UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS & SOCIAL SCIENCES
School of Social Sciences
Division of Social Statistics

ON THE USE OF DOUBLE SAMPLING SCHEMES TO CORRECT FOR MEASUREMENT ERROR IN DISCRETE LONGITUDINAL DATA

by

Nikolaos Tzavidis

Thesis for the degree of Doctor of Philosophy

January 2004

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS & SOCIAL SCIENCES

SCHOOL OF SOCIAL SCIENCES

DIVISION OF SOCIAL STATISTICS

Doctor of Philosophy

ON THE USE OF DOUBLE SAMPLING SCHEMES TO CORRECT FOR MEASUREMENT ERROR IN DISCRETE LONGITUDINAL DATA

by Nikolaos Tzavidis

Longitudinal surveys provide a key source of information for analysing dynamic phenomena. Typical examples of longitudinal data are gross flows, which are defined as transition counts between a finite number of states from one point in time to another. There are, however, a number of methodological problems associated with the use of longitudinal surveys. This thesis focuses on the measurement error problem or more naturally in a discrete framework on the misclassification problem.

We investigate the use of double sampling for correcting discrete longitudinal data for misclassification. In a double sampling context, we assume that along with the main measurement device, which is affected by misclassification, we can use a secondary measurement device (validation survey), which is free of error but more expensive to apply. Due to its higher cost, the secondary measurement device is employed only for a subset of units. Inference, using double sampling, is based on combining information from both measurement devices.

Traditional moment-based inference is reviewed and alternative moment-type estimators, which attempt to overcome the drawbacks of the traditional approach, are proposed. We subsequently argue that a more efficient parameterisation is offered in a likelihood-based framework by simultaneously modeling the true transition process and the measurement error process within the context of a missing data problem. Variants of likelihood-based inference, which allow for alternative double sampling schemes, for a complex survey design and for observed heterogeneity, are investigated. Constrained maximum likelihood estimation is also considered for relaxing some of the model assumptions. Variance estimation for the moment-type and the likelihood-based estimators is illustrated. In addition, empirical research aimed at identifying optimal design characteristics for validation surveys is presented.

The methodology is applied in the context of the UK Labour Force Survey (LFS) by estimating labour force gross flows adjusted for misclassification. Results from Monte-Carlo simulation experiments indicate that the proposed likelihood-based parameterisation offers significant gains in efficiency over the traditional moment-based parameterisation while interval estimation for the adjusted estimates can be reliably performed using the proposed variance estimators.

Contents

Chapter 2: Using Double Sampling for Misclassification Error Correction: From a Cross-sectional to a Panel Framework

Chapter 3: Likelihood-based Inference for Gross Flows in the Presence of Misclassification and Double Sampling

Chapter 4: Likelihood-based Inference for Gross Flows in the Presence of Misclassification, Double Sampling and Heterogeneity Associated with Discrete Covariates

Chapter 5: Variance Estimation for Gross Flows Estimates in the Presence of Misclassification and Double Sampling

Chapter 6: Monte-Carlo Evaluation

Chapter 7: A Note on the Design of a UK LFS Re-interview Survey: Suggestions Based on Empirical Evidence

Chapter 8: Summary and Suggestions for Further Research

List of Tables

List of Figures

Acknowledgments

I am deeply grateful to Professor Ray Chambers for his supervision, encouragement and patience during my PhD study. Apart from teaching me statistics and research methodology, Ray has also taught me the way in which to control my stress by always viewing the positive side of things.

I feel that the friendly, research oriented environment of the Division of Social Statistics and of the Southampton Statistical Sciences Research Institute at the University of Southampton has been very beneficial. Thanks to the Division and the Institute, I was able to attend the Research Student's Conference in Newcastle (2001) and deliver papers in Warwick (2002) and at the Conference of the International Statistical Institute (ISI) in Berlin (2003). Acknowledgements are also due to Professor Fred Smith and to Professor Danny Pfeffermann. Their support was really valuable.

Whilst undertaking this research, I was fortunate to visit the University of Wollongong in New South Wales, Australia and the Australian Bureau of Statistics in Canberra, Australia. The interaction with Professor David Steel (University of Wollongong) benefited my research while Dr Yan-Xia Lin (University of Wollongong) inspired me to start investigating the use of quasi-likelihood methods.

I would like to thank the UK Office for National Statistics (ONS) for providing me with data from the UK Labour Force Survey. Thanks are also due to Dr Alison Whitworth (ONS) who organised three methodological seminars (February 2002, October 2002 and September 2003) in which I presented my work. I would also like to thank Statistics Sweden for providing me with data from the Swedish Labour Force Survey re-interview programme.

Special thanks are due to my colleagues and friends both in the Division of Social Statistics and in the Southampton Statistical Sciences Research Institute. Some of them successfully managed to distract me on Friday afternoons (despite my attempts to hide). I am referring to my co-players in VIM (participating in the 5-a-side Veterans Football League and in particular to Dr David Holmes) and to the members of the Friday Night Club.

A big thank you is due to my family. My parents (Theodoros and Elpida), my sister (Pepi) and my aunt (Froso) have always supported me throughout my studies as well as my parents in law (Giorgos and Betty). But among all, the biggest THANK YOU is due to my wife Kyriaki-Pipitsa. Her support, patience and encouragement were the most valuable help during these years.

The Economic and Social Research Council (ESRC) jointly with the Division of Social Statistics, University of Southampton financially supported the work in this thesis through scholarship S42200034035.

To my wife

Kyriaki-Pipitsa Noussia

Key to Notation

U : Population

S : Sample

N : Population size

n: Sample size of the main survey (common sample of the panel survey between t and $t+1$)

n^v : Sample size of the validation survey

$Y^*_{\xi t}$: Random variable that describes the way that unit ξ is classified at time t by the fallible classifier

$Y_{\xi t}$: Random variable that describes the way that unit ξ is classified at time t by the error free classifier

$Y^*_{\xi t \to t+1}$: Random variable that describes the fallible flow of unit ξ between t and $t+1$

$Y_{\xi t \to t+1}$: Random variable that describes the true flow of unit ξ between t and $t+1$

P : Matrix that describes the probability distribution of the true classifications. A superscript u denotes this matrix at the population level

P^m : Matrix that describes the probability distribution of the true classifications in the main sample

P^v: Matrix that describes the probability distribution of the true classifications in the validation sample

Π: Matrix that describes the probability distribution of the observed classifications. A superscript u denotes this matrix at the population level

Π^m : Matrix that describes the probability distribution of the observed classifications in the main sample

Π^v: Matrix that describes the probability distribution of the observed classifications in the validation sample

Q: Matrix of misclassification probabilities. A parenthesis next to Q indicates the time periods to which the misclassification matrix refers. A superscript u denotes this matrix at the population level

$Q^{(c)}$: Matrix of misclassification probabilities. This matrix is calibrated to respect specific margins. A parenthesis next to $Q^{(c)}$ indicates the time periods to which the misclassification matrix refers. A superscript u denotes this matrix at the population level

C : Matrix of calibration probabilities. A parenthesis next to C indicates the time periods to which the matrix of calibration probabilities refers.

Φ : Total number of groups when accounting for heterogeneity

g : This symbol, used as a subscript or superscript, refers to a specific group when accounting for heterogeneity

α_g : Fraction of sample units that belong to group g

w_ξ : Survey weight for sample unit ξ

w_{mod} : Weight when using the modified estimator

w_{comp} : Weights (fixed or adaptive) when using the composite estimator

Θ : Vector of parameters

$l(\Theta)$: Log-likelihood function

D^m : Observed data from the main sample

D^v : Observed data from the validation sample

Z^m : Missing data in the main sample

Z^v : Missing data in the validation sample

D^c : Complete data

X : Design matrix

ω : Dimension of the parameter space

G : Quasi-score estimating function

H : Total number of iterations when using an iterative procedure

(h) : A specific iteration of an iterative procedure

δ : Small value that defines the convergence criterion when using iterative algorithms

$(*)$: This symbol, appearing as a superscript, denotes unobserved quantities

$\partial F(x) / \partial x$: Denotes the derivative of $F(x)$ with respect to x

∇ : Jacobian matrix

$\otimes$: Kronecker product

vec : Vector operator

E : Expectation operator

Var : Variance operator

Cov : Covariance operator

$|$: Symbol that denotes conditioning

$pr(A)$: Denotes the probability of event A

Chapter 1

Background to the Problem, Literature Review and Description of the Data Sources

1.1 Introduction, Aim and Structure of the Thesis

Gross flows are generally defined as transition counts, between a finite number of states, from one point in time to another. In the same manner, labour force gross flows represent the transition counts of the labour force population between the different labour force states. In the simplest case, labour force gross flows can be represented by a 3×3 gross flows matrix. The dimension of this matrix can be justified if we assume that a member of the labour force population can be classified only to one of the following mutually exclusive states: a) being employed, which is denoted by **(E)**, b) being unemployed, which is denoted by **(U)** and c) being economically inactive, which is denoted by **(N)**. If someone belongs to one of the first two categories then he/she is assumed as being a member of the labour force otherwise as being out of the labour force. Schematically, the gross flows table between two time points, say t and $t+1$, can be represented as follows.

Table 1.1: Labour force gross flows between t and $t+1$

	(E)	*(U)*	*(N)*
(E)	EE	EU	EN
(U)	UE	UU	UN
(N)	NE	NU	NN

The diagonal elements of the gross flows table (Table 1.1) represent the number of individuals that remain stable between t and $t+1$. The off-diagonal cells describe the number of individuals that change labour force status between t and $t+1$.

The aim of this thesis is to develop methodology for adjusting gross flows for measurement error. The application will be in the estimation of labour force gross flows. The approach we follow assumes the existence of validation data and the theory is based on the use of double sampling methods. Chapter 1 provides an overview of problems encountered in estimating gross flows. Modelling strategies to correct for measurement error are reviewed and classified into strategies that assume validation data and strategies that do not assume validation data. A description of the data sources that are used in this thesis is given and a review of the literature on validation surveys is provided. Chapter 2 describes the estimation framework of double sampling. Double sampling methods in a cross-sectional framework are contrasted with double sampling methods in a longitudinal framework. New research results on the analysis of cross-sectional misclassified data are presented. In a longitudinal framework, some new moment-type estimators are presented. Chapter 3 focuses on likelihood-based inference. The measurement error model is formulated, under alternative double sampling designs, in a missing data framework and model parameters are estimated via maximum likelihood. Alternative likelihood-based inference is examined by relaxing some of the model assumptions. The measurement error model is further extended to account for the existence of a complex survey design. The methodology is illustrated by calculating estimates of UK labour force gross flows adjusted for measurement error. In Chapter 4, the measurement error model is extended to account for heterogeneity in the gross flows mechanism and in the measurement error mechanism. The effect of measurement error on inference based on labour force gross flows is examined using data from the UK Labour Force Survey (LFS). Chapter 5 deals with variance estimation issues. Variance estimators for the moment-type and the maximum likelihood estimators are developed and illustrated using UK LFS data. In Chapter 6, we evaluate our methodology by a series of Monte-Carlo simulation experiments. In Chapter 7, we give recommendations for designing a UK LFS re-interview survey and for selecting an appropriate double sampling design. Chapter 8 summarises the research outcomes and sets directions for future research.

1.2 The Importance of Labour Force Gross Flows for Social and Economic Research

Labour force gross flows are indicators frequently used in social and economic research. In this section, we describe some of the applications with labour force gross flows. To start with, let us consider the following situation. A fall in unemployment is the net result of a larger number of individuals moving between the different labour force states. However, this net result, which is often estimated and published, is based on a series of individual gross flows. These flows can be approximated only by linking each individual's labour force activity in successive time points. Figure 1.1 shows the flows between the main labour force states of economic activity and economic inactivity. For example, the (EU) arrow refers to the number of people who moved from employment to unemployment between t and $t+1$. A more complete description of the gross flows must take into account the dynamic evolution of the population (Figure 1.2). Taking this dynamic evolution into consideration, we observe that in addition to the usual flows we have inflows and outflows, which are distributed between the different labour force states.

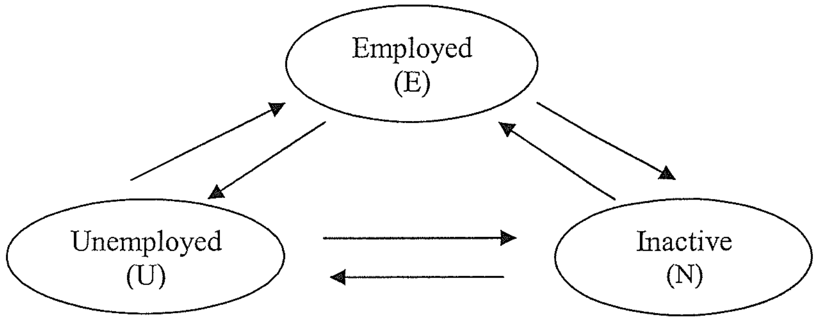


Figure 1.1: Simple model

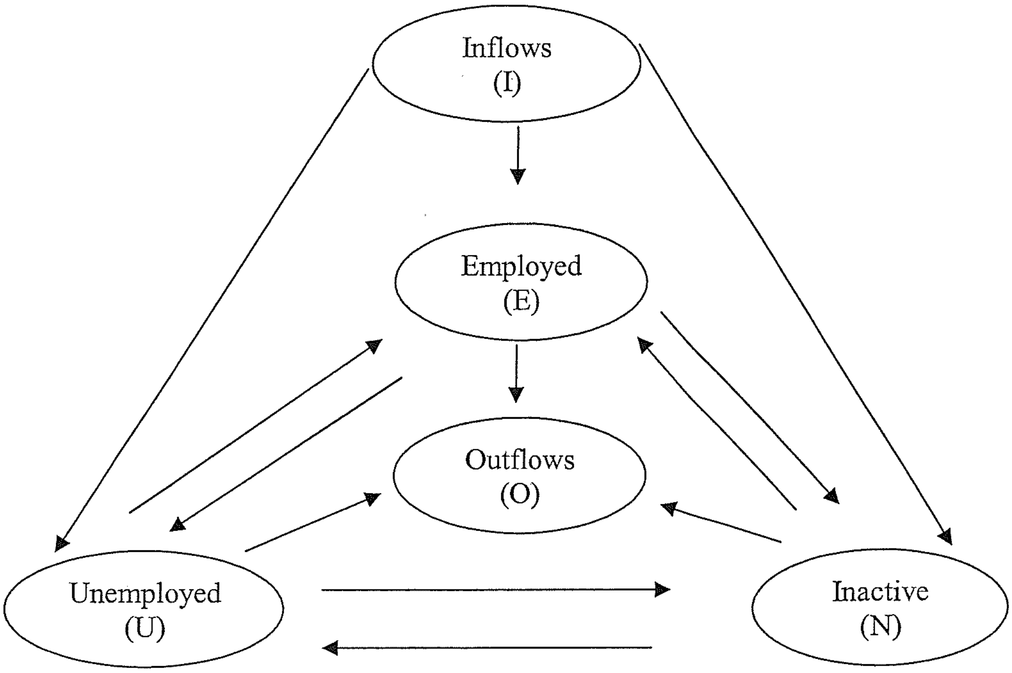


Figure 1.2: Complete model

How can these transitions be used in economic analysis? There are occasions when both the labour force participation and unemployment rise. Are these events attributable to a greater inflow of job seekers from outside the labour force or to reduced exits from the labour force? Gross flows provide a way to examine, for example, how many workers enter or leave the labour force or how many move from employment to unemployment (Barkume and Horvath 1995).

Gross flows can be interpreted as measures of the labour market condition. Consider the labour force entries to employment (NE flows). The magnitude of these flows depends on both labour force participation decisions and the demand for labour (i.e. the number of job prospects). Another example is the flows from employment to unemployment (EU flows). These transitions characterise recession periods. A further example concerns the labour force exits from unemployment (UN flows). The variations in these flows have been used to measure discouraged worker effects in business downturns (Hansen 1961).

The transition probabilities implied by labour force gross flows can be used to calculate several summary statistics for the labour market activity. Summary statistics of this type include the expected duration of a complete spell in each labour market state, the probability of an unemployment spell ending in employment and the probability of labour force withdrawal. For example, the expected duration of completed spells is calculated as the reciprocal of the exit probability from each state. The probability of an unemployment spell ending in an employment entry, given that a transition has occurred, is given by $\frac{\Pi_{UE}}{\Pi_{UE}+\Pi_{UN}}$ where Π_{UE} denotes the probability that an individual is moving from employment to unemployment and Π_{UN} denotes the probability that an individual is moving from unemployment to inactivity. The probability of a labour force withdrawal is given by $\frac{\Pi_{UN}\Pi_{U}+\Pi_{EN}\Pi_{E}}{\Pi_{U}+\Pi_{E}}$ where Π_{U} and Π_{E} denote respectively the probability that an individual is unemployed or employed (Poterba and Summers 1986).

Gross flows can be used in order to establish for how long people who have previously been in a governmental training for work scheme remain in a job. In the same field of research, gross flows allow the evaluation of different training programmes. This can be done by comparing the labour market progress of non-participants in such programmes with programme participants. In addition, by analysing the transition probabilities associated with the different job search techniques it becomes possible to evaluate these techniques and to explore the extent to which movements from unemployment into part time and temporary work act as a stepping-stone to full time employment (Atkinson and Micklewright 1991). Labour market transitions have been used in literature for studying the effects of unemployment compensation. Questions of the following type have been explored. Do cuts in unemployment benefit compensation increase the rate of exit from unemployment but cause people to leave the labour force rather than to enter employment? Does the existence of unemployment insurance lead job losers registering as unemployed rather than leaving the labour force? Does unemployment compensation provide the security that allows people to give up their jobs and acquire training? (Atkinson and Micklewright 1991). A further application of the gross flows can be in equal opportunities monitoring. This can be done by comparing people from ethnic minorities with others. Gross flows can be also used for assessing the stability over time of the labour market movements of people with disabilities and of those who receive sickness or disability benefits. For other applications using the labour force gross flows see Akerlof and Main (1980), Burda and Wyplosz (1994) and Jones and Riddell (1998).

1.3 Panel Surveys versus Retrospective Surveys and Registration Systems in Longitudinal Data Collection

Panel surveys can be regarded as the most natural way of collecting longitudinal information. Among the alternative ways of collecting longitudinal data, rotating panel designs have a prominent role. In fact, rotating panel designs have been adopted by most national Labour Force Surveys. Under a rotating panel design, sample units are interviewed in consecutive time periods usually months, quarters or years. In the simplest case, each time period corresponds to one wave of the panel survey and the interviewed sample units report their labour force state for the current period. Using such designs, one can obtain information on labour force gross flows by matching data of individuals who participate in the survey for two or more successive waves.

Alternatively, longitudinal information may be collected using retrospective surveys. This may happen by introducing retrospective questions in a cross-sectional survey. For example, in the US Survey of Income and Programme Participation (SIPP) interviews take place every four months but monthly data are collected by questioning respondents about their behaviour during the past four months. In the UK LFS (Laux and Tonks 1996), the retrospective way of collecting longitudinal information co-exists with the panel one. The UK LFS contains two sets of recall questions. The first set of recall questions is asked for respondents who move into a sampled household after the household's first interview (wave 1) and for those respondents who have worked for the same employer or who have been self-employed for less than three months. These questions are related to respondents' circumstances three months prior to the current interview and explore respondents' occupation and full-time/part-time status. The second set of recall questions is based on a twelve-month period and is asked for all respondents in the spring quarter (i.e. March to May) each year. These questions cover the same topics as the ones of the first set with the addition of information about respondents' managerial duties and the number of staff at the place of employment.

Data that are produced in a retrospective way may be affected more severely by measurement error compared to data where respondents report their status for the current period. There are a number of reasons for this to happen. Firstly, the nature of measurement error is likely to be more complex when panel data are constructed by piecing together retrospective histories. This may happen because of memory problems such as forgetting or mistiming as well as because of the respondents' or the interviewers' misunderstand. Moreover, errors in reconstructing event histories can lead to dependencies between measurement errors for example, "seam" effects where more change may be observed between measurements recorded in different interviews than for measurements recorded within the same interview (Hill 1987, Marquis and Moore 1990, Kalton and Miller 1991). Furthermore, there are significant problems regarding the consistency of definitions of the different labour force states. Assume for example the ILO definition for unemployment. This definition includes those who are out of work, available to start work within two weeks following the interview and have either looked for work in the four weeks prior to the interview or they are waiting to start a job they have already obtained. The implementation of this definition is easy in the context of a survey that collects information about activities within a reference week. However, it is unrealistic to seek information in such a detailed level about respondents' past labour activities.

The panel way of collecting longitudinal data provides a much richer source of information. The main advantage is that under a panel design the information recorded refers to the same time period that the interview takes place. Consequently, this way of collecting data minimises memory problems. For this reason, the panel way of collecting longitudinal data is the most promising one and that is why it is widely used. Nevertheless, panel surveys are affected by a number of factors that complicate estimation of the quantities of interest. One problem associated with panel surveys is the increased risk of attrition. This problem can be attributed to the increased risk of failing to follow all sample units throughout the period during which they are supposed to participate in the survey. When attrition is truly random, the only problem that is created is the loss of efficiency for the estimators. However, in many cases attrition is non-random and the resulting estimates are severely biased. Another problem associated with panel surveys is the impact of measurement error on the estimates of change. Typical examples of this problem include the overestimation of the labour market mobility and of the poverty dynamics. Last but not least, a further complication arising in panel surveys is the presence of conditioning effects. Conditioning effects occur when the behaviour of a respondent is affected by the number of times that this respondent has participated in the survey. A general review on the problems associated with panel surveys can be found in (Duncan 2000).

An alternative source of longitudinal information is registration systems. Such systems include (a) simple systems, which allow only flows in and out of the register to be recorded and (b) longitudinal systems, where individuals can be traced over periods of time as they leave and re-enter the register. Theoretically, registration systems are capable of producing true flow statistics since only registers can record every movement between the different labour force states. In addition, registers can be regarded as having perfect memories. However, these advantages are weakened if the registers are not properly updated.

Labour force gross flows statistics stem from the longitudinal character of the Labour Force Surveys (LFS). The dynamics of the labour market have presented a challenge to researchers in the United States (US) and in Canada since the 40's and 50's. Since that time researchers have recognised the importance of studying the labour market mobility and also the problems encountered in the estimation of labour force gross flows. The status of research of the labour market dynamics outside North America is described by Evans (1985). As he points out, in the 80's very few countries had published flows from household-based Labour Force Surveys. For example, in the US Current Population Survey (CPS) the base of rotation is the month and the rotation pattern is 4-8-4. This implies that the sample units are followed for four consecutive months, subsequently they are dropped out from the sample for eight months and eventually they are included again in the sample for four more months. The following table (Table 1.2) reports the countries that utilised labour force sample surveys appropriate for estimating labour force gross flows.

Table 1.2 Rotation sampling schemes of labour force sample surveys in nine countries (Evans 1985).

Countries	*Base of Rotation*	*Pattern of Rotation*
Australia	Month	8-
Canada	Month	6-
Finland	Quarter	6-
France	Year	3-
Italy	Quarter	2-2-2
Japan	Month	2-10-2
Spain	Quarter	6-
Sweden	Quarter	8-
US	Month	4-8-4

In the past years, large-scale longitudinal surveys of socio-economic conditions and behaviour of households have been established in several European countries including Belgium, Germany, Greece, Ireland, Luxembourg, The Netherlands and UK.

1.4 Problems Encountered in Estimating Labour Force Gross Flows

Labour force gross flows are estimated by linking together panel data. In this process several problems are typically encountered. Since 1953, two presidential committees in the US have recommended that the problem of gross change estimation should be studied. In 1962, the President's Committee to Appraise Employment and Unemployment Statistics under the direction of Robert A. Gordon urged that the problems discovered in gross change estimation should be thoroughly researched so that publication of the data could then be resumed. In 1978, the National Commission on Employment and Unemployment Statistics headed by

Professor Sar Levitan reviewed a paper by Ralph Smith and Jean Vanski entitled "Gross Change: The Neglected Data Base." This paper examined the potential uses of data, research that has been done using the data and errors in the data. According to this paper the main problems encountered in the estimation of labour force gross flows include (a) sampling attrition, (b) response errors and (c) rotation group bias or conditioning effects. General descriptions of these problems can be found in Hilaski (1968), Hogue (1985), Hogue and Flaim (1986), Clarke and Tate (1999) and Kristiansson (1999). Literature on non-sampling errors includes among others Bailar (1987), Barnes (1987), Trewin (1987) and Lessler and Kalsbeek (1992). In the sequel, we describe each of the problems in detail and we examine their implications for estimation of labour force gross flows.

1.4.1 Sampling Attrition

Attrition bias is one of the problems affecting the estimation of gross flows. For example, the UK LFS is based on a sample of addresses each of which is occupied by a household or, less commonly, by multiple households. The aim is to interview every eligible household member. However, when a household is approached, non-response can occur due to outright refusal, circumstantial refusal or non-contact. Outright refusal occurs when a household, or the individual from whom permission is sought, refuses to participate in the survey. Circumstantial refusal is less terminal, arising when the household does not agree to be interviewed for example, because the timing is inconvenient. The third category of interview non-response i.e. non-contact refers to the situation where it is not possible to contact an eligible household member. In the event of an outright refusal no further attempt is usually made to interview that household and consequently it is dropped out of the survey. If permission to interview a household is obtained, individuals can still non-respond because of refusal, non-contact or because other household members are unwilling or ineligible to provide proxy responses. A further complication arises due to the rotation design. This means that sampled households leave the sample after having participated in the survey for a specific number of times. Furthermore, after the first wave, families or persons may move away from the sampled addresses. The effect on the gross change data is magnified because addresses where people move in and others out cannot be used until a mover has been there for two consecutive waves. Sample units who do not match from one wave to another have generally different characteristics from the matched sample units. Hilaski (1968) points out that those who move are generally young and either married with small family or single. Empirical work (Tate 1997) verified these results by showing that a higher proportion of sample losses are associated with younger adults living in privately rented accommodation and being temporarily employed.

How can attrition affect the estimation of the gross flows? It is possible for people who move away from their addresses to be associated with a higher probability of changing labour force status. The loss of these people may lead to a downward bias in the estimates of the gross flows. For example, a family may move because an unemployed person has found job in a new location. Although this transition should have contributed to the UE cell of the transition matrix (see Table 1.1), in fact it will not since movers are not followed. Another example is students moving from a university town to other cities when looking for a job. These transitions would have contributed to the NE or NU cell. It is apparent that these losses may lead to the estimation of lower market mobility than really exists (Kristiansson 1999).

Research investigating methods of adjusting for sampling attrition can be found in Stasny and Fienberg (1985), Stasny (1986), Clarke and Chambers (1998) and Clarke and Tate (1999). Stansy and Fienberg (1985) and Stasny (1986) propose the use of models that allow for non-ignorable non-response. In the context of the UK LFS, Clarke and Chambers (1998) also propose models that account for non-ignorable non-response but, in addition, they extend their models to hold at the household level. In the same context, Clarke and Tate (1999) compare the model-based approach with the weighting approach for adjusting for non-response and conclude that the weighting approach can provide a good alternative. Based on the work by Clarke and Chambers (1998) and Clarke and Tate (1999), a weighting scheme has been developed for the UK LFS that aims also at correcting for attrition bias. The variables on which this weighting scheme is based are age, gender, tenure and region.

1.4.2 Rotation Group Bias

Rotation group bias is a further problem that complicates the estimation of labour force gross flows. Rotation group bias occurs when the number of times respondents have been exposed to the survey affects the data reported. It is hypothesised that the estimates for each of the rotation groups of a panel survey must have the same expected values. This means that it must be possible to regard the response stratum in each of the rotation groups as randomly generated from a common survey population by the same mechanism. Moreover, the measurement process must function in the same way regardless of the time that the sample members are interviewed.

Studies of the CPS have revealed systematic differences in the estimates based on different rotation groups. Empirical work (Bailar 1975) has shown that unemployment rates estimated from the CPS are higher for the first and the fifth month, decrease for the intermediate months and increase slightly for the fourth and eighth month. The effect of rotation group bias on the estimates of employment rate has been also studied by Hansen, Hurwitz, Nisselson and Steinberg (1955). Furthermore, Solon (1985) discusses different forms of rotation group bias and the effects that this phenomenon has on the estimates of unemployment. In Great Britain, Kemsley (1961) and Turner (1961) examined the presence of rotation group bias in expenditure surveys and found that the reported expenditures were higher in the first interview than in later interviews.

Rotation group bias can be attributed to the telescoping phenomenon. Telescoping means that a respondent may recall an event that happened some months ago, but state that it happened more recently. Events that are more traumatic are more likely to be reported than the real events. Think of the rotation design in the CPS. This is a 4-8-4 design. We suspect that months one and five may contain more telescoping, either because people have never been to the survey before or they have been out of the survey for eight months and they want to report something of interest. The intermediate months are probably less affected by this phenomenon since respondents have the opportunity to report on a regular basis. Other reasons that cause rotation group bias may include the change of the mode of data collection from the first to subsequent waves or the conditioning, i.e. participants in a panel survey learn a shortcut through the questionnaire, of the respondents after the initial interview.

1.4.3 Measurement Error

Every sample survey is subject to response error when the information given by the respondent is not an accurate reflection of the reality. The existence of response error implies the existence of measurement error. When dealing with discrete data, the term measurement error can be replaced by the more natural term misclassification. Hereinafter, the terms misclassification and measurement error will have an identical meaning and will imply the existence of response error. Response error may occur for a variety of reasons. A respondent may deliberately give an incorrect answer for reasons of embarrassment, prestige, or fear (Hilaski 1968, Hogue 1985). Further potential sources of error include the use of proxy respondents where one respondent answers the questions on behalf of someone else in the same household and the mode of data collection i.e. the use of face to face or telephone interviews (Tate and Clarke 1999). Other causes of misclassification reported in Kristiansson (1999) include problems in the questionnaire and difficulties in the classification of a respondent's status.

A considerable amount of literature deals with the effects of misclassification on hypothesis testing and the measures of association. Rogot (1961) studies the effects of misclassification on the Type II errors (i.e. not rejecting the null hypothesis when it should be rejected). The author restricts his interest in four specific patterns of misclassification. He concludes that the misclassification patterns of the specific type that he describes tend to increase the probability of making a Type II error. This result is consistent with the description of Kuha and Skinner (1997), (see also Buell and Dunn 1964, White 1986), that the effect of misclassification in the multivariate case is to attenuate the differences between the subclass proportions. Kuha and Skinner (1997) study the effects of misclassification also for univariate analysis. They conclude that bias is a function both of the misclassification probabilities and of the true parameters and can take any arbitrary form. Thus, an instrument with a given misclassification can lead to biased estimates in one population and unbiased estimates in another. Mote and Anderson (1965) investigate the effects of misclassification on the properties of X^2 tests for goodness of fit and for contingency tables. They conclude that the effect of misclassification is to reduce the power of these tests. Koch (1969) studies the effects of non-sampling errors on different measures of association in 2×2 tables. Chiacchierini and Arnold (1977) develop a test for independence in 2×2 tables under misclassification. They conclude that ignoring misclassification can result in erroneously rejecting the null hypothesis and vice versa. Other papers dealing with the effects of misclassification on the analysis of categorical data include Bross (1954) and Assakul and Proctor (1967) whilst two more general papers, on the effects of measurement error on the analysis of survey data, are given by Cochran (1963) and by Biemer and Trewin (1997).

It is believed that for cross-sectional data there is no particular tendency for errors to be systematic (Veevers and Macredie 1983, Lemaitre 1999, Skinner 2000). However, for longitudinal data produced by linking together data collected for the same person in different time points, this cancellation may not occur. We investigate this argument using the following approach. Denote by Q the misclassification matrix. The diagonal elements of this matrix denote the probabilities of correct classification and the off-diagonal elements the probabilities of misclassification. Denote further by P a matrix that describes the probability distribution of the true classifications, by Π a matrix that describes the probability distribution of the observed (i.e. affected by measurement error) classifications and by I the identity matrix. Under misclassification, is reasonable to assume that bias is introduced in the estimation of P. The bias can be quantified as follows:

$$Bias(P) = \Pi - P. \tag{1.1}$$

Using simple matrix operations, (1.1) can be expressed in the following way

$$Bias(P) = (Q - I)P. \tag{1.2}$$

Note that bias becomes zero when $Q = I$ i.e. when no misclassification exists. Let us assume that we are dealing with an example in the context of the UK LFS. Assume further that there are three mutually exclusive states namely Employment (E), Unemployment (U) and Inactivity (N). Following (1.2), the bias introduced in the estimation of the proportion of people that belong to each labour force category is given below

$$\begin{bmatrix} Bias(P_E) \\ Bias(P_U) \\ Bias(P_N) \end{bmatrix} = \begin{pmatrix} q_{EE}-1 & q_{EU} & q_{EN} \\ q_{UE} & q_{UU}-1 & q_{UN} \\ q_{NE} & q_{NU} & q_{NN}-1 \end{pmatrix} \begin{pmatrix} P_E \\ P_U \\ P_N \end{pmatrix}. \tag{1.3}$$

We are interested in finding whether there exists a combination of values such that (1.3) is equal to zero. A set of values for which (1.3) becomes zero is the following.

$$P_E = 0.7071, P_U = 0.0606, P_N = 0.2323 q_{EE} = 0.99, q_{UE} = 0.003, q_{NE} = 0.007$$
$$q_{EU} = 0.04, q_{UU} = 0.85, q_{NU} = 0.11, q_{EN} = 0.02, q_{UN} = 0.03, q_{NN} = 0.95.$$

This combination of values can be considered as a realistic one in a labour force framework. Thus, in a cross-sectional framework it is possible to obtain unbiased estimates of the parameters of interest even in the presence of misclassification.

In a longitudinal context, Skinner (2000) provides a similar example in which, while the marginal estimates of the gross flows matrix are unbiased, the estimates of the transition probabilities are seriously biased. For example, many researchers believe that the response errors can have serious implications for estimation of labour force gross flows. Think of the gross flows between the main labour force states (Employment, Unemployment and Inactivity). The number of people who move from one state to another during a relatively short period is small compared to the number of people who remain stable. Consequently, a response error is much more likely to lead to an apparent change when the true situation is one of stability. This implies that response errors can have an effect by upwardly biasing the flows between the different labour force states. For this reason, the response error problem and methods that attempt to correct for response error have been at the centre of research in the US and Canada during the 80's and in Europe mainly during the last decade. Generally speaking, the major aim of this thesis is to develop methodology that adjusts labour force gross flows for misclassification. As a result, from now on we will focus our interest on misclassification related issues.

1.5 Modelling Strategies to Correct for Misclassification

Misclassification can introduce bias in the estimation of the parameters of interest and as a result in the analysis and inference based on these parameters. Consequently, it is of interest to investigate modelling strategies that attempt to correct for the biasing effect of misclassification. In the following sections, we describe techniques that have been developed for both cross-sectional and longitudinal data. Before doing so, however, we should mention that due to the discrete nature of the gross flows, conventional methods of errors in variables modelling (Fuller 1987) are not applicable. Generally speaking, these modelling strategies can be placed into two broad categories (see Figure 1.3): (a) strategies that assume the existence of validation information and (b) strategies that do not require validation information.

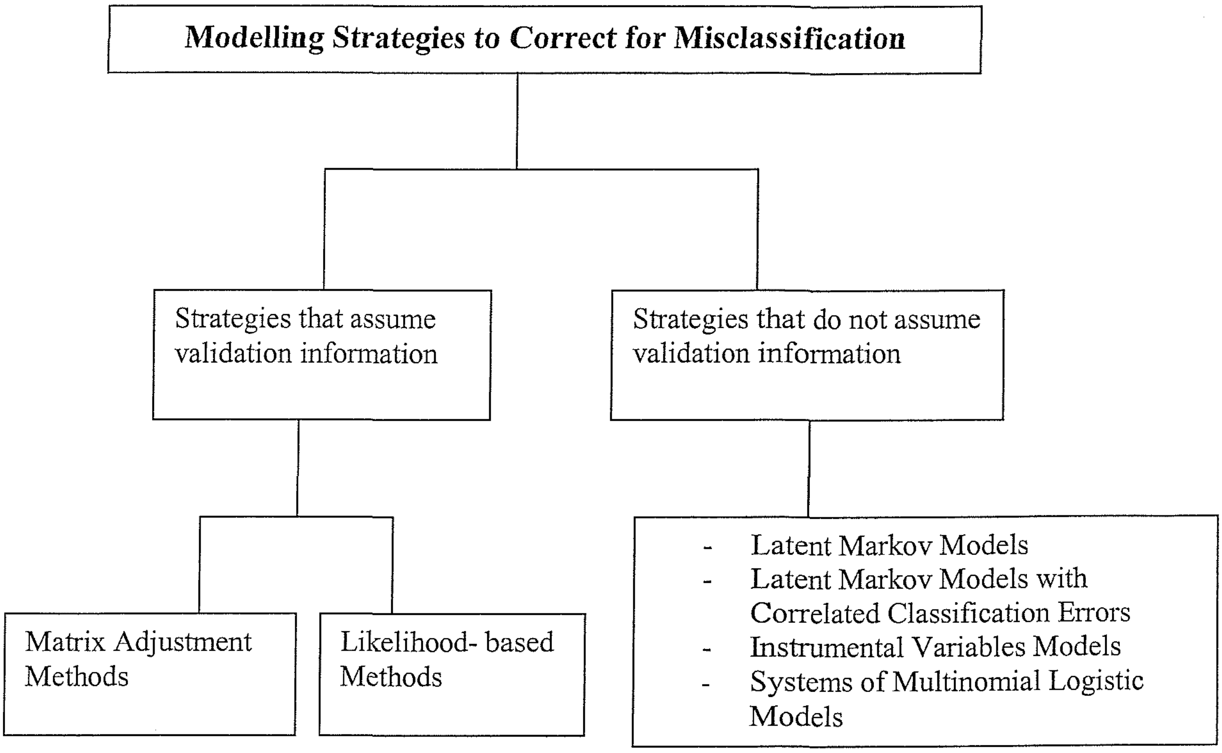

Modelling Strategies to Correct for Misclassification
Strategies that assume validation information
Strategies that do not assume validation information
Matrix Adjustment Methods
Likelihood- based Methods
- Latent Markov Models
- Latent Markov Models with Correlated Classification Errors
- Instrumental Variables Models
- Systems of Multinomial Logistic Models

Figure 1.3: Modelling strategies to correct for measurement error in a discrete framework

Although the alternative modelling strategies utilise different parameterizations, they share many common characteristics. In all approaches, the observed and the true classifications are interrelated using the misclassification mechanism. The observed classifications can then be expressed as a function of the parameters of the misclassification mechanism and of the true classifications and vice versa. As we will see later in this chapter, expressing the observed classifications as a function of the true classifications and the parameters of the misclassification mechanism is the approach that is generally adopted by the likelihood-based strategies either in the presence of validation data or when no validation data are available. The reverse approach is employed by the matrix adjustment methods. In the sequel, we describe the alternative modelling strategies.

1.5.1 Strategies that Require Validation Information

1.5.1.1 Matrix Adjustment Methods

The term matrix adjustment methods appears in Kuha and Skinner (1997) and is used to describe simple methods that provide adjustments for measurement error via matrix computations. We should mention that these methods do not specify any parametric model for estimating quantities adjusted for measurement error. Matrix adjustment techniques can be placed into the general framework of the double sampling methods. Assume that the standard measurement device (e.g. the Labour Force survey) is subject to measurement error. As a result, if the fallible measurement device is used, we will have biased results. One way of obtaining unbiased estimates is by using validation information obtained through a double sampling scheme. Literature on validation surveys is reviewed later in this chapter. Here we describe the general framework.

Denote by $Y^*_{\xi t}$ a random variable describing the observed state (i.e. affected by measurement error) of unit ξ at time t and $Y_{\xi t}$ a random variable describing the hypothetical true state of the same unit at time t. The estimation process under a double sampling scheme can be described as follows.

A random sample of n units is selected from a population of N units and

1. For the n units selected, the classifications, $Y^*_{\xi t}$, are obtained for each unit ξ using the standard measurement device, which is subject to measurement error.
2. Following this first measurement, the true classifications, $Y_{\xi t}$, are obtained for each unit ξ in a sub-sample of n^v units, selected from the n units, using the validation procedure.

Generally speaking, the double sampling methods try to combine information from both the true and the fallible classifiers in order to obtain adjusted estimates. The basic assumption of this approach is that the validation procedure identifies the true value. Consequently, using the validation information one can exogenously estimate the parameters of the misclassification mechanism and then adjust the quantities of interest for measurement error.

Matrix adjustment methods were developed initially for cross-sectional applications. Bross (1954) describes the application of such methods to adjust binomial data (proportions) for misclassification. Tenenbein (1970) derives maximum likelihood estimators and asymptotic variances for proportions adjusted for misclassification and extends these results to the multinomial case (Tenenbein 1972).

Similar methods have also been used in the analysis of longitudinal misclassified data. Literature that deals with the adjustment of labour force gross flows for misclassification using matrix adjustment methods includes Abowd and Zellner (1985), Poterba and Summers (1986), Chua and Fuller (1987), Skinner and Torelli (1993) and Singh and Rao (1995). The general set up for longitudinal applications is as follows. Assume a panel survey is conducted and a sample unit ξ is interviewed at two consecutive periods $t, t+1$. We assume that the variable of interest measured by the panel survey is subject to misclassification. Let $Y_{\xi t}^{*}$ denote an observed measurement (i.e. affected by measurement error) and $Y_{\xi t}$ an error free measurement for the same quantity. The pairs $\left(Y_{\xi t}^{*}, Y_{\xi t+1}^{*}\right)$ for different sample units are assumed to be *iid* with distribution $\Pi_{ij} = pr\left(Y_{\xi t}^{*} = i, Y_{\xi t+1}^{*} = j\right)$. The pairs $\left(Y_{\xi t}, Y_{\xi t+1}\right)$ for different sample units are assumed to be *iid* with distribution $P_{kl} = pr\left(Y_{\xi t} = k, Y_{\xi t+1} = l\right)$. We assume that we can use validation information to make inference about the probability of misclassification. However, the validation survey is conducted only at time t. Denote by $q_{ijkl} = pr\left(Y_{\xi t}^{*} = i, Y_{\xi t+1}^{*} = j \mid Y_{\xi t} = k, Y_{\xi t+1} = l\right)$ the probability that a person is observed as making a transition from state i at time t to state j at time $t+1$ when his/her actual transition is from state k to state l. We further define the matrix of misclassification probabilities Q with elements q_{ijkl}. A parenthesis next to matrix Q will be used to define the time that the misclassification refers to. For example, $Q(t, t+1)$ is used to denote the joint misclassification matrix at these two time points. Finally, we define matrix P with elements P_{kl} and matrix Π with elements Π_{ij}. The vec operator will be used throughout this thesis to define a vector obtained by stacking the columns of a matrix one on top of the other. Assuming that $Q(t, t+1)$ is invertible, the adjusted gross flows are derived using the following expression

$$vec(P) = \left[Q(t, t+1)\right]^{-1} vec(\Pi). \tag{1.4}$$

Due to the absence of panel validation information, one way to determine the joint misclassification matrix $Q(t, t+1)$, based on cross-sectional validation data, is by introducing the Independent Classification Errors (ICE) assumption, i.e.

$$pr\left(Y_{\xi t}^{*} = i, Y_{\xi t+1}^{*} = j \mid Y_{\xi t} = k, Y_{\xi t+1} = l\right) = pr\left(Y_{\xi t}^{*} = i \mid Y_{\xi t} = k\right)\ pr\left(Y_{\xi t+1}^{*} = j \mid Y_{\xi t+1} = l\right). \tag{1.5}$$

This ICE assumption implies that the observed states $Y_{\xi t}^{*}, Y_{\xi t+1}^{*}$ are conditionally independent given the true states $Y_{\xi t}, Y_{\xi t+1}$ and that the misclassification at time t depends only on the current true state and not on the previous or future true states. The ICE assumption and ways of relaxing it will be discussed in subsequent chapters. Under ICE, (1.4) becomes

$$vec(P) = \left[Q(t+1) \otimes Q(t)\right]^{-1} vec(\Pi). \qquad (1.6)$$

The main difference between articles dealing with the adjustment of gross flows for measurement error is in the estimation of the parameters of the misclassification mechanism.

Poterba and Summers (1986) Approach

While investigating validation data from the reconciled[1] sub-sample and the unreconciled[1] sub-sample of the CPS validation survey, Poterba and Summers observed an anomaly. More specifically, they found that the discrepancies between the original survey and the re-interview survey were much greater for the unreconciled sub-sample than for the reconciled one. Due to this problem, the authors assumed that the reconciled sub-sample gives information on the true labour force status while the unreconciled sub-sample can be used for estimating the actual incidence of error. Regarding individuals in the reconciled sub-sample, they estimate the probability that a respondent truly belongs in each labour force category conditional upon the initial and the unreconciled reported status. To determine these probabilities they assume that when there is an inconsistency between the two survey responses the reconciliation procedure correctly identifies the true status. Denoting by $Y_{\xi t}^{(r)}$ the classification obtained through the reconciliation process for sample unit ξ and by $Y_{\xi t}^{(u)}$ the classification obtained from the unreconciled part of the re-interview sample for the same unit, these probabilities are given by

$$b_{ijk} = pr(Y_{\xi t}^{(r)} = k \mid Y_{\xi t}^{*} = i, Y_{\xi t}^{(u)} = j) \;\; subject \; to \; b_{ijE} + b_{ijU} + b_{ijN} = 1, \quad i,j = E,U,N. \quad (1.7)$$

Subsequently, they use b_{ijk} to impute a probability distribution of the true labour force status for each individual in the unreconciled sub-sample conditional upon the responses in the initial and the unreconciled part of the re-interview survey. Using the imputed distribution of the true labour force status, two probabilities can be determined: the probability that an

[1] Usually, a validation survey is divided into reconciled and unreconciled sub-samples. The reconciled sub-sample is assumed to provide the true values. The unreconciled sub-sample replicates the survey process.

individual's recorded initial interview response is i conditional upon the imputed reconciliation status k, q_{ik} and the probability that an individual's recorded interview response in the unreconciled sample is j conditional upon the imputed reconciliation status k, q_{jk}. Using these two sets of probabilities, the error rates are computed by averaging q_{ik} and q_{jk}. The final misclassification probabilities can be substituted in expression (1.6) in order to adjust the quantities of interest for measurement error.

Abowd and Zellner (1985) Approach

The common characteristics between the Poterba and Summers approach and the Abowd and Zellner approach, for adjusting gross flows for measurement error, are the following: (a) both assume the availability of validation information and employ the independent classification errors assumption to estimate the matrix of misclassification probabilities and (b) both assume that the reconciliation process identifies the true labour force status. The difference between these two approaches is that while the Poterba-Summers approach utilises data both from the reconciled and the unreconciled sub-sample, the Abowd and Zellner approach utilises data only from the reconciled sub-sample. Furthermore, the model proposed by Abowd and Zellner simultaneously adjusts labour force gross flows for bias due to attrition and response error.

Chua and Fuller (1987) Approach

Chua and Fuller (1987) proposed a parametric approach for estimating the response error matrices using validation information that is derived from the unreconciled part of the validation sample only. However, they still use the ICE assumption in order to estimate the longitudinal structure of the misclassification mechanism.

Matrix Adjustment Methods that Attempt to Relax the ICE Assumption

Poterba and Summers (1986), Skinner and Torelli (1993) and Singh and Rao (1995) examined similar adjustment techniques. However, they extend the ICE assumption to hold within sub-populations. This is the so-called unit heterogeneity approach. Under this approach, (1.6) becomes

$$vec\left(P\right)=\left[\sum_{g=1}^{\phi}\alpha_g\left(Q\left(t+1\right)_g\otimes Q\left(t\right)_g\right)\right]^{-1}vec\left(\Pi\ \right), \qquad (1.8)$$

where ϕ denotes the total number of sub-groups, α_g denotes the fraction of people that belong to sub-group g and $Q(t)_g$ denotes the misclassification matrix for sub-group g. The use of the unit heterogeneity approach reduces the effects of the ICE assumption. As pointed out by Skinner and Torelli (1993), the effect of ignoring unit heterogeneity, when such heterogeneity is present, leads to over-adjustments of the labour force gross flows. However, they believe that the bias introduced in the estimation of the adjusted gross flows when ignoring unit heterogeneity is not large.

In the same framework, an alternative adjustment method, which has been adopted by many researchers (Poterba and Summers 1986, Singh and Rao 1995), is the unbiased margins method. Under this approach, it is assumed that the margins of the adjusted gross flows matrix must agree with the observed margins at $t, t+1$. This implies that cross-sectional estimates remain unbiased in the presence of measurement error. As we illustrated in Section 1.3.3, this is possible in a cross-sectional framework.

1.5.1.2 Likelihood-based Methods in the Presence of Validation Information

Using the matrix adjustment methods we can obtain estimates adjusted for misclassification. However, we cannot use standard methods of statistical inference (hypothesis testing, model selection) since these methods do not account for the extra uncertainty introduced by the adjustment procedure. This type of inference can be performed using likelihood-based methods.

There is a considerable literature on likelihood-based methods for adjusting for misclassification in a cross-sectional framework when validation data are available. Chen (1979) examines estimation from double sampling designs using models that are placed into a log-linear framework and specified at two levels. First, a misclassification model that specifies the relationship between the true and the misclassified variables is formulated. Using this model, we can test for the existence of a differential or a non-differential misclassification mechanism. Secondly, a model for the relationship between the true variables is specified. Using this system of models, one can utilise the misclassification structure to adjust for misclassification and then investigate the relationships among the correct classifications. Chen (1979) estimates these models using a recursive system of maximum likelihood equations.

Hochberg (1977) considers models for doubly sampled data and proposes two alternative estimation methods namely, maximum likelihood estimation and a combination of least squares and maximum likelihood estimation.

Espeland and Odoroff (1985) present models for doubly sampled data. They assume that the variable of interest is measured both by a precise and an imprecise device. They further assume that there are other variables (covariates), which are measured precisely. They specify three different types of models namely, the sampling model that describes the relationship between all variables included in the model, the misclassification model that describes the relationship between the precise and the imprecise measured variable, and the experimental model that describes the relationship between the precise variable and the precisely measured covariates. These models are estimated using the EM algorithm. Ekholm and Palmgren (1987) present also models that correct doubly-sampled data for misclassification.

The work presented by Hochberg (1977), Chen (1979), Espeland and Odoroff (1985) and Ekholm and Palmgren (1987) deals with cross-sectional data. Poterba and Summers (1995) formulate a longitudinal model for doubly sampled labour force related data that focuses on transitions from unemployment to employment and inactivity. They model the probability of an actual transition as a multinomial logistic model. The likelihood defined by the model is written as a function of the misclassification probabilities and the true transition probabilities. In the maximisation of this likelihood they assume that the misclassification probabilities are fixed at estimated values derived from the validation sample. This is equivalent to treating these misclassification probabilities as nuisance parameters. Consequently, the likelihood is maximised only with respect to a reduced set of parameters. As the authors recognise, the maximisation of the resulting conditional likelihood leads to inconsistent estimates of the standard errors since the process ignores the variability introduced from the estimation of the misclassification probabilities. An additional assumption that they impose is that the probability of misclassification is independent of the characteristics of the respondents. This is equivalent to assuming the existence of a non-differential misclassification mechanism and can be considered as quite restrictive. Their final assumption is that the observed classifications at the first time point are free of error. This might be justified given the nature of the data considered in the article. However, in most applications this assumption can not be regarded as realistic. Assuming that the observed classifications at the first time point are free of error simplifies the estimation process. This is done as follows: The observed data are derived from a panel survey whereas the validation data are derived from a cross-sectional survey. Assuming that the observed classifications at the first time point are free of error is equivalent to transforming the measurement error process from a longitudinal to a cross-sectional one. Hence, there is no need to impose the ICE assumption.

1.5.2 Strategies that do not Require Validation Information

The main objection to adjustment procedures that assume the existence of validation information is the ability to measure the truth. In the case of an external validation sample (e.g. based on administrative records), this becomes the question of how informative this source of information is about the misclassification process in the target population. For example, a validation study based on the employees in one company can provide no information on the probability of an unemployed person being classified as employed. In the case of an internal validation sample, based on re-interviewing a sub-sample of units, the main problem is the measurement of the truth via this re-interviewing process. Despite the fact that re-interviews that aim at obtaining the true values are designed to be optimal in terms of the survey procedures, they still have deficiencies.

A class of models has been developed, which does not estimate the parameters of the misclassification mechanism by attempting to measure the truth but by replicating the measurement process. Generally speaking, these models involve the combination of a true model that relates the true values at different waves and a measurement model that relates the true values to the measured (misclassified) values. Without imposing further assumptions the parameters of these models are not identified. This is because these modelling strategies do not assume validation information. To achieve identification we either have to add information provided by repeated measurements or to impose assumptions on the relationships between the variables. Such assumptions can be for example that (i) the observed measurements are conditionally independent given the true measurements, (ii) the measurement error depends only on the current state and not on previous or future true states, (iii) the distribution of error is homogeneous through time, (iv) the distribution of error is homogeneous across or within subpopulations, (v) the true values follow a Markov process within subpopulations. Given such assumptions, these models can be identified and the parameters (e.g. the transition probabilities) can be estimated using maximum likelihood estimation. Models of this type include latent Markov models, instrumental variables models and systems of multinomial logistic models.

1.5.2.1 Latent Markov Models

Latent structure analysis was developed by Lazarsfeld and Henry (1968) in order to solve problems involving unobserved variables when the data are measured at the nominal level. In one of their models they assume panel data and a latent Markov chain underpinning the observed (manifest) data. In this context, Van de Pol and De Leeuw (1986) proposed a latent Markov model to correct data from the Dutch civil servants panel survey for measurement error.

Formulating the Manifest Structure

Assume that a set of consecutive measurements is obtained by a measurement device that is affected by measurement error. Recalling the notation form previous sections, a Markov chain is specified by the initial distribution $\Pi_i = pr\left(Y^*_{\xi t} = i\right)$ and a set of transition matrices R with elements R_{ij} for transitions from i to j between t and $t+1$. The relationship between Π_i, R_{ij} and Π_{ij} is given by

$$\Pi_{ij} = \Pi_i R_{ij}, \tag{1.9}$$

where Π denotes a diagonal matrix with elements only in the diagonal and zeros elsewhere. The main assumption of the Markov model is that a transition matrix R is independent of the past states through which the process has passed. This implies that a transition matrix say for two consecutive periods $(t,t+1)$ and $(t+1,t+2)$ satisfies the following relationship

$$R(t,t+2) = R(t,t+1)R(t+1,t+2). \tag{1.10}$$

The length of periods $(t,t+1)$ and $(t+1,t+2)$ can be assumed to be the same as the time between consecutive waves of the panel survey but this is not a necessary assumption. Using relationships (1.9), (1.10) and assuming that we obtain measurements at three consecutive time points, the probability that a person belongs to manifest cell (i,j,k) is given by

$$\Pi_{ijk} = \Pi_{ij}\Pi_{jk} = \Pi_i R_{ij} R_{jk}. \tag{1.11}$$

In many cases an assumption of stationary transition probabilities is imposed.

$$R(t,t+1) = R(t+1,t+2) = \cdots = R. \tag{1.12}$$

In most of the cases, however, the Markov assumption is not met by the data. The approach adopted by the authors for relaxing the Markov assumption is to decompose the manifest data into latent data and error. This naturally leads to the formulation of the latent structure.

Formulating the Latent Structure

Corresponding to every manifest variable one latent variable is assumed. The distribution of the manifest variable Π depends on the distribution of the corresponding latent variable P and a matrix of transition probabilities Q. This matrix is equivalent to the misclassification matrix used by the strategies that require validation information. The diagonal elements of Q denote the probabilities of correct classification (i.e. the reliability by which a latent class is measured). The latent variable is related with the manifest variable, via matrix Q, by the following relationship.

$$\Pi = QP. \tag{1.13}$$

Apart from the relationship between the latent and the manifest variables, there is also a structure for the latent variables similar to the structure of the manifest variables. Consequently, on a latent level and for three time points $(t,t+1,t+2)$, there are $Y_{\xi t}, Y_{\xi t+1}, Y_{\xi t+2}$ latent observations for sample unit ξ, which are interrelated by a Markov chain. The latent transition matrix is denoted by M and the latent initial distribution by $P_a = pr\left(Y_{\xi t} = a\right)$. The probability that a person belongs to latent class (a,b,c) is given by

$$P_{abc} = P_a M_{ab} M_{bc}. \tag{1.14}$$

In order to relate the manifest variables to the latent variables, an assumption of local independence is being made. This means that the manifest variable at time t depends only on the latent variable at time t (see Figure 1.4). Thus, for a respondent in cell (a,b,c) the probability of answering (i,j,k) is given by the following expression

$$\Pi_{ijk} = \sum\nolimits_a \sum\nolimits_b \sum\nolimits_c P_a q_{ai} M_{ab} q_{bj} M_{bc} q_{ck}. \tag{1.15}$$

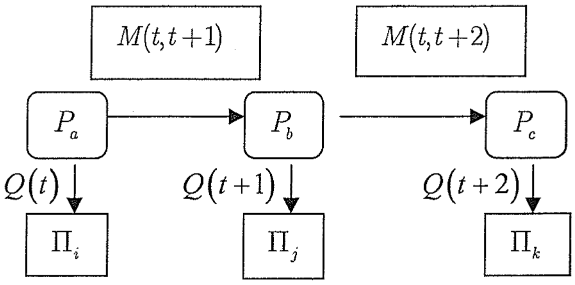


Figure 1.4: An example of a latent Markov chain in discrete time

Estimation

As proposed by Van de Pol and De Leeuw (1986), the vector of parameters Θ of the latent Markov model can be estimated using the EM algorithm. Assuming a multinomial model and denoting by n_{abcijk} the data for each cell of the cross-classification of the manifest and the latent variable, the log-likelihood for the latent Markov model is given by

$$l(\Theta) = \sum\nolimits_a \sum\nolimits_b \sum\nolimits_c \sum\nolimits_i \sum\nolimits_j \sum\nolimits_k n_{abcijk} \log\left(P_a q_{ai} M_{ab} q_{bj} M_{bc} q_{ck}\right). \tag{1.16}$$

In the E-step, the latent observations are replaced by their conditional expectations given the current vector of parameter values and the observed data. In the M-step, the likelihood is maximised and new parameter values are computed. The authors provide the steps for the EM algorithm under the assumption of stationary transition probabilities given in (1.12). However, the EM algorithm can be modified in order to relax this assumption. An alternative way of estimating the parameters of the latent Markov model is by attempting a direct maximisation of the likelihood function (1.16) using numerical methods (Haberman 1979). However, Hagenaars (1985) points out that Haberman's algorithm requires very good starting values otherwise it will not converge.

1.5.2.2 Latent Markov Models that Allow for Correlated Classification Errors

Bassi, Torelli and Trivellato (1998) describe latent class models for estimating labour force gross flows affected by classification errors when data are partially collected using retrospective questions. They start by describing a general estimation framework where the joint probability of the observed and the true classifications can be marginalized over the latent (true) classifications and expressed as a product of conditional and marginal probabilities. In this paper, the emphasis is on latent class models that allow also for correlated classification errors i.e. relaxing the ICE assumption. A suitable approach for handling such models is the so-called modified LISREL approach proposed by Hagenaars (1990).

The authors present two case studies; one from the Survey of Income and Programme Participation (SIPP) and another from the French LFS. The common characteristic of these surveys is that data are collected partially using retrospective questions. Gross flows can be separated into Within-Wave (WW) flows, which are estimated using the retrospective part of the survey, and Between-Wave (BW) flows. While for the (BW) transitions ICE can be regarded as a reasonable assumption, for the (WW) transitions it is more reasonable to assume that they are affected by correlated classification errors. The authors noticed that (BW) flows show a lower stability while (WW) flows show higher stability. The higher stability of (WW) flows can be attributed to "seam effects" i.e. more change is observed when data are collected in different interviews than when they are collected in the same interview. The proposed model corrects the (WW) transitions towards higher mobility i.e. reducing "seam effects" and (BW) transitions towards stability.

Magnac and Visser (1999) study transition models with measurement error when information is gathered partially by using retrospective questions. More specifically, they use data from the French LFS in which the sample units are interviewed for three times. At each interview the survey participants are asked to report their current labour force status and also their labour force status month by month in the preceding twelve months. The modelling assumptions that they impose are the following: (a) the labour market histories are generated using a discrete-time Markov chain, (b) the observed and the true states are related using a measurement error mechanism, (c) the current reported state is assumed to be free of error while the retrospective reported states are affected by classification error and (d) errors increase linearly with time due to recall problems. The assumption that the currently reported state is free of error is unrealistic. This is because the currently reported data are derived through an ordinary interviewing process i.e. not derived, for example, via a validation procedure. However, it might be the case that the retrospectively derived data are more severely affected by measurement error. Regarding assumption (b), Magnac and Visser used the so-called d-ICE assumption, which can be seen as a relaxed ICE assumption. The d-ICE assumption states that the misclassifications recorded at time t and $t+d$, $d>1$, are independent.

The vector of model parameters Θ is estimated by maximising a log-likelihood function. Denoting by n_{ij} the number of individuals observed in state i at time t and state j at time $t+d$, by k,l the true states and by r the total number of states, the log-likelihood function is given by the following expression

$$l(\Theta)=\sum_{t=1}^{T-d}\sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij}\log\left(\Pi_{ij}\right). \tag{1.17}$$

It can be shown that under the d-ICE assumption

$$\Pi_{ij}=\sum_{k=1}^{r}\sum_{l=1}^{r} q_{ik}P_{kl}q_{jl}, \tag{1.18}$$

where q_{ik} and q_{jl} are elements of the misclassification matrices $Q(t)$ and $Q(t+d)$ respectively. The misclassification matrices $Q(t)$ and $Q(t+d)$ are estimated under assumption (c) and the assumption that recall error increases linearly with time. After estimating $Q(t)$ and $Q(t+d)$, the log-likelihood (1.17) is maximised and maximum likelihood estimators for the true transition probabilities, P_{kl}, are derived. The model proposed by Magnac and Visser corrects the gross flows towards stability, which contradicts the findings from the model of Bassi et al. (1998). Bassi and Trivellato (2000) criticise the assumptions imposed by Magnac and Visser and re-analyse the data using the modified LISREL modelling approach, which allows for correlated classification errors. Their model corrects the retrospectively collected flows towards higher mobility.

1.5.2.3 Instrumental Variables Estimation

For measurement error in continuous variables an approach employed in the absence of auxiliary information is the method of instrumental variables estimation. An instrumental variable is one that is related to the true variable but is uncorrelated with the measurement error. Skinner and Humphreys (1997) (see also Humphreys 1996) extend the instrumental variables model to estimate flows among discrete states that are affected by classification errors. The paper focuses on the case of a binary variable. A discrete instrumental variable W is defined, which is correlated with the true variable but uncorrelated with error. The following assumptions are made: (a) the instrumental variable is conditionally independent of the observed states given the true states, (b) the instrumental variable is conditionally independent of the true state at the second time point given the true state at the first time point, (c) the classification errors at two occasions are conditionally independent given the true states, (d) the measurement errors are unbiased at each occasion in the sense that the margins of the adjusted gross flows should equal the margins of the observed gross flows matrix and (e) the error process is constant over time.

Denote by $\Pi_{ijk} = pr\left(Y_{\xi t}^{*} = i, Y_{\xi t+1}^{*} = j, W = k\right)$ and by n_{ijk} the number of individuals in the ijk combination defined by the cross-classification of the observed variable with the instrumental variable. Expressing Π_{ijk} as a product of conditional probabilities using assumptions (a)-(e), the vector model parameters Θ can be estimated by maximising the following multinomial log- likelihood

$$l(\Theta) = \sum_{i}\sum_{j}\sum_{k} n_{ijk} \log\left(\Pi_{ijk}\right). \tag{1.19}$$

The maximisation of (1.19) can be done either directly or by using software that fits latent class models.

The authors describe an application that involves the selection of an actual instrumental variable and they point out problems related to the choice of an instrumental variable that obeys both assumptions (a) and (b). They argue that it is more difficult to find an instrumental variable that satisfies the second assumption than one that satisfies the first assumption. Subsequently, they investigate models using two different instrumental variables i.e. one, which is highly correlated with the error free variable and one that is less correlated with the error free variable. The results indicate that the estimates obtained using the instrumental variable that is less related with the error free variable are associated with higher standard errors. Skinner and Humphreys (1997) investigate the trade-off between the bias of the unadjusted estimates and the increased variance of the instrumental variables estimates assuming that the instrumental variables estimates are unbiased. They point out that the variance of the instrumental variables estimates increase as the dependence between the instrumental variable and the error free variable, measured by Cramer's V, decreases. However, they conclude that in the absence of external information about the misclassification probabilities the instrumental variables approach can be very useful.

