponding estimated theoretical rate $\hat{\theta}_{i,t}$: even for the smallest Municipalities, one can be quite certain that delayed entries may significantly alter the underlying correct rates. It is important to notice that this bias exists for the larger Municipalities as well as the smaller ones, and is not subjected to the law of Large Numbers. In contrast, the difference between $\hat{\bar{y}}_{i,t}$ and $\hat{\theta}_{i,t}$ due to individual variations in $y_{ij,t}$ clearly becomes negligible once the population is large enough. Compared to the reference scenario (Table 3) there is an increase in the rmse of $\hat{\theta}_{i,t}$. The incorporation of process errors in the model thus increases the estimation uncertainty despite the classification probabilities are introduced as fixed constants.

The observed change of register employment rate (denoted by $\tilde{\bar{z}}_i$), the change estimate derived from estimated underlying levels (denoted by $\hat{\bar{y}}^{(k)}_{i,t=2} - \hat{\bar{y}}^{(k)}_{i,t=1}$) under alternative scenarios, as well as the corresponding directly estimated theoretical change (denoted by $\hat{\delta}^{(k)}_i$) using model (13) are given in Table 5, together with the associated rmse. Apparently, in absolute value, the bias varies much less across the alternative scenarios of misclassification for the change estimates than in the case of level estimation. This is seen in two respects. Firstly, the change estimates $\hat{\bar{y}}^{(k)}_{i,t=2} - \hat{\bar{y}}^{(k)}_{i,t=1}$ do not vary much from one k to another, and similarly for $\hat{\delta}^{(k)}_i$. Secondly, the estimated underlying change $\hat{\bar{y}}^{(k)}_{i,t=2} - \hat{\bar{y}}^{(k)}_{i,t=1}$ does not differ much from the observed change $\tilde{\bar{z}}_i$ even for the small

Table 4. Observed register Municipality employment rate $\bar{x}_i$, estimated underlying register rate $\hat{\bar{y}}_i^{(k)}$ and estimated theoretical rate $\hat{\theta}_i^{(k)}$ in year 2005 and 2006 in Østfold, given alternative register process errors: $(p_{i1}, p_{i0}) = (0.972, 0.027)$ if $k = 1$, $(p_{i1}, p_{i0}) = (0.950, 0.044)$ if $k = 2$, $(p_{i1}, p_{i0}) = (0.928, 0.061)$ if $k = 3$. Estimated root mean squared error (rmse) in parentheses.

