Estimation strategies in the presence of non-coverage in the German Microcensus-Panel: An evaluation using SOEP data *

Edin Basic Ulrich Rendtel

Dezember 2005

Abstract

The German Microcensus is a rotating panel with units staying in the survey for four observations. Because of the very large sample size and the mandatory participation it appears to be a valuable data base for short duration analysis. However, the German Microcensus uses area sampling that does not consider participants who left the area. Consequently, there is no information on participants after they moved. We propose two different estimation strategies in the presence of missing data caused by residential movers. To assess their performance we use the German Socio-Economic Panel. The example we analyze is the estimation of labour force flows. The results indicate that both estimation strategies lead to estimates that are less biased.

KEYWORDS: Panel survey, labour market analysis, residential mobility, non-coverage bias, log-linear modelling, inverse probability weighting. JEL C81, J69

*This work is a part of the "MC-Panel" project, sponsored by the Deutsche Forschungsgemeinschaft (DFG) under contract number RE 1445/1-1. Correspondence address: ebasic@wiwiss.fu-berlin.de

1 Introduction

The German Microcensus (MC) is a 1% survey on households carried out by the German Statistical Office. The primary goal of the MC is to collect information about the population structure, labour market behaviour and the housing situation. It has been conducted on a yearly basis, with each sample household retained for four consecutive years and one fourth of the sample replaced each year. The MC was designed to produce cross-sectional data, but it can also produce longitudinal data by linking together the data on each individual across years, see Heidenreich (2002). Thus, it has the potential to provide data for short duration analysis, the maximum length of longitudinal information on one individual being four time points. Furthermore, the MC is characterized by mandatory participation. This feature reduces the nonresponse to a minimum level. However, a methodological problem of the longitudinal use of the MC arises from the fact that residential movers are not traced in the MC. This is due to the fact that the MC uses area sampling, where the dwellings are sampled and residential movers are not followed to their new homes. Instead, new persons who move into the dwellings of the residential movers enter the MC sample. The missing information about the mobile persons in the MC might lead to some systematic non-coverage bias in the analysis of interest. For example, if we are interested in changes from unemployment to employment, then a move to a different place may be prompted by a new job. However, due to the lack of corresponding data for residential movers, the influence of current changes in the labour market status on mobility behaviour can not be analyzed. Moreover, the performance of correction methods in the presence of non-coverage bias is normally not available.

In this article, we use data from the German Socio-Economic Panel (SOEP) that covers residential mobility in order to assess the non-coverage bias and to evaluate to what extent our correction methods reduce this non-coverage bias. The potential caveat of using the SOEP data for assessing the non-coverage bias in the MC have been analyzed by Basic et al. (2005). They found that the control for some design variables and attrition makes it possible to overcome this remedy.

With respect to correction methods, there are three broad strategies to reduce the non-coverage bias in statistical estimation: Weighting, modelling and imputation of missing data. We focus on the performance of the first two, i.e. the inverse probability weighting (IPW) and the modelling approach. The idea of the IPW approach is to assign weights to immobile persons according to their inverse probability of being immobile in order to reduce biases. This IPW approach was proposed by Robins et al. (1995).

The modelling approach aims at formulating a model for the incomplete data and the mechanism that leads to missing data. Several papers proposed models for cross-tabulations with missing values, such as Baker/Laird (1988), Chambers/Welsh (1993), Little (1985).

The data we use for our analysis is taken from the SOEP and the MC and covers the years 1996 to 1999.

The paper is organized as follows: first, we estimate the size and significance of the non-coverage bias for labour force flows. Then we introduce the inverse probability weighting approach and compare the non-coverage biases with the corresponding corrected estimates and their biases. In the next section, we describe a modelling approach for estimation in the presence of non-coverage and assess the corrective power of this approach. In the last section we summarize our findings.

2 Non-Coverage Bias

One main objective of our analysis is the size of potential biases caused by the non-coverage of residential movers. We study the impact of non-coverage with the example of labour force flows. Since the SOEP not only provides information on immobile persons but also on mobile persons, we can directly assess to what extent the exclusion of mobile persons from the analysis affects the estimates of the labour force flows.
Let $\hat{P}_{IMMO}(B|A)$ be the estimate of the transition from labour force status A to labour force status B for immobile persons and $\hat{P}_{ALL}(B|A)$ the same estimate based on all persons, i.e. based on immobile plus mobile persons. Then the bias due to the missing information on mobile persons can be calculated as $\hat{P}_{ALL}(B|A) - \hat{P}_{IMMO}(B|A)$ since $\hat{P}_{ALL}(B|A)$ is a consistent estimator. The significance of the difference can be tested by the Hausman-test, see Fitzgerald et al. (1998). If this difference is not significant, the mechanism that leads to missing data is said to be ignorable, see Rubin (1976). In this case, the weighting approach, described in the next section, may be used to correct the non-coverage bias, see Robins et al. (1995). However, such an approach does not satisfactorily correct the non-coverage bias if the propensity to be mobile depends directly on the unobserved labour force status of the individual. In this case, the mechanism that leads to missing data is said to be non-ignorable. As Little (1982) noted, if the missingness mechanism is non-ignorable, one can eliminate bias only by constructing "a model that correctly represents the missingness mechanism" (p.246).
Table 1 presents the estimates between the three labour force states: employed (E), unemployed (U) and being not in labour force (N). We estimate labour force flows for different time intervals, i.e. for 1996/97, 1996/98 and 1996/99. The idea of enlarging the time interval is to detect possible trends in the difference of estimates between immobile plus mobile persons and immobile persons only. The first column (All) in