1.5.2.4 Systems of Multinomial Logistic Models

An alternative approach to adjusting labour force gross flows for measurement error is proposed by Pfeffermann, Skinner and Humphreys (1998) (see also Pfeffermann and Tsibel 1998). This approach utilises multinomial logistic models that are specified at the unit level for both the transition and the classification probabilities. The combination of these models yields an overall model for the observed flows, which permits the identification of the true transitions. The advantages of this approach are that no validation data are required and that the ICE assumption can be relaxed by including the previously observed states as covariates in the models. The following assumptions are made: (a) the observed classifications at different time points are dependent given the corresponding true values and covariates, (b) the true classifications are dependent but they do not depend on past observed states and (c) the initial state probabilities do not depend on past observed states. Assumptions (a) and (b) impose a Markovian structure on the observed and the true state probabilities. Utilising the previous assumptions and denoting by $X_{\xi t}$ the covariate information for sample unit ξ at time t, the initial state probabilities are given by $P_k = pr\left(Y_{\xi 1} = k \mid X_{\xi 1}\right)$, the misclassification probabilities are given by $q_{jl} = pr\left(Y^*_{\xi t} = j \mid Y_{\xi t} = l, Y^*_{\xi t-1} = i, X_{\xi t}\right)$, and the true transition probabilities are given by $P_{kl} = pr\left(Y_{\xi t} = l \mid Y_{\xi t-1} = k, X_{\xi t}\right)$. The joint distribution of the observed and the true states can now be expressed as a function of the misclassification probabilities and the true transition probabilities. The authors postulate separate multinomial logistic models for the misclassification probabilities, the true transition probabilities and the initial state probabilities. Generally speaking, the different parameters are expressed as follows:

$$q_{jl} = \frac{\exp(Xa)}{\{1+\exp(Xa)\}},\ P_{kl} = \frac{\exp(X\beta)}{\{1+\exp(X\beta)\}},\ P_k = \frac{\exp(X\gamma)}{\{1+\exp(X\gamma)\}}, \qquad (1.20)$$

where X denotes the design matrix and α, β, γ denote the vector of parameters of the logistic models. Note also that there is no restriction for the different models to include the same set of covariates. The authors suggest that direct maximisation of the likelihood should be avoided due to the possibly high dimensionality of the problem. Instead, they propose the use of the EM algorithm. The approach proposed by Pfeffermann, Skinner and Humphreys (1998) is a very important one since the ICE assumption is directly relaxed by including the previously observed states as covariates in the models. One difficulty with this approach is the computation of standard errors for the parameters of interest.

1.6 A Critical Comparison of the Alternative Modelling Strategies

For each of the alternative modelling strategies we need to specify the structure of the observed classifications, the structure of the true classifications and the way that the observed classifications are related to the true classifications via the measurement error mechanism. The alternative methods differ with respect to the approach they choose to estimate the measurement error mechanism. This depends on the availability of validation information. Nevertheless, in terms of modelling assumptions, the different methods share common characteristics.

Comparing the matrix adjustment methods with the latent Markov approach, we see that both approaches use the local independence assumption to estimate the parameters of the measurement error mechanism. However, the latent Markov approach results in the computation of maximum likelihood estimates, which may be considered as more efficient than the estimates obtained via the matrix adjustment methods. On the other hand, the lack of validation information in the latent Markov approach imposes some extra constraints. For example, in order that the parameters of the latent Markov model are identified, we need to utilise linked data for at least three quarters and to impose assumptions about stationary transition probabilities.

Comparing the instrumental variables model with the matrix adjustment methods, we also find common modelling assumptions. For example, assumption (c) in the instrumental variables model is equivalent to the ICE assumption of the matrix adjustment methods and assumption (d) is equivalent to the unbiased margins assumption that is also utilised by the

matrix adjustment methods (Poterba and Summers 1986, Singh and Rao 1995). However, the instrumental variables model also results in maximum likelihood estimates.
The latent Markov models with correlated classification errors and the systems of multinomial logistic models can be viewed as a separate group of methods since they attempt to relax the local independence assumption in the specification of the measurement error model. These methods can be seen as similar, in the sense of trying to relax the local independence assumption, to the unit heterogeneity approach and the unbiased margins approach of the matrix adjustment methods.

One of the major aims of this thesis is to develop likelihood-based methods for adjusting gross flows data in the presence of validation information. Generally speaking, the models induced by these methods can be parameterised in the same way as the models that do not require validation information (e.g. the latent Markov model). This means that the observed transition probabilities can be expressed as a function of the true transition probabilities and the misclassification probabilities and estimation can be performed within the context of a missing data problem. However, when validation information is available, one can avoid introducing the whole range of assumptions utilised by the modelling strategies that do not require validation information.

1.7 The UK Labour Force Survey (LFS)

In this section, we describe the main source of data that we will use for illustrating the theory throughout this thesis namely, the UK LFS.

1.7.1 Historical Notes and Purposes of the UK LFS

The UK LFS is a survey of households living at private addresses, which is conducted by the Social Survey Division (SSD) of the Office for National Statistics (ONS) in Great Britain and by the Central Survey Unit of the Department of Finance and Personnel in Northern Ireland. The first LFS in UK was conducted in 1973. Between 1973 and 1983 the survey took place every two years in the spring quarter. Between 1984 and 1991 the survey was carried out annually and consisted of two elements: (a) a quarterly survey of approximately 15000 private households and (b) a "boost" survey, in the quarter between March and May, of over

44000 private households in Great Britain and 5200 households in Northern Ireland. Quarterly LFS estimates for Great Britain became possible in 1992 when the sample was increased to cover 60000 households every quarter. The LFS quarters refer to the seasonal quarters March-May (spring), June-August (summer), September-November (autumn) and December-February (winter). Whilst the quarterly LFS is built on the annual one, there are some differences mainly regarding the response rates (response rates in quarterly LFS are lower due to the cumulative refusal across waves), the sampling design (introduction of an un-clustered design) and the target population (inclusion of people in two categories of non-private accommodation i.e. in National Health Service (NHS) accommodation and students in halls of residence).

The main purpose of the quarterly LFS is to provide information needed to develop, manage, evaluate and report on labour market policies. One potential use of the quarterly LFS is in macro-economic monitoring. Main indicators regularly published from the LFS include total employment, the unemployment rate and the economic activity rate. A further important use of the LFS is for the production of regional statistics. Based on regional data, governmental offices can assess the local labour markets and design future labour market policies. Further purposes of the LFS include the monitoring of the characteristics of the unemployed people, the gathering of information related to training and qualifications, the monitoring of the youth labour market, the gathering of information on income related variables, the monitoring of working conditions and working related accidents and also the gathering of information related to participation in trade unions.

1.7.2 Survey Design Issues

Coverage and Sampling Design

The LFS results refer to persons of working age i.e. women aged 15 to 59 and men aged 15 to 64 who are residents in private houses and in NHS accommodation in UK. The sampling frame, from which most (99%) of the Great Britain sample is taken, is the Postcode Address File (PAF). The PAF is a computer list, prepared by the Post Office, of all the addresses to which mail is delivered. In addition to the PAF, another frame is the NHS accommodation sampling frame, which was specially developed for the LFS by utilising information from district health authorities and NHS trusts. In sparsely populated areas random samples are selected from the published telephone directory while for Northern Ireland the Valuation List is used.

The LFS utilises a two stage sampling procedure. The first stage is a stratified random sample of areas and the second stage is a systematic sample of addresses selected from the PAF. The country is split into 110 interviewing areas. Each of these areas is then split into 13 "stints". These 13 stint areas are randomly allocated to the 13 weeks of a quarter. The same stint area is covered in the same week of each quarter by an LFS interviewer. A systematic sample of addresses is selected for each quarter throughout the country and is distributed between the stint areas to provide a list of addresses to be interviewed each week. The sample currently consists of about 59000 responding households in Great Britain every quarter, representing 0.3% of the population. A sample of approximately 2000 responding households in Northern Ireland is added to this, representing 0.4% of the Northern Ireland population, allowing UK level analyses.

Rotating Design

Each quarter the LFS sample of UK households is made up of five waves each of approximately 12000 households. Each wave is interviewed in five successive quarters such that in any quarter sample units belonging to the first wave will have their first interview, sample units belonging to the second wave will have their second interview etc. Thus, there is an 80% overlap in the samples for each successive quarter.

Weighting

The UK LFS includes longitudinal survey weights. These weights serve two purposes. They compensate for differential non-response and also produce estimates at the national level. As described in ONS (2000), the computation of weights for the two-quarter linked datasets involves the following stages:

(i) Initial prior weights are calculated such that they reproduce the distribution of the cross-sectional sample from the first quarter according to the tenure/landlord categories: owned, rented from local authority/housing association, privately rented.

(ii) These initial prior weights are then multiplied by a single grossing factor such that the weighted sample cases sum to an overall population control total. This process results in the derivation of prior weights used in the calculation of the final weights.

(iii) A process of calibration weighting (also known as generalised raking) is then applied to the sample using CALMAR software (see Elliot 1997). This process minimises the distance between the prior and final weights while constraining the final weights simultaneously to several marginal distributions or control totals. For the production of the weighting factors in the UK LFS, four sets of control totals are utilised (see ONS 1999, ONS 2000).

As mentioned in Section 1.4.1, the UK weighting system accounts for the sampling attrition problem. Hence, by incorporating the survey weights in the analysis we account for one of the major sources of bias affecting the estimation of labour force gross flows.

Other Design Characteristics

Households belonging to the first wave are interviewed face to face while interviews for the remaining waves are carried out by telephone. The LFS design allows interviewers to receive answers from proxy respondents (about 30% of the LFS responses are collected by proxy). A proxy respondent is usually another related adult who is a member of the same household. The LFS interviews are carried out using Computer Assisted Interviewing (CAI), which ensures improved speed from fieldwork to the analysis of the data and also better data quality (e.g. automatic check of inconsistencies).

1.7.3 Estimating Labour Force Gross Flows Using LFS Linked Datasets

The design of the UK LFS enables estimates of levels such as the number of people in employment, which are representative of the national labour force population, to be produced for any period of three consecutive months. However, due to its panel character, the LFS also allows estimates of change to be produced. This can be achieved by linking the responses of sample units that belong to consecutive quarters i.e. that belong to the common sample (see

ONS 1999, Kristiansson and Mirza 2000). These parameters of change will be the main parameters of interest in this thesis.

A labour force gross flows matrix between t and $t+1$, taking into account the dynamic evolution of the population, is presented in Table 1.3. The inflows include persons who have turned 16 or who have immigrated to the country between t and $t+1$. The outflows include persons who have turned 65, have died or have left the country between t and $t+1$.

Table 1.3: Complete labour force gross flows between t and $t+1$.

	(E)	***(U)***	***(N)***	***Outflows***	***Total at*** t
(E)	EE	EU	EN	EO	E.
(U)	UE	UU	UN	UO	U.
(N)	NE	NU	NN	NO	N.
Inflows	IE	IU	IN		
Total at $t+1$	. E	. U	. N		

The margins of Table 1.3 give the quantities that are regularly estimated by the UK LFS. The column 'Total at t' describes the distribution of labour force states for the population in working age (i.e. 16-64) at t. Similarly, the row 'Total at $t+1$' describes the distribution of labour force states for the population in working age (i.e. 16-64) at $t+1$. Table 1.3 shows also the distribution of labour force states for those who leave (Outflows) the working population and for those who enter (Inflows) the working population between t and $t+1$. Denote by O the outflows and by I the inflows. The relation between the population at t (U_t) and the population at $t+1$ (U_{t+1}) can be expressed as follows:

$$U_{t+1} = U_t - O + I \text{ or } U_{t+1} = U_t \cap U_{t+1} + I. \tag{1.21}$$

We define the following notation

Population Level

- $U_t = \{1,2,\dots,\xi,\dots,N_t\}$ denotes the working population at t consisting of N_t units.
- $U_{t+1} = \{1,2,\dots,\xi,\dots,N_{t+1}\}$ denotes the working population at $t+1$ consisting of N_{t+1} units.

- $U_{t,t+1} = U_t \cap U_{t+1} = \{1,2,\ldots,\xi,\ldots,N\}$ denotes the population units that belong both to U_t and U_{t+1} consisting of N units.

Sample Level

- $S_t = \{1,2,\ldots,\xi,\ldots,n_t\} \subset U_t$ consisting of n_t units.
- $S_{t+1} = \{1,2,\ldots,\xi,\ldots,n_{t+1}\} \subset U_{t+1}$ consisting of n_{t+1} units.
- $S_t' = S_t - Outflows$ consisting of n_t', $n_t' < n_t$ units.
- $S_{t+1}' = S_{t+1} - Inflows$ consisting of n_{t+1}', $n_{t+1}' < n_{t+1}$ units.
- $S_{t,t+1} = S_t' \cap S_{t+1}' = \{1,2,\ldots,\xi,\ldots,n\}$. Sample members who belong to $U_{t,t+1}$ and also to the sample both at t and $t+1$ consisting of n units, $n < n_t' < n_t, n < n_{t+1}' < n_{t+1}$.

In a quarterly survey with five rotation groups and 80% overlap, the difference between samples S_t', $S_{t+1}'$and $S_{t,t+1}$ is that while S_t' and S_{t+1}' are based on all five rotation groups, $S_{t,t+1}$ is based on four rotation groups.

Estimates of Level

Denote by w_ξ the cross-sectional survey weight for sample unit ξ and by $Y_{\xi t}^*$ a random variable that describes the labour force status of the same unit at time t. Denote further by T_t the total number of persons in the population with the specific labour force characteristic at time t. An estimator of T is given by

$$\hat{T}_t = \sum_{\xi \in S_t} w_\xi Y_{\xi t}^*. \tag{1.22}$$

Estimates of Change

Denote by $Y_{\xi t \to t+1}^*$ a random variable that describes a specific labour force flow of sample unit ξ between t and $t+1$ and by w_ξ the longitudinal weight for sample unit ξ. Let $T_{t \to t+1}$ denote the total number of population units that belong to a specific internal cell of Table 1.3. Utilising the common sample $S_{t,t+1}$, an estimator of $T_{t \to t+1}$ is given by

$$\hat{T}_{t \to t+1} = \sum_{\xi \in S_{t,t+1}} w_{\xi} Y^{*}_{\xi t \to t+1}. \tag{1.23}$$

Using (1.23), we obtain estimates of the labour force gross flows (i.e. our parameters of interest) as well as of the margins of the gross flows table.

1.8 Inference about the Misclassification Mechanism

Techniques that adjust the quantities of interest for misclassification require the specification of the structure of the measurement error. One way of doing this in a latent class context is by specifying the relationship between the manifest and the latent variables. An alternative way is by exogenously estimating the parameters of the misclassification mechanism using information derived from validation surveys. In the upcoming sections we focus our interest on the second approach. After a general overview of some of the validation procedures that can be used, we focus our interest to validation studies with preferred procedures (Kuha and Skinner 1997) or to what Forsman and Schreiner (1991) refer to as re-interview surveys.

1.8.1 On the Definition of True Values

The definition of what is a true value has caused large debates in the statistical community. Generally, two approaches exist. One approach assumes that true values exist independently of the survey conditions. The second approach adopts a more operational definition and assumes true or preferred values only in relation to the survey conditions. According to the first approach (Hansen et al. 1951) three criteria exist for the definition of a true value.

1. The true value is uniquely defined.
2. The true value is defined in such a way that the purposes of the survey are met.
3. Where it is possible to be consistent with the first two criteria, the true value should be defined in terms of operations that can actually be carried through despite the fact that these procedures can be expensive or difficult to perform.

However, as pointed out by Hansen et al. (1951), it may be impossible to define a true value that meets all three criteria above. Consequently, they propose to define a value that satisfies the first two criteria and an operation whose expected value under a large number of replications will give a satisfactory approximation to the true value. This approach influenced subsequent work by Kish (1965), Raj (1968) and Moser and Kalton (1972).

In contrast to this approach, which defines the true values separately from the survey conditions, Deming (1944) defines the true values as a function of the survey conditions. He states that there is no true value and we have the liberty to define and to accept a specific set of operations as preferred. However, due to cost or other reasons, these operations are not always easy to be adopted. In the same framework, Zarcovich (1966) defines true values in the context of an adopted system that consists of chosen measurement methods, concepts and definitions, tabulation plans and data collection instructions. The true values can be obtained if the system is implemented without error. A thorough literature review on the definition of the true values is given in Lessler and Kalsbeek (1992).

Assume that the standard survey process gives a contaminated measurement and the preferred procedure gives an error free measurement. Then the quantities $q_{ik} = pr\left(Y^*_{\xi t} = i \mid Y_{\xi t} = k\right)$ can be determined exogenously using preferred procedures. We consider these probabilities as the parameters that describe the misclassification mechanism. It is apparent that the preferred procedures play a key role in studying the misclassification mechanism and in evaluating the quality of the survey measurements. Examples of preferred procedures are (a) judgments of experts e.g. Swires-Hennessy and Thomas (1987) describe an application of this kind in some surveys of housing in Wales and Chen (1977) compares survey data to data derived from a physician's examination, (b) checks against administrative records e.g. Greenland (1988) describes an application where re-interview data on antibiotic use are compared against medical records and (c) re-interview programmes as part of large scale sample surveys that attempt to identify the true values e.g. the Swedish LFS re-interview programme (Kristiansson 1999).

1.8.2 Preferred Procedures: The Case of Re-interview Surveys

Survey models have been developed to meet the need for an integrated treatment of sampling errors and response errors. An example of a model of this kind is the US Bureau of the Census survey model (Hansen, et al. 1951). In this model, the mean squared error is decomposed into sampling components and response error components. Two major methodologies exist for measuring the response error components: (1) The method of interpenetrated sub-samples, which is designed for estimating the correlated component of the response variance and (2) the method of replicated measurement, which is designed for estimating the response variance and the response bias.

The method of replicated measurement is known as the re-interview method. Re-interview methodology was developed in the US and India during the 1940's and since then it has been used in a number of countries. By the term re-interview we mean a new interview that takes place some time after the original survey but refers to the same point in time as the original interview. In other words, the re-interview does not include interviews of the same persons in two or more waves of a panel survey since different waves refer to different time points. Re-interviews are important tools for estimating and reducing response errors in surveys. Response errors may be caused for a variety of reasons i.e. imperfect instructions to the interviewers, badly designed questions and questionnaires, coincidental factors that affect the interviewer or the respondent, deliberate errors from the respondent and deliberate falsification of interview results from the interviewer. There are two basic reasons for designing a re-interview survey (1) to evaluate fieldwork and (2) to estimate the error components in a survey model. As far as the first reason is concerned, a re-interview may seek to identify interviewers who falsify data or misunderstand the interview procedure and as a result require further training. With regard to the second reason, a re-interview survey may seek to estimate the response bias or the response variance. For the purposes of our work, we are mainly interested in re-interview procedures that aim at estimating components of the response error.

A considerable amount of the literature dealing with measurement error in labour force gross flows utilises re-interview data under the assumption that the re-interview responses represent an error free measurement. This implies that the re-interview survey is treated as a perfect instrument. However, this may not be the case. In what follows we will identify re-interview design characteristics that allow these assumptions to more closely reflect reality.

When the aim is to estimate the response variance component the crucial assumption is that the re-interview survey is an independent replication of the original survey. This implies that the re-interview survey must be repeated independently of the original survey but under the same survey conditions. An example of such a re-interview survey is the unreconciled part of the re-interview programme of the CPS. This sub-sample represents the 25% percent of the total re-interview sample. If we aim at obtaining the truth, we will need to conduct the re-interview survey under different survey conditions from those of the original survey. When we use the term "different survey conditions" we mean conditions of better quality such that the true value is identified. It is apparent that in this case the two measurements (i.e. the original and the re-interview measurement) are not identically distributed since the second measurement is assumed to be of higher quality. However, the assumption of independence between the original and the re-interview survey is crucial and must still hold. An example of a re-interview survey that aims at estimating the response bias component is the reconciled part of the CPS re-interview sample, which represents the 75% of the total re-interview sample of the CPS. Independence between the original and the re-interview survey can be viewed at two levels i.e. at the respondent level and at the re-interviewer level. Independence at the respondent level means that there are no recall effects between the original and the re-interview survey. If this is not the case, serial correlation will be introduced. This implies that the respondents might recall the original response and simply replicate it during the re-interview. Consequently, if the original response is erroneous, the re-interview response will be erroneous too. In such a case no discrepancy between the two measurements is observed and no attempt for reconciliation takes place. Independence at the re-interviewer level means that the re-interviewer has no access to the original responses and reconciliation is conducted using an independent method. While independence at the re-interviewer level can be achieved by elaborating the re-interview survey conditions, independence at the respondent level is more difficult to achieve and does not depend only on the re-interview survey conditions.

For the purposes of adjusting labour force gross flows for measurement error we are interested in identifying the true labour force status of each respondent by means of a re-interview survey. In the sequel, we will examine suitable re-interview survey conditions such that the assumption that we estimate the truth is close to reality. The re-interview survey design characteristics that we study are: (a) the type of reconciliation, (b) the questionnaire design, (c) the time lag between the original and the re-interview survey, (d) field implementation issues and (e) the use of computerised assisted techniques (e.g. CATI) in re-interview surveys.

Method of Reconciliation

By the term reconciliation we mean the attempt, made by the re-interviewer, to obtain the true value. Assume that during the original survey the respondent gives a specific answer while in the re-interview survey the same respondent provides a different response. The re-interviewer must find which of the two different responses reflect reality. It is apparent that when no discrepancy exists, reconciliation is not conducted. Consequently, when the aim is to estimate the response bias, reconciliation should always be carried out. The crucial decision concerns the method of reconciliation.

There are two ways of carrying out the reconciliation process: (a) the re-interviewer is supplied with the original answers and reconciliation takes place at the same time as the re-interview survey and (b) the re-interviewer is not provided with the original answers and reconciliation is conducted either by a third contact with the household or by using another independent method.

There are serious objections regarding the first method of reconciliation. These concern the fact that the assumption of independence between the original and the re-interview survey is violated when the re-interviewers are provided with the original responses. In order to understand more the consequences of the violation of this assumption, we describe the following situation. As we already mentioned, the re-interview sample of the CPS is divided into two parts: in one part of the sample differences are reconciled by providing the re-interviewer with the original responses whereas in the other part no reconciliation takes place. Theoretically, the discrepancies from both sub-samples should be the same. However, this not the case. The US Bureau of Census (1963) and O'Muircheartaigh (1986) showed that there are substantial differences in the number of discrepancies reported by the two sub-samples. Biemer and Forsman (1992) provide similar evidence by comparing estimates of the response variance from the two sub-samples. More specifically they found that the reconciled sub-sample shows fewer discrepancies than the unreconciled one. Bailar (1968) investigated the effect of reconciliation by comparing results from different re-interview strategies. She concluded that results from a re-interview sample where the reconciliation is made at the same time as the re-interview ("on the spot" reconciliation) and where the re-interviewers had access to the original answers exhibit more dependence. Similar results, in the context of estimating labour force gross flows, are reported by Poterba and Summers (1986). The US

Bureau of Census (1963) points out the difficulty of conducting an independent reconciliation when re-interviewers are provided with the original responses.

The question is whether by using an independent reconciliation we can assure the assumption of independence between the original and the re-interview survey. In this context, Schreiner (1980) conducted an independent reconciliation experiment, which did not reveal any significant differences like those revealed under a dependent reconciliation. Biemer and Forsman (1992) used also data from an independent reconciliation experiment. They concluded that the independent reconciliation offers a better solution. However, in their opinion the serial correlation does not disappear and as a result the hypothesis that we identify the true value is highly suspect. We therefore conclude that independent reconciliation of data seems to be more consistent with the assumption that the reconciliation process identifies the true values.

Questionnaire Design

The design of the questionnaire plays a crucial role in a re-interview survey since it can be directly connected with the assumption that the re-interview survey identifies the truth. According to Forsman and Schreiner (1991), two alternative questionnaire designs exist: (1) the original question(s) may be repeated and differences between the two responses are reconciled or (2) there may be a series of questions replacing the original question in an effort to obtain the truth.

The first design is used mainly when we want to estimate the response variance by conducting a re-interview that is an identical replicate of the original interview. The second design seems to be more appropriate when the aim is to discover the truth. This is because the re-interviewer is then not restricted to replicating the same question as in the original survey but is free to conduct the re-interview in a way that attempts to identify the true value.

Time Lag between the Original Interview and the Re-interview Survey

The time lag is also an important factor when trying to estimate the error components using a re-interview survey. As we have already stated, a crucial assumption when attempting to estimate the error components is the assumption of independence between the original and the re-interview survey. The time lag between the two surveys is directly connected with this assumption. Assume a situation where the time lag is not sufficiently enough and as a result respondents remember the answers they gave in the original survey. This introduces serial correlation and consequently, respondents may repeat an erroneous answer in which case no reconciliation is conducted. As a result, the time lag should neither be so large so the respondents forget what their actual status was during the reference period nor so short as to have recall effects. Furthermore, the time lag also depends on the nature of the data gathered i.e. the more the data are subject to variation the shorter the time lag should be. For example, the CPS focuses on labour force items (i.e. mobility items) and so a one-week time lag is used. However, for other surveys dealing with less volatile variables like race, gender and education the time lag can be several months. Palmer (1943) concluded that the greater lapse of time between the two surveys implies greater variability for responses related to employment status. In the same context, Bailar (1968) compared re-interview surveys with different time lags and concluded that for certain response items, like mobility items, the shorter time lag is preferable.

Field Implementation Issues

One issue associated with the fieldwork is the choice of the interviewer who is going to conduct the re-interview survey. The choices vary between the original interviewer and the better interviewer. When the target is to identify the true value, the better interviewer should be used. Another important issue is the mode used to conduct the re-interview. Usually, the telephone is used in order to reduce the costs of the re-interview survey. However, when the original survey has been conducted by a face-to-face interview, using the telephone in the re-interview survey may have a significant impact on the results.

Use of Computer Assisted Interviewing (CATI) in Re-interview Surveys

Under CATI, the perspective of re-interview changes since interviewing, which is conducted at a centralised telephone facility, can be monitored. This allows for the focus to be estimation of error components and not evaluation of the fieldwork. For the estimation of the response bias, CATI has the following advantages:

a) It is possible to conduct the re-interview using the best interviewer and the most knowledgeable respondent.

b) There is much more flexibility in deciding the "optimal" time lag between the original and the re-interview survey since in a centralised telephone facility the re-interviews can be conducted much more quickly.

c) The re-interviewer has no access to the original interview data until the end of the re-interview.

d) It is not possible for the re-interviewer to alter the re-interview responses once the re-interview has finished.

e) Identification of when there is a difference and why this difference occurs can be made automatically.

Discussion

A well-designed re-interview survey can be an extremely useful tool for estimating and reducing measurement error and consequently for improving survey quality. However, re-interview surveys have certain disadvantages. Firstly, the method is considered to be fairly expensive. In addition, re-interview surveys are multipurpose surveys. This implies that the characteristics of a re-interview survey designed to serve one purpose are not necessarily optimal for another purpose. Furthermore, the model assumptions that we impose in order to estimate the different components (i.e. the response variance and the response bias) using a re-interview survey are not always satisfied. For example, we assume that the re-interview survey is independent of the original survey and thus there are no recall effects. However, this may not be the case since the respondents may remember their prior responses and simply replicate them (i.e. serial correlation is introduced).

With regard to the cost considerations, re-interview surveys can maximise the use of the telephone so as to have reduced costs. This advantage is reinforced with computer assisted interviewing (CATI) in a centralised setting. Using CATI, costs can be kept minimal while the usefulness of a re-interview is increased.

For the purposes of our research, we are interested in a re-interview survey designed to obtain true values. In order to estimate the response bias, we need to impose two assumptions: (a) the re-interview is independent from the original survey and (b) the re-interview identifies the truth. An optimal re-interview survey, in the sense that the above assumptions become more realistic, must have certain design characteristics. We are in favour of a re-interview survey that (a) uses an independent (i.e. not "on the spot") reconciliation procedure where the re-interviewers are not provided with the original responses, (b) is carried out in a more conversational form, (c) is conducted by the best interviewers and (d) utilises a computer assisted interviewing system (CATI).

1.8.3 Validation Studies in UK

The UK LFS has not yet developed a re-interview programme that will allow the estimation of the parameters of the misclassification mechanism. However, there are other examples of validation studies in UK. For example, the UK Census Validation Survey (CVS) (Heady, Smith and Avery 1991) had as main targets to assess the coverage of the Census, to evaluate how prone to error Census questions were and to identify possible sources of error. An example of a panel (two-wave) validation study is described in the context of the Panel Study of Income Dynamics (PSID) in Hill (1992).

In a UK LFS framework, there are suggestions that one could compare the UK Census results on labour market related topics with corresponding UK LFS results. Our objection to this comparison is that the UK Census cannot be considered as a survey of higher quality for labour market related issues for example, the UK Census is a self-reported survey. Furthermore, while the UK Census is conducted every ten years, the UK LFS is a panel survey making the comparison more difficult. Other examples of validation experiments, in a UK LFS framework, include the linkage of UK LFS responses with administrative records about claimants of unemployment related benefits (ONS 1997). The purpose of this linkage study is to obtain adjusted LFS estimates for the claimants of unemployment related benefits. The disadvantage of this study is that it is restricted to specific groups of the labour force population.

1.8.4 The Swedish LFS Re-interview Programme

Sweden is one of the few countries that uses re-interview survey programmes in order to assess the impact of the measurement error on the estimation of labour force estimates (Kristiansson 1999). The first important evidence for the assessment of this measurement error came from the results of a re-interview survey in January 1978. This re-interview program was designed to obtain the true values. For this reason, the re-interviews were performed using a group of specially trained re-interviewers who, after the standard questions, asked further questions about the persons' employment status in a more conversational form. Approximately 3600 additional re-interviews were carried out in connection with the introduction of computer assisted interviewing (CATI) in the Labour Force Survey in 1989-1990 and 2100 further re-interviews were conducted during the period from October 1994 to April 1995. The aim of these more recent re-interview surveys was also to obtain the true values. Consequently, the quality characteristics of the Swedish re-interview programme seem close to those required when the target is to estimate the response bias component i.e. use of experienced interviewers, use of probing and utilisation of computer assisted interviewing to facilitate independent reconciliation.

In the absence of validation information for the UK LFS, this thesis utilises mainly Swedish validation data. The assumption is that the Swedish misclassification probabilities can be used as proxies for corresponding UK misclassification probabilities. This can be regarded as a fairly restrictive assumption. However, the methodology we develop is not data specific. The Swedish validation data offer one possible scenario for the UK measurement error process. There is evidence suggesting that the Swedish validation data may show less measurement error than what really exists in the UK LFS. For example, while the Swedish LFS allows for 3% of proxy response, the UK LFS allows for 30% proxy response. Nevertheless, the utilisation of Swedish validation data can provide a useful insight into the measurement error process in the UK LFS and experience for developing a UK LFS validation survey.

Chapter 2

Using Double Sampling for Misclassification Error Correction: From a Cross-sectional to a Panel Framework

2.1 Introduction

Methods for adjusting for measurement error via double sampling were first developed in a cross-sectional framework by Bross (1954) and Tenenbein (1970, 1972). However, recent literature on adjustment of gross flows for measurement error using validation information does not link adjustment procedures with double sampling theory. Following the suggestion of Kuha and Skinner (1997), in the first part of this chapter we investigate the links between the use of double sampling designs in a longitudinal and in a cross-sectional framework. We start by describing alternative double sampling schemes and moment-based inference in a cross-sectional framework and we investigate the impact of these schemes on the efficiency of the derived adjusted estimates. Generalising from the cross-sectional case, we extend double sampling designs and associated point estimators to a longitudinal framework and we investigate the impact of these designs on the efficiency of the resulting adjusted gross flows. There are two new features in our development. Firstly, we utilise an alternative parameterisation to the one proposed by Tenenbein (1972) for deriving maximum likelihood estimators of adjusted for misclassification quantities. A similar parameterisation is discussed in Espeland and Odoroff (1985). Secondly, we propose a parameterisation of the measurement error model in a quasi-likelihood framework as an alternative to maximum likelihood estimation.

In the second part of this chapter, we describe the disadvantages of moment-based inference in a longitudinal framework and we study some alternative moment-type estimators. In this context, we investigate the unbiased margins estimator (Poterba and Summers 1986, Singh and Rao 1995) and we propose three alternative estimators i.e. the modified estimator, the composite estimator with fixed weights and the composite estimator with adaptive weights.

2.2 Alternative Double Sampling Schemes Utilised to Correct for Misclassification in a Discrete Framework

We consider the general framework of double sampling methods described by Bross (1954) and Tenenbein (1970, 1972). Assume that the standard measurement device that we use is subject to measurement error. As a result, we have biased results. However, unbiased estimates can be obtained by using preferred procedures. Unfortunately, these procedures are costly to implement. The aim of double sampling methods is to combine information from both the true and the fallible classifier in order to obtain estimates that are adjusted for measurement error.

The sample where the preferred (validation) procedure is applied can be either internal or external. Kuha and Skinner (1997) make this distinction following literature on misclassification in the context of bio-statistical applications (see also Greenland 1988). From our point of view, the basic characteristic that distinguishes an internal 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. The validation sample is characterised as internal if it is a random sub-sample of n^v units from the main sample of n units obtained via a randomised double sampling scheme. Alternatively, the validation sample can be regarded as internal if it is selected independently from the main sample and from the same target population. On the other hand, a validation sample is external if it is derived from an external source of information (Hill 1992). The parameters of the misclassification mechanism estimated from an external validation sample are assumed to be informative of the misclassification process in the target population. However, the fallible classifications from the external validation sample cannot be combined with the fallible classifications from the main sample. Sometimes, it may be preferable to use an external validation sample or an internal validation sample that is selected independently from the main sample. An example is when our main measurement instrument is a panel survey and we wish to avoid additional measurements on the sample units that already participate in the panel survey.

2.2.1 The Cross-sectional Case

We start by introducing the basic notation. Denote by $\Pi_i = pr\left(Y^*_{\xi t} = i\right)$ the probability that unit ξ is classified in state i by the standard measurement device, which is subject to measurement error. Denote further by $P_k = pr\left(Y_{\xi t} = k\right)$ the probability that unit ξ truly belongs in state k, by $q_{ik} = pr\left(Y^*_{\xi t} = i \mid Y_{\xi t} = k\right)$ the misclassification probabilities and by $Q(t)$ the matrix of misclassification probabilities with elements q_{ik}. Recall that $Y^*_{\xi t}$, $Y_{\xi t}$ are random variables that describe the way that unit ξ is classified at time t using the fallible classifier and the perfect classifier respectively. Define now a matrix Π with elements Π_i and a matrix P with elements P_k. Generally speaking, the cross-sectional measurement error model with r mutually exclusive states can be described as follows:

$$pr\left(Y^*_{\xi t} = i, Y_{\xi t} = k\right) = pr(Y^*_{\xi t} = i \mid Y_{\xi t} = k) pr\left(Y_{\xi t} = k\right) \Rightarrow$$

$$\sum_{k=1}^{r} pr\left(Y^*_{\xi t} = i, Y_{\xi t} = k\right) = \sum_{k=1}^{r} pr(Y^*_{\xi t} = i \mid Y_{\xi t} = k) pr\left(Y_{\xi t} = k\right) \Rightarrow$$

$$pr\left(Y^*_{\xi t} = i\right) = \sum_{k=1}^{r} pr(Y^*_{\xi t} = i \mid Y_{\xi t} = k) pr\left(Y_{\xi t} = k\right) \Rightarrow \Pi_i = \sum_{k=1}^{r} q_{ik} P_k.$$

Expressing the previous relationship in matrix notation, assuming that $Q(t)$ is invertible and solving the equation with respect to P we derive the following expression

$$\underset{r\times 1}{P} = \underset{r\times r}{\left[Q(t)\right]^{-1}} \underset{r\times 1}{\Pi}. \qquad (2.1)$$

Expression (2.1) has been used extensively in literature to adjust discrete data for measurement error in a cross-sectional framework. Unknown quantities involved in (2.1) are typically estimated using a double sampling scheme. Below, we describe three such schemes. This parameterisation of the measurement error model leads to a moment-type estimator of the adjusted for misclassification quantities.

<u>Double Sampling Scheme 1</u>

A simple random sample of $n - n^v$ units is selected from a population of N units and the fallible classifications are obtained for each sample unit. For another simple random sample of n^v units, independently selected from the $n - n^v$ units and from the same target population, the fallible classifications are also obtained. At a second stage, the true classifications are obtained for each of the n^v units. Under this scheme we obtain information on the fallible classifications for n units i.e. $(n - n^v + n^v)$ and further information on the true classifications for n^v units. Figure 2.1 illustrates this double sampling scheme.

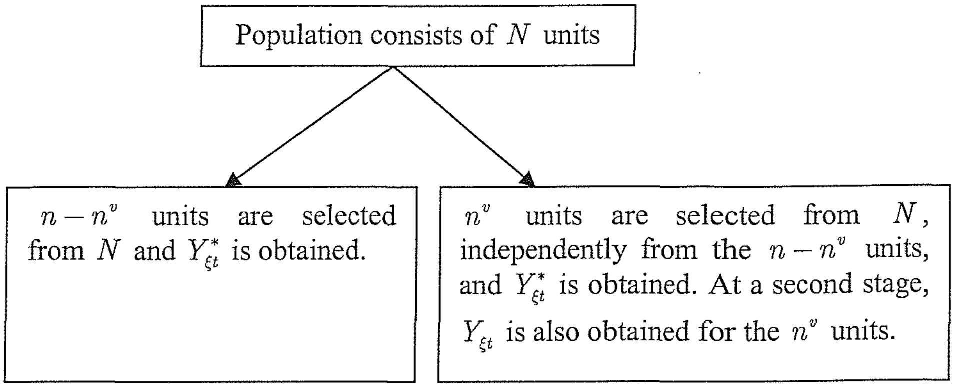

Population consists of N units
$n - n^v$ units are selected from N and $Y^*_{\xi t}$ is obtained.
n^v units are selected from N, independently from the $n - n^v$ units, and $Y^*_{\xi t}$ is obtained. At a second stage, $Y_{\xi t}$ is also obtained for the n^v units.

Figure 2.1: Double sampling scheme 1-Cross-sectional case

Double Sampling Scheme 2

A simple random sample of n units is selected from a population of N units and the fallible classifications are obtained for each sample unit. At the second stage, a sub-sample of n^v units is selected from the n units that already belong to the main sample and the true classifications are obtained for each of these n^v units. Figure 2.2 illustrates this double sampling scheme.

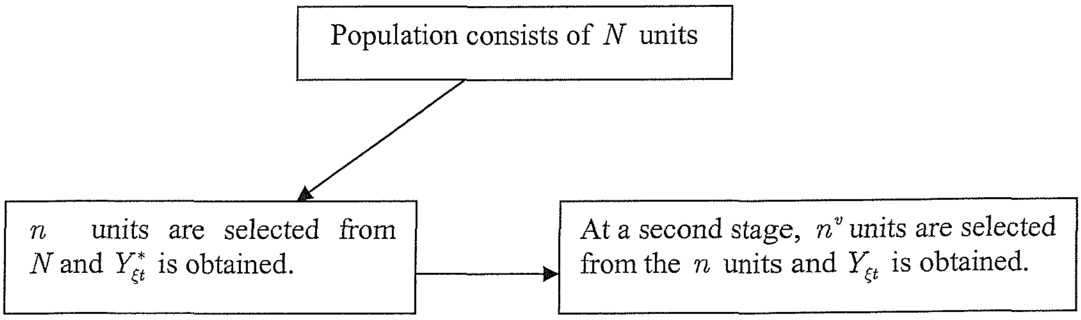
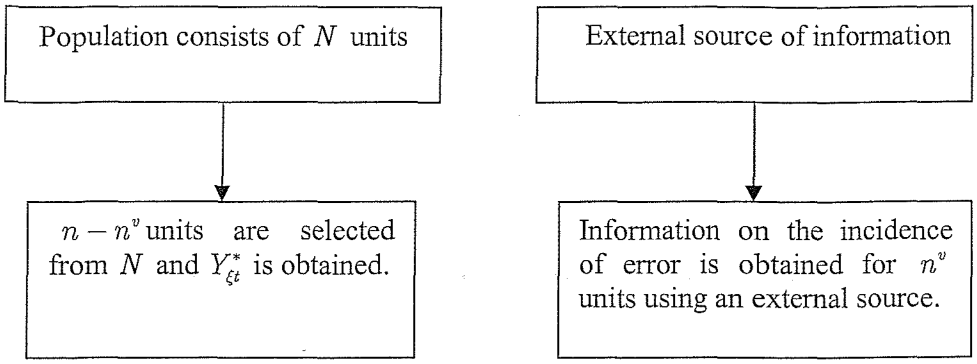

Population consists of N units
n units are selected from N and $Y^*_{\xi t}$ is obtained.
At a second stage, n^v units are selected from the n units and $Y_{\xi t}$ is obtained.

Figure 2.2: Double sampling scheme 2-Cross-sectional case

Double Sampling Scheme 3

A simple random sample of $n - n^v$ units is selected from a population of N units and the fallible classifications are obtained for each sample unit. Information about the incidence of error is derived for n^v units from an external source of information (e.g. administrative records). Figure 2.3 illustrates this double sampling scheme.



Population consists of N units

$n - n^v$ units are selected from N and $Y^*_{\xi t}$ is obtained.

External source of information

Information on the incidence of error is obtained for n^v units using an external source.

Figure 2.3: Double sampling scheme 3-Cross-sectional case

2.2.1.1 Review of Alternative Double Sampling Schemes in a Cross-sectional Framework

For double sampling scheme 1 (see Figure 2.1), the validation sample includes additional information on the observed classifications that can be combined with information on the observed classifications from the main sample. In fact, the first and the second double sampling schemes are identical. This is because under the second double sampling scheme the sample can be divided into $n - n^v$ units that participate only in the main survey and n^v units that participate both in the main and in the validation survey.

For the third scheme, the validation sample can be regarded as an external since it is selected from an external source of information. This implies that the fallible classifications from this validation sample cannot be combined with the fallible classifications from the main sample. Nevertheless, here we argue that the external validation sample can be transformed into an internal validation sample. Since the misclassification probabilities estimated from this external validation sample are assumed to be informative 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 the Iterative Proportional Fitting (IPF) algorithm (Deming and Stephan 1940). After transforming the external validation sample into an internal validation sample, the fallible classifications from the validation sample can be combined with the fallible classifications from the main sample.

Assuming that all relevant quantities can be estimated using information from the main and the validation sample, an estimator of (2.1) under double sampling scheme 1 is given by the following expression

$$\underset{r\times1}{\hat{P}^{(1)}} = \frac{1}{n-n^v+n^v}\left\{\left[\underset{r\times r}{\hat{Q}(t)}\right]^{-1}\left[(n-n^v)\underset{r\times1}{\hat{\Pi}^{m}} + n^v \underset{r\times1}{\hat{\Pi}^{v}}\right]\right\}, \quad \hat{\Pi}_i^{m} = \frac{\sum_{\xi=1}^{n-n^v} Y_{\xi t}^{*}}{n-n^v}, \hat{\Pi}_i^{v} = \frac{\sum_{\xi=1}^{n^v} Y_{\xi t}^{*}}{n^v}. \quad (2.2)$$

Note that $\hat{\Pi}^{m}$ denotes the matrix, with elements $\hat{\Pi}_i^{m}$, of estimated probabilities based on data from the main sample, $\hat{\Pi}^{v}$ denotes the corresponding estimate, with elements $\hat{\Pi}_i^{v}$, based on data from the validation sample. Combining $\hat{\Pi}^{m}$ with $\hat{\Pi}^{v}$, yields the matrix of estimated probabilities $\hat{\Pi}$, with elements $\hat{\Pi}_i$, based on both samples.

Under double sampling scheme 2, an estimator of (2.1) is given below

$$\underset{r\times1}{\hat{P}^{(2)}} = \left[\underset{r\times r}{\hat{Q}(t)}\right]^{-1}\underset{r\times1}{\hat{\Pi}}, \quad \hat{\Pi}_i = \frac{\sum_{\xi=1}^{n} Y_{\xi t}^{*}}{n}. \quad (2.3)$$

Note that by dividing the sample of the second double sampling scheme into units that participate only in the main survey and units that participate both in the main and in the validation survey, estimators (2.2) and (2.3) become identical.

For double sampling scheme 3, the validation sample is external. Here, it is not logical to combine information on the fallible classifications from the validation sample with information on the fallible classifications from the main sample. As a result, an estimator of (2.1) takes the following form

$$\underset{r\times1}{\hat{P}^{(3)}} = \left[\underset{r\times r}{\hat{Q}(t)}\right]^{-1}\underset{r\times1}{\hat{\Pi}^{m}}, \quad \hat{\Pi}_i^{m} = \frac{\sum_{\xi=1}^{n-n^v} Y_{\xi t}^{*}}{n-n^v}. \quad (2.4)$$

Comparing estimators (2.2), (2.3) and (2.4), we conclude that estimators that are based on an internal validation sample i.e. (2.2) and (2.3) are more efficient than estimator that is based on an external validation sample i.e. (2.4). However, if an external validation sample is transformed into an internal validation sample all three estimators become equivalent.

2.2.1.2 Calibration Probabilities versus Misclassification Probabilities and Maximum Likelihood Estimation in a Cross-sectional Framework

Estimators (2.2),(2.3) and (2.4) utilise the misclassification probabilities q_{ik} in order to describe the misclassification mechanism. Another way of making inferences about the misclassification mechanism is by using what Carroll (1992) refers to as calibration probabilities. The calibration probabilities are defined as $c_{ki} = pr\left(Y_{\xi t} = k \mid Y^*_{\xi t} = i\right)$. Thus, while the misclassification probabilities condition on the true classifications, the calibration probabilities condition on the observed classifications. Denote by $C(t)$ the matrix of calibration probabilities with elements c_{ki}. The measurement error model under the calibration probabilities becomes

$$pr\left(Y^*_{\xi t} = i, Y_{\xi t} = k\right) = pr(Y_{\xi t} = k \mid Y^*_{\xi t} = i)pr\left(Y^*_{\xi t} = i\right) \Rightarrow$$
$$\sum_{i=1}^{r} pr\left(Y^*_{\xi t} = i, Y_{\xi t} = k\right) = \sum_{i=1}^{r} pr(Y_{\xi t} = k \mid Y^*_{\xi t} = i)pr\left(Y^*_{\xi t} = i\right) \Rightarrow$$
$$pr\left(Y_{\xi t} = k\right) = \sum_{i=1}^{r} pr(Y_{\xi t} = k \mid Y^*_{\xi t} = i)pr\left(Y^*_{\xi t} = i\right) \Rightarrow P_k = \sum_{i=1}^{r} c_{ki}\Pi_i.$$

In matrix notation,

$$\underset{r\times 1}{P} = \underset{r\times r}{C(t)}\underset{r\times 1}{\Pi}. \tag{2.5}$$

Unknown quantities involved in (2.5) can also be estimated using a double sampling scheme. However, the measurement error model that utilises calibration probabilities can be used only in the case of an internal validation sample. In contrast to calibration probabilities that condition on the observed classifications, misclassification probabilities condition on the true classifications. The true classifications can be thought of as representative of a universal truth. This implies that unlike calibration probabilities, misclassification probabilities can be regarded as transportable to the population of interest (Kuha and Skinner 1997) and can be used also in the case of an external validation sample.

Utilising similar notation as in the case of the model defined in terms of misclassification probabilities, an estimator of (2.5) under the first double sampling scheme is defined as

$$\underset{r\times 1}{\hat{P}^{(1)}} = \underset{r\times r}{\hat{C}(t)}\left[\frac{\left(n - n^v\right)}{n}\underset{r\times 1}{\overset{\wedge}{\Pi}^m} + \frac{n^v}{n}\underset{r\times 1}{\overset{\wedge}{\Pi}^v}\right], \quad \overset{\wedge}{\Pi}{}^m_i = \frac{\sum_{\xi=1}^{n-n^v} Y^*_{\xi t}}{n - n^v}, \quad \overset{\wedge}{\Pi}{}^v_i = \frac{\sum_{\xi=1}^{n^v} Y^*_{\xi t}}{n^v}. \tag{2.6}$$

Under the second double sampling, an estimator of (2.5) is given by

$$\underset{r\times 1}{\hat{P}^{(2)}} = \underset{r\times r}{\hat{C}(t)}\underset{r\times 1}{\hat{\Pi}}, \quad \hat{\Pi}_i = \frac{\sum_{\xi=1}^{n} Y^*_{\xi t}}{n}. \tag{2.7}$$

Estimator (2.6) is identical to estimator (2.7). Tenenbein (1972) proved that the estimator defined either by (2.6) or (2.7) is the maximum likelihood estimator of the adjusted for misclassification proportions. He also provided an expression for its asymptotic variance. Assume that we utilise double sampling scheme 1 or 2 and that a sample unit can be classified in r mutually exclusive states. Denote by n^v_{ik} the count for each cell of the cross-classification of the observed by the true classifications in the validation sample and by $n_{i\cdot}, n^v_{i\cdot}$ the total number of sample units classified in state i by the fallible measurement device in the main sample and in the validation sample respectively. In order to obtain maximum likelihood estimates for the parameters of interest, Tenenbein (1972) maximised the log-likelihood function

$$\begin{aligned} l(\Theta) = \sum_{k=1}^{r}\sum_{i=1}^{r} n^v_{ik}\log c_{ki} + \sum_{k=1}^{r}\sum_{i=1}^{r}(n^v_{i\cdot} - n^v_{ik})\log(1-c_{ki}) + \sum_{i=1}^{r-1}(n_{i\cdot} + n^v_{i\cdot})\log(\Pi_i) \\ + (n_{r\cdot} + n^v_{r\cdot})\log\left(1 - \sum_{i=1}^{r-1}\Pi_i\right). \end{aligned} \tag{2.8}$$

As noted by Marshall (1990) and Kuha and Skinner (1997), the maximum likelihood estimator (2.7) will be more efficient than the moment-type estimator based on (2.1). However, this assumes internal validation data. When only external validation data are available, the moment-type estimator must be used and its poor performance is an important problem. One way to overcome this problem is by transforming the external validation sample into an internal validation sample.

2.2.1.3 An Alternative Parameterisation for Maximum Likelihood Estimation in a Cross-sectional Framework

In what follows, we present an alternative parameterisation of the measurement model presented by Tenenbein (1972). More specifically, we argue that an alternative way of obtaining maximum likelihood estimators is by using misclassification probabilities instead of calibration probabilities. The general set up is as follows. For the main sample of n units the classifications are made using only the fallible classifier. For a smaller sample of n^v units, selected independently from the main sample and from the same target population, the classifications are made using both the perfect and the fallible classifier. The problem can be described schematically as follows:

Table 2.1: Validation sample

		True Classifications			
		(1)	…	(r)	Margins
Fallible Classifications	(1)	n_{11}^{v}	…	n_{1r}^{v}	$n_{1\bullet}^{v}$
	⋮	⋮	…	⋮	⋮
	(r)	n_{r1}^{v}	…	n_{rr}^{v}	$n_{r\bullet}^{v}$
	Margins	$n_{\bullet 1}^{v}$	…	$n_{\bullet r}^{v}$	n^v

Table 2.2: Main sample

		True Classifications			
		(1)	…	(r)	Margins
Fallible Classifications	(1)	$n_{11}^{(*)}$	…	$n_{1r}^{(*)}$	$n_{1\bullet}$
	⋮	⋮	…	⋮	⋮
	(r)	$n_{r1}^{(*)}$	…	$n_{rr}^{(*)}$	$n_{r\bullet}$
	Margins	$n_{\bullet 1}^{(*)}$	…	$n_{\bullet r}^{(*)}$	n

The key concept of the parameterisation, as shown in the tables above, is that both the main sample and the validation sample have a similar structure. However, for the validation sample full information exists while for the main sample we have only marginal information about the observed classifications. Consequently, this parameterisation will lead to an optimisation problem that involves missing data. This is due to the fact that the validation procedure is not applied to the units of the main sample. We need to combine information from both the main and the validation survey. In order to do so, we make the basic assumption that the main and the validation samples share common parameters because both are assumed to be representative of the same population. Assuming independence between the main sample and the validation sample and denoting by $(*)$ any unobserved quantities, the likelihood function of the augmented data defined in terms of the misclassification probabilities is

$$L(\Theta) = \prod_{k=1}^{r}\prod_{i=1}^{r}(P_k q_{ik})^{n_{ik}^{v}}\prod_{k=1}^{r}\prod_{i=1}^{r}(P_k q_{ik})^{n_{ik}^{(*)}} \Rightarrow L(\Theta) = \prod_{k=1}^{r}\prod_{i=1}^{r} q_{ik}{}^{\left(n_{ik}^{(*)}+n_{ik}^{v}\right)} P_k^{\left(n_{ik}^{(*)}+n_{ik}^{v}\right)}. \quad (2.9)$$

Recall that $P_k = pr\left(Y_{\xi t} = k\right)$ denotes the probability of a correct classification in state k and $q_{ik} = pr\left(Y_{\xi t}^{*} = i \mid Y_{\xi t} = k\right)$ denotes the probability of misclassification. Taking the logarithms in both sides of (2.9) and imposing the additional constraints that

$$\sum_{k=1}^{r} P_k = 1 \tag{2.10}$$

$$\sum_{i=1}^{r} q_{ik} = 1 \text{ for fixed } k, \tag{2.11}$$

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

$$\begin{aligned} l(\Theta) = &\sum_{k=1}^{r}\sum_{i=1}^{r-1}\left(n_{ik}^{(*)} + n_{ik}^{v}\right)\log\left(q_{ik}\right) + \left(n_{rk}^{(*)} + n_{rk}^{v}\right)\log\left(1 - \sum_{i=1}^{r-1} q_{ik}\right) + \sum_{k=1}^{r-1}\left(n_{\bullet k}^{v} + n_{\bullet k}^{(*)}\right)\log\left(P_k\right) \\ &+ \left(n_{\bullet r}^{v} + n_{\bullet r}^{(*)}\right)\log\left(1 - \sum_{k=1}^{r-1} P_k\right). \end{aligned} \tag{2.12}$$

Estimation

The log-likelihood function (2.12) assumes the availability of unobserved data. One way of using this likelihood function to maximise the likelihood of the observed data is via the EM algorithm (Chen and Fienberg 1976, Dempster, Laird and Rubin 1976). The EM algorithm is based on two steps, the expectation (E-step) and the maximisation (M-step). Generally speaking, the algorithm is initialised using a set of arbitrarily selected starting values for the parameters involved in the model. Based on these starting values, in the E-step sufficient statistics defined by the complete data likelihood (e.g. equation (2.12)) are replaced by their conditional expectations given the observed data and the current set of parameter estimates. Having estimated these conditional expectations, the full data likelihood can now be maximised to produce a new set of maximum likelihood estimates. Using this new set of maximum likelihood estimates, new conditional expectations are estimated in the E-step and new maximum likelihood estimates are derived in the M-step. The E and M step are iterated until a convergence criterion is satisfied. For the measurement error model, these steps are described below.

E-step

We start by taking the conditional expectation of the log-likelihood of the augmented data given the observed data and the current estimates. We denote by D^c the complete data, defined both by the observed and the missing data, by D^v the observed data derived from the validation sample, by D^m the observed data derived from the main sample, by (h) the current EM iteration and by $\Theta^{(h)}$ the vector of parameters in the (h) EM iteration. The form of the log-likelihood of the augmented data after taking the conditional expectations becomes

$$\begin{aligned} E\left[l(\Theta; D^c) \mid D^v, D^m, \Theta^{(h)}\right] &= \sum_{k=1}^{r}\sum_{i=1}^{r-1} E\left[\left(n_{ik}^{(*)} + n_{ik}^{v}\right) \mid D^m, D^v, \Theta^{(h)}\right]\log\left(q_{ik}\right) \\ &+E\left[\left(n_{rk}^{(*)} + n_{rk}^{v}\right) \mid D^m, D^v, \Theta^{(h)}\right]\log\left(1-\sum_{i=1}^{r-1} q_{ik}\right) + \sum_{k=1}^{r-1} E\left[\left(n_{\bullet k}^{v} + n_{\bullet k}^{(*)}\right) \mid D^m, D^v, \Theta^{(h)}\right]\log\left(P_k\right) \quad (2.13) \\ &+E\left[\left(n_{\bullet r}^{v} + n_{\bullet r}^{(*)}\right) \mid D^m, D^v, \Theta^{(h)}\right]\log\left(1-\sum_{k=1}^{r-1} P_k\right). \end{aligned}$$

The conditional expectations are only for missing data. Under this parameterisation, missing data exist in the main sample. The expectation step (E-step) can be performed using the following result.

Result 2.1

The conditional expectations of the missing data in the main sample are estimated using the following expressions

$$\hat{E}\left(n_{ik}^{(*)} \mid D^m, \Theta^{(h)}\right) = n_{i\bullet}\left[\frac{\hat{P}_k^{(h)} \hat{q}_{ik}^{(h)}}{\sum_{k=1}^{r} \hat{P}_k^{(h)} \hat{q}_{ik}^{(h)}}\right] \text{ and } \hat{E}\left(n_{\bullet k}^{(*)} \mid D^m, \Theta^{(h)}\right) = \sum_{i=1}^{r} \hat{E}\left(n_{ik}^{(*)} \mid D^m, \Theta^{(h)}\right). \quad (2.14)$$

Proof

The number of sample units that belong in the ik cell of the cross-classification of the observed by the true classification is denoted by $n_{ik}^{(*)}$. Note that while a superscript $(*)$ refers to the unobserved quantities, a superscript $*$ refers to the observed classifications. The expectation of an unobserved quantity is given by

$$E\left(n_{ik}^{(*)}\right) = nE\left(Y_{\xi t}^{*} = i, Y_{\xi t} = k\right). \quad (2.15)$$

Equation (2.15) can be re-expressed as follows

$$E\left(n_{ik}^{(*)}\right) = nE\left(Y_{\xi t}^{*} = i \mid Y_{\xi t} = k\right)E\left(Y_{\xi t} = k\right).$$

From the main sample we have information about the observed classifications. This information can be expressed by summing the unobserved quantities within row i in Table 2.2

$$n_{i\bullet} = n\sum_{k=1}^{r} E\left(Y_{\xi t}^{*} = i \mid Y_{\xi t} = k\right) E\left(Y_{\xi t} = k\right).$$

Given the data constraints, the conditional expectations of the missing data can now be expressed as follows

$$E\left(n_{ik}^{(*)} \mid D^{m}\right) = n_{i\bullet}\left[\frac{E\left(Y_{\xi t}^{*} = i \mid Y_{\xi t} = k\right) E\left(Y_{\xi t} = k\right)}{\sum_{k=1}^{r} E\left(Y_{\xi t}^{*} = i \mid Y_{\xi t} = k\right) E\left(Y_{\xi t} = k\right)}\right]. \tag{2.16}$$

The expectations of the random variables involved in the expression above can be determined using well known results for binomial random variables. More specifically,

$$E\left(Y_{\xi t}^{*} = i \mid Y_{\xi t} = k\right) = q_{ik},\ E\left(Y_{\xi t} = k\right) = P_{k}. \tag{2.17}$$

Substituting (2.17) in (2.16) we obtain the required result

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

It follows that

$$\hat{E}\left(n_{\bullet k}^{(*)} \mid D^{m}, \Theta^{(h)}\right) = \sum_{i=1}^{r} \hat{E}\left(n_{ik}^{(*)} \mid D^{m}, \Theta^{(h)}\right).$$

□

M-step

For the maximisation step (M-step), we need to obtain the score functions defined by (2.13). These score functions are obtained by computing the partial derivatives of the log-likelihood of the augmented data with respect to the vector of parameters. The maximum likelihood estimators are then obtained by setting these derivatives equal to zero, i.e.

$$\frac{\partial E\left[l\left(\Theta; D^{c}\right) \mid D^{v}, D^{m}, \Theta^{(h)}\right]}{\partial \Theta} = 0 \tag{2.18}$$

and solving for Θ.

Result 2.2

The maximum likelihood estimators are given by the following expressions.

$$\hat{q}_{ik} = \frac{\hat{E}\left(n_{ik}^{(*)} \mid D^{m}, \Theta^{(h)}\right) + n_{ik}^{v}}{\hat{E}\left(n_{\bullet k}^{(*)} \mid D^{m}, \Theta^{(h)}\right) + n_{\bullet k}^{v}} \text{ and } \hat{P}_{k} = \frac{\hat{E}\left(n_{\bullet k}^{(*)} \mid D^{m}, \Theta^{(h)}\right) + n_{\bullet k}^{v}}{\sum_{k=1}^{r} \hat{E}\left(n_{\bullet k}^{(*)} \mid D^{m}, \Theta^{(h)}\right) + n_{\bullet k}^{v}}. \tag{2.19}$$

Proof

The system of normal equations we need to solve is defined by setting the score functions equal to zero.

$$\frac{\partial E\left[l\left(\Theta; D^c\right) \mid D^v, D^m, \Theta^{(h)}\right]}{\partial \Theta} = 0.$$

The $\left(r^2 - r\right) \times \left(r^2 - r\right)$ system of normal equations and the corresponding maximum likelihood estimator for q_{ik} is given below.

$$\left.\begin{array}{l} \dfrac{E\left(n_{11}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{11}^v}{q_{11}} - \dfrac{E\left(n_{r1}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{r1}^v}{\left(1 - q_{11} - \cdots - q_{r-11}\right)} = 0 \\ \vdots \\ \dfrac{E\left(n_{r-1r}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{r-1r}^v}{q_{r-1r}} - \dfrac{E\left(n_{rr}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{rr}^v}{\left(1 - q_{1r} - \cdots - q_{r-1r}\right)} = 0 \end{array}\right\} \tag{2.20}$$

$$\hat{q}_{ik} = \frac{\hat{E}\left(n_{ik}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{ik}^v}{\hat{E}\left(n_{\bullet k}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{\bullet k}^v}$$

Similarly, the $(r-1) \times (r-1)$ system of normal equations and the corresponding maximum likelihood estimator for P_k is given below.

$$\left.\begin{array}{l} \dfrac{E\left(n_{\bullet 1}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{\bullet 1}^v}{P_1} - \dfrac{E\left(n_{\bullet r}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{\bullet r}^v}{\left(1 - P_1 - \cdots - P_{r-1}\right)} = 0 \\ \vdots \\ \dfrac{E\left(n_{\bullet r-1}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{\bullet r-1}^v}{P_{r-1}} - \dfrac{E\left(n_{\bullet r}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{\bullet r}^v}{\left(1 - P_1 - \cdots - P_{r-1}\right)} = 0 \end{array}\right\} \tag{2.21}$$

$$\hat{P}_k = \frac{\hat{E}\left(n_{\bullet k}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{\bullet k}^v}{\sum_{k=1}^{r} \hat{E}\left(n_{\bullet k}^{(*)} \mid D^m, \Theta^{(h)}\right) + n_{\bullet k}^v}. \qquad \square$$

Identification of the Model Parameters and Convergence of the EM-algorithm

Identification of the model parameters can be checked by initialising the EM algorithm from different starting values and by seeing whether the algorithm converges to the same solution. Conditional expectations are estimated using Result 2.1. For these conditional expectations, new maximum likelihood estimates are obtained in the maximisation step (M-step) using

Result 2.2. The E and M steps are iterated until convergence. We assume that convergence is achieved when the difference between the maximum likelihood estimates obtained from two successive iterations is less than a small value δ. Denote by $\omega = r^2 - 1$ the dimension of the parameter space. The convergence criterion that we use is the L^2-norm of the vector of parameters obtained in two successive iterations $\Theta^{(h)}$ and $\Theta^{(h+1)}$. This is defined by the following expression

$$\left\|\Theta^{(h)} - \Theta^{(h+1)}\right\| = \sqrt{\sum_{i=1}^{r^2-1}\left(\theta_i^{(h)} - \theta_i^{(h+1)}\right)^2}. \tag{2.22}$$

The parameterisation presented in this section is not specific to the first double sampling scheme. Assume that the validation sample is obtained by sub-sampling n^v units from the main sample of n units. The main sample and the validation sample now share common units. Thus, independence between the main sample and the validation sample is not directly implied. However, the main sample can be divided in two parts. There are $n - n^v$ sample units that participate only in the main survey and n^v sample units that participate both in the main and in the validation survey. We now have independence between the $n - n^v$ units and the n^v units. Therefore, under both double sampling schemes the model can be formulated in exactly the same way. Variance estimation for the maximum likelihood estimates under this parameterisation will be discussed in Chapter 5.

__Application 2.1:__ Comparing the Alternative Parameterisations of the Measurement Error Model in a Cross-sectional Framework

We contrast the parameterisation of the measurement error model in a missing data framework with the parameterisation given by Tenenbein (1972). To facilitate the comparison, we utilise the numerical example that appears in Tenenbein (1972 p.197). The data of this example are given below.

Table 2.3: Validation sample derived from Tenenbein (1972 p.197)

		True Classifications			
		Defective (1)	Satisfactory (2)	Superior (3)	Margins
Fallible Classifications	Defective (1)	12	6	0	18
	Satisfactory (2)	0	20	0	20
	Superior (3)	0	1	19	20
	Margins	12	27	19	n^v =58

Table 2.4: Main sample derived from Tenenbein (1972 p.197)

		True Classifications			
		Defective (1)	Satisfactory (2)	Superior (3)	Margins
Fallible Classifications	Defective (1)	$n_{11}^{(*)}$	$n_{12}^{(*)}$	$n_{13}^{(*)}$	47
	Satisfactory(2)	$n_{21}^{(*)}$	$n_{22}^{(*)}$	$n_{23}^{(*)}$	53
	Superior (3)	$n_{31}^{(*)}$	$n_{32}^{(*)}$	$n_{33}^{(*)}$	49
	Margins	$n_{\cdot 1}^{(*)}$	$n_{\cdot 2}^{(*)}$	$n_{\cdot 3}^{(*)}$	$n = 149$

The algorithm is initialised using arbitrarily selected parameter values. For the specific application, a difference between successive values of the parameters in the order of $\delta = 10^{-4}$ can be achieved within 40 iterations. The results from the application of the EM algorithm are given below along with the results that appear in Tenenbein (1972).

