Quality evaluation of employment status in register-based census

Fosen, Johan
Statistics Norway, Divison for Statistical Methods and Standards
Kongens gate 6, P.O.Box 8131 Dep.
N-0033 Oslo, Norway
E-mail: johan.fosen@ssb.no

Zhang, Li-Chun
Statistics Norway, Divison for Statistical Methods and Standards
Kongens gate 6, P.O.Box 8131 Dep.
N-0033 Oslo, Norway
E-mail: li.chun.zhang@ssb.no

1 Introduction

The register-based employment statistics of Norway is a statistics disseminated annually by Statistics Norway. The employment variable measures whether or not a person is employed during the third week of November, and is constructed for each person in the target population by means of micro-integration. During the micro-integration, several administrative registers are linked on micro-level and the information is harmonised before the integration ends in a classification of a person as employed or not employed. The micro-integration is described in Fosen (2010).

We shall evaluate the accuracy of register-based employment statistics, *REG-employment*, for small areas such as municipalities, of which there are about 430 in Norway, and more than half of them have less than 5000 inhabitants. We will compare the mean squared error (MSE) of REG-employment, *MSE-REG*, and of LFS-employment, *MSE-LFS*. Our approach does not require linkage of the two data sources on the individual level. This can be useful in situations where such individual-level linkages are either impossible, difficult, or prohibited by legal reasons.

The main source of MSE for the REG-employment proportion is bias, and for simplicity we will assume no variance at all. For LFS we will assume no non-sampling error such as e.g. nonresponse error. Then, the contribution to MSE of LFS-employment proportion comes only from variance, and only from bias in the case of REG-employment proportion. For short we will just write *LFS-variance* and *REG-bias*, being the LFS-MSE and square root of REG-MSE respectively.As opposed to the LFS-variance, the REG-bias is not by nature a function of the sample size n. On national level, the squared REG-bias is expected to behigher than the LFS-variance. However, if we consider smaller and smaller regions, the LFS-variance can be expected to eventually dominate over the REG-bias.

In Section 2, we estimate LFS-MSE using a smoothing method. The estimation of MSE-REG is devoted to Section 3 where we will use small area modelling. In Section 4 we look at the results of the MSE comparison for the two models, using two different approaches. Finally in Section 5 we look at some properties of the bias estimators and suggest some improvements.

2 The estimated standard deviation of the LFS-proportion

Let i denote the municipality, and Y_i the number of employed according to LFS in this sub population. We consider the simple sample average $\overline{y}_i$ as the LFS-employment proportion estimate. The standard deviation of $\overline{y}_i$ is given by $\sqrt{\psi_i} = \sqrt{\theta_i(1-\theta_i)/n_i}$, where the parameter θ_i is the true employment proportion. Due to small municipalities, the estimates $\hat{\theta}_i$ have a large variation and then also the direct estimator $\hat{sd}_{dir}(\overline{y}_i) = \sqrt{\hat{\theta}_i(1-\hat{\theta}_i)/n_i}$. Instead we use the generalised variance function GVF

$$\hat{\mathrm{sd}}_{\mathrm{GVF}}(\overline{y}_i) = e^{-0.749}\mathrm{n_i}^{-1.030/2}, \tag{1}$$

found by regressing $\log\left[\hat{\mathrm{sd}}_{dir}(\overline{y}_i)\right]$ onto $\log(\sqrt{n_i})$.

3 Modelling the bias of the register-based statistics

We can write

$$\overline{y}_i = \theta_i + e_i,\ i = 1,2,\ldots m, \tag{2}$$

where $e_i = \overline{y}_i - \theta_i$. We notice that $\mathrm{Var}(e_i) = \mathrm{Var}(\overline{y}_i) = \psi_i$. Let Z_i be the number of employed according to REG in municipality i. We assume the model

$$\overline{Z}_i = \theta_i + b_i,\ i = 1,2,\ldots m, \tag{3}$$

where $\overline{Z}_i$ is the employment proportion, b_i is the bias of $\overline{Z}_i$, and m is the number of municipalities in the data set. With $\overline{X}_i = \overline{Z}_i - \overline{y}_i$ we have, when combining (2) and (3),

$$\overline{X}_i = b_i + \tilde{e}_i, \tag{4}$$

where $\tilde{e}_i = -e_i$. We assume the bias to be a linear model

$$b_i = \beta + v_i, \tag{5}$$

of the *underlying bias* β and the random variable v_i representing the unexplained variation between the biases. When not otherwise specified we will by *bias* refer to b_i. Putting (4) and (5) together we then have the linear mixed model

$$\overline{X}_i = u_i^T\beta + v_i + \tilde{e}_i, \quad u_i = (1,\ldots 1)^T. \tag{6}$$

3.1 An alternative model

For each person, we have register information about the register source for being classified as employed or not employed (Fosen 2010). Based on prior knowledge on the quality of different register sources, we divide the population into a high quality group 1, and a group 2 containing the rest of the population.