Table 1: **Labour force flows estimates**

		E			U			N		
AB		All	Immo	Δ	All	Immo	Δ	All	Immo	Δ
	97	91.02	91.16	0.14	4.92	4.86	0.06	4.05	3.97	0.08
E	98	87.82	88.03	0.21	6.32	6.04	0.28	5.86	5.93	0.07
	99	87.01	86.37	0.64	6.04	6.30	0.26	6.96	7.33	0.37
	97	32.83	30.85	1.98	48.39	49.83	1.44	18.78	19.32	0.54
U	98	34.92	31.79	3.13	40.13	41.20	1.07	24.95	27.01	2.06
	99	41.37	37.46	3.91	28.91	29.10	0.19	29.71	33.44	3.73
	97	12.74	11.64	1.10	5.48	4.97	0.51	81.77	83.39	1.62
N	98	19.66	16.07	3.59	5.09	4.40	0.69	75.25	79.54	4.29
	99	25.89	21.13	4.76	4.53	3.71	0.82	69.58	75.15	5.57

Source: Authors' calculations, Data base: SOEP, Waves: 1996-1999

Table 1 displays the estimates for immobile plus mobile persons, the second column (Immo) the estimates for immobile persons while the third column (Δ) displays the difference between the first and the second column.
As expected, there are considerable differences for some labour force flows. For example, for labour force flows from unemployment/being not in labour force to employment (UE and NE), these differences range between 1.98 and 4.76 percentage points. There are some plausible explanations for these differences. For example, unemployed

persons becoming employed might cause a residential move and children becoming employed move out of their parental home.
In case of other estimates, the effect of non-coverage is only moderate. To assess the significance of these non-coverage biases, we use the Hausman test, see Hausman (1978). The Hausman test makes use of the fact that the estimator $\hat{P}_{ALL}(B|A)$ is efficient because it is the ML-estimate based on all available information. The estimator $\hat{P}_{IMMO}(B|A)$ which is based only on the subsample of immobile persons is consistent under the null-hypothesis of no bias due to residential moves. According to the Hausman test, we find that all differences for estimates from unemployment/being not in labour force to employment are significant (p-value$<$0.05).

3 Inverse probability weighting (IPW) approach

Inverse probability weighted estimators are based on immobile persons. The observations are weighted with the inverse of their estimated probabilities to be immobile. The transition between different labour force states can be expressed as a logit model with labour force states A_i at time point t-1 as a covariate and an indicator variable Y_i for the considered transition. Express this model as

$$\ln \frac{P(Y_i = 1|A_i)}{1 - P(Y_i = 1|A_i)} = A_i^{'}\beta \tag{1}$$

Then the first derivative of the log-likelihood gives the score equations

$$\sum_i A_i \left(Y_i - \frac{\exp(A_i^{'}\beta)}{1 + \exp(A_i^{'}\beta)} \right) = 0 \tag{2}$$

Thus, if all values of Y and A are observed, solving (2) gives maximum-likelihood estimates. Moreover, at the true population parameter values, β_{true}, the expected value of the left-side of (2) is zero. This property makes the parameter estimates consistent, see Cox and Hinkley (1974). Now let

$$R_i = \begin{cases} 1 & \text{if } Y_i \text{ is observed} \\ 0 & \text{if } Y_i \text{ is not observed} \end{cases}$$

Then, the observed data score equations can be written

$$\sum_i R_i A_i \left(Y_i - \frac{\exp(A_i^{'}\beta)}{1 + \exp(A_i^{'}\beta)} \right) = 0 \tag{3}$$

In general, the left-side of (3) no longer have expectation zero when $\beta = \beta_{\text{true}}$, so parameter estimates are inconsistent. However, the use of weights Π_i can lead to consistent estimation.