Table 2.5: Contrasting the alternative parameterisations of the measurement error model

Parameters	***Results from the application of the EM algorithm using the misclassification probabilities (3 decimal places)***	***Results reported in Tenenbein (1972) using the calibration probabilities***
P_1	0.209	0.209
P_2	0.474	0.474
P_3	0.317	0.317
q_{11}	1	1
q_{21}	0	0
q_{31}	0	0
q_{12}	0.221	0.221
q_{22}	0.745	0.745
q_{32}	0.034	0.034
q_{13}	0	0
q_{23}	0	0
q_{33}	1	1

As expected, the maximum likelihood estimates obtained under the two parameterisations are the same. The application presented here will serve as a basis when attempting to develop maximum likelihood estimators for adjusted gross flows. However, in a longitudinal context we need to introduce additional assumptions in order to identify the measurement error model. The assumptions we need to impose will depend on whether we specify the model using the calibration or the misclassification probabilities.

2.2.1.4 Quasi-likelihood Estimation for Discrete Cross-sectional Data in the Presence of Misclassification and Double Sampling

In this section, we propose a quasi-likelihood parameterisation of the measurement error model as an alteranative to maximum likelihood estimation. The approach we follow was introduced by Wedderburn (1974) as a basis for fitting generalised linear regression models. As described in Heyde (1997), Wedderburn observed that from a computational point of view the only assumptions for fitting such a model are the specification of the mean and of the relationship between the mean and the variance and not necessarily a fully specified likelihood. Under this approach, Wedderburn replaced the assumptions about the underlying probability distribution by assumptions based solely on the mean variance relationship, leading to an estimating function with properties similar to those of the derivative of a log-likelihood. This estimating function is usually referred to as the quasi-score estimating function. The quasi-likelihood estimator is then defined as the solution of the system of equations defined by the quasi-score estimating function. To illustrate, consider the following model

$$Y = \mu(\Theta) + \varepsilon \qquad (2.23)$$

where Y is a $n \times 1$ data vector and $E(\varepsilon) = 0$. The quasi-score estimating function is then defined (see Heyde 1997 Theorem 2.3) as

$$G(\Theta) = \left(\frac{\partial\mu(\Theta)}{\partial\Theta}\right)^T [Var(\varepsilon)]^{-1}\varepsilon. \qquad (2.24)$$

The quasi-score estimating function defined by (2.24) is also referred to in the literature as Wedderburn's quasi-score estimating function. Here, a quasi-likelihood parameterisation of the measurement error model offers an alternative to the EM algorithm way of resolving a missing data problem. The advantage of this approach is that it does not require any explicit definition of the likelihood function.

Formulating the Model

Denote by P_k^v the probability of correct classification in category k for units in the validation sample, by q_{ik}^v the probability of misclassification for units in the validation sample, by $n_{i.}$ the number of units in the main survey classified in category i by the standard measurement device and by n the sample size of the main survey. Recall that a superscript v is used to denote quantities that are estimated using the data from the validation sample. Without loss of generality, we describe the model for the case of two mutually exclusive states to which a sample unit can be classified. Instead of specifying the form of the likelihood function (2.12), the model can now be described by a system of equations. The number of equations we need is defined by the smallest possible set of independent and unbiased estimating equations that can be established for the underlying problem. For the two-state cross-sectional measurement error model a possible system of equations is

$$\left.\begin{aligned} \overset{\wedge v}{P_1} &= P_1 + \varepsilon_1 \\ \overset{\wedge v}{q_{11}} &= q_{11} + \varepsilon_2 \\ \overset{\wedge v}{q_{12}} &= q_{12} + \varepsilon_3 \\ n_{1.} &= n\left[P_1 q_{11} + (1-P_1) q_{12}\right] + \varepsilon_4 . \end{aligned}\right\} \qquad (2.25)$$

Note that in (2.25) $n_{1.} = n\,\overset{\wedge}{pr}\left(Y_{\xi t}^* = 1\right)$. The left hand side of the equations given in (2.25) describes estimates obtained from the main sample and the validation sample whereas the right hand side describes the unknown parameters of interest plus an error term. Equations described by (2.25) incorporate the extra constraints that are also utilised by the maximum likelihood approach. For the current model, $P_2 = 1 - P_1, q_{21} = 1 - q_{11}$ and $q_{22} = 1 - q_{12}$. As in the maximum likelihood approach, we assume that the main and the validation sample share common parameters due to the fact that both are representative of the same population. Instead of (2.25), one can define another set of independent equations. For example, we can define the first three equations of (2.25) in count and not in probability terms. The estimation process, however, will be invariant under such transformations.

Assuming the general form of the model defined by (2.23), denote by $\mu(\Theta)$ the vector of means and by $\Theta = (P_1, q_{11}, q_{12})$ the vector of parameters. Following Heyde (1997), Wedderburn's quasi-score estimating function is then defined as

$$G(\Theta)=\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\right)^{T}\left[Var(\varepsilon)\right]^{-1}\varepsilon. \tag{2.26}$$

Setting the quasi-score estimating function equal to zero and solving for Θ, we obtain the quasi-score normal equations. The target is to solve the system of the quasi-score normal equations and obtain estimates of the unknown parameters. This can be achieved using numerical techniques. In terms of (2.25), equation (2.26) for the two-state model can be expressed as follows:

$$G(\Theta)=\begin{pmatrix}1 & 0 & 0 & n(q_{11}-q_{12})\\ 0 & 1 & 0 & nP_1\\ 0 & 0 & 1 & n(1-P_1)\end{pmatrix}\begin{pmatrix}\sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14}\\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24}\\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34}\\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2\end{pmatrix}^{-1}\begin{pmatrix}\hat{P}_1^{v}-P_1\\ \hat{q}_{11}^{v}-q_{11}\\ \hat{q}_{12}^{v}-q_{12}\\ n_{1\bullet}-n\left[P_1q_{11}+(1-P_1)q_{12}\right]\end{pmatrix}. \tag{2.27}$$

In (2.27) the middle term denotes the covariance matrix of the error terms. This is defined in the next sub-section. Setting (2.27) equal to zero, leads to three quasi-score normal equations.

Estimating the Covariance Matrix of the Error Terms

In the system of quasi-score normal equations defined by (2.27), the elements of the covariance matrix of the error terms are unknown and need to be estimated using the sample data. In this sub-section, we provide an approximation to elements of this covariance matrix. The variance components (i.e. the diagonal elements) are given by the following expressions

$$\begin{aligned}\sigma_1^2 &= Var\left(\hat{P}_1^{v}\right)\\ \sigma_2^2 &= Var\left(\hat{q}_{11}^{v}\right)\\ \sigma_3^2 &= Var\left(\hat{q}_{12}^{v}\right)\\ \sigma_4^2 &= Var\left(n_{1\bullet}\right).\end{aligned} \tag{2.28}$$

Under simple random sampling, σ_1^2, σ_4^2 can be estimated by

$$\left.\begin{aligned}\hat{\sigma}_1^2 &= \frac{\hat{P}_1^{v}\left(1-\hat{P}_1^{v}\right)}{n^{v}}\\ \hat{\sigma}_4^2 &= n\,\hat{pr}\left(Y_{\xi t}^{*}=1\right)\left[1-\hat{pr}\left(Y_{\xi t}^{*}=1\right)\right].\end{aligned}\right\} \tag{2.29}$$

In order to estimate the covariance matrix of the estimates of the misclassification probabilities, we denote by n_{ik}^{v} the number of sample units in the validation sample classified by the standard measurement device in state i when they truly belong in state k. The misclassification probabilities can be estimated by $\hat{q}_{ik} = \frac{n_{ik}^{v}}{\sum_{i=1}^{r} n_{ik}^{v}}$. While $n^{v} = \sum_{i=1}^{r}\sum_{k=1}^{r} n_{ik}^{v}$ can be considered as fixed, $\sum_{i=1}^{r} n_{ik}^{v}$ must be considered as random. Consequently, in the computation of the covariance matrix of the estimates of the misclassification probabilities we must take into account the non-linearity introduced by the fact that both the numerator and the de-numerator of $\hat{q}_{ik}$ are random quantities. Thus, we apply the Delta method (Bishop, Fienberg and Holland 1975, Agresti 1990). Let $\overset{\wedge}{\Theta}^{*} = \left(n_{11}^{v}, n_{21}^{v}, n_{12}^{v}, n_{22}^{v}\right)$ and $vec\left[Q\left(\overset{\wedge}{\Theta}^{*}\right)\right] = \left[f_1\left(\overset{\wedge}{\Theta}^{*}\right), \ldots, f_{r^2}\left(\overset{\wedge}{\Theta}^{*}\right)\right]^{T}$ be a $r^2 \times 1$ vector of functions of $\overset{\wedge}{\Theta}^{*}$. Note that for simplicity we drop the parenthesis next to the misclassification matrix Q that is specific of the time periods to which the misclassification matrix refers. Applying the delta method to $vec\left[Q\left(\overset{\wedge}{\Theta}^{*}\right)\right]$, we obtain the following approximation

$$vec\left[Q\left(\overset{\wedge}{\Theta}^{*}\right)\right] - vec\left[Q\left(\Theta^{*}\right)\right] \approx \nabla_{\Theta^{*}}\left(\overset{\wedge}{\Theta}^{*} - \Theta^{*}\right), \quad \nabla_{\Theta^{*}} = \frac{\partial vec\left[Q\left(\Theta^{*}\right)\right]}{\partial \Theta^{*}} \Big|_{\Theta^{*}=\overset{\wedge}{\Theta}^{*}} \cdot \qquad (2.30)$$

Taking the variance operator on both sides of (2.30) leads to

$$Var\left\{vec\left[Q\left(\overset{\wedge}{\Theta}^{*}\right)\right]\right\} \approx \nabla_{\Theta^{*}} Var\left(\overset{\wedge}{\Theta}^{*}\right)\left(\nabla_{\Theta^{*}}\right)^{T}. \qquad (2.31)$$

Under simple random sampling, $Var\left(\overset{\wedge}{\Theta}^{*}\right)$ can be estimated using the following results

$$\begin{cases} \overset{\wedge}{Var}\left(n_{ik}^{v}\right) = n^{v}\, \overset{\wedge}{pr}\left(Y_{\xi t}^{*} = i, Y_{\xi t} = k\right)\left[1 - \overset{\wedge}{pr}\left(Y_{\xi t}^{*} = i, Y_{\xi t} = k\right)\right] \\ \overset{\wedge}{Cov}\left(n_{ik}^{v}, n_{i^{*}k^{*}}^{v}\right) = -n^{v}\, \overset{\wedge}{pr}\left(Y_{\xi t}^{*} = i, Y_{\xi t} = k\right)\overset{\wedge}{pr}\left(Y_{\xi t}^{*} = i^{*}, Y_{\xi t} = k^{*}\right) \quad (ik) \neq \left(i^{*}k^{*}\right) \end{cases} \qquad (2.32)$$

while the general expression of the Jacobian matrix ∇_{Θ^*} is

$$\nabla_{\Theta^*} = \left(\begin{array}{cc|cc} \frac{n^v_{21}}{\left(n^v_{21}+n^v_{11}\right)^2} & \frac{-n^v_{11}}{\left(n^v_{21}+n^v_{11}\right)^2} & 0 & 0 \\ \frac{-n^v_{21}}{\left(n^v_{21}+n^v_{11}\right)^2} & \frac{n^v_{11}}{\left(n^v_{21}+n^v_{11}\right)^2} & 0 & 0 \\ \hline 0 & 0 & \frac{n^v_{22}}{\left(n^v_{12}+n^v_{22}\right)^2} & \frac{-n^v_{12}}{\left(n^v_{12}+n^v_{22}\right)^2} \\ 0 & 0 & \frac{-n^v_{22}}{\left(n^v_{12}+n^v_{22}\right)^2} & \frac{n^v_{12}}{\left(n^v_{12}+n^v_{22}\right)^2} \end{array} \right). \tag{2.33}$$

Substituting (2.32) and (2.33) into (2.31), we obtain estimates for σ_2^2, σ_3^2 and $\sigma_{23} = \sigma_{32}$.

Next we observe that due to the double sampling design, we can further assume independence between the main and the validation sample. This implies that

$$\sigma_{14} = \sigma_{41} = \sigma_{24} = \sigma_{42} = \sigma_{34} = \sigma_{43} = 0.$$

It only remains to estimate the following covariance terms: $\sigma_{12} = \sigma_{21}$ and $\sigma_{13} = \sigma_{31}$. These covariance terms can be generally estimated as follows:

$$\hat{Cov}\left(\hat{q}^{v}_{ik}, \hat{P}^{v}_{k}\right) = \hat{Cov}\left(\frac{n^v_{ik}}{\sum_{i=1}^{r} n^v_{ik}}, \frac{\sum_{i=1}^{r} n^v_{ik}}{n^v} \right). \tag{2.34}$$

Result 2.3

Assume that X, Y, A are three random variables and n is fixed. An approximation for $Cov\left(\frac{X}{Y}, \frac{A}{n}\right)$ is given by

$$Cov\left(\frac{X}{Y}, \frac{A}{n}\right) \approx \frac{1}{nE(Y)}\left[Cov(A,X) - \frac{E(X)}{E(Y)}Cov(A,Y)\right]. \tag{2.35}$$

Proof

Proof of this result is given in Chapter 5 that deals with variance estimation issues.

□

Setting $X = n^v_{ik}$, $Y = \sum_{i=1}^{r} n^v_{ik}$, $A = \sum_{i=1}^{r} n^v_{ik}$ and $n = n^v$ in Result 2.3, we can then estimate the remaining covariance terms of interest.

Solving the System of Quasi-score Normal Equations

Having obtained estimates for the variance terms, the final step in evaluating the quasi-likelihood estimators is to solve the system of equations defined by (2.27). This can be done using a Newton-Raphson algorithm. Define by Θ the vector of parameters of dimension $\omega \times 1$ and by A a $\omega \times \omega$ matrix with elements $A_{ij} = \dfrac{\partial G_i(\Theta)}{\partial \vartheta_j}$ $i, j = 1, \cdots, \omega$. The system of quasi-score normal equations defined by (2.27) can be now solved numerically. Assume a vector of initial solutions $\overset{\wedge}{\Theta}^{(0)}$. The vector of initial solutions can be updated using

$$\overset{\wedge}{\Theta}^{(1)} = \overset{\wedge}{\Theta}^{(0)} - A^{-1}\left[\overset{\wedge}{\Theta}^{(0)}\right] G\left[\overset{\wedge}{\Theta}^{(0)}\right]. \tag{2.36}$$

The iterations continue until a pre-specified convergence criterion is satisfied. This is when the difference between the solutions obtained from two successive iterations of the algorithm, as defined by (2.22), is less than a pre-specified small value δ. Variance estimation for the quasi-likelihood estimates is discussed in Chapter 5. Some of the practical advantages offered by the quasi-likelihood approach are also discussed there. Properties of the maximum likelihood and the quasi-likelihood estimators are empirically compared, using a Monte-Carlo simulation study, in Chapter 6.

Application 2.2: Comparing the Maximum Likelihood Approach with the Quasi-likelihood Approach

We illustrate the quasi-likelihood approach and we compare it with the maximum likelihood approach using the following fictitious example. A firm wishes to assess the quality of the units that it produces. The units can be classified into two categories i.e. either as defective or satisfactory. The firm suspects that a number of satisfactory units are classified as defective. The management team is interested in investigating the trade-off between the loss of satisfactory units and the extra cost of improving the current classifier. There are two classification methods. One, which is currently used, is not very costly but is subject to measurement error (main survey). Altenatively, the firm can use an accurate but more expensive classification method (validation survey). A sample of $n = 60000$ production units is selected and the units are classified using the inexpensive classification method. In order to validate the inexpensive classifier, another sample of $n^v = 10000$ production units is selected and these units are classified using both the expensive and the inexpensive classifier. The data for this numerical example are summarised below.

Table 2.6: Validation sample

		True Classifications		
		Defective (1)	Satisfactory (2)	Margins
Fallible Classifications	Defective (1)	672	918	1590
	Satisfactory (2)	28	8382	8410
	Margins	700	9300	$n^{(v)} = 10000$

Table 2.7: Main sample

		True Classifications		
		Defective (1)	Satisfactory (2)	Margins
Fallible Classifications	Defective (1)	$n_{11}^{(*)}$	$n_{12}^{(*)}$	9000
	Satisfactory(2)	$n_{21}^{(*)}$	$n_{22}^{(*)}$	51000
	Margins	$n_{.1}^{(*)}$	$n_{.2}^{(*)}$	$n = 60000$

Table 2.8: Estimated parameters under the alternative parameterisations of the measurement error model

Alternative Fitting Methods	$\hat{P}_1$	$\hat{q}_{11}$	$\hat{q}_{12}$
MLE (Tenenbein 1972)	0.0667	0.9586	0.0936
MLE via EM	0.0667	0.9576	0.0936
Quasi-likelihood	0.0669	0.9580	0.0932

The model parameters are identified. This is checked by initialising the EM algorithm from different starting points and by seeing whether the algorithm converges to the same solution. Figures 2.4 and 2.5 illustrate this idea. The Newton-Raphson algorithm is also invariant to the choice of starting values. The convergence criterion for the EM algorithm and for the Newton-Raphson algorithm is $\delta = 10^{-4}$. The quasi-likelihood parameterisation produces reasonable estimates that are almost identical to the maximum likelihood estimates. When the underlying distribution is from the exponential family (here we assume a multinomial distribution), the quasi-likelihood estimates will be the same as the maximum likelihood estimates (see Wedderburn 1974). However, the quasi-likelihood approach only requires that one specifies the mean and the variance structure, thus avoiding any explicit definition of the likelihood function.

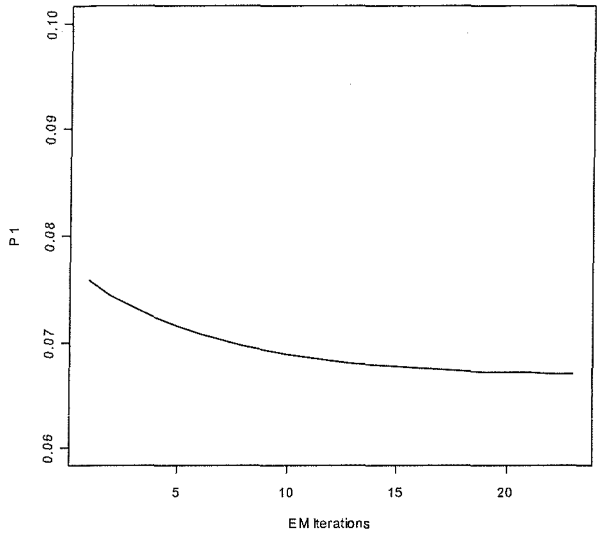


Figure 2.4: Tracing the convergence of the EM algorithm. Starting values close to the maximum likelihood point, convergence criterion $\delta = 10^{-4}$.

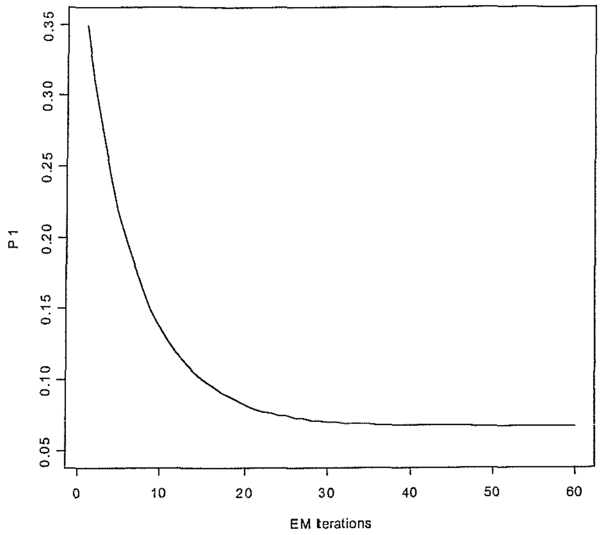


Figure 2.5: Tracing the convergence of the EM algorithm. Starting values further from the maximum likelihood point, convergence criterion $\delta = 10^{-4}$.

2.2.2 The Longitudinal Case

We now turn our attention to the longitudinal case and start by introducing the basic notation. 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, measured by the panel survey, is subject to misclassification. Denote by P_{kl} the probability that sample unit ξ truly belongs in state k at t and state l at $t+1$ and by Π_{ij} the probability that sample 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 sample unit ξ we define the random variables $Y^*_{\xi t}, Y^*_{\xi t+1}$, which describe the way that the ξ^{th} sample unit is classified at t and $t+1$ by the standard measurement device. We also define the random variables $Y_{\xi t}, Y_{\xi t+1}$, which describe the true status of the ξ^{th} sample unit 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 also assume that we can use a cross-sectional validation procedure via which we can make inference about the misclassification process.

Denote by $q_{ijkl} = pr(Y_{\xi t}^{\ *} = i\ \ Y_{\xi t+1}^{*} = j \mid Y_{\xi t} = k\ \ Y_{\xi t+1} = l)$ the misclassification probabilities and by $Q(t,t+1)$ the matrix of misclassification probabilities. Generally speaking, the measurement error model in the longitudinal case 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.

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

and

$$\Pi_{ij} = \sum_{k=1}^{r}\sum_{l=1}^{r} q_{ijkl} P_{kl}\ . \tag{2.37}$$

Writing (2.37) in matrix notation, assuming that $Q(t,t+1)$ is invertible and solving this system of equations with respect to P, we obtain the following expression for the adjusted gross flows

$$\underset{r^2 \times 1}{vec(P)} = \underset{r^2 \times r^2}{\left[Q(t,t+1)\right]^{-1}} \underset{r^2 \times 1}{vec(\Pi)}\ . \tag{2.38}$$

This parameterisation of the measurement error model leads to a moment-type estimator of the adjusted for misclassification quantities. However, estimation of the misclassification matrix $Q(t,t+1)$ is not straightforward. To see this, note that the number of free parameters when attempting to estimate $Q(t,t+1)$ is equal to $r^2\left(r^2 - 1\right)$ i.e. $\left(\underbrace{r^2 \times r^2}_{Total\ number\ of\ parameters} - \underbrace{r^2}_{Available\ Information}\right)$. This implies that information obtained from a cross-sectional validation sample is not sufficient to determine $Q(t,t+1)$. In a longitudinal context, we therefore need to introduce additional assumptions that enable us to estimate the longitudinal misclassification matrix. An assumption that has been used widely in this context is the Independent Classification Errors (ICE) assumption. The ICE assumption is defined as follows

$$pr\left(Y_{\xi t}^{*} = i, Y_{\xi t+1}^{*} = j \mid Y_{\xi t} = k, Y_{\xi t+1} = l\right) = pr\left(Y_{\xi t}^{*} = i \mid Y_{\xi t} = k\right) pr\left(Y_{\xi t+1}^{*} = j \mid Y_{\xi t+1} = l\right). \tag{2.39}$$

From (2.39) we can say that the ICE assumption embodies the following two assumptions.

a) The observed states $Y^*_{\xi t}, Y^*_{\xi t+1}$ are conditionally independent given the true states $Y_{\xi t}, Y_{\xi t+1}$.

b) The misclassification at time t depends only on the current true state and not on the previous or future true states.

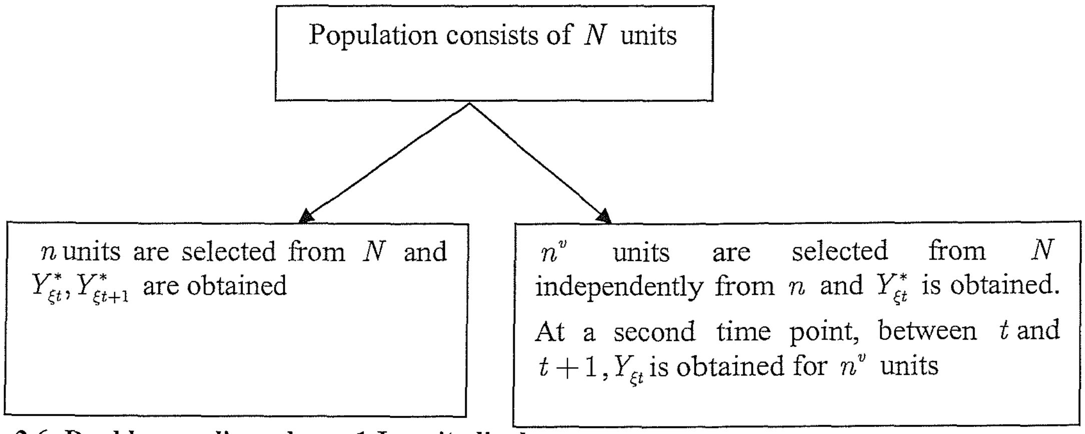
Define by $Q(t)$ the cross-sectional matrix of misclassification probabilities at t with elements q_{ik} and by $Q(t+1)$ the cross-sectional matrix of misclassification probabilities at $t+1$ with elements q_{jl}. An implication of ICE is that the longitudinal misclassification matrix can be expressed as follows

$$Q(t,t+1) = Q(t+1) \otimes Q(t) \ . \qquad (2.40)$$

However, $Q(t+1)$ is not known. We therefore assume that $Q(t) = Q(t+1)$. We now investigate three alternative double sampling schemes that can be used for estimation purposes in a longitudinal framework.

Double Sampling Scheme 1

A simple random sample of n units is selected from a population of N units and the fallible classifications at two time points, $Y^*_{\xi t}, Y^*_{\xi t+1}$, are obtained for each sample unit. For another simple random sample of n^v units, independently selected from the main sample and from the same target population, cross-sectional information on the fallible classifications is also obtained. At a time point, between t and $t+1$, information on the true classifications is obtained for these n^v units. Under this scheme, we obtain panel information on the fallible classifications for n units and further cross-sectional information on the fallible and true classification for n^v units. This double sampling scheme is set out in Figure 2.6.



Population consists of N units

n units are selected from N and $Y^*_{\xi t}, Y^*_{\xi t+1}$ are obtained

n^v units are selected from N independently from n and $Y^*_{\xi t}$ is obtained. At a second time point, between t and $t+1$, $Y_{\xi t}$ is obtained for n^v units

Figure 2.6: Double sampling scheme 1-Longitudinal case

Double Sampling Scheme 2

A simple random sample of n units is selected from a population of N units and the fallible classifications at two time points, $Y^*_{\xi t}, Y^*_{\xi t+1}$, are obtained for each sample unit. At a second time point, between t and $t+1$, the true classifications, $Y_{\xi t}$, are also obtained for a sub-sample of n^v units selected from the n units that already belong to the main sample. This double sampling scheme is set out in Figure 2.7.

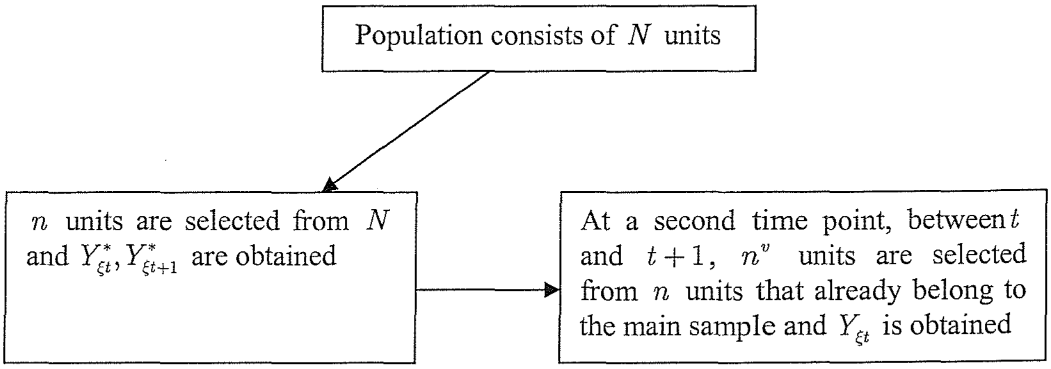

Population consists of N units
n units are selected from N and $Y^*_{\xi t}, Y^*_{\xi t+1}$ are obtained
At a second time point, between t and $t+1$, n^v units are selected from n units that already belong to the main sample and $Y_{\xi t}$ is obtained

Figure 2.7: Double sampling scheme 2-Longitudinal case

Double Sampling Scheme 3

A simple random sample of n units is selected from a population of N units and the fallible classifications at two time points, $Y^*_{\xi t}, Y^*_{\xi t+1}$, are obtained for each sample unit. Using an external source, we obtain cross-sectional information on the incidence of error for n^v units. This double sampling scheme is set out in Figure 2.8.

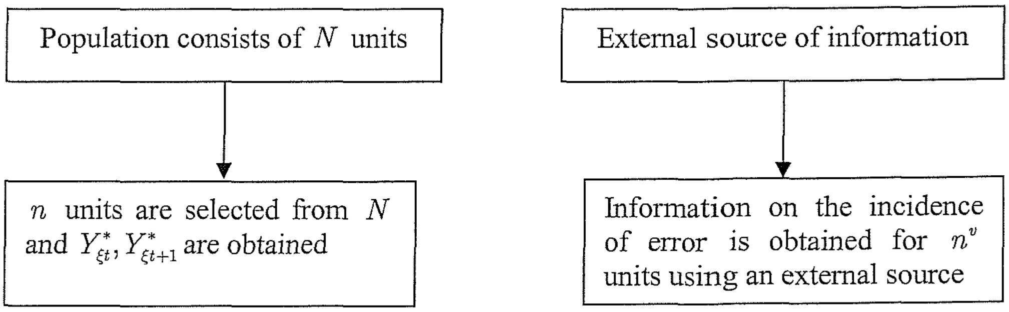

Population consists of N units
External source of information
n units are selected from N and $Y^*_{\xi t}, Y^*_{\xi t+1}$ are obtained
Information on the incidence of error is obtained for n^v units using an external source

Figure 2.8: Double sampling scheme 3-Longitudinal case

2.2.2.1 Review of the Alternative Double Sampling Schemes in a Longitudinal Framework

The absence of a panel validation sample plays a key role in the longitudinal case. If a panel validation sample is available, conclusions from the cross-sectional case can be directly extended to a longitudinal framework. Under the first double sampling scheme, although the validation and the main samples are representative of the same population, information on the fallible classifications from the validation sample cannot be directly combined with information on the fallible classifications from the main sample. We can only make inferences about the cross-sectional incidence of error from the validation sample. This is also true for the other two double sampling schemes. Furthermore, the different double sampling designs have different costs. Under the first scheme and the third scheme, we obtain fallible classifications at two time points for n units and true and fallible classifications at the first time point for n^v units. Thus, under these schemes we have cross-sectional information on the fallible classifications for $n+n^v$ units. Under the second scheme, we obtain fallible classifications at two time points for n units and true classifications for n^v units selected from the n units that already belong to the main sample. As a result, for this scheme we have cross-sectional information on the fallible classifications only for n units. This implies that the first and the third double sampling scheme may be associated with an increased cost compared to the second double sampling scheme. However, the second double sampling scheme can increase the response burden, which is something we may wish to avoid in a longitudinal study.

Recalling (Section 1.7.3) that $Y^*_{\xi t\to t+1}$ denotes a random variable that describes a specific flow of sample unit ξ between t and $t+1$ and assuming that all quantities involved in the measurement error model can be estimated by utilising a double sampling scheme and the ICE assumption, a moment-type estimator, hereinafter conventional or standard point estimator, of (2.38) is given by the following expression.

$$\underset{r^2\times 1}{vec\left(\hat{P}^{st}\right)} = \underset{r^2\mathrm{x}r^2}{\left[\hat{Q}(t)\otimes\hat{Q}(t)\right]^{-1}}\ \underset{r^2\times 1}{vec\left(\hat{\Pi}\right)},\quad \hat{\Pi}_{ij} = \frac{\sum_{\xi=1}^{n} Y^*_{\xi t\to t+1}}{n}\ . \tag{2.41}$$

Estimation of $Q(t,t+1)$ is based on information from the validation sample and the ICE assumption while estimation of Π is based only on the main sample. Unlike in a cross-sectional framework, the choice of the double sampling scheme does not affect the efficiency of the adjusted estimates in a longitudinal framework.

2.2.2.2 Calibration Probabilities versus Misclassification Probabilities and Maximum Likelihood Estimation in a Longitudinal Framework

We now extend the previous discussion about calibration and misclassification probabilities to a longitudinal framework. Meyer (1988) compared two adjustment procedures for correcting labour force gross flows for measurement error. The first procedure is one that utilises misclassification probabilities and has been described in Section 2.2.2.1. The second procedure has been developed by Statistics Canada (1979) and Wong (1983) and is discussed in Stasny (1983). This method aims at correcting gross flows for misclassification by utilising calibration probabilities. Under this approach, the joint distribution of the observed and the true classifications can be expressed as a product of the calibration probabilities and the observed transition probabilities.

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

Denote by $C\left(t,t+1\right)$ the matrix of calibration probabilities. In matrix notation, we obtain

$$vec\left(\underset{r^2\text{x}1}{P}\right)=\underset{r^2\text{x}r^2}{\left[C\left(t,t+1\right)\right]}vec\left(\underset{r^2\text{x}1}{\Pi}\right). \tag{2.42}$$

In order to estimate $C\left(t,t+1\right)$, an Independent Classification Errors assumption is imposed. However, unlike the ICE that conditions on the true classifications, this new conditional independence assumption conditions on the observed classifications and is defined as follows:

$$pr\left(Y_{\xi t}=k,Y_{\xi t+1}=l\mid Y_{\xi t}^{*}=i,Y_{\xi t+1}^{*}=j\right)=pr\left(Y_{\xi t}=k\mid Y_{\xi t}^{*}=i\right)pr\left(Y_{\xi t+1}=l\mid Y_{\xi t+1}^{*}=j\right). \tag{2.43}$$

Using (2.43), expression (2.42) becomes

$$vec\left(\underset{r^2\text{x}1}{P}\right)=\underset{r^2\text{x}r^2}{\left[C\left(t+1\right)\otimes C\left(t\right)\right]}vec\left(\underset{r^2\text{x}1}{\Pi}\right). \tag{2.44}$$

Since $C\left(t+1\right)$ is not known, we further assume that $C\left(t+1\right)=C\left(t\right)$.

Meyer (1988) points out some theoretical deficiencies associated with the conditional independence assumption that utilises the calibration probabilities. More specifically, this type of conditional independence assumption embodies the following two assumptions.

a) The true classifications, $Y_{\xi t}, Y_{\xi t+1}$, are conditionally independent given the observed classifications, $Y^*_{\xi t}, Y^*_{\xi t+1}$.

b) The misclassification at time t depends only on the current observed state and not on previous or future observed states.

Meyer (1988) argues that the main difference between the adjustment procedure that utilises misclassification probabilities and the adjustment procedure that utilises calibration probabilities is in the second assumption. Think of the following example in the context of estimating labour force gross flows. Assume that a sample unit ξ can be classified as employed (E) or unemployed (U). Using (2.43) we define the following conditional probabilities.

$$pr\left(Y_{\xi t+1} = U \mid Y^*_{\xi t} = E, Y^*_{\xi t+1} = E\right) = pr\left(Y_{\xi t+1} = U \mid Y^*_{\xi t+1} = E\right)$$
$$pr\left(Y_{\xi t+1} = U \mid Y^*_{\xi t} = U, Y^*_{\xi t+1} = E\right) = pr\left(Y_{\xi t+1} = U \mid Y^*_{\xi t+1} = E\right).$$

The probability of misclassification at the second time point for someone who is observed to remain stable is the same as the probability of misclassification for someone who is observed to change status between t and $t+1$. However, an observed transition can happen either because the transition is true or because the respondent is misclassified at one time point. As a result, we expect the probability of misclassification of someone who is observed to change status to be higher than the probability of misclassification of someone who is observed to be stable. Hence, the method that utilises calibration probabilities will predict a lower number of spurious transitions than the approach that utilises misclassification probabilities. It has been shown numerically by Meyer (1988) that the adjustments under this method are in the wrong direction i.e. lower diagonal elements and higher off-diagonal elements than actually observed. A further problem, not pointed out by Meyer (1988), is that the approach that utilises the calibration probabilities can be used only in the case of an internal validation sample. This is due to the fact that for an external validation sample only the misclassification probabilities can be regarded as transportable to the population of interest.

One advantage of the method that uses calibration probabilities is that it always produces positive adjustments. On the other hand, the method that utilises misclassification probabilities can lead to negative adjusted estimates due to the inversion of the misclassification matrix. A further advantage is that the distribution theory of an estimator that utilises calibration probabilities is simpler than the distribution theory of an estimator that utilises misclassification probabilities. This will be illustrated in Chapter 5.

In a cross-sectional framework, we have already showed that the parameterisation of the measurement error model using either misclassification or calibration probabilities will produce identical results. Likelihood-based inference in a longitudinal framework is the sole focus of Chapter 3. For the time being, all we can say is that the measurement error models based on *misclassification* or *calibration probabilities will be different*. This is due to the two different forms of conditional independence utilised by these alternative models.

2.3 A Comparison of the Double Sampling Methods in a Cross-sectional and in a Longitudinal Framework

A comparison of the different double sampling methods in a cross-sectional and in a longitudinal framework leads to some interesting findings. Before starting this comparison, we should point out that the main cause of the differences is the lack of a panel validation sample. However, the use of a cross-sectional validation sample in a panel framework is justifiable if we think of the costs associated with a validation survey. The first major difference that exists in the estimation process is the introduction of the ICE assumption. This allows estimation of the measurement error mechanism at two time points based on cross-sectional validation information. While this assumption is not imposed in the cross-sectional case, its implications for the longitudinal case can be quite important. If ICE is not valid, misclassification will be over-predicted and thus we will tend to over-adjust (Skinner and Torelli 1993). This over-correction, under the model that uses misclassification probabilities, can lead to negative adjusted estimates due to the inversion of the misclassification matrix involved in expression (2.41) (Poterba and Summers 1986).

Next, consider the impact of the alternative double sampling schemes on the efficiency of the adjusted estimates. In a cross-sectional framework, the first two double sampling designs produce more efficient estimates than the third design. However, by transforming the external validation sample into an internal validation sample, the estimates produced under the third double sampling scheme will become as efficient as the estimates produced under the first double sampling scheme and the second double sampling scheme. In a longitudinal context, the moment-type (conventional) estimator (see (2.41)) of the adjusted gross flows remains the same under all double sampling schemes.

Finally, we note that another difference between cross-sectional analysis and longitudinal analysis is in the use of calibration probabilities. In a cross-sectional framework, calibration probabilities are used when an internal validation sample is available and lead to maximum likelihood estimates (see (2.7)). These estimates are more efficient than the estimates produced by the moment-type estimator that utilises misclassification probabilities (see for example, (2.3)). Nevertheless, maximum likelihood estimates can be also derived using misclassification probabilities and the alternative parameterisation that we presented in Section 2.2.1.3. In a longitudinal framework, the use of calibration probabilities in conjunction with the Independent Classification Errors assumption, defined by (2.43), is inferior to the approach that utilises misclassification probabilities.

2.4 Alternative Moment-type Estimators for Gross Flows in the Presence of Misclassification

As discussed in Section 2.3, the absence of a panel validation sample plays a key role in the process of estimating adjusted for measurement error gross flows. The main consequence is the introduction of the ICE assumption. The consequences of using the ICE assumption can be quite important. In this section, we focus our interest on the study and development of alternative moment-type estimators. We first study the unbiased margins estimator (Poterba and Summers 1986, Chua and Fuller 1987, Singh and Rao 1995, Skinner 1998) and subsequently we describe a modified and a composite estimator (with fixed and adaptive weights). The reason for looking at these alternative estimators is that we are reluctant to accept the ICE assumption. From our point of view, a more reasonable scenario is that there is a dependence structure in the measurement error mechanism between two time points. All alternative point estimators, described in the following sections, assume the existence of homogeneous gross flows and measurement error mechanisms. Later in this thesis, we relax this assumption and allow for heterogeneity in both mechanisms.

2.4.1 The Unbiased Margins Estimator

The unbiased margins estimator (Poterba and Summers 1986, Chua and Fuller 1987, Singh and Rao 1995, Skinner 1998) is defined by constraining the margins of the adjusted, under the conventional estimator, gross flows matrix to equal the published stocks at each time point. These imposed constraints can be achieved using a raking (IPF) algorithm (Deming and Stephan 1940). The unbiased margins estimator is an alternative to the conventional estimator if we believe that ICE is not valid. The assumption underlying the unbiased margins estimator is that cross-sectional estimates remain unbiased in the presence of measurement error. As we illustrated in Section 1.4.3, this assumption may not be far from reality.

Raking Methodologies

Two raking approaches for obtaining the unbiased margins estimator have been proposed. Poterba and Summers (1986) suggest that raking be applied to the final adjusted gross flows matrix. The main disadvantage of this approach is that if one of the adjusted gross flows is negative, the raking algorithm cannot be used. An alternative raking approach is described in Singh and Rao (1995). They suggest that the cross-sectional misclassification matrix be raked before the final adjustment is carried out. Under the Singh and Rao methodology, $\hat{Q}$ is raked twice. The first raking produces $\hat{Q}^{(c)}(t)$, which is consistent with the published stocks at t and the second raking produces $\hat{Q}^{(c)}(t+1)$, which is consistent with the published stocks at $t+1$. Under this approach and by using (2.41) and properties of vec operators (Harville 1997), the unbiased margins estimator of the adjusted gross flows is given by

$$vec\left(\hat{P}^{um}\right) = \left\{\left[\hat{Q}^{(c)}(t+1)\right]^{-1} \otimes \left[\hat{Q}^{(c)}(t)\right]^{-1}\right\} vec\left(\hat{\Pi}\right). \qquad (2.45)$$

2.4.2 A Modified Estimator for Gross Flows in the Presence of Misclassification

Assume that a double sampling scheme is employed, with the validation sample selected independently from the main sample and from the same target population (i.e. scheme 1 in Section 2.2.2). This scheme can be regarded as reasonable when the main measurement instrument is a panel survey and we want to avoid additional measurements on the same sample units. We now propose a modified estimator based on the following assumptions:

a) The Independent Classification Errors assumption (ICE) is used to estimate longitudinal misclassification probabilities based on estimated cross-sectional misclassification probabilities.
b) The unconditional independence assumption is used to estimate the observed flows of the units in the validation sample based on the cross-sectional observed classifications. Using this assumption, we ignore the correlation structure implied by the longitudinal nature of gross flows.

Under the first double sampling scheme, information on the observed flows of the units in the validation sample is not available. The modified estimator makes use of this absence of panel information. Denote by $\hat{\Pi}^{m}$ the matrix of estimated observed transition probabilities based on data from the main sample. Denote further by $\hat{\Pi}^{v}$ the corresponding estimate based on data from the validation sample and the unconditional independence assumption. The modified estimator is, then, defined as follows

$$vec\left(\hat{P}^{mod}\right)=\underbrace{\left[\hat{Q}(t)\otimes\hat{Q}(t)\right]^{-1}}_{A}\left[\underbrace{w_{mod}\ vec\left(\hat{\Pi}^{m}\right)}_{B}+\underbrace{\left(1-w_{mod}\right)\ vec\left(\hat{\Pi}^{v}\right)}_{Ridge}\right],\ w_{mod}=\frac{n}{n+n^{v}}. \quad (2.46)$$

Unlike the unbiased margins estimator that modifies the measurement error structure (i.e. component A in (2.46)), the modified estimator modifies the observed flows structure. One can view the second term in square brackets, in (2.46), as a ridge component. If ICE is erroneously assumed, the observed flows will be overcorrected (i.e. diagonal flows will be over-increased and off-diagonal flows over-decreased). The effect of the unconditional independence assumption, which underpins the modified estimator, is to overestimate the probability of transition from state i to state j and consequently underestimate the probability of stability. This is because the unconditional independence assumption ignores the correlation structure implied by the longitudinal nature of gross flows. The modified observed flows that are produced by combining component B with the ridge component in (2.46) will have the following pattern. Diagonal elements of the final gross flows matrix will be underestimated compared to diagonal elements of the gross flows matrix derived when using only the main sample. Off-diagonal elements of the final gross-flows matrix will be overestimated compared to off-diagonal elements of the gross flows matrix derived when using only the main sample. As noted by Skinner and Torelli (1993), under ICE we expect to derive an upper bound of the adjustments. Adjustments produced by the modified estimator will be less severe than adjustments produced by the conventional estimator. This is due to the effect of the unconditional independence assumption. Therefore, adjustments under the modified estimator can be regarded as more reasonable if there is a doubt about the validity of the ICE assumption. If ICE is assumed to be valid, the modified estimator will be biased. In such a case, the only gain from using the modified estimator is that it protects against the occurrence of negative adjusted estimates. Thus, under ICE the modified estimator resembles a ridge procedure. On the other hand, if we believe that ICE is not valid, the modified estimator can provide more efficient adjusted estimates than the adjusted estimates derived under the conventional estimator. Gains from using the modified estimator, instead of the conventional estimator, will depend on how severe the problem of misclassification is. For example, if misclassification is not very severe, it may be more reasonable to use the conventional estimator since the effect of the ICE assumption may not be so pronounced. If severe misclassification exists, the impact of the ICE assumption becomes more important and the modified estimator can be considered as an alternative approach.

The main disadvantage of the modified estimator is that it is sample size dependent. Estimation of the observed flows under the modified estimator is based on two parts. One part uses panel information from the main survey i.e. part B in (2.46). The other part uses the unconditional independence assumption and cross-sectional data from the validation survey i.e. the ridge component in (2.46). Normally, the main sample is much larger than the validation sample. However, if the validation sample is large, the modified estimator will depend on the ridge component and the resulting estimates will be unstable.

2.4.3 A Composite Estimator for Gross Flows in the Presence of Misclassification

In previous sections, we investigated alternative moment-type estimators for obtaining adjusted gross flows in the presence of misclassification. Each of these estimators has certain drawbacks. On the one hand, the conventional estimator is based solely on the ICE assumption and therefore can give over-adjustments if ICE is invalid. On the other hand, the modified estimator, proposed above, can give very unstable results if the ridge component of this estimator dominates. The question is whether we can do better by combining the conventional estimator with the modified estimator. One way of achieving this, is by defining a composite estimator of the adjusted gross flows, $\hat{P}^{comp}$, defined as a linear combination of the conventional estimator and the modified estimator. We start by briefly reviewing the general theory of composite estimation and subsequently we focus our interest on the use of composite estimation for adjusting gross flows for measurement error.

General Theory of Composite Estimation

Composite estimation is often used in conjunction with rotation sampling schemes to reduce variability of survey estimators. In composite estimation, we derive a more precise point estimator by borrowing strength from a class of alternative point estimators (Wolter 1979, Tam 1985). Following Kuo (1989), assume that we have two independent and unbiased estimators $\hat{P}_1$ and $\hat{P}_2$ for the same parameter P with known variances σ^2_1, σ^2_2 respectively. Generally speaking, a composite estimator based on these two unbiased estimators is defined as

$$\hat{P}^{comp} = w_{comp}\hat{P}_1 + \left(1 - w_{comp}\right)\hat{P}_2 . \tag{2.47}$$

Two options for defining w_{comp} exist: (a) select a fixed w_{comp} and (b) select w_{comp} such that the mean squared error of $\hat{P}^{comp}$ is minimised. For the second case, the minimum variance unbiased estimator of P is given by

$$\hat{P}^{comp} = w_{comp}\hat{P}_1 + \left(1 - w_{comp}\right)\hat{P}_2, w_{comp} = \frac{\sigma^2_2}{\left(\sigma^2_1 + \sigma^2_2\right)} . \tag{2.48}$$

This result is derived as follows: The general form of the composite estimator is given in (2.47). The mean squared error of $\hat{P}^{comp}$, taking into account that both $\hat{P}_1$ and $\hat{P}_2$ are independent and unbiased estimators of P, is given by

$$MSE\left(\hat{P}^{comp}\right) = Var\left(\hat{P}^{comp}\right) + \left[Bias\left(\hat{P}^{comp}\right)\right]^2 = Var\left(w_{comp}\hat{P}_1 + \left(1-w_{comp}\right)\hat{P}_2\right) \Rightarrow$$

$$MSE\left(\hat{P}^{comp}\right) = \left(w_{comp}\right)^2 Var\left(\hat{P}_1\right) + \left(1-w_{comp}\right)^2 Var\left(\hat{P}_2\right) \Rightarrow$$

$$MSE\left(\hat{P}^{comp}\right) = \left(w_{comp}\right)^2 \sigma_1^2 + \left(1-w_{comp}\right)^2 \sigma_2^2. \tag{2.49}$$

The aim is to find the value of w_{comp} that minimises expression (2.49). We proceed as follows:

$$\frac{\partial MSE\left(\hat{P}^{comp}\right)}{\partial w_{comp}} = 2w_{comp}\sigma_1^2 - 2\sigma_2^2 + 2w_{comp}\sigma_2^2 = 0 \Rightarrow w_{comp} = \frac{\sigma_2^2}{\left(\sigma_1^2 + \sigma_2^2\right)}$$

and,

$$\frac{\partial^2 MSE\left(\hat{P}^{comp}\right)}{\partial^2 w_{comp}} = 2\left(\sigma_1^2 + \sigma_2^2\right) > 0. \tag{2.50}$$

Expression (2.50) states that the value of w_{comp} that minimises the mean square error of the composite estimator is given by $w_{comp} = \frac{\sigma_2^2}{\left(\sigma_1^2 + \sigma_2^2\right)}$. Usually, $\sigma^2{}_1, \sigma^2{}_2$ are unknown and are estimated from the sample data.

A Composite Estimator for Gross Flows in the Presence of Misclassification

We now utilise the idea of the composite estimation for estimating gross flows in the presence of misclassification. A composite estimator can be defined as a linear combination of the conventional estimator (see (2.41)) and the modified estimator (see (2.46)). The general form of this estimator is

$$vec\left(\hat{P}^{comp}\right) = w_{comp} vec\left(\hat{P}^{mod}\right) + \left(1-w_{comp}\right)\ vec\left(\hat{P}^{st}\right). \tag{2.51}$$

Replacing $\hat{P}^{mod}$ by its equivalent using (2.46), the composite estimator becomes

$$vec\left(\hat{P}^{comp}\right) = \left[w_{comp}w_{mod} + \left(1-w_{comp}\right)\right] vec\left(\hat{P}^{st}\right) + w_{comp}\left(1-w_{mod}\right) vec\left(\hat{P}^{v}\right). \tag{2.52}$$

Note that $vec\left(\hat{P}^{v}\right)$ denotes the vector of adjusted gross flows in the validation sample derived by multiplying component A with the ridge component in (2.46).

Regarding the choice of composite weights, w_{comp}, we investigate two options (a) selection of fixed weights and (b) selection of adaptive weights that minimise the mean squared error of the composite estimator.

Composite Estimator with Fixed Weights

For this case, we can choose weights at will. We believe that the conventional estimator should receive a higher weight than the modified estimator. This choice is reasonable given that the modified estimator is based on a much stricter assumption i.e. the unconditional independence assumption, which under certain conditions can have a larger impact on the final adjustments than the impact of the ICE assumption. A set of possible weights is $w_{comp} = 0.3, 0.2, 0.1$.

Composite Estimator with Adaptive Weights

Adaptive weights have to be determined via a minimisation process. The composite estimator with adaptive weights is defined as

$$vec\left(\hat{P}^{comp-ad}\right) = \left[w_{comp}w_{mod} + \left(1 - w_{comp}\right)\right] vec\left(\hat{P}^{st}\right) + w_{comp}\left(1 - w_{mod}\right) vec\left(\hat{P}^{v}\right). \quad (2.53)$$

The mean squared error of $\hat{P}^{comp-ad}$ can be written as follows.

$$\begin{aligned} MSE\left(\hat{P}^{comp-ad}\right) &= \left[w_{comp}w_{mod} + \left(1 - w_{comp}\right)\right]^2 Var\left(\hat{P}^{st}\right) + \left[w_{comp}\left(1 - w_{mod}\right)\right]^2 Var\left(\hat{P}^{v}\right) \\ &+ 2\left[w_{comp}w_{mod} + \left(1 - w_{comp}\right)\right] w_{comp}\left(1 - w_{mod}\right) Cov\left(\hat{P}^{st}, \hat{P}^{v}\right) \\ &+ Bias\left\{\left[w_{comp}w_{mod} + \left(1 - w_{comp}\right)\right]\hat{P}^{st} + w_{comp}\left(1 - w_{mod}\right)\hat{P}^{v}\right\}^2 . \end{aligned} \quad (2.54)$$

The target is to derive composite weights w_{comp} such that (2.54) is minimised.

Result 2.4

An estimated approximate value of w_{comp}, which minimises the mean squared error of the composite estimator, is given by

$$\hat{w}_{comp} \approx \frac{\widehat{Var}\left(\hat{P}^{st}\right)}{\left(1-w_{mod}\right)\left\{\widehat{Var}\left(\hat{P}^{st}\right)+\widehat{Var}\left(\hat{P}^{v}\right)+\left[\hat{E}\left(\hat{P}^{st}\right)-\hat{E}\left(\hat{P}^{v}\right)\right]^{2}\right\}}. \tag{2.55}$$

Proof

Denoting by P the true flows, the mean squared error of the composite estimator with adaptive weights can be expressed in the following form

$$\begin{aligned} MSE\left(\hat{P}^{comp-ad}\right) &= \left[w_{comp}w_{mod}+\left(1-w_{comp}\right)\right]^{2} Var\left(\hat{P}^{st}\right)+\left[w_{comp}\left(1-w_{mod}\right)\right]^{2} Var\left(\hat{P}^{v}\right) \\ &+2\left[w_{comp}w_{mod}+\left(1-w_{comp}\right)\right]w_{comp}\left(1-w_{mod}\right)Cov\left(\hat{P}^{st},\hat{P}^{v}\right) \\ &+Bias\left\{\left[w_{comp}w_{mod}+\left(1-w_{comp}\right)\right]\hat{P}^{st}+\left[w_{comp}\left(1-w_{mod}\right)\right]\hat{P}^{v}\right\}^{2}. \end{aligned} \tag{2.56}$$

We first evaluate the bias term in (2.56)

$$\begin{aligned} &Bias\left\{\left[w_{comp}w_{mod}+\left(1-w_{comp}\right)\right]\hat{P}^{st}+w_{comp}\left(1-w_{mod}\right)\hat{P}^{v}\right\}^{2}= \\ &\quad=\left\{\left[w_{comp}w_{mod}+\left(1-w_{comp}\right)\right]E\left(\hat{P}^{st}\right)+w_{comp}\left(1-w_{mod}\right)E\left(\hat{P}^{v}\right)-P\right\}^{2}= \\ &\quad=\left\{w_{comp}w_{mod}\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]-w_{comp}\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]+E\left[\left(\hat{P}^{st}\right)-P\right]\right\}^{2} \Rightarrow \end{aligned}$$

$$\begin{aligned} \left[Bias\left(\hat{P}^{comp-ad}\right)\right]^{2} &= \left\{w_{comp}\left(w_{mod}-1\right)\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]+\left[E\left(\hat{P}^{st}\right)-P\right]\right\}^{2}= \\ &w_{comp}^{2}\left(w_{mod}-1\right)^{2}\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]^{2}+\left[E\left(\hat{P}^{st}\right)-P\right]^{2} \\ &+2w_{comp}\left(w_{mod}-1\right)\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]\left[E\left(\hat{P}^{st}\right)-P\right]. \end{aligned} \tag{2.57}$$

Next, we substitute (2.57) into (2.56) and we minimise (2.56) with respect to w_{comp}.

$$\begin{aligned}
&2w_{comp}\left[w_{mod}^{2}Var\left(\hat{P}^{st}\right)+Var\left(\hat{P}^{st}\right)+Var\left(\hat{P}^{v}\right)-2w_{mod}Var\left(\hat{P}^{st}\right)+w_{mod}^{2}Var\left(\hat{P}^{v}\right)\right.\\
&-2w_{mod}Var\left(\hat{P}^{v}\right)+2w_{mod}\left(1-w_{mod}\right)Cov\left(\hat{P}^{st},\hat{P}^{v}\right)-2\left(1-w_{mod}\right)Cov\left(\hat{P}^{st},\hat{P}^{v}\right)\\
&\left.+\left(w_{mod}-1\right)^{2}E\left(\hat{P}^{st}-\hat{P}^{v}\right)^{2}\right]=2Var\left(\hat{P}^{st}\right)-2w_{mod}Var\left(\hat{P}^{st}\right)-2\left(1-w_{mod}\right)Cov\left(\hat{P}^{st},\hat{P}^{v}\right)\\
&\qquad\qquad\qquad\qquad-2\left(w_{mod}-1\right)\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]\left[E\left(\hat{P}^{st}\right)-P\right]\Rightarrow
\end{aligned}$$

$$\frac{\partial MSE\left(\hat{P}^{comp-ad}\right)}{\partial w_{comp}}=0\Rightarrow$$

$$\begin{aligned}
&2w_{comp}w_{mod}^{2}Var\left(\hat{P}^{st}\right)+2w_{comp}Var\left(\hat{P}^{st}\right)-2Var\left(\hat{P}^{st}\right)+2w_{mod}Var\left(\hat{P}^{st}\right)-4w_{comp}w_{mod}Var\left(\hat{P}^{st}\right)\\
&+2w_{comp}Var\left(\hat{P}^{v}\right)+2w_{comp}w_{mod}^{2}Var\left(\hat{P}^{v}\right)-4w_{comp}w_{mod}Var\left(\hat{P}^{v}\right)\\
&+4w_{comp}w_{mod}\left(1-w_{mod}\right)Cov\left(\hat{P}^{st},\hat{P}^{v}\right)+2\left(1-w_{mod}\right)Cov\left(\hat{P}^{st},\hat{P}^{v}\right)\\
&-4w_{comp}\left(1-w_{mod}\right)Cov\left(\hat{P}^{st},\hat{P}^{v}\right)2w_{comp}\left(w_{mod}-1\right)^{2}\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]^{2}\\
&+2\left(w_{mod}-1\right)\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]\left[E\left(\hat{P}^{st}\right)-P\right]=0\Rightarrow
\end{aligned}$$

$$\begin{aligned}
&w_{comp}\left(1-w_{mod}\right)^{2}\left[Var\left(\hat{P}^{st}\right)+Var\left(\hat{P}^{v}\right)-2Cov\left(\hat{P}^{st},\hat{P}^{v}\right)+E\left(\hat{P}^{st}-\hat{P}^{v}\right)^{2}\right]=\\
&\left(1-w_{mod}\right)\left\{Var\left(\hat{P}^{st}\right)-Cov\left(\hat{P}^{st},\hat{P}^{v}\right)+\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]\left[E\left(\hat{P}^{st}\right)-P\right]\right\}\Rightarrow
\end{aligned}$$

$$w_{comp}=\frac{\left\{Var\left(\hat{P}^{st}\right)-Cov\left(\hat{P}^{st},\hat{P}^{v}\right)+\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]\left[E\left(\hat{P}^{st}\right)-P\right]\right\}}{\left(1-w_{mod}\right)\left\{Var\left(\hat{P}^{st}\right)+Var\left(\hat{P}^{v}\right)+\left[E\left(\hat{P}^{v}\right)-E\left(\hat{P}^{st}\right)\right]^{2}-2Cov\left(\hat{P}^{st},\hat{P}^{v}\right)\right\}}. \qquad (2.58)$$

We approximate (2.58) assuming that $Cov\left(\hat{P}^{st},\hat{P}^{v}\right)=0$. This assumption may not be far from reality. As we will see later on, the variance of the composite estimator with fixed weights under this assumption provides a good approximation to the true variance of this estimator. Therefore, there is little impact from assuming that $Cov\left(\hat{P}^{st},\hat{P}^{v}\right)=0$

Consequently, an approximately optimal value of w_{comp} is given by

$$w_{comp} \approx \frac{Var\left(\hat{P}^{st}\right)+\left[E\left(\hat{P}^{st}\right)-P\right]\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]}{\left(1-w_{mod}\right)\left\{Var\left(\hat{P}^{st}\right)+Var\left(\hat{P}^{v}\right)+\left[E\left(\hat{P}^{st}\right)-E\left(\hat{P}^{v}\right)\right]^{2}\right\}}. \qquad (2.59)$$

Furthermore,

$$\frac{\partial^{2}MSE\left(\hat{P}^{comp-ad}\right)}{\partial^{2}w_{comp}}=\left(1-w_{mod}\right)^{2}\left[Var\left(\hat{P}^{st}\right)+Var\left(\hat{P}^{v}\right)+E\left(\hat{P}^{st}-\hat{P}^{v}\right)^{2}\right]>0.$$

One way of estimating w_{comp} is by assuming, for example, that $P=E\left(\hat{P}^{st}\right)$. Under this scenario, we favour the conventional estimator since we assume that ICE is valid. This can be considered as a "worst" case scenario for the modified estimator. The composite weights can now be estimated by

$$\hat{w}_{comp} \approx \frac{\hat{Var}\left(\hat{P}^{st}\right)}{\left(1-w_{mod}\right)\left\{\hat{Var}\left(\hat{P}^{st}\right)+\hat{Var}\left(\hat{P}^{v}\right)+\left[\hat{E}\left(\hat{P}^{st}\right)-\hat{E}\left(\hat{P}^{v}\right)\right]^{2}\right\}}. \qquad (2.60)$$

□

In Chapter 5, we develop variance estimators that make it possible to compute these adaptive weights.

2.5 Summary

In the first part of this chapter, we compared alternative double sampling designs within a cross-sectional framework and a longitudinal framework. We also presented some new results for the analysis of misclassified data in a cross-sectional framework. More specifically, we contrasted the parameterisation of the measurement error model presented by Tenenbein (1972) with an alternative parameterisation for maximum likelihood estimation within a missing data framework. The parameterisation of the measurement error model as a missing data problem offers a robust basis for extending the model to handle more complex situations for example, extending the measurement error model to a longitudinal framework. We further proposed a quasi-likelihood approach to fitting the cross-sectional measurement error model. This approach offers an alternative, to the EM algorithm, way for resolving the missing data problem implicit in the maximum likelihood approach.

In the last part of this chapter, we describe alternative moment-type estimators and argue that they provide solutions for problems affecting the conventional point estimator. However, each of the alternative estimators has disadvantages. The unbiased margins estimator is based on the assumption that the cross-sectional estimates are not affected by measurement error. The modified estimator can be very unstable if too much emphasis is placed on the unconditional independence assumption. The composite estimator with fixed weights depends on the subjective choice of these weights while the estimation of adaptive weights is also not free of assumptions. Our aim is to develop an estimator that performs reasonably under ICE but also performs better than the conventional estimator under reasonable departures from ICE. Consequently, it is of interest to investigate the performance of these alternative moment-type estimators under ICE and under departures from ICE. We use Monte-Carlo simulation experiments to examine research questions of this kind later in this thesis.

Chapter 3

Likelihood-based Inference for Gross Flows in the Presence of Misclassification and Double Sampling

3.1 Introduction

One of the main objectives of this thesis is to develop likelihood-based gross flows estimates when auxiliary information obtained via a validation procedure, for example a re-interview survey, is available. Literature on the adjustment of gross flows statistics for misclassification has focused on two approaches. In a double sampling framework, moment-type estimators have been proposed. These estimators alongside with some new moment-type estimators were described in Chapter 2. When validation information is not available, the model-based approaches described in Section 1.5.2 can be utilised. Developing maximum likelihood estimators in a double sampling context will serve two main purposes: Firstly, to improve upon the efficiency of the moment-type estimators and secondly to create a competing approach to the modelling strategies that do not assume validation information.

The structure of this chapter is as follows. We start by presenting a model for gross flows in the presence of misclassification. The model is formulated in a missing data framework and maximum likelihood estimates are derived via the EM algorithm. Although the focus of likelihood-based inference, in this chapter and throughout this thesis, will be on a double sampling scheme where the validation sample is selected independently from the main sample and from the same target population (Section 3.2.1), we also describe the model in the case that the validation sample is selected by sub-sampling units that already participate in the main survey (Section 3.2.2). In an attempt to relax the ICE assumption, a constrained maximum likelihood estimator is also presented. The constrained estimator can be seen as a maximum likelihood analogue of the unbiased margins estimator. The measurement error model is further extended to account for the existence of a complex survey design. This is achieved by utilising the survey weights and the pseudo-maximum likelihood approach. In this context, a pseudo-maximum likelihood estimator and a constrained pseudo-maximum likelihood estimator are also presented. Due to the nature of the UK LFS weights (see discussion in Section 1.7.2), a weighted analysis offers a bias correction to unweighted estimates. The methodology is illustrated in the context of the UK LFS by deriving adjusted for misclassification labour force gross flows.

3.2 Maximum Likelihood Estimation for Gross Flows in the Presence of Misclassification and Double Sampling

In this section, we formulate a measurement error model for gross flows and obtain maximum likelihood estimates for the parameters of interest under the alternative double sampling schemes that we presented in Chapter 2.

3.2.1 Maximum Likelihood Estimation When the Validation Sample is Selected Independently from the Main Sample

Stating the Assumptions and Formulating the Model

Assume a double sampling scheme under which a validation sample of n^v units is selected independently from the main sample of n units and from the same population as the main sample has been also selected (i.e. double sampling scheme 1 in Section 2.2.2). This scheme implies that the main sample and the validation sample do not share common units. The main survey is a panel survey and provides information about the flows of people 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 errors related to these states at time 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. Using as an example the UK LFS, category (1) of the true classifications in Table 3.1 denotes units in the validation sample who were truly employed at t and $t+1$. Category (1) of the fallible classifications in the same table denotes units in the validation sample who are reported to be employed at t and $t+1$. Consequently, respondents that are in category (1) of both classifications correctly classified themselves as employed at both occasions.