We use subscript g to denote group. Thus, Z_{gi} and Y_{gi} are the number of REG-employed and LFS-employed in group g in municipality i. The number of persons in this group is N_{gi}, and the proportion being REG-employed and LFS-employed is $\overline{Z}_{gi}$ and $\overline{Y}_{gi}$.

For group 1 we have two special properties: firstly, it contains only persons being classified as employed, i.e. $Z_{1i} = N_{1i}$ and thus $\overline{Z}_{1i} = 1$. Secondly, we assume no bias in this group: the group contains only persons being employed according to LFS. Then we can estimate Y_{1i} by

$$\hat{Y}_{1i} = Z_{1i}, \tag{7}$$

and further that model (6) for group 1 reduces to $\overline{X}_{1i} = \tilde{e}_{1i}$.

For group 2, we assume model (6), i.e.

$$\overline{X}_{2i} = \beta_2 + v_{2i} + \tilde{e}_{2i}. \tag{8}$$

We then have

$$\overline{X}_i = \overline{Z}_i - \overline{Y}_i = \frac{N_{1i}}{N_i} + \overline{Z}_{2i}\frac{N_{2i}}{N_i} - \overline{Y}_{1i}\frac{N_{1i}}{N_i} - \overline{Y}_{2i}\frac{N_{2i}}{N_i}, \tag{9}$$

which simplifies into

$$\overline{X}_i = \frac{N_{1i}}{N_i} - \overline{Y}_{1i}\frac{N_{1i}}{N_i} + \overline{X}_{2i}\frac{N_{2i}}{N_i}, \tag{10}$$

since $\overline{X}_{2i} = \overline{Z}_{2i} - \overline{Y}_{2i}$.

We model $\overline{Z}_{2i}$ in the same way as before and let $\overline{Z}_{2i} = \overline{Y}_{2i} + b_{2i}$, where $b_{2i} = \beta_2 + v_{2i}$, and v_{2i} is a random effect at the municipality level. Using (7), an estimator of $\overline{Y}_{2i}$ is given by

$$\hat{\overline{Y}}_{2i} = (\hat{Y}_i - Z_{1i}) / N_{2i} = (N_i \overline{y}_i - Z_{1i}) / N_{2i}, \tag{11}$$

where $\overline{y}_i$ is the LFS-employment level in the sample and $\hat{Y}_i = N_i \overline{y}_i$. If we add and subtract $\overline{Y}_{2i}$, we get

$$\hat{\overline{Y}}_{2i} = \overline{Y}_{2i} + \frac{\hat{Y}_i - Y_{2i} - Z_{1i}}{N_{2i}} = \overline{Y}_{2i} + \frac{\hat{Y}_i - Y_i}{N_{2i}} = \overline{Y}_{2i} + e_{2i}, \tag{12}$$

where e_{2i} is the associated sampling error $\left(\hat{Y}_i - Y_i\right) / N_{2i}$, which is our best estimate since we are unable to identify the groups within LFS. The expected value of e_{2i} is zero and the variance is $V(\overline{y}_i)(N_i / N_{2i})^2$.

From (11) we now have

$$\begin{aligned} &\overline{Z}_{2i} - \hat{\overline{Y}}_{2i} = (Z_i - Z_{1i}) / N_{2i} - (\hat{Y}_i - Z_{1i}) / N_{2i} = (Z_i - N_i \overline{y}_i) / N_{2i} = (\overline{Z}_i - \overline{y}_i)(N_i / N_{2i}) \\ &= \overline{x}_i (N_i / N_{2i}) \end{aligned}, \tag{13}$$

where we notice that $\overline{x}_i$ is the observed difference between the REG-employment proportion in the municipality population and the LFS-employment proportion in the sample.