Year 2005								
No.	N_i	$\bar{x}_i$	$\hat{\bar{y}}_i^{(1)}$	$\hat{\theta}_i^{(1)}$	$\hat{\bar{y}}_i^{(2)}$	$\hat{\theta}_i^{(2)}$	$\hat{\bar{y}}_i^{(3)}$	$\hat{\theta}_i^{(3)}$
1	426	0.648	0.655 (.002)	0.639 (.017)	0.663 (.003)	0.648 (.017)	0.672 (.005)	0.657 (.018)
2	961	0.602	0.610 (.002)	0.618 (.014)	0.618 (.003)	0.626 (.014)	0.627 (.004)	0.635 (.015)
3	2320	0.634	0.642 (.001)	0.640 (.010)	0.651 (.002)	0.648 (.010)	0.660 (.003)	0.658 (.011)
4	2427	0.587	0.593 (.001)	0.601 (.010)	0.601 (.002)	0.609 (.010)	0.609 (.003)	0.617 (.011)
5	2850	0.598	0.605 (.001)	0.610 (.009)	0.613 (.002)	0.617 (.010)	0.621 (.003)	0.626 (.010)
6	2899	0.664	0.673 (.001)	0.665 (.009)	0.682 (.002)	0.674 (.010)	0.692 (.003)	0.685 (.010)
7	3244	0.645	0.654 (.001)	0.650 (.009)	0.663 (.002)	0.659 (.009)	0.673 (.003)	0.669 (.010)
8	3453	0.631	0.639 (.001)	0.637 (.008)	0.647 (.002)	0.646 (.009)	0.657 (.003)	0.655 (.009)
9	3514	0.607	0.614 (.001)	0.617 (.008)	0.622 (.002)	0.624 (.009)	0.631 (.002)	0.633 (.009)
10	4664	0.645	0.653 (.001)	0.650 (.007)	0.662 (.002)	0.659 (.008)	0.672 (.002)	0.669 (.008)
11	5134	0.595	0.602 (.001)	0.605 (.007)	0.609 (.001)	0.612 (.008)	0.617 (.002)	0.620 (.008)
12	7184	0.614	0.621 (.001)	0.622 (.006)	0.629 (.001)	0.630 (.006)	0.638 (.002)	0.639 (.007)
13	9571	0.644	0.652 (.001)	0.651 (.005)	0.661 (.001)	0.660 (.006)	0.671 (.002)	0.670 (.006)
14	10049	0.620	0.627 (.001)	0.627 (.005)	0.635 (.001)	0.636 (.005)	0.644 (.002)	0.645 (.006)
15	19316	0.610	0.617 (.000)	0.618 (.004)	0.625 (.001)	0.626 (.004)	0.634 (.001)	0.634 (.004)
16	20315	0.623	0.630 (.000)	0.630 (.004)	0.639 (.001)	0.639 (.004)	0.648 (.001)	0.648 (.004)
17	35477	0.615	0.622 (.000)	0.622 (.003)	0.630 (.001)	0.630 (.003)	0.639 (.001)	0.639 (.003)
18	49975	0.617	0.625 (.000)	0.625 (.002)	0.633 (.000)	0.633 (.002)	0.642 (.001)	0.642 (.003)
Year 2006								
No.	N_i	$\bar{x}_i$	$\hat{\bar{y}}_i^{(1)}$	$\hat{\theta}_i^{(1)}$	$\hat{\bar{y}}_i^{(2)}$	$\hat{\theta}_i^{(2)}$	$\hat{\bar{y}}_i^{(3)}$	$\hat{\theta}_i^{(3)}$
1	426	0.646	0.654 (.002)	0.648 (.017)	0.662 (.004)	0.657 (.018)	0.672 (.005)	0.666 (.019)
2	961	0.637	0.645 (.002)	0.644 (.014)	0.654 (.003)	0.653 (.015)	0.664 (.004)	0.663 (.015)
3	2320	0.639	0.648 (.001)	0.647 (.010)	0.657 (.002)	0.656 (.011)	0.667 (.003)	0.666 (.011)
4	2427	0.596	0.603 (.001)	0.612 (.010)	0.611 (.002)	0.620 (.011)	0.620 (.003)	0.628 (.012)
5	2850	0.594	0.601 (.001)	0.608 (.010)	0.609 (.002)	0.616 (.010)	0.617 (.003)	0.625 (.011)
6	2899	0.679	0.688 (.001)	0.681 (.010)	0.698 (.002)	0.691 (.010)	0.709 (.003)	0.702 (.011)
7	3244	0.658	0.667 (.001)	0.664 (.009)	0.677 (.002)	0.673 (.009)	0.687 (.003)	0.684 (.010)
8	3453	0.648	0.656 (.001)	0.654 (.008)	0.666 (.002)	0.664 (.009)	0.676 (.003)	0.674 (.009)
9	3514	0.616	0.623 (.001)	0.626 (.009)	0.632 (.002)	0.635 (.009)	0.641 (.003)	0.644 (.010)
10	4664	0.654	0.663 (.001)	0.661 (.007)	0.673 (.002)	0.671 (.008)	0.683 (.002)	0.681 (.008)
11	5134	0.609	0.616 (.001)	0.619 (.007)	0.624 (.001)	0.627 (.008)	0.633 (.002)	0.636 (.008)
12	7184	0.629	0.637 (.001)	0.637 (.006)	0.646 (.001)	0.646 (.006)	0.655 (.002)	0.656 (.007)
13	9571	0.658	0.667 (.001)	0.666 (.005)	0.677 (.001)	0.676 (.006)	0.688 (.002)	0.687 (.006)
14	10049	0.635	0.644 (.001)	0.644 (.005)	0.653 (.001)	0.653 (.005)	0.662 (.002)	0.662 (.006)
15	19316	0.630	0.638 (.000)	0.638 (.004)	0.647 (.001)	0.647 (.004)	0.656 (.001)	0.656 (.004)
16	20315	0.636	0.644 (.000)	0.644 (.004)	0.653 (.001)	0.653 (.004)	0.663 (.001)	0.663 (.004)
17	35477	0.634	0.642 (.000)	0.642 (.003)	0.651 (.001)	0.651 (.003)	0.661 (.001)	0.661 (.003)
18	49975	0.634	0.643 (.000)	0.643 (.002)	0.651 (.000)	0.651 (.002)	0.661 (.001)	0.661 (.003)

Table 5. Observed change $\tilde{\bar{z}}_i$ of register employment rate from year 2005 ($t = 1$) to 2006 ($t = 2$), estimated underlying change $\hat{\bar{y}}_{i,t=2}^{(k)} - \hat{\bar{y}}_{i,t=1}^{(k)}$, and directly estimated theoretical change $\hat{\delta}_i^{(k)}$, given alternative process errors: $(p_{i1}, p_{i0}) = (0.972, 0.027)$ if $k = 1$, $(p_{i1}, p_{i0}) = (0.950, 0.044)$ if $k = 2$, $(p_{i1}, p_{i0}) = (0.928, 0.061)$ if $k = 3$. Estimated root mean squared error (rmse) in parentheses.

