Investigation of Estimation Methods for AWE

Authors: Neil Parkin, Markus Šova, John Wood and Philip Lewis.

1. Executive summary

The ONS has completed the work carried out in response to recommendation 1 of the review of the Average Earnings Index and Average Weekly Earnings (Weale, 2008 p 5) "*AWE should not become a National Statistic until further work has been carried out on the possible use of matched pairs. This work needs to compare the use of matched pairs or a combination of imputation and matched pairs with the existing AWE methodology to see which produces more reliable estimates of annual growth rates.*"

This note describes that work and supersedes the previous update (Parkin et al., 2009). The result of the work is that:

the ONS recommends that AWE continue to be calculated using the full sample from the Monthly Wages and Salaries Survey. This is recommended because the advantage of using matched pairs, being small and extremely short lived, is insufficient to outweigh the cost of changing the estimator.

It has been calculated that the standard error of one month growth in AWE, using the current method, would be 36% larger than the standard error of the matched sample estimator. The difference in standard error of growth over two months (lag 2) is 11%, which is much smaller than the difference at lag 1, the difference is smaller still for lag 3. There is no large difference between the standard errors of the two estimators after lag 3. These points are summarised in table 1, which shows the mean[1] of a measure of the *advantage* of matched pairs. The advantage is defined as the ratio of the standard error of change of the full sample estimator to the standard error of change of the matched sample estimator. This means that when the matched sample estimator is more precise the advantage is greater than one, and when the full sample estimator is more precise the advantage is less than one. When the advantage is one then the two estimators have the same precision.

lag	1	2	3	6	12	18
advantage	1.36	1.11	1.05	1.01	1.00	0.99

Table 1. The average advantage of the matched sample AWE for selected lags in the time span Aug. 2000 to Dec. 2008.

It should be noted that due to resource constraints no account has been made of bias in the matched sample estimator[2]. It is clear from earlier work on the estimator for retail sales (Kokic and Jones, 1998) that the likely bias will increase the mean square error of the matched sample estimator, and more so at larger lags. The reader can find an extended justification for the recommendation in the conclusion, section 4.

The slight advantage of the matched sample estimator, where it exists, is due to: (a) a higher correlation between successive estimates; and (b) being less susceptible to sample rotation and outliers (when outliers are not treated). These points are explained in more detail in the results section, but there is first a brief description of the methods used to calculate the standard error ratios.

2. Method

2.1 Theory

From theory, one would anticipate that estimating the series using the matched sample would result in smaller variances of short-term movement than from using the full sample, because of a higher correlation between successive estimates. However, one would anticipate that the variances of longer-term movement would be larger using a matched sample, because the size of the matched sample will be smaller than the full sample and the diminishing correlations will have little impact over longer time periods.

The research described in this note was carried out to determine the effect, on the variance of growth, of using the matched sample estimator for AWE. Due to the practical difficulties in estimating the bias, no attempt was made to estimate the effect of the matched sample on the bias.

[1] In order to avoid bias in calculating the mean of ratios, the mean has been calculated by taking the exponent of the arithmetical mean of the log standard error ratio.

[2] This would have involved the simulation described in Weale's report (pp42-3).

The theoretical work, carried out by Wood, 2008, was technically challenging and a number of simplifying assumptions were made in order to make progress, details can be found in annex A. In brief, the variance of the difference in the levels of AWE was calculated using the following formulae, for the full sample AWE,

$$V_{F,L} = \mathrm{Var}\left(\hat{\mu}_t - \hat{\mu}_{t-L}\right) = \mathrm{Var}\left(\hat{\mu}_t\right) + \mathrm{Var}\left(\hat{\mu}_{t-L}\right) - 2\,\mathrm{Cov}\left(\hat{\mu}_t, \hat{\mu}_{t-L}\right) \tag{1}$$

where $\hat{\mu}_t$ is the estimate of level of AWE at time t, based on the full sample, $\hat{\mu}_{t-L}$ is the estimate of level of AWE at time t-L, based on the full sample, where L is the time lag (the lag in the span of time over which differences are being compared). For the AWE calculated on the matched sample, the estimate of variance of the same difference is,

$$V_{M,L} = \mathrm{Var}\left(\hat{\mu}_t - \hat{\mu}_{t-L}\right) = \mathrm{Var}\left(\hat{\mu}_{t,t-1}\right) + \mathrm{Var}\left(\hat{\mu}_{t-L,t-L+1}\right) - 2\,\mathrm{Cov}\left(\hat{\mu}_{t,t-1}, \hat{\mu}_{t-L,t-L+1}\right) \tag{2}$$

where $\hat{\mu}_{t,t-1}$ is the estimate of the level of AWE at time t, based on the sample matched at times t and t-1, $\hat{\mu}_{t-L,t-L+1}$ is the estimate of the level of AWE at time t-L, based on the sample matched at times t and t-L+1. The formulae used to calculate the variances can be found in annex B.

The square root of the ratio of these two variances, $\sqrt{V_{F,L} / V_{M,L}}$, was calculated as a measure of the advantage, at lag L, of the matched sample estimator relative to the full sample estimator.

2.2 Practical issues - Outliers

Outliers have a huge effect on estimates of variance, therefore the method of identification and treatment of outliers has the potential to significantly alter the results and the conclusions drawn from them. It was therefore necessary to treat outliers for the two estimators in a fair way, so that valid inferences would result from comparisons of the two sets of variance estimates.

For the current AWE it seems reasonable to calculate variances *after* applying the current method for identifying and treating outliers. However, the question of how to identify and treat outliers for a matched sample type AWE is open, since that estimator does not exist. It is certainly not reasonable to assume that the outlier methods used currently on AWE are appropriate.

Consequently, it was necessary to choose to either (a) speculate, and develop some outlier methods for a matched sample type estimator, or (b) not treat outliers in either the current AWE or the matched sample type estimator. Choice (a) was rejected because the effort required to produce a good method would be considerable, and it would not be fair to compare a less than good method with that of the current AWE, whose outlier methods have been demonstrated to be excellent.

Thus, it was decided that no routine outlier identification and treatment would be applied to either the current AWE or the matched sample AWE. The assumption being made when making this choice is that the outlier methods used for a matched sample type AWE would be as effective as the current methods used on AWE. This choice had the added benefit of making our work independent of the work, being carried out in parallel to ours, to investigate possible improvements to the outlier methods for the current AWE (Finselbach et al., 2009).

It was found that there were a small number of firms that were having a large effect on the variances of both the estimators, and making it impossible to make inferences. Some of these firms would have been identified as outliers using the current method, some would not. The history and behaviour of these firms was investigated, and where appropriate their returns were treated. The treatment consisted either of re-labelling the firm identifiers or excluding the firm from the calculations, there were only 29 firms treated in this way.

3. Results

The graph in figure 1 shows the estimated standard error of one month growth in AWE, for both the current and matched sample estimators, for weekly pay excluding bonuses. The standard error is in units of percentage points. Both series are volatile but it can be seen that the variance of the matched pair estimator is typically smaller than that of the current AWE, though there are some exceptions especially towards the end of the time span considered. It is also noticeable that the estimates of variance for the matched sample estimator are less affected by the changes to the AWE sample that occur in December each year. The matched sample estimator is less affected by outliers because some firms with extremes of pay are in the full sample but not the matched sample.