We now need to describe the information available from the main survey and from the validation survey. The validation survey provides cross-sectional information on the observed and the true classifications. The main survey provides information on the observed flows. Consider the cross-classification of the fallible with the true classifications in the main and in the validation sample (see Table 3.1 and Table 3.2). The information available from the validation survey can be described schematically by forming all possible $r \times r$ adjacent squares and by summing the elements in each of these squares. For example, in the case that $r = 2$ one can form 4 different adjacent squares each of dimension 2×2 (see Table 3.3). The sum of the elements of the first square denotes the number of people in the validation survey that were reported to be employed and were truly employed at the first time point. The information available from the main survey can be described by summing the elements in each column of Table 3.2. For example, the sum of the elements of the first column represents the number of people in the main survey that reported to be employed at both time points.

Despite the different kind of information that is contained in the main sample and in the validation sample, the way we formulate the model implies a similar structure for both data sources. This structure consists of the observed flows, the true flows and a misclassification mechanism that relates the observed flows to the true flows. The basic idea is to 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 survey and the validation survey. While the main survey is panel, the validation survey is cross-sectional. The other source of missing data is due to the fact that people participating in the main survey do not participate in the validation survey. Unlike the parameterisation presented in Section 2.2.1.3, under the current parameterisation missing data exist both in the main and in the validation sample.

The final assumption that we use is the ICE assumption. This is an identifying assumption and is used in order to estimate longitudinal misclassification probabilities based on cross-sectional misclassification probabilities. Recalling the notation from Chapter 1, ICE is defined as

$$pr\left(Y^*_{\xi t} = i, Y^*_{\xi t+1} = j \mid Y_{\xi t} = k, Y_{\xi t+1} = l\right) = pr\left(Y^*_{\xi t} = i \mid Y_{\xi t} = k\right) pr\left(Y^*_{\xi t+1} = j \mid Y_{\xi t+1} = l\right). \quad (3.1)$$

Denote by n_{ij}, n_{ij}^{v} the number of sample units classified in cell ij defined by the cross-classification of the true with the fallible classifications in the main sample and in the validation sample respectively. Note that a $(*)$ superscript denotes unobserved quantities.

Table 3.1: Validation sample

		Fallible Classifications			
		(1)	$\cdots$	(r^2)	Margins
True Classifications	(1)	$n_{11}^{v(*)}$	$\cdots$	$n_{1r^2}^{v(*)}$	$n_{1\bullet}^{v(*)}$
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	(r^2)	$n_{r^2 1}^{v(*)}$	$\cdots$	$n_{r^2 r^2}^{v(*)}$	$n_{r^2 \bullet}^{v(*)}$
	Margins	$n_{\bullet 1}^{v(*)}$	$\cdots$	$n_{\bullet r^2}^{v(*)}$	n^{v}

Table 3.2: Main sample

		Fallible Classifications			
		(1)	$\cdots$	(r^2)	Margins
True Classifications	(1)	$n_{11}^{(*)}$	$\cdots$	$n_{1r^2}^{(*)}$	$n_{1\bullet}^{(*)}$
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
	(r^2)	$n_{r^2 1}^{(*)}$	$\cdots$	$n_{r^2 r^2}^{(*)}$	$n_{r^2 \bullet}^{(*)}$
	Margins	$n_{\bullet 1}$	$\cdots$	$n_{\bullet r^2}$	n

Denote by P_i the probability that a respondent truly belongs in category i, by q_{ij} the probability that a respondent is classified in category j given that he/she truly belongs in category i and by Θ the vector of parameters. The probability that a sample unit belongs in cell ij is expressed as a product of the true transition probabilities and the misclassification probabilities. Assuming independence between the main sample and the validation sample, the likelihood function of the augmented data for the model described by Tables 3.1 and 3.2 is given by

$$L(\Theta) = \prod_{i=1}^{r^2} \prod_{j=1}^{r^2} \left(P_i q_{ij}\right)^{n_{ij}^{(*)}} \prod_{i=1}^{r^2} \prod_{j=1}^{r^2} \left(P_i q_{ij}\right)^{n_{ij}^{v(*)}}, \tag{3.2}$$

which can be expressed as follows:

$$L(\Theta) = \prod_{i=1}^{r^2}\prod_{j=1}^{r^2}\left(q_{ij}\right)^{n_{ij}^{v(*)}+n_{ij}^{(*)}} \left(P_i\right)^{n_{ij}^{v(*)}+n_{ij}^{(*)}} . \tag{3.3}$$

Taking the logarithms on both sides and imposing the following constraint

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

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

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

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 here 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 process and $r^2 - 1$ gross flows-specific parameters. Thus, under this parameterization we finally need to estimate $2r^2 - r - 1$ parameters.

Estimation

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 (for a general description see Section 2.2.1.3). In the sequel, we describe the expectation step and the maximisation step.

E-step

Recalling the notation from Chapter 2, denote by D^c the complete data, by D^v the observed data from the validation sample, by D^m the observed data from the main sample, by (h) the current iteration of the EM algorithm and by $\Theta^{(h)}$ the vector of parameters in the (h)EM iteration. Taking the conditional expectation of the augmented log-likelihood given the observed data and the current vector of parameters, the augmented log-likelihood is given by

$$E\left[l\left(\Theta;D^{c}\right)\mid D^{m},D^{v},\Theta^{(h)}\right]=\sum_{i=1}^{r^2-1}\left[E\left(n_{i\bullet}^{v(*)}\mid D^{v},\Theta^{(h)}\right)+E\left(n_{i\bullet}^{(*)}\mid D^{m},\Theta^{(h)}\right)\right]\log P_i$$
$$+\left[E\left(n_{r^2\bullet}^{v(*)}\mid D^{v},\Theta^{(h)}\right)+E\left(n_{r^2\bullet}^{(*)}\mid D^{m},\Theta^{(h)}\right)\right]\log\left(1-\sum_{i=1}^{r^2-1}P_i\right) \qquad (3.5)$$
$$+\sum_{i=1}^{r^2}\sum_{j=1}^{r^2}\left[E\left(n_{ij}^{v(*)}\mid D^{v},\Theta^{(h)}\right)+E\left(n_{ij}^{(*)}\mid D^{m},\Theta^{(h)}\right)\right]\log\left(q_{ij}\right).$$

In (3.5), the longitudinal misclassification probabilities, q_{ij}, need to be replaced under ICE by products of the cross-sectional misclassification probabilities and the additional constraint that the sum of the misclassification probabilities for a given true classification must add up to one. In order to perform the E-step, we need to estimate the unobserved quantities in (3.5). This can be done using the following two results:

Result 3.1

The conditional expectations of the missing data in the main sample are estimated using the following expression

$$\hat{E}\left(n_{ij}^{(*)}\mid D^{m},\Theta^{(h)}\right)=n_{\bullet j}\left(\frac{\hat{q}_{ij}^{(h)}\,\hat{P}_i^{(h)}}{\sum_{i=1}^{r^2}\hat{q}_{ij}^{(h)}\,\hat{P}_i^{(h)}}\right). \qquad (3.6)$$

Proof

The number of sample units that belong in cell ij defined by the cross-classification of the observed with the true classifications in the main sample is denoted by $n_{ij}^{(*)}$. Recall that $Y_{\xi t\to t+1}^{*}$ denotes an indicator random variable, which takes value 1 if the ξ^{th} sample unit is classified by the fallible measurement device as making a specific transition between t and $t+1$ and 0 otherwise. Denote further by $Y_{\xi t\to t+1}$ an indicator random variable, which takes value 1 if the ξ^{th} sample unit is classified by the "perfect" measurement device as making a specific transition between t and $t+1$ and 0 otherwise. Note that while a superscript $(*)$ refers to unobserved quantities, a superscript $*$ refers to observed classifications. Using these two random variables, the expectations of the missing data can now be expressed as follows

$$E\left(n_{ij}^{(*)}\right)=nE\left(Y_{\xi t\to t+1}=i,Y_{\xi t\to t+1}^{*}=j\right). \qquad (3.7)$$

Expression (3.7) is re-defined below

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

From the main sample we have information about the observed flows. This information is summarised by summing the unobserved quantities within column j in Table 3.2 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 data constraints, the conditional expectations of the missing data can now be expressed as follows

$$E\left(n_{ij}^{(*)} \mid D^{m}\right) = n_{\bullet j}\left[\frac{E\left(Y^{*}_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right)E\left(Y_{\xi t\to t+1} = i\right)}{\sum\limits_{i=1}^{r^2} E\left(Y^{*}_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right)E\left(Y_{\xi t\to t+1} = i\right)}\right]. \qquad (3.8)$$

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

$$E\left(Y^{*}_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) = q_{ij},\ E\left(Y_{\xi t\to t+1} = i\right) = P_i\ . \qquad (3.9)$$

Substituting (3.9) in (3.8) we obtain the required result

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

It follows that

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

We now proceed to the computation of the conditional expectations of the missing data in the validation sample. As we mentioned at the beginning of this section, the information available from the validation sample can be summarised by forming r^2 adjacent squares each of dimension $r \times r$ and by denoting by n_k^v, k=1,2,$\cdots r^2$ the sum of the elements of each square. This is schematically illustrated in Table 3.3. For example, the sum of the elements of the first square, n_1^v, represents the number of sample units that were observed to be employed and were truly employed at the first time point. In the same way, the sum of the elements of the last square represents the number of sample units that were observed to be unemployed and were truly unemployed at the first time point. For the 4-state model there are four such summations defined by

$$n_1^v = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \, , n_2^v = \sum_{i=1}^{2}\sum_{j=3}^{4} n_{ij} \, , \; n_3^v = \sum_{i=3}^{4}\sum_{j=1}^{2} n_{ij} \, , n_4^v = \sum_{i=3}^{4}\sum_{j=3}^{4} n_{ij} \; . \tag{3.10}$$

Table 3.3: Validation sample in the 4-state model

		Fallible Classifications				
		(1)	(2)	(3)	(4)	Margins
True Classifications	(1)	$n_{11}^{v(*)}$	$n_{12}^{v(*)}$	$n_{13}^{v(*)}$	$n_{14}^{v(*)}$	$n_{1\bullet}^{v(*)}$
	(2)	$n_{21}^{v(*)}$	$n_{22}^{v(*)}$	$n_{23}^{v(*)}$	$n_{24}^{v(*)}$	$n_{2\bullet}^{v(*)}$
	(3)	$n_{31}^{v(*)}$	$n_{32}^{v(*)}$	$n_{33}^{v(*)}$	$n_{34}^{v(*)}$	$n_{3\bullet}^{v(*)}$
	(4)	$n_{41}^{v(*)}$	$n_{42}^{v(*)}$	$n_{43}^{v(*)}$	$n_{44}^{v(*)}$	$n_{4\bullet}^{v(*)}$
	Margins	$n_{\bullet 1}^{v(*)}$	$n_{\bullet 2}^{v(*)}$	$n_{\bullet 3}^{v(*)}$	$n_{\bullet 4}^{v(*)}$	n^v

Result 3.2

The conditional expectations of the missing data in the validation sample are estimated using the expression below

$$\hat{E}\left(n_{ij}^{v(*)} \mid D^v, \Theta^{(h)}\right) = n_k^v \left(\frac{\hat{q}_{ij}^{(h)} \, \hat{P}_i^{(h)}}{\sum_i \sum_j \hat{q}_{ij}^{(h)} \, \hat{P}_i^{(h)}} \right) \tag{3.11}$$

where i, j are running over the rows and columns of the square we are working with.

Proof

The number of sample units that belong in cell ij defined by the cross-classification of the observed with the true classifications in the validation sample is denoted by $n_{ij}^{v(*)}$. Using the same notation as in Result 3.1, the expectations of the missing data in the validation sample are expressed as

$$E\left(n_{ij}^{v(*)}\right) = n^v E\left(Y_{\xi t \to t+1}^{*} = j, Y_{\xi t \to t+1} = i\right) \; . \tag{3.12}$$

Expression (3.12) 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. This information can be expressed as follows:

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

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

$$E\left(n_{ij}^{v(*)} \mid D^v\right) = n_k^v \left[\frac{E\left(Y^*_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) E\left(Y_{\xi t\to t+1} = i\right)}{\sum_{j=1}^{r^2} \sum_{i=1}^{r^2} E\left(Y^*_{\xi t\to t+1} = j \mid Y_{\xi t\to t+1} = i\right) E\left(Y_{\xi t\to t+1} = i\right)} \right]. \qquad (3.13)$$

The expectations of the random variables involved in the expression above can be determined using (3.9). Substituting (3.9) in (3.13) we obtain the required result

$$\hat{E}\left(n_{ij}^{v(*)} \mid D^v, \Theta^{(h)}\right) = n_k^v \left(\frac{\hat{q}{}_{ij}^{(h)} \hat{P}{}_i^{(h)}}{\sum_i \sum_j \hat{q}{}_{ij}^{(h)} \hat{P}{}_i^{(h)}} \right).$$

It follows that

$$\hat{E}\left(n_{i.}^{v(*)} \mid D^v, \Theta^{(h)}\right) = \sum_{j=1}^{r^2} \hat{E}\left(n_{ij}^{v(*)} \mid D^v, \Theta^{(h)}\right).$$

□

Note that in Results 3.1 and 3.2 the parameters q_{ij} need to be replaced under ICE by products of the cross-sectional misclassification probabilities.

M-step

For the maximisation step (M-step), we need to derive the score functions. These score functions are obtained by computing the partial derivatives of the log-likelihood of the augmented data with respect to the vector of parameters. The maximum likelihood estimators are then obtained by setting these derivatives equal to zero i.e.

$$\frac{\partial E\left[l(\Theta; D^c) \mid D^m, D^v, \Theta^{(h)}\right]}{\partial \Theta} = 0 \qquad (3.14)$$

and solving for Θ. For the model described here, the maximisation step is performed numerically using a Newton –type algorithm (Dennis and Schnabel 1983).

Identification of the Model Parameters and Convergence of the EM algorithm

An important requirement for fitting a model is that the model parameters are identified. There are many tests available for checking identifiability empirically. Relevant literature can be found in Goodman (1974). A first test is offered by computing the eigenvalues of the information matrix. If all eigenvalues are positive, the model parameters are identified. In Chapter 5, we provide an approximation to the information matrix. Based on this approximation, this test can be implemented. An alternative solution is offered by initialising the EM algorithm from different sets of starting values. If the algorithm converges to the same region, it is reasonable to assume that the parameters are identified (see also Section 2.2.1.3). This test is implemented in this chapter (see application 3.1). As a convergence diagnostic we use the L^2-norm of the vector of parameters derived from two successive iterations of the EM algorithm defined as

$$\left\|\Theta^{(h)}-\Theta^{(h+1)}\right\|=\sqrt{\sum_{i=1}^{2r^2-r-1}\left(\theta_i^{(h)}-\theta_i^{(h+1)}\right)^2} \; . \tag{3.15}$$

3.2.2 Maximum Likelihood Estimation When the Validation Sample is Selected by Sub-sampling Units from the Main Sample

In Section 3.2.1, we formulated the measurement error model under a double sampling scheme where the validation sample is selected independently from the main sample and from the same target population. In this section, we formulate the measurement error model assuming that the validation sample of n^v units is selected by sub-sampling units from the main sample of n units. Under this scheme, 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.

After 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, 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 $n-n^v$ units between r mutually exclusive states at t and $t+1$. On the other hand, the validation survey provides information on the cross-sectional incidence of misclassification errors related to these states at time t and the observed flows of the n^v units. One can now notice the difference between this model and the model that we presented in Section 3.2.1. If the validation sample is selected independently from the main sample (see Section 3.2.1), the validation sample will provide information only on the cross-sectional incidence of misclassification errors. On the other hand, if the validation sample is selected by sub-sampling units that already participate in the main survey (Section 3.2.1), the validation sample will provide information both on the cross-sectional incidence of misclassification errors and the observed flows.

We form a model similar to the model that we described in Section 3.2.1. The target is to obtain maximum likelihood estimates for the adjusted gross flows. Assuming independence between the units in the main sample and in the validation sample, the augmented data log-likelihood of the model is described by (3.4). Since this likelihood involves missing data, we can maximise it via the EM algorithm. In the E-step we need to estimate the conditional expectations of the missing data in the main sample and in the validation sample using the information available from these two samples. The E-step is described below.

E-step

For the main sample, we have information on the observed flows of the units that participate only in the main survey. Therefore, the conditional expectations of the missing data in the main sample can be simply estimated using Result 3.1.

For the validation survey, we now have information on the cross-sectional misclassification probabilities and the observed flows. As a result, estimating the conditional expectations for the validation sample cannot be based only on Result 3.2. Instead, we use a two-step procedure. For simplicity, we describe this procedure for the 4-state model. The two steps for estimating the conditional expectations of the missing data in the validation sample can be illustrated using Table 3.3.

Step a

In this first 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. This is achieved using the second step.

Step b

Based on these provisional conditional expectations, we compute the following quantities:

$$\begin{aligned} &a = n_{11}^{v(*)} + n_{21}^{v(*)},\ b = n_{12}^{v(*)} + n_{22}^{v(*)},\ \ c = n_{31}^{v(*)} + n_{41}^{v(*)},\ d = n_{32}^{v(*)} + n_{42}^{v(*)} \\ &e = n_{13}^{v(*)} + n_{23}^{v(*)},\ \ f = n_{14}^{v(*)} + n_{24}^{v(*)},\ \ g = n_{33}^{v(*)} + n_{43}^{v(*)}, h = n_{34}^{v(*)} + n_{44}^{v(*)}. \end{aligned} \tag{3.16}$$

We then form two 2×2 tables defined by $\{a,b,c,d\}$ and $\{e,f,g,h\}$ respectively. One can realise that the margins of these tables summarise the information available from the validation sample. 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 use the IPF algorithm to rake the internal cells of these matrices to the information available from the validation survey. The newly derived internal cells are denoted by $\{a^*,b^*,c^*,d^*\}$ and $\{e^*,f^*,g^*,h^*\}$. For example, $a^* + c^*$ will respect $n_{\cdot 1}^{v}$ and $a^* + b^*$ will respect n_{1}^{v}. 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, a 2×1 vector is defined by $n_{11}^{v(*)}, n_{21}^{v(*)}$ such that $a^* = n_{11}^{v(*)} + n_{21}^{v(*)}$. For the 4-state model one can form 8 such vectors. Following the same logic as for Results 3.1 and 3.2, the conditional expectations are then estimated within each of the 2×1 vectors. For example,

$$\hat{E}\left(n_{11}^{v(*)} \mid D^v, \Theta^{(h)}\right) = a^* \left(\frac{\hat{q}_{11}^{(h)} \hat{P}_1^{(h)}}{\sum_{i=1}^{2} \hat{q}_{i1}^{(h)} \hat{P}_i^{(h)}} \right) \text{and } \hat{E}\left(n_{21}^{v(*)} \mid D^v, \Theta^{(h)}\right) = a^* \left(\frac{\hat{q}_{21}^{(h)} \hat{P}_2^{(h)}}{\sum_{i=1}^{2} \hat{q}_{i1}^{(h)} \hat{P}_i^{(h)}} \right). \tag{3.17}$$

These final conditional expectations will respect both the cross-sectional validation information and the observed flows of the units in the validation sample. Having estimated the conditional expectations of the unobserved quantities in the main sample and in the validation sample, the M-step is performed numerically using the procedures that we described in Section 3.2.1. The convergence of the EM algorithm is checked using (3.15).

A “naïve” alternative for estimating the conditional expectations of the missing data in the validation sample, when the validation sample is selected by sub-sampling units from the main sample, is to ignore the observed flows of the units that participate both in the main survey and in the validation survey. This implies that these conditional expectations can be estimated by satisfying only the cross-sectional validation information using the results from Section 3.2.1. The assumption underlying this procedure is that the observed flows of the units that participate both in the main survey and in the validation survey are not different from the observed flows of the units that participate only in the main survey. It is of interest to investigate what we actually lose by ignoring this extra piece of information. This is examined in Chapter 6. The theory we describe in the rest of this chapter allows for a validation sample that is selected independently from the main sample and from the same target population. However, this theory can be easily modified to allow for an alternative double sampling scheme.

3.3 A Constrained Maximum Likelihood Estimator for Gross Flows in the Presence of Misclassification and Double Sampling

In Section 3.2.1, we presented a maximum likelihood estimator for gross flows in the presence of misclassification and double sampling. This estimator utilizes the ICE assumption for estimating longitudinal misclassification probabilities based on cross-sectional misclassification probabilities. However, as discussed in Chapter 2, the consequences of using the ICE assumption can be quite important. In this section, we develop a maximum likelihood estimator that attempts to relax the ICE assumption by imposing an unbiased margins constraint. From now on we will refer to this point estimator as the constrained maximum likelihood estimator. The relaxation to the ICE occurs because under the unbiased margins constraint we use two distinct misclassification matrices $Q(t)$, $Q(t+1)$instead of assuming, as under ICE, that $Q(t)=Q(t+1)$. In order to obtain this estimator, we need to impose constraints on the P_i parameters. The most natural way of resolving the constraint maximisation problem is to impose the raking constraints directly into the log-likelihood function. This is equivalent to full maximum likelihood. However, this approach introduces complexities since the constraints are non-linear functions. Instead, we follow an alternative approach namely, the estimated likelihood approach (Gong and Samaniego 1981, Pawitan 2001). The estimated likelihood approach offers one way of dealing with nuisance parameters. Generally speaking, this method replaces nuisance parameters by their estimates and then treats them as fixed. Here, we treat the parameters of the misclassification mechanism, q_{ij}, as nuisance parameters and gross flows-specific parameters as the parameters of primary interest. Constraints on P_i are imposed implicitly via q_{ij}. It is apparent that this procedure is not equivalent to full maximum likelihood since estimates for the parameters of interest are obtained by maximising only a part of the full likelihood.

Without loss of generality, let us consider the observed labour force gross flows matrix as obtained from a Labour Force Survey. Let us assume that the margins of this matrix represent the published stocks[1] at t and $t+1$. In the first step, we are implementing the raking approach proposed by Singh and Rao (1995). We rake the cross-sectional misclassification matrix, estimated using data from the validation survey, twice. The first raking produces $Q^{(c)}(t)$, which is consistent with the observed estimates at time t and the second raking produces $Q^{(c)}(t+1)$, which is consistent with the observed estimates at time $t+1$. The produced raked misclassification matrices can be seen as two different sources of data. The elements of the first matrix represent cross-sectional validation data such that the first set of constraints is satisfied, whereas the elements of the second matrix represent cross-sectional validation data such that the second set of constraints is satisfied.

The second step involves two maximization problems. We maximize the log-likelihood function (3.4) using as information from the validation sample the data from the first raked matrix. This maximization step will produce maximum likelihood estimates for P_i and the cross-sectional misclassification probabilities under the first set of constraints. We denote these maximum likelihood estimates for the cross-sectional misclassification probabilities by $\overset{\wedge(c)}{q_{ij}}(t)$. Subsequently, we maximize the log-likelihood function (3.4) using as information from the validation sample the data from the second raked matrix. This maximization step will produce maximum likelihood estimates for P_i and the cross-sectional misclassification probabilities under the second set of constraints. We denote these maximum likelihood

[1] Note that usually the published stocks are computed taking into account the survey weights. However, for the time being we are only concerned with unweighted estimates.

estimates for the cross-sectional misclassification probabilities by $\overset{\wedge(c)}{q_{ij}}(t+1)$. Both maximisations are performed using the EM algorithm and Results 3.1 and 3.2.

In the third step, we bring the unbiased margins constraint into P_i via $\overset{\wedge(c)}{q_{ij}}(t)$ and $\overset{\wedge(c)}{q_{ij}}(t+1)$. This is done as follows. Assume that $\overset{\wedge(c)}{q_{ij}}(t)$ and $\overset{\wedge(c)}{q_{ij}}(t+1)$ are fixed at their maximum likelihood values as these are obtained from the second step. The usual likelihood function of the augmented data that is given below

$$L(\Theta)=\prod_{i=1}^{r^2}\prod_{j=1}^{r^2}(P_i)^{n_{ij}^{v(*)}+n_{ij}^{(*)}}\left(q_{ij}\right)^{n_{ij}^{v(*)}+n_{ij}^{(*)}} \tag{3.18}$$

becomes now

$$L(\Theta)=c\prod_{i=1}^{r^2}(P_i)^{n_{i\bullet}^{v(*)}+n_{i\bullet}^{(*)}} \; . \tag{3.19}$$

Term c denotes a constant term resulting from the second component of the likelihood function (3.18). It follows that

$$L(\Theta)\ \alpha\ \prod_{i=1}^{r^2}(P_i)^{n_{i\bullet}^{v(*)}+n_{i\bullet}^{(*)}} \; . \tag{3.20}$$

Taking the logarithms on both sides of (3.20) and imposing the additional constraint that

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

we obtain the following log-likelihood function

$$l(\Theta)\ \alpha\ \sum_{i=1}^{r^2-1}\left(n_{i\bullet}^{v(*)}+n_{i\bullet}^{(*)}\right)\log P_i+\left(n_{r^2\bullet}^{v(*)}+n_{r^2\bullet}^{(*)}\right)\log\left(1-\sum_{i=1}^{r^2-1}P_i\right) \; . \tag{3.21}$$

In the final step, the log-likelihood function given in (3.21) is maximized assuming that $\overset{\wedge(c)}{q_{ij}}(t)$ and $\overset{\wedge(c)}{q_{ij}}(t+1)$ are fixed at their maximum likelihood values as these obtained from the second step of the estimation process. This can be done using the EM algorithm. Taking the conditional expectation of the a log-likelihood (3.21), we have that

$$\begin{aligned}&E\left[l(\Theta;D^c)\mid D^m,D^v,\Theta^{(h)}\right]\ \alpha\\&\sum_{i=1}^{r^2-1}E\left[\left(n_{i\bullet}^{v(*)}+n_{i\bullet}^{(*)}\right)\mid D^m,D^v,\Theta^{(h)}\right]\log P_i+E\left[\left(n_{r^2\bullet}^{v(*)}+n_{r^2\bullet}^{(*)}\right)\mid D^m,D^v,\Theta^{(h)}\right]\log\left(1-\sum_{i=1}^{r^2-1}P_i\right).\end{aligned} \tag{3.22}$$

The E-step can be performed using the following two results.

Result 3.3

The conditional expectations of the missing data in the main sample are estimated using the expressions below

$$\hat{E}\left(n_{ij}^{(*)} \mid D^m, \Theta^{(h)}\right) = n_{\bullet j}\left[\frac{\hat{q}_{ij}^{(c)}(t+1)\ \hat{q}_{ij}^{(c)}(t)\hat{P}_i^{(h)}}{\sum_{i=1}^{r^2}\hat{q}_{ij}^{(c)}(t+1)\ \hat{q}_{ij}^{(c)}(t)\hat{P}_i^{(h)}}\right]. \tag{3.23}$$

It follows that

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

Proof

The proof is identical to the proof that is given for Result 3.1. □

Result 3.4

The conditional expectations of the missing data in the validation sample are estimated using the expressions below

$$\hat{E}\left(n_{ij}^{v(*)} \mid D^v, \Theta^{(h)}\right) = n_k^v\left[\frac{\hat{q}_{ij}^{(c)}(t+1)\ \hat{q}_{ij}^{(c)}(t)\hat{P}_i^{(h)}}{\sum_i\sum_j\hat{q}_{ij}^{(c)}(t+1)\ \hat{q}_{ij}^{(c)}(t)\hat{P}_i^{(h)}}\right]. \tag{3.24}$$

Furthermore, it follows that

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

Proof

The proof is identical to the proof that is given for Result 3.2. □

In Results 3.3 and 3.4, the data constraints, $n_{\bullet j}$, are derived from the main sample. The data constraints, n_k^v, are obtained from the original (i.e. not raked) cross-sectional validation sample. The misclassification probabilities do not have a superscript (h) since they are assumed to be fixed at their maximum likelihood values under the first and the second set of constraints. We also note that in the expressions of the conditional expectations (3.23) and (3.24) the products of the cross-sectional misclassification probabilities, under the first and second set of constraints, appear. These products replace the classical ICE assumption by a "modified" ICE assumption that is also utilised by the unbiased margins estimator (see Section 2.4.1). Having estimated the conditional expectations in the E-step, the M-step is performed numerically. The effect of fixing the cross-sectional misclassification probabilities at their maximum likelihood values is to restrict the estimates derived from the maximisation of the log-likelihood function (3.21) so that they satisfy the marginal constraints.

Before closing this section, we need to make two additional comments. One way to simplify the second step in computing the constrained maximum likelihood estimator, is by simply replacing the cross-sectional misclassification probabilities by their sample estimates as these are obtained from the two raked misclassification matrices. Furthermore, in this thesis we do not discuss variance estimation for the unbiased margins estimator. Approaches for computing the variance of an estimator in the presence of raking can be found in survey literature (Deville and Sarndal 1992). In addition, inference based on the estimated likelihood approach must account for the extra variability introduced by the estimation of the nuisance parameters. An approach for accounting for this additional variability is described in Gong and Samaniego (1981).

3.4 Accounting for the Complex Survey Design

The previously described methodology has been developed within a simple random sampling framework. However, in most of the cases survey data are collected by utilizing complex survey designs. In the following sections, we attempt to account for the existence of a complex survey design.

3.4.1 Pseudo-Maximum Likelihood Estimation: A General Framework

Under simple random sampling, the general framework for maximum likelihood estimation is as follows. Let y_i, i=1,$\cdots$,n denote n independent and identically distributed random variables with known probability density function $f_i(y_i;\theta)$. Assume now that we are interested in making inference about the unknown parameter θ. In order to do so, we need to use the likelihood function given below

$$L(\theta)=\prod_{i=1}^{n} f_i(y_i;\theta). \tag{3.25}$$

The maximum likelihood estimator of θ is given by the value that maximises the logarithm of the likelihood function. This value can be obtained by setting the score function equal to zero and solving the resulting normal equation with respect to the unknown parameter θ

$$\frac{\partial l(\theta)}{\partial \theta}=\sum_{i=1}^{n}\frac{\partial \log\left[L(\theta)\right]}{\partial \theta}=0. \tag{3.26}$$

When the sampling design is complex, the density functions $f_i(y_i;\theta)$ become the conditional densities of the population given the sampling design. If we want to apply maximum likelihood estimation with complex samples, we will need to define the structure of these conditional distributions. This process can be highly complicated since it requires modelling the relationship between y_i and the design variables. An alternative approach, avoiding the complications of defining these conditional distributions, is offered by the Pseudo-Maximum Likelihood approach, hereinafter PML approach. For a general description of the PML approach see Skinner (1989).

Denote by U the population of interest consisting of N units. If population information is available, we can write the log-likelihood function at the population level as follows:

$$l(\theta)=\sum_{i=1}^{N}\log\left[f_i(y_i;\theta)\right]. \tag{3.27}$$

The population level maximum likelihood estimator can be obtained by setting the population score function equal to zero and solving the equation with respect to the unknown parameter θ

$$\frac{\partial l(\theta)}{\partial \theta}=\sum_{i=1}^{N}\frac{\partial \log\left[L(\theta)\right]}{\partial \theta}=0. \tag{3.28}$$

In practice, census information is not available. In the absence of auxiliary information an estimator of the population parameter of interest can be obtained by employing the survey weights. In the simplest case, the survey weights are inversely proportional to the probability of selecting a unit in the sample. Denote by w_ξ the survey weights for the ξ^{th} sampled unit. The PML approach works by replacing the population level score function by a consistent estimate

$$\frac{\partial l(\theta)}{\partial \theta}=\sum_{\xi=1}^{n}w_\xi\frac{\partial \log\left[L(\theta)\right]}{\partial \theta}=0. \tag{3.29}$$

Solving equation (3.29) with respect to the unknown parameter θ, we obtain the pseudo-maximum likelihood estimator $\overset{\wedge}{\theta}^{pml}$. We should, however, note that $\overset{\wedge}{\theta}^{pml}$ is not unique since many consistent estimators of the population score function may exist.

3.4.2 The Gross Flows Model and Pseudo-Maximum Likelihood in the Presence of Misclassification and Double Sampling

We now extend the measurement error model presented in Section 3.2.1 to incorporate survey weights using the PML approach. We assume a double sampling design similar to the one described in Section 3.2.1. We start by formulating the measurement error model pretending that population information is available. Denote by N_{ij}, N_{ij}^{v} the population counts for the main survey and for the validation survey respectively. Before proceeding with the description of the model, we need to explain the choice of notation. When the validation sample refers to the same population as the main sample, N_{ij}, N_{ij}^{v} are the same. In our developments, however, we distinguish between these two quantities. When the validation sample is selected independently from the main sample and from the same target population (Section 3.2.1), the data we use for estimating the conditional expectations of the missing data in the main sample are the observed flows from the main sample. The data that we use for estimating the conditional expectations of the missing data in the validation sample are the cross-sectional validation data. Under this approach, we ignore the information from the main sample when estimating the conditional expectations in the validation sample. This is equivalent to treating N_{ij} differently from N_{ij}^{v}. Nevertheless, one may argue that since both samples refer to the same population, the information from the main sample must be also used in the validation sample otherwise this will impact on the efficiency of the maximum likelihood estimator. If we wish to include this extra piece of information for fitting the model, we will need to use the results of Section 3.2.2.

The log-likelihood function of the augmented data at the population level is defined by replacing sample counts with population counts in (3.4). This log-likelihood function is given by

$$l(\Theta) = \sum_{i=1}^{r^2-1} \left(N_{i\bullet}^{v(*)} + N_{i\bullet}^{(*)}\right) \log P_i + \left(N_{r^2\bullet}^{v(*)} + N_{r^2\bullet}^{(*)}\right) \log\left(1 - \sum_{i=1}^{r^2-1} P_i\right) + \sum_{i=1}^{r^2} \sum_{j=1}^{r^2} \left(N_{ij}^{v(*)} + N_{ij}^{(*)}\right) \log\left(q_{ij}\right) . \quad (3.30)$$

In (3.30), the q_{ij} parameters need to be replaced, under ICE, by the products of cross-sectional misclassification probabilities. In addition, we need to include the extra constraint that the sum of the cross-sectional misclassification probabilities for a given true classification must add up to one.

Estimation

The maximisation of the likelihood function (3.30) is performed using the EM algorithm. Implementation of the EM algorithm into a pseudo-maximum likelihood estimation framework is also considered in Pfeffermann (1988) but for solving a different problem. The author proves that the estimates obtained via the weighted EM algorithm are unbiased and consistent. Denote by $\Theta^{(h)pml}$ the vector of pseudo-maximum likelihood estimates in the (h)EM iteration. We start by writing the conditional expectation of the augmented log-likelihood as in (3.5) but by replacing the sample by population counts.

$$\begin{aligned}E\left[l(\Theta;D^c)\mid D^m,D^v,\Theta^{(h)pml}\right]=&\sum_{i=1}^{r^2-1}\left[E\left(N_{i\bullet}^{v(*)}\mid D^v,\Theta^{(h)pml}\right)+E\left(N_{i\bullet}^{(*)}\mid D^m,\Theta^{(h)pml}\right)\right]\log P_i\\&+\left[E\left(N_{r^2\bullet}^{v(*)}\mid D^v,\Theta^{(h)pml}\right)+E\left(N_{r^2\bullet}^{(*)}\mid D^m,\Theta^{(h)pml}\right)\right]\log\left(1-\sum_{i=1}^{r^2-1}P_i\right)\\&+\sum_{i=1}^{r^2}\sum_{j=1}^{r^2}\left[E\left(N_{ij}^{v(*)}\mid D^v,\Theta^{(h)pml}\right)+E\left(N_{ij}^{(*)}\mid D^m,\Theta^{(h)pml}\right)\right]\log\left(q_{ij}\right).\end{aligned}\tag{3.31}$$

The conditional expectations involved in (3.31) are estimated using the following two results.

Result 3.5

Denote by $w_{\xi j}$ the survey weights for individual ξ performing an observed transition j. The conditional expectations of the missing data in the main sample are estimated using the expression below

$$\hat{E}\left(N_{ij}^{(*)}\mid D^m,\Theta^{(h)pml}\right)=n\left(\frac{\sum_{\xi=1}^{n}w_{\xi j}}{N}\right)\left[\frac{\hat{q}_{ij}^{(h)}\hat{P}_i^{(h)}}{\sum_{i=1}^{r^2}\hat{q}_{ij}^{(h)}\hat{P}_i^{(h)}}\right].\tag{3.32}$$

Proof

Replacing sample counts by population counts and utilising Result 3.1, the conditional expectations of the unobserved quantities in the main sample are estimated using the expression below

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

We note that the data we pretend to have are the population observed gross flows $N_{\bullet j}$. In reality, the $N_{\bullet j}$'s are unknown. We estimate $N_{\bullet j}$ using the survey weights as follows:

$$\hat{N}_{\bullet j} = \sum_{\xi=1}^{n} w_{\xi j}. \qquad (3.33)$$

Replacing the $N_{\bullet j}$'s by their estimates from (3.33), the expression for estimating the conditional expectations of the missing data is now given by

$$\hat{E}\left(N_{ij}^{(*)} \mid D^m, \Theta^{(h)pml}\right) = \sum_{\xi=1}^{n} w_{\xi j}\left[\frac{\hat{q}_{ij}^{(h)}\,\hat{P}_i^{(h)}}{\sum_{i=1}^{r^2}\hat{q}_{ij}^{(h)}\,\hat{P}_i^{(h)}}\right]. \qquad (3.34)$$

In (3.34) we can further replace the survey weights by normalised survey weights. In this case expression (3.34) takes the following form

$$\hat{E}\left(N_{ij}^{(*)} \mid D^m, \Theta^{(h)pml}\right) = n\left(\frac{\sum_{\xi=1}^{n} w_{\xi j}}{N}\right)\left[\frac{\hat{q}_{ij}^{(h)}\,\hat{P}_i^{(h)}}{\sum_{i=1}^{r^2}\hat{q}_{ij}^{(h)}\,\hat{P}_i^{(h)}}\right].$$

It follows that

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

□

Result 3.6

The conditional expectations of the unobserved quantities in the validation sample are estimated using the expression below

$$\hat{E}\left(N_{ij}^{v(*)} \mid D^v, \Theta^{(h)pml}\right) = n^v\left(\frac{\sum_{\xi=1}^{n^v} w_{\xi k}}{N}\right)\left[\frac{\hat{q}_{ij}^{(h)}\,\hat{P}_i^{(h)}}{\sum_{i}\sum_{j}\hat{q}_{ij}^{(h)}\,\hat{P}_i^{(h)}}\right]. \qquad (3.35)$$

Proof

Replacing the sample counts by population counts and utilising Result 3.2, the conditional expectations of the unobserved quantities in the validation survey are estimated by

$$\hat{E}\left(N_{ij}^{v(*)} \mid D^v, \Theta^{(h)pml}\right) = N_k^v \left[\frac{\hat{q}{}_{ij}^{(h)} \hat{P}{}_i^{(h)}}{\sum_i \sum_j \hat{q}{}_{ij}^{(h)} \hat{P}{}_i^{(h)}} \right].$$

The data that we pretend to have are the cross-sectional population counts on the incidence of error. In reality, these N_k^v's are unknown and need to be estimated. This can be done using the survey weights of the validation survey and the following expression

$$\hat{N}{}_k^v = \sum_{\xi=1}^{n^v} w_{\xi k}. \tag{3.36}$$

Replacing the N_k^v's by their estimates from (3.36), we obtain the following expression

$$\hat{E}\left(N_{ij}^{v(*)} \mid D^v, \Theta^{(h)pml}\right) = \sum_{\xi=1}^{n^v} w_{\xi k} \left[\frac{\hat{q}{}_{ij}^{(h)} \hat{P}{}_i^{(h)}}{\sum_i \sum_j \hat{q}{}_{ij}^{(h)} \hat{P}{}_i^{(h)}} \right].$$

Replacing survey weights by normalised survey weights, we obtain the required results

$$\hat{E}\left(N_{ij}^{v(*)} \mid D^v, \Theta^{(h)pml}\right) = n^v \left(\frac{\sum_{\xi=1}^{n^v} w_{\xi k}}{N} \right) \left[\frac{\hat{q}{}_{ij}^{(h)} \hat{P}{}_i^{(h)}}{\sum_i \sum_j \hat{q}{}_{ij}^{(h)} \hat{P}{}_i^{(h)}} \right].$$

It follows that

$$\hat{E}\left(N_{i\cdot}^{v(*)} \mid D^v, \Theta^{(h)pml}\right) = \sum_{j=1}^{r^2} \hat{E}\left(N_{ij}^{v(*)} \mid D^v, \Theta^{(h)pml}\right).$$

□

After estimating the conditional expectations of the unobserved quantities in the main sample and in the validation sample, full information is obtained that is used to maximise the log-likelihood function (3.30). The M-step is performed numerically. Note that the score functions are now the weighted score functions. These provide estimates of the population level score functions. The algorithm runs until the convergence criterion defined by (3.15) is satisfied. At the end of this algorithm a vector of pseudo maximum likelihood estimators, Θ^{pml}, is derived.

3.4.3 A Note on the Estimation of Conditional Expectations in the Case of an External Validation Sample

In this section, we present a procedure for obtaining cross-sectional population estimates for the incidence of error when weights for the validation survey are not available or not appropriate for the population of interest. This can happen, for example, in the case that we employ an external validation sample. The procedure is as follows.

Population cross-sectional estimates at t are derived using the weights of the main survey. Assume now that the misclassification process in the external validation sample is informative of the misclassification process in the target population. Under this assumption, the misclassification probabilities from the validation sample can be employed to correct population cross-sectional estimates at t for measurement error. This can be achieved by using one of the estimators that we presented in Section 2.2.1.1. At the end of this process we obtain cross-sectional observed and adjusted for misclassification population estimates. The final step in this process involves calibrating $pr\left(Y^*_{\xi t} = i, Y_{\xi t} = k\right)$, from the external validation sample, to these two estimated population margins. The internal cells of the raked matrix can then be multiplied by the population size to produce cross-sectional estimated population counts for the incidence of error. These estimates are employed to estimate the conditional expectations of the missing data in the validation sample when using the EM algorithm and the PML approach.

The same procedure can be used also for unweighted analysis. The only difference now is that the marginal observed and adjusted estimates are derived without utilising the weights of the main survey, but simply by using unweighted data from the main survey. This is equivalent to transforming an external validation sample into an internal validation sample. Both procedures will be illustrated later when employing misclassification probabilities from an external validation sample.

3.4.4 A Constrained Pseudo-Maximum Likelihood Estimator

A natural extension to the pseudo-maximum likelihood estimator is the constrained pseudo-maximum likelihood estimator. In this section, we briefly describe this estimator, which can be derived by a straightforward extension of the results presented in Sections 3.3 and 3.4.2. The constrained pseudo-maximum likelihood estimator must approximately respect the marginal population estimates at t and $t+1$. Pretending that census information is available, the log-likelihood function given in (3.30) is utilised. The constrained pseudo-maximum estimator is obtained using the procedure described in Section 3.3. The only modification is the following: We now rake the cross-sectional misclassification matrix twice. The first raking produces $Q^{(c)u}(t)$, which is consistent with the population observed marginal estimates at the first time point and the second raking produces $Q^{(c)u}(t+1)$, which is consistent with the population observed marginal estimates at the second time point. The rest of the estimation process remains the same. The second step involves two maximisation problems under the first and the second set of constraints. These maximisation problems are solved using the EM algorithm and Results 3.5 and 3.6. In the final step, the log-likelihood function (3.30) is maximised only with respect to P_i assuming that q_{ij} are fixed at their maximum likelihood values. This final maximisation problem is also solved using the EM algorithm.

3.4.5 Weighted Moment-type Estimators

Before illustrating the methodology of the previous sections, we need to provide expressions for the weighted versions of some of the moment-type estimators presented in Chapter 2.

The Weighted Conventional Estimator

Denote by $\hat{\Pi}^{u}$ the matrix of population level estimates of the observed transition probabilities. Denote further by $\hat{Q}^{u}(t)$ the weighted cross-sectional misclassification matrix. The weighted equivalent of estimator (2.41) is given by

$$vec\left(\hat{P}^{u-st}\right) = \left[\hat{Q}^{u}(t) \otimes \hat{Q}^{u}(t)\right]^{-1} vec\left(\hat{\Pi}^{u}\right). \tag{3.37}$$

The Weighted Unbiased Margins Estimator

Denote by $\hat{Q}^{(c)u}(t+1)$ the misclassification matrix produced by raking the weighted misclassification matrix, $\hat{Q}^{u}(t)$, to the population observed marginal estimates at $t+1$.

Denote also by $\hat{Q}^{(c)u}(t)$ the misclassification matrix produced by raking the weighted misclassification matrix , $\hat{Q}^{u}(t)$, to the population observed marginal estimates at t. A weighted equivalent of the unbiased margins estimator (see Section 2.4.1) is given by

$$vec\left(\hat{P}^{u-um}\right)=\left\{\left[\hat{Q}^{(c)u}(t+1)\right]^{-1}\otimes\left[\hat{Q}^{(c)u}(t)\right]^{-1}\right\}vec\left(\overset{\wedge}{\Pi}^{u}\right). \qquad (3.38)$$

3.5 Deriving UK Labour Force Gross Flows Adjusted for Misclassification

In this section, we present five applications. The first two applications illustrate the adjustment of UK labour force gross flows for misclassification. This is done using a number of alternative point estimators that were presented in Chapter 2 and Chapter 3. The third application aims at contrasting the conventional (moment-type) estimator with the maximum likelihood estimator in the presence of intense misclassification. In the fourth application, we conduct a sensitivity analysis of the adjusted UK labour force gross flows using different validation datasets. Finally, in the fifth application we derive maximum likelihood estimates of the adjusted gross flows when the validation sample is selected by sub-sampling units that already participate in the main survey. Observed labour force gross flows are estimated by utilising the common sample between two quarters from the UK LFS (see discussion in Section 1.7.3). Due to the absence of a UK validation survey, we utilise external validation data. The joint distribution for the incident of error $pr\left(Y_{\xi t}^{*}=i,Y_{\xi t}=k\right)$, estimated from the external validation survey, is raked to the UK marginal labour force estimates at time t using the procedures described in Section 3.4.3. This is done both for unweighted analysis and for weighted analysis.

<u>**Application 3.1:**</u> Adjusting UK Labour Force Gross Flows - Unweighted Analysis

For this application, we utilise UK labour force gross flows between summer-autumn 1997 and a smoothed version of reconciled validation data from the Swedish (October 1994 - April 1995) LFS re-interview programme (see Section 1.8.4). The estimators we consider are the following: The unadjusted point estimator, the conventional estimator (Section 2.2.2.1), the unbiased margins estimator (Section 2.4.1), the maximum likelihood estimator (Section 3.2.1) and the constrained maximum likelihood estimator (Section 3.3). The matrix of misclassification probabilities is given below. The columns of this matrix denote true states whereas the rows denote observed states. The convergence criterion for the EM algorithm, as this is defined by (3.15), is $\delta = 10^{-4}$. Convergence was achieved within 43 iterations. An empirical investigation of the identifiability of the model is provided by initialising the EM algorithm using different sets of starting values. Two scenarios are examined. Under the first scenario, the EM is initialised using values close to the maximum likelihood point. Under the second scenario, the EM is initialised using values further from the maximum likelihood point. The algorithm always arrived at the same convergence region. This is illustrated by producing figures that trace the convergence of the EM algorithm, for the different gross flows parameters, under the two scenarios (see figures in Appendix II).

$$\hat{Q} = \begin{array}{c|ccc} & E & U & N \\ \hline E & 0.981 & 0.017 & 0.032 \\ U & 0.008 & 0.951 & 0.027 \\ N & 0.011 & 0.032 & 0.941 \end{array} .$$

The UK observed marginal labour force estimates for summer-autumn 1997 are given below. These marginal estimates must be approximately respected both by the unbiased margins estimator and by the constrained maximum likelihood estimator.

Unweighted observed labour force marginal estimates at t

$E = 0.741$, U=0.052, N=0.207

Unweighted observed labour force marginal estimates at $t+1$

$E = 0.748$, U=0.046, N=0.206

Table 3.4: Adjusted UK labour force gross flows for summer-autumn 1997 using the alternative moment-type and maximum likelihood estimators - Unweighted analysis

Flow	*Observed Flows*	*Conventional (ICE)*	*Maximum Likelihood (ICE)*	*Unbiased Margins*	*Constrained Maximum Likelihood*
EE	0.716	0.7420	0.7410	0.7325	0.7330
EU	0.009	0.0028	0.0033	0.0063	0.0060
EN	0.016	0.0024	0.0026	0.0020	0.0025
UE	0.016	0.0102	0.0106	0.0138	0.0132
UU	0.027	0.0292	0.0291	0.0330	0.0327
UN	0.009	0.0033	0.0036	0.0051	0.0046
NE	0.016	0.0026	0.0028	0.0019	0.0024
NU	0.010	0.0045	0.0050	0.0068	0.0064
NN	0.181	0.2030	0.2020	0.1986	0.1992

Application 3.2: Adjusting UK Labour Force Gross Flows - Weighted Analysis

In this application, we bring into the analysis the weights of the UK LFS by using weighted UK labour force gross flows between summer-autumn 1997. As discussed in Section 1.7.2, the weights of the UK LFS serve two purposes i.e. they produce population level estimates and at the same time compensate for the bias due to sampling attrition. Therefore, by including the survey weights into the measurement error model we implicitly provide a bias correction to labour force gross flows estimates also for sampling attrition. The estimators we consider are the weighted unadjusted point estimator, the weighted conventional estimator and the weighted unbiased margins estimator (Section 3.4.5), the pseudo-maximum likelihood estimator (Section 3.4.2) and the constrained pseudo-maximum likelihood estimator (Section 3.4.4). The convergence criterion for the EM algorithm is $\delta = 10^{-4}$. Convergence was achieved within 22 iterations. The UK weighted observed marginal labour force estimates for summer-autumn 1997 are given below. These marginal estimates must be approximately respected both by the weighted unbiased margins estimator and by the constrained pseudo-maximum likelihood estimator.

Weighted observed marginal estimates at t

$E = 0.734$, U=0.059, N=0.207

Weighted observed marginal estimates at $t+1$

$E = 0.737$, U=0.052, N=0.211

Table 3.5: Adjusted UK labour force gross flows for summer-autumn 1997 using the alternative moment-type and maximum likelihood estimators - Weighted analysis

Flow	*Weighted Observed Flows*	*Weighted Conventional (ICE)*	*Pseudo Maximum Likelihood (ICE)*	*Weighted Unbiased Margins*	*Constrained Pseudo Maximum Likelihood*
EE	0.705	0.7312	0.7304	0.7225	0.7226
EU	0.010	0.0038	0.0041	0.0068	0.0067
EN	0.019	0.0049	0.0052	0.0043	0.0051
UE	0.017	0.0113	0.0114	0.0143	0.0139
UU	0.032	0.0349	0.0347	0.0390	0.0380
UN	0.010	0.0042	0.0044	0.0060	0.0057
NE	0.015	0.0005	0.0015	0.0001	0.0013
NU	0.010	0.0046	0.0047	0.0067	0.0065
NN	0.182	0.2046	0.2036	0.2003	0.2002

Application 3.3: Comparing the Moment-type Estimators with the Maximum Likelihood Estimators in the Presence of Intense Misclassification

One of the main disadvantages associated with the use of the conventional point estimator is that it can result in estimates that lie outside the parameter space. This can happen due to the inversion of the misclassification matrix involved in the computation of this estimator. In Chapter 2, we investigated ways to overcome this problem by defining alternative moment-type estimators. An alternative solution can be offered by the maximum likelihood estimator. In this application, we use the original (i.e. not smoothed) misclassification matrix estimated from the Swedish validation survey (October 1994 - April 1995). Some of the elements of this matrix are associated with intense misclassification for example, $q_{EN} = 0.041$. We compare estimates derived when using the moment-type and the maximum likelihood estimators. The matrix of misclassification probabilities is given by

$$\hat{Q} = \begin{array}{cccc} & E & U & N \\ E & 0.980 & 0.016 & 0.041 \\ U & 0.008 & 0.950 & 0.023 \\ N & 0.012 & 0.034 & 0.936 \end{array} .$$

Table 3.6: Comparing the moment-type estimators with the maximum likelihood estimators in the presence of intense misclassification

Flow	*Observed Flows*	*Conventional (ICE)*	*Maximum Likelihood (ICE)*	*Unbiased Margins*	*Constrained Maximum Likelihood*
EE	0.716	0.7450	0.7419	0.7365	0.7346
EU	0.009	0.0025	0.0028	0.0059	0.0056
EN	0.016	-0.0009	0.0016	-0.0014	0.0019
UE	0.016	0.0101	0.0103	0.0135	0.0129
UU	0.027	0.0294	0.0293	0.0331	0.0320
UN	0.009	0.0038	0.0040	0.0053	0.0052
NE	0.016	-0.0012	0.0013	-0.0020	0.0017
NU	0.010	0.0051	0.0052	0.0069	0.0069
NN	0.181	0.2062	0.2036	0.2022	0.1992

The results indicate that when intense misclassification exists, the conventional point estimator can easily produce awkward estimates (in this case negative probabilities) that lie outside the parameter space. On the other hand, the maximum likelihood estimator overcomes this problem by constraining the estimates to lie within the parameter space.

Furthermore, although the unbiased margins estimator is designed to relax ICE, it can still lead to negative adjusted flows (see for example, Table 3.6). In contrast, the constrained maximum likelihood estimator produces estimates that lie within the boundaries of the parameter space. Therefore, we propose that both the unconstrained maximum likelihood estimator and the constrained maximum likelihood estimator should be preferred over the corresponding moment-type estimators.

<u>Application 3.4:</u> Sensitivity Analysis of the Adjusted UK Labour Force Gross Flows

In applications 3.1 and 3.3, we used different versions of the Swedish misclassification probabilities in order to adjust UK labour force gross flows for measurement error. In this application, we conduct a sensitivity analysis. More specifically, we investigate the impact that alternative sets of misclassification matrices have on the unweighted adjusted UK labour force gross flows. For the purposes of this application, we employ the misclassification probabilities from the LFS re-interview survey in Canada (Singh and Rao 1995) and from the CPS re-interview survey in the US (Poterba and Summers 1986). We further use the smoothed version of the Swedish misclassification probabilities (see application 3.1), the original Swedish misclassification probabilities (see application 3.3) and a weighted version, using the weights from the Swedish re-interview survey (October 1994 - April 1995), of the Swedish misclassification matrix. Adjusted gross flows are derived using the maximum likelihood estimator. The smoothed version and the original version of the Swedish misclassification matrices are reported in applications 3.1 and 3.3 respectively. The new misclassification matrices are defined as follows:

$$\hat{Q}^{Canadian\ LFS} = \begin{array}{c} \\ E \\ U \\ N \end{array} \begin{array}{c} \begin{array}{ccc} E & U & N \end{array} \\ \begin{pmatrix} 0.993 & 0.024 & 0.007 \\ 0.002 & 0.90 & 0.008 \\ 0.005 & 0.076 & 0.985 \end{pmatrix} \end{array} \qquad \hat{Q}^{Weighted\ Swedish} = \begin{array}{c} \\ E \\ U \\ N \end{array} \begin{array}{c} \begin{array}{ccc} E & U & N \end{array} \\ \begin{pmatrix} 0.981 & 0.023 & 0.035 \\ 0.004 & 0.907 & 0.007 \\ 0.015 & 0.070 & 0.958 \end{pmatrix} \end{array}$$

$$\hat{Q}^{CPS} = \begin{array}{c} \\ E \\ U \\ N \end{array} \begin{array}{c} \begin{array}{ccc} E & U & N \end{array} \\ \begin{pmatrix} 0.981 & 0.035 & 0.02 \\ 0.003 & 0.83 & 0.01 \\ 0.016 & 0.135 & 0.97 \end{pmatrix} \end{array}.$$

Table 3.7: Sensitivity analysis of the adjusted UK labour force gross flows using the alternative sources of validation data

Flow	*Observed Flows*	*Original Swedish*	*Smoothed Swedish*	*Weighted Swedish*	*CPS*	*Canadian LFS*
EE	0.716	0.7419	0.7410	0.7391	0.7395	0.7254
EU	0.009	0.0028	0.0033	0.0058	0.0067	0.0076
EN	0.016	0.0016	0.0026	0.0012	0.0019	0.0103
UE	0.016	0.0103	0.0106	0.0136	0.0151	0.0154
UU	0.027	0.0293	0.0291	0.0326	0.0385	0.0330
UN	0.009	0.0040	0.0036	0.0062	0.0036	0.0059
NE	0.016	0.0013	0.0028	0.0011	0.0013	0.0101
NU	0.010	0.0052	0.0050	0.0075	0.0050	0.0071
NN	0.181	0.2036	0.2020	0.1929	0.1884	0.1852

To quantify the effect of the different misclassification matrices, we compute the sum of the off-diagonal adjusted flows. This sum represents the overall adjusted probability of changing labour force status between two quarters. We further compute the sum of the off-diagonal unadjusted flows. The closer the adjusted sum is to the unadjusted sum, the less the impact of the adjustment procedure. These sums appear in the table below.

Table 3.8: Investigating the impact of the alternative misclassification matrices

	Unadjusted	*Original Swedish*	*Smoothed Swedish*	*Weighted Swedish*	*CPS*	*Canadian LFS*
Sum	0.076	0.025	0.028	0.035	0.034	0.056

The Canadian set of misclassification probabilities provides the less severe set of adjustments while the original Swedish misclassification probabilities provide the most severe set of adjustments. Using these results, one can construct a range of adjusted UK labour force gross flows.

Application 3.5: Maximum Likelihood Estimation When the Validation Sample is Selected by Sub-sampling Units from the Main Sample - Unweighted Analysis

In this application, we allow for a double sampling scheme under which the validation sample is selected by sub-sampling units that already participate in the main survey. The size of the main survey is $n = 60000$. Between t and $t+1$ we select a sub-sample of 10000 units out of the 60000 units. The units of this sub-sample participate in the cross-sectional validation survey. The information we have consists of the observed flows for $n - n^v = 50000$ units and the observed flows and the cross-sectional misclassification probabilities for

$n^v = 10000$ units. We compute maximum likelihood estimates of the adjusted gross flows using the theory of Section 3.2.2. For comparison reasons, we also include the conventional estimator. Without loss of generality, the theory is illustrated for the 2-state model i.e. Employed and Unemployed or Inactive. The convergence criterion for the EM algorithm is $\delta = 10^{-4}$. Convergence was achieved within 58 iterations. The observed labour force gross flows are estimated from the UK LFS (summer-autumn 1997). The matrix of misclassification probabilities is estimated using the smoothed version of the Swedish re-interview data and is given by

$$\hat{Q} = \begin{array}{cc} & \begin{array}{cc} E & U+N \end{array} \\ \begin{array}{c} E \\ U+N \end{array} & \begin{pmatrix} 0.99 & 0.053 \\ 0.01 & 0.947 \end{pmatrix} \end{array}.$$

Table 3.9: Adjusted labour force gross flows (4-state model) when the validation sample is selected by sub-sampling units from the main sample

Flow	*Observed*	*Conventional (ICE)*	*Maximum Likelihood (ICE)*
E,E	0.716	0.7284	0.7270
E, U+N	0.025	0.0051	0.0054
U+N, E	0.032	0.0131	0.0134
U+N, U+N	0.227	0.2532	0.2542

3.5.1 Discussion on the Adjustments Derived from the Alternative Estimators

The existence of measurement error, when estimating labour force gross flows, results in the overestimation of the labour market mobility. The effect of adjusting labour force gross flows is to increase the diagonal elements and decrease the off-diagonal elements of the unadjusted gross flows matrix. Adjustments derived under ICE are more severe than those produced when relaxing ICE. For example, a relaxation of ICE is provided by the unbiased margins assumption. To see the effect of this assumption, one can compare the sum of the diagonal adjusted flows when using the unbiased margins estimator (or the weighted unbiased margins estimator) with the sum of the diagonal adjusted flows when using the conventional estimator (or the weighted conventional estimator). The sum of the diagonal adjusted flows produced under the former group of estimators is lower than the sum of the diagonal adjusted flows produced under the latter group of estimators. The same holds also when comparing the constrained maximum likelihood estimator with the maximum likelihood estimator.

When incorporating the survey weights into the estimation process, the adjustments remain in the same direction. However, it is of interest to investigate the impact of weighting. Comparing the weighted with the unweighted observed flows, we note that in most of the cases the weighted off-diagonal elements increase compared to their unweighted equivalents. This seems reasonable. The UK LFS weights account for sampling attrition. Sampling attrition is related to more volatile sample units i.e. units associated with a higher probability of changing labour force status between t and $t+1$. Therefore, the survey weights appear to correctly modify the unweighted estimates.

3.6 Summary

In this chapter, we presented a model for adjusting gross flows estimates for misclassification. The model is formulated in a missing data framework and under alternative double sampling schemes. Maximum likelihood estimates are derived using maximum likelihood estimation via the EM algorithm. A constrained maximum likelihood estimator relaxes ICE and protects against the misspecification of the model assumptions. The model has been extended to account for the existence of a complex survey design. This is achieved by using the survey weights and the pseudo-maximum likelihood approach. Adjusted UK labour force gross flows are derived using alternative point estimators and re-interview data. Accounting for measurement error, results in estimating a less dynamic labour market. Use of the maximum likelihood estimator offers some practical advantages over the use of the conventional (moment-type) estimator. The current model assumes a non-differential measurement error mechanism and a non-differential gross flows mechanism. In Chapter 4, we extend the model to allow for heterogeneity in both mechanisms. In Chapter 5, we derive variance estimators for the moment-based and the likelihood-based adjusted gross flows. Finally, in Chapter 6 we contrast the likelihood-based approach with the moment-type approach using Monte-Carlo simulation experiments.

Chapter 4

Likelihood-based Inference for Gross Flows in the Presence of Misclassification, Double Sampling and Heterogeneity Associated with Discrete Covariates

4.1 Introduction

In Chapter 3, we presented a likelihood-based approach for adjusting gross flows for measurement error. However, the underlying model assumes the existence of a homogeneous measurement error and gross flows mechanism. It may be more realistic to assume that respondents with different socio-demographic characteristics have different probabilities of misclassification and different gross flows patterns. For example, younger respondents can be regarded as being more volatile and more prone to misclassification than older respondents. After giving some basic definitions, the model for adjusting gross flows for measurement error (see Chapter 3) will be extended to allow for heterogeneity. The use of discrete covariates implies that we account for heterogeneity by fitting the measurement error model within the post-strata defined by these covariates. The constrained maximum likelihood estimator and the pseudo-maximum likelihood (PML) estimator that allow for heterogeneity are also presented. Since most of the theory in this chapter is derived by using a straightforward extension of the theory in Chapter 3, the focus will be on applications. The effect of introducing heterogeneity is examined by contrasting estimators that allow for heterogeneity with estimators that ignore heterogeneity. The impact of measurement error on summary statistics of the labour market activity and on the probabilities of transition for different socio-demographic groups is also investigated. In the final sections, we discuss the limitations of the post-stratification parameterisation and we sketch an alternative, more flexible parameterisation.

4.2 Definitions and Previous Work

We start by giving the following two definitions.

Definition 4.1

Assume a $p \times z$ cross-classification defined by p categorical variables with z levels each. The misclassification mechanism is defined as non-differential if the following holds

$$pr\left(Y^*_{\xi t} = i, Y^*_{\xi t+1} = j \mid Y_{\xi t} = k, Y_{\xi t+1} = l\right) = q_{ijkl} \ \forall \text{ groups defined by } p \times z. \qquad (4.1)$$

The assumption of non-differential misclassification states that the all groups defined by the $p \times z$ cross-classification have the same proneness to error.

Definition 4.2

Assume a $p \times z$ cross-classification defined by p categorical variables with z levels each. The gross flows mechanism is defined as non-differential if the following holds

$$pr\left(Y^*_{\xi t} = i, Y^*_{\xi t+1} = j\right) = \Pi_{ij} \ \forall \text{ groups defined by } p \times z. \qquad (4.2)$$

The assumption of a non-differential gross flows mechanism states that the all groups defined by the $p \times z$ cross-classification have the same gross flows pattern.

In the presence of cross-sectional validation information, the existing literature accounts for heterogeneity by allowing ICE to hold within the groups defined by the $p \times z$ cross-classification. This is the so-called unit heterogeneity assumption since heterogeneity is assumed to exist between units that belong in different groups. Let us assume that there are Φ groups defined by the $p \times z$ cross-classification. Denote by $\overset{\wedge}{\Pi}$ the matrix of estimated observed transition probabilities, by $\hat{Q}_g$ the estimated misclassification matrix for group g and by α_g the fraction of sample units that belong in group g. The unit heterogeneity estimator is defined by

$$vec\left(\hat{P}^{unit}\right) = \left[\sum_{g=1}^{\Phi} \alpha_g \left(\hat{Q}_g(t) \otimes \hat{Q}_g(t)\right)\right]^{-1} vec\left(\overset{\wedge}{\Pi}\right). \qquad (4.3)$$

This is a moment-type estimator that allows for heterogeneity. However, this is only in the measurement error mechanism. Skinner and Torelli (1993) provide an expression for the bias introduced when ignoring heterogeneity. Denoting by λ the overall probability of correct classification and by $diag$ a matrix with elements only in the main diagonal and zeros elsewhere, the bias of $\hat{\lambda}$ when ignoring heterogeneity is given by

$$Bias\left(\hat{\lambda}\right) = tr\left\{\left[\hat{Q}(t) \otimes \hat{Q}(t) - \sum_{g=1}^{\Phi} a_g \hat{Q}_g(t) \otimes \hat{Q}_g(t)\right] diag\left(\hat{P}\right)\right\}. \qquad (4.4)$$

The authors suggest that ignoring heterogeneity will result in underestimating the overall probability of correct classification or, equivalently, in overcorrecting for measurement error. Unlike some of the methods that do not use validation information (see for example, Pfeffermann, Skinner and Humphreys 1998), the unit heterogeneity approach does not allow for heterogeneity in the gross flows mechanism. In the following sections, we present a more flexible way of incorporating heterogeneity than the one that is proposed in a moment-based framework.

