Freie Universität Berlin

Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft
der Freien Universität Berlin

Volkswirtschaftliche Reihe

2007/6

Assessing the bias due to non-coverage of residential movers in the German Microcensus Panel: An evaluation using data from the Socio-economic Panel

Ulrich Rendtel und Edin Basic

3-938369-54-X

Assessing the bias due to non-coverage of residential movers in the German Microcensus Panel: An evaluation using data from the Socio-economic Panel.[*][†][‡]

Edin Basic Ulrich Rendtel

Abstract

The German Microcensus (MC) is a large scale rotating panel survey over three years. The MC is attractive for longitudinal analysis over the entire participation duration because of the mandatory participation and the very high case numbers (about 200 thousand respondents). However, as a consequence of the area sampling that is used for the MC , residential mobility is not covered and consequently statistical information at the new residence is lacking in the MC sample. This raises the question whether longitudinal analyses, like transitions between labour market states, are biased and how different methods perform that promise to reduce such a bias.

Based on data of the German Socio-Economic Panel (SOEP), which covers residential mobility, we analysed the effects of missing data of residential movers by the estimation of labour force flows. By comparing the results from the complete SOEP sample and the results from the SOEP, restricted to the non-movers, we concluded that the non-coverage of the residential movers can not be ignored in Rubin's sense.

With respect to correction methods we analysed weighting by inverse mobility scores and loglinear models for partially observed contingency tables. Our results indicate that weighting by inverse mobility scores reduces the bias to about 60 percent whereas the official longitudinal weights obtained by calibration result in a bias reduction of about 80 percent. The estimation of loglinear models for non-ignorable nonresponse leads to very unstable results.

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

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

[*] This is an extended version of the paper accepted to be published in: Advances in Statistical Analysis - AStA.

[†] 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

[‡] We gratefully acknowledge valuable comments from three anonymous referees that helped to improve the first version of manuscript.

population structure, labour market behaviour and the housing situation. It is conducted on a yearly basis, with each sample household retained for four consecutive years and one fourth of the sample replaced each year. Originally 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 support an analysis of change, 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.

A methodological problem of the longitudinal use of the MC arises from the fact that residential movers are not traced. This is due to the fact that the MC uses area sampling, where the dwellings within the area 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 bias in the analysis of interest, if the residential movers differ systematically from persons who stay at their home. 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, see Wagner/Mulder (2000). However, due to the lack of corresponding data for residential movers, the impact of current changes of the labour force status on the mobility behaviour can not be analyzed. Moreover, the performance of correction methods in the presence of a bias is not available.

The problem of the longitudinal use of area samples is not restricted to the German MC. A similar problem occurs in the British Labour Force Survey[1] (LFS), which is also a rotating panel based on area sampling and is also used for longitudinal analysis, see Clarke/Tate (1999, 2002) and Tate (1999). However, here the time intervals between successive interviews is only three months and the total duration of households in the sample is 15 months. Thus, the extent of residential mobility is much smaller. Furthermore, participation in the LFS is not mandatory, resulting in substantial non-response rates. Therefore Clarke and Tate have collapsed all sources of missingness and created a weighting scheme based on an analysis of the collapsed non-response. For the evaluation of this weighting scheme they used a simulation where they generated labour force states and non-response pattern from statistical models, see Clarke/Tate (2002).

Here we use a different approach, that avoids the generation of labour force states and carefully reflects the nonresponse mechanism by residential mobility. Our approach is based on data of the German Socio-Economic Panel (SOEP) which covers residential mobility. In order to assess to what extent the missing information about residential movers affect the estimates, we compare results based on the full SOEP sample with estimates based on the SOEP subsample of immobile persons, which mimics the MC. However, the two surveys differ with respect of their sampling design and the questionnaire. Also the SOEP is subject to panel attrition. These differences make it necessary to evaluate carefully the comparability of results from these surveys.

With respect to correction methods, we used standard methods in the analysis of gross changes, namely weighting of transition tables by inverse propensity scores of mobility (IPW approach) see, for example, Robins et al. (1995), Miller et al. (2001), and loglinear modelling of incomplete contingency tables. The loglinear approach uses a model for completely cross-classified contingency tables. The table results from the indicators of interest and the corresponding response indicators. Several papers proposed models for cross-tabulations with missing values, such as Baker/Laird (1988),

[1] The same holds also for the US Current Population Survey (CPS).

Chambers/Welsh (1993), Little (1985). Our methodology is evaluated for labour force flows, which serves as a standard example in the analysis of change, see for example Stasny (1986) and Clarke/Chambers (1998). While Stasny (1996) has used the observed labour force state as an indicator for a missing observation, Clarke and Chambers (1998) used the unobserved labour force state for this purpose. Here we will go one step further and will investigate whether special transitions between labour force states should be used as indicators for residential mobility. Besides its use as a standard example, labour force flows are a key topic in both surveys, the MC and the SOEP.

The paper is organized as follows: first, we investigate the comparability of the SOEP and the MC and display the extent of residential mobility in both surveys. Then we estimate the size and significance of the non-coverage bias for labour force flows. In section 4 we introduce the inverse probability weighting approach and assess the corrective power of the IPW estimates. In the next section, we describe a loglinear approach and assess the performance of this approach to reduce biases in the estimates of labour force flows. In the last section we summarize our findings.

2 Measurement of residential mobility by the SOEP and the MC

As we are going to use the SOEP as a means to control the non-coverage bias in the MC we have to check the comparability of these two surveys with respect to residential mobility.

Both surveys are interviewer-based and address similar topics. There are some minor differences in the design of the questionnaire. Here, the MC uses a strict concept of a reference week (last week in April) whereas the SOEP refers to the moment of the interview (approximately the end of March).

The MC and the SOEP use different sampling designs. The SOEP uses oversampling of special sub-populations, namely foreigners, East-Germans and immigrants, see Haisken-DeNew/Frick (2005). It has been started in 1984. So in 1996 the major part of the SOEP stayed for 12 years in the survey. In contrast, the MC uses an equal probability sample for the entire population, being the resident population in Germany for both surveys[2], see Heidenreich (2002). The rotation group that is analysed here started its MC membership in 1996.

The important difference of the two surveys that is used here is in the treatment of residential mobility. According to the area sampling of the MC dwellings are sampled and residential movers are not followed. Unlike in the SOEP, where residential movers are followed, this feature leads to missing information for residential movers at their new home. Instead, new persons who move into the dwellings of the residential movers enter the MC sample. Such moves are not covered in the SOEP, where only moves into already existing households are recorded.

Furthermore, participation in the MC is mandatory. Hence non-response due to noncontact and unwillingness is reduced to a minimum level of about 3 percent. Contrary, the SOEP is based on a voluntary participation. In a panel survey one has to distinguish between non-response at the start of the SOEP in 1984 and panel attrition in subsequent waves. The initial non-response at the start of the SOEP was about 30

[2] There are minor differences with respect to the inclusion of institutions, which concerns elderly people. However, for our analysis of labour states which includes only persons from age 16 to 65, these differences can be ignored.

percent. The attrition rate between subsequent waves is about 5 percent per wave, see for example Rendtel (1995) and Kroh/Spiess (2006) for more recent figures. For panel attrition field work is of utmost importance. This does not only hold for the SOEP but also for other European household panels, see Behr et al. (2005). Here residential mobility has in general a negative impact on participation rates, even if we control for other variables like age and education level. This may raise doubts about the SOEP as a useful means to control the effect of not covering residential mobility.