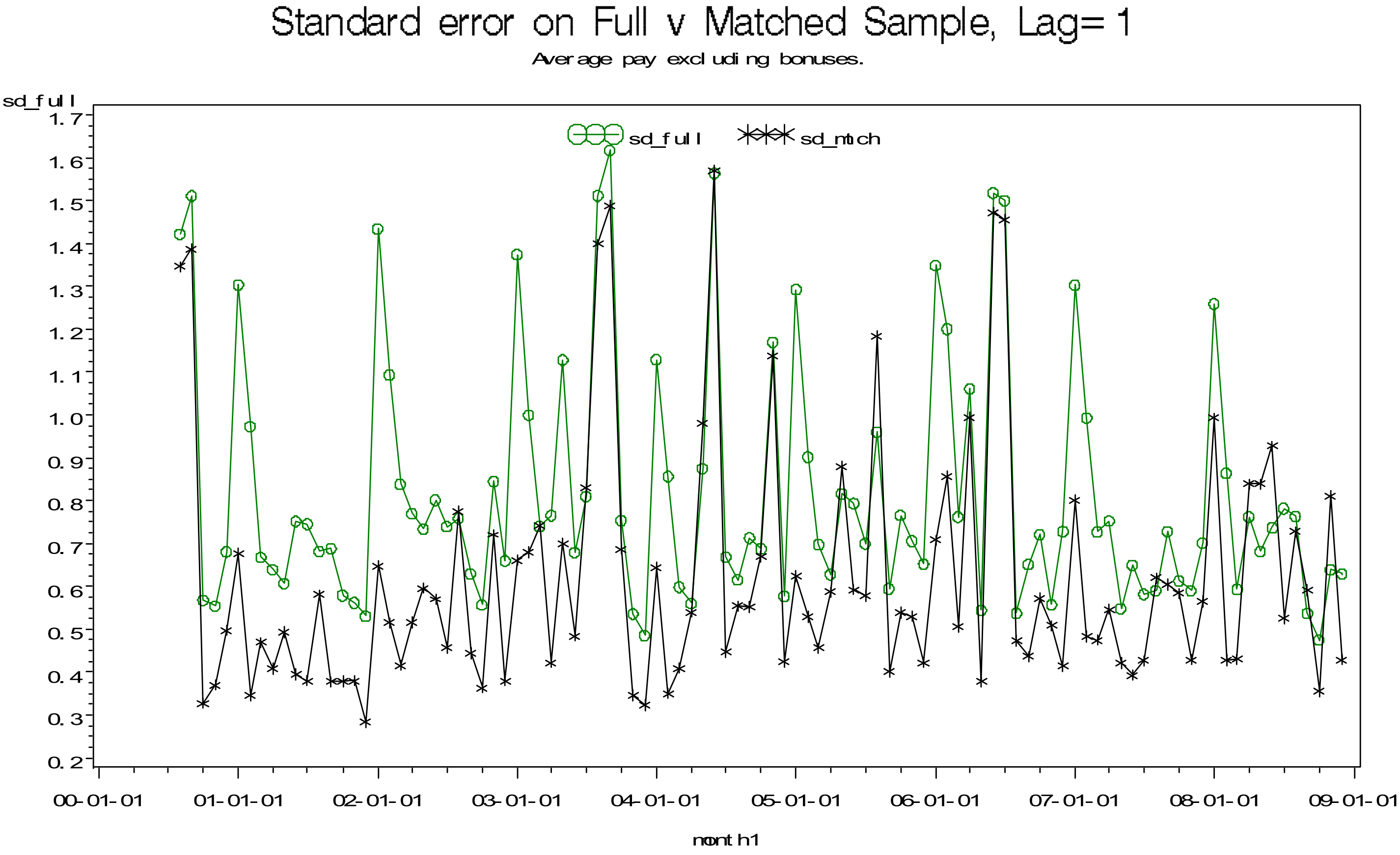

Figure 1

The graph in figure 2 shows the standard error of growth over 3 months (that is lag 3). It is clear that most of any advantage in standard error for the matched sample estimator is gone, and the two series appear remarkably similar.

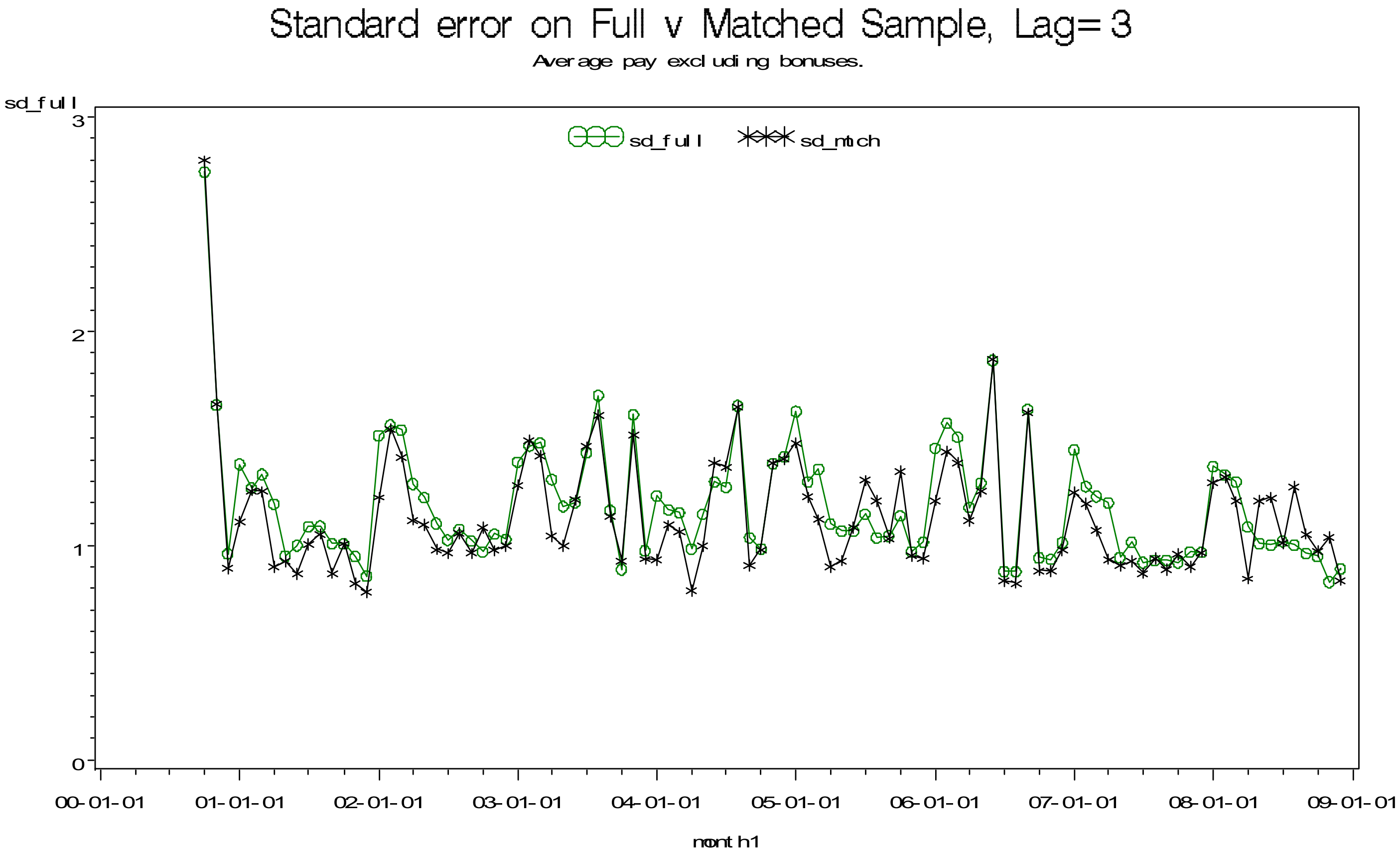

Figure 2

There is another reason for the advantage of the matched sample estimator, apart from the reduced susceptibility to sample changes and outliers (that was illustrated in figure 1.). The standard error of the matched sample estimator is smaller because the estimates in successive months are on average more highly correlated, this can be seen from an examination of figure 3. The graph shows the average, over the months Aug 2000 to Dec 2008, of correlations for the full sample estimates at each lag joined by a red line, and those for the matched sample estimates joined by a blue line. A box and whiskers is also plotted for each lag, showing the upper and lower quartiles, and minimum and maximum, of the correlations.