$$\sum_i \frac{R_i}{\Pi_i} A_i \left(Y_i - \frac{\exp(A_i^{'}\beta)}{1 + \exp(A_i^{'}\beta)} \right) = 0 \tag{4}$$

with $\Pi_i = P(R_i = 1|A_i, X_i)$, with X_i being some socio-economic covariates, which are assumed to affect residential mobility.
From the practical point of view, the weights Π_i^{-1} cannot depend on the potentially

missing variable Y_i because the estimation of the model for R, a model for missingness mechanism, requires the potentially unobservable values of Y_i.
Contrary to standard weighted least square results, where weighting does not affect consistency, here the consistency depends on the correct model for Π_i. With the law of iterated expectations and Π defined as above, we have:

$$
\begin{aligned}
& E\left[\sum_i \frac{R_i}{\Pi_i(A_i, X_i)} A_i \left(Y_i - \frac{\exp(A_i'\beta)}{1+\exp(A_i'\beta)}\right)\right] \\
=& E\left[\sum_i E\left(\frac{R_i}{\Pi_i(A_i, X_i)} A_i \left(Y_i - \frac{\exp(A_i'\beta)}{1+\exp(A_i'\beta)}\right) \middle| Y_i, X_i, A_i\right)\right] \\
=& E\left[\sum_i A_i \left(Y_i - \frac{\exp(A_i'\beta)}{1+\exp(A_i'\beta)}\right) \frac{1}{\Pi_i(A_i, X_i)} E\left(R_i | Y_i, X_i, A_i\right)\right] \\
=& E\left[\sum_i A_i \left(Y_i - \frac{\exp(A_i'\beta)}{1+\exp(A_i'\beta)}\right) \frac{P(R_i = 1 | Y_i, X_i, A_i)}{\Pi_i(X_i, A_i)}\right] \qquad (5)
\end{aligned}
$$

The last term in (5) is equal to one if $P(R_i = 1|Y_i, X_i, A_i) = \Pi_i(X_i, A_i)$, i.e. if the mobility behaviour is independent of the changes in labour force states, given the socio-economic covariates X and the observed labour force status A. In this case the IPW approach leads to the same results as in the case of complete observations. This condition is fulfilled only if the mechanism that leads to missing data is ignorable. Here the IPW approach leads to consistent estimates.
The weights Π_i are estimated by the logit model[1] for Immobility versus Mobility, conditional on a set of characteristics that are measured for all individuals at first wave, i.e. in 1996. The variables we use for explaining the mobility behaviour are: household income, region, household size and number of household members with respect to age group, education level, employment status and school level.[2] The estimates of the model parameters of the model for R are given in Table 2. The first column in Table 2 (MC) shows the results of the logit analyses based on the MC and the second column (SOEP) the results based on the SOEP data for the 1996 to 1998 period. The third column displays the differences between the two estimates. Significant estimates are indicated by bold figures. The significance level is 0.05. Concerning the estimated effects of the observable covariates in 1996, we find the following: Age plays an important role in mobility behaviour. As expected, the number of persons over 60 years in a household increases the probability of staying at the same address. On the opposite side, the number of young persons decreases the probability of staying immobile. Similarly, the household size is a good predictor for mobility behaviour. Nationality, measured by an indicator for Non-Germans in the household, has a positive impact on mobility. This is the only indicator where the MC and the SOEP differ in all three time intervals, with higher mobility in the MC. This difference probably stems from the oversampling of special nationalities in the SOEP.[3] The remaining covariates only

[1] Note that we estimate a household (im-)mobility model as there is evidence that residential mobility is not an individual decision, see Clarke and Tate (2002). Therefore, we assign the weight $\hat{P}(R = 1|X, A)^{-1}$ to each member of an immobile household and the weight zero to each member of a mobile household.

[2] An important explanatory variable of household mobility, housing tenure, is not used here because it is not available in the MC.

[3] The SOEP started with a separate foreigner sample, the so-called Sample B, that represents the immigrant workers in 1984. Furthermore, the immigrants subsample, Sample D, overrepresents immigrants from

Table 2: **Probability of residential (im-)mobility over the period 1996-1998 (Logit analysis)**