The following arguments underpin the use of the SOEP as a means for bias control:

a) The effect of panel attrition in the European Community household Panel (ECHP), which includes the SOEP as a national subsample, is generally small, if not ignorable, see Ehling/ Rendtel (2004, Section on panel attrition). Even in the case where one has perfect control on non-response and attrition, like in the Finnish subsample where information from the national register can be used for participants and non-responders, there was no evidence of a virulent attrition bias for change-analysis, see Sisto (2003).

b) There was some evidence of a bias resulting from initial non-response, however in later panel waves this initial non-response bias declines very fast, see Sisto (2003) for an example on change analysis with Finnish register data.

c) For the SOEP residential mobility was always included as a variable to control non-response via estimated response propensities, see Kroh/Spiess (2006, Section 3). The inverse of the response propensities was multiplied by the initial survey weights to produce design-based estimates of population totals and proportions. In general unweighted estimates of transition probabilities show only minor differences to the design-based estimates, as we will demonstrate below.

d) Finally, as a simple matter of scale, about 85 percent of all residential movers in the SOEP participate in the next panel wave, see Rendtel (1995, p. 219). So the potential of an attrition bias, expressed by the percentage of the sample size that is lost, is rather small.

Special care must however been given to the longitudinal treatment of attrition and the computation of the mobility status in the SOEP. As long as a SOEP sample person is in the gross sample his or her mobility status is known. This holds also in the case when no interview was realized. However, in the next year such a person is no longer in the gross sample and therefore the mobility status is no longer recorded. Thus, if we take for example the mobility status from wave 96 to 98, we know the mobility status in 98 of the attriters in wave 97 if they moved away in 97. But we don't know the mobility status of 98 if the attriter in wave 97 did not move in 97. In this case we have to know the mobility behaviour in wave 98 which is unknown in this case. Inclusion of all cases with known mobility status would therefore yield too high mobility rates. Therefore one has to skip the attriters of wave 97 from the analysis of mobility over the period 96 to 98. Similar arguments hold for the period 96 to 99, where the attriters of the waves 97 and 98 have to be skipped.

e) The total rate of residential mobility in the SOEP and the MC is very similar, see Table 1.[3] This holds for the unweighted and weighted SOEP data. According to

[3] The same conclusion holds also for an analysis at the household level, see Basic et al. (2005)

the MC, the cumulative effect from 1996 to 1999 amounts to about 26 percent of all individuals, which is a substantial part of the longitudinal sample.

Table 1: **The cumulative extent of residential mobility in the MC and the SOEP. Percentage and cases of individuals with residential mobility after 1996. (MC, SOEP = unweighted results, SOEP* = design-based results using the design weights and the attrition correction factors)**

Sample	**Transition**					
	1996-1997		**1996-1998**		**1996-1999**	
	%	cases	%	cases	%	cases
MC	11.13	12594	19.30	21719	25.87	28968
SOEP	10.51	1520	20.23	2836	26.64	3524
SOEP*	9.94		19.62		26.15	

f) Age is the most important variable for the explanation of residential mobility. The impact of age is demonstrated in Figure 1. The mobility rates of the SOEP and the MC almost coincide within the range of 16 to 65. The pattern displayed in Figure 1 exhibits a high mobility for young adults (people leaving the parental home, moving together with a new partner, etc.). Similar pattern can be found for the periods 1996–1998 and 1996–1999.

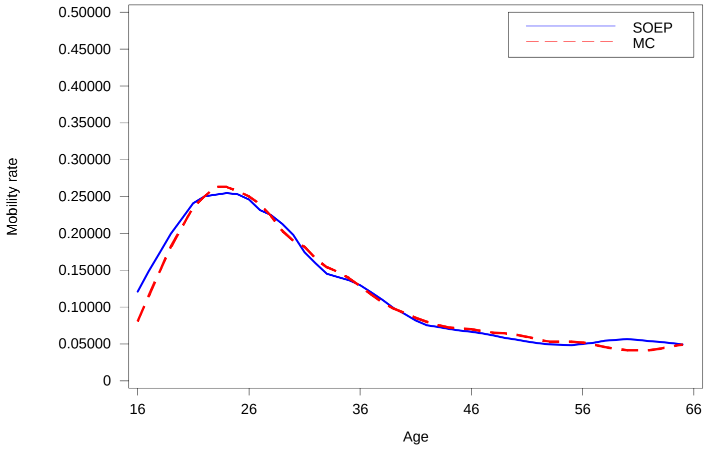

Figure 1: **Mobility rates from 1996 to 1997 calculated from the SOEP and the MC. Rates computed from a scatter plot smoother (cubic spline interpolation) according to SAS procedure LOESS.**

g) The impact of covariates in a logit model for residential mobility is almost identical, as shown below in Table 2.

In order to produce comparable results the set of covariates must be identical for the SOEP and the MC. The variables we use for explaining the mobility behaviour are: age-groups, household size, sex, region, education level, nationality, marital status, and labour force status. For the reference category, we consider an unmarried female between 50 and 65 years being not in labour force, with lower education, leaving in western part of Germany and in household that contains more than two persons. The variables are measured at the start of the period in 1996.[4] The dependent variable is the residential status in 1997.

For the estimation of the model we had to consider the dependency[5] of residential mobility at the household level, i.e. in the great majority of cases, about 85 percent, all persons in a household react in the same way. One approach[6] to cope with this dependency is the GEE approach of Liang/Zeger (1986). For the working correlation matrix we selected the equi-correlation matrix.[7]

Table 2: **Probability of residential immobility over the period 1996-1997 (GEE analysis with household clusters)**

Variable		**MC**	**SOEP**	**Diff.**
Intercept		**1.6114**	**2.0190**	**-0.4076**
		(0.0596)	(0.1265)	(0.1399)
Age ≤ 30		**-0.5132**	**-0.5354**	0.0221
		(0.0281)	(0.0657)	(0.0714)
Age > 45		**0.7191**	**0.5402**	0.1788
		(0.0338)	(0.0880)	(0.0943)
Household size	1 person	**-0.5452**	**-0.4675**	-0.0776
		(0.0388)	(0.1122)	(0.1187)
	2 persons	**-0.1379**	-0.0726	-0.0653
		(0.0377)	(0.0945)	(0.1017)
Sex	male	**0.0337**	0.0024	0.0313
		(0.0133)	(0.0308)	(0.0335)
Region	East-Germany	-0.0196	-0.0800	0.0684
		(0.0350)	(0.0879)	(0.0946)
Education	vocational	0.0251	0.1075	-0.0824
		(0.0255)	(0.0561)	(0.0616)
	tertiary level	**-0.1561**	**-0.1987**	0.0427
		(0.0290)	(0.0688)	(0.0746)
Nationality	German	**0.5114**	0.1570	**0.3543**
		(0.0415)	(0.0845)	(0.0941)
Marital Status	Married	**0.3265**	**0.3389**	-0.0124
		(0.0352)	(0.0796)	(0.0870)
Labour Force Status	Employment	**-0.1621**	**-0.1251**	-0.0371
		(0.0239)	(0.0554)	(0.0603)
	Unemployment	**-0.2631**	-0.0437	**-0.2194**
		(0.0394)	(0.0870)	(0.0955)
Observations (Individuals)		76'835	11'955	
Log Likelihood		-24'876	-3'918	
Pseudo R^2		0.1166	0.2366	

Standard deviations in paranthesis.
coefficients for logarithm of odds ratio $\mathrm{P}(R = 1)/\mathrm{P}(R = 0)$

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. The third column displays the differences between the two estimates. Significant estimates are indicated by bold figures. The significance level is 0.05.

[4] An important explanatory variable of residential mobility, housing tenure, is not used here because it is available in the MC only for the year 1998.

[5] Ignoring dependency would lead to incorrect estimation of standard errors, resulting in incorrect inferences of parameters.

[6] Alternatively, Clark/Chambers (1998) proposed to use an analysis at the household level. The individual characteristics have then to be transformed into variables that count the occurrence of certain people within the household. Basic et al. (2005) used this approach. Their conclusions were the same as reported here.

