Correcting for Misclassification Error in Gross Flows Using Double Sampling: Moment-based Inference vs. Likelihood-based Inference

Nikos Tzavidis

Abstract

Gross flows are discrete longitudinal data that are defined as transition counts, between a finite number of states, from one point in time to another. We discuss the analysis of gross flows in the presence of misclassification error via double sampling methods. Traditionally, adjusted for misclassification error estimates are obtained using a moment-based estimator. We propose a likelihood-based approach that works by simultaneously modeling the true transition process and the misclassification error process within the context of a missing data problem. Monte-Carlo simulation results indicate that the maximumlikelihood estimator is more efficient than the moment-based estimator.

S^{3}RI Methodology Working Paper M04/11

Correcting for Misclassification Error in Gross Flows Using Double Sampling: Moment-based Inference vs. Likelihood-based Inference

Nikos Tzavidis[1]

Abstract

Gross flows are discrete panel data that are generally defined as transition counts, between a finite number of states, from one point in time to another. Gross flows are typically estimated by linking panel data from consecutive waves. This process, however, is affected by the existence of non-sampling errors such as response errors that cause misclassification error. We discuss alternative approaches for correcting for misclassification error in gross flows via double sampling. Traditionally, in a double sampling context, adjusted for misclassification error estimates are obtained using a moment-based estimator. We propose a likelihood-based approach that works by simultaneously modeling the true transition process and the misclassification error process within the context of a missing data problem. The model is formulated under alternative double sampling designs and maximum likelihood estimates are derived by maximizing the likelihood of the augmented data via the EM algorithm. The issue of variance estimation for the adjusted estimates is resolved using Taylor series linearization and the Missing Information Principle. Monte-Carlo simulation results indicate that the maximum likelihood estimator is more efficient than the moment-based estimator while a real data application indicates that the maximum likelihood estimator has desirable numerical properties that are appealing to the data analyst.

KEYWORDS: Nonsampling errors; Response bias; Panel surveys; Re-interview surveys; Missing data; Labour force gross flows

1. Introduction

Gross flows are defined as transition counts, between a finite number of states, from one point in time to another. Typical examples of gross flows are labour force gross flows that represent transition counts of the labour force population between the different labour force states. Gross flows estimates are frequently derived from panel surveys by linking panel data from consecutive

[1] Nikos Tzavidis, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield Campus, Southampton, SO17 1BJ, UK

waves. This process, however, is affected by non-sampling errors such as response errors that cause misclassification error (Hogue and Flaim 1986; Kristiansson 1999).

The existence of misclassification error in data used for statistical analysis can introduce serious bias in the derived results. Methods that account for the existence of misclassification error have received great attention in the statistical literature. In the presence of misclassification error, such methods need to be employed in order to ensure the validity of the inferential process. One of the traditional approaches for adjusting for misclassification error in discrete data, such as gross flows, is by assuming the existence of validation information derived from a validation survey, which is free of error. The use of validation surveys can be placed into the framework of double sampling methods (Bross 1954). In a double sampling framework we assume that along with the main measurement device, which is affected by misclassification error, we have a secondary measurement device (validation survey), which is free of error but more expensive to apply. Due to its higher cost, the validation survey is employed only for a subset of sampling units. Inference using double sampling is based on combining information from both measurement devices. Other approaches to misclassification error correction include latent class models (Van de Pol and De Leeuw 1986) and instrumental variables models (Skinner and Humphreys 1997). However, in this paper, we will focus on the case that validation data, obtained via double sampling, are available.

We examine alternative approaches for correcting for misclassification error in gross flows when validation information is available. However, while the main measurement device is a panel survey, we allow only for cross-sectional validation data. This choice can be justified given the costs associated with conducting a validation survey. We propose a maximum likelihood estimator as an alternative to the traditional moment-based estimator. We show that in contrast to the moment-based estimator, the maximum likelihood estimator provides more efficient adjusted estimates and has numerical properties that are appealing to the data analyst.

The structure of the paper is as follows: In Section 2 we present the estimation framework of double sampling methods. We further present different double sampling designs that can be used

with panel data along with alternative specifications for quantifying the misclassification error mechanism. In Section 3 we describe two alternative approaches for misclassification error correction via double sampling. The first approach is an existing one that leads to a moment-based estimator. The second approach is what we propose and leads to a maximum likelihood estimator. Both approaches are investigated under alternative double sampling designs. In Section 4 we discuss variance estimation for the moment-based and the maximum likelihood estimators. In Section 5 a series of Monte-Carlo simulation studies are designed for empirically comparing the alternative point and variance estimators while in Section 6 the methodology is illustrated in the context of the US Current Population Survey (CPS) by estimating labour force gross flows adjusted for misclassification error. Finally, in Section 7 we summarize the main findings and provide directions for further research.

2. Using Double Sampling for Misclassification Error Correction

Suppose that the standard measurement device is subject to misclassification error. As a result we have biased results. Unbiased estimates can be obtained by utilizing more elaborate measurement tools usually referred to as preferred procedures (Forsman and Schreiner 1991; Kuha and Skinner 1997). An example of a preferred procedure is re-interview surveys (Bailar 1968). In bio-statistical applications the term “gold standard” is more commonly used (Bauman and Koch 1983). Other examples include judgments of experts or checks against administrative records (Greenland 1988). The assumption that the preferred procedure is free of error makes possible the estimation of the parameters of the misclassification error mechanism. On the other hand, the preferred procedures are considered to be fairly expensive and thus unsuitable to be used for the entire sample. Therefore, these procedures are normally applied to a smaller sample usually referred to as validation sample.

The validation sample can be either internal or external. Kuha and Skinner (1997) make this distinction following literature on misclassification error in medical applications (Greenland 1988). The characteristic that distinguishes an internal validation sample from an external

validation sample is whether the fallible classifications from the validation sample can be combined with the fallible classifications from the main sample. A validation sample is defined as internal if it is a sub-sample of n^v units from the main sample of n units obtained via a randomised double sampling design. Alternatively, the validation sample is defined as internal if it is selected independently from the main sample and from the same target population. Otherwise, the validation sample is characterised as external. The parameters of the misclassification error mechanism estimated from an external validation sample are assumed to be representative of the misclassification process in the target population but the fallible classifications from this validation sample cannot be combined with the fallible classifications from the main sample.

Initially, double sampling methods were developed to adjust cross-sectional data for misclassification error. In this context, Bross (1954) described the general framework of double sampling methods. Maximum likelihood adjusted estimates for binomial and multinomial data were derived by Tenenbein (1970;1972) respectively. These results were then extended for fitting log-linear models in the presence of misclassification error (Espeland and Odoroff 1985). It is believed that for cross-sectional data there is no particular tendency for errors to be systematic (Skinner 2000). However, for panel data produced by linking information on the same individual in different time points, this cancellation may not occur. Work on the adjustment of gross flows for misclassification error via the use of double sampling includes Abowd and Zellner (1985), Poterba and Summers (1986), Skinner and Torelli (1993), and Singh and Rao (1995).

2.1 Double Sampling Designs for Panel Data Analysis

In this section we examine three alternative double sampling designs that can be used with panel misclassified data when only cross-sectional validation data are available.

Double Sampling Design 1

A simple random sample of n units is selected from a population of N units and the classifications for these n units are obtained at time t and $t+1$ using a standard measurement device, which is affected by misclassification error. At a second time point, between t and $t+1$,

a sub-sample of n^v units is selected from the n units that already belong to the main sample and their classifications by the standard measurement device at time t are validated using more elaborate survey techniques.

Double Sampling Design 2

A simple random sample of n units is selected from a population of N units and the classifications for these n units are obtained at time t and $t+1$ using the standard measurement device, which is affected by misclassification error. For another simple random sample of n^v units, independently selected from the main sample and from the same target population, classifications are obtained only at time t using also the standard measurement device. At a second time point, between t and $t+1$, the classifications of the n^v units obtained by the standard measurement device are validated using more elaborate survey techniques.

Double Sampling Design 3

A simple random sample of n units is selected from a population of N units and the classifications for these n units are obtained at time t and $t+1$ using a standard measurement device, which is affected by misclassification error. Using an external source of information, we then obtain cross-sectional information on the incidence of error. The assumption underpinning this design is that the external source of information adequately describes the misclassification process in the target population.

Traditionally, double sampling methods are associated with double sampling design 1. However, when dealing with panel data assuming that only cross-sectional data are available, distinguishing between the different designs is important. Although under design 2 the validation and the main samples are representative of the same population, information on the fallible classifications from the validation sample cannot be combined with information on the fallible classifications from the main sample. This is because the validation sample is cross-sectional and not panel. Thus, the validation sample can be used only for making inferences about the cross-sectional incidence of misclassification error. The same is also true for the other two double

sampling designs. The different double sampling designs have also different costs. Under design 2 we conduct the main survey using n units and the validation survey using n^v different units. Therefore, under this design we have cross-sectional information on the observed classifications for $n + n^v$ units. On the other hand, under design 1 we have cross-sectional information on the observed classifications only for n units. This implies that design 2 may be associated with an increased cost compared to design 1. In this discussion, however, we need to consider one of the main disadvantages associated with design 1. Under this design, the sample units that participate in the validation survey participate also in the main, panel, survey. One may argue that this is similar to adding an extra wave to the panel survey, which effectively may increase the response burden of the respondents and therefore impact on the quality of the collected data.