Variable		MC	SOEP	Diff.
Intercept		**1.5548**	**1.8182**	-0.2635
		(0.1147)	(0.3018)	(0.3229)
Household size	1 person	**-1.2111**	**-1.1797**	-0.0313
		(0.0614)	(0.1608)	(0.1721)
	3 persons	**0.7824**	**0.6776**	0.1048
		(0.0612)	(0.1527)	(0.1645)
	4 persons	**1.3701**	**1.4752**	-0.1051
		(0.0962)	(0.2442)	(0.2624)
	≥ 5 persons	**1.4375**	**1.6101**	-0.1727
		(0.1126)	(0.2792)	(0.3011)
Age ≤ 30	1 person	**-1.4298**	**-1.3827**	-0.0471
		(0.0320)	(0.0935)	(0.0988)
	≥ 2 persons	**-2.2189**	**-2.1525**	-0.0665
		(0.0496)	(0.1258)	(0.1352)
Age > 60	1 person	**0.6612**	**0.8264**	-0.1653
		(0.0436)	(0.1472)	(0.1535)
	≥ 2 persons	**0.4538**	**0.6598**	-0.2059
		(0.0781)	(0.2586)	(0.2701)
Household income	n/a	**-0.1584**	-0.0132	-0.1452
		(0.0449)	(0.1692)	(0.1750)
	< 2200	**0.2453**	**0.2656**	-0.0203
		(0.0389)	(0.1106)	(0.1172)
	2200 to < 3000	**0.2157**	**0.2519**	-0.0362
		(0.0403)	(0.1093)	(0.1165)
	4000 to < 5500	**-0.1540**	-0.1040	-0.0500
		(0.0415)	(0.0954)	(0.1040)
	5500 and more	**-0.2532**	-0.1944	-0.0588
		(0.0463)	(0.1113)	(0.1206)
Region	East-Germany	**-0.3224**	**-0.3881**	0.0657
		(0.0279)	(0.0794)	(0.0842)
School [No.]	secondary	0.0171	0.0906	-0.0735
		(0.0334)	(0.0905)	(0.0965)
	grammar	**0.1250**	**0.1528**	-0.0278
		(0.0255)	(0.0758)	(0.0800)
Education [No.]	vocational	**0.1535**	**0.1690**	-0.0155
		(0.0221)	(0.0620)	(0.0658)
	tertiary level	**0.1618**	0.0723	0.0895
		(0.0366)	(0.0937)	(0.1006)
Nationality	$\geq$ 1foreigner	**-0.7096**	**-0.2816**	**-0.4280**
		(0.0431)	(0.0942)	(0.1037)
Employment	1 person	**0.3964**	0.0769	**0.3195**
		(0.0583)	(0.1522)	(0.1630)
	≥ 2 persons	**0.6222**	0.2395	0.3827
		(0.0990)	(0.2460)	(0.2651)
Unemployment	1 person	0.0908	-0.0020	0.0928
		(0.0616)	(0.1460)	(0.1585)
	≥ 2 person	-0.2181	-0.5496	0.3315
		(0.1530)	(0.3458)	(0.3781)
Not in labour force	1 person	**0.1943**	0.0430	0.1513
		(0.0564)	(0.1438)	(0.1545)
	≥ 2 persons	**0.2429**	0.0296	0.2132
		(0.1044)	(0.2597)	(0.2799)
Observations		53'821	6'777	

Dependent Variable: indicator of mobility
coefficients for logarithm of odds ratio $\mathrm{P}(R=0)/\mathrm{P}(R=1)$
Standard deviations in paranthesis

have a low numerical influence on mobility behaviour. By varying time interval, we find that the estimated slope coefficients are stable over time.[4] Only the intercept de-

Eastern Europe. There is no information about this group membership in the MC sample.

[4]These results are available from the authors upon request.

creases with the length of the time interval indicating a higher propensity for mobility with increasing time interval, which is plausible.

To analyze the corrective power of the IPW approach, we now compare the above computed biases (Table 1) with biases obtained when using the weights. The bias obtained when using the weights is calculated as $\hat{P}_{ALL}(B|A) - \hat{P}_{IPW}(B|A)$, where $\hat{P}_{IPW}(B|A)$ is the estimate obtained by weighting the observations of the immobile persons. However, due to the lack of knowledge of the estimate $\hat{P}_{ALL}(B|A)$ for the MC, such a direct comparison is possible for the SOEP only. To assess the corrective power of the IPW approach in the MC, there are two possibilities: First, we can take the SOEP estimate $\hat{P}_{ALL}(B|A)$ as a benchmark for the MC. However, this assumption implies that the estimates on the basis of the SOEP and the MC are equal.[5] Second, we can calculate an improvement rate (IR) which is based on the ratio between the Correction $(:= \hat{P}_{IPW}(B|A) - \hat{P}_{IMMO}(B|A))$ in the MC and the Bias $(:= \hat{P}_{ALL}(B|A) - \hat{P}_{IMMO}(B|A))$ from the SOEP, i.e. $IR = \frac{Correction}{Bias}$. The IR gives the proportion of the bias corrected by using the above weights. Here, we make the assumption that the bias in the MC and in the SOEP is equal. According to the interpretation of the improvement rate we distinguish five cases: First, an improvement rate of 1 indicates a complete bias reduction. Second, the bias will be reduced to some extent if the rate lies between zero and one. Third, a negative value of the rate implies an increase of the bias. Fourth, a rate higher than one implies an overcorrection of the bias and a rate equal zero indicates complete failure of the bias reduction.

We present the performance of the IPW approach for both cases, i.e. if we assume equal absolute values (Table 3) and equal biases (Table 4) between the SOEP and the MC.