Similarly as for (4), we now write

$$\overline{Z}_{2i} - \hat{\overline{Y}}_{2i} = b_{2i} - e_{2i}, \tag{14}$$

which we insert into the left-hand side of (13). Then we have the following linear mixed model

$$\overline{x}_i = \frac{N_{2i}}{N_i} b_{2i} - \frac{N_{2i}}{N_i} e_{2i} = \frac{N_{2i}}{N_i} b_{2i} + \varepsilon_i, \tag{15}$$

where

$$E(b_{2i}) = \beta_2 \quad \text{and} \quad V(b_{2i}) = \sigma_{v2}^2$$

and

$$E(\varepsilon_i) = 0 \quad \text{and} \quad V(\varepsilon_i) = V(\overline{y}_i)$$

Compared to the model of x_i earlier, the knowledge of $Z_{1i} = N_{1i} = Y_{1i}$ is incorporated as a shrinkage factor (i.e. always between 0 and 1) of the random effect b_{2i}. We assume the linear relation $b_{2i} = \beta_2 + v_{2i}$, similar to (5). Then we can write (15) as

$$\overline{x}_i = u_i\beta_2 + c_i v_{2i} + \varepsilon_i, \tag{16}$$

where $u_i = c_i = N_{2i} / N_i$ and $b_i = u_i\beta_2 + c_i v_{2i}$. This linear mixed model is an alternative to (6).

3.2 Fitting the multilevel model

We assume v_i and $\tilde{e}_i$ to be independent with expectation zero. In the case where the distribution of $\tilde{e}_i$ is known, we have a special case of a *basic Type A area level small area model* of Rao (2003, Chapter 5), which can be written as

$$\phi_i(\overline{X}) = u_i^T\beta + c_i v_i + \tilde{e}_i.$$

We identify our models (6) and (16) as special cases, with $\phi(\cdot) = I(\cdot)$. For the simpler model (6) we have $c_i = u_i = 1$, whereas $c_i = u_i = N_{2i} / N_i$ under model (16).

We assume the variance ψ_i of $\tilde{e}_i$ known and given by (1). We want to estimate the bias

$$b_i = u_i\beta + c_i \mathrm{v}_i \tag{17}$$

where $u_i\beta$ is the underlying bias and $c_i v_i$ is the unexplained variation between the areas. The best linear unbiased prediction (BLUP) estimator for our model becomes (Rao 2003; Section 7.1.1)

$$\tilde{b}_i = \hat{\overline{X}}_i = \gamma_i \overline{X}_i + (1-\gamma_i) u_i\hat{\beta}, \tag{18}$$

where

$$\gamma_i = \frac{\sigma_v^2 c_i^2}{\psi_i + \sigma_v^2 c_i^2} \qquad \text{and} \qquad \hat{\beta} = \sum_i \frac{u_i \overline{X}_i}{\psi_i + \sigma_v^2 c_i^2}\left[\sum_i \frac{u_i^2}{\psi_i + \sigma_v^2 c_i^2}\right]^{-1}. \tag{19}$$

The BLUP is a weighted average of the directly observed difference $\overline{X}_i$ and the model-induced bias $u_i\beta$. The relative thrust γ_i put in the direct observation equals the proportion of the variance being between-area variation $\sigma_v^2 c_i^2$. For areas with few observations n_i, the sampling variance, i.e. the within-area variation ψ_i is large, hence γ_i is small and little thrust is put into $\overline{X}_i$. By inserting the estimated variances $\hat{\psi}_i$ and $\hat{\sigma}_v^2$ into (18), we get the EBLUP estimator.

The estimation of σ_v^2 is done using the *iterative Fay-Herriot* method suggested in Rao (2003;Section 7.1.2), where the $(a+1)$-th iteration is

$$\sigma_v^{2(a+1)} = \sigma_v^{2(a)} + \frac{m - p - h\left(\sigma_v^{2(a)}\right)}{h'\left(\sigma_v^{2(a)}\right)}, \tag{20}$$

where $h\left(\sigma_\zeta^{2(a)}\right) = \sum_i \left(\overline{X}_i - u_i\hat{\beta}\right)^2 \left(\psi_i + \sigma_v^2 c_i^2\right)^{-1}$ and $h'\left(\sigma_v^{2(a)}\right) = -\sum_i c_i^2 \left(\overline{X}_i - u_i\hat{\beta}\right)^2 \left(\psi_i + \sigma_v^2 c_i^2\right)^{-2}$.
Usually 10 iterations are sufficient for convergence of the algorithm.