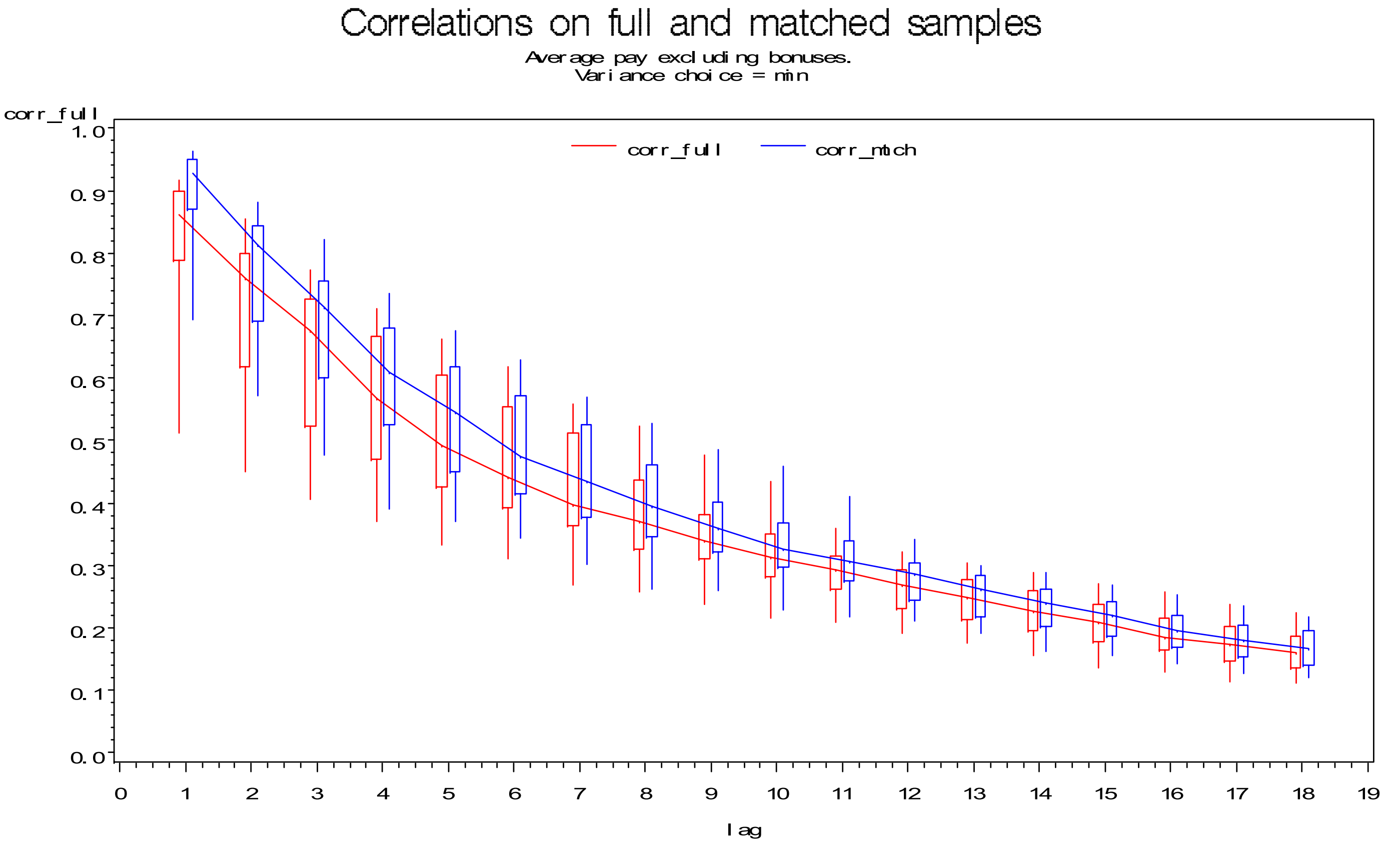

Figure 3 – Correlations between estimates at different lags.

The relative standard error of the two estimators is shown in a different way in figure 4. That plot shows the advantage of the matched sample estimator for growth over each pair of months in the span Aug. 2000 to Dec. 2008, lags 1 to 18. The ratio of standard errors is colour coded, with those most favourable to matched sample estimator being red (larger values of the ratio), and those most favourable to the full sample estimator being purple (lower values of the ratio). (Note that entries on the diagonal, representing growth over 0 months, have been set to missing. Also, the method used to make the plot performs some interpolation between values.) The graph illustrates that the matched sample estimator is better for small lags, shown by the red near the diagonal, and that this advantage is not uniform over the whole period, shown for example by blues near the diagonal around the middle of 2004 and through much of 2008.

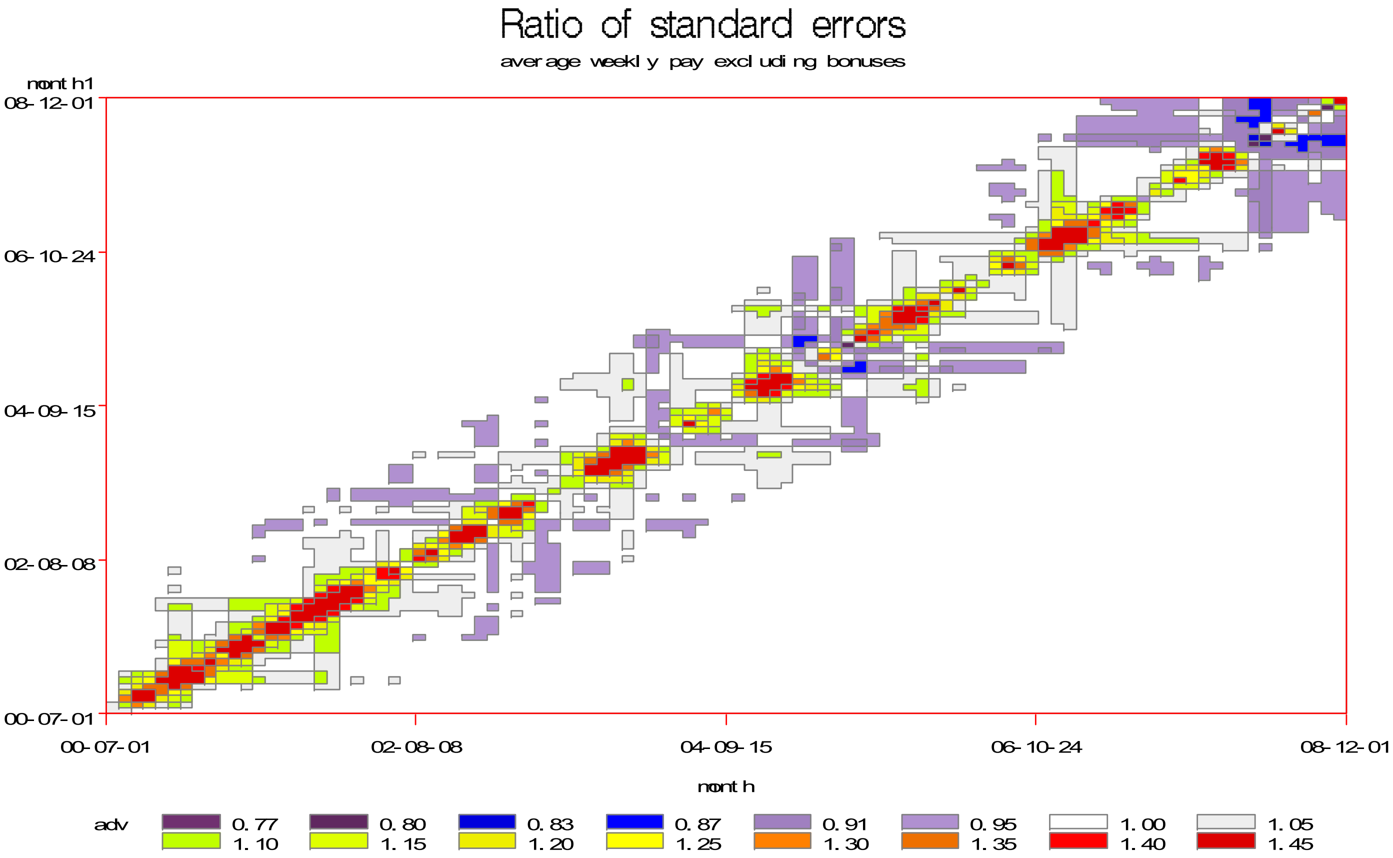

Figure 4 – Smoothed contour plot of the relative advantage of the matched sample estimator to the full sample estimator, calculated up to lag 18.

4. Conclusions

The results presented above somewhat exaggerate the case for the matched sample estimator for AWE. This is for two reasons.

Firstly, there have been theoretical and practical problems to solve in order to complete this work. In order to solve those problems assumptions have been made – and in such a way as to favour the matched sample estimator, or be neutral with respect to both estimators.