2.2 Quantifying the Misclassification Error Mechanism

Assume that a sample of n units has been selected via a randomised design from a population of N units and let ξ denote a member of this sample. Let us further assume that the variable of interest, measured by the survey, is subject to misclassification error and that a validation survey is used for identifying the true values for a subset of sample units. Define the random variables $Y^*_{\xi t}, Y_{\xi t}$ that respectively describe the observed and true classifications for the ξ^{th} sample unit at time t. One way to quantify the misclassification error mechanism is via the use of misclassification probabilities defined as $q_{ik} = \Pr(Y^*_{\xi t} = i \mid Y_{\xi t} = k)$ (see for example Tenenbein 1972). An alternative approach is by using what Carroll (1992) refers to as calibration probabilities. The calibration probabilities are defined as $c_{ki} = \Pr\left(Y_{\xi t} = k \mid Y^*_{\xi t} = i\right)$.

The misclassification probabilities can be used both with internal (designs 1 and 2 in Section 2.1) and external (design 3 in Section 2.1) validation samples. This is because the misclassification probabilities condition on the true classifications. Therefore, the misclassification probabilities can be regarded as transportable to the population of interest and can be used also in the case of an external validation sample. Unlike misclassification

probabilities, calibration probabilities can be used only with internal validation samples. This is because calibration probabilities condition on the observed classifications, which can not be considered as transportable to the population of interest when only external validation data exist. Nevertheless, here we argue that an external validation sample can be transformed into an internal validation sample. Since the misclassification process in the external validation sample is assumed to be representative of the misclassification process in the target population, we propose to calibrate $\Pr\left(Y_{\xi t}^{*}=i, Y_{\xi t}=k\right)$ on the marginal information derived from the main sample. In the simplest case, this calibration procedure can be performed using an Iterative Proportional Fitting (IPF) algorithm (Deming and Stephan 1940). This transformation will be assumed throughout this paper when employing double sampling design 3.

In a cross-sectional context Tenenbein (1970,1972) developed maximum likelihood estimators using calibration probabilities. In a recent paper, Tzavidis and Lin (2004) proposed a missing data specification that utilises misclassification probabilities for deriving maximum likelihood or quasi-likelihood estimates. All previous approaches lead to identical results.

Unlike in the cross-sectional case, the use of misclassification or calibration probabilities when adjusting panel data for misclassification error, assuming that only cross-sectional validation data are available, requires careful consideration. This is because the absence of panel validation data dictates the use of additional assumptions for identifying the parameters of the panel misclassification error mechanism. A widely used assumption is the Independent Classification Errors (ICE) assumption. The ICE assumption can be defined either with misclassification or calibration probabilities. When using misclassification probabilities, the ICE assumption embodies the following two assumptions: (a) The observed classifications are conditionally independent given the true classifications and (b) the misclassification error at the current time point depends only on the current true state and not on previous or future true states. The ICE with calibration probabilities embodies the same assumptions but conditioning now on the observed instead of the true classifications. Both assumptions were studied by Meyer (1988). The author

argues that the main difference between the use of misclassification or calibration probabilities is in assumption (b) and concludes that the use of the ICE assumption with misclassification probabilities is more reasonable. Therefore, in this paper we will consider only misclassification probabilities.

3. Two Alternative Specifications for Misclassification Error Correction via Double Sampling

In this section we present two alternative specifications for adjusting for misclassification error in gross flows. The first one leads to a moment-based estimator that has been already proposed in the literature (Poterba and Summers 1986; Singh and Rao 1995). The second specification is what we propose and is based on expressing the misclassification problem as a missing data problem, which we solve via the EM algorithm (Dempster; Laird and Rubin 1976). This second approach leads to maximum likelihood estimates.

3.1 Moment-based Inference for Gross Flows under Misclassification Error and Double Sampling

Suppose that we conduct a panel survey where a sample unit ξ is interviewed at two consecutive time points $t, t+1$. The variable of interest, i.e. the flows between r mutually exclusive states measured by the panel survey, is subject to misclassification error. Denote by P_{kl} the probability that unit ξ truly belongs in state k at t and state l at $t+1$ and by Π_{ij} the probability that unit ξ is observed in state i at t and state j at $t+1$. Let P denote the matrix with elements P_{kl} and Π the matrix with elements Π_{ij}. Corresponding to each element of Π and unit ξ we define the random variables $Y^*_{\xi t}, Y^*_{\xi t+1}$, which describe the observed (affected by misclassification error) classifications of unit ξ at t and $t+1$. We also define the random variables $Y_{\xi t}, Y_{\xi t+1}$, which describe the true classifications of unit ξ at t and $t+1$. The pairs $\left(Y^*_{\xi t}, Y^*_{\xi t+1}\right)$ and $\left(Y_{\xi t}, Y_{\xi t+1}\right)$ are assumed to be iid for different sample units. We further assume that we can use a cross-sectional validation procedure through which we can make inference about the

misclassification error process. The misclassification probabilities are denoted by $q_{ijkl} = \Pr(Y_{\xi t}^{\ *} = i \;\; Y_{\xi t+1}^{*} = j \mid Y_{\xi t} = k \;\; Y_{\xi t+1} = l)$ and the matrix of misclassification probabilities by $Q(t,t+1)$.

Generally speaking, the misclassification error model is defined by expressing the joint distribution of the observed and true classifications as a product of the misclassification probabilities times the true transition probabilities as follows

$$\Pr\left(Y_{\xi t}^{\ *} = i \;\; Y_{\xi t+1}^{*} = j\right) = \sum_{k=1}^{r}\sum_{l=1}^{r} \Pr(Y_{\xi t}^{\ *} = i \;\; Y_{\xi t+1}^{*} = j \mid Y_{\xi t} = k \;\; Y_{\xi t+1} = l) \;\; \Pr\left(Y_{\xi t} = k \;\; Y_{\xi t+1} = l\right) \quad (3.1)$$

Expressing (3.1) in vector notation, assuming that $Q(t,t+1)$ is non-singular and solving the resulting system of equations with respect to the vector of true flows P we obtain the following expression for the adjusted gross flows

$$vec(P) = \left[Q(t,t+1)\right]^{-1} vec(\Pi). \quad (3.2)$$

The estimation of the misclassification matrix $Q(t,t+1)$ is not straightforward. To see this note that the number of free parameters when estimating $Q(t,t+1)$ is equal to $r^2\left(r^2-1\right)$. This implies that information obtained from a cross-sectional validation sample is not sufficient to determine $Q(t,t+1)$. We therefore need to introduce additional assumptions that will enable us to estimate the longitudinal misclassification matrix. The assumption that we utilise is the ICE one with misclassification probabilities (Section 2.2), which is more rigorously defined as follows:

$$\Pr\left(Y_{\xi t}^{*} = i, Y_{\xi t+1}^{*} = j \mid Y_{\xi t} = k, Y_{\xi t+1} = l\right) = \Pr\left(Y_{\xi t}^{*} = i \mid Y_{\xi t} = k\right)\Pr\left(Y_{\xi t+1}^{*} = j \mid Y_{\xi t+1} = l\right).$$

Restating the ICE assumption we can say that (a) The observed classifications $Y_{\xi t}^{*}, Y_{\xi t+1}^{*}$ are conditionally independent given the true classifications $Y_{\xi t}, Y_{\xi t+1}$ and (b) The misclassification at t depends only on the current true state and not on the previous or future true states. Denote by $Q(t)$ the cross-sectional matrix of misclassification probabilities at time t with elements q_{ik}, by $Q(t+1)$ the cross-sectional matrix of misclassification probabilities at $t+1$ with elements q_{jl}

and by $\otimes$ the kronecker product. An implication of ICE is that the longitudinal misclassification matrix can now be expressed in matrix notation as follows:

$$Q(t,t+1) = Q(t+1) \otimes Q(t) \ .$$

However, since $Q(t+1)$is not known, we further assume that $Q(t) = Q(t+1)$(assumption of stationary misclassification error). Assuming now that all quantities involved in the measurement error model can be estimated by utilising a double sampling design and the ICE assumption, an estimator of (3.2) is given by the following expression

$$vec\left(\hat{P}\right) = \left[\hat{Q}(t) \otimes \hat{Q}(t)\right]^{-1} vec\left(\hat{\Pi}\right). \tag{3.3}$$

Let us now examine the effect of the choice of double sampling design on the moment-based estimator. Since we allow only for cross-sectional validation data, the validation sample is used for estimating the parameters of the misclassification error mechanism while the main sample is used for estimating gross flows. This is true under all three double sampling designs. Thus, the choice of double sampling design has no effect on point estimation performed via the moment-based estimator. Differences may be encountered in variance estimation due to the extra covariance terms introduced under double sampling design 1. This is investigated in Section 4.

A drawback associated with the use of the moment-based estimator is that under certain conditions it can produce estimates that lie outside the parameter space. This can happen due to the inversion of the misclassification matrix involved in (3.3). As an alternative to the moment-based estimator, in the upcoming section we propose a maximum likelihood estimator.