[7] For the computation we used the SAS procedure GENMOD.

Concerning the estimated effects of the observable covariates in 1996, we find the following: Age is the most important variable for mobility. This was clear from Figure 1. The corresponding estimated coefficients for the MC and the SOEP show almost no difference. It is also reasonable that 1–person households have a higher tendency for a residential move. This is well reflected for the MC and the SOEP. Furthermore, persons with tertiary level of education have a higher mobility rate. This is also plausible and holds in both surveys. The same holds for the unmarried persons. In the case of nationality we find substantive differences. This concerns the significance of the variable within each survey – for the SOEP nationality is no longer significant – but also the difference between the two estimated coefficients is significant. This difference can be related to the the origin of the SOEP foreigner sample (Sample B), which is a sample of immigrant workers that was created in 1984. It is reasonable to assume that after more than 12 years in Germany this population has adapted to some extent to the general German mobility level. The difference between the two intercepts merely reflects the different mobility rates for foreigners who constitute the reference category.[8]

Finally, we used the indicators for the labour force status as predictors of the residential mobility. The results indicate a lower mobility for those who are not in the labour force, which is plausible. In principle both surveys give results in the same direction. However, the difference between the coefficients for unemployment is significant. We do not see any argument for systematic differences of the SOEP and the MC with respect to unemployment. So the result may be simply the outcome of picking for significances.[9]

h) Finally, in order to compare the measurement of the labour force status, we compared the labour force flows for the persons without residential mobility.

Table 3 displays the results for the MC and the SOEP. We also performed a test[10] to check whether the differences between the estimates based on the MC and the SOEP are significant. Significant differences (p-value<0.05) are printed in boldface. The results in Table 3 reveal that differences between corresponding estimates of MC and SOEP are small. However, some of these differences are significant. This may be due to the very high case numbers in MC resulting in small estimated standard errors.

The only exception are the transitions from unemployment to unemployment/not being in labour force for a single time interval, the period 96–99. Here we found a minor change in the construction of the SOEP questionnaire from 1998 to 1999. This leads to a slight exchange in case numbers between the states U and N for the year 1999. Thus, one has to be careful with instabilities in the questionnaire design. However, for all other transitions and periods the results of Table 3 are quite promising.

From these findings we conclude that the process of residential mobility is comparable between the two surveys. We also conclude that the measurement of the labour

[8] Adding the intercept and the coefficient for being German results for both survey in the same value.

[9] In fact, for the time intervals 1996 – 1998 and 1996 – 1999 the difference of the coefficients of unemployment is no longer significant. These results can be obtained from the authors upon request.

[10] We performed a two group test, which is based on the fact that two samples are independent from each other. We used the following test statistic $t = \frac{(\hat{p}_{\mathrm{MZ}} - \hat{p}_{\mathrm{SOEP}})}{\sqrt{\hat{\sigma}^2_{\mathrm{MZ}} + \hat{\sigma}^2_{\mathrm{SOEP}}}} \sim \mathcal{N}(0,1)$

Table 3: **Longitudinal results for transition between labor states for persons without residential mobility (SOEP and MC data).**

							No. of cases	
Transition	E		U		N		All	
from 96	MC	SOEP	MC	SOEP	MC	SOEP	MC	SOEP
97	90.59	91.16	**4.21**	**4.86**	**5.20**	**3.97**	44164	6869
E 98	87.57	88.03	**4.95**	**6.04**	**7.48**	**5.93**	39469	5696
99	85.32	86.37	**5.16**	**6.30**	9.52	7.33	35700	4887
97	29.29	30.85	52.20	49.83	18.51	19.32	4514	885
U 98	32.53	31.79	42.39	41.20	25.08	27.01	3972	733
99	35.89	37.46	**34.60**	**29.10**	**29.51**	**33.44**	3569	622
97	**13.11**	**11.64**	**3.77**	**4.97**	83.12	83.39	18340	2234
N 98	17.69	16.07	3.41	4.40	78.90	79.54	15706	1774
99	22.14	21.13	2.98	3.71	74.89	75.15	13498	1481

force states is comparable across the two questionnaires. Therefore an estimate of a bias due to the non-coverage of the residential movers in the SOEP may be taken as a estimate of the corresponding bias in the MC. Some care has to be taken in the case of an analysis of foreigners. Here the SOEP may understate the rate of residential mobility.

3 The bias due to the missing information on residential movers

A labor force flow is defined as the probability of individuals moving from one economic activity state (E=employed, U=unemployed, N=not in the labor force) at t_1 to another state (or possibly the same state) at t_2. If A denotes the activity status at time t_1 and B the activity status at time t_2 the interest of change analysis[11] is the estimation of $P(B|A, X)$, where X is a set of observed covariates. Regional non-mobility is indicated by R. In the case of the MC $R = 0$ means that the person is not followed and B is not observed.[12] In this section we want to analyze whether conditioning on R = 1 is relevant. Thus we will estimate $P(B|A) - P(B|A, R = 1)$ or $P(B|A, X) - P(B|A, X, R = 1)$ if covariates are concerned. Note that the result may be different for turn-over tables with or without covariates.

Below we will use the nomenclature of missing data types, which was coined by Rubin (1976) for likelihood analysis. Within the given context R may be called:

a) Missing completely at random (MCAR) if:

$$P(R|A, B, X) = P(R) \tag{1}$$

[11] There is a slight difference to the analysis of Clarke/Tate (2002), who are interested in the unconditional probability $P(A, B)$. The use of A as a control variable will ease the comparability of SOEP and the MC.

[12] Remember that the situation in the MC is more complex than in the SOEP, as there appear persons who move into the dwellings left by the residential movers. This situation will be analyzed later.

b) Missing at random (MAR) if:[13]

$$P(R|A,B,X) = P(R|A,X) \tag{2}$$

c) Not missing at random (NMAR) if:

$$P(R|A,B,X) \neq P(R|A,X) \tag{3}$$

By virtue of the Bayes theorem the MAR condition is equal to:

$$P(B|A,X) = P(B|A,X,R=1) \tag{4}$$

Thus we are going to check whether missingness due to residential mobility is MAR. Note that there is no way to test the MAR hypothesis from the MC-data alone as we have no information on $P(B|A,X,R=0)$.

With the SOEP-data we use the following estimates:

$$\widehat{P}_{ALL}(B|A) = \frac{n_{B,A}}{n_A} \tag{5}$$

$$\widehat{P}_{IMMO}(B|A) = \frac{n_{B,A,R=1}}{n_{A,R=1}} \tag{6}$$

where n_A is the sum of the $n_{B,A}$ over B and $n_{B,A}$ is the number of observations with labor force status B at t_2 and A at t_1. Similar $n_{B,A,R=1}$ is the number of observations with profile A, B among the non-mobile persons (R = 1). Summing up the $n_{B,A,R=1}$ over B gives $n_{A,R,=1}$.

As $\widehat{P}_{ALL}(B|A)$ is the ML-estimate of $P(B|A)$ it is asymptotically efficient. Under the Null-hypothesis of MAR $\widehat{P}_{IMMO}(B|A)$ is a consistent estimator for $P(B|A)$. Hence the variance of the difference can be computed by the difference of the variances , see Hausman (1978). Thus in order to test the differences $\Delta = \hat{P}_{ALL} - \hat{P}_{IMMO}$ we compare Δ with its standard deviation $\sqrt{\sigma^2_{\hat{P}_{IMMO}} - \sigma^2_{\hat{P}_{ALL}}}$. Under the Null-hypothesis that both estimates are consistent, the standardized Δ^2 is χ^2 distributed with 1 DF. [14] [15] [16]

Table 4 presents the estimates of the labor force flows. The first column of Table 4 displays the estimates obtained using the full SOEP sample, the second column the estimates obtained using the subsample of immobile persons, and the third column the difference between the two estimates. Significant differences (p-value < 0.05) are indicated by bold figures.[17] In order to detect possible trends in the difference of estimates, we estimate labour force flows for the periods 1996/97, 1996/98, and 1996/99.