Secondly, the bias of the matched sample estimator has not been estimated. It is clear from earlier work on the estimator for retail sales (Kokic and Jones, 1998) that the likely bias will increase the mean square error of the matched sample estimator, and more so at larger lags.

The calculations reported in section 3 demonstrate a small advantage in standard error for the matched sample estimator at lag 1, however this advantage rapidly peters out, and by lag 3 it is tiny. This is the *best* that could be expected from the matched sample estimator, in practice it is bound not to perform so well for the two reasons given above.

The cost of changing from the current system to a matched sample estimator is likely to be large. The costs would include those for:

(1) changing the software used to routinely calculate AWE, together with that used to support the statistician in charge, and training the compilers of these statistics;

(2) developing and implementing a new outlier procedure;

(3) delay in making AWE a national statistic, and thus the need to support two short term measures of earnings.

Furthermore, the work to change to the matched sample estimator would take considerable time to complete.

No formal cost/benefit analysis has been carried out to support the recommendation because it is sufficiently clear that the benefits of a small, and a very short term, gain in efficiency could not outweigh the substantial costs of changing to the matched sample estimator. Furthermore, there is no advantage at lag 12, that is for annual growth, the target of the recommendation of Weale (2008).

5. References

1. Finselbach, H., Merad, S., and Lewis, D. "An Investigation of Outlier Treatment in the AWE", ONS internal report.
2. Kokic, P. and Jones, T. (1998) "Comparing Estimation Methods for a Monthly Business Inquiry", Proceedings of Statistics Canada Symposium 97, 269-272, Statistics Canada, Ottawa.
3. Parkin, N., Šova, M., Wood, J., Lewis, P. A. Untitled ONS internal report.
4. Weale, M. R., 2008 "The Average Earnings Index and Average Weekly Earnings" www.statistics.gov.uk/downloads/theme_labour/Wealefinalreport.pdf
5. Wood, J. (2008) "Matched Pairs versus Estimates of Levels – A Theoretical Analysis", ONS internal note.

Annex A

Extract from Wood (2008).

This document presents a theoretical analysis to compare the relative efficiency of matched sample against the use of level estimates in the estimation of growth rates. Previous empirical work (Kokic & Jones, 1998 and Smith *et al*, 2003) suggests that matched sample is better (that is, has smaller mean squared error) for short-term growth rates but that it suffers from long-term, random drift, becoming progressively worse for the estimation of longer term growth rates and for the estimation of levels. The long-term drift can be corrected by benchmarking to level estimates but this leads to extensive, backdated revisions.

This theoretical work is part of a project to address this question with the aim of deciding on the best formulation for Average Weekly Earnings (AWE), as recommended in Weale (2008):

> ***Recommendation 1. (page 43). The role of matched pairs in dealing with missing observations and sample rotation***
> **AWE should not become a National Statistic until further work has been carried out on the possible use of matched pairs. This work needs to compare the use of matched pairs or a combination of imputation and matched pairs with the existing AWE methodology to see which produces more reliable estimates of annual growth rates. (Recommendation 1).**

The analysis below inevitably includes some simplifications. However, the intention is to produce results that are as general as possible in order to be able to assess the most important influences on the variance of growth rates. The effects of two particular complications that are prevalent in ONS business surveys, namely stratification and finite population corrections, are discussed briefly at various points in the analysis but to assess their actual effects requires examining the effects of the particular sample designs and study populations concerned on the parameters presented in the analysis.

Basic Parameters

We start with the simple situation of a fixed population and a fixed sample (that is, a panel of units which was selected randomly at some time in the past). We wish to estimate the mean μ_t of some response variable for a succession of time periods *t*. For AWE, we are concerned with weekly earnings and each time period is a month. For convenience and to match the context, we shall use month as the period but the analysis applies equally to any other sampling frequency, such as quarter or year, although suitable values for the parameters in the analysis would be different.

In this situation, there is no difference in the sample between matched pairs and level estimation. We assume that the estimation methods used are coherent, in the sense that, for this simple scenario, the matched pairs estimate of ($\mu_t - \mu_{t-1}$) is equal to the difference between the level estimates $\hat{\mu}_t$ and $\hat{\mu}_{t-1}$. Note that we make no assumptions about the actual method of estimation, only that the two methods are coherent in the sense described in the previous sentence, that estimators for sub-populations may be added, with appropriate weighting, to obtain the aggregate estimator $\hat{\mu}_t$ and that is an unbiased estimator of the population mean μ_t. Note also that, for ease of analysis, we shall assess the effect of changes in sampling structure on differences between estimates rather than growth rates. This avoids the complications of adjusting for the normalising denominator in growth rates. It is unlikely that applying the analysis to growth rates would have much impact on the conclusions.

In this simple, basic scenario, we define the following parameters:

$\sigma_t^2 = \text{var}\left[\hat{\mu}_t\right]$: the variance of the level estimate $\hat{\mu}_t$ in month *t*;

$\rho_{st} = \dfrac{\text{cov}\left[\hat{\mu}_s, \hat{\mu}_t\right]}{\sigma_s \sigma_t}$: the coefficient of correlation between level estimates $\hat{\mu}_s$ and $\hat{\mu}_t$ ($s \neq t$).

We may then express the variance of change in terms of the parameters $\left\{\sigma_t^2\right\}$ and $\left\{\rho_{st}\right\}$. For month *t* and lag *l* (*l*=1,2,3,...), consider $\text{var}\left[\hat{\mu}_t - \hat{\mu}_{t-l}\right]$. We can assess this in two ways, on the level estimates themselves:

$$\text{var}\left[\hat{\mu}_t - \hat{\mu}_{t-l}\right] = \text{var}\left[\hat{\mu}_t\right] + \text{var}\left[\hat{\mu}_{t-l}\right] - 2\,\text{co v}\left[\hat{\mu}_t, \hat{\mu}_{t-l}\right] = \sigma_t^2 + \sigma_{t-l}^2 - 2\rho_{t,t-l}\sigma_t\sigma_{t-l}, \qquad (1)$$

or on the succession of monthly changes:

$$\operatorname{var}\left[\hat{\mu}_t - \hat{\mu}_{t-l}\right] = \operatorname{var}\left[\sum_{r=0}^{l-1}\left(\hat{\mu}_{t-r} - \hat{\mu}_{t-r-1}\right)\right] = \sum_{r=0}^{l-1}\sum_{q=0}^{l-1}\operatorname{cov}\left(\hat{\mu}_{t-r} - \hat{\mu}_{t-r-1}, \hat{\mu}_{t-q} - \hat{\mu}_{t-q-1}\right) \quad . \qquad (2)$$

For the constant population and fixed sample in this basic scenario, expressions (1) and (2) are merely different expressions of the same variance of change and are therefore equivalent. In the more general case, with a changing sample, this is not so: the lag 1 differences of estimators in expression (2) are based on the matched samples, not on the full samples. The implications of this are considered in the next section.

Before doing so, it is worth considering the nature of $\rho_{t,t-l}$, which plays a pivotal role in the comparison between matched pairs and level estimates. For a simple random sample, $\rho_{t,t-l}$ is simply the correlation coefficient between responses in months *t* and *t-l* for the common population. For more complicated sample designs, it may be thought of as a weighted average of the corresponding correlation coefficients for the component populations.

To illustrate this, consider a stratified, simple random sample. In this case, we may write $\hat{\mu}_t = \sum_h w_{h:t}\hat{\mu}_{h:t}$, where summation is over all component strata *h* and the $\{w_{ht}\}$ are appropriate population weights with $\sum_h w_{h:t} = 1$. We then have:

$$\rho_{t,t-l}\sigma_t\sigma_{t-l} = \operatorname{co\,v}\left[\hat{\mu}_t, \hat{\mu}_{t-l}\right] = \sum_h w_{h:t}w_{h:t-l}\operatorname{cov}\left[\hat{\mu}_{h:t}, \hat{\mu}_{h:t-l}\right] = \sum_h w_{h:t}w_{h:t-l}\rho_{h:t,t-l}\sigma_{h:t}\sigma_{h:t-l}$$

We may write $w_{h:t}\sigma_{h:t} = \left(1+\varepsilon_{h:t|t-l}\right)w_{h:t-l}\sigma_{h:t-l}$. The $\left\{\varepsilon_{h:t|t-l}\right\}$ should be small because, under normal circumstances, stratum weights and variances should be stable.