3.2 Likelihood-based Inference for Gross Flows under Misclassification Error and Double Sampling

As an alternative to the moment-based estimator, in this section we propose a likelihood-based approach for adjusting gross flows for misclassification error. A model is specified by simultaneously modeling the true transition process and the misclassification error process within the context of a missing data problem. The model parameters are estimated by maximizing the likelihood of the augmented data via the EM algorithm. Two alternative double sampling designs

are considered. In Section 3.2.1 we allow for double sampling design 2 while in Section 3.2.2 we allow for double sampling design 1. The case of double sampling design 3 is covered by double sampling design 2 using the transformation described in Section 2.2.

3.2.1 Likelihood-based Inference under Double Sampling Design 2

Let us assume that the panel survey that is utilized for estimating gross flows is affected by misclassification error and that a cross-sectional validation sample of n^v units is selected via double sampling design 2. The main survey provides information about the flows of the sample units between r mutually exclusive states at t and $t+1$. On the other hand, the validation survey provides information about the cross-sectional incidence of misclassification error related to these states at t. In what follows we define a category as a pair of states for which there is a flow, so there are r^2 such flow categories.

Consider the cross-classification of the fallible with the true classifications. Denote by n_{ij}, n_{ij}^v the number of sample units classified in cell ij defined by this cross-classification in the main and in the validation samples respectively. We formulate a model by combining information from both samples. This will eventually lead to a missing data problem. One source of missing data is attributed to the different time dimensions of the main and the validation surveys. The other source of missing data is due to the fact that individuals participating in the main survey do not participate in the validation survey.

Denote by P_i the probability that a respondent truly belongs in category i and by q_{ij} the probability that a respondent is classified in category j given that he/she truly belongs in category i. The probability that a sample unit belongs in cell ij is expressed as a product of the true transition probabilities and the misclassification probabilities. Denote further by Θ the vector of parameters, by $D^{Complete}$ the complete (augmented) data and by a superscript $(*)$ any missing data. Assuming independence between the main and the validation samples, the likelihood function of the augmented data is given by

$$L(\Theta; D^{Complete}) = \prod_{i=1}^{r^2}\prod_{j=1}^{r^2}\left(P_i q_{ij}\right)^{n_{ij}^{(*)}} \prod_{i=1}^{r^2}\prod_{j=1}^{r^2}\left(P_i q_{ij}\right)^{n_{ij}^{v(*)}}.$$

Taking the logarithms on both sides and imposing the constraint

$$\sum_{i=1}^{r^2} P_i = 1$$

we obtain the following expression for the log-likelihood function of the augmented data

$$l(\Theta; D^{Complete}) = \sum_{i=1}^{r^2-1}\left(n_{i\bullet}^{v(*)} + n_{i\bullet}^{(*)}\right)\log P_i + \left(n_{r^2\bullet}^{v(*)} + n_{r^2\bullet}^{(*)}\right)\log\left(1 - \sum_{i=1}^{r^2-1} P_i\right) + \sum_{i=1}^{r^2}\sum_{j=1}^{r^2}\left(n_{ij}^{v(*)} + n_{ij}^{(*)}\right)\log\left(q_{ij}\right). \qquad (3.4)$$

The longitudinal misclassification probabilities, q_{ij}, are unknown and are estimated using the cross-sectional misclassification probabilities and the ICE assumption. The log-likelihood function given in (3.4) is presented in its generic form i.e. without incorporating the ICE assumption. However, after incorporating ICE we need to add the extra constraint that the sum of the cross-sectional misclassification probabilities for a given true classification must add up to one. This extra constraint implies that we have to estimate $r^2 - r$ parameters that describe the misclassification error process and $r^2 - 1$ gross flows specific parameters.

Since the likelihood function involves missing data, one way of using this likelihood to maximise the likelihood of the observed data is via the EM algorithm. In the sequel we describe the expectation step (E-step) and the maximization step (M-step). Denote by D^v the observed data from the validation sample, by D^m the observed data from the main sample and by (h) the current iteration of the EM algorithm. In order to perform the E-step we need to estimate the conditional expectations of the unobserved quantities in the main sample and in the validation sample given the observed data. This can be done using the following result.

Result 3.1

Denote by $n_{\bullet j}$ the total number of sample units in the main sample classified by the standard measurement device as making transition j. The conditional expectations of the missing data in the main sample are estimated using the following expression

$$\hat{E}\left(n_{ij}^{(*)} \mid D^{m}, \Theta\right) = n_{\bullet j}\left(\frac{\hat{q}_{ij}^{(h)}\ \hat{P}_{i}^{(h)}}{\sum_{i=1}^{r^2}\hat{q}_{ij}^{(h)}\ \hat{P}_{i}^{(h)}}\right).$$

Proof

Proof of Result 3.1 is given in Appendix A

Result 3.2

Denote by n_k^v the total number of sample units in the validation sample that belong to the k^{th} cell of the misclassification matrix. The conditional expectations of the missing data in the validation sample are estimated using the following expression

$$\hat{E}\left(n_{ij}^{v(*)} \mid D^{v}, \Theta\right) = n_k^v\left(\frac{\hat{q}_{ij}^{(h)}\ \hat{P}_{i}^{(h)}}{\sum_{i}\sum_{j}\hat{q}_{ij}^{(h)}\ \hat{P}_{i}^{(h)}}\right)$$

Proof

Proof of Result 3.2 is given in Appendix A

Having performed the E-step, the missing data in log-likelihood function (3.4) are replaced by the estimated conditional expectations. The M-step can then be performed by numerically maximising (Dennis and Schnabel 1983) the log-likelihood function of the augmented data. The E and M steps are iterated until a convergence criterion, for example the L^2-norm of the vector of parameters derived from two successive iterations of the EM algorithm, is satisfied.

3.2.2 Likelihood-based Inference under Double Sampling Design 1

In Section 3.2.1 we formulated the model under double sampling design 2. In this section we formulate the model under double sampling design 1. Under this design independence between the units in the main sample and in the validation sample is not automatically guaranteed. However, independence can be imposed by dividing the main sample into units that participate only in the main survey and units that participate both in the main survey and in the validation survey. Having performed this separation, the information available from these two samples is as follows: The main survey is a panel survey and provides information on the observed flows of the

$n - n^v$ that participate only in the main survey. On the other hand, the validation survey provides now information on the cross-sectional incidence of misclassification error related to the classifications at time t and on the observed flows of the n^v units that participate both in the main and in the validation surveys. Under this design, the log-likelihood function of the augmented data is also given by (3.4) and is maximized using the EM algorithm. The E-step is described below.

For the main sample the conditional expectations of the missing data can be estimated using Result 3.1. However, for the validation sample estimating the conditional expectations of the missing data cannot be simply based on Result 3.2. This is because under design 1 we need to condition on two sets of observed data (the data from the main sample and the data from the validation sample). Therefore, a two-stage E-step is employed. For simplicity, we illustrate this E-step for the 4-state model that can be schematically described via a 4×4 cross-classification of the observed with the true classifications.

In the first stage of the E-step we estimate initial conditional expectations using Result 3.2. These provisional conditional expectations will therefore respect the cross-sectional validation information. However, we also need to respect the information about the observed flows of the units in the validation sample. This is achieved at the second stage. Based on the provisional conditional expectations, we compute the following quantities

$$a = n_{11}^{v(*)} + n_{21}^{v(*)},\ b = n_{12}^{v(*)} + n_{22}^{v(*)},\ \ c = n_{31}^{v(*)} + n_{41}^{v(*)},\ d = n_{32}^{v(*)} + n_{42}^{v(*)}$$
$$e = n_{13}^{v(*)} + n_{23}^{v(*)},\ \ f = n_{14}^{v(*)} + n_{24}^{v(*)},\ \ g = n_{33}^{v(*)} + n_{43}^{v(*)},\ h = n_{34}^{v(*)} + n_{44}^{v(*)}.$$

We then form two 2×2 tables the margins of which are defined by $\{a, b, c, d\}$ and $\{e, f, g, h\}$ respectively. It can be easily verified that the margins of these two tables summarise the information available for the units in the validation sample under double sampling design 1. More specifically, the column margins define the observed flows and the row margins define the cross-sectional validation information. Having formed these 2×2 tables, we then use the IPF (Deming and Stephan 1940) algorithm to rake the internal cells of these matrices to the data constraints that

we need to respect. The newly derived internal cells are denoted by $\{a^*, b^*, c^*, d^*\}$ and $\{e^*, f^*, g^*, h^*\}$. It remains to estimate the final conditional expectations of the unobserved quantities in the validation sample. In order to do so we form the 2×1 vectors that summarise $\{a^*, b^*, c^*, d^*\}$ and $\{e^*, f^*, g^*, h^*\}$. For example, the 2×1 vector defined by $n_{11}^{v(*)}, n_{21}^{v(*)}$ must respect the constraint that $a^* = n_{11}^{v(*)} + n_{21}^{v(*)}$. For the 4-state model one can form 8 such vectors. Using arguments analogous to the ones for Results 3.1 and 3.2, the conditional expectations are then estimated within each of these 2×1 vectors. For example,

$$\hat{E}\left(n_{11}^{v(*)} \mid D^v, \Theta\right) = a^* \left(\frac{\hat{q}_{11}^{(h)} \hat{P}_1^{(h)}}{\sum_{i=1}^{2} \hat{q}_{i1}^{(h)} \hat{P}_i^{(h)}} \right), \hat{E}\left(n_{21}^{v(*)} \mid D^v, \Theta\right) = a^* \left(\frac{\hat{q}_{21}^{(h)} \hat{P}_2^{(h)}}{\sum_{i=1}^{2} \hat{q}_{i1}^{(h)} \hat{P}_i^{(h)}} \right).$$

These estimated conditional expectations will respect both the cross-sectional validation information and the observed flows of the units in the validation sample. After estimating the conditional expectations of the unobserved quantities in the main and in the validation samples, the M-step is performed numerically.