[13] Furthermore there must be no functional dependency between $P(A,B,X)$ and $P(R|A,B,X)$.

[14] One might have compared the difference $\widehat{P}_{IMMO} - \widehat{P}_{MOVE}$, where $\widehat{P}_{MOVE}$ is an analogous estimate based on data from residential movers. However, this difference is a poor estimate for the bias resulting from the non-coverage of the residential movers.

[15] The behaviour of the efficient and the consistent estimator under the alternative is exchanged here. In econometric application the efficient estimator becomes inconsistent while the consistent estimator remains consistent under the alternative. As the Hausman-test is evaluated under the Null hypothesis this change is irrelevant here.

[16] Note that in some cases the variance of the estimator $\hat{P}_{IMMO}(B|A)$ may be smaller than the variance of $\hat{P}_{ALL}(B|A)$. This is due to the asymptotical nature of the result. In these cases the test cannot be applied.

[17] We tested the hypothesis that the bias of the flow to U and E is equal to zero. This implies also a zero bias for the flow to the remaining category N.

Table 4: **Bias estimates for flows between labour force states based on the SOEP data (unweighted results).**

Flows	E			U			N		
from 96 to	FULL	IMMO	Δ	FULL	IMMO	Δ	FULL	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

Δ = estimate of absolute bias

Boldface figures: Significant differences $\hat{P}_{ALL} - \hat{P}_{IMMO}$

The results in Table 4 indicate that the MAR assumption will not hold for some transitions. For example, the flows from unemployment to employment (U → E) are underestimated using only the information from immobile persons. This finding is plausible as the new job might have caused a change of residence. Also transitions from being not active to employment (N → E) are underestimated as the change into employment is often seen as a reason to leave the parental home. Note that we did not use covariates. As age is an important variable for mobility we use a dummy variable for age ≤ 30 in the transition analysis too. The result is shown in Table 5.

Table 5: **Bias estimates for flows between labour force states based on the SOEP data. Control by age.**

Transition	$U \rightarrow E$			$N \rightarrow E$		
	ALL	IMMO	Δ	ALL	IMMO	Δ
	Age≤30					
97	52.43	52.12	0.31	25.98	24.16	**1.82**
98	55.09	56.02	0.93	37.86	33.33	**4.53**
99	65.69	64.05	1.64	50.07	46.28	**3.79**
	Age>30					
97	24.02	22.04	**1.98**	6.36	6.13	**0.23**
98	25.90	23.25	**2.75**	10.13	8.81	1.32*
99	30.28	28.78	1.50	12.72	11.28	**1.44**
	Total					
97	32.84	30.85	**1.99**	12.74	11.64	**1.10**
98	34.92	31.79	**3.13**	19.66	16.07	**3.59**
99	41.37	37.46	**3.89**	25.89	21.13	**4.76**

Δ = estimate of absolute Bias

Boldface figures: Significant differences $\hat{P}_{ALL} - \hat{P}_{IMMO}$

* indicates: the Hausman test did not apply because of negative difference of variances

For the transition $U \to E$ the control by the age dummy is quite effective. However, for the transition $N \to E$ there seems to be no control effect in the group below 30. This happens despite the fact that age has a clear impact on the transition rate $N \to E$: the transition rates between the two age groups differ by a factor of about 4. In this case the association between a transition $N \to E$ and residential mobility seems to be restricted to the age group under 30. This appears to be plausible as for young persons entries into the labour force are very often combined with a move from the parental home. At a higher age this reasoning holds not any longer.

4 Bias correction by inverse probability weighting (IPW)

The idea of this approach has some similarity with the Horvitz-Thompson estimator of design-based survey methodology, see Särndal et al. (1992 p.42). In this approach every unit i in the finite universe has a selection probability π_i that is determined by the sampling design. The characteristic y_i of unit i is then weighted by $w_i = 1/\pi_i$ to estimate population entities like, for example, a total $t = \sum_{i \in U} y_i$. Then t is estimated by the weighted sum of the sample observations $\hat{t} = \sum_{i \in s} w_i y_i$.

As the MC-panel file is produced by German Official Statistics it is offered together with weights to be used in standard calculations. In this case the weights are the inverse of the estimated probability of $R_i = 1$, i.e. the probability to stay at the same residence. This weight is applied to the subsample of residential stayers.

Despite the formal similarity with the design-based approach this calculus cannot applied here. This is due to the fact that the selection probabilities of residential movers are zero by design and the Horvitz-Thompson estimator does not cover this case.[18]

However, the IPW approach works also in a model based setting, see for example Robins et al. (1995). The basic argument is shown here within the context of a logit model, which can be easily extended to the general GEE-approach, see Liang/Zeger (1986) for the GEE-approach and Copas et al. (2004) for an application of the IPW-approach within the framework of GEE.

In order to simplify the notation we denote the set of covariates that explain transitions between labour force states by $V = (A, X)'$. We also restrict our analysis to a special state, say $B = b_0$. Thus the dependent variable Y of the logit model, indicates whether a change from A to $B = b_0$ takes place $(Y = 1)$ or not $(Y = 0)$. Then the logit model is given by:

$$\ln \frac{P(Y_i = 1|V_i)}{1 - P(Y_i = 1|V_i)} = \beta' V_i \tag{7}$$

If there were no missing data, the estimating equations are given by:

$$U_\beta = \sum_i V_i \left(Y_i - \mu_i\right) \tag{8}$$

with $\mu_i = (1 + \exp(-\beta' V_i))^{-1}$. Thus, if all values of Y and V are observed, the maximum-likelihood estimate $\hat{\beta}$ satisfies $U_{\hat{\beta}} = 0$. Moreover, at the true population parameter values, β_{true}, the expected value of the left-side of (8) is zero. This property guarantees the consistency the parameter estimate, see Cox and Hinkley (1974).

[18] The selection probabilities must be strictly positive for <u>all</u> units in the universe U. For this reason we use the term ”non-coverage” to indicate that the sample design does not cover this part of the population.

In the presence of residential mobility the IPW estimator is defined setting the sum of weighted scores

$$U_\beta(\pi) = \sum_i \frac{R_i}{\pi_i} V_i \left(Y_i - \mu_i\right) \tag{9}$$

with $\pi_i = P(R_i = 1|Z_i)$ to zero. Here, Z_i are socio-economic covariates, which are assumed to affect residential mobility.

Denote by $E_{R|Y,V,Z}(\cdot)$ the expectation with respect to the conditional distribution of R with (Y, V, Z) being fixed. Similar $E_{Y|V,Z}(\cdot)$ denotes the expectation with respect to the conditional distribution of Y with (V, Z) being fixed. Then, by the law of iterated expectations, we have:

$$\begin{aligned}
&E_{R,Y|V,Z}\left[\sum_i \frac{R_i}{\pi_i(Z_i)} V_i \left(Y_i - \mu_i\right)\right] \\
=&E_{Y|V,Z}\left[\sum_i E_{R|Y,V,Z}\left(\frac{R_i}{\pi_i(Z_i)} V_i \left(Y_i - \mu_i\right)\middle| Y_i, Z_i, V_i\right)\right] \\
=&E_{Y|V,Z}\left[\sum_i V_i \left(Y_i - \mu_i\right) \frac{1}{\pi_i(Z_i)} E_{R|Y,V,Z}\left(R_i | Y_i, Z_i, V_i\right)\right] \\
=&E_{Y|V,Z}\left[\sum_i V_i \left(Y_i - \mu_i\right) \frac{P(R_i = 1|Y_i, Z_i, V_i)}{\pi_i(Z_i)}\right] \qquad (10)
\end{aligned}$$

If the last term in the above expression equals one then for every fixed set (V, Z) the expected value of the right hand side of the above equation is zero. Thus on average with respect to R the IWP estimator solves the ML-estimation equation for the data set without losses due to residential mobility. This guarantees the consistency of the IPW estimator.

