The impact of the Non-coverage of Residential mobility. A comparison of the German Microcensus panel with the SOEP.

Ulrich Rendtel
Freie Universität Berlin

Edin Basic
Freie Universität Berlin

Stefan Grosenick
Freie Universität Berlin

Abstract

The German Microcensus can be seen as a rotating panel, where the units stay for four observations in the survey. Because of the very high case numbers and the mandatory participation it appears as a valuable data base for short duration analysis. However, the Microcensus is by design a sample of dwellings. Consequently there is no information on participants after they moved from the dwelling. We investigate how the use of the German Socio-Economic Panel (SOEP) can help to measure the non-coverage bias of the Microcensus. The analysis is demonstrated for transitions between the labor force states: employed, unemployed and inactive. This methodological work is a necessary prerequisite for the use of the Microcensus as a longitudinal data base. Besides, our analysis can be seen as a first longitudinal evaluation of the SOEP against the Microcensus.

Paper presented at the 6th International German Socio-Economic Panel User Conference -SOEP2004- Berlin, June 24-26, 2004

1 Introduction

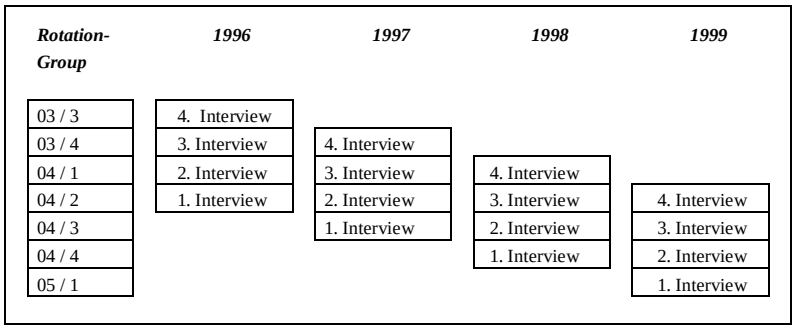
There is an increased demand on longitudinal data at the individual level. Usually such data are collected by panel surveys, which sample the same persons or households at subsequent points in time. A general overview of panel surveys and their methodology can be found in Kasprzyk et al. (1989).
The topics of the longitudinal analysis vary considerably and have a strong impact on the length of the panel and the sampled units. For example, analysis over the life cycle base on samples of birth cohorts and last in the ideal case from the birth until the death of its members. The sampling entities are persons. Things are more complicated in the analysis of welfare and the labour market. Here the natural entity for welfare analysis is the household, although households are no time-stable sampling entities. The time scope is covered by short term intervals. For example, the European Union Survey of Income and Living Conditions (EU-SILC), which is going to be launched in 2004, requires a time-interval of 4 years as a minimum. The Survey of Income and Progress Participation (SIPP), which is run by the US Census Bureau, uses a rotating panel, where the participants stay only 2 and a half year in the survey.
Household panels were mainly run by academic institutes on a voluntary basis with sample sizes of about 5000 households. This holds for example for the Panel Study of Income Dynamics (PSID) which was started in 1968 by the US Survey Research Center, the German Socio Economic Panel (SOEP), launched 1984 by the German Institute of Economic Research (DIW) and the British Household Panel Study (BHPS), run by the University of Essex since 1991.
These panels suffers from two drawbacks. First, because of the voluntary participation, as there was a substantial initial nonresponse of about one third of a sample, which was followed by non-participation in later waves, called panel attrition. The cumulative effect of panel attrition reduce the case numbers and this aggravates the problem of the representation of rare events, like drawing of social aid payments

or very high incomes, in the sample.
Thus there may be too small case numbers to come to a conclusive evidence in the analysis of interest. The voluntary participation is often regarded as an uncontrolled source of a bias in the results. Indeed, there is evidence that panel attrition is selective, see Fitzgerald et al. (1998) and Rendtel (2002).
In Germany both objections can be met by the annual micro census (MC), which can be merged to a four year rotating panel. Its longitudinal size is about ten times larger than the academic national counterpart, the SOEP. Furthermore the participation in the MC is mandatory. Hence selective non-response is reduced to a minimum level. The MC as a cross-section belongs to the standard program of German official statistics. Therefore no substantive extra costs arise from the longitudinal use of the data base. However, certain details which are due to the cross-sectional use of the MC cause problems in a simple longitudinal use of the MC.
The most serious problem arises from the fact that residential movers are not followed in the MC. This features saves field costs and is applied also in other regular repeated surveys like, for example the British Labour Force Survey (LFS), which could also be merged to a longitudinal data base. Therefore the methodological considerations to overcame the non-coverage problem in the longitudinal use of the MC are of general interest.
The paper is organized as follows: Section 2 displays the main sampling features of the MC. Section 3 describes the empirical difficulties with the longitudinal merging and displays the extent of the regional mobility. In Section 4 we present some results on the non-coverage bias from employment staten. These results base on the SOEP-data. Section 5 discusses several strategies to cope with the non-coverage of the movers. Section 6 gives an outlook.