4.3 Modelling Gross Flows in the Presence of Heterogeneity Induced by Discrete Covariates

Stating the Assumptions and Formulating the Model

Assume a double sampling scheme under which a validation sample of n^v units is selected independently from the main sample of n units and from the same population as the main sample has been also selected. This scheme implies that the main sample and the validation sample do not share common units. The main survey is a panel survey and provides information about the flows of people 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 errors related to these states at t. Let us further assume that both the validation sample and the main sample can be stratified in Φ mutually exclusive groups defined by p categorical variables with z levels each. The double sampling scheme described by Tables 3.1 and 3.2 (see Chapter 3) is now defined for each of the Φ groups. 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. Denote by n_{ijg}, n_{ijg}^v the number of sample units that belong in group g

and cell ij defined by the cross-classification of the observed with the true classifications in the main sample and in the validation sample respectively. A superscript $(*)$ is used to denote unobserved quantities. Extending the likelihood function of the augmented data (3.2) to allow for heterogeneity, we have that

$$L(\Theta)=\prod_{g=1}^{\Phi}\prod_{i=1}^{r^2}\prod_{j=1}^{r^2}\left(P_{ig}q_{ijg}\right)^{n_{ijg}^{v(*)}}\prod_{g=1}^{\Phi}\prod_{i=1}^{r^2}\prod_{j=1}^{r^2}\left(P_{ig}q_{ijg}\right)^{n_{ijg}^{(*)}}. \tag{4.5}$$

The model described by the likelihood function (4.5) assumes the existence of both a heterogeneous gross flows mechanism and of a heterogeneous misclassification mechanism. The model assumes that sample units in different groups have different misclassification and gross flows patterns. Taking the logarithms on both sides of (4.5) and imposing the additional constraint that

$$\sum_{i=1}^{r^2}P_{ig}=1 \quad \text{for fixed g}, \tag{4.6}$$

we obtain the following log-likelihood function

$$l(\Theta)=\sum_{g=1}^{\Phi}\sum_{i=1}^{r^2-1}\left(n_{i\bullet g}^{v(*)}+n_{i\bullet g}^{(*)}\right)\log P_{ig}+\left(n_{r^2\bullet g}^{v(*)}+n_{r^2\bullet g}^{(*)}\right)\log\left(1-\sum_{i=1}^{r^2-1}P_{ig}\right)+\sum_{g=1}^{\Phi}\sum_{i=1}^{r^2}\sum_{j=1}^{r^2}\left(n_{ijg}^{v(*)}+n_{ijg}^{(*)}\right)\log\left(q_{ijg}\right). \tag{4.7}$$

The group-specific misclassification probabilities, q_{ijg}, are unknown and are estimated using the group-specific cross-sectional misclassification probabilities and the ICE assumption. However, under this model ICE holds within groups. This assumption is equivalent to the unit heterogeneity assumption that is also used in moment-based framework. The likelihood function presented here does not incorporate the ICE assumption. 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 implies that the final number of group-specific parameters is $2r^2-r-1$.

Estimation

The log-likelihood function of the augmented data is maximised using the EM algorithm. The expectation step and the maximisation step are described below.

E-step

Denote by D^{vg} the group-specific observed data from the validation sample, by D^{mg} the group-specific observed data from the main sample, by (h) the current EM iteration and by $\Theta^{(h)}$ the vector of parameters in the (h) EM iteration. Taking the conditional expectation of the augmented log-likelihood (4.7) given the observed data and the parameters from the (h) EM iteration, the augmented log-likelihood is given by

$$\begin{aligned} E\left[l(\Theta;D^{c}) \mid D^{m},D^{v},\Theta^{(h)}\right] &= \sum_{g=1}^{\Phi}\sum_{i=1}^{r^2-1}\left[E\left(n_{i\bullet g}^{v(*)} \mid D^{vg},\Theta^{(h)}\right)+E\left(n_{i\bullet g}^{(*)} \mid D^{mg},\Theta^{(h)}\right)\right]\log P_{ig} \\ &+\left[E\left(n_{r^2\bullet g}^{v(*)} \mid D^{vg},\Theta^{(h)}\right)+E\left(n_{r^2\bullet g}^{(*)} \mid D^{mg},\Theta^{(h)}\right)\right]\log\left(1-\sum_{i=1}^{r^2-1}P_{ig}\right) \\ &+\sum_{g=1}^{\Phi}\sum_{i=1}^{r^2}\sum_{j=1}^{r^2}\left[E\left(n_{ijg}^{v(*)} \mid D^{vg},\Theta^{(h)}\right)+E\left(n_{ijg}^{(*)} \mid D^{mg},\Theta^{(h)}\right)\right]\log\left(q_{ijg}\right). \end{aligned} \tag{4.8}$$

The expectation step (E-step) can be performed using the following two results.

Result 4.1

Denote by $n_{\bullet jg}$ the total number of sample units in the main sample and group g that make transition j. The conditional expectations of the missing data in the main sample and group g are estimated using the following expression

$$\hat{E}\left(n_{ijg}^{(*)} \mid D^{mg},\Theta^{(h)}\right)=n_{\bullet jg}\left[\frac{\hat{q}_{ijg}^{(h)}\,\hat{P}_{ig}^{(h)}}{\sum_{i=1}^{r^2}\hat{q}_{ijg}^{(h)}\,\hat{P}_{ig}^{(h)}}\right] \quad \text{for fixed g.} \tag{4.9}$$

It follows that

$$\hat{E}\left(n_{i\bullet g}^{(*)} \mid D^{mg},\Theta^{(h)}\right)=\sum_{j=1}^{r^2}\hat{E}\left(n_{ijg}^{(*)} \mid D^{mg},\Theta^{(h)}\right) \quad \text{for fixed g.} \tag{4.10}$$

Result 4.2

Denote by n_{kg}^{v} the total number of sample units in the validation sample and group g that belong in the k^{th} combination of the true and the fallible classifications (see Table 3.3). The conditional expectations of the missing data in the validation sample and group g are estimated using the following expression

$$\hat{E}\left(n_{ijg}^{v(*)} \mid D^{vg}, \Theta^{(h)}\right) = n_{kg}^{v}\left[\frac{\hat{q}_{ijg}^{(h)}\,\hat{P}_{ig}^{(h)}}{\sum_{i}\sum_{j}\hat{q}_{ijg}^{(h)}\,\hat{P}_{ig}^{(h)}}\right] \text{ for fixed g.} \tag{4.11}$$

It follows that

$$\hat{E}\left(n_{i\bullet g}^{v(*)} \mid D^{vg}, \Theta^{(h)}\right) = \sum_{j=1}^{r^2}\hat{E}\left(n_{ijg}^{v(*)} \mid D^{vg}, \Theta^{(h)}\right) \text{ for fixed g.} \tag{4.12}$$

In Results 4.1 and 4.2, the parameters $\hat{q}_{ijg}^{(h)}$ need to be replaced by the products of group-specific cross-sectional misclassification probabilities. The proofs of Results 4.1 and 4.2 are identical to the proofs of Results 3.1 and 3.2.

M-step and Convergence of the EM Algorithm

The maximisation step is performed numerically using a Newton–type algorithm. However, unlike the model that assumes homogeneity, the maximisation step is now replaced by a series of maximisation steps i.e. one for each group g. For a fixed group g, the convergence criterion we use is the L^2-norm of the vector of parameters obtained from two successive iterations of the EM algorithm defined by

$$\left\|\Theta_{g}^{(h)} - \Theta_{g}^{(h+1)}\right\| = \sqrt{\sum_{i=1}^{2r^2-r-1}\left(\theta_{ig}^{(h)} - \theta_{ig}^{(h+1)}\right)^2}. \tag{4.13}$$

Using the previously described model, we obtain group-specific maximum likelihood estimates for the adjusted gross flows and for the misclassification probabilities. However, in many cases we are interested in obtaining overall adjusted gross flows. Combining the group-specific maximum likelihood estimates and assuming that a_g is fixed, we derive overall adjusted gross flows using the following expression

$$vec\left(\hat{P}\right) = \sum_{g=1}^{\Phi} a_g vec\left(\hat{P}_g\right). \tag{4.14}$$

The heterogeneity model can now be extended to account for the existence of a complex survey design. This can be achieved using the survey weights and the pseudo-maximum likelihood approach (see Section 3.4.2). Denote by $N_{ijg}^{(*)}, N_{ijg}^{v(*)}$ the number of population units that belong in group g and cell ij defined by the cross-classification of the observed with the true classifications. The expectation step (E-step) is performed using the following two results.

Result 4.3

Denote by $w_{\xi jg}$ the survey weights for sample unit ξ, which belongs in sub-population g and performs transition j, by N_g the size of sub-population g and by n_g the size of group g in the main sample. The conditional expectations of the missing data in the main sample and group g are estimated using the following expression

$$\hat{E}\left(N_{ijg}^{(*)} \mid D^{mg}, \Theta^{(h)pml}\right) = n_g \left(\frac{\sum_{\xi=1}^{n} w_{\xi jg}}{N_g}\right)\left[\frac{\hat{q}{}_{ijg}^{(h)} \hat{P}{}_{ig}^{(h)}}{\sum_{i=1}^{r^2} \hat{q}{}_{ijg}^{(h)} \hat{P}{}_{ig}^{(h)}}\right]. \tag{4.15}$$

It follows that

$$\hat{E}\left(N_{i\bullet g}^{(*)} \mid D^{mg}, \Theta^{(h)pml}\right) = \sum_{j=1}^{r^2} \hat{E}\left(N_{ijg}^{(*)} \mid D^{mg}, \Theta^{(h)pml}\right). \tag{4.16}$$

Result 4.4

Denote by $w_{\xi kg}$ the survey weights for sample unit ξ in the validation sample, which belongs in sub-population g and in the k^{th} combination of the true and the fallible classifications, by N_g the size of sub-population g and by n_g^v the size of group g in the validation sample. The conditional expectations of the missing data in the validation sample and group g are estimated using the following expression

$$\hat{E}\left(N_{ijg}^{v(*)} \mid D^{vg}, \Theta^{(h)pml}\right) = n_g^v \left(\frac{\sum_{\xi=1}^{n^v} w_{\xi kg}}{N_g}\right)\left[\frac{\hat{q}{}_{ijg}^{(h)} \hat{P}{}_{ig}^{(h)}}{\sum_{i}\sum_{j} \hat{q}{}_{ijg}^{(h)} \hat{P}{}_{ig}^{(h)}}\right]. \tag{4.17}$$

It follows that

$$\hat{E}\left(N^{v(*)}_{i\bullet g} \mid D^{vg}, \Theta^{(h)pml}\right) = \sum_{j=1}^{r^2} \hat{E}\left(N^{v(*)}_{ijg} \mid D^{vg}, \Theta^{(h)pml}\right). \tag{4.18}$$

The proofs of Results 4.3 and 4.4 are identical to the proofs of Results 3.5 and 3.6.

Extending the Constrained Maximum Likelihood Estimator to Account for Heterogeneity Induced by Discrete Covariates

The constrained maximum likelihood estimator (see Section 3.3) can be modified to account for the existence of heterogeneity. The stratification assumed here defines Φ gross flows matrices. The rows and columns of these matrices represent marginal constraints that need to be respected by the adjusted gross flows. Following the same approach as in Section 3.3, these constraints are imposed implicitly via the estimated likelihood approach. More specifically, after estimating group-specific misclassification probabilities $\hat{q}_{ijq}$ under the first and the second set of constraints, in the third step we impose the unbiased margins constraint by fixing q_{ijq} at their maximum likelihood values. The log-likelihood defined by (4.7) will now depend only on P_{ig} and is maximised using the EM algorithm and the following two results:

Result 4.5

The conditional expectations of the missing data in the main sample and group g are estimated using the following expression

$$\hat{E}\left(n^{(*)}_{ijg} \mid D^{mg}, \Theta^{(h)}\right) = n_{\bullet jg}\left[\frac{\hat{q}^{(c)}_{ijg}(t+1)\ \hat{q}^{(c)}_{ijg}(t)\hat{P}^{(h)}_{ig}}{\sum_{i=1}^{r^2}\hat{q}^{(c)}_{ijg}(t+1)\ \hat{q}^{(c)}_{ijg}(t)\hat{P}^{(h)}_{ig}}\right]. \tag{4.19}$$

Result 4.6

The conditional expectations of the missing data in the validation sample and group g are estimated using the following expression

$$\hat{E}\left(n^{v(*)}_{ijg} \mid D^{vg}, \Theta^{(h)}\right) = n^{v}_{kg}\left[\frac{\hat{q}^{(c)}_{ijg}(t+1)\ \hat{q}^{(c)}_{ijg}(t)\hat{P}^{(h)}_{ig}}{\sum_{i}\sum_{j}\hat{q}^{(c)}_{ijg}(t+1)\ \hat{q}^{(c)}_{ijg}(t)\hat{P}^{(h)}_{ig}}\right]. \tag{4.20}$$

The data constraints $n_{\bullet jg}$ are derived from the main sample and group g. The data constraints n_{kg}^{v} are derived from the original (i.e. not raked) cross-sectional validation sample and group g.

4.4 Adjusted UK Labour Force Gross Flows that Allow for Heterogeneity

The measurement error model that allows for heterogeneity is used to derive labour force gross flows adjusted for measurement error. The observed labour force gross flows are estimated by utilising the common UK LFS sample between summer-autumn 1997. Due to the absence of a UK validation survey, we use a smoothed version of reconciled validation data from the Swedish (October 1994 – April 1995) LFS re-interview programme (see also the applications in Chapter 3). The joint distribution $pr\left(Y_t^* = i, Y_t = k\right)$ for group g, estimated from the smoothed Swedish validation sample, is raked to corresponding UK group-specific marginal labour force estimates. This is achieved using the procedures described in Section 3.4.3. We present three applications. In the first application, we employ multinomial logistic models in an attempt to investigate the existence of heterogeneity in the gross flows and/or in the measurement error mechanism. The other two applications illustrate the methodology for adjusting labour force gross flows for measurement error. Both weighted and unweighted analysis is considered.

Application 4.1: Detecting the Existence of Heterogeneity in the Misclassification and/or in the Gross Flows Mechanism

The simplest way to detect heterogeneity is to fit multinomial logistic models. If respondents that belong to different groups are associated with different transition and/or misclassification probabilities, we will assume that heterogeneity exists. We model the probability that a unit of the main sample ξ makes a transition from state i at t to state j at $t+1$, compared to a baseline probability of transition, as a function of p categorical variables. This model is described by

$$\log\left[\frac{pr\left(Y_{\xi t}^* = i, Y_{\xi t+1}^* = j\right)}{pr\left(Y_{\xi t}^* = i', Y_{\xi t+1}^* = j'\right)}\right] = \alpha' X, \quad (ij) \neq \left(i'j'\right). \qquad (4.21)$$

We further model the probability that a unit of the validation sample ξ is classified in state i at t when he/she truly belongs in state k, compared to a baseline probability of misclassification, as a function of p categorical variables. This model is given by

$$\log\left[\frac{pr\left(Y_t^* = i, Y_t = k\right)}{pr\left(Y_t^* = i', Y_t = k'\right)}\right] = \beta' X, \quad (ik) \neq (i'k'). \qquad (4.22)$$

In (4.21) and (4.22), X denotes the design matrix and α, β are the vectors of parameters to be estimated. Note that there is no requirement for the two models to include the same set of covariates. We fit the models (4.21) and (4.22) using the common UK LFS sample between summer-autumn 1997 and the smoothed version of the Swedish validation dataset respectively. Both modelling exercises do not account for the complex survey design under which the data have been collected. For the purposes of this application, we include in the models two categorical variables with two levels each i.e. gender (1="Males, 0="Females") and age (1="16-22", 0="23-64"). The reason for using this particular partitioning of the age variable is because we are trying to maximise the dissimilarity between the age groups. However, one can use a more detailed partitioning.

Table 4.1: Anova table from modelling the unadjusted probabilities of transition

Model (Terms added sequentially)	*Deviance*	*Change in Degrees of Freedom*	*Likelihood Ratio Statistic**
Null	117611.8	8	
Gender	116017.6	8	1594.20
Age	113872.4	8	2145.20

* $X^2_{8\ df,0.05} = 15.50, X^2_{8\ df,0.01} = 20.09$

Table 4.2: Parameter estimates from modelling the unadjusted probabilities of transition-Multinomial logistic model with age and gender as covariates

Transition	*Intercept*	*Gender*	*Age*
EU	-4.69	0.19	1.31
EN	-3.92	-0.50	1.78
UE	-4.16	0.01	1.73
UU	-3.72	0.51	0.96
UN	-4.70	-0.06	1.75
NE	-3.85	-0.56	1.74
NU	-4.33	-0.40	1.47
NN	-1.05	-0.80	0.25

The parameter estimates for gender and age, reported in Table 4.2 and in Table 4.4, refer to young males. In comparison to the empty model, the anova table (Table 4.1) indicates that the inclusion of age and gender significantly improves the fit of the model that describes the probability of transition between the different labour force states. This implies the existence of heterogeneity in labour force gross flows. For example, the odds of an EU transition for young males are 3.7 times the odds of an EU transition for older males.

Table 4.3: Anova table from modelling the cross-sectional incidence of error

Model (Terms added sequentially)	*Deviance*	*Change in Degrees of Freedom*	*Likelihood Ratio Statistic**
Null	16841.20	8	
Gender	16617.46	8	223.74
Age	16287.86	8	329.60

* $X^2_{8\ df,0.05} = 15.50, X^2_{8\ df,0.01} = 20.09$

Table 4.4: Parameter estimates from modelling the cross-sectional incidence of error - Multinomial logistic model with age and gender as covariates

*Misclassification Pattern **	*Intercept*	*Gender*	*Age*
EU	-5.55	0.16	1.73
EN	-5.36	-0.64	2.49
UE	-6.94	-0.39	1.48
UU	-3.29	0.31	1.36
UN	-6.29	-0.75	1.75
NE	-4.78	-0.005	0.53
NU	-4.80	-0.17	-0.44
NN	-1.01	-0.71	0.31

*The first letter refers to the true classification while the second letter refers to the classification that contains measurement error i.e. diagonal elements (EE,UU,NN) indicate correct classification and off-diagonal elements indicate misclassification.

In comparison to the empty model, the anova table (Table 4.3) indicates that the inclusion of age and gender significantly improves the fit of the model that describes the cross-sectional misclassification process. This implies the existence of heterogeneity in the measurement error mechanism. For example, the odds of young respondents to be classified as inactive (N) when they are truly employed (E) are 12 times the odds of older respondents.

<u>Application 4.2:</u> Adjusting UK Labour Force Gross Flows for Measurement Error in the Presence of Heterogeneity-Unweighted Analysis

Having investigated the existence of heterogeneity in the measurement error and in the gross flows mechanism, we now utilise the methodology derived in this chapter to adjust labour force gross flows for measurement error. The estimators considered here are the following: The conventional non-heterogeneous estimator (Section 2.2.2.1), the unit heterogeneity estimator (Section 4.2), the non-heterogeneous maximum likelihood estimator (Section 3.2) and the maximum likelihood estimator that allows for heterogeneity (Section 4.3). We allow for heterogeneity according to age, (1=”16-22”, 0=”23-64”) and gender (1=”Males, 0=”Females”). The convergence criterion for the EM algorithm is $\delta = 10^{-4}$. Table 4.5 presents adjusted for measurement error labour force gross flows, using the alternative estimators, and contrasts them with the observed (unadjusted) labour force gross flows.

Table 4.5: Adjusted UK labour force gross flows for summer-autumn 1997 using a range of different estimators – Unweighted analysis

Flow	*Observed Flows*	*Conventional Non-Heterogeneous (ICE)*	*Unit Heterogeneity (Age,Gender)*	*MLE Non-Heterogeneous (ICE)*	*MLE Heterogeneity (Age,Gender)*
EE	0.716	0.7420	0.7415	0.7410	0.7323
EU	0.009	0.0028	0.0030	0.0033	0.0043
EN	0.016	0.0024	0.0027	0.0026	0.0065
UE	0.016	0.0102	0.0104	0.0106	0.0120
UU	0.027	0.0292	0.0291	0.0291	0.0292
UN	0.009	0.0033	0.0033	0.0036	0.0044
NE	0.016	0.0026	0.0029	0.0028	0.0067
NU	0.010	0.0045	0.0045	0.0050	0.0057
NN	0.181	0.2030	0.2026	0.2020	0.1989

Table 4.6: Anova table from fitting the measurement error model

Model	*Log (L)*	*Likelihood Ratio Statistic**
Null	-89199.30	
Gender	-87686.68	3025.24 (14)*
Gender * Age	-84767.11	5839.14 (28)*

*In brackets we report the change in the degrees of freedom as we move from the simplest model towards the more complicated model. Also, $X^2_{14\,df,0.01} = 29.14$, $X^2_{28\,df,0.01} = 48.28$

Table 4.6 indicates that the inclusion of age and gender significantly improves the fit of the measurement error model.

Application 4.3: Adjusting UK Labour Force Gross Flows for Measurement Error in the Presence of Heterogeneity-Weighted Analysis

We now derive adjusted UK labour force gross flows between summer-autumn 1997 using the alternative point estimators and the UK LFS weights. The survey weights implicitly adjust for sampling attrition. The estimators considered in the current application are the following: The weighted non-heterogeneous conventional estimator (Section 3.4.5), the weighted estimator that allows for heterogeneity (i.e. the weighted version of estimator presented in Section 4.2), the non-heterogeneous pseudo-maximum likelihood estimator and the pseudo-maximum likelihood estimator that allows for heterogeneity (Sections 3.4.2 and

4.3 respectively). We allow for heterogeneity according to age and gender. The convergence criterion for the EM algorithm is $\delta = 10^{-4}$. Table 4.7 presents adjusted for measurement error labour force gross flows, using the alternative estimators, and contrasts them with the weighted observed (unadjusted) labour force gross flows.

Table 4.7: Adjusted UK labour force gross flows for summer-autumn 1997 using a range of different estimators – Weighted analysis

Flow	*Weighted Observed Flows*	*Weighted Conventional Non-Heterogeneous (ICE)*	*Weighted Unit-Heterogeneity (Age,Gender)*	*PML Non-Heterogeneous (ICE)*	*PML Heterogeneity (Age,Gender)*
EE	0.705	0.7312	0.7307	0.7304	0.7245
EU	0.010	0.0038	0.0039	0.0041	0.0051
EN	0.019	0.0049	0.0052	0.0052	0.0081
UE	0.017	0.0113	0.0114	0.0114	0.0122
UU	0.032	0.0349	0.0347	0.0347	0.0343
UN	0.010	0.0042	0.0042	0.0044	0.0052
NE	0.015	0.0005	0.0008	0.0015	0.0038
NU	0.010	0.0046	0.0046	0.0047	0.0057
NN	0.182	0.2046	0.2045	0.2036	0.2011

4.4.1 Discussion of the Adjustments Derived from the Alternative Estimators

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. Adjustments derived when accounting for heterogeneity are less severe than adjustments derived when heterogeneity is ignored. To illustrate this, one can compute the sum of the off-diagonal adjusted gross flows derived from the alternative point estimators. For example, in Table 4.5 the sum of the off-diagonal gross flows derived from the conventional non-heterogeneous estimator is 0.0258. The same sum computed when using the conventional unit heterogeneity estimator is 0.0268. Similarly, for the non-heterogeneous maximum likelihood this sum is 0.0279 and for the maximum likelihood estimator that accounts for heterogeneity the sum is 0.0396. These results are consistent with the assumption that the effect of ignoring heterogeneity, when heterogeneity exists, results in over-adjusting gross flows for measurement error. Unlike the unit heterogeneity estimator that allows for heterogeneity only in the measurement error mechanism, the maximum likelihood estimator allows for heterogeneity both in the measurement error and in the gross flows mechanism. This explains the larger impact of the adjustments derived when using this estimator.

4.4.2 The Effect of Misclassification on Inference Based on Labour Force Gross Flows

Labour force gross flows are widely used by social scientists and economists for research and policy purposes. In this section, we analyse the effects of making inference based on the unadjusted as opposed to the adjusted labour force gross flows. As a summary statistic, we use the estimated probabilities of transition from state i to state j between t and $t+1$. More specifically, we compute two sets of probabilities. Firstly, we determine the unadjusted probabilities of transition by modelling the unweighted observed labour force gross flows. The model we employ is a multinomial logistic that includes the main effects according to age and gender and the interaction term between these two covariates. In addition, we use the adjusted probabilities of transition derived from the maximum likelihood estimator that accounts for heterogeneity (see Table 4.5). The reason for including the interaction term in the multinomial logistic model is because fitting the measurement error model within the post-strata is equivalent to a multinomial logistic model that includes all possible interaction terms. To quantify the effect of measurement error, we compute ratios of estimated transition probabilities, before and after adjustment is applied, for different age and gender groups. Moreover, we investigate the effect of measurement error on two widely used summary statistics of the labour market activity. These are the probability of a successful exit from unemployment and the probability of a successful exit from inactivity for different age and gender groups. These probabilities are defined respectively by

$$pr\left(Y_t = U, Y_{t+1} = E\right) = \frac{pr\left(Y_t = U, Y_{t+1} = E\right)}{pr\left(Y_t = U, Y_{t+1} = E\right) + pr\left(Y_t = U, Y_{t+1} = N\right)}, \qquad (4.23)$$

$$pr\left(Y_t = N, Y_{t+1} = E\right) = \frac{pr\left(Y_t = N, Y_{t+1} = E\right)}{pr\left(Y_t = N, Y_{t+1} = E\right) + pr\left(Y_t = N, Y_{t+1} = U\right)}. \qquad (4.24)$$

Table 4.8: Ratios of probabilities of transition for young males vs. old males

Flow	***Ratios based on the unadjusted labour force gross flows (multinomial logistic model with age and gender)***	***Ratios based on the adjusted labour force gross flows obtained from the MLE with heterogeneity according to age and gender***
EU	2.55	3.29
EN	5.11	12.4
UE	4.66	6.94
UN	4.00	16.5
NE	5.00	12.2
NU	3.29	13.5

Table 4.9: Ratios of probabilities of transition for young females vs. old females

Flow	***Ratios based on the unadjusted labour force gross flows (multinomial logistic model with age and gender)***	***Ratios based on the adjusted labour force gross flows obtained from the MLE with heterogeneity according to age and gender***
EU	3.16	0.63
EN	4.00	3.45
UE	3.72	3.12
UN	4.66	16.7
NE	3.80	2.81
NU	3.22	5.47

Table 4.10: Ratios of probabilities of transition for young males vs. young females

Flow	***Ratios based on the unadjusted labour force gross flows (multinomial logistic model with age and gender)***	***Ratios based on the adjusted labour force gross flows obtained from the MLE with heterogeneity according to age and gender***
EU	1.21	4.33
EN	0.82	1.16
UE	1.36	1.56
UN	1.00	1.16
NE	0.79	1.18
NU	0.80	0.86

Table 4.11: Ratios of probabilities of transition for old males vs. old females

Flow	***Ratios based on the unadjusted labour force gross flows (multinomial logistic model with age and gender)***	***Ratios based on the adjusted labour force gross flows obtained from the MLE with heterogeneity according to age and gender***
EU	1.50	0.83
EN	0.64	0.32
UE	1.09	0.70
UN	1.16	1.17
NE	0.60	0.27
NU	0.77	0.35

Table 4.12: The effect of measurement error on the probability of a successful exit from unemployment

Group	***Probability of successful exit from unemployment based on the unadjusted labour force gross flows (multinomial logistic model with age and gender)***	***Probability of successful exit from unemployment based on the adjusted labour force gross flows obtained from the MLE with heterogeneity according to age and gender***
Young males	0.66	0.64
Young Females	0.59	0.57
Old males	0.63	0.81
Old females	0.65	0.88

Table 4.13: The effect of measurement error on the probability of a successful exit from inactivity

Group	***Probability of successful exit from inactivity based on the unadjusted labour force gross flows (multinomial logistic model with age and gender)***	***Probability of successful exit from inactivity based on the adjusted labour force gross flows obtained from the MLE with heterogeneity according to age and gender***
Young males	0.66	0.51
Young Females	0.66	0.44
Old males	0.56	0.55
Old females	0.62	0.61

The previous results indicate that measurement error can have a significant impact on inference based on labour force gross flows. Two key characteristics emerge from this analysis. Firstly, the probability of transition of one group compared to another can be underestimated or overestimated but remain in the same direction when the unadjusted flows are used. For example, based on the unadjusted flows, the probability of transition from unemployment to inactivity for young males is 4 times higher than the same probability for old males. Based on the adjusted flows, young males have 16.5 times the probability of old males for making a transition from unemployment to inactivity. An example where the probability of transition is overestimated when using the unadjusted gross flows is in the transition from inactivity to employment for young females compared to old females. A second, more important consequence of measurement error is that in some cases there is a complete reversal in the direction of inference. The most obvious case is in the probability of transition from employment to unemployment for young females compared to old females. Based on the unadjusted gross flows, young females have 3.16 times the probability of older females for making a transition from employment to unemployment. Based on the adjusted gross flows, young females have 0.63 times the probability of older females for making the same transition. Other examples, where reversals in the direction of inference occur, are reported in Table 4.11. Observing further the results reported in Table 4.10, one can say that in most of the cases, based on the adjusted gross flows, young males are more volatile than young females. Based also on the adjusted gross flows and Table 4.11, in most of the cases old females are more volatile than old males. Based on the unadjusted flows, the volatility of young males is underestimated relatively to the volatility of young females and the volatility of old females is underestimated relatively to the volatility of old males.

Measurement error appears also to have a distorting effect on the summary statistics of the labour market activity. Based on the unadjusted labour force gross flows, the probability of a successful exit from unemployment is seriously underestimated for the less volatile groups (old males and old females). For the more volatile groups (young males and young females), the same probability is slightly overestimated. Poterba and Summers (1986) report similar findings in the context of the CPS. Based also on the unadjusted labour force gross flows, the probability of a successful exit from inactivity is overestimated for all different groups with the most volatile groups being affected more.

4.5 The Limitations of the Current Parameterisation of the Measurement Error Model and Extensions

In this last section, we discuss some of the limitations of the current parameterisation of the measurement error model that allows for heterogeneity and we sketch an alternative parameterisation.

Under the current parameterisation, the model parameters are estimated by fitting the model within the post-strata defined by the discrete covariates. This is equivalent to a logistic parameterisation that includes interaction terms. Let us assume that we employ two discrete covariates with two levels each. Consequently, there are four different post-strata formed and 56 parameters estimated by fitting the measurement error model within each of the post-strata. An equivalent multinomial logistic parameterisation is as follows: (a) A multinomial logistic model for modelling the probability of transition that includes the two main effects and the two-way interaction (32 parameters), (b) a multinomial logistic model for each column of the misclassification matrix that includes the two main effects and the two-way interaction (24 parameters). However, one may wish to include, for example, only main effects in the model or include a different set of covariates for modelling the transition and the misclassification probabilities. An alternative, more natural solution is offered by re-expressing the parameters of the measurement error model using a logistic formulation. For example, $P_i = \dfrac{\exp(\beta' X)}{1 + \exp(\beta' X)}$ where X denotes the design matrix and β is the vector of parameters to be estimated.

A second limitation of the current parameterisation is that it allows for heterogeneity only via discrete covariates. Unlike this model, latent Markov models can allow for heterogeneity according to discrete and continuous covariates (Humphreys 1996). In addition, the small sample size of the validation survey implies sparseness of data when attempting to estimate more complicated models (i.e. with many categorical variables).

Given appropriate validation data, one possibility for extending the measurement error model is to include as a covariate the type of response i.e. "self" or "proxy" response. Relevant theory (O'Muircheartaigh 1991) suggests that the type of response is highly related to the measurement error problem. Unfortunately, we were not able to include the response status variable into our analysis because of insufficient data. This is due to the low percentage of proxy response in the Swedish LFS (around 3%) and the small sample size of the Swedish validation survey. Finally, one can view the rotation group bias as a misclassification problem. If rotation group bias exists, the misclassification mechanism can be expected to be differential with respect to rotation group. By including rotation group as a covariate in the heterogeneity model, one can adjust also for this source of bias.

4.6 Summary

The analysis presented in this chapter indicated that heterogeneity is likely to exist both in the gross flows and in the measurement error mechanism. The model presented allows for heterogeneity in both mechanisms and can be considered as more realistic than the model that ignores heterogeneity or the moment-based approach that assumes heterogeneity only in the measurement error mechanism. The effect of ignoring heterogeneity, when heterogeneity exists, can result in an overcorrection for measurement error. For example, we show that ignoring heterogeneity, when correcting UK labour force gross flows for measurement error, results in estimating a less volatile labour market than the real one. A further result concerns the effect of measurement error on inference based on gross flows. Our analysis indicates that ignoring the measurement error problem can have a severe effect, which in some cases can result in a complete reversal of the direction of inference.

Chapter 5

Variance Estimation for Gross Flows Estimates in the Presence of Misclassification and Double Sampling

5.1 Introduction

Having discussed alternative approaches for point estimation, we now turn to development of variance estimators for the adjusted gross flows estimates. Generally speaking, variance estimation in a double sampling framework must account for the extra variability introduced by the smaller size of the second phase sample. Literature on variance estimation for cross-sectional estimates in the presence of misclassification and double sampling includes Tenenbein (1972), Selen (1986) and Greenland (1988).

The structure of this chapter is as follows. In Section 5.2, we develop a variance estimator for the conventional (moment-type) estimator (see Section 2.2.2.1). In Section 5.3, we develop variance estimators for alternative moment-type estimators that were presented in Section 2.4. In Section 5.4 variance estimation for the maximum likelihood estimator, when the validation sample is selected independently from the main sample and from the same target population (see Section 3.2.1), is considered. We further present a procedure for estimating the variance of the adjusted cross-sectional estimates when using maximum likelihood estimation via the EM algorithm (see Section 2.2.1.3). Variance estimation for the quasi-likelihood adjusted estimates (see Section 2.2.1.4) is discussed in Section 5.5. In the final section, the theory is illustrated via three applications.

5.2 Variance Estimation for the Conventional (Moment-Type) Estimator of the Adjusted Gross Flows

The conventional estimator, under ICE, is given by

$$vec\left(\underset{r\times r}{\hat{P}^{st}}\right) = vec\left[\underset{r\times r}{\hat{Q}^{-1}}\ \underset{r\times r}{\overset{\wedge}{\Pi}}\left(\underset{r\times r}{\hat{Q}^{-1}}\right)^{\mathrm{T}}\right]. \tag{5.1}$$

In order to simplify the notation, we drop the parenthesis next to Q that is specific of the time periods to which the misclassification matrix refers. A variance estimator for (5.1) can be derived by employing the δ-method (Bishop, Fienberg and Holland 1975, Agresti 1990). This involves expanding $vec\left(\hat{P}^{st}\right)$ in a Taylor series around its true value $vec(P)$. Let $vec\left[P^{st}\left(\overset{\wedge}{\Theta}\right)\right] = \left[g_1\left(\overset{\wedge}{\Theta}\right), g_2\left(\overset{\wedge}{\Theta}\right), \ldots, g_{r^2}\left(\overset{\wedge}{\Theta}\right)\right]^T$ represent a $r^2 \times 1$ vector of non-linear functions of $\overset{\wedge}{\Theta} = \left(\hat{q}_{11}, \hat{q}_{21}, \hat{q}_{31}, \ldots, \hat{q}_{rr}, \overset{\wedge}{\Pi}_{11}, \overset{\wedge}{\Pi}_{21}, \overset{\wedge}{\Pi}_{31}, \ldots, \overset{\wedge}{\Pi}_{rr}\right)$. Recall that q_{ik} denotes the misclassification probabilities and Π_{lj} denotes the observed transition probabilities between t and $t+1$. Note also that we now distinguish between the subscripts l, i for reasons that will become apparent later in this chapter. However, both subscripts refer to the observed classification at t. Expanding $vec\left(\hat{P}^{st}\right)$ around its true value using Taylor series, we have that

$$vec\left[P^{st}\left(\overset{\wedge}{\Theta}\right)\right] \approx vec[P(\Theta)] + \nabla_{\Theta}\left(\overset{\wedge}{\Theta} - \Theta\right), \nabla_{\Theta} = \frac{\partial vec\left[\underset{r\times r}{Q^{-1}}\ \underset{r\times r}{\Pi}\ \underset{r\times r}{\left(Q^{-1}\right)^{\mathrm{T}}}\right]}{\partial\Theta}\Big|_{\Theta=\overset{\wedge}{\Theta}}. \tag{5.2}$$

It follows that

$$vec\left[P^{st}\left(\overset{\wedge}{\Theta}\right)\right] - vec[P(\Theta)] \approx \nabla_{\Theta}\left(\overset{\wedge}{\Theta} - \Theta\right), \nabla_{\Theta} = \frac{\partial vec\left[\underset{r\times r}{Q^{-1}}\ \underset{r\times r}{\Pi}\ \underset{r\times r}{\left(Q^{-1}\right)^{\mathrm{T}}}\right]}{\partial\Theta}\Big|_{\Theta=\overset{\wedge}{\Theta}}. \tag{5.3}$$

Taking the variance operator on both sides of (5.3), we have that

$$Var\left\{vec\left[P^{st}\left(\overset{\wedge}{\Theta}\right)\right]\right\} \approx \nabla_{\Theta} Var\left(\overset{\wedge}{\Theta}\right)\left(\nabla_{\Theta}\right)^{T}. \tag{5.4}$$

In order to estimate (5.4), we need to evaluate the Jacobian matrices ∇_{Θ}, $(\nabla_{\Theta})^T$ and estimate the covariance matrix $Var\left(\hat{\Theta}\right)$. In the later case, we need to estimate the following components: (a) the variance-covariance structure of the unadjusted estimated probabilities of transition $\hat{\Pi}_{lj}$, (b) the variance-covariance structure of the estimated misclassification probabilities $\hat{q}_{ik}$ and (c) the covariance structure of $\hat{\Pi}_{lj}, \hat{q}_{ik}$ $i,j,k,l = 1,2,\ldots r$. Without loss of generality, we focus our interest on the case that $r = 3$, $i,j,k,l = 1,2,3$. This is due to our interest in estimating labour force gross flows that are frequently described by a 3×3 gross flows matrix. For simplicity, we denote $Var\left(\hat{\Theta}\right)$ by Σ. When $r = 3$ Σ is described by a 18×18 matrix whose general form is given below

$$\Sigma = \left[\begin{array}{ccc|ccc}
Var\left(\hat{q}_{11}\right) & \cdots & Cov\left(\hat{q}_{11}, \hat{q}_{33}\right) & Cov\left(\hat{q}_{11}, \hat{\Pi}_{11}\right) & \cdots & Cov\left(\hat{q}_{11}, \hat{\Pi}_{33}\right) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
Cov\left(\hat{q}_{33}, \hat{q}_{11}\right) & \cdots & Var\left(\hat{q}_{33}\right) & Cov\left(\hat{q}_{33}, \hat{\Pi}_{11}\right) & \cdots & Cov\left(\hat{q}_{33}, \hat{\Pi}_{33}\right) \\
\hline
Cov\left(\hat{\Pi}_{11}, \hat{q}_{11}\right) & \cdots & Cov\left(\hat{\Pi}_{11}, \hat{q}_{33}\right) & Var\left(\hat{\Pi}_{11}\right) & \cdots & Cov\left(\hat{\Pi}_{11}, \hat{\Pi}_{33}\right) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
Cov\left(\hat{\Pi}_{33}, \hat{q}_{11}\right) & \cdots & Cov\left(\hat{\Pi}_{33}, \hat{q}_{33}\right) & Cov\left(\hat{\Pi}_{33}, \hat{\Pi}_{11}\right) & \cdots & Var\left(\hat{\Pi}_{33}\right)
\end{array}\right].$$

The upper left block of Σ represents the covariance matrix of the estimates of the elements of the misclassification matrix, the upper right block and the lower left block represent the covariance structure between the estimated unadjusted probabilities of transition and the estimated misclassification probabilities and the lower right block represents the covariance matrix of the estimated unadjusted probabilities of transition.

Variance-Covariance Structure of the Unadjusted Estimated Probabilities of Transition

Result 5.1

Under simple random sampling and taking also into account that the sample size of the main sample is fixed, we can regard the cells of the observed gross flows matrix as multinomial proportions. Ignoring the finite population correction, the covariance matrix of the elements of the unadjusted gross flows matrix is estimated using standard results for the variance of binomial random variables as follows:

$$\begin{cases} \hat{Var}\left(\hat{\Pi}_{lj}\right) = \dfrac{\hat{\Pi}_{lj}\left(1-\hat{\Pi}_{lj}\right)}{n} \\ \hat{Cov}\left(\hat{\Pi}_{lj}, \hat{\Pi}_{l^*j^*}\right) = \dfrac{-\hat{\Pi}_{lj}\hat{\Pi}_{l^*j^*}}{n} \qquad (lj) \neq \left(l^*j^*\right). \end{cases} \tag{5.5}$$

Substituting (5.5) into lower right block of Σ, we obtain an estimate for the covariance matrix of the estimated unadjusted probabilities of transition.

Variance-Covariance Structure of the Estimated Misclassification Probabilities

Denote by n^v the size of the validation sample and by n_{ik}^v the number of sample units that are observed in state i at t when they truly belong in state k. The estimated misclassification probabilities are defined by $\hat{q}_{ik} = \dfrac{n_{ik}^v}{\sum_{i=1}^{r} n_{ik}^v}$. While $n^v = \sum_{i=1}^{r}\sum_{k=1}^{r} n_{ik}^v$ can be considered as fixed, $\sum_{i=1}^{r} n_{ik}^v$ must be considered as a random. Thus, $\hat{q}_{ik}$ is defined as a ratio of random quantities. Consequently, in the computation of the variance-covariance structure of the misclassification probabilities, we must take into account an extra level of non-linearity introduced by the fact that both the numerator and denominator of $\hat{q}_{ik}$ are random quantities. Therefore, we need to make a second application of the δ-method.

Denote by $\hat{Q}$ the misclassification matrix estimated from the validation sample. Let $\hat{\Theta}^* = \left(n_{11}^v, n_{21}^v, n_{31}^v, \ldots, n_{rr}^v\right)$ and $vec\left[Q\left(\hat{\Theta}^*\right)\right] = \left[f_1\left(\hat{\Theta}^*\right), \ldots, f_{r^2}\left(\hat{\Theta}^*\right)\right]^T$ be the $r^2 \times 1$ vector of non-linear functions of $\hat{\Theta}^*$. Applying the delta method to $vec\left[Q\left(\hat{\Theta}^*\right)\right]$, we derive the following

$$vec\left[Q\left(\hat{\Theta}^*\right)\right] \approx vec\left[Q\left(\Theta^*\right)\right] + \nabla_{\Theta^*}\left(\hat{\Theta}^* - \Theta^*\right), \quad \nabla_{\Theta^*} = \frac{\partial vec\left[Q\left(\Theta^*\right)\right]}{\partial \Theta^*}\Big|_{\Theta^* = \hat{\Theta}^*} . \tag{5.6}$$

It follows that

$$vec\left[Q\left(\overset{\wedge}{\Theta}^{*}\right)\right]-vec[Q(\Theta^{*})]\approx\nabla_{\Theta^{*}}\left(\overset{\wedge}{\Theta}^{*}-\Theta^{*}\right),\quad \nabla_{\Theta^{*}}=\frac{\partial vec\left[Q\left(\Theta^{*}\right)\right]}{\partial\Theta^{*}}\Big|_{\Theta^{*}=\overset{\wedge}{\Theta}^{*}}\cdot \tag{5.7}$$

Taking the variance operator on both sides of (5.7), we obtain the following

$$Var\left\{vec\left[Q\left(\overset{\wedge}{\Theta}^{*}\right)\right]\right\}\approx\nabla_{\Theta^{*}}Var\left(\overset{\wedge}{\Theta}^{*}\right)\left(\nabla_{\Theta^{*}}\right)^{T}. \tag{5.8}$$

In order to estimate (5.8), we need to evaluate the Jacobian matrices $\nabla_{\Theta^{*}}$, $\left(\nabla_{\Theta^{*}}\right)^{T}$ and the covariance matrix $Var\left(\overset{\wedge}{\Theta}^{*}\right)$. Under simple random sampling and taking into account that the sample size of the validation survey is fixed, we can regard n_{ik}^{v} as multinomial counts. Ignoring the finite population correction, the required covariance matrix can be estimated as follows:

$$\begin{cases}\overset{\wedge}{Var}\left(n_{ik}^{v}\right)=n^{v}\hat{P}_{ik}\left(1-\hat{P}_{ik}\right)\\ \overset{\wedge}{Cov}\left(n_{ik}^{v},n_{i^{*}k^{*}}^{v}\right)=-n^{v}\hat{P}_{ik}\hat{P}_{i^{*}k^{*}}\qquad (ik)\neq\left(i^{*}k^{*}\right).\end{cases} \tag{5.9}$$

It remains to evaluate the Jacobian matrices $\nabla_{\Theta^{*}}$ involved in the second application of the delta method. These matrices are evaluated using the expressions below

$$\nabla_{\Theta^{*}}=\begin{bmatrix}\dfrac{\partial\left(\dfrac{n_{11}^{v}}{n_{.1}^{v}}\right)}{\partial n_{11}^{v}} & \dfrac{\partial\left(\dfrac{n_{11}^{v}}{n_{.1}^{v}}\right)}{\partial n_{21}^{v}} & \dfrac{\partial\left(\dfrac{n_{11}^{v}}{n_{.1}^{v}}\right)}{\partial n_{31}^{v}} & \cdots & \dfrac{\partial\left(\dfrac{n_{11}^{v}}{n_{.1}^{v}}\right)}{\partial n_{33}^{v}}\\ \dfrac{\partial\left(\dfrac{n_{21}^{v}}{n_{.1}^{v}}\right)}{\partial n_{11}^{v}} & \dfrac{\partial\left(\dfrac{n_{21}^{v}}{n_{.1}^{v}}\right)}{\partial n_{21}^{v}} & \dfrac{\partial\left(\dfrac{n_{21}^{v}}{n_{.1}^{v}}\right)}{\partial n_{31}^{v}} & \cdots & \dfrac{\partial\left(\dfrac{n_{21}^{v}}{n_{.1}^{v}}\right)}{\partial n_{33}^{v}}\\ \vdots & \vdots & \vdots & \cdots & \vdots\\ \dfrac{\partial\left(\dfrac{n_{33}^{v}}{n_{.3}^{v}}\right)}{\partial n_{11}^{v}} & \dfrac{\partial\left(\dfrac{n_{33}^{v}}{n_{.3}^{v}}\right)}{\partial n_{21}^{v}} & \dfrac{\partial\left(\dfrac{n_{33}^{v}}{n_{.3}^{v}}\right)}{\partial n_{31}^{v}} & \cdots & \dfrac{\partial\left(\dfrac{n_{33}^{v}}{n_{.3}^{v}}\right)}{\partial n_{33}^{v}}\end{bmatrix}$$

and

$$\nabla_{\Theta^*} = \begin{bmatrix} \frac{n_{21}^v + n_{31}^v}{\left(n_{11}^v + n_{21}^v + n_{31}^v\right)^2} & \frac{-n_{11}^v}{\left(n_{11}^v + n_{21}^v + n_{31}^v\right)^2} & \frac{-n_{11}^v}{\left(n_{11}^v + n_{21}^v + n_{31}^v\right)^2} & \cdots & 0 \\ \frac{-n_{21}^v}{\left(n_{11}^v + n_{21}^v + n_{31}^v\right)^2} & \frac{n_{11}^v + n_{31}^v}{\left(n_{11}^v + n_{21}^v + n_{31}^v\right)^2} & \frac{-n_{21}^v}{\left(n_{11}^v + n_{21}^v + n_{31}^v\right)^2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{n_{13}^v + n_{23}^v}{\left(n_{13}^v + n_{23}^v + n_{33}^v\right)^2} \end{bmatrix} . \qquad (5.10)$$

Substituting (5.9) and (5.10) into (5.8), we obtain an estimate for the covariance matrix of the estimated misclassification probabilities.

Covariance Structure between the Estimated Unadjusted Transition Probabilities and the Estimated Misclassification Probabilities

The remaining part, in estimating $Var\left(\hat{\Theta}\right)$, is to evaluate the covariance structure between the unadjusted estimated transition probabilities and the estimated misclassification probabilities i.e. $Cov\left(\hat{\Pi}_{lj}, \hat{q}_{ik}\right)$.

We distinguish two cases:

(a) a double sampling scheme under which the misclassification probabilities are estimated either via an internal validation sample that is selected independently from the main sample and from the same target population or via an external validation sample. For this case, it is reasonable to assume that

$$Cov\left(\hat{q}_{ik,}\ \hat{\Pi}_{lj}\right) = 0. \qquad (5.11)$$

(b) a double sampling scheme under which the misclassification probabilities are estimated via an internal validation sample that is selected by sub-sampling units that already participate in the main survey. For this case, we assume that

$$Cov\left(\hat{q}_{ik,}\ \hat{\Pi}_{lj}\right) = Cov\left(\frac{n_{ik}^v}{\sum_{i=1}^{r} n_{ik}^v}, \frac{n_{lj}}{n}\right) \neq 0. \qquad (5.12)$$

In order to derive variance estimates for the second case, we need to estimate the covariance terms $Cov\left(\hat{q}_{ik,}\ \overset{\wedge}{\Pi}_{lj}\right)$.

Lemma 5.1

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

$$\begin{aligned} E[g(X,Y)] \approx g(\mu_X,\mu_Y) + \frac{1}{2}\frac{\partial^2}{\partial y^2}g(X,Y)\,|_{\mu_X,\mu_Y}\ Var(Y) + \frac{1}{2}\frac{\partial^2}{\partial x^2}g(X,Y)\,|_{\mu_X,\mu_Y}\ Var(X) \\ + \frac{\partial^2}{\partial x \partial y}g(X,Y)\,|_{\mu_X,\mu_Y}\ Cov(X,Y). \end{aligned} \tag{5.13}$$

Proof

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

□

Result 5.2

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

$$Cov\left(\frac{X}{Y},\frac{A}{n}\right) \approx \frac{1}{nE(Y)}\left[Cov(A,X) - \frac{E(X)}{E(Y)}Cov(A,Y)\right]. \tag{5.14}$$

Proof

We start the proof by expanding the covariance term of interest using the standard definition for the covariance between random variables i.e.

$$Cov\left(\frac{X}{Y},\frac{A}{n}\right) = E\left(\frac{X}{Y}\frac{A}{n}\right) - E\left(\frac{X}{Y}\right)E\left(\frac{A}{n}\right) = \frac{1}{n}\left[E\left(\frac{AX}{Y}\right) - E\left(\frac{X}{Y}\right)E(A)\right]. \tag{5.15}$$

We evaluate the different components of the expression above using Lemma 5.1. More specifically, we approximate $E\left(\frac{AX}{Y}\right)$ utilising the Taylor series expansion of $\frac{AX}{Y}$ around (μ_X,μ_Y,μ_A). This Taylor series expansion is given below

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

It follows that

$$E[g(X,Y,A)] \approx \frac{\mu_X\mu_A}{\mu_Y} + \frac{1}{2}\frac{2\mu_X\mu_A}{\mu_Y^{\,3}} Var(Y) - \frac{\mu_A}{\mu_Y^{\,2}} Cov(X,Y) + \frac{1}{\mu_Y} Cov(X,A) - \frac{\mu_X}{\mu_Y^{\,2}} Cov(A,Y). \tag{5.16}$$

Next, we approximate $E\left(\frac{X}{Y}\right)$ using a Taylor series expansion of $\frac{X}{Y}$ around (μ_X,μ_Y) as follows

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

It follows that

$$E[g(X,Y)] \approx \frac{\mu_X}{\mu_Y} + \frac{1}{2}\frac{2\mu_X}{\mu_Y^{\,3}} Var(Y) - \frac{1}{\mu_Y^{\,2}} Cov(X,Y). \tag{5.17}$$

Substituting expressions (5.16) and (5.17) into (5.15), we derive the following

$$Cov\left(\frac{X}{Y},\frac{A}{n}\right) \approx \frac{1}{n}\left[\frac{\mu_X\mu_A}{\mu_Y} + \frac{1}{2}\frac{2\mu_X\mu_A}{\mu_Y^{\,3}} Var(Y) - \frac{\mu_A}{\mu_Y^{\,2}} Cov(X,Y) + \right.$$
$$\left.\frac{1}{\mu_Y} Cov(X,A) - \frac{\mu_X}{\mu_Y^{\,2}} Cov(A,Y) - \frac{\mu_X\mu_A}{\mu_Y} - \frac{1}{2}\frac{2\mu_X\mu_A}{\mu_Y^{\,3}} Var(Y) + \frac{\mu_A}{\mu_Y^{\,2}} Cov(X,Y)\right].$$

It follows that

$$Cov\left(\frac{X}{Y},\frac{A}{n}\right) \approx \frac{1}{n}\left[\frac{1}{\mu_Y} Cov(X,A) - \frac{\mu_X}{\mu_Y^{\,2}} Cov(A,Y)\right] = \frac{1}{n\mu_Y}\left[Cov(X,A) - \frac{\mu_X}{\mu_Y} Cov(A,Y)\right].$$

Finally,

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

Result 5.3

Let $X=n^v_{ik},Y=\sum_{i=1}^{r}n^v_{ik},A=n_{lj}$ be three random variables and n fixed. Using Result 5.2, an approximate expression for $Cov\left(\hat{q}_{ik,}\ \hat{\Pi}_{lj}\right)$ is given by

$$Cov\left(\frac{n^v_{ik}}{\sum_{i=1}^{r}n^v_{ik}},\frac{n_{lj}}{n}\right)\approx\frac{1}{nE\left(\sum_{i=1}^{r}n^v_{ik}\right)}\left[Cov\left(n^v_{ik},n_{lj}\right)-\frac{E\left(n^v_{ik}\right)}{E\left(\sum_{i=1}^{r}n^v_{ik}\right)}Cov\left(n_{lj},\sum_{i=1}^{r}n^v_{ik}\right)\right].$$

An estimator for the covariance term of interest is given by

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

$$\left.-\frac{\hat{E}\left(n^v_{ik}\right)}{\hat{E}\left(\sum_{i=1}^{r}n^v_{ik}\right)}\left[n^v\,\hat{pr}\left(Y^*_{\xi t}=i,Y^*_{\xi t+1}=j,Y_{\xi t}=k\right)-\frac{\hat{E}\left(n^v_{ik}\right)\hat{E}\left(n_{ij}\right)}{n^v}-\sum_{l\neq i}\frac{\hat{E}\left(n^v_{ik}\right)\hat{E}\left(n_{lj}\right)}{n^v}\right]\right\}. \tag{5.18}$$

Proof

We start the proof by evaluating $Cov\left(n^v_{ik},n_{lj}\right)$

$$Cov\left(n^v_{ik},n_{lj}\right)=E\left(n^v_{ik},n_{lj}\right)-E\left(n^v_{ik}\right)E\left(n_{lj}\right). \tag{5.19}$$

We define the following indicator variables

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

At this point, we need to make the following comments:

1. k is known only for those units that belong to the validation sample.
2. Since the validation sample is selected by sub-sampling units from the main sample, the main sample and the validation sample will have some units in common. This implies that for all ξ units lj is known.
3. S, s denote the main and the validation sample respectively.

It follows that

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

$$=\sum_{\xi\in s} E\left(I_{\xi}J_{\xi}\right)+\sum_{\substack{\xi\in s,\xi'\in S\\ \xi\neq\xi'}} E\left(I_{\xi}J_{\xi'}\right)\Rightarrow$$

$$E\left(n_{ik}^{v} n_{lj}\right)=\sum_{\xi\in s} E\left(I_{\xi}J_{\xi}\right)+\sum_{\substack{\xi\in s,\xi'\in S\\ \xi\neq\xi'}} E\left(I_{\xi}\right)E\left(J_{\xi'}\right). \qquad (5.20)$$

Furthermore,

$$E\left(n_{ik}^{v}\right)E\left(n_{lj}\right)=\sum_{\xi\in s}\sum_{\xi'\in S} E\left(I_{\xi}\right)E\left(J_{\xi'}\right)=\sum_{\xi\in s} E\left(I_{\xi}\right)E\left(J_{\xi}\right)+\sum_{\substack{\xi\in s,\xi'\in S\\ \xi\neq\xi'}} E\left(I_{\xi}\right)E\left(J_{\xi'}\right). \qquad (5.21)$$

Substituting expressions (5.20) and (5.21) into (5.19), we obtain the following

$$Cov\left(n_{ik}^{v}, n_{lj}\right)=\sum_{\xi\in s} E\left(I_{\xi}J_{\xi}\right)+\sum_{\substack{\xi\in s,\xi'\in S\\ \xi\neq\xi'}} E\left(I_{\xi}\right)E\left(J_{\xi'}\right)-\sum_{\xi\in s} E\left(I_{\xi}\right)E\left(J_{\xi}\right)-\sum_{\substack{\xi\in s,\xi'\in S\\ \xi\neq\xi'}} E\left(I_{\xi}\right)E\left(J_{\xi'}\right)\Rightarrow$$

$$Cov\left(n_{ik}^{v}, n_{lj}\right)=\sum_{\xi\in s} E\left(I_{\xi}J_{\xi}\right)-\sum_{\xi\in s} E\left(I_{\xi}\right)E\left(J_{\xi}\right). \qquad (5.22)$$

From (5.22), it follows that an estimator of the covariance term is given by

$$\hat{C}ov\left(n_{ik}^{v}, n_{lj}\right)=\sum_{\xi\in s} \hat{E}\left(I_{\xi}J_{\xi}\right)-\sum_{\xi\in s} \hat{E}\left(I_{\xi}\right)\hat{E}\left(J_{\xi}\right)=n^{v}\,\hat{p}r\left(Y_{\xi t}^{*}=l, Y_{\xi t+1}^{*}=j, Y_{\xi t}^{*}=i, Y_{\xi t}=k\right)-$$

$$n^{v}\frac{\hat{E}\left(n_{ik}^{v}\right)}{n^{v}}\frac{\hat{E}\left(n_{lj}\right)}{n^{v}}\Rightarrow$$

$$\hat{C}ov\left(n_{ik}^{v}, n_{lj}\right)=n^{v}\,\hat{p}r\left(Y_{\xi t}^{*}=l, Y_{\xi t+1}^{*}=j, Y_{\xi t}^{*}=i, Y_{\xi t}=k\right)-\frac{\hat{E}\left(n_{ik}^{v}\right)\hat{E}\left(n_{lj}\right)}{n^{v}} \qquad (5.23)$$

where

$$\begin{cases} \hat{pr}\left(Y^*_{\xi t} = l, Y^*_{\xi t+1} = j, Y^*_{\xi t} = i, Y_{\xi t} = k\right) = 0 & if \ l \neq i \\ \hat{pr}\left(Y^*_{\xi t} = l, Y^*_{\xi t+1} = j, Y^*_{\xi t} = i, Y_{\xi t} = k\right) \neq 0 & if \ l = i \ . \end{cases} \tag{5.24}$$

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

$$\frac{E\left(n^v_{ik}\right)}{E\left(\sum_{i=1}^{r} n^v_{ik}\right)} Cov\left(n_{lj}, \sum_{i=1}^{r} n^v_{ik}\right). \tag{5.25}$$

This can be done as follows:

$$\frac{E\left(n^v_{ik}\right)}{E\left(\sum_{i=1}^{r} n^v_{ik}\right)} Cov\left(n_{lj}, \sum_{i=1}^{r} n^v_{ik}\right) = \frac{E\left(n^v_{ik}\right)}{E\left(\sum_{i=1}^{r} n^v_{ik}\right)} Cov\left[n_{lj}, \left(n^v_{1k} + n^v_{2k} + \ldots + n^v_{rk}\right)\right] =$$

$$\frac{E\left(n^v_{ik}\right)}{E\left(\sum_{i=1}^{r} n^v_{ik}\right)} \left\{E\left[n_{lj}\left(n^v_{1k} + n^v_{2k} + \ldots + n^v_{rk}\right)\right] - E\left(n_{lj}\right) E\left(n^v_{1k} + n^v_{2k} + \ldots + n^v_{rk}\right)\right\}.$$

Consequently,

$$\frac{E\left(n^v_{ik}\right)}{E\left(\sum_{i=1}^{r} n^v_{ik}\right)} Cov\left(n_{lj}, \sum_{i=1}^{r} n^v_{ik}\right) = \frac{E\left(n^v_{ik}\right)}{E\left(\sum_{i=1}^{r} n^v_{ik}\right)} \left[Cov\left(n_{lj}, n^v_{1k}\right) + Cov\left(n_{lj}, n^v_{2k}\right) + \ldots + Cov\left(n_{lj}, n^v_{rk}\right)\right]. \tag{5.26}$$

We estimate the covariance terms involved in (5.26) by employing expression (5.23) as follows:

$$\hat{Cov}\left(n^v_{ik}, n_{lj}\right) = \begin{cases} n^v \, \hat{pr}\left(Y^*_{\xi t} = i, Y^*_{\xi t+1} = j, Y^*_{\xi t} = i, Y_{\xi t} = k\right) - \dfrac{\hat{E}\left(n^v_{ik}\right) \hat{E}\left(n_{ij}\right)}{n^v} & if \ l = i \\ -\dfrac{\hat{E}\left(n^v_{ik}\right) \hat{E}\left(n_{lj}\right)}{n^v} & if \ l \neq i \ . \end{cases} \tag{5.27}$$

Consequently,

$$\frac{\hat{E}\left(n^{v}_{ik}\right)}{\hat{E}\left(\sum_{i=1}^{r} n^{v}_{ik}\right)} \hat{Cov}\left(n_{lj}, \sum_{i=1}^{r} n^{v}_{ik}\right) = \frac{\hat{E}\left(n^{v}_{ik}\right)}{\hat{E}\left(\sum_{i=1}^{r} n^{v}_{ik}\right)} \left[n^{v}\, \hat{pr}\left(Y^{*}_{\xi t} = i, Y^{*}_{\xi t+1} = j, Y_{\xi t} = k\right) - \frac{\hat{E}\left(n^{v}_{ik}\right)\hat{E}\left(n_{ij}\right)}{n^{v}} \right.$$
$$\left. - \sum_{l \neq i} \frac{\hat{E}\left(n^{v}_{ik}\right)\hat{E}\left(n_{lj}\right)}{n^{v}} \right]. \tag{5.28}$$

Combining (5.23) and (5.28), we obtain an estimator for the covariance terms that are of interest for our analysis.

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

□

<u>Evaluation of the Jacobian Matrices from the First Application of the δ-Method</u>

<u>Analytical Differentiation</u>

In order to complete the variance estimation process, we need to evaluate the Jabobian matrices from the first application of the delta method given by

$$\nabla_{\Theta} = \frac{\partial vec\left[\underset{r\times r}{Q^{-1}}\, \underset{r\times r}{\Pi}\, \underset{r\times r}{\left(Q^{-1}\right)^{\mathrm{T}}}\right]}{\partial \Theta}\Bigg|_{\Theta = \hat{\Theta}}.$$

The general form of this Jacobian is given by the following 9×18 matrix

$$\nabla_{\Theta} = \begin{bmatrix} \frac{\partial(P_{11})}{\partial q_{11}} & \frac{\partial(P_{11})}{\partial q_{21}} & \cdots & \frac{\partial(P_{11})}{\partial \Pi_{33}} \\ \frac{\partial(P_{21})}{\partial q_{11}} & \frac{\partial(P_{21})}{\partial q_{21}} & \cdots & \frac{\partial(P_{21})}{\partial \Pi_{33}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial(P_{33})}{\partial q_{11}} & \frac{\partial(P_{33})}{\partial q_{21}} & \cdots & \frac{\partial(P_{33})}{\partial \Pi_{33}} \end{bmatrix}. \tag{5.29}$$

One way of evaluating the elements of (5.29) is analytically. For $\frac{\partial vec(P)}{\partial[vec(\Pi)]^T}$ we can follow Harville (1997 p.366) and evaluate these elements using the following result

$$\frac{\partial vec(AXB)}{\partial[vec(X)]^T} = B^T \otimes A. \tag{5.30}$$

Applying this result in our case, we derive the following

$$\frac{\partial vec\left[Q^{-1}\Pi\left(Q^{-1}\right)^{\mathrm{T}}\right]}{\partial[vec(\Pi)]^T} = \left[\left(Q^{-1}\right)^{\mathrm{T}}\right]^{\mathrm{T}} \otimes Q^{-1} = Q^{-1} \otimes Q^{-1}. \tag{5.31}$$

Furthermore, $\frac{\partial vec(P)}{\partial[vec(Q)]^T}$ can be evaluated analytically but this involves more complex expressions that cannot be expressed easily in a general form.

Numerical Differentiation

Alternatively, one can employ numerical differentiation to evaluate the elements of (5.29) and (5.10). The method we utilise is the method of central differences (Dennis and Schnabel, 1983).

We employ both analytical and numerical differentiation. The numerical approach is used for validating the analytical results. Substituting the results from the evaluation of the Jacobian matrices using either analytical or numerical differentiation and expressions (5.5), (5.9), (5.10) and (5.11) (for the case of an internal validation sample that is selected independently from the main sample or for the case of an external validation sample) or (5.18) (for the case of an internal validation sample that is selected by sub-sampling units from the main sample) into (5.4), we obtain variance estimates for the adjusted gross flows estimated from the conventional point estimator.

5.3 Variance Estimation for Alternative Moment-type Estimators of the Adjusted Gross Flows

Based on the results from Section 5.2, we now develop variance estimators the alternative moment-type estimators of the adjusted gross flows.