Thus we have to assume $P(R_i = 1|Y_i, Z_i, V_i) = \pi_i(Z_i)$. Usually we would assume the V-covariates, explaining transitions into labour force state $B = b_0$ to be a subset of the Z covariates.[19] In this case the above restriction implies that the residential mobility must not directly depend on the transition into the state $B = b_0$. Note that such a statement is more general than the MAR statement in (2) as the set Z of control variables for the residential mobility may be quite general. In principle the whole set of all observed variables may be included into Z.

The idea to use the IPW estimator also in the NMAR case is as follows: suppose that R depends intrinsically on the labour force state B, which means $P(R_i = 1|A_i, B_i, X_i) = P(R_i = 1|B_i)$. Then it is supposed that (A_i, X_i, Z_i) act as reasonable proxies for B_i, i.e. we assume $P(R_i = 1|B_i) \approx P(R_i = 1|A_i, X_i, Z_i)$ for each unit.

Such a hypothesis can be directly checked with the SOEP data. Table 6 displays the SOEP mobility rates depending on A and B. It is immediately seen that neither a main effect in A nor in B will adequately describe the dependency of the residential mobility on labour force states.

[19] This can be simply achieved by using all $V = (A, X)'$ variables for the prediction of the residential mobility. If the V variables are not used for the prediction of R, it is implicitly assumed that their regression coefficients are zero.

Table 6: **Mobility rates in the SOEP according to labour force flows starting in 1996.**

State at the beginning	End of period	State at the end of the period E	 U	 N
E	1997	0.11	0.12	0.13
	1998	0.21	0.25	0.20
	1999	0.29	0.25	0.25
U	1997	0.17	0.09	0.09
	1998	0.28	0.18	0.14
	1999	0.36	0.28	0.20
N	1997	0.16	0.16	0.06
	1998	0.32	0.28	0.12
	1999	0.37	0.37	0.17

However, such a statement may change in the presence of covariates. In a first step we augmented the mobility model from Table 2 with the labour force states B (Model 2). The result is shown in Table 7. In fact the significant impact of the previous labour force state A on residential mobility vanishes immediately, while the current labour force state B turns out to have a influence on the mobility behaviour. Furthermore, Model 2 results in a better fit of the data, as indicated by the pseudo R^2. However, the impact of the other covariates remains the same. Thus $P(R_i = 1|A_i, B_i, X_i) = P(R_i = 1|B_i)$ does not hold. For those who stay in the same labour force state $P(R_i = 1|A_i, X_i)$ is a reasonable approximation for $P(R_i = 1|B_i, X_i)$. However, for a person with transition $N \rightarrow E$ the residential mobility is understated by approximately 18 percent.[20] For the transition $U \rightarrow E$ we end up with an underestimate of mobility by 14 percent.[21]

As a main effects model the Model 2 is restrictive with respect to the impact for the labour force states on the mobility rate. From Table 6 we can conclude that there are 3 groups with different mobility behaviour. There is a group with high mobility rates consisting of the transitions $U \rightarrow E$,$N \rightarrow E$ and $N \rightarrow U$. The group with low mobility is given by the transitions $N \rightarrow N$ and $U \rightarrow N$ (Model 3). However, we also tried a different grouping by putting the transition $U \rightarrow N$ to the rest category with mean mobility intensity, (Model 4). All other transitions may be collected into a group with mean mobility.

It is interesting to see that all main effects of the labour force states immediately vanish, while the estimated effects of the other covariates remain stable across the different models.[22] From Model 4 we conclude, that apart from other covariate effects the residential mobility is linked to changes of the labour force status. Note that such a model is different from the test model of Clarke/Tate (2002) which is a main effect model similar to Model 2. This test model is taken as the truth in the simulation experiments to evaluate the bias reduction in the British LFS.

To analyse the corrective power of the IPW approach, we now compare the above

[20] Comparing the value 0 in Table 2 for state N with the value -0.1811 in Table 7 for state E.

[21] Comparing the value -0.0437 in Table 2 for state U with the value -0.1811 in Table 7 for state E.

[22] The increase of the constant is due to the reference category which is the transition $N \rightarrow N$ with high immobility.

Table 7: **Alternative models for residential immobility in the SOEP. Period 1996-1997. GEE analysis with household clusters.**

Variable		**Model 2**	**Model 3**	**Model 4**
Intercept		**2.1183**	**2.1674**	**2.1961**
		(0.1310)	(0.1326)	(0.1325)
Age ≤ 30		**-0.4658**	**-0.4532**	**-0.4495**
		(0.0685)	(0.0685)	(0.0683)
Age > 45		**0.4927**	**0.4918**	**0.4918**
		(0.0741)	(0.0739)	(0.0740)
Household size	1 person	**-0.4211**	**-0.4334**	**-0.4311**
		(0.1213)	(0.1213)	(0.1212)
	2 persons	-0.0396	-0.0452	-0.0438
		(0.1016)	(0.1016)	(0.1013)
Sex	male	-0.0051	-0.0045	-0.0034
		(0.0302)	(0.0302)	(0.0301)
Region	East-Germany	-0.0944	-0.0990	-0.0978
		(0.0931)	(0.0934)	(0.0932)
Education	vocational	**0.1406**	**0.1419**	**0.1440**
		(0.0569)	(0.0570)	(0.0568)
	tertiary level	**-0.1698**	**-0.1733**	**-0.1762**
		(0.0708)	(0.0703)	(0.0703)
Nationality	German	0.1445	0.1540	0.1549
		(0.0841)	(0.0842)	(0.0842)
Marital Status	Married	**0.3242**	**0.3197**	**0.3250**
		(0.0852)	(0.0853)	(0.0850)
Labour Force Status	Employment (96)	0.0018	-0.0061	0.1638
		(0.0862)	(0.1897)	(0.1963)
	Unemployment (96)	0.0152	0.0465	0.2527
		(0.1074)	(0.1318)	(0.1410)
	Employment (97)	**-0.1811**	0.1154	0.1076
		(0.0856)	(0.1145)	(0.0982)
	Unemployment (97)	-0.1591	0.1063	0.1211
		(0.1020)	(0.1488)	(0.1353)
	Δ_{middle}		-0.2670	**-0.5229**
			(0.2307)	(0.2355)
	Δ_{high}		**-0.4856**	**-0.6008**
			(0.1578)	(0.1464)
Observations (Individuals)		11’955	11’156	11’156
Log Likelihood		-3’437	-3’427	-3’428
Pseudo R^2		0.3303	0.3322	0.3322

Model 2: recent and current labour force states included (Main effects)
Model 3+4: Model 2 + different indicators for transitions
coefficients for logarithm of odds ratio $\mathrm{P}(R=1)/\mathrm{P}(R=0)$
Standard deviations in paranthesis

computed biases (Table 4) 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. Therefore, to assess the corrective power of the IPW approach for the MC we calculate an improvement rate (IR) which is based on the ratio between the correction for the MC and the estimated bias from the SOEP, i.e.,

$$IR = \frac{\hat{P}_{\text{IPW,MC}}(B|A) - \hat{P}_{\text{IMMO,MC}}(B|A)}{\hat{P}_{\text{FULL,SOEP}}(B|A) - \hat{P}_{\text{IMMO,SOEP}}(B|A)}$$

The IR gives the proportion of the bias corrected by using the above weights. Here, we assumed that the bias in the SOEP and the MC is of the same size, i.e, $\hat{P}_{ALL}(B|A) - \hat{P}_{IMMO}(B|A)$ is the same in the SOEP and the MC. According to the interpretation of the IR we distinguish following 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 equal zero indicates complete failure of the bias reduction. Finally a rate larger than 1 implies a correction in the right direction but this as too big amount. We call this over-correction. An over-correction of 2 corresponds to the same absolute bias with no correction, but it has the opposite sign.