In Table 3 we display the absolute bias only for the labour force flows in which the substantial bias occurred, i.e. UE and NE. The bias is estimated without the weights (uncorrected) and with the above weights (corrected). The weighting leads to improved estimates if the corrected bias is smaller than the uncorrected bias. The comparison of

Table 3: **Absolute biases of labour force flows UE and NE**

Transition	**1996-1997**		**1996-1998**		**1996-1999**	
	uncorrected	corrected	uncorrected	corrected	uncorrected	corrected
	SOEP					
UE	1.98	1.37	3.13	2.02	3.98	0.75
NE	1.10	0.66	3.59	2.49	4.56	2.21
	MC					
UE	3.54	2.61	2.39	0.26	5.48	3.55
NE	-0.37	-0.71	1.97	0.86	3.75	2.20

the corrected and uncorrected biases shows that all biases are reduced to some extent. This holds for the SOEP (heading SOEP) as well as for the MC (heading MC). Similar results for the correction of attrition effects in the European Community Household Panel (ECHP) were shown by Neukirch (2002).

Table 4 presents the estimated improvement rates for the SOEP and the MC. The first column in Table 4 (Bias) shows the original bias, the second column (SOEP) the

[5]Basic et al. (2005) found that the difference of the corresponding labour force flows estimates for the immobile persons between the SOEP and the MC range between 0.5 and 1.5 percent.

Table 4: **Bias reduction expressed by ratio correction/bias (SOEP and MC data)**

t	Bias	Biascorrection/Bias	
		SOEP	MC
		UE	
1997	1.98	0.31	0.47
1998	3.13	0.35	0.68
1999	3.91	0.81	0.49
		NE	
1997	1.10	0.40	0.31
1998	3.59	0.31	0.31
1999	4.56	0.56	0.35

corrected proportion of the bias for the SOEP and the third column (MC) the corrected proportion of the bias for the MC. The results in Table 4 reveal that all the biases are reduced to some extent. For example, in the case of the flow UE the resulting correction lies between 31 percent (1997) and 81 percent (1999) for the SOEP and between 47 percent (1997) and 68 percent (1999) for the MC.

4 Log-Linear models for missing data

In Section 3 we presented the IPW approach for the estimation of labour force flows in the presence of missing data for residential movers. This approach leads to unbiased results if the movers are a random sample of all individuals. However, the results in Table 1 indicate that, for some labour force flows, the mobility is related to the future labour force status. Mobility of this nature is an example of non-ignorable mobility because information about the missing labour force flow is unobserved. In this case it is necessary to specify a joint model for flows and missingness mechanism. This procedure ensures that the estimates of the labour force flows are adjusted in accordance with the model for missingness mechanism.
In what follows, we present the use of log-linear models for labour force flows subject to non-ignorable missingness, see Baker and Laird (1988), Fay (1986).
Let A and B denote labour force states at time t_1 and t_2 respectively, and R and S two response indicators representing whether A or B are missing or not (0 for missing; 1 for not missing). Since both variables, A and B, can be missing, our data consists of four different types of contingency tables: a table completely classified by A and B, two marginal tables classified by A or B only, in which the cell frequencies are the marginal sums across the missing cells and a marginal table classified by neither A nor B, in which the cell frequency is the sum across the missing cells of A and B. Table 5 provides an example of this data structure with A and B having three categories (E, employed; U, unemployed; N, not in labour force). If persons are immobile at both times, the observed data are the cells of this 2-way table. However, if the persons move at t_1 or t_2 the observed data correspond to the margins of the table: $r(E+)$, $r(U+)$, $r(N+)$ are the observed data if persons are immobile at t_1, but move out at t_2; and $s(+E)$, $s(+U)$, $s(+N)$ are the observed data if persons move in at t_1 but are immobile at t_2.

Table 5: **A 3 $\times$ 3 Table with Data partially classified on both variables**

Status		t_2			
		E	U	N	
t_1	E	$n(EE)$	$n(EU)$	$n(EN)$	$r(E+)$
	U	$n(UE)$	$n(UU)$	$n(UN)$	$r(U+)$
	N	$n(NE)$	$n(NU)$	$n(NN)$	$r(N+)$
		$s(+E)$	$s(+U)$	$s(+N)$	

As the table is incomplete, a fully saturated log-linear model is not identifiable. However, various restricted models can be estimated, including those where R and S depend on A and B.

We now consider possible models for the variables A, B, R and S. The joint distribution $P(A, B, R, S)$ can be factorized as $P(A, B)P(R, S|A, B)$. As A and B represent the labour force states at time t_1 and t_2, we shall specify a saturated log-linear model for $P(A, B)$ and just consider alternative specifications of $P(R, S|A, B)$ which determine the missingness mechanism, i.e reasons why persons move. Since A and B are both missing for some cases, all models that associate R or S with A or B are non-ignorable. Thus the only ignorable model is one, which assumes that mobility behaviour is independent of A and B.