5.3.1 Variance of the Modified Estimator

Utilising the results from Section 5.2, we now derive a variance estimator for the modified estimator of the adjusted gross flows (see Section 2.4.2). These variance estimates are also required for computing the set of adaptive weights of the composite estimator (see Section 2.4.3). Recall that the modified estimator was developed by employing a double sampling scheme under which the validation sample is selected independently from the main sample and from the same target population. The modified estimator is defined as follows:

$$vec\left(\hat{P}^{mod}\right) = w_{mod}\; vec\left(\hat{P}^{st}\right) + \left(1 - w_{mod}\right) vec\left(\hat{P}^{v}\right), \quad w_{mod} = \frac{n}{n + n^{v}}. \quad (5.32)$$

Taking the variance operator on both sides of (5.32), ignoring $Cov\left(\hat{P}^{st}, \hat{P}^{v}\right)$ and taking into account that w_{mod} is fixed we have that

$$Var\left[vec\left(\hat{P}^{mod}\right)\right] = w_{mod}^{2} Var\left[vec\left(\hat{P}^{st}\right)\right] + \left(1 - w_{mod}\right)^{2} Var\left[vec\left(\hat{P}^{v}\right)\right]. \quad (5.33)$$

An estimate of $Var\left[vec\left(\hat{P}^{st}\right)\right]$ can be found by utilising the results from Section 5.2 for the case of an external validation sample. An estimate of $Var\left[vec\left(\hat{P}^{v}\right)\right]$ can be found by utilising the results from Section 5.2 for the case of an internal validation sample that is selected by sub-sampling units from the main sample.

5.3.2 Variance of the Composite Estimator

Utilising the results from Section 5.2, we derive a variance estimator for the composite estimator (see Section 2.4.3). The general form of the composite estimator is given by

$$vec\left(\hat{P}^{comp}\right) = \left[w_{comp}w_{mod} + \left(1 - w_{comp}\right)\right] vec\left(\hat{P}^{st}\right) + w_{comp}\left(1 - w_{mod}\right) vec\left(\hat{P}^{v}\right). \quad (5.34)$$

Taking the variance operator on both sides of (5.34) we have that

$$Var\left[vec\left(\hat{P}^{comp}\right)\right] = Var\left\{\left[w_{comp}w_{mod} + \left(1 - w_{comp}\right)\right] vec\left(\hat{P}^{st}\right) + w_{comp}\left(1 - w_{mod}\right) vec\left(\hat{P}^{v}\right)\right\}. \quad (5.35)$$

We distinguish two cases: (a) the composite weights w_{comp} are fixed and (b) the composite weights w_{comp} are adaptive, i.e. random, since they are estimated by minimising the mean squared error of the composite estimator. For the first case, the variance of the composite estimator, ignoring $Cov\left(\hat{P}^{st}, \hat{P}^{v}\right)$, is given below

$$Var\left[vec\left(\hat{P}^{comp-fx}\right)\right] = \left[w_{comp}w_{mod} + \left(1 - w_{comp}\right)\right]^2 Var\left[vec\left(\hat{P}^{st}\right)\right] + w_{comp}^2\left(1 - w_{mod}\right)^2 Var\left[vec\left(\hat{P}^{v}\right)\right]. (5.36)$$

An estimate of $Var\left[vec\left(\hat{P}^{st}\right)\right]$ is derived by employing the results from Section 5.2 for the case of an external validation sample and an estimate of $Var\left[vec\left(\hat{P}^{v}\right)\right]$ is derived by employing the results of Section 5.2 for the case of an internal validation sample that is selected by sub-sampling units from the main sample. Variance estimation in the case of the composite estimator with adaptive weights becomes more complicated since the composite weights can no longer be considered as fixed quantities. One possible solution, for approximating the variance of a composite estimator with adaptive weights, is to use the jackknife method (Kuo 1989).

5.4 Variance Estimation for the Maximum Likelihood Estimator of the Adjusted Gross Flows

In this section, we develop a variance estimator for the maximum likelihood estimator when the validation sample is selected independently from the main sample and from the same target population (see Section 3.2.1). Variance estimation for the maximum likelihood estimates of the adjusted gross flows can be placed into the general framework of maximum likelihood estimation. This implies the use of the inverse of the information matrix. However, due to the parameterisation of the measurement error model in a missing data framework, variance estimation must reflect the additional variability introduced by the existence of missing data. One way of obtaining variance estimates for the parameters of interest in an EM framework is by using the Missing Information Principle (Woodbury 1977, Efron and Hinkley 1978, Louis 1982).

Denote by $\overset{\wedge}{\Theta}$ the vector of maximum likelihood estimates, by Z^m, Z^v the missing data in the main and in the validation sample respectively and by D^m, D^v the observed data in the main and in the validation sample respectively. The missing data and the observed data define the complete data denoted by D^c. The Missing Information Principle is defined as

$$Observed\ Information = Complete\ Information - Missing\ Information \quad (5.37)$$

Lemma 5.2

The complete information matrix is evaluated using the following expression

$$Complete\ Information = E\left[-\frac{\partial^2 l(\Theta; D^c)}{\partial\Theta\partial\Theta^T} \mid D^m, D^v\right]. \quad (5.38)$$

Proof

Proof of this lemma can be found in Tanner (1996, p.75).

□

Lemma 5.3

The missing information matrix is evaluated using the following expression

$$Missing\ Information = Var\left[\frac{\partial l(\Theta; D^c)}{\partial\Theta} \mid D^m, D^v\right]. \quad (5.39)$$

Proof

Proof of this lemma can be found in Tanner (1996, p.75).

□

When full information exists, the second component of (5.37) disappears and variance estimates for the maximum likelihood estimates are derived by employing the inverse of the complete information matrix. In the presence of missing data, the effect of the missing information matrix is to reduce the available information and, thus, introduce extra variability. In this section, we derive estimates of the complete information matrix and of the missing information matrix at $\hat{\Theta}$. Having derived these estimates, we can then apply the Missing Information Principle to derive the observed information matrix and the inverse of the observed information matrix to compute appropriate variance estimates.

Lemma 5.4

Conditionally on the information available from the validation sample, there are r^2 multinomial distributions defined.

Proof

Before selecting the validation sample, the only fixed quantity is the size of this sample n^v. This implies that n_k^v is random. The EM algorithm conditions on the information available from the validation sample. Thus, conditionally on this information, n_k^v is considered to be fixed. Consequently, there are r^2 multinomial distributions defined (see Section 3.2.1).

□

Lemma 5.5

Conditionally on the information available from the main sample, there are r^2 multinomial distributions defined by the r^2 columns of the cross-classification of the observed with the true classifications.

Proof

Before selecting the main sample, the only fixed quantity is the size of this sample n. This implies that $n_{\bullet j}$ is random. The EM algorithm conditions on the information available from the main sample. Thus, conditionally on this information, $n_{\bullet j}$ is considered to be fixed. Consequently, there are r^2 multinomial distributions defined by the r^2 columns of the cross-classification of the observed with the true classifications (see Section 3.2.1).

□

Evaluating the Complete Information Matrix

The first step in the application of the Missing Information Principle involves the evaluation of the complete information matrix. Some of the second order derivatives required for computing this information matrix can be found in Appendix III. These quantities are evaluated at the last step of the EM algorithm.

Evaluating the Missing Information Matrix

The second step in the application of the Missing Information Principle involves the evaluation of the missing information matrix. This is achieved by computing the variance of the score functions.

Definition 5.1

Let X denote a $d \times 1$ vector of random variables. It follows that the variance of X is given by the following $d \times d$ covariance matrix

$$Var(X) = E\left(XX^T\right) - E(X)E\left(X^T\right). \tag{5.40}$$

Let Θ denote the vector of parameters with elements θ_i, $i = 1, \cdots, \omega$. Utilising definition 5.1, the covariance matrix of the score functions will be of dimension $\left(2r^2 - r - 1\right) \times \left(2r^2 - r - 1\right)$ with diagonal and off-diagonal elements given respectively by the following general expressions

$$\begin{cases} V_{ij} = Var\left[\dfrac{\partial l(\Theta; D^c)}{\partial \theta_i} \mid D^m, D^v\right], & i = j \\ V_{ij} = Cov\left[\dfrac{\partial l(\Theta; D^c)}{\partial \theta_i}, \dfrac{\partial l(\Theta; D^c)}{\partial \theta_j} \mid D^m, D^v\right] & i \neq j\ . \end{cases} \tag{5.41}$$

Lemma 5.6

For the main sample, under simple random sampling, the following holds

$$\overset{\wedge}{Var}\left(n_{i\bullet}^{(*)} \mid D^m, \overset{\wedge}{\Theta}\right) = \sum_{j=1}^{r^2} \left[n_{\bullet j} \frac{\hat{E}\left(n_{ij}^{(*)} \mid D^m, \overset{\wedge}{\Theta}\right)}{n_{\bullet j}}\right]\left[1 - \frac{\hat{E}\left(n_{ij}^{(*)} \mid D^m, \overset{\wedge}{\Theta}\right)}{n_{\bullet j}}\right]. \tag{5.42}$$

Proof

We start the proof using the standard definition for the variance of a sum of random variables

$$Var\left(n_{i\bullet}^{(*)} \mid D^m, \Theta\right) = \sum_{j=1}^{r^2} Var\left(n_{ij}^{(*)} \mid D^m, \Theta\right) + 2\sum_{j=1}^{r} Cov\left(n_{ij}^{(*)}, n_{ij'}^{(*)} \mid D^m, \Theta\right).$$

Using Lemma 5.5, there are r^2 multinomial distributions defined. This implies that random variables that refer to different conditional distributions are independent. Thus, the covariance terms of the above expression are equal to zero. Using results for the variance and the covariance of binomial random variables (see also (5.9)) and information available from the main sample, we derive the variance component of interest.

□

Lemma 5.7

For the validation sample, under simple random sampling, the following holds

$$\hat{Var}\left(n_{i\bullet}^{v(*)} \mid D^v, \hat{\Theta}\right) =$$

$$\sum_{j=1}^{r}\left[n_k^v \frac{\hat{E}\left(n_{ij}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v}\right]\left[1 - \frac{\hat{E}\left(n_{ij}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v}\right] \text{-}2\sum_{j=1}^{r-1}\sum_{j'=r-1}^{r}\left[n_k^v \frac{\hat{E}\left(n_{ij}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v} \frac{\hat{E}\left(n_{ij'}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v}\right]$$

$$+2\left[n_k^v \frac{\hat{E}\left(n_{ir-1}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v} \frac{\hat{E}\left(n_{ir-1}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v}\right] + \cdots + \sum_{j=r^2-r+1}^{r^2}\left[n_k^v \frac{\hat{E}\left(n_{ij}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v}\right]\left[1 - \frac{\hat{E}\left(n_{ij}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v}\right]$$

$$\text{-}2\sum_{j=r^2-r+1}^{r^2-1}\sum_{j'=r^2-1}^{r^2}\left[n_k^v \frac{\hat{E}\left(n_{ij}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v} \frac{\hat{E}\left(n_{ij'}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v}\right] + 2\left[n_k^v \frac{\hat{E}\left(n_{ir^2-1}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v} \frac{\hat{E}\left(n_{ir^2-1}^{v(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v}\right]. \qquad (5.43)$$

Proof

We start the proof using the standard definition for the variance of a sum of random variables and we decompose this sum into a sum with r components. This is because the covariance terms involved in the summation exist only for random variables that refer to the same conditional distribution (see Lemma 5.4). Consequently,

$$Var\left(n_{i\bullet}^{v(*)} \mid D^v, \Theta\right) = \sum_{j=1}^{r} Var\left(n_{ij}^{v(*)} \mid D^v, \Theta\right) + 2\sum_{j=1}^{r-1} \sum_{j'=r-1}^{r} Cov\left(n_{ij}^{v(*)}, n_{ij'}^{v(*)} \mid D^v, \Theta\right) + \cdots +$$

$$\sum_{j=r^2-r+1}^{r^2} Var\left(n_{ij}^{v(*)} \mid D^v, \Theta\right) + 2 \sum_{j=r^2-r+1}^{r^2-1} \sum_{j'=r^2-1}^{r^2} Cov\left(n_{ij}^{v(*)}, n_{ij'}^{v(*)} \mid D^v, \Theta\right).$$

Replacing the variance and the covariance terms using results for binomial random variables, subtracting the covariance terms between identical random variables resulting from the double summation and utilising information from the validation sample, we end up with the required result.

□

Lemma 5.8

For the main sample, under simple random sampling, the following holds

$$\hat{Cov}\left(n_{i\bullet}^{(*)} n_{i'\bullet}^{(*)} \mid D^m, \hat{\Theta}\right) = -\sum_{j=1}^{r^2} n_{\bullet j} \frac{\hat{E}\left(n_{ij}^{(*)} \mid D^m, \hat{\Theta}\right)}{n_{\bullet j}} \frac{\hat{E}\left(n_{i'j}^{(*)} \mid D^m, \hat{\Theta}\right)}{n_{\bullet j}}. \tag{5.44}$$

Proof

Using Lemma 5.5, the required covariance term can be expanded as follows

$$Cov\left(n_{i\bullet}^{(*)}, n_{i'\bullet}^{(*)} \mid D^m, \Theta\right) = Cov\left[\left(\sum_{j=1}^{r^2} n_{ij}^{(*)} \mid D^m, \Theta\right), \left(\sum_{j=1}^{r^2} n_{i'j}^{(*)} \mid D^m, \Theta\right)\right].$$

This is due to the fact that these covariance terms exist only for random variables that refer to the same conditional distribution. Substituting results for the covariance between binomial random variables into the expression above, we derive the required result.

□

Lemma 5.9

For the validation sample, under simple random sampling, the following holds

$$\hat{Cov}\left(n_{i\bullet}^{v\,(*)}, n_{i'\bullet}^{v\,(*)} \mid D^v, \hat{\Theta}\right) = \left[\sum_{j=1}^{r} \sum_{j'=1}^{r} -n_k^v \frac{\hat{E}\left(n_{ij}^{v\,(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v} \frac{\hat{E}\left(n_{i'j'}^{v\,(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v} + \cdots + \right.$$
$$\left. \sum_{j=r^2-r+1}^{r^2} \sum_{j'=r^2-r+1}^{r^2} -n_k^v \frac{\hat{E}\left(n_{ij}^{v\,(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v} \frac{\hat{E}\left(n_{i'j'}^{v\,(*)} \mid D^v, \hat{\Theta}\right)}{n_k^v}\right]. \tag{5.45}$$

Proof

In order to prove this lemma, we need to utilise Lemma 5.4. Using Lemma 5.4, we realise that the requested covariance terms exist only for random variables that refer to the same conditional distribution. For example, when $r = 3$ the following covariance terms exist

$$\begin{gathered}
Cov\left(n_{1\cdot}^{v(*)}, n_{2\cdot}^{v(*)} \mid D^v, \Theta\right), Cov\left(n_{1\cdot}^{v(*)}, n_{3\cdot}^{v(*)} \mid D^v, \Theta\right), Cov\left(n_{2\cdot}^{v(*)}, n_{3\cdot}^{v(*)} \mid D^v, \Theta\right) \\
Cov\left(n_{4\cdot}^{v(*)}, n_{5\cdot}^{v(*)} \mid D^v, \Theta\right), Cov\left(n_{4\cdot}^{v(*)}, n_{6\cdot}^{v(*)} \mid D^v, \Theta\right), Cov\left(n_{5\cdot}^{v(*)}, n_{6\cdot}^{v(*)} \mid D^v, \Theta\right) \\
Cov\left(n_{7\cdot}^{v(*)}, n_{8\cdot}^{v(*)} \mid D^v, \Theta\right), Cov\left(n_{7\cdot}^{v(*)}, n_{9\cdot}^{v(*)} \mid D^v, \Theta\right), Cov\left(n_{8\cdot}^{v(*)}, n_{9\cdot}^{v(*)} \mid D^v, \Theta\right).
\end{gathered}$$

Each one of the covariance terms above can be decomposed into a sum with r terms as follows

$$Cov\left(n_{i\cdot}^{v(*)}, n_{i'\cdot}^{v(*)} \mid D^v, \Theta\right) = \left[\underbrace{\sum_{j=1}^{r}\sum_{j'=1}^{r} Cov\left(n_{ij}^{v(*)}, n_{i'j'}^{v(*)} \mid D^v, \Theta\right) + \cdots + \sum_{j=r^2-r+1}^{r^2}\sum_{j'=r^2-r+1}^{r^2} Cov\left(n_{ij}^{v(*)}, n_{i'j}^{v(*)} \mid D^v, \Theta\right)}_{r-terms}\right].$$

Replacing the covariance terms using results for binomial random variables and information from the validation sample, we end up with the required result.

□

An estimate of the covariance matrix of the score functions can be obtained using the first order derivatives of the augmented log-likelihood (see Appendix III), Lemmas 5.6-5.9 and standard definitions for the variance and the covariance of sums of random variables. In the sequel, we present general expressions for computing some of the elements of this covariance matrix.

Result 5.4

$$Var\left[\frac{\partial l(\Theta; D^c)}{\partial P_i} \mid D^m, D^v\right] =$$

$$\frac{1}{P_i^2}\left[Var\left(n_{i\cdot}^{(*)} \mid D^m, \Theta\right) + Var\left(n_{i\cdot}^{v(*)} \mid D^v, \Theta\right)\right] + \frac{1}{\left(1 - \sum_{i=1}^{r^2-1} P_i\right)^2}\left[Var\left(n_{r^2\cdot}^{(*)} \mid D^m, \Theta\right) + Var\left(n_{r^2\cdot}^{v(*)} \mid D^v, \Theta\right)\right]$$

$$-\frac{2}{P_i\left(1 - \sum_{i=1}^{r^2-1} P_i\right)}\left[Cov\left(n_{i\cdot}^{(*)}, n_{r^2\cdot}^{(*)} \mid D^m, \Theta\right) + Cov\left(n_{i\cdot}^{v(*)}, n_{r^2\cdot}^{v(*)} \mid D^v, \Theta\right)\right]$$

(5.46)

Proof

Taking the variance operator on both sides of the first order derivative with respect to P_i (see Appendix III), we obtain the following

$$Var\left[\frac{\partial l(\Theta;D^c)}{\partial P_i} \mid D^m, D^v\right] = Var\left[\frac{\left(n_{i\bullet}^{(*)} + n_{i\bullet}^{v(*)}\right)}{P_i} - \frac{\left(n_{r^2\bullet}^{(*)} + n_{r^2\bullet}^{v(*)}\right)}{1 - \sum_{i=1}^{r^2-1} P_i} \mid D^m, D^v, \Theta\right].$$

The required result is obtained by applying the definition for the variance of a sum of random variables and taking into account the fact that the main sample and the validation sample are independent. An estimate of this variance component is obtained using Lemmas 5.6-5.9.

□

Result 5.5

$$\begin{aligned}
Cov\left[\frac{\partial l(\Theta;D^c)}{\partial P_i}, \frac{\partial l(\Theta;D^c)}{\partial P_{i'}} \mid D^m, D^v\right] &= \frac{1}{P_i P_{i'}}\left[Cov\left(n_{i\bullet}^{v(*)}, n_{i'\bullet}^{v(*)} \mid D^v, \Theta\right) + Cov\left(n_{i\bullet}^{(*)}, n_{i'\bullet}^{(*)} \mid D^m, \Theta\right)\right] \\
&- \frac{1}{P_i P_{r^2}}\left[Cov\left(n_{i\bullet}^{v(*)}, n_{r^2\bullet}^{v(*)} \mid D^v, \Theta\right) + Cov\left(n_{i\bullet}^{(*)}, n_{r^2\bullet}^{(*)} \mid D^m, \Theta\right)\right] \\
&- \frac{1}{P_{i'} P_{r^2}}\left[Cov\left(n_{i'\bullet}^{v(*)}, n_{r^2\bullet}^{v(*)} \mid D^v, \Theta\right) + Cov\left(n_{i'\bullet}^{(*)}, n_{r^2\bullet}^{(*)} \mid D^m, \Theta\right)\right] \\
&+ \frac{1}{P_{r^2}}\left[Var\left(n_{r^2\bullet}^{v(*)} \mid D^v, \Theta\right) + Var\left(n_{r^2\bullet}^{(*)} \mid D^m, \Theta\right)\right].
\end{aligned}$$

(5.47)

Proof

Using the first order derivative of the augmented log-likelihood with respect to P_i (see Appendix III), the covariance term of interest is expressed as follows

$$Cov\left[\frac{\left(n_{i\bullet}^{(*)} + n_{i\bullet}^{v(*)}\right)}{P_i} - \frac{\left(n_{r^2\bullet}^{(*)} + n_{r^2\bullet}^{v(*)}\right)}{1 - \sum_{i=1}^{r^2-1} P_i}, \frac{\left(n_{i'\bullet}^{(*)} + n_{i'\bullet}^{v(*)}\right)}{P_{i'}} - \frac{\left(n_{r^2\bullet}^{(*)} + n_{r^2\bullet}^{v(*)}\right)}{1 - \sum_{i=1}^{r^2-1} P_i} \mid D^m, D^v, \Theta\right].$$

However, we note that the covariance term above has the following general form

$$Cov(A - B, C - B) = Cov(A, C) - Cov(A, B) - Cov(B, C) + Var(B). \quad (5.48)$$

Applying (5.48) to the covariance term of interest, we derive the required result. An estimate of this covariance component is obtained using Lemmas 5.6-5.9.

□

Using also the definitions for the variance and the covariance of sums of random variables and Lemmas 5.6-5.9, the following quantities can be evaluated analytically

$$Cov\left[\frac{\partial l(\Theta;D^c)}{\partial P_i},\frac{\partial l(\Theta;D^c)}{\partial q_{ij}} \mid D^m,D^v\right], \tag{5.49}$$

$$Var\left[\frac{\partial l(\Theta;D^c)}{\partial q_{ij}} \mid D^m,D^v\right], \tag{5.50}$$

$$Cov\left[\frac{\partial l(\Theta;D^c)}{\partial q_{ij}},\frac{\partial l(\Theta;D^c)}{\partial q_{i'j'}} \mid D^m,D^v\right] \quad (ij)\neq(i'j'). \tag{5.51}$$

However, these are more complex expressions that are not easily expressed in a general form. Note also that q_{ij} refers to the longitudinal misclassification probabilities. These probabilities need to be replaced, under ICE, by products of cross-sectional misclassification probabilities. This reduces the dimensionality of the problem. The elements of the covariance matrix of the score functions are evaluated at the last step of the EM algorithm. After evaluating the complete information matrix and the missing information matrix, the observed information matrix is defined as the difference of these two matrices. Inverting the observed information matrix, results in an estimate of the covariance matrix of the maximum likelihood estimates.

5.4.1 Evaluating the Complete Information Matrix and the Missing Information Matrix Using Simulation

In spite of being able to derive general expressions for the complete information matrix and for the missing information, it is tedious to evaluate these expressions analytically. The main problem arises in evaluating of the covariance matrix of the score functions. An alternative solution for approximating the components of the Missing Information Principle is offered by means of simulation. The idea is described in Tanner (1996). The algorithm is as follows. Having arrived at the maximum likelihood estimates, we generate H complete datasets by drawing

$$Z_1^v,Z_2^v,\ldots,Z_H^v \overset{iid}{\sim} p\left(Z^v \mid D^v,\hat{\Theta}\right); \tag{5.52}$$

$$Z_1^m,Z_2^m,\ldots,Z_H^m \overset{iid}{\sim} p\left(Z^m \mid D^m,\hat{\Theta}\right) \tag{5.53}$$

where $p\left(Z^v \mid D^v, \overset{\wedge}{\Theta}\right), p\left(Z^m \mid D^m, \overset{\wedge}{\Theta}\right)$ denote the conditional distributions of the missing data in the validation sample and in the main sample respectively given the observed data and the maximum likelihood estimates and H denotes the total number of simulations. The conditional distributions are defined by Lemma 5.4 and Lemma 5.5. This first step of the simulation can be viewed as the imputation step. Having replaced the missing data with imputed values in simulation (h), we derive complete data $D^{c(h)}$ 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 over simulations defined as

$$E\left[-\frac{\partial^2 l(\Theta; D^c)}{\partial\Theta\,\partial\Theta^T} \mid D^m, D^v\right] = \frac{1}{H}\sum_{h=1}^{H} -\frac{\partial^2 l\left(\Theta; D^{c(h)}\right)}{\partial\Theta\,\partial\Theta^T}, \tag{5.54}$$

$$Var\left[\frac{\partial l(\Theta; D^c)}{\partial\Theta} \mid D^m, D^v\right] = \frac{1}{H}\sum_{h=1}^{H}\left\{\frac{\partial l\left(\Theta; D^{c(h)}\right)}{\partial\Theta} - E\left[\frac{\partial l\left(\Theta; D^{c(h)}\right)}{\partial\Theta}\right]\right\}^2. \tag{5.55}$$

5.4.2 Variance Estimation for the Likelihood-based Adjusted Estimates in a Cross-sectional Framework

In Section 2.2.1.3, we parameterised the cross-sectional measurement error model in a missing data framework and maximum likelihood estimates were derived via the EM algorithm. Variance estimates for the cross-sectional maximum likelihood estimates can be also derived using the Missing Information Principle. The crucial difference between the longitudinal framework and the cross-sectional framework is that in the latter case missing data exist only in the main sample. The complete information matrix is evaluated at the last step of the EM algorithm using the second order derivatives of the augmented log-likelihood (see Appendix I). The covariance matrix of the score functions is evaluated also at the last step of the EM using the first order derivatives of the augmented log-likelihood (see Appendix I), results for the variance of binomial random variables and the lemma below.

Lemma 5.10

Conditionally on the information available from the main sample, there are r multinomial distributions defined by the r columns of the cross-classification between the observed and the true classifications.

Proof

The proof is identical to the proof of Lemma 5.5.

□

Alternatively, one can use the simulation approach described in Section 5.4.1 along with the expressions given in Appendix I. The implementation of the simulation approach requires sampling from the conditional distributions of the missing data given the observed data in the main sample and the maximum likelihood estimates. These conditional distributions are defined by Lemma 5.10.

5.5 Variance Estimation for the Quasi-likelihood Adjusted Estimates

In this section, we develop variance estimates for the parameters of the cross-sectional measurement error model when using a quasi-likelihood approach (see Section 2.2.1.4).

Result 5.6

Variance estimates for the parameters of the cross-sectional measurement error model when using the quasi-likelihood approach are derived using the expression below

$$\hat{Var}\left(\hat{\Theta}\right) \approx \left\{\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^{T}\left[\hat{Var}(\varepsilon)\right]^{-1}\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)\right\}^{-1}. \qquad (5.56)$$

Proof

Let $\hat{\Theta}$ denote the vector of quasi-likelihood estimates. The quasi-score estimating function is defined by

$$G(\Theta) = \left(\frac{\partial \mu(\Theta)}{\partial \Theta}\right)^{T}\left[Var(\varepsilon)\right]^{-1}\varepsilon. \qquad (5.57)$$

It follows that

$$Var\left[G\left(\hat{\Theta}\right)\right] = Var\left\{\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^{T}\left[Var(\varepsilon)\right]^{-1}\varepsilon\right\} = \left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^{T}\left[Var(\varepsilon)\right]^{-1} Var(\varepsilon) \qquad (5.58)$$

$$\left\{\left(\frac{\partial \mu(\Theta)}{\partial \Theta}\Big|_{\Theta=\hat{\Theta}}\right)^{T}\left[Var(\varepsilon)\right]^{-1}\right\}^{T}.$$

Taken into account that $[Var(\varepsilon)]^{-1}$ is symmetric, it follows that

$$Var\left[G\left(\hat{\Theta}\right)\right] = \left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1}\left[\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T . \qquad (5.59)$$

Now, $G\left(\hat{\Theta}\right)$ can be expanded as follows

$$G\left(\hat{\Theta}\right) \approx G(\Theta) + \left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1}\left[\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T \left(\hat{\Theta}-\Theta\right). \qquad (5.60)$$

Thus,

$$Var\left[G\left(\hat{\Theta}\right)\right] \approx Var\left\{\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1}\left[\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T \left(\hat{\Theta}-\Theta\right)\right\}. \qquad (5.61)$$

It follows that

$$Var\left[G\left(\hat{\Theta}\right)\right] \approx \left\{\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1}\left[\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T\right\}\left[Var\left(\hat{\Theta}\right)\right]$$
$$\left\{\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1}\left[\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T\right\}^T = \left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T [Var(\varepsilon)]^{-1}\left[\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T\right]^T$$
$$(5.62)$$

Solving equation (5.62) with respect to $Var\left(\hat{\Theta}\right)$ and replacing the unknown quantities by their estimates, we obtain the required result

$$\hat{Var}\left(\hat{\Theta}\right) \approx \left\{\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)^T \left[\hat{Var}(\varepsilon)\right]^{-1}\left(\frac{\partial\mu(\Theta)}{\partial\Theta}\Big|_{\Theta=\hat{\Theta}}\right)\right\}^{-1}.$$

□

The evaluation of the covariance matrix of the quasi-likelihood estimates is now straightforward since it requires the utilization of matrices that have been already used during the estimation process. Note also that $\hat{Var}(\varepsilon)$ is computed using the results from Section 2.2.1.4. Unlike in the case of the EM algorithm, variance estimation in a quasi-likelihood framework does not imply the use of computer intensive methods. This practical advantage offers an additional justification for preferring the quasi-likelihood approach, instead of the maximum likelihood approach, when analyzing cross-sectional misclassified data.

5.6 Applications

The methodology for variance estimation is illustrated in three applications. In the first application, we obtain variance estimates for the adjusted labour force gross flows estimated by the conventional point estimator. In the second application, we derive variance estimates for the likelihood-based adjusted labour force gross flows. The third application illustrates variance estimation for the parameters of the cross-sectional measurement error model estimated using either the EM algorithm (see Section 2.2.1.3) or the quasi-likelihood approach (see Section 2.2.1.4).

Application 5.1: Variance estimation for the Conventional (Moment-type) Estimator of the Adjusted Gross Flows

Variance estimation for the conventional point estimator is performed using the results from Section 5.2. We employ gross flows data from the UK LFS (summer –autumn 1997) and the smoothed version of the validation data from the Swedish (October 1994 – April 1995) LFS re-interview programme. The estimated observed labour force gross flows and the adjusted labour force gross flows, using the conventional estimator, are reported below. The matrix of misclassification probabilities we use is the same as the matrix used in application 3.1. The variance of the observed labour force gross flows is computed assuming a multinomial distribution for these flows.

Table 5.1: Variance estimation for the adjusted labour force gross flows derived from the conventional estimator, standard deviations in parenthesis

Flow	*Observed Labour Force Gross Flows*	*Adjusted Labour Force Gross Flows*
EE	0.716 (1.84E-03)	0.7420 (3.09E-03)
EU	0.009 (3.85E-04)	0.0028 (9.16E-04)
EN	0.016 (5.12E-04)	0.0024 (1.31E-03)
UE	0.016 (5.12E-04)	0.0102 (9.77E-04)
UU	0.027 (6.61E-04)	0.0292 (9.86E-04)
UN	0.009 (3.85E-04)	0.0033 (9.13E-04)
NE	0.016 (5.12E-04)	0.0026 (1.37E-03)
NU	0.010 (4.06E-04)	0.0045 (9.18E-04)
NN	0.181 (1.57E-03)	0.2030 (2.87E-03)

$n = 60000, n^v = 10000$

Application 5.2: Variance Estimation for the Maximum Likelihood Estimator of the Adjusted Gross Flows

In order to evaluate the variance of the likelihood-based adjusted estimates, we use the Missing Information Principle. We utilise gross flows data from the UK LFS (summer – autumn 1997) and the smoothed version of the validation data from the Swedish (October 1994 – April 1995) LFS re-interview programme. Due to the large number of computations involved, we derive variance estimates for the 2-state model i.e. Employed and Unemployed or Inactive. The observed labour force gross flows, the adjusted labour force gross flows, using the likelihood-based approach, and the matrix of misclassification probabilities are reported below. The variance of the observed labour force gross flows is computed assuming a multinomial distribution. The variance of the adjusted labour force gross flows is evaluated using the Missing Information Principle and the simulation approach. More specifically, we generated 20000 complete datasets using the conditional distributions of the missing data given the observed data and the maximum likelihood estimates in the main and in the validation sample. For each generated dataset, we computed the complete information matrix and the score functions. Subsequently, we evaluated the expectation of the complete information matrix and the variance of the score functions using (5.54) and (5.55) respectively. Finally, we employed the Missing Information Principle to determine the observed information matrix and the inverse of the observed information matrix to determine the covariance matrix of the adjusted likelihood-based estimates.

Misclassification Matrix

$$\begin{array}{c|cc} & E & U+N \\ E & 0.99 & 0.053 \\ U+N & 0.01 & 0.947 \end{array}.$$

Table 5.2: Variance estimation for the maximum likelihood estimates (4-state model), standard deviations in parenthesis

Flow	*Observed Labour Force Gross Flows*	*Adjusted Labour Force Gross Flows*
E,E	0.716 (1.84E-03)	0.730 (2.30E-03)
E, U+N	0.025 (6.37E-04)	0.006 (1.52E-03)
U+N, E	0.032 (7.18E-04)	0.014 (1.54E-03)
U+N, U+N	0.227 (1.71E-03)	0.250 (2.69E-03)

$n = 60000, n^v = 10000$

Application 5.3: Variance Estimation for the Maximum Likelihood and the Quasi-likelihood Cross-sectional Adjusted Estimates

In Chapter 2, we parameterised the cross-sectional measurement error model in a missing data framework and maximum likelihood estimates were derived via the EM algorithm. As an alternative approach, we further presented a quasi-likelihood parameterisation of the cross-sectional measurement error model. Variance estimation, under these two parameterisations, is illustrated using the data from application 2.2 in Section 2.2.1.4. For the maximum likelihood adjusted estimates, we employed the Missing Information Principle and the results from Section 5.4.2. The components of the Missing Information Principle are approximated by means of simulation. More specifically, we generated 10000 complete datasets using the conditional distributions of the missing data given the observed data and the maximum likelihood estimates in the main sample. For the quasi-likelihood approach, we utilised the results from Section 5.5.

Table 5.3: Variance estimation for the maximum likelihood and the quasi-likelihood cross-sectional adjusted estimates, standard deviations in parenthesis

Estimate	*MLE (EM Algorithm)*	*Quasi-likelihood*
$\hat{P}_1$	0.0667 (0.0021)	0.0669 (0.0022)

5.7 Summary

In this chapter, we developed variance estimators for some of the alternative point estimators of the adjusted gross flows. More specifically, we presented variance estimators for the conventional (moment-type) estimator under alternative double sampling schemes, for the modified estimator, for the composite estimator with fixed weights and for the maximum likelihood estimator under a validation sample that is selected independently from the main sample. Variance estimation for the maximum likelihood estimates when the validation sample is selected by sub-sampling units from the main survey (see Section 3.2.2), is more complex. The complexity arises due to the approach we follow for estimating the conditional expectations of the missing data in the validation sample. For the time being, we will rely on Monte-Carlo simulation for computing the variance of the maximum likelihood estimates under the specific double sampling scheme. We further developed variance estimators for the cross-sectional maximum likelihood estimates, using the parameterisation presented in

Section 2.2.1.3, and for the cross-sectional quasi-likelihood estimates. The quasi-likelihood parameterisation offers a practical advantage over the EM parameterisation by providing an easier way of performing variance estimation. The variance estimators account for the extra variability introduced by the adjustment for measurement error. The variance estimator of the maximum likelihood estimator accounts for the existence of missing data via the missing information matrix. Using the missing information matrix, we can now quantify the loss of information due to the missing data. In addition, the existence of a positive definite covariance matrix, obtained from the application of the Missing Information Principle, can be used as a diagnostic for checking whether the parameters of the measurement error model are identified. Having derived variance estimates, one can further examine the trade off between the increased variance of the adjusted estimates and the bias, due to measurement error, of the unadjusted estimates. It remains to evaluate the empirical properties of the different variance estimators. This is tackled in Chapter 6 using Monte-Carlo simulation.

Chapter 6

Monte-Carlo Evaluation

6.1 Introduction

In previous chapters we developed tools for point and interval estimation of gross flows statistics in the presence of misclassification and double sampling. However, it remains to assess the properties of these inference tools. In this chapter, we perform this assessment by designing a series of Monte-Carlo simulation experiments. In Section 6.2, we design a general simulation algorithm that can be employed with any type of flows data in the presence of misclassification and double sampling. As a special case, a simulation algorithm for cross-sectional inference is also presented. In Section 6.3, we describe a procedure that aims at relaxing the ICE assumption in the simulation. This is achieved by introducing dependence structure in the measurement error mechanism. In Section 6.4, we provide detailed information about the Monte-Carlo simulation studies and the data used in the context of the UK LFS. Sections 6.5 to 6.8 are devoted to reporting and commenting on the results.

6.2 Description of the Simulation Algorithm

Gross flows are estimated using information on the same individuals from at least two time points. This common sample consists of n sample units. From now on, we will refer to the sample that we use as the basis for our simulation as the original sample. We denote by H the total number of simulations that we perform and by (h) a specific simulation. Generally speaking, the simulation is performed in a reverse mechanics way. We start by generating true flows and then generate observed flows by introducing measurement error to these true flows.

For iteration (h)

Step 1: Generating True Flows

In this step, we generate true flows. This is done by employing the probability distribution function (defined explicitly in Section 6.4) of the true flows between two time points (e.g. months, quarters) say t and $t+1$. From this probability distribution function we draw a with replacement sample of size n. Recalling the notation from Chapter 2, $Y^*_{\xi t}, Y^*_{\xi t+1}$ are random variables that describe the observed status of the ξ^{th} unit at t and $t+1$ and $Y_{\xi t}, Y_{\xi t+1}$ are random variables that describe the true status of the ξ^{th} unit at t and $t+1$. Consequently, in this first step we generate values k, l such that $\left(Y_{\xi t} = k, Y_{\xi t+1} = l\right)$ for each sample unit ξ.

Step 2: Generating Cross-sectional Measurement Error

Having generated true flows, we now assume the existence of a cross-sectional measurement error model that is described by the misclassification probabilities q_{ik}. Using these misclassification probabilities, we generate the observed status at t given the true status at t for each sample unit ξ i.e. $\left(Y^*_{\xi t} = i \mid Y_{\xi t} = k\right)$.

Step 3: Generating Longitudinal Measurement Error

Having generated the observed status at t, we then 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 i.e. $\left(Y^*_{\xi t+1} = j \mid Y^*_{\xi t} = i, Y_{\xi t} = k, Y_{\xi t+1} = l\right)$. The theory we develop assumes the availability of the cross-sectional misclassification probabilities. Therefore, in order to generate the longitudinal measurement error we need to introduce additional assumptions. Initially, we generate longitudinal measurement error assuming that ICE is valid. However, we will also later investigate approaches that relax the ICE assumption.

To facilitate the description of the simulation algorithm, we present an example in the context of the labour force gross flows. The true flows, generated from Step 1, can be represented by the 1×9 vector shown in Table 6.1, where the cells represent the number of sample units that belong to the different labour force flows categories.

Table 6.1: Data generated after Step 1 of the simulation process

$Y_{\xi t}, Y_{\xi t+1}$	*EE*	*EU*	*EN*	*UE*	*UU*	$\cdots$	*NN*
	n_{EE}	n_{EU}	n_{EN}	n_{UE}	n_{UU}	$\cdots$	n_{NN}

At Step 2 we generate the observed status for sample unit ξ at t $Y^*_{\xi t}$, given his/her true status at t, using the cross-sectional misclassification probabilities q_{ik} (see Table 6.2). As a result, at this step we introduce cross-sectional measurement error. For example, a sample unit ξ that is truly employed both at t and $t+1$ is allocated to (EEE, UEE, NEE) according to q_{iE}.

Table 6.2: Data generated after Step 2 of the simulation process

$Y^*_{\xi t} \mid Y_{\xi t}, Y_{\xi t+1}$	*EE*	*EU*	*EN*	*UE*	*UU*	$\cdots$	*NN*
E	n_{EEE}	n_{EEU}	n_{EEN}	n_{EUE}	n_{EUU}	$\cdots$	n_{ENN}
U	n_{UEE}	n_{UEU}	n_{UEN}	n_{UUE}	n_{UUU}	$\cdots$	n_{UNN}
N	n_{NEE}	n_{NEU}	n_{NEN}	n_{NUE}	n_{NUU}	$\cdots$	n_{NNN}

At Step 3 we generate the observed status for sample unit ξ at $t+1$, $Y^*_{\xi t+1}$, given the information from Steps 1 and 2. For example, a sample unit ξ that is truly employed both at t and at $t+1$ and is also observed to be employed at t is allocated to (EEEE, EUEE and ENEE). This leads to the counts shown in Table 6.3.

Table 6.3: Data generated after Step 3 of the simulation process

$Y^*_{\xi t+1} \mid Y^*_{\xi t}, Y_{\xi t}, Y_{\xi t+1}$	*EEE*	*UEE*	*NEE*	*EEU*	*UEU*	$\cdots$	*NNN*
E	n_{EEEE}	n_{UEEE}	n_{NEEE}	n_{EEEU}	n_{UEEU}	$\cdots$	n_{NENN}
U	n_{EUEE}	n_{UUEE}	n_{NUEE}	n_{EUEU}	n_{UUEU}	$\cdots$	n_{NUNN}
N	n_{ENEE}	n_{UNEE}	n_{NNEE}	n_{ENEU}	n_{UNEU}	$\cdots$	n_{NNNN}

Therefore, in Steps 1-3 we generate $\left(Y^*_{\xi t} = i, Y^*_{\xi t+1} = j, Y_{\xi t} = k, Y_{\xi t+1} = l\right)$. The observed flows correspond to certain margins of Table 6.3. More specifically, the observed flow from state i at t to state j at $t+1$ can be extracted using the following summation

$$\sum_{k=E}^{N}\sum_{l=E}^{N}\sum_{\xi=1}^{n}\left(Y^*_{\xi t} = i, Y^*_{\xi t+1} = j, Y_{\xi t} = k, Y_{\xi t+1} = l\right). \qquad (6.1)$$

Step 4: Simulating an Internal or an External Second Phase Sample

In order to simulate the availability of validation information, derived from a smaller validation sample of n^v units $(n^v < n)$, we distinguish two cases:

(a) An internal validation sample is simulated by selecting a sub-sample of n^v units using the generated data of Table 6.3. The generated cross-sectional validation information can be extracted using the following summation

$$\sum_{j=E}^{N}\sum_{l=E}^{N}\sum_{\xi=1}^{n^v}\left(Y_{\xi t}^{*} = i, Y_{\xi t+1}^{*} = j, Y_{\xi t} = k, Y_{\xi t+1} = l\right). \qquad (6.2)$$

(b) An external validation sample of n^v units is simulated independently of the data generated in Step 3.

Hereinafter, when referring to an internal validation sample we will imply a validation sample that is generated using procedure 4a. An external validation sample will refer to a validation sample that is generated using procedure 4b. Note that throughout this chapter we assume that the independently selected validation sample (using 4b) is drawn from the same target population as the main sample. In that respect, this validation sample can be also regarded as internal.

Step 5: Estimation Step

Having generated observed (unadjusted) gross flows and cross-sectional validation information in Steps 1-4, we then utilise the generated data for estimation purposes i.e. for computing the alternative point and variance estimators.

Extending the Simulation Algorithm to Allow for Heterogeneity in the Gross Flows Mechanism and/or in the Measurement Error Mechanism

The simulation algorithm can be modified to allow for heterogeneity in the measurement error mechanism and/or in the gross flows mechanism. Generally speaking, this can be achieved by employing group-specific information to generate this heterogeneity. Two different scenarios are investigated: (a) we allow for heterogeneity both in the gross flows mechanism and in the measurement error mechanism and (b) we allow for heterogeneity in the measurement error mechanism while assuming a homogeneous gross flows mechanism. Scenario (a) is implemented by generating group-specific information at all stages of the simulation algorithm. This is equivalent to introducing stratification in the simulation. Scenario (b) requires the generation of group-specific data only for the measurement error mechanism.

A Simulation Algorithm for Cross-sectional Inference

The simulation algorithm can be modified in order to be suitable for cross-sectional inference. The cross-sectional algorithm is also performed in a reverse mechanics way. We start by generating cross-sectional true classifications. These true classifications are then contaminated with cross-sectional measurement error to produce cross-sectional observed classifications. Validation information is obtained by simulating an internal or an external second phase sample. Estimation is performed at the final step using the generated data. This algorithm will be used for comparing the maximum likelihood with the quasi-likelihood and the moment-based approach (see Chapter 2).

6.3 Relaxing the ICE Assumption by Introducing Dependence Structure in the Measurement Error Mechanism

The key assumption for estimating gross flows adjusted for measurement error, when only cross-sectional validation information is available, is the ICE. From our point of view, this is a rather strong assumption since we should expect some carry-over effects from the classification at the first time point. We believe that a scenario where a dependence structure in the errors exists is more realistic. One possibility for relaxing the ICE assumption arises when allowing for heterogeneity. However, this approach still assumes that ICE holds but now within the different sub-groups. An alternative proposal for relaxing the ICE assumption is given by Kristiansson (1983) and is described also in Hoem (1985). Under ICE, the following holds for each sample unit ξ

$$q_{ijkl} = pr\left(Y^*_{\xi t} = i, Y^*_{\xi t+1} = j \mid Y_{\xi t} = k, Y_{\xi t+1} = l\right) = q_{ik} q_{jl} \quad i,j,k,l = 1,2,\cdots r. (6.3)$$

Kristiansson (1983) proposed to replace (6.3) with an expression of the following form

$$\begin{cases} q_{ijkl} = q_{ik} q_{jl} \;\; k \neq l \\ q_{ijkk} = q_{ik} \Phi_{ijk} \end{cases} \quad where\ \Phi_{iii} > q_{ii} > \Phi_{ikk}. \tag{6.4}$$

The idea is that (a) a change in the real status should make the classifications recorded at two time points independent and (b) the classifications recorded at two time points should be conditionally dependent, given the true states, when no change in the true status has occurred. Kristiansson's proposal seems reasonable and can be justified using the memory effect. For example, when an individual's true labour force status remains stable between two time points we can assume that there is a memory effect, which is stronger compared to the case where the true labour force status changes between the two time points. For the latter case Kristiansson assumes that the ICE assumption is valid. However, for the former case he imposes a dependence structure defined by (6.4). Think of the following three examples. Assume that an individual is truly employed at both time points. According to (6.4) the following holds

$$q_{EEEE} = pr\left(Y^*_{\xi t} = E, Y^*_{\xi t+1} = E \mid Y_{\xi t} = E, Y_{\xi t+1} = E\right) = q_{EE}\Phi_{EEE} \text{ where } \Phi_{EEE} > q_{EE}.$$

This means that the probability of correct classification at the second time point, given that the individual is correctly classified at the first time point and the true labour force is stable between the two time points, is reinforced compared to the probability of correct classification at the second time point predicted under ICE. Assume now that an individual who is truly employed at both time points is correctly classified at the first time point and misclassified as unemployed at the second time point. According to (6.4)

$$q_{EUEE} = pr\left(Y^*_{\xi t} = E, Y^*_{\xi t+1} = U \mid Y_{\xi t} = E, Y_{\xi t+1} = E\right) = q_{EE}\Phi_{EUE} \quad \text{where } \Phi_{EUE} < q_{UE}.$$

This means that the probability of misclassification at the second time point of an individual whose true labour force status is stable at both time points and who has been correctly classified at the first time point is lower than the probability of misclassification predicted under ICE. Assume finally that an individual who is truly employed at both time points is correctly classified at the second time point and misclassified as unemployed at the first time point. According to (6.4)

$$q_{UEEE} = pr\left(Y^*_{\xi t} = U, Y^*_{\xi t+1} = E \mid Y_{\xi t} = E, Y_{\xi t+1} = E\right) = q_{UE}\Phi_{UEE} \quad \text{where } q_{EE} > \Phi_{UEE} .$$

This means that the probability of correct classification at the second time point of an individual who is truly employed at both time points and is misclassified at the first time point is lower than the probability of correct classification at the second time point predicted under ICE.

Based on Kristiansson's idea, we investigate the robustness of the alternative estimators under ICE and under departures from ICE. With respect to the choice of alternative to the ICE error models, we investigate two scenarios. These scenarios are defined by modifying the probabilities of correct classification and misclassification, predicted under ICE, for individuals that remain truly stable between t and $t+1$ while preserving the probabilities of correct classification and misclassification, predicted under ICE, for individuals who truly change their status between t and $t+1$. These modified probabilities are then used to generate the data in the simulation.

6.4 Describing the Simulation Studies and the Data Sources in the Context of the UK LFS

The methodology we develop in this thesis is targeted at flows data obtained from the UK LFS. The UK LFS is a quarterly panel survey and labour force gross flows are estimated using information on the same sampled individuals at two successive quarters. This common sample consists of approximately 60000 individuals. We now describe a series of Monte-Carlo simulation studies based on these data. Table 6.4 summarises the information about the simulation studies we conducted. Tables 6.5 and 6.6 summarise the notation for the different point and variance estimators that are included in the simulation studies.

As we described in Section 6.2, after the first three steps of the simulation algorithm we can compute the generated observed gross flows. The UK LFS is used implicitly by ensuring that the generated observed labour force gross flows are close to the un-weighted UK labour force gross flows defined by the common LFS sample between summer-autumn 1997. This is achieved as follows. Utilising the unadjusted UK labour force gross flows between summer-autumn 1997 and a set of misclassification probabilities, we estimate UK labour force gross flows (summer-autumn 1997) adjusted for misclassification using one of the alternative estimators. The probability distribution function defined by these estimated adjusted labour force gross flows is then used to generate the true flows at Step 1 of the simulation. For simulation studies I-V (Tables 6.7-6.31 and 6.42-6.51), VIII (Table 6.52) and X (Tables 6.54-6.61), the probability distribution function that we use to generate true flows is estimated using the conventional (moment-type) estimator. For simulation study VI (Tables 6.32-6.36), the probability distribution function is estimated using the moment-type unit heterogeneity estimator. For simulation study VII (Tables 6.37-6.41), the probability distribution function is estimated using the post-stratified version of the conventional estimator. Finally, for simulation study IX (Table 6.53) the probability distribution function is estimated using the maximum likelihood estimator.

In Step 2, the algorithm requires the specification of a set of misclassification probabilities that will be used for inflating the generated true flows with cross-sectional measurement error. For simulation studies I-III and VIII, we define the misclassification probabilities by modifying slightly the unweighted Swedish (October 1994 – April 1995) misclassification probabilities. This modification was performed in order to avoid, under ICE, problems with negative adjusted flows. For simulation study X, we use the misclassification probabilities as these appear in application 5.2. For simulation studies IV-V, we further modify the Swedish (October 1994 – April 1995) misclassification probabilities. This modification is performed in order to avoid boundary values when fitting the EM algorithm. For simulation studies VI-VII, the group-specific Swedish (October 1994 – April 1995) misclassification probabilities are also modified for the same reason. Note that the misclassification probabilities, which we use in Step 2, are the same as the misclassification probabilities we use to estimate the probability distribution function of the true flows that is utilised in Step 1.

In Step 3, we need to generate longitudinal measurement error based on cross-sectional measurement error. The simulations are conducted both under ICE and under relaxed-ICE scenarios. The relaxed-ICE scenarios are defined using (6.4). The cross-sectional misclassification matrices utilised in Step 2 along with the misclassification probabilities either under ICE or under a relaxed–ICE scenario are reported in Appendix IV.

In Step 4, we simulate the availability of validation information. More specifically, we simulate (a) a validation sample that is selected by sub-sampling units that already participate in the main survey and (b) an independently selected from the main sample validation sample that refers to the same target population.

In simulation study X, we compare the maximum likelihood estimator with the conventional point estimator when the validation sample is selected by sub-sampling units from the main survey. For simplicity, we compare these two estimators for the 4-state model (i.e. Employed and Unemployed or Inactive). In Section 3.2.2, we also described a "naïve" approach for estimating the conditional expectations of the missing data in the validation sample when the validation sample is selected by sub-sampling units from the main sample. This "naïve" approach attempts to simplify the E-step of the EM algorithm (see Section 3.2.2). The "naïve" approach is compared with the approach that utilises full information also in simulation study X.

In Section 6.2, we presented an extension to the simulation algorithm that allows for heterogeneity. Two scenarios for heterogeneity are investigated. Under the first scenario, we allow for heterogeneity both in the measurement error and in the gross flows mechanism. Under the second scenario, we allow for heterogeneity in the measurement error mechanism while assuming a homogeneous gross flows mechanism. Here, we assume the existence of moderate heterogeneity only according to gender. However, the algorithm can be easily extended to accommodate heterogeneity according to more variables. Note also that the simulation studies that allow for heterogeneity are designed to preserve the group-specific labour force gross flows patterns of the original sample. The matrices of the misclassification probabilities (i.e. for males and for females) used in simulation studies VI-VII are reported in Appendix IV.

In simulation study XI, (Tables 6.62-6.63) we contrast the alternative point estimators used for cross-sectional inference. The estimators we consider are the following: (a) the moment-type estimator, (b) the maximum likelihood estimator with calibration probabilities (Tenenbein 1972), (c) the maximum likelihood estimator with misclassification probabilities (using the EM algorithm) and (d) the quasi-likelihood estimator. For the purposes of this simulation study, we use an artificial dataset. The set of probabilities we used to generate the data is the following: $P_1 = 0.606$, $q_{11} = 0.98$, $q_{12} = 0.04$.

Details of each simulation study i.e. the sample size of the main survey, the sample size of the validation survey, the number of iterations and the type of the second phase sample are reported along with the results from the specific simulation study.

Table 6.4: Information about the alternative simulation studies

Simulation Study	*Description*
I	Comparing alternative moment-type estimators and variance estimators under ICE
II	Comparing alternative moment-type estimators under relaxed-ICE 1
III	Comparing alternative moment-type estimators and variance estimators under relaxed-ICE 2
IV	Comparing alternative moment-type estimators with the maximum likelihood estimators under ICE
V	Comparing alternative moment-type estimators with the maximum likelihood estimators under a relaxed-ICE scenario
VI	Comparing alternative point estimators when allowing for heterogeneity only in the measurement error mechanism
VII	Comparing alternative point estimators when allowing for heterogeneity in the gross flows and in the measurement error mechanism
VIII	Evaluating the performance of the variance estimator of the conventional estimator in the case of an internal validation sample
IX	Evaluating the performance of the variance estimator of the maximum likelihood estimator under an external validation sample
X	Comparing the moment-type estimator with the maximum likelihood estimator under ICE and an internal validation sample
XI	Comparing the maximum likelihood with the quasi-likelihood and the moment-based approach in a cross-sectional framework

Table 6.5: Notation for the point estimators appearing in the simulation studies

Point Estimator	*Notation*
Observed flows (Section 1.7.3)	P-OBS
Conventional (Moment-type) (Section 2.2.2.1)	P-ST
Modified (Section 2.4.2)	P-MOD
Unbiased margins (Section 2.4.1)	P-UM
Composite with fixed weights $w_{comp} = 0.3$ (Section 2.4.3 – 1st set of weights)	P-CF1
Composite with fixed weights $w_{comp} = 0.2$ (Section 2.4.3 - 2nd set of weights)	P-CF2
Composite with fixed weights $w_{comp} = 0.1$ (Section 2.4.3 – 3rd set of weights)	P-CF3
Composite with adaptive weights (Section 2.4.3)	P-CAD
Maximum likelihood (Section 3.2)	P-MLE
Constrained maximum likelihood (Section 3.3)	P-UMLE
Moment-type unit heterogeneity (Section 4.2)	P-UNIT
Maximum likelihood that allows for heterogeneity (Section 4.3)	P-UNMLE
Moment-type (for cross-sectional inference) (Section 2.2.1.1)	Moment-type
Quasi-likelihood (for cross-sectional inference) (Section 2.2.1.4)	Quasi-likelihood
Maximum likelihood with calibration probabilities-Tenenbein (1972) (for cross-sectional inference) (Section 2.2.1.2)	MLE (Tenenbein 1972)
Maximum likelihood with misclassification probabilities-EM algorithm (for cross-sectional inference) (Section 2.2.1.3)	MLE (EM algorithm)

Table 6.6: Notation for the variance estimators appearing in the simulation studies

Variance Estimator	*Notation*
Variance estimator of the conventional (Moment-type) estimator under an external double sampling scheme (Section 5.2)	$\hat{Var}\left(\hat{P}^{st-ext}\right)$
Variance estimator of the conventional (Moment-type) estimator under an internal double sampling scheme (Section 5.2)	$\hat{Var}\left(\hat{P}^{st-int}\right)$
Variance estimator of the modified estimator (Section 5.3)	$\hat{Var}\left(\hat{P}^{mod}\right)$
Variance estimator of the composite estimator with fixed weights (Section 5.3)	$\hat{Var}\left(\hat{P}^{comp-fx}\right)$
Variance estimator of the maximum likelihood estimator under an external double sampling scheme (Section 5.4)	$\hat{Var}\left(\hat{P}^{mle}\right)$

6.5 Evaluating the Performance of the Alternative Point and Variance Estimators

The performance of the different point and variance estimators is assessed using the following evaluation criteria:

1. Relative bias of a point estimator.
2. Standard deviation of a point estimator.
3. Root Mean Squared Error (RMSE) of a point estimator.
4. Relative bias of a variance estimator.
5. Coverage rate.

Bias and Relative Bias of a Point Estimator $\hat{P}$

$$Bias\left(\hat{P}\right) = E\left(\hat{P}\right) - P\,,\ E\left(\hat{P}\right) = \frac{1}{H}\sum_{h=1}^{H}\hat{P}^{(h)}, \tag{6.5}$$

where $\hat{P}^{(h)}$ denotes the point estimator in simulation (h)

$$RB\left(\hat{P}\right) = \frac{E\left(\hat{P}\right) - P}{P} \times 100. \tag{6.6}$$

Simulation (Empirical) Variance of a Point Estimator

$$V\left(\hat{P}\right) = \frac{1}{H-1}\sum_{h=1}^{H}\left[\hat{P}^{(h)} - E\left(\hat{P}\right)\right]^2. \tag{6.7}$$

The standard deviation is derived by taking the square root of (6.7)

Root Mean Squared Error of a Point Estimator $\hat{P}$

$$RMSE\left(\hat{P}\right) = \sqrt{V\left(\hat{P}\right) + \left[Bias\left(\hat{P}\right)\right]^2}. \tag{6.8}$$

Bias and Relative Bias of a Variance Estimator $\hat{Var}\left(\hat{P}\right)$

$$Bias\left[\hat{Var}\left(\hat{P}\right)\right]=E\left[\hat{Var}\left(\hat{P}\right)\right]-V\left(\hat{P}\right),\quad E\left[\hat{Var}\left(\hat{P}\right)\right]=\frac{1}{H}\sum_{h=1}^{H}\hat{Var}\left(\hat{P}^{(h)}\right). \tag{6.9}$$

$$RB\left[\hat{Var}\left(\hat{P}\right)\right]=\frac{E\left[\hat{Var}\left(\hat{P}\right)\right]-V\left(\hat{P}\right)}{V\left(\hat{P}\right)}\times 100. \tag{6.10}$$

Coverage Rate

For each replication (h) we calculate the 95% confidence interval for each estimator $\hat{P}^{(h)}$ given by

$$CI^{(h)}=\hat{P}^{(h)}\pm 1.96\sqrt{\hat{Var}\left(\hat{P}^{(h)}\right)}. \tag{6.11}$$

The coverage rate is defined as the total number of times that $CI^{(h)}$ contains the true value P divided by the total number of simulations H. Ideally, the coverage rate given by (6.11), should be close to 95%.

6.6 Results

In this section, we report the results from the different simulation studies.