2 The German Micro Census (MC)

In this section we give a comprehensive description of the German MC. For a detailed description see Rendtel/Schimpl-Neimanns (2001).
The MC is a one percent sample of the German population. The participation is mandatory although the response to some items is optional. The survey sampling uses extensive regional stratification and, within regional strata, stratification according to house size. Within strata clusters of small areas were sampled, containing on average 9 households. The sampling plan is an equal probability sample of clusters within a stratum.
The original sample is divided into 4 rotation groups. Each rotation group is four times interviewed with a one year interval. At each year one rotation group is replaced by a new rotation group. The rotation scheme is displayed in Figure 1.

Figure 1: The rotation scheme of the German micro census



The rotation groups are numbered by sample and rotation number. For example, rotation group 2 of sample 4 remains in the MC from 1996 to 1999. The motivation for the use of the rotation scheme is two-fold: First, it reduces the costs for sampling and field work, as each sample is used for 4 occasions. Second, by virtue of the 75 percent overlap of subsequent cross-sectional samples the sampling variance is reduced if changes of population totals are estimated.

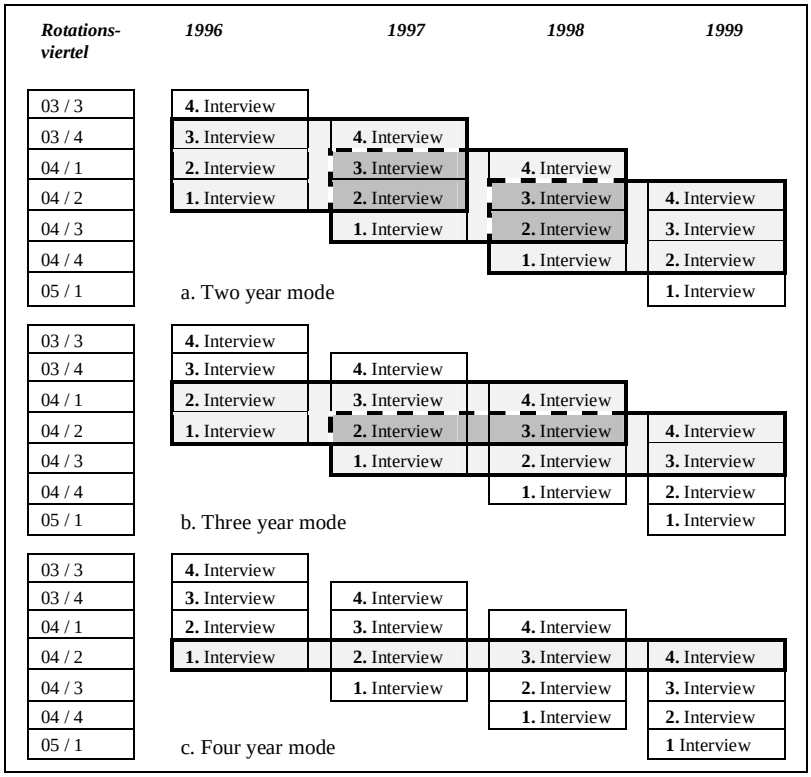
Note, the longitudinal analysis was not the argument for the use of the rotation scheme.
An important feature of the MC is the area sampling. According to area sampling the dwellings are sampled and residential movers are not followed to their new homes. Instead, new persons who move into the dwellings of the residential movers enter the MC sample. The reason for the application of this strategy is the cost reduction that is achieved by the preservation of the local sampling clusters and the saving of follow-up costs of residential movers.

3 The MC as a longitudinal data base

Although the MC has the potential to produce longitudinal information over four measurements this use was blocked by German legislation[1] until 1996. Depending on the length of the longitudinal reference period different samples of different sample sizes can be constructed from the rotation scheme, as shown in Figure 2.

[1]The argument was that the merge of the cross-sections generates more information than the sample persons were obliged to give by law.