4. Variance Estimation

Having investigated alternative approaches for point estimation, in this section we develop tools for variance estimation. Variance estimation for the moment-based estimator is discussed in Section 4.1. Variance estimation for the maximum likelihood estimator is discussed in Section 4.2.

4.1 Variance Estimation for the Moment-based Estimator

Using properties of vec operators, the moment-based estimator, under ICE, is given by

$$vec\left(\hat{P}\right) = \left[\hat{Q}^{-1} \hat{\Pi} \left(\hat{Q}^{-1} \right)^T \right] \tag{4.1}$$

In order to simplify the notation, we drop the parenthesis next to Q that is time specific. A variance estimator for (4.1) can be derived by employing the δ-method (Bishop, Fienberg and Holland 1975). This involves expanding $vec\left(\hat{P}\right)$ in a Taylor series around its true value $vec(P)$.

Let $vec\left[P\left(\hat{\Theta}\right)\right]=\left[g_1\left(\hat{\Theta}\right),g_2\left(\hat{\Theta}\right),...,g_{r^2}\left(\hat{\Theta}\right)\right]^T$ represent a $r^2\times 1$ vector of non-linear functions of a vector $\hat{\Theta}=\left(\hat{q}_{11},\hat{q}_{21},\hat{q}_{31},...,\hat{q}_{rr},\hat{\Pi}_{11},\hat{\Pi}_{21},\hat{\Pi}_{31},...,\hat{\Pi}_{rr}\right)$. Recall that q_{ik} denotes the misclassification probabilities and Π_{lj} denotes the observed transition probabilities between t and $t+1$. Note also that now we distinguish between the subscripts l,i. However, both subscripts refer to the observed classifications at t. Expanding $vec\left(\hat{P}\right)$ around its true value using Taylor series we have that

$$vec\left[P\left(\hat{\Theta}\right)\right]-vec\left[P\left(\Theta\right)\right]\approx\nabla_\Theta\left(\hat{\Theta}-\Theta\right),\nabla_\Theta=\frac{\partial vec\left[P\left(\Theta\right)\right]}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}. \qquad (4.2)$$

Taking the variance operator on both sides of (4.2) we have that

$$Var\left\{vec\left[P\left(\hat{\Theta}\right)\right]\right\}\approx\nabla_\Theta Var\left(\hat{\Theta}\right)\left(\nabla_\Theta\right)^T. \qquad (4.3)$$

In order to estimate (4.3), we need to evaluate the Jacobian matrices ∇_Θ, $\left(\nabla_\Theta\right)^T$ and estimate the covariance matrix $Var\left(\hat{\Theta}\right)$. For the later case, we need to estimate the following components: (a) the covariance matrix of the unadjusted estimated probabilities of transition $\hat{\Pi}_{lj}$ (b) the covariance matrix of the estimated misclassification probabilities $\hat{q}_{ik}$ and (c) the covariance of $\hat{\Pi}_{lj},\hat{q}_{ik}$.

Under simple random sampling, component (a) can be estimated using standard results for the variance of binomial random variables. Component (b) requires a second application of the δ-method. This is because the estimated misclassification probabilities are defined as ratios of random variables. Let $\hat{\Theta}^*=\left(n_{11}^v,n_{21}^v,n_{31}^v,...,n_{rr}^v\right)$. Applying the δ-method to $vec\left[Q\left(\hat{\Theta}^*\right)\right]$ we derive the following

$$vec\left[Q\left(\hat{\Theta}^{*}\right)\right]-vec\left[Q\left(\Theta^{*}\right)\right]\approx\nabla_{\Theta^{*}}\left(\hat{\Theta}^{*}-\Theta^{*}\right),\quad\nabla_{\Theta^{*}}=\frac{\partial vec\left[Q\left(\Theta^{*}\right)\right]}{\partial\Theta^{*}}\Big|_{\Theta^{*}=\hat{\Theta}^{*}}\cdot\text{(4.4)}$$

Taking the variance operator on both sides of (4.4) we obtain the following

$$Var\left\{vec\left[Q\left(\hat{\Theta}^{*}\right)\right]\right\}\approx\nabla_{\Theta^{*}}Var\left(\hat{\Theta}^{*}\right)\left(\nabla_{\Theta^{*}}\right)^{T}.\qquad(4.5)$$

In order to estimate (4.5) we need to evaluate the Jacobian matrices $\nabla_{\Theta^{*}}$, $\left(\nabla_{\Theta^{*}}\right)^{T}$ and the covariance matrix $Var\left(\hat{\Theta}^{*}\right)$. Under simple random sampling and taking into account that the sample size of the validation survey is fixed, we can treat n_{ik}^{v} as multinomial counts. Therefore $Var\left(\hat{\Theta}^{*}\right)$can be estimated using standard results for the variance of binomial random variables.

For component (c) i.e. the covariance between the unadjusted estimated probabilities of transition and the estimated misclassification probabilities we distinguish two cases. Under double sampling design 2 and double sampling design 3 we assume that $Cov\left(\hat{q}_{ik,}\,\hat{\Pi}_{lj}\right)=0.$ Under double sampling design 1 we assume that $Cov\left(\hat{q}_{ik,}\,\hat{\Pi}_{lj}\right)\neq 0.$ For the latter case we estimate this covariance term using the following result

Result 4.1

An estimator for the covariance term of interest is given by

$$\hat{Cov}\left(\frac{n_{ik}^{v}}{\sum_{i=1}^{r}n_{ik}^{v}},\frac{n_{lj}}{n}\right)\approx\frac{1}{n\hat{E}\left(\sum_{i=1}^{r}n_{ik}^{v}\right)}\left\{n^{v}\hat{\Pr}\left(Y_{\xi t}^{*}=i,Y_{\xi t+1}^{*}=j,Y_{\xi t}=k\right)-\frac{\hat{E}\left(n_{ik}^{v}\right)\hat{E}\left(n_{lj}\right)}{n^{v}}\right.$$

$$\left.-\frac{\hat{E}\left(n_{ik}^{v}\right)}{\hat{E}\left(\sum_{i=1}^{r}n_{ik}^{v}\right)}\left[n^{v}\hat{\Pr}\left(Y_{\xi t}^{*}=i,Y_{\xi t+1}^{*}=j,Y_{\xi t}=k\right)-\frac{\hat{E}\left(n_{ik}^{v}\right)\hat{E}\left(n_{lj}\right)}{n^{v}}-\sum_{l\neq i}\frac{\hat{E}\left(n_{ik}^{v}\right)\hat{E}\left(n_{lj}\right)}{n^{v}}\right]\right\}.$$

Proof

Proof of Result 4.1 is given in Appendix B

4.2 Variance Estimation for the Maximum Likelihood Estimator

In this section we perform variance estimation for the maximum likelihood estimator under double sampling design 2. Variance estimation for the maximum likelihood estimator implies the use of the inverse of the information matrix. However, due to the formulation of the model in a missing data framework, variance estimation must reflect the additional variability introduced by the existence of missing data. One way to obtain variance estimates for the parameters of interest when using the EM algorithm is by application of the Missing Information Principle (Louis 1982).

Denote by Z^m, Z^v the missing data in the main and in the validation samples respectively and by D^m, D^v the observed data from the main and the validation samples. The Missing Information Principle is defined as follows

$$Observed\ Information = Complete\ Information - Missing\ Information.$$

Following Louis (1982), the complete information matrix can be obtained using the second order derivatives of the log-likelihood function evaluated at the last step of the EM algorithm. The missing information matrix can be obtained by estimating the variance of the score functions. In spite of being able to derive general expressions for the expectation of the complete information matrix and the variance of the score functions, it is tedious to evaluate these expressions analytically. The main problem arises in evaluating the missing information matrix. An alternative solution is offered by means of Monte-Carlo simulation. The simulation algorithm is described in Tanner (1996). Having arrived at the maximum likelihood estimates (last step of the EM), we generate H complete datasets by drawing

$$Z_1^v, Z_2^v, \ldots, Z_H^v \overset{iid}{\sim} \Pr\left(Z^v \mid D^v, \hat{\Theta}_{mle}\right),$$

$$Z_1^m, Z_2^m, \ldots, Z_H^m \overset{iid}{\sim} \Pr\left(Z^m \mid D^m, \hat{\Theta}_{mle}\right)$$

where $\Pr\left(Z^v \mid D^v, \hat{\Theta}_{mle}\right), \Pr\left(Z^m \mid D^m, \hat{\Theta}_{mle}\right)$ denote the conditional distributions of the missing data in the validation and in the main samples respectively given the observed data and the maximum likelihood estimates and H denotes the total number of simulations. This first step can

be viewed as the imputation step. Having replaced the missing data with imputed values in simulation (h), we derive complete data $D^{complete(h)}$ that are employed for evaluating the complete information matrix and the missing information matrix. This is done by using the simulation-based (empirical) estimators for the complete information matrix and for the variance of the score functions defined respectively by

$$E\left[-\frac{\partial^2 l\left(\Theta; D^{complete}\right)}{\partial\Theta\ \partial\Theta^T} \mid D^m, D^v\right] = \frac{1}{H}\sum_{h=1}^{H} -\frac{\partial^2 l\left(\Theta; D^{complete(h)}\right)}{\partial\Theta\ \partial\Theta^T},$$

$$Var\left[\frac{\partial l\left(\Theta; D^{complete}\right)}{\partial\Theta} \mid D^m, D^v\right] = \frac{1}{H}\sum_{h=1}^{H}\left\{\frac{\partial l\left(\Theta; D^{complete(h)}\right)}{\partial\Theta} - E\left[\frac{\partial l\left(\Theta; D^{complete(h)}\right)}{\partial\Theta}\right]\right\}^2.$$

Having derived the complete information matrix and the missing information matrix, variance estimates are derived by inverting the matrix resulting from the difference of these two matrices.