Simulation Study I: Non–differential flows – Non–differential misclassification

ICE True Scenario – External Validation Sample
$n^v = 2150,\ n = 60000, H = 20000$

Table 6.7: True flows

EE	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
0.7459	0.0102	0.0004	0.0015	0.0295	0.005	0.0008	0.0027	0.204

Table 6.8: Point estimates, Averages over simulations

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	0.7180	0.0160	0.0160	0.0080	0.0270	0.010	0.0160	0.0080	0.1810
P-ST	0.7460	0.0102	0.00038	0.00149	0.0295	0.0049	0.00072	0.00260	0.2042
P-MOD	0.7395	0.0109	0.00577	0.00250	0.0286	0.0051	0.00610	0.00285	0.1986
P-UM	0.7361	0.0139	0.00002	0.00526	0.0328	0.0069	0.00048	0.00410	0.2005
P-CF1	0.7441	0.0104	0.00198	0.00178	0.0292	0.0050	0.00233	0.00270	0.2025
P-CF2	0.7447	0.0103	0.00144	0.00170	0.0293	0.0050	0.00179	0.00267	0.2031
P-CF3	0.7454	0.0102	0.00089	0.00157	0.0294	0.0050	0.00125	0.00266	0.2036
P-CAD	0.7458	0.0102	0.00041	0.00157	0.0294	0.0053	0.00076	0.00300	0.2036

Table 6.9: Relative bias of point estimators (%)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	-3.74	57.9	3900	433.3	-8.47	100	1900	196	-11.3
P-ST	0.01	0.01	-5	-0.66	0.01	-2	-10	-3.70	0.10
P-MOD	-0.86	6.86	1343	66.6	-3.05	2	662	5.55	-2.64
P-UM	-1.31	36.2	-95	251	11.2	38	-40	51.9	-1.71
P-CF1	-0.24	1.96	395	18.7	-1.02	0.01	191	0.01	-0.73
P-CF2	-0.16	0.98	260	13.3	-0.68	0.01	124	-1.11	-0.44
P-CF3	-0.07	0.01	123	4.66	-0.34	0.01	56.2	-1.48	-0.19
P-CAD	-0.01	0.01	2.50	4.66	-0.34	6	-5	11.1	-0.19

Table 6.10: Standard deviation of point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	1.83	0.51	0.51	0.36	0.66	0.41	0.51	0.51	0.54
P-ST	5.72	1.86	2.91	1.84	1.66	1.68	2.90	1.70	1.67
P-MOD	5.40	1.78	2.67	1.76	1.58	1.63	2.77	1.64	1.62
P-UM	3.13	1.24	2.86	1.29	1.93	1.42	2.80	1.62	1.04
P-CF1	5.64	1.84	2.86	1.83	1.63	1.67	2.86	1.69	1.65
P-CF2	5.67	1.85	2.90	1.83	1.63	1.67	2.88	1.70	1.66
P-CF3	5.70	1.85	2.92	1.84	1.64	1.68	2.90	1.71	1.67
P-CAD	5.65	1.82	2.88	1.81	1.65	1.49	2.87	1.61	1.66

Table 6.11: RMSE of point estimators (*10^5)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	8.86	1.85	4.93	2.07	0.80	1.58	4.82	1.69	7.31
P-ST	1.81	0.59	0.92	0.58	0.52	0.53	0.92	0.54	1.67
P-MOD	2.66	0.61	1.90	0.65	0.57	0.52	1.91	0.52	2.37
P-UM	3.27	1.24	0.92	1.27	1.23	0.73	0.89	0.68	1.53
P-CF1	1.88	0.59	1.04	0.58	1.65	0.53	1.03	0.54	1.73
P-CF2	1.84	0.58	0.97	0.58	0.52	0.53	0.97	0.54	1.70
P-CF3	1.81	0.59	0.89	0.58	0.52	0.53	0.93	0.54	1.68
P-CAD	1.79	0.57	0.91	0.57	0.52	0.49	0.91	0.52	1.67

Simulation Study II: Non–differential flows – Non–differential misclassification

Relaxed ICE Scenario 1– External Validation Sample

$n^v = 2150, n = 60000, H = 20000$

Table 6.12: True flows

EE	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
0.744	0.0103	0.0012	0.0018	0.0294	0.006	0.0015	0.003	0.2028

Table 6.13: Point estimates, Averages over simulations

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	0.7178	0.0161	0.01610	0.00760	0.0271	0.0103	0.01560	0.00810	0.1809
P-ST	0.7458	0.0102	0.00049	0.00108	0.0297	0.0054	0.00033	0.00280	0.2042
P-MOD	0.7394	0.0110	0.00590	0.00214	0.0287	0.0055	0.00570	0.0030	0.1986
P-UM	0.7358	0.0133	0.00064	0.00480	0.0328	0.0081	0.00108	0.00450	0.1989
P-CF1	0.7439	0.0104	0.00211	0.00140	0.0293	0.0054	0.00195	0.00286	0.2024
P-CF2	0.7445	0.0104	0.00157	0.00130	0.0294	0.0054	0.00141	0.00283	0.2031
P-CF3	0.7452	0.0103	0.00100	0.00120	0.0295	0.0054	0.00087	0.00281	0.2036
P-CAD	0.7457	0.0104	0.00052	0.00120	0.0295	0.0058	0.00036	0.00310	0.2039

Table 6.14: Relative bias of point estimators (%)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	-3.52	56.3	1241	322	-7.82	71.6	940	170	-10.8
P-ST	0.24	-0.97	-59.2	-40	1.02	-10	-78	-6.66	0.69
P-MOD	-0.62	6.79	392	18.9	-2.38	-8.33	280	0.01	-2.07
P-UM	-1.10	29.1	-46.6	167	11.6	35	-28	50	-1.92
P-CF1	-0.01	0.97	75.8	-22.2	-0.34	-10	30	-4.67	-0.19
P-CF2	0.07	0.97	30.8	-27.7	0.01	-10	-6	-5.66	0.15
P-CF3	0.16	0.01	-16.6	-33.3	0.34	-10	-42	-6.33	0.39
P-CAD	0.23	0.97	-56.6	-33.3	0.34	-3.33	-76	3.33	0.54

Table 6.15: Standard deviation of point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	1.84	0.51	0.51	0.35	0.66	0.41	0.51	0.36	1.50
P-ST	5.72	1.85	2.92	1.84	1.64	1.67	2.91	1.70	5.24
P-MOD	5.54	1.79	2.82	1.77	1.58	1.62	2.81	1.64	5.07
P-UM	2.93	1.26	2.83	1.34	1.93	1.36	2.79	1.60	3.11
P-CF1	5.67	1.84	2.89	1.82	1.62	1.66	2.88	1.68	5.20
P-CF2	5.68	1.84	2.90	1.82	1.63	1.66	2.88	1.69	5.22
P-CF3	5.70	1.85	2.91	1.83	1.63	1.67	2.89	1.69	5.22
P-CAD	5.69	1.83	2.91	1.81	1.60	1.56	2.89	1.65	5.22

Table 6.16: RMSE of point estimators (*10^5)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	8.30	1.84	4.71	1.84	0.75	1.36	4.46	1.62	6.92
P-ST	1.88	0.58	0.95	0.62	0.52	0.57	0.99	0.54	1.71
P-MOD	2.27	0.61	1.73	0.57	0.55	0.53	1.61	0.52	2.09
P-UM	2.78	1.01	0.91	1.04	1.24	0.79	0.89	0.69	1.58
P-CF1	1.79	0.58	0.95	0.58	0.51	0.55	0.92	0.53	1.65
P-CF2	1.81	0.58	0.92	0.60	0.51	0.56	0.91	0.54	1.65
P-CF3	1.85	0.58	0.92	0.61	0.52	0.56	0.94	0.54	1.67
P-CAD	1.88	0.57	0.94	0.60	0.51	0.50	0.98	0.52	1.69

Simulation Study III: Non–differential flows – Non–differential misclassification

Relaxed ICE Scenario 2 – External Validation Sample

$n^v = 2150$, $n = 60000, H = 20000$

Table 6.17: True flows

EE	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
0.7425	0.0103	0.0018	0.0027	0.0291	0.0063	0.0027	0.0031	0.2015

Table 6.18: Point estimates, Averages over simulations

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	0.7179	0.0161	0.01590	0.0080	0.0270	0.0102	0.01560	0.00810	0.1810
P-ST	0.7460	0.0102	0.00025	0.0015	0.0296	0.0052	0.00037	0.00270	0.2042
P-MOD	0.7395	0.0109	0.00570	0.0025	0.0286	0.0053	0.00570	0.00300	0.1986
P-UM	0.7358	0.0132	0.00067	0.0048	0.0328	0.0081	0.00111	0.00450	0.1989
P-CF1	0.7440	0.0104	0.00188	0.0018	0.0292	0.0052	0.00199	0.00280	0.2024
P-CF2	0.7447	0.0104	0.00133	0.0017	0.0294	0.0052	0.00145	0.00277	0.2031
P-CF3	0.7453	0.0103	0.00080	0.0016	0.0295	0.0052	0.00091	0.00274	0.2036
P-CAD	0.7458	0.0104	0.00029	0.0016	0.0294	0.0056	0.00041	0.00310	0.2040

Table 6.19: Relative bias of point estimators (%)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	3.31	56.3	783	196	-7.2	61.9	478	161	-10.2
P-ST	0.47	-0.97	-86	-44	1.72	-17.4	-86	-12.9	1.34
P-MOD	-0.40	5.82	217	-7.40	-1.72	-15.8	111	-3.22	-1.44
P-UM	-0.90	28.1	-62.7	77.7	12.7	28.5	-58.8	45.1	-1.29
P-CF1	0.20	0.97	4.44	-33	0.34	-17	-26.3	-9.67	0.44
P-CF2	0.29	0.97	-26.1	-37	1.03	-17.5	-46.3	-10.6	0.79
P-CF3	0.37	0.01	-55.5	-40.7	1.37	-17.4	-66.3	-11.6	1.04
P-CAD	0.44	0.97	-83.9	-40.7	1.03	-11.1	-84.8	0.01	1.24

Table 6.20: Standard deviation of point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	1.84	0.51	0.51	0.36	0.66	0.41	0.51	0.36	1.72
P-ST	5.68	1.85	2.95	1.83	1.66	1.71	2.93	1.73	5.29
P-MOD	5.51	1.79	2.86	1.77	1.60	1.65	2.84	1.68	5.11
P-UM	3.14	1.26	2.82	1.35	1.82	1.39	2.77	1.59	3.15
P-CF1	5.63	1.83	2.92	1.81	1.64	1.69	2.90	1.72	5.23
P-CF2	5.65	1.84	2.93	1.82	1.64	1.70	2.91	1.72	5.25
P-CF3	5.67	1.85	2.94	1.82	1.65	1.71	2.92	1.73	5.27
P-CAD	5.67	1.83	2.94	1.81	1.62	1.60	2.92	1.68	5.26

Table 6.21: RMSE of point estimators (*10^5)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	7.80	1.84	4.47	1.68	0.69	1.24	4.09	1.58	6.50
P-ST	2.12	0.58	1.05	0.69	0.54	0.64	1.18	0.56	1.87
P-MOD	1.97	0.60	1.52	0.57	0.53	0.61	1.33	0.53	1.85
P-UM	2.37	1.00	0.96	0.79	1.30	0.72	1.01	0.68	1.30
P-CF1	1.85	0.58	0.93	0.64	0.52	0.64	0.94	0.55	1.69
P-CF2	1.92	0.58	0.94	0.66	0.53	0.64	1.00	0.56	1.73
P-CF3	2.01	0.58	0.98	0.68	0.53	0.65	1.08	0.56	1.80
P-CAD	2.08	0.57	1.04	0.67	0.52	0.56	1.17	0.53	1.84

Simulation Study IV: Non–differential flows – Non–differential misclassification

ICE True Scenario – External Validation Sample

$n^v = 10000$, $n = 60000, H = 100$

Table 6.22: True flows

EE	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
0.7316	0.0131	0.0091	0.0047	0.0283	0.0071	0.0093	0.0049	0.1919

Table 6.23: Point estimates, Averages over simulations

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	0.7177	0.0160	0.0161	0.0080	0.0267	0.010	0.0161	0.0081	0.1813
P-ST	0.7318	0.0131	0.0090	0.0047	0.0281	0.0070	0.0093	0.0048	0.1922
P-UM	0.7262	0.0148	0.0087	0.0064	0.0301	0.0083	0.0092	0.0059	0.1904
P-CF2	0.7268	0.0136	0.0132	0.0055	0.0274	0.0071	0.0134	0.0050	0.1880
P-CF3	0.7293	0.0133	0.0111	0.0051	0.0278	0.0071	0.0113	0.0049	0.1901
P-CAD	0.7318	0.0130	0.0090	0.0047	0.0281	0.0072	0.0092	0.0049	0.1921
P-MLE	0.7320	0.0130	0.0089	0.0046	0.0281	0.0071	0.0091	0.0049	0.1923
P-UMLE	0.7276	0.0140	0.0088	0.0058	0.0295	0.0083	0.0093	0.0059	0.1908

Table 6.24: Relative bias of point estimators (%)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	-1.89	22.1	76.9	70.2	-5.65	40.8	73.1	65.3	-5.52
P-ST	0.03	0.01	-1.09	0.01	-0.71	-1.41	0.01	-2.42	0.16
P-UM	-0.74	12.9	-4.39	36.2	6.36	16.9	-1.07	20.4	-0.78
P-CF2	-0.66	3.81	45	17	-3.18	0.01	44.1	2.04	-2.03
P-CF3	-0.31	1.53	21.9	8.51	-1.76	0.01	21.5	0.01	-0.94
P-CAD	0.03	-0.76	-1.09	0.01	-0.71	1.41	-1.07	0.01	0.10
P-MLE	0.05	-0.76	-2.19	2.12	-0.71	0.01	-2.15	0.01	0.21
P-UMLE	-0.55	6.87	-3.29	23.4	4.24	16.9	0.01	20.4	-0.57

Table 6.25: Standard deviation of point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	1.92	0.48	0.52	0.33	0.80	0.41	0.47	0.35	1.55
P-ST	2.41	0.73	1.09	0.62	1.01	0.71	1.05	0.69	2.17
P-UM	2.09	0.66	1.04	0.54	1.10	0.60	0.98	0.61	1.96
P-CF2	2.36	0.71	1.07	0.62	0.98	0.69	0.33	0.66	2.11
P-CF3	2.38	0.72	0.35	0.62	0.99	0.70	0.33	0.68	2.14
P-CAD	2.40	0.73	1.09	0.62	1.01	0.72	0.33	0.69	2.17
P-MLE	2.03	0.60	0.82	0.48	0.90	0.60	0.78	0.54	1.82
P-UMLE	1.67	0.47	0.44	0.30	0.77	0.42	0.40	0.33	1.45

Table 6.26: RMSE of point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	14.1	2.95	7.00	3.33	1.71	3.07	6.84	3.24	3.45
P-ST	2.42	0.73	1.09	0.62	1.01	0.71	1.05	0.69	2.17
P-UM	5.78	1.85	1.10	1.79	2.18	1.39	0.98	1.16	2.60
P-CF2	5.30	0.41	4.24	1.04	1.30	0.69	4.27	0.67	4.63
P-CF3	3.27	0.76	2.29	0.75	1.11	0.70	2.31	0.68	2.92
P-CAD	2.41	0.73	1.09	0.62	1.02	0.72	1.05	0.69	2.17
P-MLE	2.12	0.61	0.85	0.49	0.91	0.61	0.80	0.55	1.82
P-UMLE	4.29	1.04	0.48	1.16	1.41	1.25	0.40	0.99	1.86

Simulation Study V: Non–differential flows – Non–differential misclassification

ICE Relaxed Scenario – External Validation Sample

$n^v = 10000$, $n = 60000, H = 100$

Table 6.27: True flows

EE	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
0.7283	0.0131	0.01	0.0059	0.028	0.008	0.0117	0.0052	0.1898

Table 6.28: Point estimates, Averages over simulations

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	0.7180	0.0160	0.0160	0.0081	0.0270	0.010	0.0160	0.0080	0.1809
P-ST	0.7320	0.0133	0.0088	0.0049	0.0284	0.0070	0.0090	0.0049	0.1917
P-UM	0.7266	0.0149	0.0086	0.0065	0.0303	0.0082	0.0090	0.0058	0.1901
P-CF2	0.7269	0.0139	0.0130	0.0057	0.0276	0.0071	0.0132	0.0050	0.1876
P-CF3	0.7294	0.0136	0.0110	0.0053	0.0280	0.0071	0.0111	0.0049	0.1896
P-CAD	0.7319	0.0133	0.0088	0.0049	0.0284	0.0072	0.0090	0.0050	0.1915
P-MLE	0.7320	0.0131	0.0088	0.0048	0.0284	0.0071	0.0090	0.0049	0.1919
P-UMLE	0.7280	0.0140	0.0089	0.0057	0.030	0.0081	0.0091	0.0058	0.1904

Table 6.29: Relative bias of point estimators (%)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	-1.41	22.1	60	37.3	-3.57	25	36.7	53.8	-4.69
P-ST	0.51	1.53	-12	-16.9	1.43	-12.5	-23.1	-5.77	1
P-UM	-0.23	13.7	-14	10.2	8.21	2.5	-23.1	11.5	0.16
P-CF2	-0.19	6.10	30	-3.39	-1.43	-11.2	12.8	-3.84	-1.16
P-CF3	0.15	3.81	10	-10.2	0.01	-11.2	-5.12	-5.77	-0.11
P-CAD	0.49	1.53	-12	-16.9	1.42	-10	-23.1	-3.84	0.89
P-MLE	0.51	0.01	-12	-18.6	1.42	-11.2	-23.1	-5.77	1.10
P-UMLE	-0.04	6.87	-11	-3.39	7.14	1.25	-22.2	11.5	0.32

Table 6.30: Standard deviation of point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	1.77	0.50	0.52	0.37	0.60	0.42	0.49	0.39	1.66
P-ST	2.65	0.74	1.14	0.71	0.82	0.62	1.00	0.64	2.28
P-UM	2.02	0.66	1.10	0.61	0.89	0.58	0.97	0.59	2.08
P-CF2	2.60	0.73	1.11	0.69	0.80	0.60	0.99	0.62	2.22
P-CF3	2.62	0.73	1.13	0.70	0.81	0.61	0.99	0.62	2.25
P-CAD	2.65	0.74	1.14	0.71	0.82	0.62	1.00	0.65	2.28
P-MLE	2.27	0.68	0.84	0.59	0.77	0.53	0.73	0.49	2.02
P-UMLE	1.68	0.58	0.41	0.33	0.58	0.41	0.40	0.35	1.65

Table 6.31: RMSE of point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	10.63	2.99	6.11	2.31	1.30	2.09	4.42	2.87	9.19
P-ST	4.55	0.77	1.66	1.23	0.92	1.17	2.88	0.71	2.97
P-UM	2.64	1.91	1.78	0.85	2.47	0.61	2.87	0.84	2.10
P-CF2	2.95	1.08	3.19	0.72	0.89	1.08	1.80	0.65	3.13
P-CF3	2.84	0.88	1.51	0.92	0.81	1.09	1.16	0.69	2.26
P-CAD	4.46	0.77	1.66	1.22	0.92	1.01	2.89	0.67	2.84
P-MLE	4.34	0.68	1.47	1.24	0.87	1.04	2.80	0.57	2.92
P-UMLE	1.71	1.07	1.17	0.39	2.08	0.42	2.63	0.69	1.75

Simulation Study VI: Differential measurement error – Non–differential flows

External validation sample, Heterogeneity according to gender

$n^v = 10000$, $n = 60000, H = 100$

Table 6.32: True flows

EE	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
0.7353	0.0128	0.0080	0.0056	0.0280	0.0081	0.0077	0.0070	0.1875

Table 6.33: Point estimates, Averages over simulations

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	0.7154	0.0160	0.0167	0.0091	0.0268	0.0101	0.0163	0.0091	0.1805
P-ST	0.7356	0.0127	0.0080	0.0056	0.0280	0.0081	0.0077	0.0070	0.1873
P-UM	0.7243	0.0146	0.0092	0.0075	0.0296	0.0090	0.0090	0.0077	0.1891
P-UNIT	0.7355	0.0128	0.0080	0.0056	0.0280	0.0081	0.0077	0.0070	0.1873
P-MLE	0.7363	0.0127	0.0076	0.0055	0.0280	0.0080	0.0073	0.0070	0.1876
P-UNMLE	0.7359	0.0127	0.0079	0.0055	0.0280	0.0082	0.0076	0.0071	0.1871
P-UMLE	0.7259	0.0141	0.0093	0.0072	0.0293	0.0089	0.0091	0.0076	0.1886

Table 6.34: Relative Bias (%) of the estimators

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	-2.70	25	108	62.5	-4.28	25	112	30	-3.73
P-ST	0.04	-0.78	0.01	0.01	0.01	0.01	0.01	0.01	-0.11
P-UM	-1.49	14.1	15	33.9	5.71	11.1	16.8	10	0.85
P-UNIT	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	-0.11
P-MLE	0.13	-0.78	-5	-1.78	0.01	-1.23	-5.19	0.01	0.05
P-UNMLE	0.08	-0.78	-1.25	-1.78	0.01	1.23	-1.30	1.43	-0.21
P-UMLE	-1.28	10.2	16.2	28.6	4.64	9.87	18.2	8.57	0.59

Table 6.35: Standard deviation of the point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	1.72	0.56	0.53	0.41	0.66	0.39	0.54	0.39	1.54
P-ST	2.84	0.85	0.99	0.71	0.73	0.57	1.05	0.52	2.08
P-UM	2.03	0.71	1.04	0.55	0.80	0.49	1.08	0.54	1.89
P-UNIT	2.84	0.85	0.98	0.71	0.73	0.57	1.04	0.52	2.07
P-MLE	2.31	0.73	0.70	0.59	0.67	0.52	0.79	0.47	1.92
P-UNMLE	2.12	0.69	0.61	0.56	0.67	0.50	0.66	0.45	1.85
P-UMLE	1.80	0.63	0.85	0.52	0.71	0.47	0.91	0.50	1.75

Table 6.36: RMSE of point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	19.9	3.26	8.71	3.52	1.37	2.00	8.64	2.17	7.18
P-ST	2.85	0.85	0.99	0.71	0.74	0.57	1.05	0.52	2.08
P-UM	11.2	1.96	1.55	1.95	1.75	0.98	1.73	0.93	2.50
P-UNIT	2.84	0.85	0.98	0.71	0.75	0.57	1.04	0.52	2.07
P-MLE	2.51	0.73	0.84	0.61	0.68	0.54	0.90	0.47	1.98
P-UNMLE	2.21	0.70	0.63	0.57	0.68	0.50	0.68	0.45	1.84
P-UMLE	9.63	1.48	1.55	1.64	1.51	0.82	1.72	0.74	2.15

Simulation Study VII: Differential measurement error – Differential flows

External validation sample, Heterogeneity according to gender

$n^v = 10000$, $n = 60000, H = 100$

Table 6.37: True flows

EE	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
0.7353	0.0128	0.0080	0.0056	0.0280	0.0081	0.0077	0.0070	0.1875

Table 6.38: Point estimates, Averages over simulations

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	0.7163	0.0160	0.0160	0.0091	0.0269	0.0102	0.0156	0.0091	0.1808
P-ST	0.7363	0.0127	0.0072	0.0056	0.0281	0.0082	0.0068	0.0070	0.1881
P-UM	0.7247	0.0144	0.0091	0.0074	0.0296	0.0092	0.0090	0.0079	0.1887
P-UNIT	0.7362	0.0127	0.0072	0.0057	0.0281	0.0082	0.0069	0.0071	0.1879
P-MLE	0.7368	0.0125	0.0073	0.0055	0.0280	0.0083	0.0070	0.0072	0.1874
P-UNMLE	0.7362	0.0125	0.0077	0.0055	0.0279	0.0085	0.0074	0.0074	0.1869
P-UMLE	0.7265	0.0140	0.0092	0.0071	0.0293	0.0090	0.0090	0.0077	0.1882

Table 6.39: Relative Bias (%) of the estimators

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	-2.58	25	100	62.5	-3.93	25.9	103	30	-3.57
P-ST	0.14	-0.78	-10	0.01	0.36	1.23	-11.7	0.01	0.32
P-UM	-1.44	12.5	13.7	32.1	5.71	13.6	16.9	12.8	0.64
P-UNIT	0.12	-0.78	-10	1.78	0.35	1.23	-10.4	1.43	0.21
P-MLE	0.20	-2.34	-8.75	-1.78	0.01	2.47	-9.09	2.85	-0.05
P-UNMLE	0.12	-2.34	-3.75	-1.78	-0.35	4.94	-3.89	5.71	-0.32
P-UMLE	-1.19	9.37	15	26.8	4.64	11.1	16.9	10	0.37

Table 6.40: Standard deviation of the point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	1.93	0.49	0.48	0.36	0.72	0.45	0.48	0.37	1.72
P-ST	2.87	0.78	1.13	0.69	0.89	0.66	1.10	0.66	2.23
P-UM	2.20	0.63	0.99	0.56	0.94	0.59	1.04	0.56	2.05
P-UNIT	2.87	0.78	1.13	0.69	0.89	0.63	1.10	0.66	2.22
P-MLE	2.27	0.68	0.79	0.57	0.78	0.57	0.77	0.53	1.67
P-UNMLE	2.04	0.62	0.66	0.50	0.77	0.54	0.66	0.49	1.67
P-UMLE	1.86	0.55	0.83	0.53	0.84	0.58	0.84	0.51	1.67

Table 6.41: RMSE of point estimators (*10^6)

Estimators	*EE*	*UE*	*NE*	*EU*	*UU*	*NU*	*EN*	*UN*	*NN*
P-OBS	19.0	3.14	7.95	3.52	1.33	2.07	7.86	2.20	6.72
P-ST	3.05	0.79	0.45	0.69	0.90	0.66	1.39	0.66	2.32
P-UM	10.8	1.73	1.47	1.87	1.83	1.22	1.61	1.11	2.43
P-UNIT	3.04	0.79	1.40	0.69	0.89	0.66	1.37	0.66	2.30
P-MLE	2.77	0.72	1.06	0.58	0.78	0.60	1.05	0.57	1.68
P-UNMLE	2.25	0.67	0.75	0.52	0.78	0.65	0.74	0.62	1.78
P-UMLE	8.98	1.26	1.40	1.57	1.57	1.05	1.37	0.91	1.88

Simulation Study I: Non–differential flows – Non–differential misclassification

ICE True Scenario – External Validation Sample

$n^v = 2150,\ n = 60000, H = 20000$

Table 6.42: Performance of the variance estimator for the conventional estimator

Flow	$E\left[\hat{Var}\left(\hat{P}^{st-ext}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{st-ext}\right)$ *(*10⁶)*	***Absolute Relative Bias (%)***	***Coverage Rate***
EE	32.6	32.7	0.30	0.945
UE	3.47	3.48	0.28	0.934
NE	8.55	8.44	1.30	0.949
EU	3.42	3.44	0.58	0.934
UU	2.82	2.80	0.71	0.924
NU	2.89	2.87	0.69	0.939
EN	8.48	8.36	1.43	0.948
UN	2.96	2.88	2.77	0.935
NN	27.9	27.6	1.08	0.943

Table 6.43: Performance of the variance estimator for the modified estimator

Flow	$E\left[\hat{Var}\left(\hat{P}^{mod}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{mod}\right)$ *(*10⁶)*	***Absolute Relative Bias (%)***	***Coverage Rate***
EE	30.5	30.6	0.33	0.751
UE	3.24	3.20	1.25	0.895
NE	8.08	7.95	1.63	0.514
EU	3.20	3.15	1.58	0.864
UU	2.62	2.57	1.94	0.826
NU	2.70	2.68	0.75	0.931
EN	7.75	7.86	1.40	0.502
UN	2.76	2.73	1.10	0.927
NN	25.9	26.1	0.77	0.764

Table 6.44: Performance of the variance estimator for the composite estimator with fixed weights (1st set of weights, see Table 6.5)

Flow	$E\left[\hat{Var}\left(\hat{P}^{comp-fx1}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{comp-fx1}\right)$ *(*10⁶)*	***Absolute Relative Bias (%)***	***Coverage Rate***
EE	31.9	31.4	1.59	0.918
UE	3.40	3.36	1.19	0.930
NE	8.37	8.30	0.84	0.881
EU	3.34	3.29	1.52	0.922
UU	2.75	2.69	2.23	0.902
NU	2.83	2.81	0.71	0.937
EN	8.14	8.25	1.33	0.877
UN	2.90	2.87	1.04	0.935
NN	27.2	27.4	0.73	0.914

Table 6.45: Performance of the variance estimator for the composite estimator with fixed weights (2nd set of weights, see Table 6.5)

Flow	$E\left[\hat{Var}\left(\hat{P}^{comp-fx2}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{comp-fx2}\right)$ *(*10⁶)*	*Absolute Relative Bias (%)*	*Coverage Rate*
EE	32.1	31.7	1.26	0.931
UE	3.42	3.39	0.88	0.934
NE	8.42	8.36	0.85	0.911
EU	3.37	3.32	0.72	0.928
UU	2.77	2.71	2.21	0.910
NU	2.86	2.83	1.06	0.937
EN	8.20	8.31	1.32	0.905
UN	2.92	2.89	1.04	0.937
NN	27.4	27.6	0.72	0.926

Table 6.46: Performance of the variance estimator for the composite estimator with fixed weights (3rd set of weights, see Table 6.5)

Flow	$E\left[\hat{Var}\left(\hat{P}^{comp-fx3}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{comp-fx3}\right)$ *(*10⁶)*	*Absolute Relative Bias (%)*	*Coverage Rate*
EE	32.4	31.9	1.56	0.941
UE	3.44	3.41	0.88	0.937
NE	8.48	8.41	0.83	0.930
EU	3.39	3.34	1.50	0.933
UU	2.79	2.73	2.20	0.917
NU	2.88	2.85	1.05	0.938
EN	8.25	8.36	1.31	0.927
UN	2.94	2.91	1.03	0.938
NN	27.6	27.8	0.72	0.936

Simulation Study III: Non–differential flows – Non–differential misclassification

Relaxed ICE Scenario 2 – External Validation Sample

$n^v = 2150,\ n = 60000, H = 20000$

Table 6.47: Performance of the variance estimator for the conventional estimator

Flow	$E\left[\hat{Var}\left(\hat{P}^{st-ext}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{st-ext}\right)$ *(*10⁶)*	*Absolute Relative Bias (%)*	*Coverage Rate*
EE	32.6	32.7	0.31	0.928
UE	3.46	3.44	0.58	0.942
NE	8.54	8.66	1.38	0.939
EU	3.41	3.37	1.10	0.936
UU	2.81	2.78	1.19	0.944
NU	2.89	2.84	1.76	0.946
EN	8.48	8.36	1.43	0.908
UN	2.96	2.91	1.72	0.950
NN	27.8	28.1	1.07	0.945

Table 6.48: Performance of the variance estimator for the modified estimator

Flow	$E\left[\hat{Var}\left(\hat{P}^{mod}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{mod}\right)$ *(*10⁶)*	*Absolute Relative Bias (%)*	*Coverage Rate*
EE	30.5	30.7	0.65	0.890
UE	3.25	3.22	0.93	0.900
NE	8.09	8.10	0.12	0.678
EU	3.20	3.15	1.58	0.942
UU	2.62	2.59	1.16	0.876
NU	2.70	2.65	1.88	0.949
EN	7.79	7.81	0.26	0.749
UN	2.76	2.72	1.47	0.942
NN	25.9	26.2	1.14	0.870

Table 6.49: Performance of the variance estimator for the composite estimator with fixed weights (1st set of weights, see Table 6.5)

Flow	$E\left[\hat{Var}\left(\hat{P}^{comp-fx1}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{comp-fx1}\right)$ *(*10⁶)*	*Absolute Relative Bias (%)*	*Coverage Rate*
EE	31.9	32.1	0.62	0.950
UE	3.39	3.37	0.59	0.933
NE	8.37	8.48	1.29	0.940
EU	3.34	3.30	1.21	0.946
UU	2.76	2.72	1.47	0.932
NU	2.83	2.78	1.79	0.947
EN	8.15	8.19	0.49	0.945
UN	2.90	2.85	1.75	0.948
NN	27.2	27.5	1.09	0.949

Table 6.50: Performance of the variance estimator for the composite estimator with fixed weights (2nd set of weights, see Table 6.5)

Flow	$E\left[\hat{Var}\left(\hat{P}^{comp-fx2}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{comp-fx2}\right)$ *(*10⁶)*	***Absolute Relative Bias (%)***	***Coverage Rate***
EE	32.1	32.3	0.62	0.947
UE	3.42	3.39	0.88	0.936
NE	8.42	8.54	1.40	0.945
EU	3.36	3.32	1.20	0.946
UU	2.78	2.74	1.46	0.936
NU	2.85	2.80	1.78	0.947
EN	8.21	8.25	0.48	0.942
UN	2.91	2.87	1.39	0.948
NN	27.4	27.7	1.08	0.951

Table 6.51: Performance of the variance estimator for the composite estimator with fixed weights (3rd set of weights, see Table 6.5)

Flow	$E\left[\hat{Var}\left(\hat{P}^{comp-fx3}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{comp-fx3}\right)$ *(*10⁶)*	***Absolute Relative Bias (%)***	***Coverage Rate***
EE	32.3	32.5	0.61	0.939
UE	3.44	3.41	0.88	0.940
NE	8.48	8.60	1.39	0.947
EU	3.38	3.35	0.89	0.941
UU	2.80	2.76	1.45	0.942
NU	2.87	2.82	1.77	0.946
EN	8.26	8.31	0.60	0.931
UN	2.94	2.89	1.73	0.949
NN	27.6	27.9	1.07	0.949

Simulation Study VIII: Non–differential flows – Non–differential misclassification

ICE True Scenario – Internal Validation Sample

$n^v = 2150,\ n - n^v = 57850, H = 20000$

Table 6.52: Performance of the variance estimator for the conventional estimator

Flow	$E\left[\hat{Var}\left(\hat{P}^{st-int}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{st-int}\right)$ *(*10⁶)*	*Absolute Relative Bias (%)*	*Coverage Rate*
EE	31.4	31.3	0.32	0.944
UE	3.23	3.29	1.82	0.931
NE	8.21	8.29	0.96	0.942
EU	3.38	3.44	1.74	0.930
UU	2.72	2.73	0.36	0.920
NU	2.87	2.90	1.03	0.933
EN	8.20	8.31	1.32	0.943
UN	2.79	2.84	1.76	0.933
NN	27.0	27.2	0.73	0.942

Simulation Study IX: Non–differential flows – Non-differential misclassification

ICE True Scenario – External Validation Sample

$n^v = 10000,\ n = 60000, H = 100$

Table 6.53: Performance of the variance estimator for the maximum likelihood estimator (4-state model) under the Missing Information Principle with 50000 simulations (Section 5.4.1)

Flow	$E\left[\hat{Var}\left(\hat{P}^{mle}\right)\right]$ *(*10⁶)*	$V\left(\hat{P}^{mle}\right)$ *(*10⁶)*	*Absolute Relative Bias (%)*	*Coverage Rate*
EE	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

Simulation Study X: Non–differential flows – Non-differential misclassification

ICE Scenario – Internal Validation Sample

$n^v = 10000$, $n - n^v = 50000, H = 500$

Table 6.54: True flows

EE	*U+N,E*	*E,U+N*	*U+N,U+N*
0.7288	0.0129	0.0054	0.2529

Table 6.55: Point estimates, Averages over simulations

Estimator	*EE*	*U+N,E*	*E,U+N*	*U+N,U+N*
P-OBS	0.7160	0.0320	0.0250	0.2270
P-ST	0.7290	0.0128	0.0053	0.2529
P-MLE	0.7286	0.0132	0.0057	0.2525

Table 6.56: Relative bias of point estimators (%)

Estimator	*EE*	*U+N,E*	*E,U+N*	*U+N,U+N*
P-OBS	-1.756	148.0	362.9	-10.24
P-ST	0.027	-0.775	-1.851	0.039
P-MLE	-0.027	2.325	5.555	-0.158

Table 6.57: Standard deviation of point estimators (*10^6)

Estimator	*EE*	*U+N,E*	*E,U+N*	*U+N,U+N*
P-OBS	2.58	1.09	0.87	2.54
P-ST	3.21	1.86	1.68	3.45
P-MLE	3.03	1.62	1.39	3.08

Table 6.58: RMSE point estimators (*10^5)

Estimator	*EE*	*U+N,E*	*E,U+N*	*U+N,U+N*
P-OBS	4.13	6.05	6.20	8.23
P-ST	1.02	1.87	1.69	1.09
P-MLE	0.96	1.64	1.41	0.98

Table 6.59: Point estimates, Averages over simulations (“Naïve” Vs. Full information)

Estimator	*EE*	*U+N,E*	*E,U+N*	*U+N,U+N*
P-ST	0.7290	0.0128	0.0053	0.2529
P-MLE (“Naïve”)	0.7285	0.0133	0.0059	0.2522
P-MLE (Full Information)	0.7286	0.0132	0.0057	0.2525

Table 6.60: Standard deviation of point estimators (*10^6) (“Naïve” Vs. Full information)

Estimator	*EE*	*U+N,E*	*E,U+N*	*U+N,U+N*
P-ST	3.21	1.86	1.68	3.45
P-MLE (“Naïve”)	3.09	1.83	1.62	3.28
P-MLE (Full Information)	3.03	1.62	1.39	3.08

Table 6.61: RMSE of point estimators (*10^5) (“Naïve” Vs. Full information)

Estimator	*EE*	*U+N,E*	*E,U+N*	*U+N,U+N*
P-ST	1.02	1.87	1.69	1.09
P-MLE (“Naïve”)	0.98	1.87	1.69	1.06
P-MLE (Full Information)	0.96	1.64	1.41	0.98

Simulation Study XI: Comparing alternative parameterisations for cross-sectional inference

External Validation Sample

$n = 60000,\ n^v = 3000,\ H = 1000$

Table 6.62: Point Estimates, Averages of 1000 simulations

Point Estimators	$\hat{P}_1$
Moment-type	0.6061
MLE (Tenenbein 1972)	0.6059
MLE (EM algorithm)	0.6059
Quasi-likelihood	0.6059

Table 6.63: Empirical comparison of the alternative point estimators

Point Estimators	*Relative Bias (%)*	*Standard Deviation*	*RMSE*
	($*10^4$)	*($*10^5$)*	*($*10^5$)*
Moment-type	1.65	1.18	1.18
MLE (Tenenbein 1972)	-1.65	1.13	1.13
MLE (EM algorithm)	-1.65	1.13	1.13
Quasi-likelihood	-1.65	1.13	1.13

6.7 The Performance of the Alternative Point Estimators

Based on the results derived from the different simulation studies, we now compare the alternative point estimators.

The Performance of the Unadjusted Estimator

Measurement error has a significant effect on the estimated labour force gross flows. Ignoring measurement error and estimating labour force gross flows without any further adjustment results in the overestimation of the probabilities of transition. Deciding whether to use the unadjusted estimates or the adjusted estimates will depend on the intensity of the misclassification problem and on the trade-off between the variance of the adjusted gross flows and the bias of the unadjusted gross flows. Based on the simulation results, the variance of the unadjusted gross flows appears to be smaller than the variance of the adjusted gross flows (see for example, Tables 6.10, 6.15, 6.20). This is due to the extra variability introduced by the adjustment procedure. However, the bias of the unadjusted gross flows is always larger than the bias of the adjusted gross flows (see for example, Tables 6.9, 6.14, 6.19). This is the case both under ICE and under the relaxed-ICE scenarios. Based on a mean squared error criterion (see for example, Tables 6.11, 6.16, 6.21), we conclude that the point estimators that adjust gross flows for measurement error should be preferred over the point estimator that ignores measurement error.

The Performance of the Alternative Moment-type Estimators

As expected, the effect of the conventional estimator is to correct the observed flows towards higher stability i.e. decrease the off diagonal observed flows and increase the diagonal observed flows. This estimator performs very well under ICE (see for example, Table 6.11), but starts deteriorating under the relaxed-ICE scenarios. However, it appears to be quite robust to departures from ICE (see for example, Tables 6.16 and 6.21).

The modified estimator adjusts the observed flows in the same direction as the conventional estimator. However, it tends to reduce the adjustments produced under ICE. Consequently, the performance of the modified estimator is in the reverse direction from that of the conventional estimator (see Tables 6.11, 6.16 and 6.21). Given the disadvantages associated with the modified estimator (see Section 2.4.2) and the robustness of the conventional estimator, we propose the use of the conventional estimator instead of the modified estimator.

The composite estimator with fixed weights also reduces the adjustments produced by the conventional estimator. However, this estimator performs reasonably well under ICE (see Table 6.11) and very well under the relaxed ICE scenarios (see Tables 6.16 and 6.21). Thus, we propose the use of the composite estimator with fixed weights, as an alternative to the conventional estimator, in situations where the effect of the ICE assumption is very pronounced.

The composite estimator with adaptive weights provides less severe adjustments than the adjustments derived by the conventional estimator. The composite estimator with adaptive weights performs very well under ICE (see Table 6.11). In fact, under ICE this estimator has the minimum mean squared error. In addition, the composite estimator with adaptive weights is robust to departures from the ICE. We suggest that this estimator provides a promising moment-based alternative to the conventional estimator.

In contrast to the conventional estimator that provides an upper bound for the adjustments, the unbiased margins estimator provides a lower bound for the adjustments. The unbiased margins estimator behaves like the moment-type estimators that attempt to relax the ICE assumption. It is biased under ICE (see Table 6.9) but it improves under the relaxed ICE scenarios (see Tables 6.16 and 6.21). This estimator provides an alternative to the conventional estimator when the impact of the ICE assumption is very pronounced.

In summary, the composite estimator with fixed or adaptive weights and the unbiased margins estimator provide reasonable moment-type alternatives to the conventional estimator.

Contrasting the Moment-type Estimators with the Maximum Likelihood Estimators

In Chapter 3, we developed a maximum likelihood estimator and a constrained maximum likelihood estimator as alternatives to the conventional estimator and to the unbiased margins estimator respectively. Our simulation results indicate that when the validation sample is selected independently from the main sample and from the same target population, the proposed maximum likelihood estimators are more efficient than the moment-type estimators (see for example, Tables 6.26 and 6.31). A validation sample that is selected by sub-sampling units from the main (panel) survey may increase the response burden of these units. Using an independently selected validation sample in a panel framework may be more reasonable. However, such an independently selected validation sample is also associated with higher costs. This is because, when using an independently selected validation sample, we conduct an additional cross-sectional survey of individuals that do not participate in the main survey. The conventional 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 X we contrast the maximum likelihood estimator with the conventional estimator when the validation sample is selected by sub-sampling units from the main sample. Under this double sampling scheme, both estimators use the same information.

Again, our results indicate that the maximum likelihood estimator is more efficient (see Table 6.58). Based on these results, we therefore recommend the maximum likelihood estimator and the constrained maximum likelihood estimator instead of the conventional estimator and the unbiased margins estimator.

In Section 3.2.2, we also presented two alternative approaches (i.e. the "naïve" approach and the full information approach) for performing the E-step of the EM algorithm when the validation sample is selected by sub-sampling units from the main sample. In Tables 6.59-6.61 we contrast these two approaches. We conclude that in performing the E-step of the EM algorithm under the specific double sampling scheme, the full information approach is more efficient than the "naïve" approach (see Table 6.61) and therefore should be preferred.

Comparing the Alternative Estimators in the Presence of Heterogeneity

In simulation studies VI and VII, we allowed for moderate gender-based heterogeneity in the gross flows mechanism and/or in the measurement error mechanism. The maximum likelihood estimator that allows for heterogeneity is generally more efficient than the corresponding moment-type (unit heterogeneity) estimator (see Tables 6.36 and 6.41). In addition, since the maximum likelihood that allows for heterogeneity incorporates stratification, we expect that it will be superior to the maximum likelihood estimator that ignores heterogeneity. The results verify this assumption (see Table 6.36 and Table 6.41). We conclude that even in the presence of moderate heterogeneity, the maximum likelihood estimator that allows for heterogeneity should be preferred over the maximum likelihood estimator that ignores heterogeneity.

6.8 The Performance of the Alternative Variance Estimators

In this section, we assess the performance of the variance estimators. Two evaluation criteria are used. These are the relative bias of the variance estimator and the coverage rate when using the variance estimator.

The Performance of the Variance Estimator for the Conventional Estimator

Under ICE, the variance estimator of the conventional estimator works very well. More specifically, in the case of an external validation sample the absolute relative bias ranges between 0.28 % and 2.77 % and the coverage rate ranges between 92.4% and 94.9% (see Table 6.42). For the case of an internal validation sample, the absolute relative bias ranges between 0.32% and 1.82% and the coverage rate ranges between 92.0% and 94.4% (see Table 6.52). Under the second relaxed-ICE scenario, the variance estimator for the conventional estimator performs well with low relative bias. The coverage rates are not affected (see Table 6.47). In only one case does the coverage rate drop from 94.8% to 90.8% (EN flow). The preservation of coverage rates close to 95% indicates that the conventional estimator is robust to departures from ICE.

The Performance of the Variance Estimator for the Modified Estimator

Under ICE, the variance estimator of the modified estimator works well with absolute relative bias that ranges between 0.33% and 1.94%. However, the coverage rates range between 50% and 93% (see Table 6.43). This under-coverage can be attributed to the bias of the modified estimator under ICE. Under the second relaxed-ICE scenario, the modified estimator preserves its good performance (see Table 6.48). The coverage rates increase but there are still cases of under-coverage due to the bias of this estimator.

The Performance of the Variance Estimator for the Composite Estimator with Fixed Weights

Under ICE, the variance estimator of the composite estimator with fixed weights works well with absolute relative bias that ranges between 0.71% and 2.23%. The coverage rates come closer to 95% as we reduce the weight of the modified estimator (see Tables 6.44, 6.45 and 6.46). This is expected, since under the third set of weights (Table 6.5) the composite estimator is closer to the conventional estimator and thus approximately unbiased when ICE is valid. Under a relaxed-ICE scenario, the variance approximations for this estimator work well with coverage rates close to 95 % (see Tables 6.49, 6.50 and 6.51).

The Performance of the Variance Estimator for the Maximum Likelihood Estimator

The variance estimator of the maximum likelihood estimator (under a validation sample that is selected independently from the main sample and from the same target population) appears to be conservative since it overestimates the true variance (see Table 6.53). This overestimation occurs mainly in the off-diagonal elements of the gross flows matrix. Despite being conservative, two positive outcomes emerge from the use of this variance estimator. Firstly, we feel confident that we capture the variability due to the missing data. Secondly, we derive reasonable coverage rates that range between 90%-94%. Given the complexity of the problem, we believe that this variance estimator provides a reasonable approximation to the variance of the maximum likelihood estimator.

6.9 Summary

We now summarise the main findings from the evaluation of the methodology presented in this chapter. Among the alternative moment-type estimators, we propose the use of the composite estimator with fixed or with adaptive weights and of the unbiased margins estimator as alternatives to the conventional estimator. The maximum likelihood estimator and the constrained maximum likelihood estimator are more efficient than the conventional estimator and the unbiased margins estimator respectively. The higher efficiency of the maximum likelihood estimator is preserved under a validation sample that is selected by sub-sampling units from the main sample. When using the EM algorithm under this double sampling scheme, it is preferable to employ the full information approach instead of the "naïve" approach for estimating the conditional expectations of the missing data in the validation sample (see Section 3.2.2). The gains from accounting for heterogeneity seem also to be quite significant. The variance estimators of the moment-type estimators provide reasonable approximations to the true variance of these estimators. The variance estimator of the maximum likelihood estimator, under a validation sample that is selected independently from the main sample, appears to be conservative. However, since this estimator captures the variability due to the missing data and leads to reasonable coverage rates, we argue that it can provide a reliable solution.

Chapter 7

A Note on the Design of a UK LFS Re-interview Survey: Suggestions Based on Empirical Evidence

7.1 Introduction

The methodology developed in this thesis assumes the availability of validation data that are derived from a re-interview survey. However, the UK LFS has not yet developed a re-interview survey for estimating the parameters of the measurement error mechanism. For this reason, throughout this thesis we have relied on external re-interview data (mainly Swedish, and also data from Canada and the US) calibrated to information derived from the UK LFS. In this chapter, we provide some recommendations for the design of a re-interview survey for the UK LFS. The emphasis is on identifying optimal design characteristics for conducting a re-interview survey. In Section 7.2, we empirically compare re-interview surveys with different reconciliation strategies. In Section 7.3, we compare alternative double sampling schemes. We give suggestions for the selection of an appropriate double sampling scheme based on the following three criteria: (a) the cost of implementing this scheme, (b) the implications for the quality of the main (ongoing) survey and of the validation survey and (c) the implications for point and interval estimation.

7.2 Comparing Re-interview Surveys with Different Reconciliation Strategies

A crucial design characteristic of a re-interview survey is the method of reconciling the original response with the re-interview response (see Section 1.8.2). In this section, we investigate the effect of different reconciliation strategies using information from two validation surveys namely, the CPS re-interview survey as described in Poterba and Summers (1986) and the Swedish re-interview survey (Kristiansson 1999). The CPS re-interview survey consists of two samples. In one sample, consisting of the 25% of the total re-interview sample, re-interviews are carried out without any attempt for reconciliation and without access to the original responses. For the remaining 75% of the total re-interview sample, re-interviewers are provided with the original responses and attempt reconciliation if there are discrepancies between the original and the re-interview responses. The Swedish validation survey is described in Section 1.8.4.

One way to quantify the comparison between validation surveys with different reconciliation procedures is by contrasting the misclassification rates from the unreconciled and the reconciled samples and examine whether these are significantly different. One indication of the violation of the assumption that the re-interview survey identifies the true value is that the reconciled sample shows a smaller number of discrepancies compared to the unreconciled sample. A problem of this kind has been reported by Poterba and Summers (1986) using data from the CPS re-interview programme. Unfortunately, similar comparisons between the reconciled and the unreconciled data for the Swedish re-interview programme (October 1994 - April 1995) are difficult because the design of this re-interview programme is different from that of the CPS re-interview survey. In the Swedish case, after the first interview a sub-sample of units is re-interviewed using computer assisted telephone interviewing. Furthermore, due to the computerised nature of the Swedish re-interview programme, re-interviewers have no access to the original data before the re-interview survey. In case a discrepancy between the original and the re-interview occurs, reconciliation takes place. Note that this reconciliation process is not just a clerical check where the results from the re-interview are considered as the true values. For example, there are discrepancies between the re-interview and the reconciled results in the Swedish re-interview dataset.

In order to examine whether the reconciliation process in the Swedish case introduces any problems, we simulate a design similar to the design that is used by the CPS re-interview survey. We use re-interview data from the Swedish LFS re-interview programme (October 1994 - April 1995). The Swedish re-interview programme consists of approximately $n^v = 2150$ individuals. In each simulation, we randomly select a sample of size $n^{v(u)} = 538$ (i.e. 25% of the total re-interview sample) from the Swedish re-interview data set. For these units, we compare the original with the re-interview responses and we construct the misclassification matrix. For the non-sampled units $n^{v(r)} = 1613$ (i.e. the 75% of the total re-interview sample), we compare the original responses with the reconciled responses and we construct a second misclassification matrix. We conduct a total of $H = 1000$ simulations. The first sample can be regarded as equivalent to the unreconciled sample of the CPS re-interview survey while the second sample can be regarded as equivalent to the reconciled sample of the CPS re-interview survey. Furthermore, the responses in the first sample are independent from the responses in the second sample since the two samples do not share common units. Our primary target is to compare corresponding off-diagonal elements of the misclassification matrices estimated from these two samples.

One way to perform these comparisons is by constructing confidence intervals for the differences between corresponding off-diagonal elements of the misclassification matrices. Treating the elements of the misclassification matrices as multinomial proportions, one can make inferences by constructing simultaneous confidence intervals. Literature on simultaneous confidence intervals for multinomial proportions includes Quesenberry and Hurst (1964), Goodman (1964, 1965), and Fitzpatrick and Scott (1987). For the purposes of our application we follow Goodman's approach (1964). Assume that for each re-interview programme we have two samples. In one sample reconciliation takes place while in the other sample no reconciliation is conducted. Previously we used q_{ik} to denote the cross-sectional misclassification probabilities. However, q_{ik} will now refer to the joint and not to the conditional probabilities of misclassification. Recalling the notation from Chapter 1, a superscript (u) will denote quantities from the unreconciled sample while a superscript (r) will denote quantities from the reconciled sample. Denote also by $n^{v(u)}$ the sample size of the unreconciled sample and by $n^{v(r)}$ the sample size of the reconciled sample. Define a contrast θ_{ik} between the two samples to be a linear function of q_{ik} satisfying

$$\theta_{ik} = \sum_{j=u}^{r} c_{ik}^{(j)} q_{ik}^{(j)}, \ \sum_{j=u}^{r} c_{ik}^{(j)} = 0 . \tag{7.1}$$

Let $\hat{\theta}_{ik} = \sum_{j=u}^{r} c_{ik}^{(j)} \hat{q}_{ik}^{(j)}$ denote the estimated contrast for a fixed cell ik. Under a multinomial assumption and the assumption of independence between the reconciled sample and the unreconciled sample, we have that

$$\hat{Var}\left(\hat{\theta}_{ik}\right) = \hat{Var}\left[\sum_{j=u}^{r} c_{ik}^{(j)} \hat{q}_{ik}^{(j)}\right] = \left(c_{ik}^{(u)}\right)^2 \frac{\hat{q}_{ik}^{(u)}\left(1-\hat{q}_{ik}^{(u)}\right)}{n^{v(u)}} + \left(c_{ik}^{(r)}\right)^2 \frac{\hat{q}_{ik}^{(r)}\left(1-\hat{q}_{ik}^{(r)}\right)}{n^{v(r)}}. \tag{7.2}$$

For the purposes of our application, we are not interested in all possible combinations of contrasts but in comparisons between corresponding off-diagonal elements of the misclassification matrices. Goodman (1964) gives an expression for constructing simultaneous confidence intervals in case one is interested to a specific set of G contrasts

$$\hat{\theta}_{ik} - \sqrt{\hat{Var}\left(\hat{\theta}_{ik}\right)}\, Z \leq \theta_{ik} \leq \hat{\theta}_{ik} - \sqrt{\hat{Var}\left(\hat{\theta}_{ik}\right)}\, Z, \tag{7.3}$$

where $\hat{Var}\left(\hat{\theta}_{ik}\right)$ is defined by (7.2) and Z is the $100(1-\alpha/2G)th$ percentile point of the unit normal distribution. For the CPS re-interview programme, we directly apply (7.3) using the data from Poterba and Summers (1986 p.1323). For the Swedish re-interview programme, we evaluate the quantities involved in (7.3) using the simulation study. The expectation (over simulations) of an off-diagonal element of the misclassification matrix from re-interview sample j is given by

$$E\left(\hat{q}_{ik}^{(j)}\right) = \sum_{h=1}^{H} \frac{1}{H} \hat{q}_{ikh}^{(j)}. \tag{7.4}$$

The empirical (simulation-based) variance of an off-diagonal element of the misclassification matrix from re-interview sample j is given by

$$V\left(\hat{q}_{ik}^{(j)}\right) = \frac{1}{H-1} \sum_{h=1}^{H} \left[\hat{q}_{ikh}^{(j)} - E\left(\hat{q}_{ik}^{(j)}\right)\right]^2. \tag{7.5}$$

Using (7.4) and (7.1), one can estimate the expectation (over simulations) of contrast θ_{ik}. Using (7.5), one can estimate the variance of contrast θ_{ik}. The simultaneous confidence intervals are then determined using (7.3).

Table 7.1: Simultaneous confidence intervals for contrasts between corresponding off-diagonal elements derived from the unreconciled sample and the reconciled sample of the CPS re-interview survey, $\alpha = 0.05$, $G = 6$.

	CPS Re-interview Survey	
Contrast	***Lower Bound***	***Upper Bound***
θ_{EU}	0.0015	0.0069
θ_{EN}	0.0052	0.0110
θ_{UE}	0.0001	0.0049
θ_{UN}	0.0006	0.0068
θ_{NE}	0.0001	0.0088
θ_{NU}	-0.0018	0.0051

Table 7.2: Simultaneous confidence intervals for contrasts between corresponding off-diagonal elements derived from the unreconciled sample and the reconciled sample of the Swedish LFS re-interview survey, $\alpha = 0.05$, $G = 6$, $H = 10000$.

Swedish Re-interview Survey		
Contrast	***Lower Bound***	***Upper Bound***
θ_{EU}	-0.0056	0.0082
θ_{EN}	-0.0115	0.0096
θ_{UE}	-0.0089	0.0127
θ_{UN}	-0.0080	0.0097
θ_{NE}	-0.0120	0.0127
θ_{NU}	-0.0079	0.0108

The results for the CPS re-interview survey (Table 7.1) show that in all cases but one the probabilities of misclassification estimated from the unreconciled sample are significantly higher than the probabilities of misclassification estimated from the reconciled sample. This suggests that the reconciliation process of the CPS re-interview survey invalidates the assumption that reconciliation identifies the true value. For the Swedish re-interview survey, the confidence intervals indicate that in all cases there is no significant difference in the probabilities of misclassification estimated from the reconciled and the unreconciled samples (Table 7.2).

One can object that the simultaneous confidence intervals are conservative. However, the results are in the same direction even by constructing the less conservative pair-wise confidence intervals (each at $\alpha = 0.05$). One can further object that these differences are a consequence of the different sample sizes of the CPS re-interview survey and of the Swedish re-interview survey. The sample size of the Swedish re-interview survey is smaller than the sample size of the CPS re-interview survey. Consequently, the fact that we don't detect any differences between the unreconciled sample and the reconciled sample of the Swedish re-interview survey is due to the higher variability (smaller sample size) in this survey. For this reason, we constructed simultaneous confidence intervals assuming that the sample size of the Swedish re-interview survey is the same as the sample size of the CPS re-interview survey. The results indicated that in all cases there is no significant difference in the probabilities of misclassification estimated from the reconciled and the unreconciled samples. Unlike the CPS reconciliation process, the reconciliation process of the Swedish re-interview survey allows unbiased estimation of the misclassification rates. This can be attributed to the design characteristics of the Swedish re-interview survey, which are close to the optimal characteristics when the aim is to estimate the response bias component (see Section 1.8.2).

7.3 Selecting a Double Sampling Scheme

In designing a re-interview survey, one key decision relates to the choice of a double sampling scheme. In this section, we compare different double sampling schemes for selecting a validation sample. The criteria we use to assess these schemes are the following: (a) the cost of implementing the scheme, (b) the implications for the quality of the main (ongoing) survey and of the validation survey and (c) the implications for point and interval estimation.

One option is to select a validation sample independently from the main sample (double sampling scheme 1). This can be either an external validation (transformed into an internal validation sample using the ideas in Section 2.2.1.1) or a validation sample that is selected independently from the main sample and from the same target population. The second option is to select a validation sample by sub-sampling units from the main sample (double sampling scheme 2).

A validation sample that is selected independently from the main sample or an external validation sample may be associated with higher costs than a validation sample that is selected by sub-sampling units from the main sample. This is because, when using an independently selected validation sample, we conduct an additional cross-sectional survey of individuals that do not participate in the main survey.

When comparing double sampling schemes, we need to account for the fact that the main survey instrument, for example the UK LFS, is a panel survey. Sub-sampling units that already participate in the main survey is equivalent to introducing an extra wave into the panel survey. These additional measurements can lead to an increase in the response burden. Therefore, this double sampling scheme may have implications for the quality both of the main (ongoing) survey and of the validation survey.

In Section 3.2, we formulated the measurement error model both under double sampling scheme 1 and double sampling scheme 2. Point estimation is performed via the EM algorithm. However, when the validation sample is selected by sub-sampling units from the main sample (i.e. double sampling scheme 2), the expectation step in the EM algorithm becomes complicated (see Section 3.2.2). This implies that interval estimation under the specific double sampling scheme will also be complicated.

Based on this comparison, we conclude the following. A validation sample that is selected independently from the main sample is associated with higher costs. However, this double sampling scheme does not impact on the quality of the main survey and of the validation survey. Moreover, under this double sampling scheme, point and interval estimation is simpler. This comparison is summarised in Table 7.3.

Table 7.3: Comparing alternative double sampling schemes

Double Sampling Scheme	*Cost*	*Risk for Quality of Surveys*	*Inference*
1	High	Low risk	Easy
2	Lower	High risk	Complicated

In some cases, an efficient way of selecting a validation sample is offered by the use of administrative databases. The use of an administrative data source can be placed into the context of double sampling scheme 1 or 2. Using an administrative data source may not be as expensive as conducting an interview-based validation survey. In addition, under this approach, we don't prejudice the quality of the main survey. As an alternative, one could include in the re-interview survey only the sample units that participate for the last time in the panel survey (double sampling scheme 2). The advantage of this approach is that there is no risk to the quality of the ongoing survey, since these units participate for a last time in this survey. However, we may risk the quality of the validation survey due to possible existence of conditioning effects in the respondents. Finally, an alternative but more costly solution is offered by selecting the validation sample from a cross-sectional survey that collects information of similar nature to the information that is collected by the main (panel) survey (double sampling scheme 1).

7.4 Summary

In this chapter, we provide some suggestions for designing a UK LFS re-interview survey based on empirical evidence. We suggest an independent reconciliation process, where the re-interviewers have no access to the original responses, and a double sampling scheme based on a validation sample that is selected independently from the main sample. Although this scheme is considered to be less cost efficient, it has advantages in the sense that it does not affect the quality of the ongoing survey and it provides an easier way of performing inference.

Chapter 8

Summary and Suggestions for Further Research

8.1 Summary

The purpose of this thesis is to develop methodology for correcting gross flows estimates for measurement error. The application we focus on is estimation of labour force gross flows, and the approach we follow assumes the availability of cross-sectional validation data that provide information about the measurement error process. We now summarise our basic findings and give directions for future research.

In Chapter 2, we define a general estimation framework and contrast the use of alternative double sampling schemes in both cross-sectional and longitudinal situations. In a cross-sectional situation, we present two alternative parameterisations of the measurement error model that work by combining information both from the main sample and the validation sample. Under the first parameterisation, we formulate the measurement error model as a missing data problem and maximum likelihood estimation is performed via the EM algorithm. Under the second parameterisation, the measurement error model is formulated in a quasi-likelihood framework. There are two advantages offered by the quasi-likelihood parameterisation. Firstly, under this approach we avoid an explicit definition of the likelihood function. Secondly, the quasi-likelihood approach offers an alternative to the EM algorithm, when tackling a missing data problem, and an easier method, compared with the application of the Missing Information Principle, for computing the variances of the adjusted cross-sectional estimates. The results from a Monte-Carlo simulation study (Chapter 6) indicate that the quasi-likelihood approach leads in estimates that are as efficient as the estimates derived from the maximum likelihood approach. In the rest of Chapter 2, we extend the double sampling schemes to the longitudinal situation and we present alternative moment-based estimators to the conventional (moment-based) estimator. These alternative estimators attempt to relax the ICE assumption and to overcome some of the practical problems affecting the conventional estimator for example, the possibility of deriving negative adjusted estimates. Based on the simulation results reported in Chapter 6, we propose the use of the composite estimator with fixed or adaptive weights and the unbiased margins estimator as alternatives to the conventional estimator. We further relaxed the ICE assumption in our simulations by introducing dependence structure in the measurement error mechanism. Our results indicate that the conventional estimator is robust to departures from ICE.

Having parameterised the measurement error model as a missing data problem, in Chapter 3 we extend this idea from the cross-sectional into the longitudinal framework. The lack of a panel validation survey introduces extra complications since missing data exist now in both the main sample and the validation sample. In order to identify the parameters of the measurement error model, the ICE assumption is utilised and estimation is based on the EM algorithm. The model is formulated under two alternative double sampling schemes. Under the first scheme, the validation sample is selected independently from the main sample and from the same target population. Under the second scheme, the validation sample is selected by sub-sampling units from the main sample. In addition, we further propose a constrained maximum likelihood estimator that imposes an unbiased margins constraint on the estimation of the adjusted gross flows and therefore relaxes ICE. The survey weights are allowed for estimation via the pseudo-maximum likelihood approach. In the context of the UK LFS, the survey weights are constructed to adjust for sampling attrition. Thus, the inclusion of the survey weights into the measurement error model contributes implicitly towards the adjustment for sampling attrition and provides a bias correction to the unweighted results. From the simulation results (Chapter 6), we conclude that the maximum likelihood estimators are more efficient than the moment-type estimators. The higher efficiency of the maximum likelihood estimators is preserved under the alternative double sampling schemes. A further advantage of the maximum likelihood estimators is that they constrain estimates to lie within the boundaries of the parameter space. Unlike the maximum likelihood estimators, the moment-type estimators can produce estimates that lie outside the boundaries of the parameter space for example, negative proportions. Based on the outcomes of this research, we argue in favour of the use of maximum likelihood estimators.

In Chapter 4, the longitudinal measurement error model is extended to account for the existence of heterogeneity associated with discrete covariates. The measurement error model allows for heterogeneity through post-strata defined by the cross-classification between these covariates. The results indicate that this type of heterogeneity can have an effect on the adjustments for measurement error. Also the gains in efficiency, even when allowing for moderate heterogeneity (Chapter 6), can be quite significant. One important outcome of this modelling exercise is that we can naturally quantify the effect of misclassification. To explore this effect, we compared the predicted probabilities of transition from a multinomial logistic model that utilises the unadjusted data to the predicted probabilities of transition from the model that accounts for the measurement error process for different age by gender groups. The results indicate that ignoring measurement error can have a severe effect, which in some cases can result in a complete reversal of the direction of the flows.

The central theme of Chapter 5 is variance estimation. A variance estimator for the conventional (moment-type) estimator, under the alternative double sampling schemes, is derived using Taylor series linearization. Based on these results, we further develop variance estimators for the modified estimator and for the composite estimator with fixed weights. Variance estimation for the maximum likelihood estimator, when the validation sample is selected independently from the main sample and from the same target population, is also derived using the Missing Information Principle. General expressions for applying the Missing Information Principle are provided. However, due to the large number of computations involved in the evaluation of this estimator, we follow a simulation-based approach. This algorithm is based on sampling from the conditional distribution of the missing data given the observed data and the maximum likelihood estimates. The simulation results from Chapter 6 indicate that the variance estimators of the moment-type estimators give reasonable approximations to the true variance of these estimators. The variance estimator of the maximum likelihood estimator appears to be conservative. However, since it captures the variability due to the missing data and results in reasonable coverage rates it can be regarded as providing a reliable solution.

In Chapter 7, we provide some suggestions for the design of a re-interview survey for the UK LFS. An empirical comparison of re-interview surveys with different reconciliation strategies (CPS vs. Swedish) indicated that the Swedish method of reconciliation is superior to the CPS reconciliation procedure. A validation sample that is selected by sub-sampling units from the main survey appears to be more cost efficient than a validation sample that is selected independently from the main sample. However, we suggest that the inclusion in the validation survey of sample units that already participate in a panel survey should be avoided since this can contribute to an increase in the response burden. Moreover, inference under this double sampling scheme appears to be more complex than under a scheme where the validation sample is selected independently from the main sample. Some approaches for selecting a validation sample are discussed below. In the case that an administrative database exists that is representative of the target population, we recommend that it should be preferred over an interview-based validation survey. Alternatively, given that there are no conditioning effects, one may use the sample units that are due to be rotated out of the survey. Finally, one may design a validation survey based on a cross-sectional survey that collects information of similar nature to the information that is collected by the panel survey. As a first step towards the design of a UK LFS validation survey, we propose the use of a small-scale experimental survey that will attempt to identify the basic characteristics of the measurement error process in the UK LFS.

8.2 Further Ongoing Research

In this thesis, we have demonstrated that measurement error can introduce severe bias in the estimation of gross flows. The implementation of this methodology in the context of the UK LFS requires the development of a validation survey by the UK Office for National Statistics. Many practical and theoretical issues will undoubtedly arise during a possible implementation. However, we believe that the methodology we developed provides a reliable and efficient solution for adjusting gross flows data for measurement error. Nevertheless, there remain many issues associated with adjusting gross flows for measurement error that are not tackled in this thesis. Firstly, variance estimation taking into account the survey weights requires investigation. One option is to treat these weights as random and use the jackknife method for variance estimation. Variance estimation for the constrained maximum likelihood estimator and for the maximum likelihood estimator, when the validation sample is selected by sub-sampling units from the main survey, also requires further investigation.

In this final section, we describe some ongoing research as well as potential research that extends the methodology developed in this thesis. In Chapter 1 (Section 1.8.3), we described the UK 1991 Census Validation Survey (CVS) (Heady, Smith and Avery 1991). This survey had as its main target an evaluation of how prone to error Census questions were. Among the questions that were tested was one asking about the labour force status of the respondents. Treating the CVS as an error free source of information and linking the CVS to the Census responses, one can estimate a misclassification matrix. However, the Census may contain more measurement error than the LFS. This is because the Census is a self-reported survey. Consequently, the misclassification matrix estimated by linking the CVS to the Census responses may overestimate the measurement error problem in the UK LFS. The question is how we can use the UK CVS in order to approximate the parameters of the misclassification mechanism in the UK LFS. Currently, we are investigating two alternative approaches. Under the first approach, the misclassification probabilities estimated by linking the CVS to the Census responses are employed in order to adjust the Census-based cross-sectional labour force distribution for misclassification. This initial adjustment can be achieved using for example the moment-type estimator described in Section 2.2.1.1. Recalling the notation from Chapter 2, one can then use the following identity that relates the observed, LFS-based, cross-sectional labour force distribution to the adjusted, Census-based, cross-sectional labour force distribution.

$$\underset{r\times 1}{\Pi} = \underset{r\times r}{Q(t)} \underset{r\times 1}{P}. \tag{8.1}$$

We know Π (i.e. the LFS-based observed cross-sectional labour force distribution) and P (i.e. the adjusted Census-based cross-sectional labour force distribution). The unknown quantity in (8.1) is $Q(t)$. Unfortunately, the system of equations defined by (8.1) does not have a unique solution. Consequently, in order to solve this system of equations, we need to introduce additional constraints. One such constraint is offered by minimising the Euclidian distance between the CVS-based misclassification matrix and $Q(t)$. This minimisation can be achieved using Lagrange multipliers. However, initial results have indicated that this approach can lead to negative misclassification probabilities. For this reason, we investigate a second approach. At the first step, the misclassification probabilities estimated by linking the CVS to the Census responses are employed in order to adjust the Census-based cross-sectional labour force distribution for misclassification. In addition, from the LFS we derive an estimate of the observed cross-sectional labour force distribution. The margins of the matrix estimated by linking the CVS to the Census responses represent the cross-sectional observed and adjusted for misclassification labour force distributions. Consequently, one can use the IPF algorithm in order to rake this matrix to the information from LFS (observed cross-sectional labour force distribution) and to the Census-adjusted cross-sectional labour force distribution. This approach provides another solution to the system of equations defined by (8.1). It remains to evaluate both methods in the context of the UK LFS.

In Chapter 2 (Section 2.2.1.4), we parameterised the cross-sectional measurement error model in a quasi-likelihood framework. This needs to be extended by parameterisation of the longitudinal measurement error model in a quasi-likelihood framework. The advantage of the quasi-likelihood parameterisation is that it provides an easier approach to variance estimation (see Section 5.5). This is particularly important in a longitudinal framework where variance estimation for the adjusted maximum likelihood estimates is tedious both analytically and computationally.