4 Results

When fitting the multilevel models (6) and (16) we use the data set of all municipalities where the LFS net sample size is at least two persons.

4.1 Parameter estimates

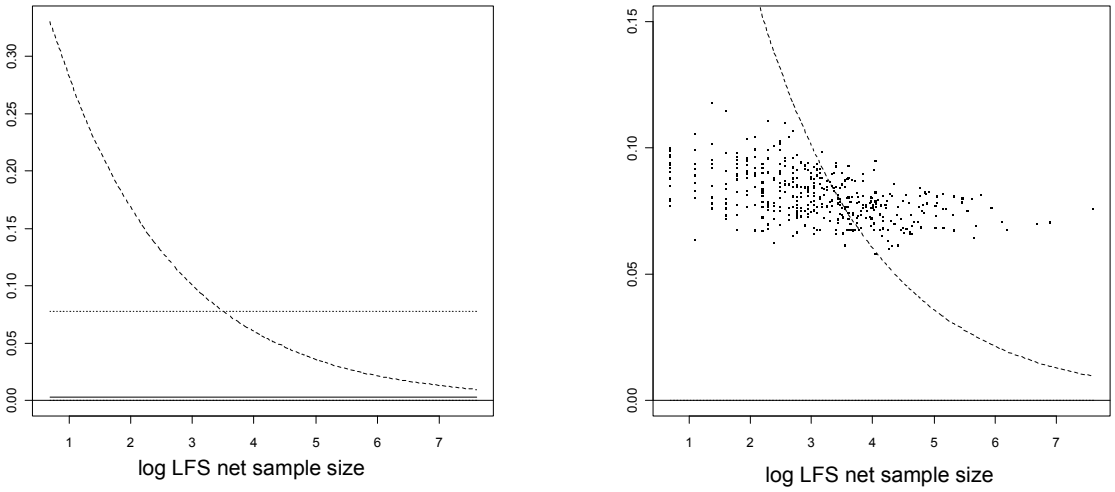
For model (6) we have $u_i = c_i = 1$, and the parameter estimates become $\hat{\beta} = 0.00302$ and $\hat{\sigma}_v$=0.03825. Since this model has no covariates, the underlying bias $u_i\beta$ of the bias (17) is $\hat{\beta} = 0.00302$.

For the heterogeneity model (16) we have $u_i = c_i = N_{2i} / N_i$. We use superscript H to distinguish from model (6), and get $\beta_2^H = 0.00914$ and $\sigma_{2v}^H = 0.09228$. The estimated underlying bias $\beta_2 u_i$ of the bias (17) is averagely 0.0039 (and median 0.0038), thus 30 percent higher than the underlying bias above of model (6).

4.2 Comparison of REG-employment and REG-employment using the distribution of the underlying bias estimator

The small area model assumes that the underlying REG-bias $u_i\beta$ (as well as the single biases b_i) does not depend on sample size. However, the LFS-variance is decreasing with sample size. If we decrease the sample size, we expect at some point that REG-employment outperforms LFS-employment in terms of MSE. For model (6) this is illustrated in the left hand panel of Figure 1, and represents one approach to MSE-comparisons. The estimated underlying REG-bias $u_i\hat{\beta} = \hat{\beta}$ is positive and its 95 percent confidence interval is also a 95 percent confidence interval of the square root of the REG-MSE. On the other hand, the estimated LFS-employment standard deviation is the estimated square root of LFS-MSE. We see that when log sample size is less than 3.5, i.e. sample size less than 33, we can be 95 percent certain that REG-employment is better since MSE of LFS-employment is then higher than the 95% confidence interval of MSE of REG-employment. For larger sample sizes, we can not conclude one way or the other by this plot.

Figure 1. Square root of MSE for LFS-employment (dashed line), upper 95% confidence interval for square root of MSE for REG-employment (dotted lines), and the estimated underlying REG-bias $u_i\hat{\beta}$ ***(solid line); lower confidence interval truncated to zero. Model (6) in left panel and model (16) in right panel.***