Variance estimation for the maximum likelihood estimator under double sampling design 1 is more complex and is not tackled in this paper. The complexity arises due to the stepwise approach we follow for estimating the conditional expectations of the missing data in the validation sample (see Section 3.2.2). In order to tackle this problem, one may consider using computer intensive methods such as bootstrap or jackknife. In this paper, however, we will solely rely on empirical variance estimates derived via Monte-Carlo simulation.

5. Simulation Study

In this section we evaluate the performance of the alternative point and variance estimators using Monte-Carlo simulation. The simulation algorithm is designed as follows. In the first step we generate error free (true) gross flows. This is done by employing the probability distribution function of the true flows between two time points say t and $t+1$ and by drawing from this distribution a with replacement sample of size n. Having generated true flows, in the second step we assume the existence of a cross-sectional misclassification error model described by the misclassification probabilities. Using these probabilities, we generate the observed status at t given the true status at t for each sample unit ξ. Having generated the observed status at t, in the

third step we generate the observed status at $t+1$ given the observed status at t, the true status at t and the true status at $t+1$ for each sample unit ξ. This is equivalent to introducing panel misclassification error. Since all developments in this paper are based on the ICE assumption that uses misclassification probabilities, the panel misclassification error mechanism is simulated under ICE. After all three previous steps of the simulation have been completed, the joint distribution of the observed and the true classifications is constructed. From this distribution one can extract the marginal distribution that refers to the observed gross flows. Using the joint distribution, one can also extract the marginal distribution that refers to the cross-sectional incidence of misclassification error.

In order to simulate the availability of validation information derived from a validation sample of n^v units $(n^v < n)$, we distinguish two cases: (a) Double sampling design 1 is simulated by selecting a sub-sample of n^v units from the marginal distribution that describes the cross-sectional incidence of misclassification error derived after the first three steps of the simulation and (b) double sampling design 2 is simulated independently of the data generated by the first three steps of the simulation. The case of double sampling design 3 is covered by double sampling design 2 using the transformation described in Section 2.2.

We implement the simulation study within the context of estimating labour force gross flows. More specifically, the target is to estimate gross flows between the main labour force states i.e. Employment (**E**), Unemployment (**U**) and Inactivity (**N**) in the presence of misclassification error. We contrast the following three estimators (a) the estimator of the observed (unadjusted) flows denoted by **P-OBS**, (b) the moment-based estimator used to adjust for misclassification error (Section 3.1) denoted by **P-ST** and (c) the maximum likelihood estimator used to adjust for misclassification error (Section 3.2) denoted by **P-MLE**.

In simulation study I (Tables 1 and 2) we compare the estimator of the observed flows (**P-OBS**) with the moment-based estimator (**P-ST**) and the maximum likelihood estimator (**P-MLE**) under double sampling design 2. In simulation study II (Tables 3 and 4) we compare the estimator of the

observed flows (**P-OBS**) with the moment-based estimator (**P-ST**) and the maximum likelihood estimator (**P-MLE**) under double sampling design 1. For easing the computations, this comparison is performed only for a reduced model that only allows for flows between "Employment" and "Unemployment" or "Inactivity". In simulation study III (Table 5) we evaluate the performance of the variance estimator of the moment-based estimator under double sampling design 2, in simulation study IV (Table 6) we evaluate the performance of the variance estimator of the moment-based estimator under double sampling design 1 and in Simulation study V (Table 7) we evaluate the performance of the variance estimator of the maximum likelihood estimator under double sampling design 2. Due to the computer intensive methods required for computing this last variance estimator (Section 4.2), we consider its performance also relatively to the reduced model that allows for flows between "Employment" and "Unemployment" or "Inactivity".

The properties of the various point and variance estimators are assessed using the following criteria: (a) Relative bias of a point or a variance estimator, (b) standard deviation of a point estimator, (c) Root Mean Squared Error (**RMSE**) of a point estimator, (d) relative efficiency (**RE**) of the moment-based estimator compared to the maximum likelihood estimator defined as the ratio between the RMSE's of these two estimators and (e) coverage rate of a variance estimator.

Simulation Study I: $n = 60000, n^v = 10000$, Double sampling design 2

Table 1: True flows and point estimates (Averages over simulations)

Flow	True Flows	P-OBS	P-ST	P-MLE
E → E	0.7316	0.7174	0.7313	0.7318
U → E	0.0131	0.0160	0.0131	0.0130
N → E	0.0091	0.0162	0.0092	0.0090
E → U	0.0047	0.0079	0.0047	0.0046
U → U	0.0283	0.0268	0.0281	0.0281
N → U	0.0071	0.0100	0.0069	0.0071
E → N	0.0093	0.0162	0.0094	0.0092
U → N	0.0049	0.0080	0.0047	0.0049
N → N	0.1919	0.1815	0.1926	0.1923

Table 2: Comparing the alternative estimators of the labour force gross flows

	P-OBS	P-ST	P-MLE	P-OBS	P-ST	P-MLE	P-OBS	P-ST	P-MLE	
Flow	Relative Bias (%)			Standard Deviation ($*10^3$)			RMSE ($*10^3$)			RE
E → E	-1.94	-0.04	0.03	1.98	2.63	2.28	14.29	2.64	2.29	**1.15**
U → E	22.14	0.001	-0.76	0.37	0.77	0.66	2.96	0.77	0.66	**1.16**
N → E	78.02	1.10	-1.10	0.49	0.88	0.66	7.14	0.89	0.67	**1.33**
E → U	68.09	0.001	-2.13	0.52	0.67	0.55	3.30	0.67	0.55	**1.22**
U → U	-5.30	-0.71	-0.71	0.72	0.87	0.80	1.61	0.88	0.81	**1.09**
N → U	40.85	-2.82	0.001	0.35	0.68	0.55	3.01	0.69	0.55	**1.25**
E → N	74.19	1.08	-1.08	0.46	0.93	0.71	6.98	0.94	0.71	**1.32**
U → N	63.27	-4.08	0.001	0.39	0.72	0.57	3.17	0.73	0.57	**1.28**
N → N	-5.42	0.36	0.21	1.59	2.20	1.85	10.94	2.23	1.85	**1.21**

Simulation Study II: $n = 60000, n^{v} = 10000$, Double sampling design 1

Table 3: True flows and point estimates (Averages over simulations)

Flow	True Flows	P-OBS	P-ST	P-MLE
E → E	0.7288	0.7161	0.7293	0.7288
U+N → E	0.0129	0.0319	0.0127	0.0130
E → U+N	0.0054	0.0249	0.0052	0.0056
U+N → U+N	0.2529	0.2271	0.2528	0.2526

Table 4: Comparing the alternative estimators of the labour force gross flows

Flow	P-OBS	P-ST	P-MLE	P-OBS	P-ST	P-MLE	P-OBS	P-ST	P-MLE	
	Relative Bias (%)			Standard Deviation ($*10^3$)			RMSE ($*10^3$)			RE
E → E	-1.74	0.07	-0.002	2.46	3.23	3.04	12.9	3.27	3.04	**1.07**
U+N → E	147	-1.55	0.77	1.04	1.79	1.57	19.0	1.80	1.58	**1.14**
E → U+N	361	-3.70	3.70	0.95	1.78	1.48	19.5	1.79	1.49	**1.20**
U+N → U+N	-10.2	-0.04	-0.12	2.42	3.54	3.13	25.9	3.54	3.14	**1.13**

Simulation Study III: $n = 60000, n^{v} = 2150$, Double sampling design 2

Table 5: Performance of the variance estimator of the moment-based estimator

Flow	$\mathrm{E}\left[\hat{\mathrm{V}}\left(\hat{\mathrm{P}}\right)\right]$ ($*10^6$)	$\mathrm{V}\left(\hat{\mathrm{P}}\right)$ ($*10^6$)	Absolute Relative Bias (%)	Coverage Rate (95%)
E → E	32.6	32.7	0.30	0.945
U → E	3.47	3.48	0.28	0.934
N → E	8.55	8.44	1.30	0.949
E → U	3.42	3.44	0.58	0.934
U → U	2.82	2.80	0.71	0.924
N → U	2.89	2.87	0.69	0.939
E → N	8.48	8.36	1.43	0.948
U → N	2.96	2.88	2.77	0.935
N → N	27.9	27.6	1.08	0.943