The measurement error model we presented in Chapter 3 adjusts implicitly for sampling attrition via the survey weights and the pseudo-maximum likelihood approach. However, due to the formulation of the measurement error model as a missing data problem, it will be interesting to examine whether a simultaneous modelling of the measurement error process and the sampling attrition process is feasible. Preliminary research has indicated that in order for the parameters of the simultaneous model to be identified, we need to impose fairly strong assumptions. Furthermore, the likelihood-based approach that we developed in Chapter 3 opens the possibility of directly contrasting the modelling strategies that assume validation information with the modelling strategies that do not assume validation information for example, the latent class approach.

It is also of some interest to identify other research areas where this methodology can be applied. For example, in demographic applications one of the most common problems is heaping. In reporting the age of death, heaping occurs when the respondents round the reporting age. A graphical representation of the frequency of deaths by age (grouped in 5-year bands) will reveal peaks at 0 and 5 years. This problem is currently tackled using smoothing techniques (Benjamin and Pollard 1986). An alternative adjustment may be possible by viewing heaping as a misclassification problem. Validation data, regarding the year of birth, can be derived from administrative data sources and misclassification probabilities can be estimated by comparing the true to the reported age of death.

Two additional areas of research in official statistics are related to the research conducted in this thesis. In statistical disclosure control, one way of protecting the data is by deliberately misclassifying them and then providing the data analyst with the misclassified data and the misclassification probabilities (Van den Hout and Van der Heijden 2002). The basic difference between the approach utilised by the statistical disclosure control and our methodology is that in the former case the misclassification probabilities are treated as fixed and known whereas in the latter case the misclassification probabilities are unknown and are estimated from a validation survey. Nevertheless, the approach followed in statistical disclosure control shares many similarities with our approach.

The second area of potential application regards adjustments in the Census. For example, the existence of inaccurate addresses can result in the erroneous estimation of the population size in an area. This problem can be modelled in a misclassification context. In some countries, for example in Israel, there are administrative lists that provide information about the address of a population unit. Treating the Census as the main survey and the administrative list as the validation survey, one can employ a model that combines information from both sources to provide adjustments to the Census-based estimates.

Appendix I

First and Second Order Derivatives Involved in the Application of the Missing Information Principle in a Cross-sectional Framework

$$\frac{\partial E\left[l(\Theta;D^c)\mid D^m,D^v\right]}{\partial P_k}=\frac{E\left(n_{\bullet k}^{(*)}\mid D^m,\Theta\right)+n_{\bullet k}^{v}}{P_k}-\frac{E\left(n_{\bullet r}^{(*)}\mid D^m,\Theta\right)+n_{\bullet r}^{v}}{1-\sum_{k=1}^{r-1}P_k} \tag{I1}$$

$$\frac{\partial E\left[l(\Theta;D^c)\mid D^m,D^v\right]}{\partial q_{ik}}=\frac{E\left(n_{ik}^{(*)}\mid D^m,\Theta\right)+n_{ik}^{v}}{q_{ik}}-\frac{E\left(n_{rk}^{(*)}\mid D^m,\Theta\right)+n_{rk}^{v}}{1-\sum_{i=1}^{r-1}q_{ik}} \tag{I2}$$

$$-\frac{\partial^2 E\left[l(\Theta;D^c)\mid D^m,D^v\right]}{\partial P_k^2}=\frac{E\left(n_{\bullet k}^{(*)}\mid D^m,\Theta\right)+n_{\bullet k}^{v}}{P_k^2}+\frac{E\left(n_{\bullet r}^{(*)}\mid D^m,\Theta\right)+n_{\bullet r}^{v}}{\left(1-\sum_{k=1}^{r-1}P_k\right)^2} \tag{I3}$$

$$-\frac{\partial^2 E\left[l(\Theta;D^c)\mid D^m,D^v\right]}{\partial P_k\partial P_{k'}}=\frac{E\left(n_{\bullet r}^{(*)}\mid D^m,\Theta\right)+n_{\bullet r}^{v}}{\left(1-\sum_{k=1}^{r-1}P_k\right)^2} \tag{I4}$$

$$-\frac{\partial^2 E\left[l(\Theta;D^c)\mid D^m,D^v\right]}{\partial P_k\partial q_{ik}}=0 \tag{I5}$$

$$-\frac{\partial^2 E\left[l(\Theta;D^c)\mid D^m,D^v\right]}{\partial q_{ik}\partial q_{ik}}=\frac{E\left(n_{ik}^{(*)}\mid D^m,\Theta\right)+n_{ik}^{v}}{q_{ik}^{\;2}}+\frac{E\left(n_{rk}^{(*)}\mid D^m,\Theta\right)+n_{rk}^{v}}{\left(1-\sum_{i=1}^{r-1}q_{ik}\right)^2} \tag{I6}$$

$$-\frac{\partial^2 E\left[l(\Theta;D^c)\mid D^m,D^v\right]}{\partial q_{ik}\partial q_{i'k}}=\frac{E\left(n_{rk}^{(*)}\mid D^m,\Theta\right)+n_{rk}^{v}}{\left(1-\sum_{i=1}^{r-1}q_{ik}\right)^2},\quad i\neq i' \tag{I7}$$

$$-\frac{\partial^2 E\left[l(\Theta;D^c)\mid D^m,D^v\right]}{\partial q_{ik}\partial q_{i'k'}}=0,\quad i=i'\text{ or }i\neq i'\text{ and }k\neq k' \tag{I8}$$

Appendix II

Tracing the Convergence of the EM Algorithm in Application 3.1

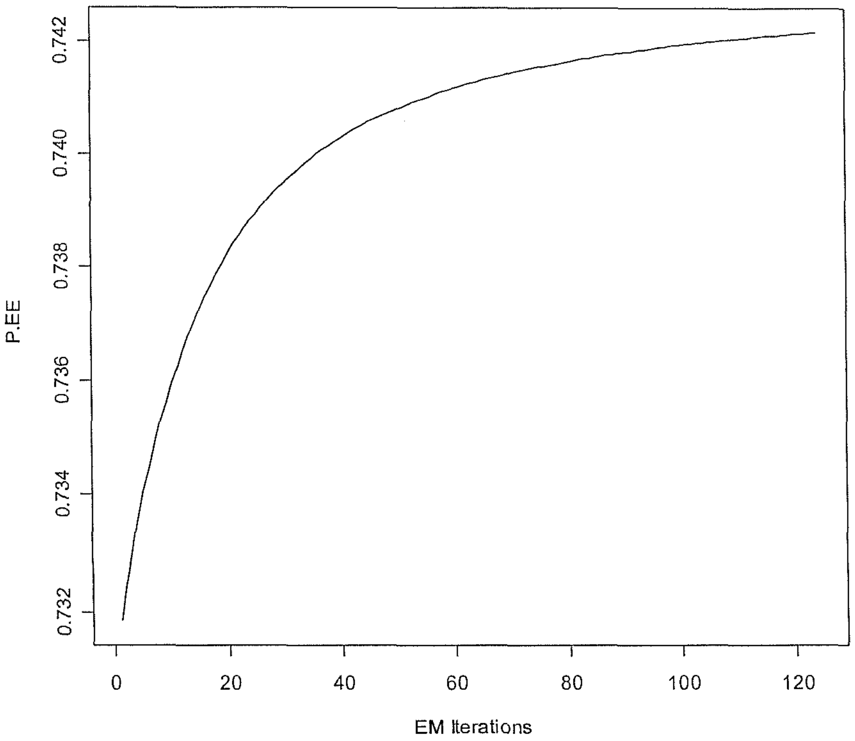


Figure II.1: Tracing the convergence of the EM algorithm for the EE flow. Starting values close to the maximum likelihood point, convergence criterion 0.00001.

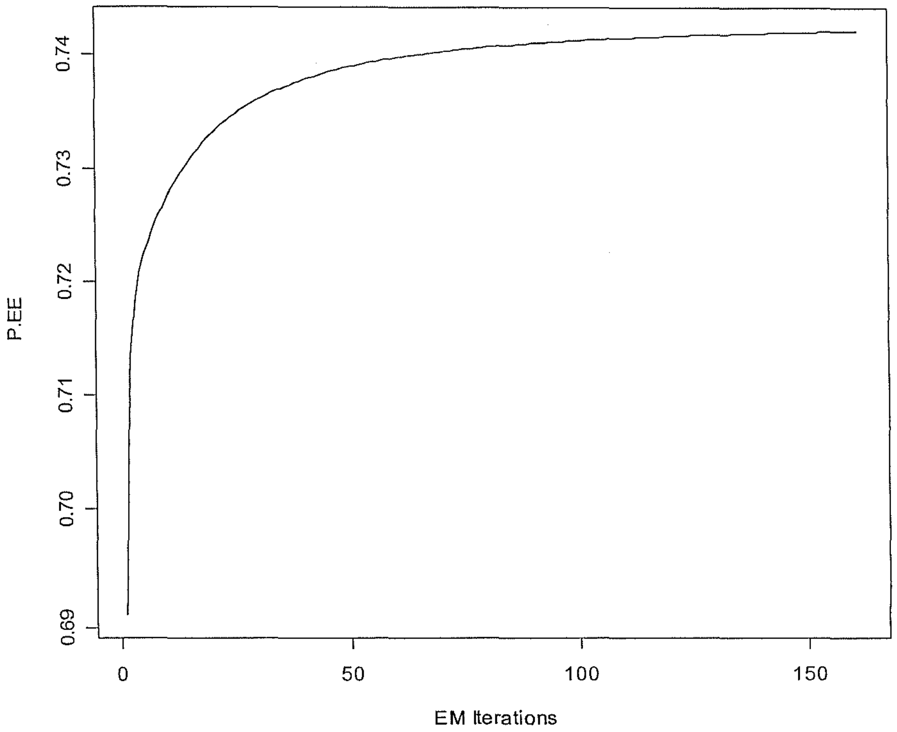


Figure II.2: Tracing the convergence of the EM algorithm for the EE flow. Starting values further from the maximum likelihood point, convergence criterion 0.00001.

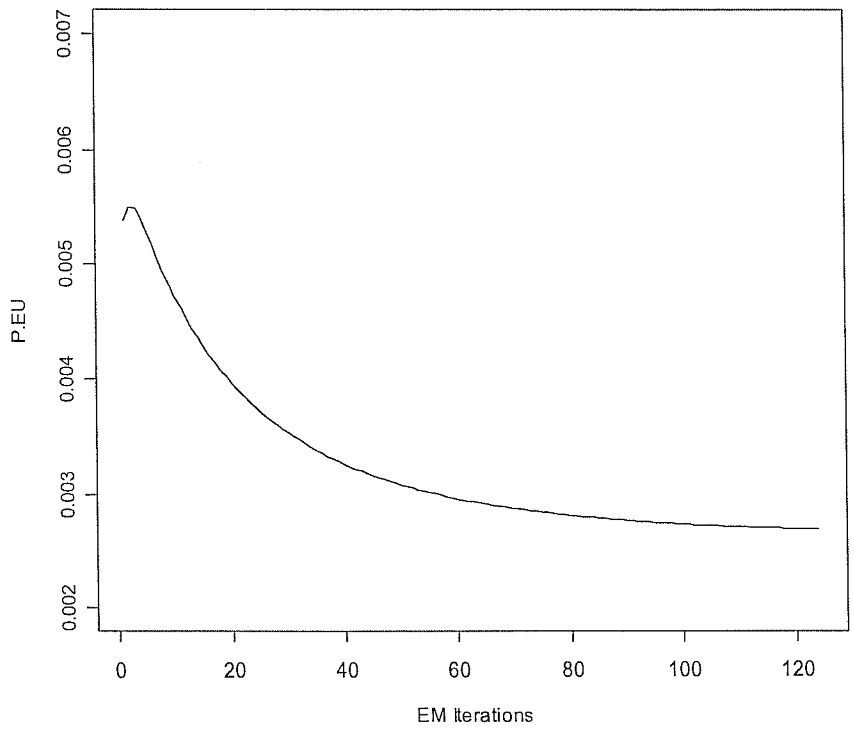


Figure II.3: Tracing the convergence of the EM algorithm for the EU flow. Starting values close to the maximum likelihood point, convergence criterion 0.00001.

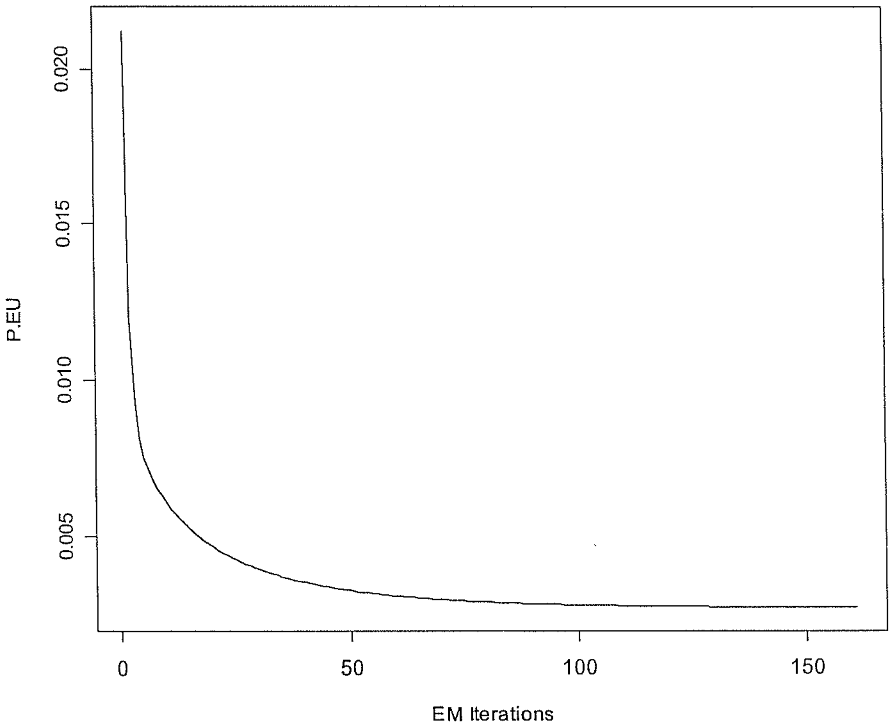


Figure II.4: Tracing the convergence of the EM algorithm for the EU flow. Starting values further from the maximum likelihood point, convergence criterion 0.00001.

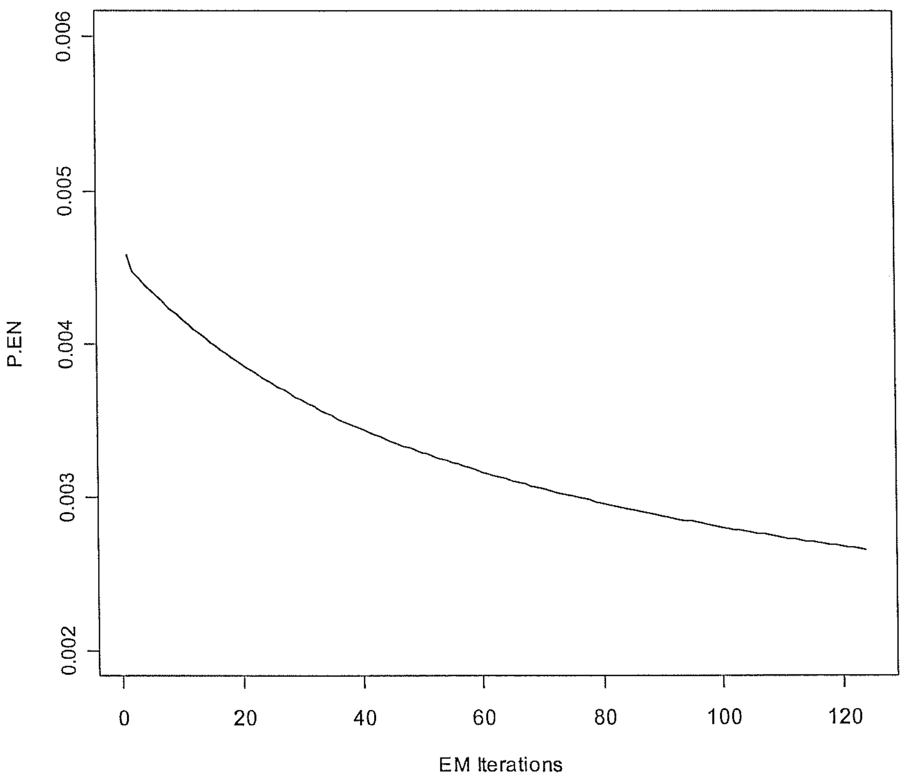


Figure II.5: Tracing the convergence of the EM algorithm for the EN flow. Starting values close to the maximum likelihood point, convergence criterion 0.00001.

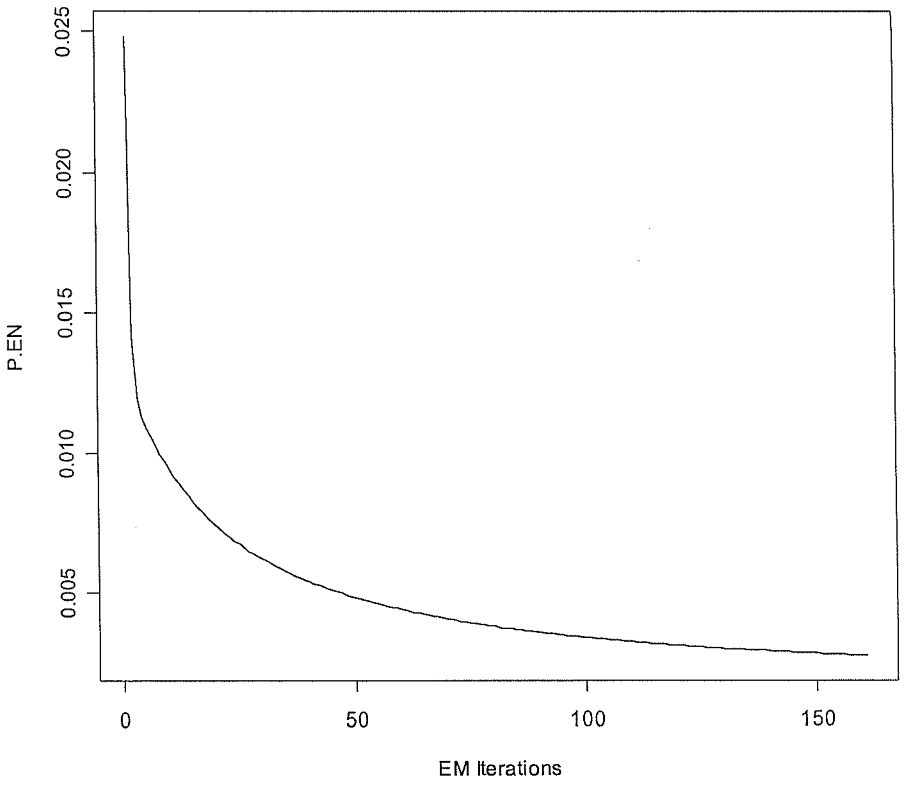


Figure II.6: Tracing the convergence of the EM algorithm for the EN flow. Starting values further from the maximum likelihood point, convergence criterion 0.00001.

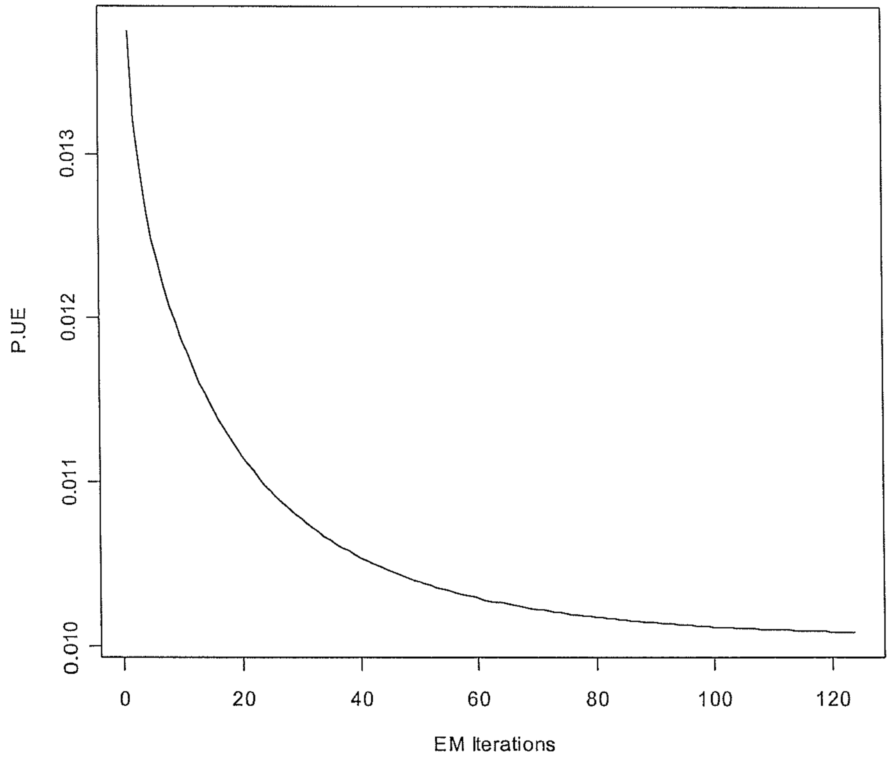


Figure II.7: Tracing the convergence of the EM algorithm for the UE flow. Starting values close to the maximum likelihood point, convergence criterion 0.00001.

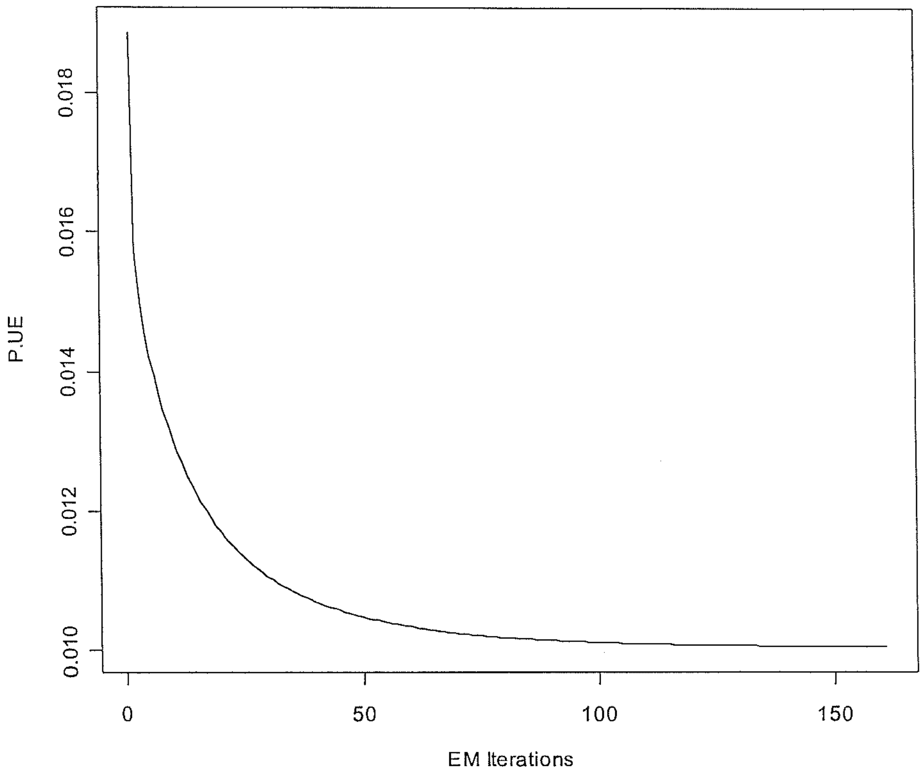


Figure II.8: Tracing the convergence of the EM algorithm for the UE flow. Starting values further from the maximum likelihood point, convergence criterion 0.00001.

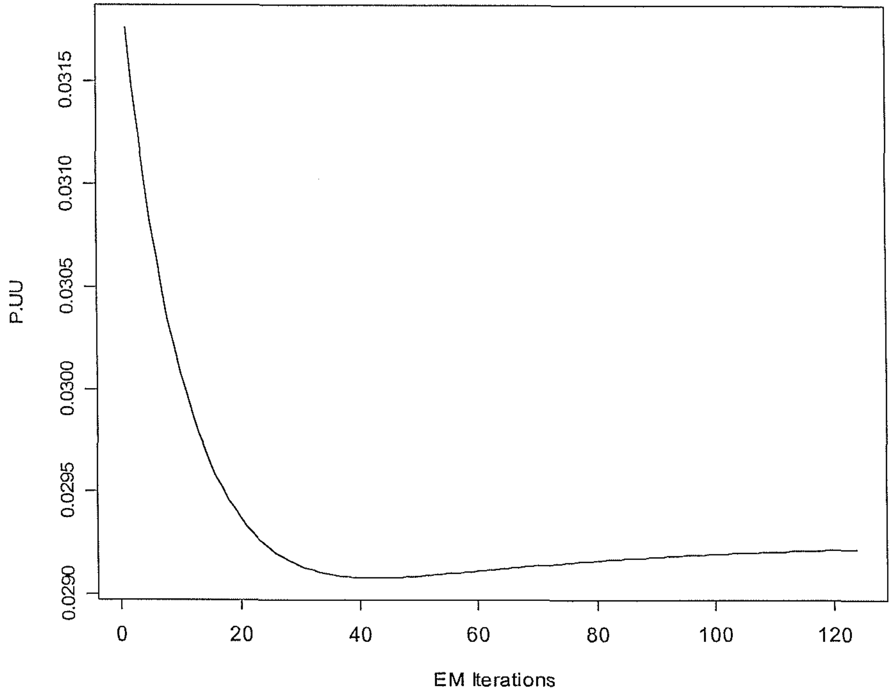


Figure II.9: Tracing the convergence of the EM algorithm for the UU flow. Starting values close to the maximum likelihood point, convergence criterion 0.00001.

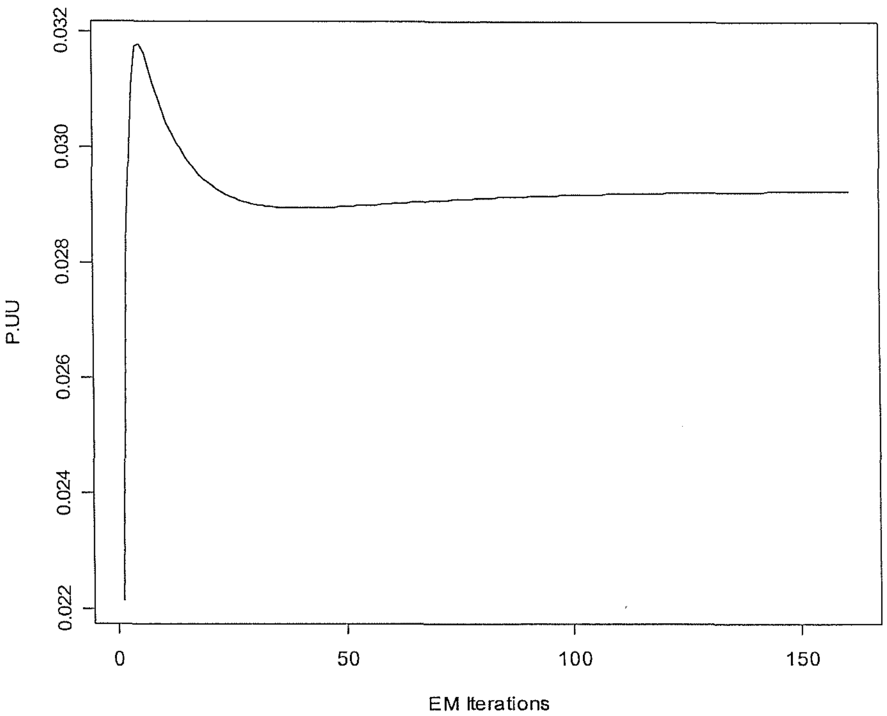


Figure II.10: Tracing the convergence of the EM algorithm for the UU flow. Starting values further from the maximum likelihood point, convergence criterion 0.00001.

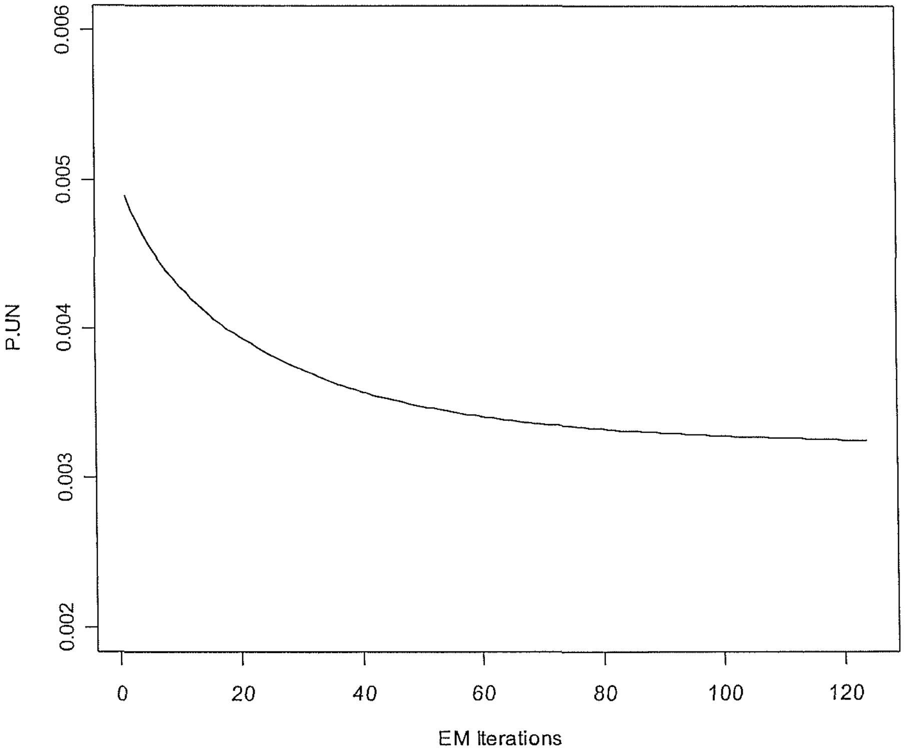


Figure II.11: Tracing the convergence of the EM algorithm for the UN flow. Starting values close to the maximum likelihood point, convergence criterion 0.00001.

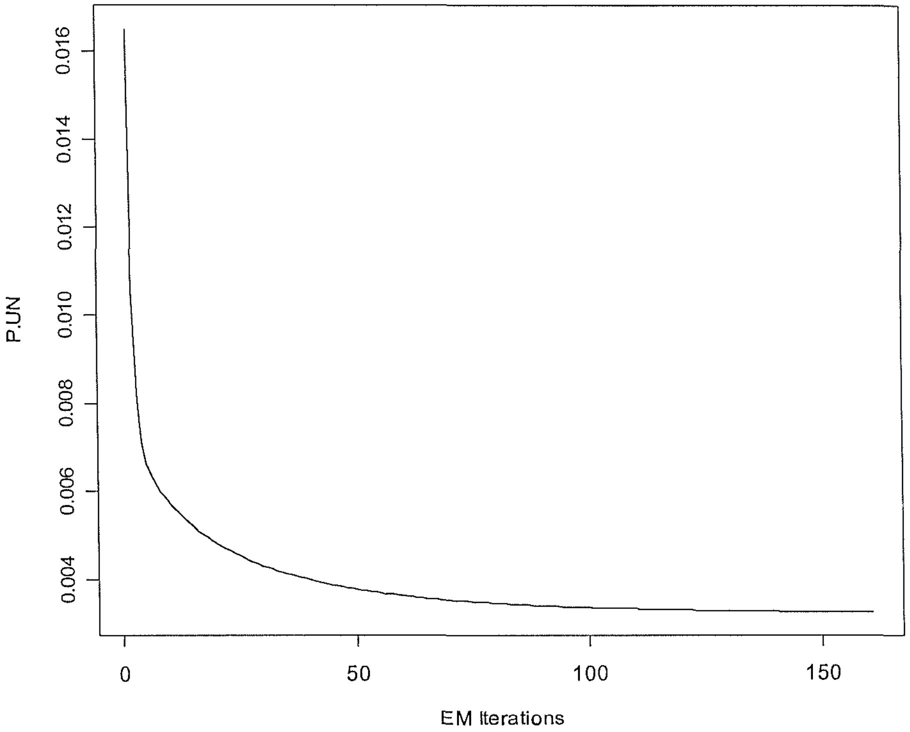


Figure II.12: Tracing the convergence of the EM algorithm for the UN flow. Starting values further from the maximum likelihood point, convergence criterion 0.00001.

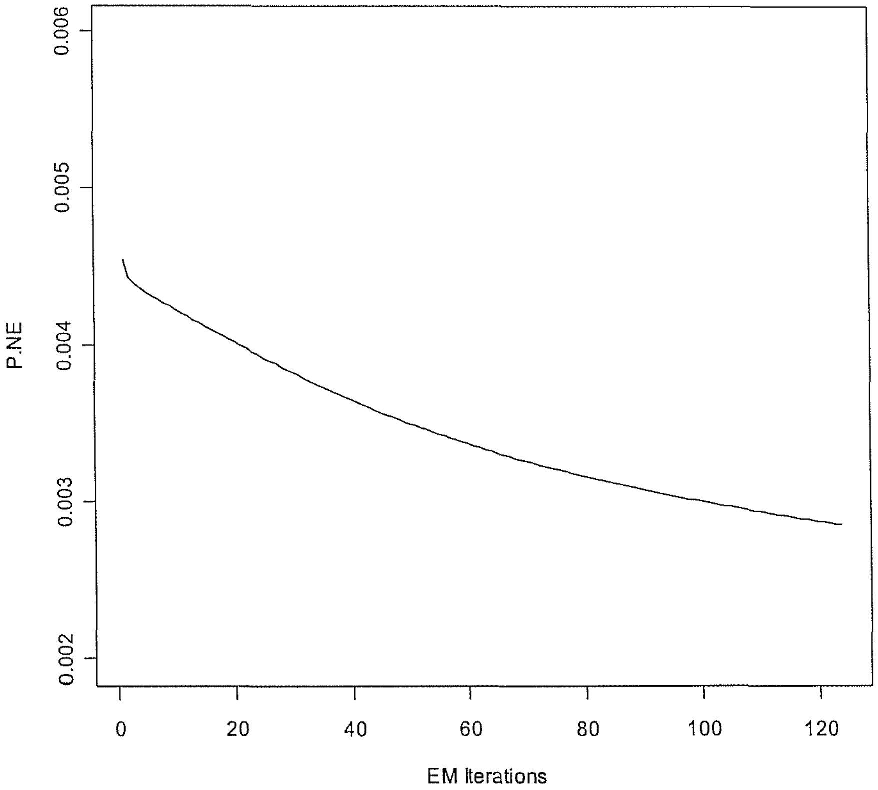


Figure II.13: Tracing the convergence of the EM algorithm for the NE flow. Starting values close to the maximum likelihood point, convergence criterion 0.00001.

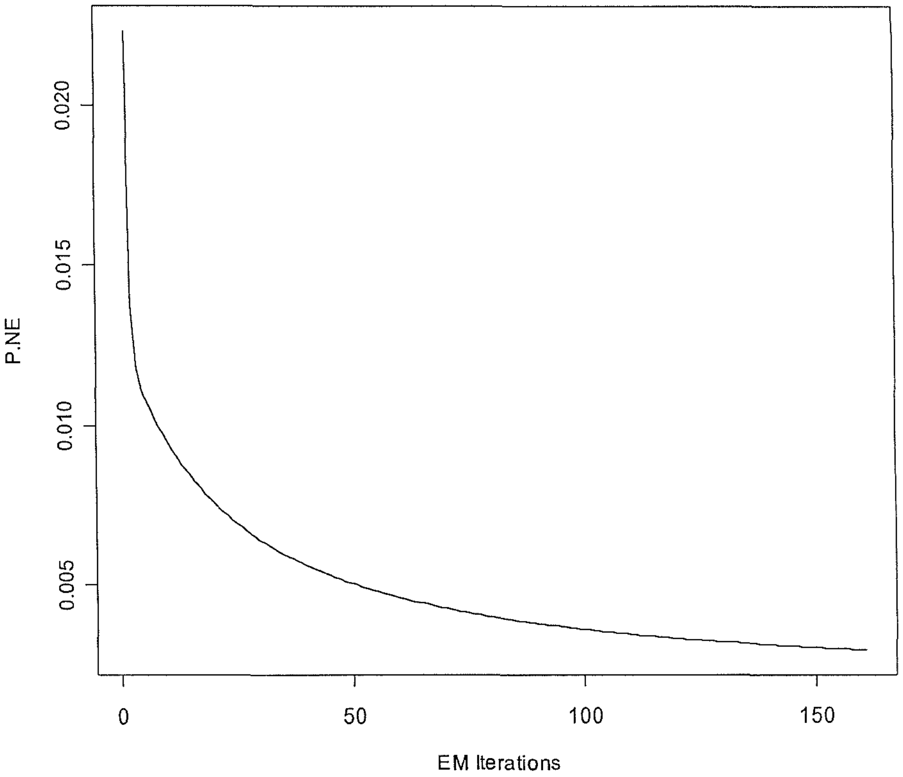


Figure II.14: Tracing the convergence of the EM algorithm for the NE flow. Starting values further from the maximum likelihood point, convergence criterion 0.00001.

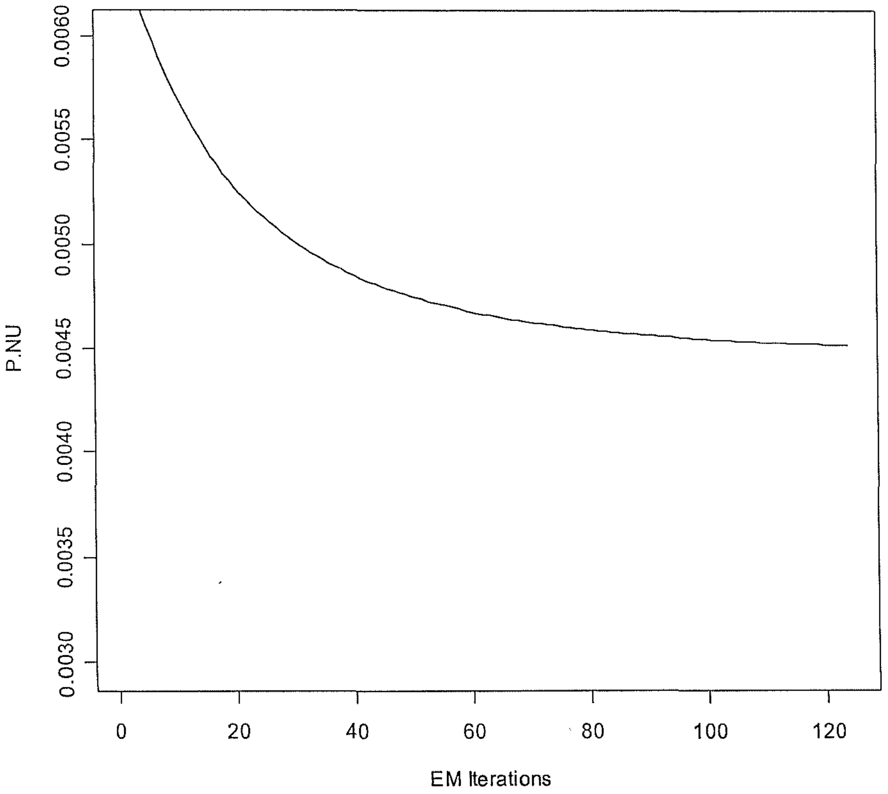


Figure II.15: Tracing the convergence of the EM algorithm for the NU flow. Starting values close to the maximum likelihood point, convergence criterion 0.00001.

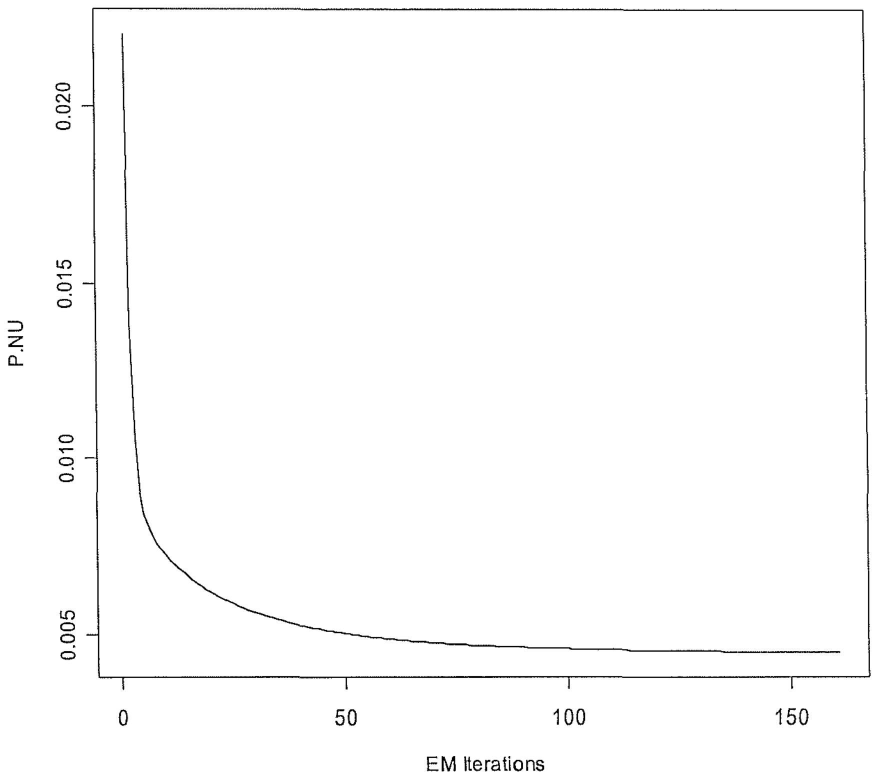


Figure II.16: Tracing the convergence of the EM algorithm for the NU flow. Starting values further from the maximum likelihood point, convergence criterion 0.00001.

Appendix III

First and Second Order Derivatives Involved in the Application of the Missing Information Principle in a Longitudinal Framework

$$\frac{\partial E\left[l(\Theta;D^c) \mid D^m,D^v\right]}{\partial P_i} = \frac{E\left[\left(n_{i\bullet}^{(*)} + n_{i\bullet}^{v(*)}\right) \mid D^m,D^v,\Theta\right]}{P_i} - \frac{E\left[\left(n_{r^2\bullet}^{(*)} + n_{r^2\bullet}^{v(*)}\right) \mid D^m,D^v,\Theta\right]}{1-\sum_{i=1}^{r^2-1} P_i} \qquad \text{(III1)}$$

$$-\frac{\partial^2 E\left[l(\Theta;D^c) \mid D^m,D^v\right]}{\partial P_i^2} = \frac{E\left[\left(n_{i\bullet}^{(*)} + n_{i\bullet}^{v(*)}\right) \mid D^m,D^v,\Theta\right]}{P_i^2} + \frac{E\left[\left(n_{r^2\bullet}^{(*)} + n_{r^2\bullet}^{v(*)}\right) \mid D^m,D^v,\Theta\right]}{\left(1-\sum_{i=1}^{r^2-1} P_i\right)^2} \qquad \text{(III2)}$$

$$-\frac{\partial^2 E\left[l(\Theta;D^c) \mid D^m,D^v\right]}{\partial P_i \partial P_{i'}} = \frac{E\left[\left(n_{r^2\bullet}^{(*)} + n_{r^2\bullet}^{v(*)}\right) \mid D^m,D^v,\Theta\right]}{\left(1-\sum_{i=1}^{r^2-1} P_i\right)^2} \qquad \text{(III3)}$$

$$-\frac{\partial^2 E\left[l(\Theta;D^c) \mid D^m,D^v\right]}{\partial P_i \partial q_{ij}} = 0 \qquad \text{(III4)}$$

The first and second order derivatives with respect to the misclassification parameters can be also computed analytically but due to the introduction of the ICE assumption this involves more complicated expressions.

Appendix IV

Misclassification Probabilities Employed in Simulation Studies

Matrix of Misclassification Probabilities Used in Simulation Studies I-III and VIII

$$Q = \begin{array}{c|ccc|} & E & U & N \\ E & 0.981 & 0.016 & 0.036 \\ U & 0.008 & 0.950 & 0.023 \\ N & 0.011 & 0.034 & 0.941 \end{array}$$

Matrix of Misclassification Probabilities Used in Simulation Studies IV-V

$$Q = \begin{array}{c|ccc|} & E & U & N \\ E & 0.99 & 0.01 & 0.015 \\ U & 0.004 & 0.97 & 0.015 \\ N & 0.006 & 0.02 & 0.97 \end{array}$$

Matrices of Misclassification Probabilities Used in Simulation Studies VI-VII

$$Q_{Males} = \begin{array}{c|ccc|} & E & U & N \\ E & 0.991 & 0.01 & 0.015 \\ U & 0.005 & 0.98 & 0.015 \\ N & 0.004 & 0.01 & 0.97 \end{array}$$

$$Q_{Females} = \begin{array}{c|ccc|} & E & U & N \\ E & 0.981 & 0.01 & 0.005 \\ U & 0.004 & 0.97 & 0.005 \\ N & 0.015 & 0.02 & 0.99 \end{array}$$

Table IV.1: Sets of misclassification probabilities at $t+1$ used in simulation studies I-III under the different scenarios for the longitudinal measurement error mechanism. The relaxed-ICE scenarios are based on Kristiansson's (1983) proposal

Misclassification Probability	*ICE*	*Relaxed-ICE 1*	*Relaxed-ICE 2*
q_{EEEE}	0.981	0.983	0.985
q_{EUEE}	0.008	0.007	0.0063
q_{ENEE}	0.011	0.010	0.0087
q_{UEEE}	0.981	0.979	0.976
q_{UUEE}	0.008	0.0088	0.01
q_{UNEE}	0.011	0.0122	0.014
q_{NEEE}	0.981	0.979	0.976
q_{NUEE}	0.008	0.0088	0.01
q_{NNEE}	0.011	0.0122	0.014
q_{EEUU}	0.016	0.0166	0.0188
q_{EUUU}	0.95	0.9444	0.9375
q_{ENUU}	0.034	0.039	0.0437
q_{UEUU}	0.016	0.0135	0.012
q_{UUUU}	0.95	0.955	0.96
q_{UNUU}	0.034	0.0315	0.028
q_{NEUU}	0.016	0.0166	0.0188
q_{NUUU}	0.95	0.9444	0.9375
q_{NNUU}	0.034	0.039	0.0437
q_{EENN}	0.036	0.041	0.0463
q_{EUNN}	0.023	0.025	0.0287
q_{ENNN}	0.941	0.934	0.9250
q_{UENN}	0.036	0.041	0.0463
q_{UUNN}	0.023	0.025	0.0287
q_{UNNN}	0.941	0.934	0.9250
q_{NENN}	0.036	0.033	0.030
q_{NUNN}	0.023	0.021	0.018
q_{NNNN}	0.941	0.946	0.952

Table IV.2: Sets of misclassification probabilities at $t+1$ used in simulation studies IV-V under the different scenarios for the longitudinal measurement error mechanism. The relaxed-ICE scenario is based on Kristiansson's (1983) proposal

Misclassification Probability	*ICE*	*Relaxed-ICE*
q_{EEEE}	0.99	0.995
q_{EUEE}	0.004	0.0025
q_{ENEE}	0.006	0.0025
q_{UEEE}	0.99	0.98
q_{UUEE}	0.004	0.01
q_{UNEE}	0.006	0.01
q_{NEEE}	0.99	0.98
q_{NUEE}	0.004	0.01
q_{NNEE}	0.006	0.01
q_{EEUU}	0.01	0.015
q_{EUUU}	0.97	0.96
q_{ENUU}	0.02	0.025
q_{UEUU}	0.01	0.01
q_{UUUU}	0.97	0.98
q_{UNUU}	0.02	0.01
q_{NEUU}	0.01	0.015
q_{NUUU}	0.97	0.96
q_{NNUU}	0.02	0.025
q_{EENN}	0.015	0.02
q_{EUNN}	0.015	0.02
q_{ENNN}	0.97	0.96
q_{UENN}	0.015	0.02
q_{UUNN}	0.015	0.02
q_{UNNN}	0.97	0.96
q_{NENN}	0.015	0.01
q_{NUNN}	0.015	0.01
q_{NNNN}	0.97	0.98

References

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

Agresti, A. (1990). Categorical Data Analysis, Wiley.

Akerlof, G. and Main, B. (1980). Unemployment Spells and Unemployment Experience, *American Economic Review*, **70**, 885-893.

Assakul, K. and Proctor, C.H. (1967). Testing Independence in Two-way Contingency Tables with Data Subject to Misclassification, *Psychometrica*, **32**, 67-76.

Attkinson, B.A. and Mickelwright, J. (1991). Unemployment Compensation and Labour Market Transitions: A Critical Review, *Journal of Economic Literature*, **XXIX**, 1679-1727.

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

Bailar, A.B. (1975). The Effects of Rotation Group Bias on Estimates from Panel Surveys, *Journal of the American Statistical Association*, **70**, 23-30.

Bailar, A.B. (1987). Nonsampling Errors, *Journal of Official Statistics*, **3**, 323-325.

Barkume, J.A., and Hovarth, W.F. (1995). Using Gross Flows to Explore Movements in the Labour Force, *Monthly Labour Review (April 1995)*, 28-35.

Barnes, R. (1987). Non-sampling Errors: Some Approaches Adopted in Major Government Surveys in Britain, *Journal of Official Statistics*, **3**, 329-333.

Bassi, F., Torelli, N. and Trivellato, U. (1998). Data and Modelling Strategies in Estimating Labour Force Gross Flows Affected by Classification Errors, *Survey Methodology*, **24(2)**, 109-122.

Bassi, F. and Trivellato, U. (2000). Gross Flows from the French Labour Force Survey: A Re-analysis, *Proceedings of the Conference on Methodological Issues in Official Statistics*, October 12-13 2000, Stockholm.

Benjamin, B. and Pollard, J.M. (1986). The Analysis of Mortality and Other Actuarial Statistics, 2^{nd} Edition, Heinemann.

Biemer, P.P. and Forsman, G. (1992). On the Quality of Re-interview Data with Application to the Current Population Survey, *Journal of the American Statistical Association*, **87**, 915-923.

Biemer, P.P. and Trewin, D. (1997). A Review of Measurement Error Effects on the Analysis of Survey Data, in *Survey Measurement and Process Quality, Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, Trewin (eds)*, 603-632 Wiley.

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.

Buell, P. and Dunn, J.E. (1964). The Dilution Effect of Misclassification, *American Journal of Public Health*, **54**, 598-602.

Burda, M. and Wyplosz, C. (1994). Gross Worker and Job Flows in Europe, *European Economic Review*, **38**, 1287-1315.

Caroll, 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.

Chen, T.T. and Fienberg, S. (1976). The Analysis of Contingency Tables with Incompletely Classified Data, *Biometrics*, **32**, 133-144.

Chen, T.T. (1979). Log-linear Models for Categorical Data with Misclassification and Double Sampling, *Journal of the American Statistical Association*, **74**, 481-488.

Chiacchierini, P.R. and Arnold, C.J. (1977). A Two-Sample Test for Independence in 2×2 Tables with Both Margins Subject to Misclassification, *Journal of the American Statistical Association*, **72**, 170-174.

Chua, T.C. and Fuller, W.A. (1987). A Model for Multinomial Response Error, *Journal of the American Statistical Association*, **82**, 46-51.

Clarke, P.S. and Tate, P. (1999). Methodological Development of a Proposed Procedure to Adjust Longitudinal Data from the Labour Force Survey for Non-Response Bias, *Unpublished Manuscript.*

Clarke, P.S. and Chambers, R.L. (1998). Estimating Gross Flows from Household Based Surveys Subject to Non-ignorable Non-response, *Survey Methodology*, **24(2)**, 123-129.

Cochran, G.W. (1963). Errors of Measurement in Statistics, *Technometrics*, **10**, 637-666.

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.

Deming, W.E. (1944). On Errors in Surveys, *American Sociological Review*, **9**, 359-369.

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 Unconstrianed Optimisation and Nonlinear Equations, Precentice-Hall Series in Computational Mathematics.

Deville, J.C. and Sarndal, C.E. (1992). Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, **87**, 376-382.

Duncan, G. (2000). Using Panel Studies to Understand Household Behaviour and Well-being, *in Researching Social and Economic Changes - The Uses of Household Panel Surveys, David Rose (eds),* 54-75, Routledge.

Efron, B. and Hinkley, V.D. (1978). Assessing the Accuracy of the Maximum Likelihood Estimator: Observed versus Expected Fisher Information, *Biometrika*, **65**, 457-487.

Ekholm, A. and Palmgren, J. (1987). Correction for Misclassification Using Doubly Sampled Data, *Journal of Official Statistics*, **3**, 419-429.

Elliot, D. (1997). Software to Weight and Gross Survey Data, *Government Statistical Service Methodology Series*, **1,** Office for National Statistics, UK.

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.

Evans, M.J. (1985). Gross Flows Statistics Outside North America: Construction and Use, *Proceedings of the Conference on Gross Flows in Labour Force Statistics*, 119-129, Washington DC, US Government Printing Office.

Fitzpatrick, S. and Scott, A. (1987). Quick Simultaneous Confidence Intervals for Multinomial Proportions, *Journal of the American Statistical Association*, **82**, 875-878.

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.

Fuller, W.A. (1987). Measurement Error Models, Wiley.

Gong, G. and Samaniengo, F.J. (1981). Pseudo Maximum Likelihood Estimation: Theory and Applications, *Annals of Statistics*, **9**, 861-869.

Goodman, A.L. (1964). Simultaneous Confidence Intervals for Contrasts among Multinomial Populations, *Annals of Mathematical Statistics*, **35**, 716-725.

Goodman, A.L. (1965). On Simultaneous Confidence Intervals for Multinomial Proportions, *Technometrics*, **7**, 247-254.

Goodman, A.L. (1974). Exploratory Latent Structure Analysis Using both Identifiable and Unidentifiable Models, *Biometrika*, **61**, 215-231.

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

Haberman, S.J. (1979). Analysis of Qualitative Data, **Vol 2**, New York Academic Press.

Hagenaars, J.A. (1985). LCAG-Log-linear Modelling with Latent Variables, A Modified LISREL Approach, *Proceedings of the Conference on Methodological Research*, Amsterdam.

Hagenaars, J.A. (1990). Categorical Longitudinal Data: Log-linear, Panel, Trend and Cohort Analysis, Sage.

Hansen, H.M., Hurwitz, N.W., Marks, S.E. and Mauldin, W.P. (1951). Response Errors in Surveys, *Journal of the American Statistical Association*, **46**, 147-190.

Hansen, H.M., Hurwitz, N.W., Nisselson, H. and Steinberg, J. (1955). The Redesign of the Census Current Population Survey, *Journal of the American Statistical Association*, **50**, 701-719.

Hansen, W.L (1961). The Cyclical Sensitivity of the Labour Supply, *American Economic Review*, **51**, 299-309.

Harville, A.D. (1997). Matrix Algebra from a Statistician's Perspective, Springer.

Heady, P., Smith, S., and Avery, V. (1991). Census Validation Survey, *Quality Report*, Office for National Statistics, UK.

Heyde, C.C. (1997). Quasi-Likelihood and Its Application, A General Approach to Optimal Parameter Estimation, Springer Series in Statistics.

Hilasky, H.J. (1968). The Status of Research on Gross Changes in Labour Force, *Employment and Earnings*, **14**, 6-13.

Hill, D. (1987). Response Errors Around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods, *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 210-215.

Hill, M.S. (1992). The Panel Study of Income Dynamics, A User's Guide to Major Science Data Bases, **2**, Sage.

Hochberg, Y. (1977). On the Use of Double Sampling Schemes in Analysing Categorical Data with Misclassification Errors, *Journal of the American Statistical Association*, **72**, 914-921.

Hoem, M.J. (1985). Weighting, Misclassification, and Other Issues in the Analysis of Survey Samples of Life Histories, in *Longitudinal Analysis of Labour Market Data*, H*eckman and Singer (eds)*, 249-293, Cambridge University Press.

Hogue, C.R. (1985). History and Problems Encountered in Estimating Gross Flows, *Proceedings of the Conference on Gross Flows in Labour Force Statistics*, 1-5, Washington DC, US Government Printing Office.

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.

Humphreys, K. (1996). Latent Variable Models for Discrete Longitudinal Data with Measurement Error, *Ph.D. Thesis*, Department of Social Statistics, University of Southampton, UK.

Humphreys, K. and Skinner, C.J. (1997). Instrumental Variable Estimation of Gross Flows in the Presence of Measurement Error, *Survey Methodology*, **23(1)**, 53-60.

Jones, R.G.S. and Riddell, C.W. (1998). Gross Flows of Labour in Canada and the United States, *Canadian Public Policy*, **XXIV**, 103-120.

Kalton, G. and Miller, M.W. (1991). The Seam Effect with Social Security Income in the Survey of Income and Programme Participation, *Journal of Official Statistics*, **7**, 235-245.

Kemsley, W.F.F. (1961). The Household Expenditure Enquiry of the Ministry of Labour, *Applied Statistics*, **10**, 117-135.

Kish, L. (1965). Survey Sampling, Wiley.

Koch, G.G. (1969). The Effect of Non-Sampling Errors on Measures of Association in 2×2 Contingency Tables, *Journal of the American Statistical Association*, **64**, 852-863.

Kristianson, K-E. (1983). Gross Flows Estimates in the Swedish Labour Force Surveys, *Paper Prepared for the Meeting on Manpower Statistics*, Geneva, May 1983.

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

Kristiansson, K-E. and Mirza, H. (2000). Estimation of Gross Flows in Swedish LFS: A Rough Outline, *Internal Report*, 1-7, Statistics Sweden.

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

Kuo, L. (1989). Composite Estimation of Totals for Livestock Surveys, *Journal of the American Statistical Association*, **84**, 421-429.

Laux, R. and Tonks, E. (1996). Longitudinal Data from the Labour Force Survey, *Methods and Quality Papers*, 1-45, Office for National Statistics, UK.

Lazarsfeld, P.F. and Henry, N.W. (1968). Latent Structure Analysis, Houghton-Mifflin.

Lemaitre, G. (1994). Current Statistics on Labour Dynamics, *Internal Report*, 1-28 OECD.

Lessler, T.J. and Kalsbeek, D.W. (1992). Nonsampling Errors in Surveys, 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.

Magnac, T. and Visser, M. (1999). Transition Models with Measurement Errors, *The Review of Economics and Statistics*, **81(3)**, 466-474.

Marquis, K.H. and Moore, J.C. (1990). Measurement Errors in the Survey of Income and Program Participation (SIPP) Program Reports, *Proceedings of the Annual Research Conference*, 721-745, US Bureau of the Census.

Marshall, R.J. (1990). Validation Study Methods for Estimating Exposure Proportions and Odds Ratios with Misclassified Data, *Journal of Clinical Epidemiology*, **43**, 95-109.

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.

Moser, C.A. and Kalton, G. (1972). Survey Methods in Social Investigation, Basic Books.

Mote, V.L. and Anderson, R.L. (1965). An Investigation of the Effect of Misclassification on the Properties of Chi-Square Tests in the Analysis of Categorical Data, *Biometrika*, **52**, 95-109.

O'Muircheartaigh, C. (1986). Correlates of Re-interview Response Inconsistency in the Current Population Survey, *Proceedings of the Annual Research Conference*, 208-234, US Bureau of the Census.

O'Muircheartaigh, C. (1991). Simple Response Variance: Estimation and Determinants, in *Measurement Error in Surveys, P.P. Biemer, R.M. Grooves, L.E. Lyberg, N.A. Mathiowetz, S. Sudman (eds)*, 551-574, Wiley.

ONS (1997). JUVOS Matching Exercise, *Internal Report*, Social Survey Division, Office for National Statistics, UK.

ONS (1999). Labour Force Survey, *User Guide*, **Vol 1** (Background and Methodology), Office for National Statistics, UK.

ONS (2000). Guide to LFS Data, *Internal Report*, 1-7, Office for National Statistics, UK.

Palmer, L.G. (1943). Factors in the Variability of Response in Enumerative Studies, *Journal of the American Statistical Association*, **38**, 143-152.

Pawitan, Y. (2001). In All Likelihood: Statistical Modelling and Inference Using Likelihood, Oxford University Press.

Pfeffermann, D. (1988). The Effect of Sampling Design and Response Mechanism on Multivariate Regression-Based Predictors, *Journal of the American Statistical Association*, **83**, 824-833.

Pfeffermann, D., Skinner, C.J. and Humphreys, K. (1998). The Estimation of Gross Flows in the Presence of Measurement Error Using Auxiliary Information, *Journal of the Royal Statistical Society Series A*, **161**, 13-22.

Pfeffermann, D. and Tsibel, N. (1998). Estimation of Gross Flows from Complex Surveys, Adjusting for Missing Data, Classification Errors and Informative Sampling, *Proceedings of the Statistics Canada Symposium on Longitudinal Data Analysis for Complex Surveys*, 75-83.

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

Poterba, J.M. and Summers, L.H. (1995). Unemployment Benefits and Labour Market Transitions: A Multinomial Logit Model with Errors in Classification, *The Review of Economics and Statistics*, **LXXVII**, 207-216.

Quesenberry, P.C. and Hurst, C.D. (1964). Large Sample Simultaneous Confidence Intervals for Multinomial Proportions, *Technometrics*, **6**, 191-195.

Raj, D. (1968). Sampling Theory, McGraw-Hill.

Rogot, E. (1961). A Note on Measurement Errors and Detecting Real Differences, *Journal of the American Statistical Association*, **56**, 314-319.

Schreiner, I. (1980). Re-interview Results from the CPS Independent Reconciliation Experiment, *Internal Report,* US Bureau of the Census.

Selen, J. (1986). Adjusting for Errors in Classification and Measurement in the Analysis of Partly and Purely Categorical, *Journal of the American Statistical Association*, **81**, 75-81.

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.J. (1989). Domain Means, Regression and Multivariate Analysis, in *Analysis of Complex Surveys Skinner, C. J., Holt, D. and Smith, T. M. F. (eds)*, 59-87**,** Wiley.

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

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.

Smith, R.E. and Vansky, J. (1979). Gross Change Data: The Neglected Data Base, in *Report of the National Commission of Employment and Unemployment*, Appendix **II**, Washington DC, US Government Printing Office.

Solon, G. (1985). Effects of Rotation Group Bias on Estimation of Unemployment, *Proceedings of the Conference on Gross Flows in Labour Force Statistics*, 15-21, Washington DC, US Government Printing Office.

Stasny, A.E. and Fienberg, S. (1985). Stochastic Models for Estimating Gross Flows in the Presence of Non-random Non-response, *Proceedings of the Conference on Gross Flows in Labour Force Statistics*, 25-39, Washington DC, US Government Printing Office.

Stasny, A.E. (1986). Estimating Gross Flows Using Panel Data with Non-response: An Example From the Canadian Labour Force Survey, *Journal of the American Statistical Association*, **81**, 42-47.

Statistics Canada (1979). Estimation of Labour Force Gross Flows from the Canadian Labour Force Survey, *Working Paper*, Statistics Canada.

Swires-Hennessy, E. and Thomas, G.W. (1987). The Good, the Bad and the Ugly: Multiple Stratified Sampling in the 1986 Welsh House Condition Survey, *Statistical News*, **79**, 24-26.

Tam, S. M. (1985). On Labour Force Estimators, *Applied Statistics*, **34(3)**, 264-272.

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

Tate, P.F. (1997). Utilising Longitudinally Linked Data From the Labour Force Survey, *Paper Presented to the Labour Market Statistics User Group Semina*r, 8 July 1997, Office for National Statistics, UK.

Tate, P.F. and Clarke, P.S. (1999). Methodological Issues in the Production and Analysis of Longitudinal Data From the Labour Force Survey, *Government Statistical Service Methodology Series*, **17**, Office for National Statistics, UK.

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

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

Trewin, D. (1987). How Do We Reduce Non-sampling Errors?, Journal *of Official Statistics*, **3**, 343-347.

Turner, R. (1961). Inter-week Variations in Expenditure Recorded During a Two Week Survey of Family Expenditure, *Applied Statistics*, **10**, 136-146.

US Bureau of the Census (1963). The Current Population Survey Re-interview Program, Some Notes and Discussions, *Technical Paper*, **6**, Washington DC, US Government Printing Office.

Van den Hout, A. and Van der Heijden, M.G. (2002). Randomised Response, Statistical Disclosure Control and Misclassification: A Review, *International Statistical Review*, **70**, 269-288.

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.

Veevers, R. and Macredie, I. (1983). Estimating Gross Flows from the Canadian Labour Force Survey, *Internal Report*, 1-13, Statistics Canada.

Wedderburn, R.W.M. (1974). Quasi-likelihood Functions, Generalised Linear Models and Gauss-Newton Method, *Biometrika*, **61**, 439-447.

White, E. (1986). The Effects of Misclassification of Disease Status in Follow-up Studies: Implication for Selecting Disease Classification Criteria, *American Journal of Epidemiology*, **124**, 816-825.

Wolter, M.K. (1979). Composite Estimation in Finite Populations, *Journal of the American Statistical Association*, **74**, 604-613.

Wong, F. (1983). A Technique to Correct the Response Bias in the 4×4 Labour Force Gross Flow Matrix, *Technical Report*, Statistics Canada.

Woodbury, M.A. (1977). Discussion of Paper by Hartley and Hocking, *Biometrics*, **27**, 808-817.

Zarcovich, S.S. (1966). Quality of Statistical Data, Food and Agricultural Organisation of the United Nations.