The observed likelihood corresponding to the above model is given by

$$
\begin{aligned}
L = & \prod_{i \in S_{11}} P(A, B)P(R = 1, S = 1|A, B) \\
& \times \prod_{i \in S_{10}} \sum_{A} P(A, B)P(R = 1, S = 0|A, B) \\
& \times \prod_{i \in S_{01}} \sum_{B} P(A, B)P(R = 0, S = 1|A, B) \\
& \times \prod_{i \in S_{00}} \sum_{A} \sum_{B} P(A, B)P(R = 0, S = 0|A, B) \qquad (6)
\end{aligned}
$$

where S_{RS} is the set of individuals with response pattern (R, S). The first term in equation (6) is the contribution to the likelihood from the completely classified individuals, the second term the contribution to the likelihood of individuals only observed on A, analogously the third term is the contribution to the likelihood of the individuals only observed on B and finally the last term is the contribution to the likelihood of the individuals neither observed on A nor on B.

To ensure that the model parameters can be estimated, the model $P(R, S|A, B)$ must be constrained in accordance with some assumption about the reasons why persons move. Constraints are first placed on the probability that an individual moves in both time periods. We assume that there are no persons who are mobile on both time points, i.e. $P(R = 0, S = 0|A, B) = 0$. Next, the breakdown of mobile persons into movers out and movers in is due to the randomization of the MC sample. As each move out is equal to a move in it seems reasonable to assume $P(R = 1, S = 0|A, B) = P(R =$

$0, S = 1|A, B)$, i.e. the probability that the person moves out is equal the probability that the person moves in. Further constraints can be derived from the SOEP as in the SOEP the mobile persons are followed to their new homes. Thus, we can check whether the considered constraints are correct or not. However, this validation is only possible for the persons moving out from the sample, as the persons who move into the empty dwellings of the sample are not recorded in the SOEP. The SOEP only records moves into already existing households.[6]

First, we can test whether the mobility behaviour depends on neither A nor B, on A, on B and on A and B together. This can be seen as a test on ignorability or non-ignorability of the mobility process. The first two models are ignorable[7] and the other two nonignorable, since R is associated with variable B which is sometimes missing. To decide which model is the best, we use the AIC and BIC indices.[8] According to these two indices the model $P(R|AB)$ is chosen.

The following constraints, derived from the SOEP, are made to give an estimable model considered here, which is described below.

The probabilities of moving from labour force states E and U are constant and the same for all persons, the only exception being the transition from U to E. The probabilities of moving for the transitions U to E, N to E and N to U are constant and the same for all persons.

In Table 6 we compare the absolute biases for the labour force flows UE and NE obtained without modelling approach (uncorrected) and by using the above log-linear modell (corrected). In the case of the SOEP (heading SOEP) the bias reduction is only moderate. This is due to the fact that for the SOEP we estimate a log-linear model with one response indicator only.[9] On the other hand the bias reduction for the MC is much higher and for the flow NE the bias is even overcorrected.[10] The tendency for overcorrection is a typical feature of the non-ignorable models, see, for example, Chambers and Welsh (1993) or Little (1985). These findings are also supported by the results in Table 7, which displays the estimated improvement rates. As in the case of absolute biases we observe only moderate bias reduction for the SOEP in Table 7. But for the MC the bias reduction is much larger. For example, in the case of the flow UE the correction lies between 0.76 (1996) percent and 1.10 (1999) percent of the original bias. These results reveal the importance of additional information of movers in for the bias reduction. Here we also observe some overcorrection of the bias. Especially for the flow NE.

Up to this point, we have presented the modelling strategy only for two time points. As the MC is a four year panel, we now extend the approach to four time points. Let A, B, C and D represent labour force status measured at four subsequent time points. Since each variable, A, B, C and D, can be missing, we have four indicator variables R, S, U and V, showing whether A, B, C or D is observed or not. In the MC people can return to the sample once they moved out. This is due to the fact that the MC is

[6] Thus, we cannot test whether the assumption $P(R = 1, S = 0|A, B) = P(R = 0, S = 1|A, B)$ is correct or not.

[7] Contrary to the MC in which both variables, A and B, can be missing, in the SOEP only B can be missing. Thus A is always observed and any model that associates R with A is ignorable.

[8] The AIC measure (Akaike, 1974) includes, beside the loglikelihood $\ln L$, the number of parameters to be estimated $(-2 \ln L + 2 * \text{parameters})$. The BIC measure (Schwarz, 1978) considers the loglikelihood $\ln L$ in relation to the number of subjects and the number of parameters to be estimated $(-2 \ln L + \ln(n) * \text{paramters})$. The lower the AIC and BIC indices the better the model.