Table 8: **Bias reduction expressed by ratio (bias – correction)/bias (SOEP and MC data).**

t	Bias	(Bias–correction)/Bias		
		SOEP	MC	MC*
			$U \to E$	
1997	1.98	0.49	0.46	0.59
1998	3.13	0.54	0.64	0.80
1999	3.91	0.87	0.69	0.80
			$N \to E$	
1997	1.10	0.55	0.52	1.00
1998	3.59	0.41	0.48	0.70
1999	4.56	0.67	0.69	0.97

Table 8 presents the performance of the IPW approach only for the labour force flows, where the substantial bias occurred, i.e. $U \to E$ and $N \to E$. The weights were computed by the estimated non-mobility scores according to the model from Table 2. The first column in Table 8 (Bias) shows the original bias in percentage points, the second column (SOEP) the corrected proportion of the bias for the SOEP and the third column (MC) the corrected proportion of the bias for the MC. In column MC* we present the results by using the longitudinal weights produced by the statistical office.[23] These weights were generated by calibration to population totals at the beginning and

[23] Because of the large number of calibration constraints we were not able to produce the same results for the SOEP. Results for the SOEP with a simplified calibration scheme can be found in Marek (2005).

the end of the reference period, including age groups, sex, regional strata and foreigner status. Furthermore there was a calibration to the number of births, deceased persons, divorces and marriages to took place in between, see MC panel User's Guide (2006).

The results in Table 8 reveal that all the biases are reduced to some extent. For example, in the case of the flow $U \rightarrow E$ the resulting correction lies between 49 percent (1997) and 87 percent (1999) for the SOEP and between 46 percent (1997) and 69 percent (1999) for the MC. Note that the correction rates for the SOEP and the MC are similar, re–affirming our general approach to choose the SOEP as an evaluation instrument for the MC. The calibration approach appeared to be very effective in reducing the bias, resulting in a reduction rate of about 80 percent. As this approach uses a different methodology[24] it does not compare directly with the IPW estimator. However, these weights were constructed for a use independent of the variable of interest, i.e. they were not designed with special reference to labour force transitions. So the user of the MC panel may be well off by using the calibration weights also for other analyses.

Finally we may ask whether it is helpful to use control variables like age in tabulations, like in Table 5, together with the weighting variable. However, as age was already used for the construction of weights there was no additional bias reduction.[25]

5 Loglinear Approach

In contrast to the weighting approach where only information of residential stayers are used in the analysis, the loglinear approach makes also use of information of mobile persons. Here we introduce two response indicators, R and S, representing whether A or B are missing or not (0 for missing; 1 for not missing). If we use the SOEP to indicate missingnes due to residential mobility only variable B can be missing. Thus, in the SOEP we have only one response indicator S. However, in the MC we have persons that move into the dwellings of residential movers. For them A will be missing.

The data then can be represented by the cell frequencies of a hypothetical complete data contingency table, cross-classified by A, B, R, and S for the MC and by A, B, and S for the SOEP. Table 9 presents this data structure for the MC with A and B having three categories (E,U,N). In the case of the SOEP we only have a table completely classified by A and B and a marginal table classified by A only.

For residential stayers the observed data are the cells of the A*B 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 have not moved at t_1, but move away 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. As the table is incomplete, a fully saturated log-linear model is not identifiable.

It is the basic idea of this section to use information from the SOEP to gain restrictions for the impact of A and B on R and S. In principle it would be possible to analyse the two data sets jointly. But this might be regarded as a too mixed use of two different surveys. However it might be accepted to use the functional relationship of labour force states on the residential mobility known from the SOEP also for the MC.

The joint distribution[26] $P(A, B, R, S)$ can be factorized as $P(A, B)P(R, S|A, B)$.

[24] We used the GREG estimator in the framework of the design-based approach, see Särndal et al. (1992, Chapter 6).

[25] The results may be obtained from the authors upon request.

[26] In the case of a covariate X we have $P(A, B, X, R, S) = P(B|A, X)P(A, X)P(R, S|A, B, X)$. Here we have to specify a model for $P(R, S|A, B, X)$.

Table 9: **A 3 × 3 Table with data partially classified on both variables.**

R=1 S=1 Status		t_2			R=1 S=0
		E	U	N	
	E	$n(EE)$	$n(EU)$	$n(EN)$	$r(E+)$
t_1	U	$n(UE)$	$n(UU)$	$n(UN)$	$r(U+)$
	N	$n(NE)$	$n(NU)$	$n(NN)$	$r(N+)$
R=0	S=1	$s(+E)$	$s(+U)$	$s(+N)$	

The observed likelihood corresponding to the above model is given for the MC by

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

and for the SOEP by

$$\begin{aligned} L = &\prod_{i\in S_{11}} P(A,B)P(R=1,S=1|A,B) \\ \times &\prod_{i\in S_{10}} \sum_A P(A,B)P(R=1,S=0|A,B) \end{aligned} \tag{12}$$

where S_{RS} refers to the set of individuals with response pattern (R, S).

To ensure that the model parameters can be estimated[27], the model $P(R, S|A, B)$ must be constrained in accordance with some assumption about the reasons why persons move. Here, we assume that the breakdown of residential movers into persons who move out of MC homes and persons who move into MC homes is due to the randomization of the MC sample. As each move out is also 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 for the $P(R = 1, S = 0|A, B)$ from the SOEP. From Table 6 we conclude that R depends on the flows from A to B with three groups of transitions for high, low and middle mobility. We will use three alternative sets of restrictions. For $alt1$ and $alt2$ we use the same grouping for high mobility. As in the previous section these were the transitions $U \rightarrow E$,$N \rightarrow E$ and $N \rightarrow U$. The group with low mobility is different for $alt1$ and $alt2$. For $alt1$ we used the transitions $N \rightarrow N$ and $U \rightarrow N$ to indicate low mobility rates. In $alt2$ the transition $U \rightarrow N$ was attributed to the group with mean mobility. Finally, in order to refer to the results

[27] For the estimation of the Loglinear model we used the software package LEM (1997), which is freely available.

from the IPW-approach, we used a main effect model for the current labour force state B, indicated by $alt3$. Note, that the model with a main effect of state A, results in the MAR assumption. In this case we obtain no correction of the transition probability estimates from the non-mobile persons.

Table 10 displays the estimated improvement rates for the two transitions with the largest bias. In order to demonstrate the effect of the two restrictions we used the MC data in two versions. In column MC we used MC data in the same way as SOEP data, we ignored the information from the marginal table $s(+E), s(+U), s(+N)$ that arises from persons moving into MC areas. In column MC* we used the full information for all movers including also the persons moving into MC homes.