It then follows that:

$$\begin{aligned}
\rho_{t,t-l} &= \frac{\sum_h w_{h:t}w_{h:t-l}\rho_{h:t,t-l}\sigma_{h:t}\sigma_{h:t-l}}{\sigma_t\sigma_{t-l}} \\
&= \frac{\sum_h w_{h:t}w_{h:t-l}\rho_{h:t,t-l}\sigma_{h:t}\sigma_{h:t-l}}{\sqrt{\sum_h w_{h:t}^2\sigma_{h:t}^2\sum_g w_{g:t-l}^2\sigma_{g:t-l}^2}} \\
&= \left\{\begin{array}{l}
\dfrac{\sum_h w_{h:t}w_{h:t-l}\rho_{h:t,t-l}\sigma_{h:t}\sigma_{h:t-l}}{\sum_h w_{h:t}w_{h:t-l}\sigma_{h:t}\sigma_{h:t-l}} \\
\times\sqrt{\dfrac{\sum_h w_{h:t}w_{h:t-l}\sigma_{h:t}\sigma_{h:t-l}}{\sum_h w_{h:t}w_{h:t-l}\sigma_{h:t}\sigma_{h:t-l}\left(1+\varepsilon_{h:t|t-l}\right)}\dfrac{\sum_g w_{g:t}w_{g:t-l}\sigma_{g:t}\sigma_{g:t-l}}{\sum_g \dfrac{w_{g:t}w_{g:t-l}\sigma_{g:t}\sigma_{g:t-l}}{\left(1+\varepsilon_{g:t|t-l}\right)}}}
\end{array}\right.
\end{aligned} \qquad (2a)$$

Thus $\rho_{t,t-l}$ is a weighted average of the $\left\{\rho_{h:t,t-l}\right\}$ with a small adjustment expressed in the square root term. The expression within the square root is the ratio between the weighted harmonic mean of the factors $\left\{1+\varepsilon_{h:t|t-l}\right\}$ and the corresponding, weighted arithmetic mean. This ratio depends on the dispersion of the $\left\{\varepsilon_{h:t|t-l}\right\}$ and is always less than or equal to 1: the less the $\left\{\varepsilon_{h:t|t-l}\right\}$ are dispersed, the closer the ratio is to one.

Note that the weights for this weighted average depend on the stratum standard errors as well as the stratum weights. In the usual context of finite population sampling, this means that greater weight would be applied to those strata with small sampling fractions. In the extreme case, zero weight applies to fully enumerated strata. In practice, the precise weighting may not be important because the population correlation coefficients for different strata, over the same time lag, are likely to be similar.

General Results for Matched Pairs

We now consider the effect of changes to the sample on the variance of change. We assume that the sample is subject to depletion by deaths (units leaving the population) or by units being rotated out of the sample and to augmentation by births (units entering the population) or by units being rotated into the sample. The samples in months *t* and *t-l* may also be subject to non-response, which reduces further the number of units that are present in the sample in both month *t* and month *t-l*.

Let:

$\hat{\mu}_{t|t-l}$ be the estimated, mean response in month *t* evaluated using only those sampled units that respond in both month *t* and month *t-l*;

$\hat{\mu}_{t\rangle\langle t-l}$ be the estimated, mean response in month *t* evaluted using those sampled units that respond in month *t* but not in month *t-l* (that is, excluding births and units rotated in between months *t-l* and *t* and newly responding units);

$p_{t\rangle\langle t-l}$ be the weight of component $\hat{\mu}_{t\rangle\langle t-l}$ in the aggregate estimator $\hat{\mu}_t$.

We then have:

$$\hat{\mu}_t = \left(1 - p_{t\rangle\langle t-l}\right)\hat{\mu}_{t|t-l} + p_{t\rangle\langle t-l}\hat{\mu}_{t\rangle\langle t-l} \tag{3}$$

Similarly, let:

$\hat{\mu}_{t-l|t}$ be the estimated, mean response in month *t-l* evaluated using only those sampled units that respond in both month *t* and month *t-l*;

$\hat{\mu}_{t-l\rangle\langle t}$ be the estimated, mean response in month *t-l* evaluated using those sampled units that respond in month *t-l* but not in month *t* (that is, excluding deaths and units rotated out between months *t-l* and *t* and newly non-responding units);

$p_{t-l\rangle\langle t}$ be the weight of component $\hat{\mu}_{t-l\rangle\langle t}$ in the aggregate estimator $\hat{\mu}_{t-l}$.

We then have:

$$\hat{\mu}_{t-l} = \left(1 - p_{t-l\rangle\langle t}\right)\hat{\mu}_{t-l|t} + p_{t-l\rangle\langle t}\hat{\mu}_{t-l\rangle\langle t} \tag{4}$$

Notes on expressions (3) and (4)

We assume that $\hat{\mu}_{t|t-l}, \hat{\mu}_{t-l|t}$ are unbiased estimators of the population means $\mu_{t|t-l}, \mu_{t-l|t}$, for which commonality relates to the population common to months *t* and *t-l*, not only to the common sample.

In general, $\mathrm{E}\left[\hat{\mu}_{t|t-l}\right] = \mu_{t|t-l} \neq \mu_t = \mathrm{E}\left[\hat{\mu}_t\right]$ and $\mathrm{E}\left[\hat{\mu}_{t-l|t}\right] = \mu_{t-l|t} \neq \mu_{t-l} = \mathrm{E}\left[\hat{\mu}_{t-l}\right]$, so the matched pairs estimator of growth

$$\sum_{r=0}^{l-1}\left(\hat{\mu}_{t-r|t-r-1} - \hat{\mu}_{t-r-1|t-r}\right) \tag{5}$$

is usually a biased estimator of the population growth $\left(\mu_t - \mu_{t-l}\right)$. In the analysis below, however, we shall concentrate on variance because of the practical difficulties in estimating the magnitude and direction of any bias.

Because the estimators $\hat{\mu}_{t|t-l}, \hat{\mu}_{t-l|t}$ are based on the same panel of units, they are correlated with correlation coefficient $\rho_{t,t-1}$, as discussed above.

We assume that the expected values of $\hat{\mu}_{t\rangle\langle t-l}, \hat{\mu}_{t-l\rangle\langle t}$ are determined by the relationships:

$$\mu_t = \left(1 - p_{t\rangle\langle t-l}\right)\mu_{t|t-l} + p_{t\rangle\langle t-l}\mu_{t\rangle\langle t-l}$$

$$\mu_{t-l} = \left(1 - p_{t-l\rangle\langle t}\right)\mu_{t-l|t} + p_{t-l\rangle\langle t}\mu_{t-l\rangle\langle t}$$

and their variances and covariances by the relationships:

$$\sigma_t^2 = \operatorname{var}\left[\left(1 - p_{t\rangle\langle t-l}\right)\hat{\mu}_{t|t-l} + p_{t\rangle\langle t-l}\hat{\mu}_{t\rangle\langle t-l}\right]$$

$$\sigma_{t-l}^2 = \operatorname{var}\left[\left(1 - p_{t-l\rangle\langle t}\right)\hat{\mu}_{t-l|t} + p_{t-l\rangle\langle t}\hat{\mu}_{t-l\rangle\langle t}\right]$$

$$\operatorname{cov}\left[\hat{\mu}_t, \hat{\mu}_{t-l}\right] = \operatorname{cov}\left[\left(1 - p_{t\rangle\langle t-l}\right)\hat{\mu}_{t|t-l} + p_{t\rangle\langle t-l}\hat{\mu}_{t\rangle\langle t-l}, \left(1 - p_{t-l\rangle\langle t}\right)\hat{\mu}_{t-l|t} + p_{t-l\rangle\langle t}\hat{\mu}_{t-l\rangle\langle t}\right]$$

In ONS, the inclusion of births and deaths in the sample and the application of rotation is controlled through the use of Permanent Random Numbers. We may also assume that non-response is a random process. So the weights $p_{t\rangle\langle t-l}$ and $p_{t-l\rangle\langle t}$ are not fixed but are random variables. However, the contributions to variance from the randomness of these proportions is proportional to $\left(\mu_{t|t-1} - \mu_{t\rangle\langle t-l}\right)^2$ or $\left(\mu_{t-1|t} - \mu_{t-l\rangle\langle t}\right)^2$. These squared differences of population mean earnings are likely to be much smaller than the variance of individual earnings within each sample, so these contributions to variance should be negligible. We shall therefore assume that the weights $p_{t\rangle\langle t-l}$ and $p_{t-l\rangle\langle t}$ are fixed at the respective, expected values for the relevant populations.