Simulation Study IV: $n = 60000, n^v = 2150$, Double sampling design 1

Table 6: Performance of the variance estimator of the moment-based estimator

Flow	$\mathrm{E}\left[\hat{\mathrm{V}}\left(\hat{\mathrm{P}}\right)\right]$ **(*10^6)**	$\mathrm{V}\left(\hat{\mathrm{P}}\right)$ **(*10^6)**	**Absolute Relative Bias (%)**	**Coverage Rate (95%)**
E → E	31.4	31.3	0.32	0.944
U → E	3.23	3.29	1.82	0.931
N → E	8.21	8.29	0.96	0.942
E → U	3.38	3.44	1.74	0.930
U → U	2.72	2.73	0.36	0.920
N → U	2.87	2.90	1.03	0.933
E → N	8.20	8.31	1.32	0.943
U → N	2.79	2.84	1.76	0.933
N → N	27.0	27.2	0.73	0.942

Simulation Study V: $n = 60000, n^v = 10000$, Double sampling design 2

Table 7: Performance of the variance estimator of the maximum likelihood estimator

Flow	$\mathrm{E}\left[\hat{\mathrm{V}}\left(\hat{\mathrm{P}}\right)\right]$ **(*10^6)**	$\mathrm{V}\left(\hat{\mathrm{P}}\right)$ **(*10^6)**	**Absolute Relative Bias (%)**	**Coverage Rate (95%)**
E → E	5.19	5.00	3.80	0.94
E → U+N	2.95	2.02	46.0	0.90
U+N → E	2.95	2.00	47.5	0.92
U+N → U+N	7.70	6.80	13.2	0.94

5.1 Discussion

We start by comparing the moment-based estimator with the maximum likelihood estimator under alternative double sampling designs. Results from simulation study I (Table 2) indicate that using the maximum likelihood estimator instead of the moment-based estimator under double sampling design 2 leads to gains in relative efficiency that range between 9% and 33% (see last column of Table 2). Double sampling design 2 may be more reasonable with panel data. This is because a validation sample that is selected by sub-sampling units from the main (panel) survey (i.e. using double sampling design 1) may increase the response burden of these units. However, design 2 is also associated with higher costs. This is because when using an independently selected validation sample we conduct an additional cross-sectional survey on individuals that do not participate in the main survey. The moment-based estimator uses information from the cross-sectional validation sample only for estimating the misclassification probabilities. On the other

hand, the maximum likelihood estimator makes optimal use of the cross-sectional validation information leading to an increase of the effective sample size. One could object that in order to gain this increased efficiency, we pay the price of conducting an expensive validation survey. For this reason, in simulation study II we contrasted the maximum likelihood estimator with the moment-based estimator under double sampling design 1. Under this design both estimators use the same information. Again our results indicate that the maximum likelihood estimator is more efficient with relative efficiency gains now ranging between 7% and 20% (see last column of Table 4).

In simulation studies III and IV we evaluate the performance of the variance estimator of the moment-based estimator under alternative double sampling designs. The results in Tables 5 and 6 indicate that the variance estimators of the moment-based estimator work well with low relative bias and coverage rates close to 95%. In simulation study V we assess the variance estimator of the maximum likelihood estimator. Results from this simulation (Table 7) indicate that the variance estimator of the maximum likelihood estimator is conservative since it overestimates the true variance. This overestimation occurs mainly in the off-diagonal elements of the gross flows matrix. Despite being conservative, this variance estimator captures the variability due to the missing data and results in reasonable coverage rates that range between 90% and 94%.

6. Application: Adjusting for Misclassification Error in Labour Force Gross Flows Estimated from the US Current Population Survey (CPS)

In this section we employ US CPS labour force gross flows that have been previously analysed by Poterba and Summers (1986) using the moment-based estimator. In addition to the Poterba and Summers analysis, we further present maximum likelihood estimation but for the reduced model that allows for flows only between employment (**E**) and Unemployment (**U**) or Inactivity (**N**).

Two applications are presented. In the first application we use the misclassification matrix of Poterba and Summers (1986 p.1323) assuming double sampling design 2 (Section 2.1). The diagonal elements of the matrix of misclassification probabilities are reported in Appendix C

(Table C.1, column entitled "original"). Adjustment for misclassification error is performed using the moment-based and the maximum likelihood estimators. Variance estimates are also provided. More specifically, for the observed (unadjusted) labour force gross flows variance estimates are derived using multinomial results whereas for the moment-based and the maximum likelihood estimators variance estimates are derived using the results of Section 4. Results from this application are reported in Table 8. In the second application we compare the moment-based estimator with the maximum likelihood estimator when "intense" misclassification exists. In order to perform this comparison, we modify the misclassification matrix used in the first application. The diagonal elements of this modified misclassification matrix are also reported in Appendix C (Table C.1, column entitled "modified"). Results from this application are given in Table 9. For both applications the convergence criterion for the EM algorithm, as this is defined by the L^2-norm, is $\varepsilon = 10^{-8}$. Identification of the model parameters is checked by initializing the EM algorithm using different sets of starting values and examining whether the algorithm converges to the same point. For both applications the EM algorithm converges. The sample sizes of the main and the validation surveys are $n = 163907$ and $n^v = 20000$ respectively.

Table 8: Observed and adjusted, using the moment-based and the maximum likelihood estimators, labour force gross flows from the US CPS. Estimated standard deviations in parenthesis

Flow	Observed	Moment-based	Maximum Likelihood
E → E	0.560 ($1.22*10^3$)	0.5814 ($2.63*10^3$)	0.5815 ($2.16*10^3$)
E → U+N	0.029 ($4.11*10^4$)	0.0107 ($1.24*10^3$)	0.0106 ($1.21*10^3$)
U+N → E	0.028 ($4.07*10^4$)	0.0097 ($1.23*10^3$)	0.0097 ($1.21*10^3$)
U+N → U+N	0.383 ($1.19*10^3$)	0.3982 ($2.40*10^3$)	0.3982 ($1.82*10^3$)

Table 9: Observed and adjusted, using the moment-based and the maximum likelihood estimators, labour force gross flows from the US CPS under intense misclassification

Flow	Observed	Moment-based	Maximum Likelihood
E → E	0.560	0.581	0.5791
E → U+N	0.029	-0.0027	0.000097
U+N → E	0.028	-0.0017	0.000903
U+N → U+N	0.383	0.4234	0.4199

The existence of measurement error when estimating labour force gross flows leads to an overestimation of the labour market mobility. The effect of adjusting labour force gross flows for measurement error is to increase the diagonal elements and decrease the off-diagonal elements of the unadjusted gross flows matrix. This is consistent with the results of previous research (Poterba and Summers 1986; Singh and Rao 1995). The higher efficiency of the maximum likelihood estimator, compared to the moment-based estimator, is further illustrated in Table 8 by examining the estimated standard deviations. However, here we should also account for the fact that the variance estimator of the maximum likelihood estimator, using the Missing Information Principle, overestimates the true variance of this estimator (see Simulation V in Section 5). Last but not least, assuming that the adjusted estimates are unbiased, both the moment-based and the maximum likelihood estimators outperform the unadjusted estimator in Mean Squared Error terms.

In the second application we contrasted the moment-based estimator with the maximum-likelihood estimator in the presence of "intense" misclassification. Even a relatively small change in an entry of the original misclassification matrix is capable of causing the moment-based estimator to produce estimates that lie outside the boundaries of the parameter space (Table 9). This is partially due to the inversion of the misclassification matrix involved in deriving the moment-based estimates (Section 3.1). A further reason, however, is the use of the ICE assumption. The effect of the ICE assumption is to overestimate the panel misclassification error compared to a case where serial correlation in the misclassification error exists. Unlike the moment-based estimator, the maximum likelihood estimator constrains the adjusted estimates to lie within the boundaries of the parameter space.

7. Summary

In this paper we present alternative approaches for inference when gross flows subject to misclassification error and cross-sectional validation information is available. We argue that, compared to the traditional moment-based approach, a more efficient solution is offered by simultaneously modeling the true transition process and the misclassification error process within

the context of a missing data problem. Monte-Carlo simulation results verify that the likelihood-based approach offers significant gains in efficiency over the moment-based method. This is true under alternative double sampling designs. Variance estimation is considered and the proposed variance estimators appear to have good coverage properties. Using a real data application we illustrate that under certain conditions the moment-based estimator can produce estimates that lie outside the boundaries of the parameter space. Unlike the moment-based estimator, the maximum likelihood estimator constraints the adjusted estimates to lie within the boundaries of the parameter space. Based on the increased efficiency and the desirable numerical properties of the maximum likelihood estimator, we propose that this estimator should be preferred over the moment-based estimator.

Currently, we investigate the application of this methodology in other areas of statistical research such as in demographic applications for tackling the problem of heaping and in statistical disclosure control for protecting sensitive data via the introduction of misclassification error.