Table 10: **Bias reduction expressed by ratio (bias – correction)/bias (SOEP and MC data).**

t	Bias	(Bias – correction)/Bias								
		SOEP			MC			MC*		
		$alt1$	$alt2$	$alt3$	$alt1$	$alt2$	$alt3$	$alt1$	$alt2$	$alt3$
		Transition $U \rightarrow E$								
1997	1.98	0.62	0.20	-0.09	1.80	1.30	-0.37	1.00	0.76	-0.81
1998	3.13	0.89	0.15	-0.07	1.78	1.09	-0.59	1.29	0.95	-0.66
1999	3.91	1.03	0.07	-0.05	1.68	0.94	-0.61	1.74	1.27	-0.51
		Transition $N \rightarrow E$								
1997	1.10	0.82	0.52	0.32	2.26	1.78	0.21	1.37	1.35	0.30
1998	3.59	0.68	0.33	0.27	1.29	0.93	0.08	1.04	1.05	0.20
1999	4.56	0.85	0.35	0.28	1.26	0.89	0.01	1.47	1.58	0.23

MC only with indicator R (move away)
MC* with both indicators R (move away) and S (move in)
$alt1$ indicates transition $U \rightarrow N$ being attributed to the low mobility group
$alt2$ indicates transition $U \rightarrow N$ being attributed to the mean mobility group
$alt3$ indicates a main effect model for the current labour force state B

Restriction set $alt1$ works well for the SOEP where the bias reduction is higher as in the case of the IPW approach, see column SOEP$_{alt1}$ in Table 10. However, these figures are too optimistic as we have chosen the model from the same data set. Furthermore the second model, designated by column SOEP$_{alt2}$ in Table 10, appears to be even more reasonable from Table 6, if we restrict our attention to the first period 1996–1997. However the consequences for the bias reduction in the SOEP are substantive. For $alt2$ the loglinear model performs even worse than the IPW approach. For example, for the transition $U \rightarrow E$ and period 1996 – 1999 we have an almost perfect bias reduction under $alt1$ while there is no bias reduction at all under $alt2$. From Table 7, which indicates that the current labour force state B is a more powerful predictor than the previous labour state A, we might be motivated to use $alt3$. However the resulting bias reduction is even worse than in the case of the IPW-approach.

This sensitivity with respect to the selection of restrictions applies also for the MC. However, here alternative 1 results in all cases in an over-correction of the bias, if we use only the MC data with move–away's (column MC$_{alt1}$). The additional information on the marginal distribution of the persons that move into the MC homes does correct this over-correction for the states E and U, but not for the state N, where the over-correction is enlarged, see column MC*$_{alt1}$. Finally alt_3 corrects for the transition $U \to E$ the bias is even in the wrong direction.

The tendency for over–correction is a typical feature of the non-ignorable models, see, for example, Chambers and Welsh (1993) or Little (1985). This is found also here. But also the ranking of the different alternatives is different for the SOEP and the MC*: for the SOEP $alt1$ seems to be the better choice while for the MC* $alt2$ should be preferred.

Finally we investigate whether the use of control variables stabilizes the behaviour of the NMAR models. As in the IPW case we used the age dummy, indicating persons with age above 30 years. This leads to the formulation of a model for $P(R|A, B, \text{Age} \leq 30)$ and $P(R|A, B, \text{Age} > 30)$. In both subgroups we use the same equality constraints of the model alternatives 1 and 2. However we did not assume the two probabilities to be equal across the age groups, which is not reasonable. Therefore, we have doubled the number of parameters in the mobility model. The high number of model parameters is an intrinsic problem of the treatment of covariates in loglinear modeling. One alternative is to use a main effect model. Here, we used a model with the main effects of Age and B which is indicated in Table 11 as $alt3$.

The result is given in Table 11. Because of the increased number of model parameters the exaggerations of the NMAR model estimates remain. For example, for the transition $U \to E$ in the period 1996 – 1999 the differences between the two model alternatives become even more pronounced for the age group older than 30 years compared to the total sample. Besides, for the total sample $alt1$ gives a perfect bias correction, while $alt2$ results in almost no correction at all. However in the age group > 30 $alt2$ results in a perfect bias correction while $alt1$ results in a tremendous over–correction. The main effect model works reasonably well for the transition $N \to E$, but for the transition $U \to E$ the corrections go into the wrong direction. Here the ”corrections” are stronger than in the case with no covariates. However in one case we have a strong overcorrection.[28]

All models fit the observed data quite well.[29] This indicates that the likelihood is quite flat over the different models. As a consequence the standard errors for the estimated probabilities $P(B|X, A)$ can be substantial. These standard errors are displayed in Table 11 in parenthesis under the estimate. When we compare the standard errors in column IMMO with the standard errors for the three alt-models we notice a substantial increase. This happens despite the fact that the number of observations that is used for the IMMO–model is smaller than the observations that enter the alt-models. Thus we notice that the switch to NMAR-models decreases the precision of estimates while offering consistency under the model. Furthermore the standard errors become as big as the bias itself. Therefore, the bias is regarded as statistically small.

[28] $age \leq 30$, period 1996 – 1998, transition $U \to E$.

[29] The models $alt1$ and $alt2$ result in a perfect fit with no degrees of freedom. Model $alt3$ results in a goodness of fit test with a p-value of 0.41 and two degrees of freedom.

Table 11: **Estimation of flows between labour force states. Control by age. Correction of estimates by different model alternatives.**

96	$U \rightarrow E$					$N \rightarrow E$				
to	ALL	IMMO	alt_1	alt_2	alt_3	ALL	IMMO	alt_1	alt_2	alt_3
	Age≤30									
97	52.43	52.12	52.39	51.46	53.09	25.98	24.16	25.33	24.88	25.67
	(2.84)	(3.10)	(3.53)	(3.46)	(3.64)	(1.56)	(1.64)	(2.35)	(2.20)	(1.82)
98	55.09	56.02	57.78	55.29	57.64	37.86	33.33	37.50	35.89	37.42
	(2.95)	(3.59)	(4.50)	(4.22)	(5.04)	(1.80)	(2.06)	(4.45)	(3.78)	(2.68)
99	65.69	64.05	65.65	62.49	68.39	50.07	46.28	51.37	48.91	53.57
	(2.87)	(3.88)	(5.31)	(4.62)	(6.87)	(1.92)	(2.44)	(6.85)	(5.34)	(3.85)
	Age>30									
97	24.02	22.04	24.19	23.41	21.63	6.36	6.13	6.97	6.75	6.24
	(1.63)	(1.66)	(1.96)	(1.91)	(1.65)	(0.60)	(0.61)	(0.82)	(0.77)	(0.62)
98	25.90	23.25	27.45	24.98	22.66	10.13	8.81	11.14	10.14	9.19
	(1.74)	(1.81)	(2.48)	(2.31)	(1.84)	(0.81)	(0.80)	(1.46)	(1.21)	(0.85)
99	30.28	28.78	36.04	30.70	27.79	12.72	11.28	15.73	13.40	12.13
	(1.87)	(2.09)	(2.85)	(2.64)	(2.19)	(0.95)	(0.97)	(2.06)	(1.56)	(1.06)
	Total									
97	32.84	30.85	32.08	31.24	30.68	12.74	11.64	12.54	12.21	11.99
	(1.49)	(1.55)	(1.83)	(1.78)	(1.60)	(0.68)	(0.68)	(0.93)	(0.87)	(0.71)
98	34.92	31.79	34.55	32.26	31.57	19.66	16.07	18.48	17.26	16.89
	(1.57)	(1.72)	(2.41)	(2.16)	(1.86)	(0.86)	(0.87)	(1.71)	(1.39)	(0.95)
99	41.37	37.46	41.74	37.74	37.25	25.89	21.13	25.19	22.78	22.48
	(1.66)	(1.94)	(2.46)	(2.45)	(2.24)	(1.00)	(1.06)	(2.54)	(1.84)	(1.20)

alt_1 : transitions $U \rightarrow N$ attributed to the low mobility group
alt_2 : transitions $U \rightarrow N$ attributed to the mean mobility
alt_3 : Main effect model for B
Standard deviations in parenthesis.