From the matched pairs estimator given in expression (5), we have:

$$\sum_{r=0}^{l-1}\left(\hat{\mu}_{t-r|t-r-1} - \hat{\mu}_{t-r-1|t-r}\right) = \sum_{r=0}^{l-1}\hat{\mu}_{t-r|t-r-1} - \sum_{r=1}^{l}\hat{\mu}_{t-r|t-r+1}$$

$$= \hat{\mu}_{t|t-1} - \hat{\mu}_{t-l|t-l+1} + \sum_{r=1}^{l-1}\left(\hat{\mu}_{t-r|t-r-1} - \hat{\mu}_{t-r|t-r+1}\right)$$

Hence:

$$\operatorname{var}\left[\sum_{r=0}^{l-1}\left(\hat{\mu}_{t-r|t-r-1} - \hat{\mu}_{t-r-1|t-r}\right)\right] = \operatorname{var}\left[\hat{\mu}_{t|t-1} - \hat{\mu}_{t-l|t-l+1} + \sum_{r=1}^{l-1}\left(\hat{\mu}_{t-r|t-r-1} - \hat{\mu}_{t-r|t-r+1}\right)\right] \qquad (6)$$

The term $\sum_{r=1}^{l-1}\left(\hat{\mu}_{t-r|t-r-1} - \hat{\mu}_{t-r|t-r+1}\right)$ is the sum of the differences between estimates of the same target mean from adjacent matched samples. These differences are determined by the occurrence of births, deaths, rotation and non-response, which arise randomly or are assumed to do so. Because of the large overlap between adjacent matched samples, the variance of these differences should be small relative to the variance of the aggregate change and, because changes in the sample arise, or are assumed to arise, independently of changes in the response values for the population, the covariance with the aggregate change should be close to zero. This gives:

$$\operatorname{var}\left[\sum_{r=0}^{l-1}\left(\hat{\mu}_{t-r|t-r-1} - \hat{\mu}_{t-r-1|t-r}\right)\right] >\approx \operatorname{var}\left[\hat{\mu}_{t|t-1} - \hat{\mu}_{t-l|t-l+1}\right] \qquad (7)$$

The relative insignificance of the term $\sum_{r=1}^{l-1}\left(\hat{\mu}_{t-r|t-r-1}-\hat{\mu}_{t-r|t-r+1}\right)$ is more easily seen by using equations (3) and (4) to clarify the nature of these differences in estimates. From equation (3), we have:

$$\hat{\mu}_{t|t-l}=\frac{\hat{\mu}_t}{\left(1-p_{t\rangle\langle t-l}\right)}-\frac{p_{t\rangle\langle t-l}\hat{\mu}_{t\rangle\langle t-l}}{\left(1-p_{t\rangle\langle t-l}\right)} \qquad (8)$$

and from equation (4), we have:

$$\hat{\mu}_{t-l|t}=\frac{\hat{\mu}_{t-l}}{\left(1-p_{t-l\rangle\langle t}\right)}-\frac{p_{t-l\rangle\langle t}\hat{\mu}_{t-l\rangle\langle t}}{\left(1-p_{t-l\rangle\langle t}\right)} \qquad (9)$$

Setting *l*=1 in equations (8) and (9) and applying them to the term $\sum_{r=1}^{l-1}\left(\hat{\mu}_{t-r|t-r-1}-\hat{\mu}_{t-r|t-r+1}\right)$ then gives:

$$\begin{aligned}\sum_{r=1}^{l-1}\left(\hat{\mu}_{t-r|t-r-1}-\hat{\mu}_{t-r|t-r+1}\right)&=\sum_{r=1}^{l-1}\left\{\begin{array}{l}\dfrac{\hat{\mu}_{t-r}}{\left(1-p_{t-r\rangle\langle t-r-1}\right)}-\dfrac{p_{t-r\rangle\langle t-r-1}\hat{\mu}_{t-r\rangle\langle t-r-1}}{\left(1-p_{t-r\rangle\langle t-r-1}\right)}\\-\dfrac{\hat{\mu}_{t-r}}{\left(1-p_{t-r\rangle\langle t-r+1}\right)}+\dfrac{p_{t-r\rangle\langle t-r+1}\hat{\mu}_{t-r\rangle\langle t-r+1}}{\left(1-p_{t-r\rangle\langle t-r+1}\right)}\end{array}\right\}\\&=\sum_{r=1}^{l-1}\left\{\begin{array}{l}\dfrac{\left(p_{t-r\rangle\langle t-r-1}-p_{t-r\rangle\langle t-r+1}\right)\hat{\mu}_{t-r}}{\left(1-p_{t-r\rangle\langle t-r-1}\right)\left(1-p_{t-r\rangle\langle t-r+1}\right)}\\-\left[\dfrac{p_{t-r\rangle\langle t-r-1}\hat{\mu}_{t-r\rangle\langle t-r-1}}{\left(1-p_{t-r\rangle\langle t-r-1}\right)}-\dfrac{p_{t-r\rangle\langle t-r+1}\hat{\mu}_{t-r\rangle\langle t-r+1}}{\left(1-p_{t-r\rangle\langle t-r+1}\right)}\right]\end{array}\right\}\end{aligned} \qquad (10)$$

The weights $\left\{p_{t-r\rangle\langle t-r-1},p_{t-r\rangle\langle t-r+1}\right\}$ represent only one month's change in the sample, so the coefficients of the $\left\{\hat{\mu}_{t-r}\right\}$, based on the differences between these small weights, are very small. The terms in $\left\{\hat{\mu}_{t-r}\right\}$ are also directly offset by similarly sized terms in $\left\{\hat{\mu}_{t-r\rangle\langle t-r-1} \text{ and } \hat{\mu}_{t-r\rangle\langle t-r+1}\right\}$, further reducing the impact of expression (10). Expression (10) matters only if there are consistent differences between $p_{t-r\rangle\langle t-r-1}$ and $p_{t-r\rangle\langle t-r+1}$, implying an expanding or contracting sample or population, or between $\hat{\mu}_{t-r\rangle\langle t-r-1}$ and $\hat{\mu}_{t-r\rangle\langle t-r+1}$, implying that the response values for new units to the sample have different means from the response values for units leaving the sample (this is likely to be important for the relatively small proportions of births and deaths in the population).

As noted above, the effect of expression (10) is to add to the variance of the matched pairs estimator. So use of approximation (7) understates the variance of the matched pairs estimator and is therefore biased in favour of matched pairs.

Using approximation (7), we therefore have the following condition for the matched pairs estimator of growth to have a smaller variance than the estimator based on levels:

$$\begin{aligned}
&\operatorname{var}\left[\sum_{r=0}^{l-1}\left(\hat{\mu}_{t-r|t-r-1}-\hat{\mu}_{t-r-1|t-r}\right)\right]<\operatorname{var}\left[\hat{\mu}_t-\hat{\mu}_{t-l}\right] \\
&\Rightarrow \operatorname{var}\left[\hat{\mu}_{t|t-1}-\hat{\mu}_{t-l|t-l+1}\right]<\operatorname{var}\left[\hat{\mu}_t-\hat{\mu}_{t-l}\right] \\
&\Rightarrow \begin{cases}\operatorname{var}\left[\hat{\mu}_{t|t-1}\right]+\operatorname{var}\left[\hat{\mu}_{t-l|t-l+1}\right]-2\operatorname{cov}\left[\hat{\mu}_{t|t-1},\hat{\mu}_{t-l|t-l+1}\right] \\ <\operatorname{var}\left[\hat{\mu}_t\right]+\operatorname{var}\left[\hat{\mu}_{t-l}\right]-2\operatorname{cov}\left[\hat{\mu}_t,\hat{\mu}_{t-l}\right]\end{cases} \\
&\Rightarrow \begin{cases}\operatorname{cov}\left[\hat{\mu}_{t|t-1},\hat{\mu}_{t-l|t-l+1}\right]-\operatorname{cov}\left[\hat{\mu}_t,\hat{\mu}_{t-l}\right] \\ >\frac{1}{2}\left(\operatorname{var}\left[\hat{\mu}_{t|t-1}\right]-\operatorname{var}\left[\hat{\mu}_t\right]+\operatorname{var}\left[\hat{\mu}_{t-l|t-l+1}\right]-\operatorname{var}\left[\hat{\mu}_{t-l}\right]\right)\end{cases}
\end{aligned} \qquad (11)$$

That is, for matched pairs to have the lower variance, the covariance between estimators of levels for the initial and final matched pairs samples needs to exceed the covariance between estimators of levels for the initial and final full samples by more than the mean excess in the corresponding variances.