Figure 2: Different longitudinal files built from German Micro Census



The largest reduction in sample size is achieved for the four years sample. However, this is still a 0.25 percent sample of the population.

The German Statistical Office has merged the cross-sections 1996 to 1999 to a longitudinal sample see Heidenreich (2002). There appeared several technical problems from the cross-sectional nature of the data base and the questionnaire. A prerequisite to a use of the MC as a longitudinal data base is a one to one matching of persons and households across the cross-sections. However, in the past no time consistent individual identification numbers for persons and households have been used. Only the identification number of the cluster was used in a time consistent fashion. Such practice is due to the separation of the questionnaire and the information on the name and the address of the respondent which is due to privacy legislation. So it was tried to match the persons by cluster number, their

household membership, their year of birth and gender and their sequence number in the household. This problem is aggravated by the fact that between two waves persons may leave household or enter it. In the extreme case, all persons left a dwelling and the new people moved into dwelling. Also demographic losses by death and births may effect the household composition. Nevertheless the simple above identification rule yielded satisfactory results. For the rotation group that remained in the MC from 1996 to 1999 the balance looks as follows: Out of the 154 000 persons at the start (1996) 41 900 persons were not present in the 1999 wave. 80 percent of these losses occurred because the entire household moved away. In 15 percent persons moved out of an ongoing household. Finally 5 percent of the persons deceased. This outflow analysis is to be set against an in-flow analysis: There are 155 000 persons at the end of the reference period. 42 800 persons did not occur at the starting period. Out of these gains 82 percent moved into a MC dwelling as a completely new household. 10 percent of the gains moved into existing households and 8 percent were born into existing households. Thus in this rotation group there are 196 900 persons: 112 200 stayers, 41 900 move-out persons and 42 800 move-in persons. The stayers amount only 57 percent of all persons in this rotation group. This clearly marks the importance of the mover group.

4 The bias due to the non-coverage of residential movers

The bias due to the non-coverage of the residential movers cannot be derived from the MC alone. Therefore the German Socio economic panel (SOEP), an academic household panel started in 1984 covering residential mobility, was used; see SOEP Group (2001) for a description of a SOEP. Here we compare the transitions between labor force status for residential movers and non-movers. The focus is on the difference of transition rates including residential movers

and stayers, called ”All” in the subsequent tabulations, and transition rates based only on residential stayers. The difference between the two rates is an estimate of the bias due to the non-coverage of the residential mobility in the MC.
In Table 1 we compare transitions between employment (E), unemployment (U) and not being in the labour force (N). The first

Table 1: **Transition between different types of labour market status. (Source: SOEP, 1996/97, 1996/98, 1996/99)**

Transitions		E			U			N	
from 96 to	All	Stayers	Movers	All	Stayers	Movers	All	Stayers	Movers
	97	87.97	88.43	83.60	5.30	5.15	6.78	6.72	6.42	9.62
E	98	85.45	86.03	82.96	5.82	5.63	6.63	8.73	8.34	10.41
	99	83.45	82.89	85.10	5.75	5.82	5.54	10.80	11.29	9.36
	97	36.97	33.80	61.64	47.74	49.82	31.51	15.29	16.37	6.85
U	98	40.37	36.67	53.38	38.54	41.58	27.82	21.10	21.75	18.80
	99	39.12	34.43	51.32	32.91	33.16	32.24	27.97	32.41	16.45
	97	9.09	8.33	20.62	1.54	1.46	2.75	89.37	90.21	76.63
N	98	13.60	11.61	28.05	1.44	1.42	1.55	84.96	86.97	70.41
	99	15.39	12.77	28.27	1.26	0.93	2.89	83.35	86.30	68.84

E: employment U: unemployment N: not in labour force

column in Table 1 displays the transition rates for all persons, the second column the transition rates for residential movers while the third column displays the transition rates for the stayers. In order to find out if there are any trends in transition rates we considered the transitions between 1996/97, 1996/98 and 1996/99. For some transitions there are substantial differences. For example, the transition from unemployment to employment is more frequent among residential movers (61.64 percent) than among residential stayers (33.80), which is plausible as the new employment might have caused a residential move. Thus, the resulting bias from the omission of the movers is 3.17 percent points. This trend is stable over time, if not increasing. Also for the transition from inactivity to employ-