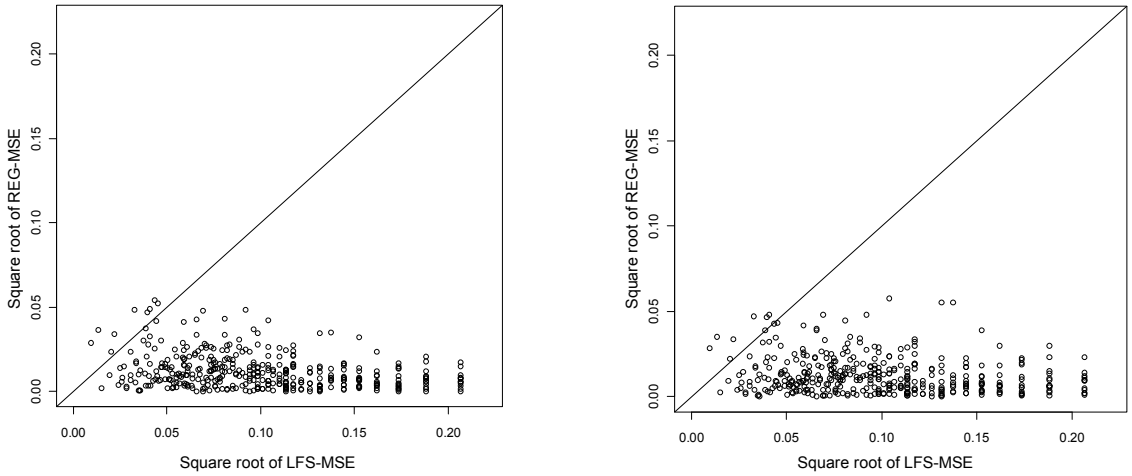


For model (16) we see in the right panel of Figure 1 that when $\log(n_i) < 3.1$, i.e. sample size up to 22 and municipality size roughly below 3500, we can conclude that REG-employment has a lower MSE than LFS-employment. For $3.1 < \log(n_i) < 3.8$, i.e. sample size between 22 and 44 (municipality size roughly below 6500) we can draw this conclusion for some of the municipalities, whereas for larger sizes this way of comparison is inconclusive. For model (6) we remember that $\log(n_i) < 3.5$ makes REG better, otherwise this comparison is inconclusive.

4.3 Comparison of REG-employment and REG-employment using the EBLUP estimator

We now compare the individual municipality EBLUP estimates of REG-bias against the LFS GVF-function. The left panel of Figure 2 shows that when we use the EBLUP estimator $\tilde{b}_i$ of model (6) for the REG-employment bias, the REG- MSE is smaller than LFS-MSE for almost all the municipalities. Similarly for the right panel for model (16).

Figure 2. The square root of MSE of REG-employment against that of LFS-employment. MSE of REG-employment based on EBLUP estimates. Left panel is model (6) and right panel is model (16).



5 Adjusting the EBLUP estimator

Under the assumed model, which in this section will be (6), the EBLUP estimator $\tilde{b}_i$ gives the best *area specific* estimates among all linear estimates, i.e. the best estimates for the municipalities when regarded one by one. According to the same model, the bias of REG-employment has expectation β and variance σ_v^2. However, the EBLUPs in general do not possess this ensemble property, in which we often may be interested as in the comparison of MSE between REG and LFS in Figure 2. The EBLUP estimator (18) shrinks $\overline{X}_i$ towards the underlying common bias β, causing the empirical variance of $\tilde{b}_i$ to be smaller than σ_v^2, for which reason the problem is known as *overshrinkage.*

Assume that $b_i \sim N(\beta, \sigma_\zeta)$. We can construct the overshrinkage-adjusted estimator $\tilde{b}_i^{*G}$ by first sorting $\tilde{b}_i$ into $\tilde{b}_{(i)}, i = 1, \ldots m$. Then we replace the i-th smallest $\tilde{b}_{(i)}$ of the $\left\{\tilde{b}_j\right\}$ by the $\frac{i}{m}$-quantile in the $N(\hat{\beta}, \hat{\sigma}_v)$-distribution, giving a new set of predicted biases $\left\{\tilde{b}_i^{*G}\right\}$ having the desired distribution. Such a simultaneous approach has been considered by Zhang (2003). This algorithm limits the deviations $\tilde{b}_i^{*G} - \tilde{b}_i$ from the best area-specific estimate, by keeping the order of the municipality-estimates before and after the adjustment.

Since the $\tilde{b}_i$ are over-shrunk towards the global expectation, the largest REG-MSEs are underestimated, i.e. in favour of the REG-employment in the comparison with LFS-MSE in such cases. After overshrinkage reduction, we see in Figure 3 that there are some more municipalities where LFS-MSE becomes smaller than REG-MSE, despite the overall conclusion remains.