[9] In the SOEP we do not observe the marginal table for persons who move into the empty dwellings.

[10] The negative value of the estimated absolute bias implies $\hat{P}_{LOG}(B|A) > \hat{P}_{ALL}(B|A)$, with $\hat{P}_{LOG}(B|A)$ being the estimate based on the log-linear modelling strategy.

Table 6: **Absolute biases of labour force flows UE and NE**

Transition	**1996-1997**		**1996-1998**		**1996-1999**	
	uncorrected	corrected	uncorrected	corrected	uncorrected	corrected
	SOEP					
UE	1.98	1.59	3.13	2.66	3.98	3.63
NE	1.10	0.53	3.59	2.40	4.56	3.11
	MC					
UE	3.54	2.04	2.39	-0.10	5.48	1.16
NE	0.37	-1.85	1.97	-1.42	3.75	-2.73

Table 7: **Bias reduction expressed by ratio correction/bias (SOEP and MC data)**

t	Bias	Biascorrection/Bias	
		SOEP	MC
		UE	
1997	1.98	0.20	0.76
1998	3.13	0.15	0.80
1999	3.91	0.07	1.10
		NE	
1997	1.10	0.52	1.35
1998	3.59	0.33	0.94
1999	4.56	0.35	1.36

an area sampling. However, such re-entries cannot be detected by linking together data across the years, see Herther-Eschweiler (2003). Thus, we have a monotone pattern of missing data (Little, 1982).[11] The joint distribution $P(A, B, C, D, R, S, U, V)$ can be factorized as $P(A, B, C, D)P(R, S, U, V|A, B, C, D)$. As in the case of two time points the fully saturated log-linear model is not identified. For the model representing the missingness mechanism the following restriction seems to be plausible:

$$P(R, S, U, V|A, B, C, D) = P(R|A)P(S|A, B, R)P(U|B, C, S)P(V|C, D, U)$$

This assumption implies that the decision whether someone moves at time t depends on the labour force states in t-1 and t and on the mobility status in the previous wave. The probability of being mobile is equal to one for already mobile households, because we have assumed monotone patterns of missing data.
Furthermore, we assume population homogeneity. That is, all subjects are characterized by the same set of parameters describing the reasons why subjects move.

$$P(S|A, B, R = 0) = P(U|B, C, S = 0) = P(V|C, D, U = 0)$$

The model $P(A, B, C, D)$ can be factorized as $P(A)P(B|A)P(C|A, B)P(D|A, B, C)$. Here, we assume a first order Markov process. That is, the category someone belongs

[11] A monotone pattern of missing data means that the variables can be ordered in such a way that a missing score on one particular variable implies having missing scores on all subsequent variables too.

to at time point t depends on the category he or she belonged to in the most recent time point t-1 only, but not at earlier points of time.

$$P(A, B, C, D) = P(A)P(B|A)P(C|B)P(D|C)$$

These restrictions together with restrictions from above model for two time points are assumed to give an estimable model considered here.
Table 8 presents the estimated absolute biases and Table 9 the estimated improvement rates according to this model. The results in Table 8 reveal that the pattern valid for

Table 8: **Absolute biases of labour force flows UE and NE, four time points (SOEP and MC data)**

Transition	**1996-1997**		**1997-1998**		**1998-1999**	
	uncorrected	corrected	uncorrected	corrected	uncorrected	corrected
	SOEP					
UE	4.93	3.21	3.24	4.41	3.24	2.74
NE	2.84	2.26	2.55	3.22	2.26	2.17
	MC					
UE	5.33	-0.64	1.06	-3.10	4.49	-2.37
NE	0.48	-3.11	2.17	-1.22	3.16	-2.93

the model of two time points also transfers to the model of four time points. Thus, the bias reduction for the SOEP is only moderate and, for the transition 97/98, the bias is even enlarged. In the case of the MC, the bias reduction is much larger and also the overcorrection of biases increased.
In Table 9 we observe the same patterns as in Table 8, i.e. overcorrection of the bias

Table 9: **Bias reduction expressed by ratio correction/bias, four time points (SOEP and MC data)**

t	Bias	Biascorrection/Bias	
		SOEP	MC
		UE	
96 to 97	4.93	0.35	1.21
97 to 98	3.24	-0.36	1.28
98 to 99	3.24	0.15	2.12
		NE	
96 to 97	2.84	0.20	1.26
97 to 98	2.55	-0.26	1.33
98 to 99	2.26	0.04	2.69

for the MC and only moderate bias reduction for the SOEP.
Furthermore, we can estimate a model, with labour force flows that are assumed to be time homogenous , i.e. $P(B|A) = P(C|B) = P(D|C)$. Tables 10 and 11 display the estimated absolute biases and the estimated improvement rates for this model. Here, we also observe the same patterns as in the case of the previous models, i.e. larger bias reduction for the MC and some overcorrection of the biases in the MC.