ACKNOWLEDGEMENT

The author would like to thank Ray Chambers for his guidance and helpful comments. The work in this paper was supported by the Economic and Social Research Council through award S42200034035.

APPENDIX A: PROOFS OF RESULTS IN SECTION 3

Proof of Result 3.1

Recalling the notation from Section 3, the expectations of the missing data can be expressed as follows:

$$E\left(n_{ij}^{(*)}\right) = nE\left(Y_{\xi t\to t+1} = i, Y^{*}_{\xi t\to t+1} = j\right). \tag{A.1}$$

Expression (A.1) is re-defined below

$$E\left(n_{ij}^{(*)}\right) = nE\left(Y^{*}_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) E\left(Y_{\xi t\to t+1} = i\right).$$

The observed data from the main sample are expressed as follows

$$n_{\bullet j} = n\sum_{i=1}^{r^2} E\left(Y^*_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) E\left(Y_{\xi t\to t+1} = i\right).$$

Given the observed data, the conditional expectations of the missing data are expressed as follows

$$E\left(n_{ij}^{(*)} \mid D^m\right) = n_{\bullet j}\left[\frac{E\left(Y^*_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) E\left(Y_{\xi t\to t+1} = i\right)}{\sum_{i=1}^{r^2} E\left(Y^*_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) E\left(Y_{\xi t\to t+1} = i\right)}\right]. \qquad \text{(A.2)}$$

The expectations of the random variables involved in the expression above are determined using results for binomial random variables. More specifically,

$$E\left(Y^*_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) = q_{ij},\ E\left(Y_{\xi t\to t+1} = i\right) = P_i\ . \qquad \text{(A.3)}$$

Substituting (A.3) in (A.2) we obtain the required result

$$\hat{E}\left(n_{ij}^{(*)} \mid D^m, \Theta^{(h)}\right) = n_{\bullet j}\left(\frac{\hat{q}{}_{ij}^{(h)}\ \hat{P}{}_i^{(h)}}{\sum_{i=1}^{r^2} \hat{q}{}_{ij}^{(h)}\ \hat{P}{}_i^{(h)}}\right).$$

Proof of Result 3.2

Using the same notation as in Result 3.1, the expectations of the missing data in the validation sample are expressed as

$$E\left(n_{ij}^{v(*)}\right) = n^v E\left(Y_{\xi t\to t+1} = i, Y^*_{\xi t\to t+1} = j\right). \qquad \text{(A.4)}$$

Expression (A.4) is re-defined below

$$E\left(n_{ij}^{v(*)}\right) = n^v E\left(Y^*_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) E\left(Y_{\xi t\to t+1} = i\right).$$

For the validation sample we have information about the cross-sectional incidence of error

$$n_k^v = n^v \sum_i \sum_j E\left(Y^*_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) E\left(Y_{\xi t\to t+1} = i\right).$$

Given the observed data, the conditional expectations of the missing data are expressed as follows

$$E\left(n_{ij}^{v(*)} \mid D^v\right) = n_k^v\left[\frac{E\left(Y^*_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) E\left(Y_{\xi t\to t+1} = i\right)}{\sum_{j=1}^{r^2}\sum_{i=1}^{r^2} E\left(Y^*_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) E\left(Y_{\xi t\to t+1} = i\right)}\right]. \qquad \text{(A.5)}$$

Replacing the conditional expectations in (A.5) using binomial results, we derive the final result

$$\hat{E}\left(n_{ij}^{v(*)} \mid D^{v}, \Theta^{(h)}\right) = n_{k}^{v}\left(\frac{\hat{q}_{ij}^{(h)}\,\hat{P}_{i}^{(h)}}{\sum_{i}\sum_{j}\hat{q}_{ij}^{(h)}\,\hat{P}_{i}^{(h)}}\right) .$$

APPENDIX B: PROOFS OF RESULTS IN SECTION 4

Lemma B.1

An approximate expression for the expectation of a function $g(X,Y)$ of two random variables X,Y using a Taylor's series expansion around (μ_X,μ_Y) is given by

$$E[g(X,Y)] \approx g(\mu_X,\mu_Y) + \frac{1}{2}\frac{\partial^2}{\partial y^2} g(X,Y)\,|_{\mu_X,\mu_Y}\ Var(Y) + \frac{1}{2}\frac{\partial^2}{\partial x^2} g(X,Y)\,|_{\mu_X,\mu_Y}\ Var(X) + \frac{\partial^2}{\partial x \partial y} g(X,Y)\,|_{\mu_X,\mu_Y}\ Cov(X,Y).$$

Proof

Proof of this Lemma can be found in Mood et al. (1963 p.181).

Result B.1

Let X,Y,A denote three random variables and n is fixed. An approximate expression for $Cov\left(\frac{X}{Y},\frac{A}{n}\right)$ is given by

$$Cov\left(\frac{X}{Y},\frac{A}{n}\right) \approx \frac{1}{nE(Y)}\left[Cov(A,X) - \frac{E(X)}{E(Y)}Cov(A,Y)\right].$$

Proof

We start by expressing the covariance term of interest using the standard covariance definition

$$Cov\left(\frac{X}{Y},\frac{A}{n}\right) = E\left(\frac{X}{Y}\frac{A}{n}\right) - E\left(\frac{X}{Y}\right)E\left(\frac{A}{n}\right) = \frac{1}{n}\left[E\left(\frac{AX}{Y}\right) - E\left(\frac{X}{Y}\right)E(A)\right]. \qquad \text{(B.1)}$$

We approximate $E\left(\frac{AX}{Y}\right)$ and $E\left(\frac{X}{Y}\right)$ via a Taylor series expansion of $\frac{AX}{Y}$, $\frac{X}{Y}$ around (μ_X,μ_Y,μ_A), (μ_X,μ_Y) respectively. The first Taylor series expansion, using Lemma B.1, is given below

$$E\big[g(X,Y,A)\big] \approx g(\mu_X,\mu_Y,\mu_A) + \frac{1}{2}\frac{\partial^2}{\partial Y^2} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\ Var(Y) + \frac{1}{2}\frac{\partial^2}{\partial X^2} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\ Var(X)$$
$$+\frac{1}{2}\frac{\partial^2}{\partial A^2} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\ Var(A) + \frac{\partial^2}{\partial X \partial Y} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\ Cov(X,Y)$$
$$+\frac{\partial^2}{\partial X \partial A} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\ Cov(X,A) + \frac{\partial^2}{\partial A \partial Y} g(X,Y,A)\,|_{\mu_X,\mu_Y,\mu_A}\ Cov(Y,A).$$

and

$$E\big[g(X,Y,A)\big] \approx \frac{\mu_X \mu_A}{\mu_Y} + \frac{1}{2}\frac{2\mu_X \mu_A}{{\mu_Y}^3} Var(Y) - \frac{\mu_A}{{\mu_Y}^2} Cov(X,Y) + \frac{1}{\mu_Y} Cov(X,A) - \frac{\mu_X}{{\mu_Y}^2} Cov(A,Y). \quad \text{(B.2)}$$

The second Taylor series expansion, using Lemma B.1, is given below

$$E\big[g(X,Y)\big] \approx g(\mu_X,\mu_Y) + \frac{1}{2}\frac{\partial^2}{\partial Y^2} g(X,Y)\,|_{\mu_X,\mu_Y}\ Var(Y) + \frac{1}{2}\frac{\partial^2}{\partial X^2} g(X,Y)\,|_{\mu_X,\mu_Y}\ Var(X)$$
$$+\frac{\partial^2}{\partial X \partial Y} g(X,Y)\,|_{\mu_X,\mu_Y}\ Cov(X,Y).$$

It follows that

$$E\big[g(X,Y)\big] \approx \frac{\mu_X}{\mu_Y} + \frac{1}{2}\frac{2\mu_X}{{\mu_Y}^3} Var(Y) - \frac{1}{{\mu_Y}^2} Cov(X,Y). \quad \text{(B.3)}$$

Substituting expressions (B.2) and (B.3) into (B.1), we derive the required result

Proof of Result 4.1

Let $X = n_{ik}^v, Y = \sum_{i=1}^{r} n_{ik}^v, A = n_{lj}$ be three random variables and n fixed. Result 4.1 can be obtained by direct application of Result B.1. We first evaluate the following covariance term

$$Cov\left(n_{ik}^v, n_{lj}\right) = E\left(n_{ik}^v, n_{lj}\right) - E\left(n_{ik}^v\right) E\left(n_{lj}\right). \quad \text{(B.4)}$$

We define the following indicator variables

$$I_\xi = \begin{cases} 1 \ if \ individual \ \xi \ has \ status \ ik \\ 0 \quad otherwise \end{cases} \qquad J_{\xi'} = \begin{cases} 1 \ if \ individual \ \xi' \ has \ status \ lj \\ 0 \quad otherwise \end{cases}.$$

We also define by S, s the indicators for the main and the validation samples respectively. It follows that

$$E\left(n_{ik}^v n_{lj}\right) = E\left(\sum_{\xi \in s} I_\xi \sum_{\xi' \in S} J_{\xi'}\right) = E\left(\sum_{\xi \in s} I_\xi J_\xi + \sum_{\substack{\xi \in s, \xi' \in S \\ \xi \neq \xi'}} I_\xi J_{\xi'}\right) = \sum_{\xi \in s} E\left(I_\xi J_\xi\right) + \sum_{\substack{\xi \in s, \xi' \in S \\ \xi \neq \xi'}} E\left(I_\xi J_{\xi'}\right) \quad \text{(B.5)}$$