Annex B - Calculation of Covariances for AWE

Definitions

Time s is assumed to be before, or the same as, time t.

U_{hs} is the population at time s in stratum h,

U_{ht} is the population at time t in stratum h,

S_{hs} is the sample at time s in stratum h,

S_{ht} is the sample at time t in stratum h,

$\overline{S}_{hs}$ is the complement of the sample in the population at time s in stratum h, $\overline{S}_{hs} = U_{hs} - S_{hs}$,

$\overline{S}_{ht}$ is the complement of the sample in the population at time t in stratum h, $\overline{S}_{ht} = U_{ht} - S_{ht}$,

M_{hst} is the set of firms in the sample at time s and time t in stratum h,

n_{hst} is the size of the matched sample M_{hst},

RO_{hst} is the set of those firms in the sample at time s, in the population at time t but not the sample at time t, that is those firms that have rotated out of the sample,

RI_{hst} is the set of those firms in the sample at time t, in the population at time s but not in the sample at time s, that is those firms that have rotated into the sample,

N_{hst} is the set of those firms in the common population at times s and t, but in neither sample, that is

$N_{hst} = (U_{hs} \cap U_{ht}) - (S_{hs} \cup S_{ht})$,

the remaining definitions apply to both times s and t, only those for time s are given.

μ_s the population value for average pay, at the all industry level.

μ_{hs} the population value for average pay for stratum h.

$\hat{\mu}_s$ the estimate for the population value of average pay, at the all industry level.

$\hat{\mu}_{hs}$ the estimate for the population value of average pay for stratum h.

r_{hsi} is the average earnings for firm i, at time s, in stratum h,

$\overline{r}_{hs(t)} = \frac{1}{n_{hst}} \sum_{i \in M_{hst}} r_{hsi}$ and $\overline{r}_{ht(s)} = \frac{1}{n_{hst}} \sum_{i \in M_{hst}} r_{hti}$, are the averages of the average earnings at time s and t on the matched sample,

x_{hsi} is the returned employment for firm i, at time s, in stratum h,

ψ_{hsi} is the weight for the average pay for firm i, at time s, in stratum h, referred to the sample S_s,

$$\psi_{hst} = \frac{x_{hsi}}{\sum_{i \in S_s} x_{hsi}}.$$

χ_{hsi} is the weight for the average pay for firm i, at time s, in stratum h, referred to the population U_s,

$$\chi_{hst} = \frac{x_{hsi}}{\sum_{i \in U_s} x_{hsi}}.$$

$c_{hst} = \mathrm{Cov}(r_{hsi}, r_{hti}) \quad \forall i \in U_s$ is the proposed model variance. Note that $\mathrm{Cov}(r_{hsi}, r_{htj}) = 0 \quad \text{when } i \neq j$.

$\hat{c}_{hst} = \frac{1}{n_{hst}-1}\sum_{i\in M_{hst}}\left(r_{hsi}-\overline{r}_{hs(t)}\right)\left(r_{hti}-\overline{r}_{ht(s)}\right)$ is the estimated value of c_{hst}.

Calculations

The population value, μ_s, for average earnings at time s is given by

$$\begin{aligned}\mu_s &= \frac{\sum_h \sum_{i\in U_{hs}} y_{hsi}}{\sum_h \sum_{i\in U_{hs}} x_{hsi}} \\ &= \sum_h v_{hs} \sum_{i\in U_{hs}} \chi_{hsi} r_{hsi}\end{aligned},$$

where $\chi_{hsi} = \frac{x_{hsi}}{\sum_{i\in U_{hs}} x_{hsi}}$ and $v_{hs} = \frac{\sum_{i\in U_{hs}} x_{hsi}}{\sum_h \sum_{i\in U_{hs}} x_{hsi}}$, with a similar expression for μ_t. Also, the estimate, $\hat{\mu}_s$, for the population value at time s is given by

$$\begin{aligned}\hat{\mu}_s &= \frac{\sum_h \sum_{i\in S_{hs}} \phi_{hs} y_{hsi}}{\sum_h \sum_{i\in S_{hs}} \phi_{hs} x_{hsi}} \\ &= \sum_h w_{hs} \sum_{i\in S_{hs}} \psi_{hsi} r_{hsi}\end{aligned},$$

where $\psi_{hsi} = \frac{\phi_{hs} x_{hsi}}{\sum_{i\in S_{hs}} \phi_{hs} x_{hsi}} = \frac{x_{hsi}}{\sum_{i\in S_{hs}} x_{hsi}}$ and $w_{hs} = \frac{\phi_{hs}\sum_{i\in S_{hs}} x_{hsi}}{\sum_h \phi_{hs} \sum_{i\in S_{hs}} x_{hsi}}$, with a similar expression for $\hat{\mu}_t$.

ϕ_{hs} is the design weight for stratum *h* at time *s*: $\phi_{hs} = \frac{\sum_{i\in U_{hs}} z_{hsi}}{\sum_{i\in S_{hs}} z_{hsi}} = \frac{1}{f_{hs}}$,

where z_{hsi} is the register employment for firm *i* at time *s* in stratum *h.*

The proposal is to calculate the covariances $K_{st} = \mathrm{Cov}\left(\hat{\mu}_s, \hat{\mu}_t\right)$ in the following way

$$K_{st} = \sum_h w_{hs} w_{ht}\,\mathrm{Cov}\left(\hat{\mu}_{hs}-\mu_{hs}, \hat{\mu}_{ht}-\mu_{ht}\right).$$

Thus, it is necessary to calculate $K_{hst} = \mathrm{Cov}\left(\hat{\mu}_{hs}-\mu_{hs}, \hat{\mu}_{ht}-\mu_{ht}\right)$. Note that

$$\begin{aligned}\hat{\mu}_{hs}-\mu_{hs}&=\sum_{i\in S_{hs}}\psi_{hsi}r_{hsi}-\sum_{i\in U_{hs}}\chi_{hsi}r_{hsi}\\&=\sum_{i\in S_{hs}}\left(\psi_{hsi}-\chi_{hsi}\right)r_{hsi}+\sum_{i\in\bar{S}_{hs}}\chi_{hsi}r_{hsi}\\&=\sum_{i\in M_{hst}}\left(\psi_{hsi}-\chi_{hsi}\right)r_{hsi}+\sum_{i\in S_{hs}-M_{hst}}\left(\psi_{hsi}-\chi_{hsi}\right)r_{hsi}+\sum_{i\in\bar{S}_{hs}}\chi_{hsi}r_{hsi}\\&=A_{hs}+B_{hs}+C_{hs}\end{aligned},$$

and similarly,

$$\hat{\mu}_{ht}-\mu_{ht}=A_{ht}+B_{ht}+C_{ht}.$$

There are no firms in common between the sets: M_{hst} and $S_{ht}-M_{hst}$; M_{hst} and $\bar{S}_{ht}$; M_{hst} and $S_{hs}-M_{hst}$; M_{hst} and $\bar{S}_{hs}$; or $S_{hs}-M_{hst}$ and $S_{ht}-M_{hst}$, hence

$$\begin{aligned}K_{hst}&=\operatorname{Cov}\left(A_{hs},A_{ht}\right)+\operatorname{Cov}\left(B_{hs},C_{ht}\right)+\operatorname{Cov}\left(C_{hs},B_{ht}\right)+\operatorname{Cov}\left(C_{hs},C_{ht}\right)\\&=W_h+X_h+Y_h+Z_h\end{aligned}.$$