No.	N_i	$\tilde{\bar{z}}_i$	$\hat{\bar{y}}_{i,t=2}^{(k)} - \hat{\bar{y}}_{i,t=1}^{(k)}$			$\hat{\delta}_i^{(k)}$		
			$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
1	426	–0.001	0.000	0.000	0.001	0.013 (.006)	0.013 (.006)	0.015 (.006)
2	961	0.034	0.035	0.036	0.037	0.019 (.005)	0.020 (.005)	0.019 (.004)
3	2320	0.005	0.005	0.006	0.006	0.011 (.006)	0.011 (.006)	0.014 (.006)
4	2427	0.009	0.010	0.011	0.011	0.013 (.005)	0.013 (.005)	0.015 (.005)
5	2850	–0.004	–0.003	–0.003	–0.003	0.005 (.006)	0.006 (.007)	0.010 (.007)
6	2899	0.015	0.016	0.016	0.017	0.015 (.005)	0.016 (.005)	0.017 (.005)
7	3244	0.013	0.013	0.014	0.015	0.014 (.005)	0.015 (.005)	0.016 (.005)
8	3453	0.017	0.018	0.018	0.019	0.016 (.005)	0.017 (.005)	0.017 (.005)
9	3514	0.009	0.009	0.010	0.010	0.012 (.005)	0.012 (.005)	0.014 (.005)
10	4664	0.010	0.010	0.011	0.011	0.012 (.005)	0.013 (.005)	0.014 (.005)
11	5134	0.014	0.014	0.015	0.016	0.014 (.004)	0.015 (.004)	0.016 (.005)
12	7184	0.015	0.016	0.016	0.017	0.015 (.004)	0.016 (.004)	0.017 (.004)
13	9571	0.014	0.015	0.016	0.017	0.015 (.003)	0.016 (.003)	0.016 (.004)
14	10049	0.016	0.017	0.017	0.018	0.016 (.003)	0.017 (.003)	0.017 (.004)
15	19316	0.020	0.021	0.021	0.022	0.020 (.002)	0.021 (.003)	0.021 (.004)
16	20315	0.013	0.014	0.015	0.016	0.014 (.002)	0.015 (.003)	0.016 (.003)
17	35477	0.019	0.020	0.021	0.022	0.020 (.002)	0.020 (.002)	0.021 (.003)
18	49975	0.017	0.018	0.019	0.019	0.018 (.002)	0.018 (.002)	0.019 (.002)

Municipalities. It seems that, provided the misclassification probabilities are stable over time, the change in register employment rates is less affected by the amount of delayed entries. It is interesting to take another look at the two "outlying" Municipalities No. 1 and 5 in Table 3. Allowing for misclassification does not change the impression that the ostensible negative developments in these two Municipalities may be attributed to individual variations, and does not necessarily represent any underlying negative trend in the respective labor market. In fact, as the misclassification probabilities increase from $k = 1$ to $k = 3$, the directly estimated theoretical change becomes increasingly positive in both areas. The sensitivity analysis thus strengthens the conclusion that has been reached in the absence of misclassification errors. Of course, the sensitivity analysis here has its own premises, and one must avoid drawing overly extensive conclusions. For instance, one possibility is to vary the misclassiûcation probabilities from one year to another in a plausible way, and to extend the above sensitivity analysis approach to cover more complex alternative scenarios.

4. CONCLUSION

The production of statistics at detailed spatial-demographic levels can greatly benefit from the use of administrative register data. In this paper we proposed a modeling approach that accounts for different kinds of random variations in the register data, so as to assess the 'signals' of interest from the measurement 'noises' that are ever present. Distinction is made between

randomness at the population level, the individual level as well as along the registration/production process. Using the data from the Norwegian Employer/ Employee register, we demonstrated the existence of two types of process errors in terms of the measurement of a register-based employment status. Of these the delayed entries have a much greater amount than the recurred entries in the data of this study. Moreover, the delayed entries are non-ignorable because they do affect the classification of the employment status. We outlined a sensitivity analysis approach that covers a plausible range of alternative scenarios of the misclassification mechanism based on historic data. This provides us with helpful uncertainty measures of the *unadjusted* register-based statistics. For future research it will be interesting to develop more elaborated models, which allow for differential error mechanisms both over time and across the population domains.

ACKNOWLEDGEMENT

We would like to thank Inge Aukrust and Per Svein Aurdal for their help in understanding the data.

REFERENCES

Aukrust, I., Aurdal, P.S., Bråthen, M. and Køber, T. (2010). Register-based Employement Statistics. Technical Report "Notater 8/2010", Statistics Norway, in Norwegian.

Groves, R.M., Fowler Jr., F.J., Couper, M., Lepkowski, J.M., Singer, E. and Tourrangeau, R. (2004). *Survey Methodology*. Wiley, New York.

Hedlin, D., Fenton, T., McDonald, J.W., Pont, M. and Wang, S. (2006). Estimating the undercovergae of a sampling frame due to reporting delays. *J. Official Statist.*, **22**, 53–70.

Holt, D. (2007). The official statistics Olympic challenge: Wider, deeper, quicker, better, cheaper. (With discussions). *Amer. Statist.*, **61**, 1–15.

Linkletter, C.D. and Sitter, R.R. (2007). Predicting natural gas production in Texas: An application of nonparametric reporting lag dsitribution estimation. *J. Official Statist.*, **23**, 239–251.

Lohr, S. and Rao, J.N.K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, **96**, 457–468.

Rao, J.N.K. (2003). *Small Area Estimation.* Wiley, New York.

Zhang, L.-C. (2011). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, to appear.