Furthermore,

$$E\left(n_{ik}^{v}\right)E\left(n_{lj}\right)=\sum_{\xi\in s}\sum_{\xi'\in S}E\left(I_{\xi}\right)E\left(J_{\xi'}\right)=\sum_{\xi\in s}E\left(I_{\xi}\right)E\left(J_{\xi}\right)+\sum_{\substack{\xi\in s,\xi'\in S\\ \xi\neq\xi'}}E\left(I_{\xi}\right)E\left(J_{\xi'}\right). \quad \text{(B.6)}$$

Substituting expressions (B.5) and (B.6) into (B.4), we obtain the following result

$$Cov\left(n_{ik}^{v},n_{lj}\right)=\sum_{\xi\in s}E\left(I_{\xi}J_{\xi}\right)-\sum_{\xi\in s}E\left(I_{\xi}\right)E\left(J_{\xi}\right). \quad \text{(B.7)}$$

From (B.7) it follows that an estimator of the covariance term is given by

$$\hat{Cov}\left(n_{ik}^{v},n_{lj}\right)=n^{v}\,\hat{\Pr}\left(Y_{\xi t}^{*}=l,Y_{\xi t+1}^{*}=j,Y_{\xi t}^{*}=i,Y_{\xi t}=k\right)-n^{v}\frac{\hat{E}\left(n_{ik}^{v}\right)}{n^{v}}\frac{\hat{E}\left(n_{lj}\right)}{n^{v}} \quad \text{(B.8)}$$

where

$$\begin{cases}\hat{\Pr}\left(Y_{\xi t}^{*}=l,Y_{\xi t+1}^{*}=j,Y_{\xi t}^{*}=i,Y_{\xi t}=k\right)=0 & if\ \ l\neq i\\ \hat{\Pr}\left(Y_{\xi t}^{*}=l,Y_{\xi t+1}^{*}=j,Y_{\xi t}^{*}=i,Y_{\xi t}=k\right)\neq 0 & if\ \ l=i\ .\end{cases}$$

In order to complete the proof, we further need to evaluate the following expression

$$\frac{E\left(n_{ik}^{v}\right)}{E\left(\sum_{i=1}^{r}n_{ik}^{v}\right)}Cov\left(n_{lj},\sum_{i=1}^{r}n_{ik}^{v}\right).$$

This can be done as follows

$$\frac{E\left(n_{ik}^{v}\right)}{E\left(\sum_{i=1}^{r}n_{ik}^{v}\right)}Cov\left(n_{lj},\sum_{i=1}^{r}n_{ik}^{v}\right)=\frac{E\left(n_{ik}^{v}\right)}{E\left(\sum_{i=1}^{r}n_{ik}^{v}\right)}\left[Cov\left(n_{lj},n_{1k}^{v}\right)+Cov\left(n_{lj},n_{2k}^{v}\right)+\ldots+Cov\left(n_{lj},n_{rk}^{v}\right)\right]. \quad \text{(B.9)}$$

The covariance terms in (B.9) are estimated using the result below

$$\hat{Cov}\left(n_{ik}^{v},n_{lj}\right)=\begin{cases}n^{v}\,\hat{\Pr}\left(Y_{\xi t}^{*}=i,Y_{\xi t+1}^{*}=j,Y_{\xi t}^{*}=i,Y_{\xi t}=k\right)-\dfrac{\hat{E}\left(n_{ik}^{v}\right)\hat{E}\left(n_{lj}\right)}{n^{v}} & if\ \ l=i\\ -\dfrac{\hat{E}\left(n_{ik}^{v}\right)\hat{E}\left(n_{lj}\right)}{n^{v}} & if\ \ l\neq i\ .\end{cases} \quad \text{(B.10)}$$

Combining (B.8) with (B.9) and (B.10) and using Result B.1 we obtain the required result

APPENDIX C: DIAGONAL ELEMENTS OF Q USED IN SECTION 6

Table C.1: Probabilities of correct classification used in applications of Section 6

Probabilities of Correct Classification*	**Original**	**Modified ("Intense Misclassification")**
q_{EE}	0.981	0.981
q_{U+NU+N}	0.978	0.95

***** The probabilities of misclassification are defined as $1-\Pr\left(Correct\ \ Classsification\right)$.

References

Abowd, M.J. and Zellner, A. (1985). Estimating Gross Labour Force Flows, *Journal of Business and Economic Statistics*, 3, 254-283.

Bailar, A.B. (1968). Recent Research in Re-interview Procedures, *Journal of the American Statistical Association*, 63, 41-63.

Bauman, K.E. and Koch, G.C. (1983). Validity of Self-reports and Descriptive and Analytical Conclusions: The Case of Cigarette Smoking by Adolescents and Their Mothers, *American Journal of Epidemiology*, 118, 90-98.

Bishop, Y., Fienberg, S., and Holland, P. (1975). Discrete Multivariate Analysis, MIT Press.

Bross, I. (1954). Misclassification in 2×2 Tables, *Biometrics*, 10, 478-486.

Carroll, R.J. (1992). Approaches to Estimation with Error in Predictors, in *Advances in GLIM and Statistical Modelling, Fahrmeir, L., Francis, B., Gilchrist, R. and Tutz, G. (eds)*, 40-47, Springer-Verlag.

Deming, W.E. and Stephan, F.F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known, *Annals of Mathematical Statistics*, 11, 427-444.

Dempster, P.A., Laird, M.N. and Rubin, B.D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society Series B*, 39, 1-38.

Dennis, E.J. and Schnabel, B.R. (1983). Numerical Methods for Unconstrained Optimisation and Nonlinear Equations, Precentice-Hall Series in Computational Mathematics.

Espeland, A.M. and Odoroff, L.C. (1985). Log-Linear Models for Doubly Sampled Categorical Data Fitted by the EM Algorithm, *Journal of the American Statistical Association*, 80, 663-670.

Forsman, G. and Schreiner, I. (1991). The Design and Analysis of Re-interview: An Overview, in *Measurement Error in Surveys, P.P. Biemer, R.M. Grooves, L.E. Lyberg, N.A. Mathiowetz, S. Sudman (eds)*, 279-301, Wiley.

Greenland, S. (1988). Variance Estimation for Epidemiologic Effect Estimates Under Misclassification, *Statistics in Medicine*, 7, 745-757.

Hogue, C.R. and Flaim, P.O. (1986). Measuring Gross Flows in the Labour Force: An Overview of a Special Conference, *Journal of Business and Economic Statistics*, 4, 111-121.

Kristianson, K-E. (1999). Estimation of Gross Flows in LFS, *Internal Report*, 1-24 Statistics Sweden.

Kuha, J. and Skinner, C. (1997). Categorical Data Analysis and Misclassification, in *Survey Measurement and Process Quality, Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, Trewin (eds)*, 633-670, Wiley.

Louis, T.A. (1982). Finding the Observed Information Matrix When Using the EM Algorithm, *Journal of the Royal Statistical Society Series B*, 44, 226-233.

Meyer, D.B. (1988). Classification Error Models and Labour-Market Dynamics, *Journal of Business and Economic Statistics*, 6, 385-390.

Mood, A.M., Graybill, A.F. and Boes, C.D. (1963). "Introduction to the Theory of Statistics", McGraw-Hill.

Poterba, J.M. and Summers, L.H. (1986). Reporting Errors and Labour Market Dynamics, *Econometrica*, 54, 1319-1338.

Singh, A.C. and Rao, J.N.K. (1995). On the Adjustment of Gross Flows Estimates for Classification Error with Application to Data from the Canadian Labour Force Survey, *Journal of the American Statistical Association*, 90, 478-488.

Skinner, C. and Torelli, N. (1993). Measurement Errors and the Estimation of Gross Flows From Longitudinal Economic Data, *Statistica*, 3, 391-405.

Skinner, C.J. and Humphreys, K. (1997). Instrumental Variable Estimation of Gross Flows in the Presence of Measurement Error, *Survey Methodology*, 23, 53-60.

Skinner, C.J. (2000). Dealing with Measurement Error in Panel Analysis, in Researching Social and Economic Change –The Uses of Household Panel Surveys, Rose, D. (eds), 113-125, Routledge.

Tanner, M.A. (1996). Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, Springer.

Tenenbein, A. (1970). A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications, *Journal of the American Statistical Association*, 65, 1350-1361.

Tenenbein, A. (1972). A Double Sampling Scheme for Estimating from Misclassified Multinomial Data with Applications to Sampling Inspection, *Technometrics*, 14, 187-202.

Tzavidis, N. and Lin, Y-X. (2004). Estimating from Cross-sectional Categorical Data Subject to Misclassification and Double Sampling: Moment-based, Maximum Likelihood and Quasi-Likelihood Approaches, *Methodology Series Working Papers, S3RI, University of Southampton*, M04/03, (http://www.s3ri.soton.ac.uk/publications/methodology.php).

Van de Pol, F. and De Leeuw, J. (1986). A Latent Markov Model to Correct for Measurement Error, *Sociological Methods and Research*, 15, 118-141.