Now,

$$\begin{aligned}W_h&=\operatorname{Cov}\left(\sum_{i\in M_{hst}}\left(\psi_{hsi}-\chi_{hsi}\right)r_{hsi},\sum_{j\in M_{hst}}\left(\psi_{htj}-\chi_{htj}\right)r_{htj}\right)\\&=\sum_{i\in M_{hst}}\left(\psi_{hsi}-\chi_{hsi}\right)\left(\psi_{hti}-\chi_{hti}\right)c_{hst}\end{aligned},$$

$$\begin{aligned}X_h&=\operatorname{Cov}\left(\sum_{i\in S_{hs}-M_{hst}}\left(\psi_{hsi}-\chi_{hsi}\right)r_{hsi},\sum_{j\in\bar{S}_{hst}}\chi_{htj}r_{htj}\right)\\&=\sum_{i\in RO_{hst}}\chi_{hti}\left(\psi_{hsi}-\chi_{hsi}\right)c_{hst}\end{aligned},$$

$$\begin{aligned}Y_h&=\operatorname{Cov}\left(\sum_{i\in\bar{S}_{hst}}\chi_{hsi}r_{hsi},\sum_{j\in S_{hs}-M_{hst}}\left(\psi_{htj}-\chi_{htj}\right)r_{htj}\right)\\&=\sum_{i\in RI_{hst}}\chi_{hsi}\left(\psi_{hti}-\chi_{hti}\right)c_{hst}\end{aligned}, \text{ and}$$

$$\begin{aligned}Z_h&=\operatorname{Cov}\left(\sum_{i\in\bar{S}_{hs}}\chi_{hsi}r_{hsi},\sum_{j\in\bar{S}_{ht}}\chi_{htj}r_{htj}\right)\\&=\sum_{i\in N_{hst}}\chi_{hsi}\chi_{hti}c_{hst}\end{aligned}.$$

Therefore,

$$K_{hst}=c_{hst}\left(\sum_{i\in M_{hst}}\left\{\psi_{hsi}-\chi_{hsi}\right\}\left\{\psi_{hti}-\chi_{hti}\right\}+\sum_{i\in RO_{hst}}\chi_{hti}\left\{\psi_{hsi}-\chi_{hsi}\right\}+\sum_{i\in RI_{hst}}\chi_{hsi}\left\{\psi_{hti}-\chi_{hti}\right\}+\sum_{i\in N_{hst}}\chi_{hsi}\chi_{hti}\right).$$

Note that,

$\psi_{hsi}-\chi_{hsi}=x_{hsi}\dfrac{\sum_{i\in U_{hs}}x_{hsi}-\sum_{i\in S_{hs}}x_{hsi}}{\sum_{i\in U_{hs}}x_{hsi}\sum_{i\in S_{hs}}x_{hsi}}$, let $X_{U_{hs}}=\sum_{i\in U_{hs}}x_{hsi}$, and $X_{S_{hs}}=\sum_{i\in S_{hs}}x_{hsi}$, then

$\psi_{hsi} - \chi_{hsi} = x_{hsi} \dfrac{1}{X_{U_{hs}}} \left(\dfrac{X_{U_{hs}}}{X_{S_{hs}}} - 1 \right)$, and similarly $\psi_{hti} - \chi_{hti} = x_{hti} \dfrac{1}{X_{U_{ht}}} \left(\dfrac{X_{U_{ht}}}{X_{S_{ht}}} - 1 \right)$, and it follows that,

$$\begin{aligned} K_{hst} &= c_{hst} \frac{1}{X_{U_{hs}} X_{U_{ht}}} \left(\left\{ \frac{X_{U_{hs}}}{X_{S_{hs}}} - 1 \right\} \left\{ \frac{X_{U_{ht}}}{X_{S_{ht}}} - 1 \right\} \sum_{i \in M_{hst}} x_{hsi} x_{hti} + \left\{ \frac{X_{U_{hs}}}{X_{S_{hs}}} - 1 \right\} \sum_{i \in RO_{hst}} x_{hsi} x_{hti} + \left\{ \frac{X_{U_{ht}}}{X_{S_{ht}}} - 1 \right\} \sum_{i \in RI_{hst}} x_{hsi} x_{hti} + \sum_{i \in N_{hst}} x_{hsi} x_{hti} \right) \\ &= c_{hst} \left\{ \left(1 - \frac{X_{S_{hs}}}{X_{U_{hs}}} \right) \left(1 - \frac{X_{S_{ht}}}{X_{U_{ht}}} \right) \frac{\sum_{i \in M_{hst}} x_{hsi} x_{hti}}{X_{S_{hs}} X_{S_{ht}}} + \left(1 - \frac{X_{S_{hs}}}{X_{U_{hs}}} \right) \frac{\sum_{i \in RO_{hst}} x_{hsi} x_{hti}}{X_{S_{hs}} X_{U_{ht}}} + \left(1 - \frac{X_{S_{ht}}}{X_{U_{ht}}} \right) \frac{\sum_{i \in RI_{hst}} x_{hsi} x_{hti}}{X_{U_{hs}} X_{S_{ht}}} + \frac{\sum_{i \in N_{hst}} x_{hsi} x_{hti}}{X_{U_{hs}} X_{U_{ht}}} \right\} \end{aligned}.$$

In order to estimate K_{hst} we propose to substitute register employment for returned employment in those ratios involving unknown population totals, so that

$$\begin{aligned} \hat{K}_{hst} &= \frac{1}{n_{hst} - 1} \sum_{i \in M_{hst}} \left(r_{hsi} - \bar{r}_{hs(t)} \right) \left(r_{hti} - \bar{r}_{ht(s)} \right) \left\{ \left(1 - \frac{Z_{S_{hs}}}{Z_{U_{hs}}} \right) \left(1 - \frac{Z_{S_{ht}}}{Z_{U_{ht}}} \right) \frac{\sum_{i \in M_{hst}} x_{hsi} x_{hti}}{X_{S_{hs}} X_{S_{ht}}} + \left(1 - \frac{Z_{S_{hs}}}{Z_{U_{hs}}} \right) \frac{\sum_{i \in RO_{hst}} x_{hsi} z_{hti}}{X_{S_{hs}} Z_{U_{ht}}} + \left(1 - \frac{Z_{S_{ht}}}{Z_{U_{ht}}} \right) \frac{\sum_{i \in RI_{hst}} z_{hsi} x_{hti}}{Z_{U_{hs}} X_{S_{ht}}} + \frac{\sum_{i \in N_{hst}} z_{hsi} z_{hti}}{Z_{U_{hs}} Z_{U_{ht}}} \right\} \\ &= \frac{1}{n_{hst} - 1} \sum_{i \in M_{hst}} \left(r_{hsi} - \bar{r}_{hs(t)} \right) \left(r_{hti} - \bar{r}_{ht(s)} \right) \left\{ \left(1 - f_{hs} \right) \left(1 - f_{ht} \right) \frac{\sum_{i \in M_{hst}} x_{hsi} x_{hti}}{X_{S_{hs}} X_{S_{ht}}} + \left(1 - f_{hs} \right) \frac{\sum_{i \in RO_{hst}} x_{hsi} z_{hti}}{X_{S_{hs}} Z_{U_{ht}}} + \left(1 - f_{ht} \right) \frac{\sum_{i \in RI_{hst}} z_{hsi} x_{hti}}{Z_{U_{hs}} X_{S_{ht}}} + \frac{\sum_{i \in N_{hst}} z_{hsi} z_{hti}}{Z_{U_{hs}} Z_{U_{ht}}} \right\} \end{aligned}$$

Note that the proposal is to use this formula to estimate variances $\hat{K}_{hs} = \hat{K}_{hss}$ also, in which case the formula reduces to

$$\hat{K}_{hs} = \frac{1}{n_{hs} - 1} \sum_{i \in S_{hs}} \left(r_{hsi} - \bar{r}_{hs} \right)^2 \left\{ \left(1 - f_{hs} \right)^2 \frac{\sum_{i \in S_{hs}} x_{hsi}^2}{X_{S_{hs}}^2} + \frac{\sum_{i \in N_{hs}} z_{hsi}^2}{Z_{U_{hs}}^2} \right\}.$$