Table 10: **Absolute biases of labour force flows UE and NE, time homogenous labour force flows**

Transition	uncorrected	corrected
	SOEP	
UE	3.83	3.50
NE	2.54	2.47
	MC	
UE	3.59	-1.87
NE	1.96	-2.18

Table 11: **Bias reduction expressed by ratio correction/bias, time homogenous transition probabilities (SOEP and MC data)**

Bias	Biascorrection/Bias	
	SOEP	MC
	UE	
3.83	0.09	1.42
	NE	
2.54	0.03	1.63

5 Conclusion

Our starting point was the question to what extent the non-coverage of residential movers in the MC affects the estimates of labour force flows. Here, we used the SOEP, which covers residential mobility to asses these affects. The results show that the labour force flows from unemployment/being not in labour force to employment are underestimated on the basis of the immobile persons only.

Next, we presented two approaches which are used to reduce the effects of non-coverage in the estimation of labour force flows: Weighting and modelling approach. The weighting approach leads to systematic improvements in the case of all biases. The longer the time interval, the higher the bias reduction due to the application of weights. However, the bias reduction range between 30 and 70 percent of the original bias.

The modelling approach leads to a higher bias reduction than the weighting approach. However, this approach tends to overcorrect the bias. This feature is typical for the non-ignorable models .

Beside the difference in the reduction of biases, the two approaches, weighting and modelling, differ in the use of the available information. The weighting approach only makes use of the information of immobile persons. The information of movers enters indirectly the estimation through the estimated probabilities of being immobile. In the case of the modelling approach all available information is directly incorporated into the estimation. Thus, the modelling approach is more efficient than the weighting approach.

A recommendation regarding the use of either weighting or log-linear approach to correct for the non-coverage bias needs knowledge about the missingness mechanism. Since, the missing data mechanism is not verifiable, we recommend to use both approaches, the weighting and the log-linear approach. If the estimates are robust to both approaches, this could be seen as an indication of ignorable missing data mechanism.

If the estimates are not robust, this suggests that mechanism that leads to missing data is non-ignorable. In this case one may be cautious about interpreting the results.

References

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716–723.

BAKER, S.G., LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association* **83** 62–69.

BASIC, E., MAREK, I., RENDTEL, U. (2005). The German Microcensus as a tool for longitudinal data analysis: An evaluation using SOEP data. *Schmoller's Jahrbuch - Journal of Applied Social Science Studies* **Vol. 125** Number 1.1–16

CHAMBERS, R.L., WELSH, A.H. (1993). Log-linear Models for Survey Data with Non-ignorable Non-response. *Journal of the Royal Statistical Society B* **53** 157–170.

CLARKE, P.S., TATE, P.F. (2002). An Application of Non-Ignorable Non-Response Models for Gross Flows Estimation in the British Labour Force Survey. *Australian & New Zealand Journal of Statistics* **44** 413-425.

COX, D.R., HINKLEY, D.V. (1974). *Theoretical Statistics.* Chapman and Hall, New York.

FAY, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association* **81** 354–365.

FITZGERALD, J., GOTTSCHALK, P., MOFFITT, R. (1998). An Analysis of sample Attrition in Panel Data - The Michigan Panel Study of Income Dynamics. *Journal of Human Resources* **33** 251–259.

HAUSMAN, J. (1978). Specification tests in econometrics. *Econometrica* **46** 1251–1271.

HEIDENREICH, H.-J. (2002). Längsschnittdaten aus dem Mikrozensus. Basis für neue Analysemöglichkeiten. [Longitudinal Data on the Basis of the German Microcensus. A new data base for analysis.]. *Allgemeines Statistisches Archiv* **86** 213–231.

HERTER-ESCHWEILER, R. (2003). Längsschnittdaten aus dem Mikrozensus: Basis für neue Analysemöglichkeiten. Statistisches Bundesamt, Bonn.

LITLLE, R.J.A (1982). Models for Nonresponse in Sample Surveys. *Journal of the American Statistical Association* **77** 237–250.

LITLLE, R.J.A (1985). Nonresponse Adjustments in Longitudinal Surveys: Models for Categorical Data. *Bulletin of the International Statistical Institute* **15** 1–15.

LITLLE, R.J.A, RUBIN, D.B. (2002). *Statistical Analysis with Missing Data.* Wiley, New York.

NEUKIRCH, T. (2002). Nonignorable attrition and selectivity biases in the Finnish subsample of the ECHP. CHINTEX Working Paper #5.URL: http://www.destatis.de/chintex/download/paper5.pdf

ROBINS, J., ROTNITZKY, A., ZHAO, L. (1995). Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90** 106–121.

RUBIN, D.B. (1976). Inferences and missing data. *Biometrika* **63** 581–592.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6** 461–464.