Figure 3. The square root of MSE of REG-employment against that of LFS-employment, for municipalitites. REG-bias based on the Gaussian approach, with bias estimates $\tilde{b}_i^{*G}$. Model (6)

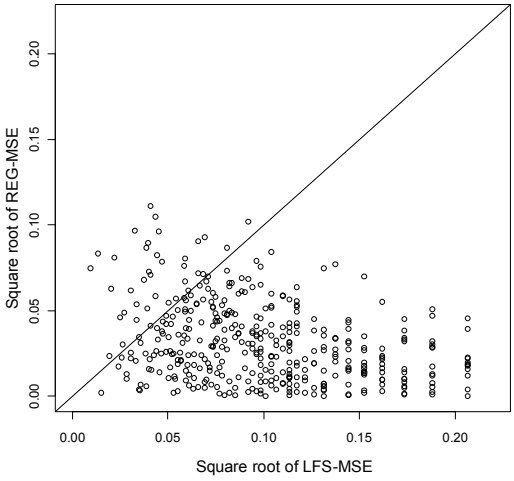


Figure 4. The bias of REG-employment based on the EBLUP estimates. Model (6).

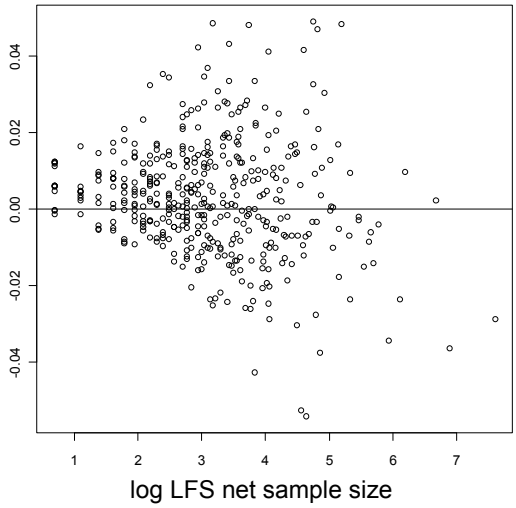


Figure 4 shows that the variation of $\tilde{b}_i$ increases with the municipality sample size. However, we have no reason to believe that the true variation should depend on the sample size in this way. From (18) we see that γ_i is an increasing function in the municipality sample size. Therefore, to make the variation of $\tilde{b}_i$ increase less rapidly, we may use a tranformed γ_i, such as $\hat{\gamma}_i^{of} = \left(\hat{\gamma}_i\right)^{1/\lambda}$. By selecting $\lambda = 2$, we get an estimator $\tilde{b}_i^{of}$ whose empirical variance $S^2_{\tilde{b}^{of}}$ becomes almost identical to the estimated variance $\sigma_v^{\,2}$, hence also this approach results in an overshrinkage corrected estimator. A constrained empirical Bayesian (CEB) justification of this particular value $\lambda = 2$ was given by Spjøtvoll and Thomsen (1987). We notice that $\hat{\gamma}_i^{of}$ is larger than $\hat{\gamma}_i$ for all sample sizes, and greater emphasis is put on the direct estimator $\overline{X}_i$ compared to (18) for *all* municipalities.

Figure 5. Bias of REG-employment using EBLUP estimates ('1') and when using adjusted ('2'). For the following methods: Gaussian-based overshrinkage-adjusted $\tilde{b}_i^{*G}$ (left panel), and constrained empirical Bayesian (CEB) overshrinkage-adjusted $\tilde{b}_i^{of}$ (right panel). Model (6).