6 Conclusions

We used a methodology that carefully reflects the special type of missingness that is of interest here, namely the drop-out of persons due to residential mobility. In contrast to the approaches that use information entirely from the survey of interest we are able to make statements about the nature of missing data, even if it is not MAR. This is achieved by the use of a similar survey which covers residential mobility.

However there are some reasons why the residential mobility behaviour differs between the two surveys for certain subgroups. For example, we observed that the label ”foreigners” is mainly used in the SOEP to indicate migrant workers who have been staying in Germany for at least 12 years. It is reasonable that such a label does not reflect the mobility behaviour of foreigners in general, which is the usage in the MC. However, this is the only difference we found in the mobility behaviour of the SOEP and the MC.

With the help of the SOEP we could show that residential mobility depends intrinsically on changes of the labour force status even if we control for other variables like age, sex, education, marital status, size of the household and region. Thus the missingness pattern is intrinsically not MAR. Such a finding is not feasible with an analysis of the MC alone. This finding is also different from the standard assumptions of a main effect model either in A (see for example Stasny (1986)) or in B (see for example Clarke/Tate (2002)).

This analysis is only possible for those characteristics that are measured in both surveys in a comparable fashion. One may object then that the MC information is useless at all as it brings no new information that could not be retrieved from the SOEP. One answer is that such a bias approximation is meaningful also for variables which are slightly different from the variables with full comparability. For example, the case numbers in the SOEP may be too small to estimate labour force changes in certain areas or population subgroups while this can be done with the MC data. Furthermore we are able to evaluate the performance of different bias correction methods in a realistic fashion.

We have gained some evidence that the bias for the SOEP data and the MC data is very similar. We conclude this from the similarity of figures on residential mobility and from similar corrections for the IPW approach. The IPW approach results in a bias reduction of aboout 60 percent.

In our analysis age is the most important predictor for residential mobility. Therefore an age dummy for age ≤ 30 may help to reduce the immobility bias. However, this did not help to reduce the undercount for the transition $N \rightarrow E$ for young persons. In the majority of cases these transitions are induced by children entering the labor force after school. Often such an event is connected with a move from the parental home. In this case the parents will probably live in the same place and it will be possible to ask them for a proxy interview for their children. Proxy interviewing seems to be a reasonable strategy for many items of interest. Besides proxy interviewing is frequently used for the German MC.[30]

The most powerful correction procedure is the use of SOEP restrictions of NMAR–type in loglinear models for the MC. However, these estimators are under high risk of over-corrections. Therefore, this approach is regarded as hazardous. The use of control variables in loglinear modeling inflates the number of parameters and does not help to stabilize the estimates from the NMAR models.

What we recommend here, is to take a small subsample of residential movers in the MC. With such a subsample at hand it is possible to identify all parameters of the loglinear model without using restrictions from a different survey. If the sample size of such a mobility subsample is low, such an approach may be regarded as still convenient with respect to field work effort. However it can be realized only for future realizations of the MC.

For the MC panel of the past years it may be worthwhile to use the IPW approach as a robust method.

We demonstrated the gains from our approach only for two bias reduction routines. We did not evaluate all possible approaches for bias reduction, for example we did

[30] About 25 – 30 percent of all interviews in the MC are proxy interviews. For pupils being in upper secondary level education in percentage of proxy interviews is about 80 percent, see Schimpl-Neimanns (2006).

not discuss the imputation strategies as described in Little/Rubin (2002, Chapter 10). However the weighting approach and the loglinear models for incomplete observed contingency tables are the first choice in our context, see for example Little/Rubin (2002, Section 15.7). Besides there are some reservations to impute large proportions of a survey produced by official statistics.

Within the framework of Official Statistics calibration is a standard routine for bias reduction, see Lundström/Särndal (2005). The theoretical framework of this approach is the design-based approach. In our example this approach proved to be very effective in reducing the non-coverage bias. As these weights were produced without reference to a certain analysis the users may be well off to use it also for other analyses.

References

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

BEHR, A., BELLGARDT, E., RENDTEL, U. (2005). Extent and Determinants of Panel Attrition in the European Community Household Panel. *European Sociological Review* **21(5)** 489–512.

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. , CHAMBERS, R. (1998). Estimating Labour Force gross Flows From Surveys Subject to Household-level Nonignorable Nonresponse. *Survey Methodology* **24** 123-129.

CLARKE, P.S., TATE, P.F. (1999). Methodological Issues in the Production and Analysis of Longitudinal Data from the Labour Force Survey. GSS Methodology Series No.17
www.statistics.gov.uk/downloads/theme_other/GSSMethodology_No_17_v2.pdf

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.

COPAS, A.J., FAREWELL, V.T., MERCER, C.H., YAO, G. (2004). The sensitivity of estimates of the change in population behaviour to realistic changes in bias in repeated surveys.. *Journal of the Royal Statistical Society A* **167** 579–595.

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

EHLING, M., RENDTEL, U. (2004). *Harmonisation of Panel Surveys and Data Quality.* Statistisches Bundesamt.Wiesbaden

GERMAN STATISTICAL OFFICE (2006). MC-Panel User's Guide 1996 - 1999.
URL: www.destatis.de/download/mv/MZP9699_Handbuch.pdf.

HAISKEN-DENEW, J., FRICK, J. (2005). Desktop Companion to the German Socio-Economic Panel Study (GSOEP). Version 8.0 - Updated to Wave 21, DIW, Berlin.

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.

KROH, M., SPIESS, M. (2006). Sample Sizes and Panel Attrition in the German Socio-Economic Panel (GSOEP) (1984 until 2005). Data Documentation 15.

LEM (1997). A general program for the analysis of categorial data. Tilburg University. http://spitswww.uvt.nl/web/fsw/mto/lem/manual.pdf

LIANG, K.Y., ZEGER, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

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

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

LUNDSTRÖM, S., SÄRNDAL, C.-E. (2005). *Estimation in Surveys with Nonresponse.* Wiley, New York.

MAREK, I. (2005). Weighting adjustments in the presence of non-coverage due to residential mobility in the German Microcensus-Panel. Arbeitspapier Nr.10 Projektgruppe MZ Panel. URL: www.destatis.de/download/d/mv/arbeitspapier10.pdf

MILLER, M.E., TEN HAVE, T.R., REBOUSSIN, B.A., LOHMAN, K.K., REJESKI, W.J. (2001). A marginal model for analysing discrete outcomes from longitudinal surveys with outcomes subject to multiple-cause nonresponse. *Journal of the American Statistical Association* **96** 844–857.

RENDTEL, U. (1995). *Panelausfälle und Panelrepräsentativität.* Campus Verlag, Frankfurt, New York.

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.

SÄRNDAL, C-E., SWENSSON, B., WRETMAN, J. (1992). *Model assisted Survey Sampling.* Springer Verlag, New York.

SCHIMPL-NEIMANNS, BERNHARD (2006). Zur Datenqualität der Bildungsangaben im Mikrozensus.. ZUMA-Arbeitsbericht Nr. 2006/03. Mannheim: ZUMA

SISTO, J. (2003). Attrition Effects on the Design Based Estimates of Disposable Household Income. Chintex Working Paper No.9.
URL: www.destatis.de/chintex/download/paper9.pdf

STASNY, E. A. (1986). Estimating Gross Flows Using Panel Data With Nonresponse: An Example From the Canadian Labour Force Survey. *Journal of the American Statistical Association* **81** 42–47.

TATE, P.F. (1999). Utilizing Longitudinally Linked Data from the British Labour Force Survey. *Survey Methodology* **25** 99–103.

WAGNER, M., MULDER, C.H. (2000). Wohneigentum im Lebenslauf: Kohortendynamik, Familiengründung und sozioökonomische Ressourcen. *Zeitschrift für Soziologie* **29** 44–59.