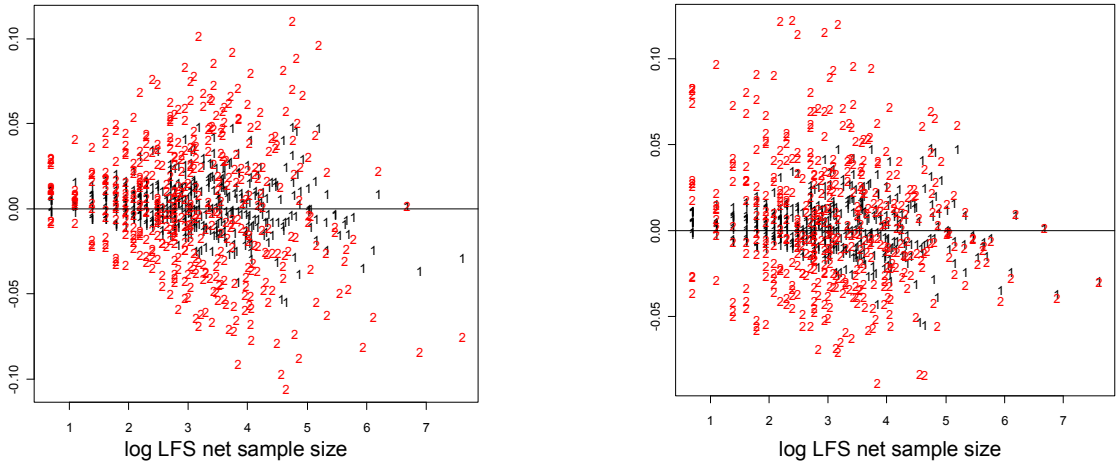
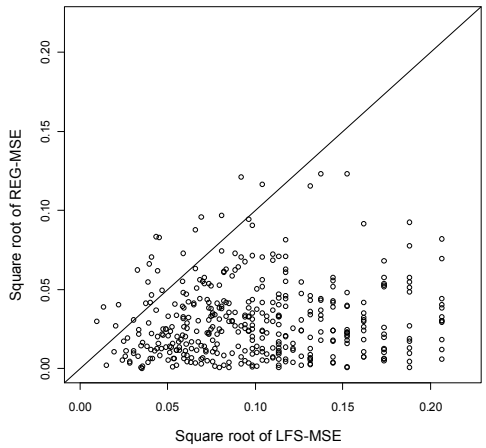


Figure 5 shows that $\hat{b}_i^{of}$ is more uniformly distributed with regard to the municipality sample size (left panel), compared to $\tilde{b}_i^{*G}$ (right panel).

We notice from the right panel of Figure 5 that $\hat{b}_i^{of}$ hardly adjusts the EBLUP estimator $\tilde{b}_i$ for the 15-20 largest municipalities. This is in contrast to $\tilde{b}_i^{*G}$ which clearly adjusts $\tilde{b}_i$ also for these municipalities, and even for the largest municipality Oslo. Intuitively, we would want the adjustment for Oslo and other larger municipalities to be limited since the best area specific estimates $\tilde{b}_i$ are more precise for these municipalities. For Oslo, $\tilde{b}_i$ takes as much as 96 percent of its value from the direct estimator and only 4 percent from the common $\hat{\beta}$. The emphasis on the direct estimator is even stronger for $\hat{b}_i^{of}$ which for Oslo takes 99 percent of its value from the direct estimate. Figure 5 reveals that the overshrinkage adjustment method $\tilde{b}_i^{*G}$ has a drawback in that $\tilde{b}_i$ is modified with no regard to the sample size of the municipalities. Meanwhile, in Figure 6 we see that the number of municipalities where REG-MSE exceeds LFS-MSE is approximately by the CEB overshrinkage adjustment method as by the approach underlying Figure 3.

Figure 6. The square root of MSE of REG-employment against that of LFS-employment, for municipalitites. REG-bias based on the CEB overshrinkage adjusted bias estimate $\hat{\beta}_i^{of}$. Model (6).



6 Conclusions

We have described an approach for comparing register-based statistics with survey-based statistics that does not require linkage of the data across the sources on the individual level. Essentially, this comes down to the trade-off between the bias of the register data and the sampling variance of the survey data. Small area estimation techniques are used to estimate the bias of the register-based statistics. Adjustment for overshrinkage of the area-specific best estimates has been considered, which may affect the comparison. The methodology was illustrated using the Norwegian register-based employment and LFS data.

Acknowledgements

This work has been a part of the ESSnet Data Integration project for which Eurostat provides the major part of the funding. The participating National Statistical Institutes are Italy, the Netherlands, Norway, Poland, Spain and Switzerland.

REFERENCES

Fosen, J. (2011). Register-based employment statistics. A case of micro-integration. To appear as a part of an Essnet Data-Integration report on case studies.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley Series in Survey Methodology, Wiley.

Spjøtvoll, E. and Thomsen, I. (1987). Application of some empirical Bayes methods to small area statistics. *Bulletin of the International Statistical Institute,* vol. 2, 435-449.

Zhang, L.-C. (2003). Simultaneous estimation of the mean of a binary variable from a large number of small areas. *Journal of Official Statistics,* vol. 19, 253-263.