ment we observe large differences between residential stayers and movers. Here we also observe an increasing trend in the bias.
These results indicate that there is an apparent tendency to overestimate stability in the labour market status if only stayers are regarded. Thus, we observed that stayers seem more likely to remain unemployed or not in labour force, and less likely to exit once they are unemployed or not in labour force.
To asses the non-coverage bias, we carry out the Hausman-test to test whether the difference between the estimates using only the information of stayers (p_{immo}) and the estimates using the information of attriters as well as respondents (p_{all}) is significant. The estimate of the bias is

$$\hat{b}(p) = \hat{p}_{\text{immo}} - \hat{p}_{\text{all}}$$

The hypothesis $b(p) = 0$ is tested against the alternative $b(p) \neq 0$ making use of the asymptotic result that the covariance matrix of the difference Σ_{diff} between a consistent estimator under the null-hypothesis (p_{immo}) and an efficient estimator (p_{all}) is given by their difference:

(1) $$\Sigma_{\text{diff}} = \Sigma_{\text{immo}} - \Sigma_{\text{all}}$$

The Hausman-test statistic is then calculated as

$$t = (\hat{p}_{\text{immo}} - \hat{p}_{\text{all}})' \Sigma_{\text{diff}}^{-1} (\hat{p}_{\text{immo}} - \hat{p}_{\text{all}}) \sim \chi^2_{\text{k}}$$

Table 2 displays the results of the Hausman test. Since empirically the standard deviation of the less efficient estimated parameters are in some cases smaller than the standard deviation of the estimation using the full sample, the Hausman test can not be applied. These cases were marked by ”n.a.” in Table 2. According to the Hausman-test in Table 2 we found that especially for transitions from employment to employment, from unemployment to employment and from not active to employment exhibit a non-coverage bias.

Table 2: **Results of the Hausman test**

HAUSMAN TEST

status96	status97	All	σ_{all}	Stayers	$\sigma_{stayers}$	chi-square	p-value
E	E	0.8797	0.0040	0.8843	0.0041	19.47	0.0000
E	U	0.0530	0.0027	0.0515	0.0028	4.30	0.0381
U	E	0.3697	0.0190	0.3380	0.0198	33.06	0.0000
U	U	0.4774	0.0197	0.4982	0.0210	8.54	0.0034
N	E	0.0909	0.0042	0.0833	0.0042	67.95	0.0000
N	U	0.0154	0.0018	0.0146	0.0018	17.53	0.0000
status96	status98						
E	E	0.8545	0.0045	0.8603	0.0049	8.65	0.0033
E	U	0.0582	0.0030	0.0563	0.0033	2.02	0.1554
U	E	0.4036	0.0200	0.3667	0.0222	14.30	0.0002
U	U	0.3854	0.0198	0.4158	0.0223	7.42	0.0064
N	E	0.1360	0.0053	0.1160	0.0052	n.a	n.a.
N	U	0.0144	0.0018	0.0142	0.0019	0.05	0.8164
status96	status99						
E	E	0.8345	0.0049	0.8289	0.0058	3.45	0.0632
E	U	0.0575	0.0031	0.0582	0.0036	0.148	0.7002
U	E	0.3912	0.0209	0.3443	0.0240	16.17	0.0000
U	U	0.3291	0.0201	0.3316	0.0237	0.04	0.8372
N	E	0.1539	0.0058	0.1277	0.0059	682.20	0.0000
N	U	0.0126	0.0018	0.0093	0.0017	n.a.	n.a.

In the following, we are interested whether the use of some control variables reduces this non-coverage bias. We present the results of several logit analyses for both groups, namely the group of stayers as well as movers and the group of stayers only. The key idea of these analysis is that above differences in transitions between residential movers and stayers might arise from differences in other observed characteristics (i.e. socio-demographics structure). The sets of explanatory variables are those commonly used in this context: Age, sex, region, school/education level, duration of unemployment and

number of children in household. The estimates of model parameters are summarized in Tables 3, 4 and 5. The first column (All) in Tables 3, 4 and 5 shows the results obtained using movers as well as stayers since second column (Immo) shows the results obtained using stayers only. The third column (Δ) in Tables 3, 4 and 5 displays the results of the Hausman test. Significant estimates are printed in boldface. In Table 3 we observe that the younger persons have a significantly higher probability of making transition from inactive to employment state in comparison to persons older than 30 years. Men are significantly more likely to move from inactive to employment state than their female counterparts. The higher the school/education level the more likely individuals are to make a transition from inactive to employment state. The residence in the western part of the Germany has a negative impact of making transition from inactive to employment state although this result is not statistically significant. Finally, the number of children in household has a positive effect on transition from inactive to employment state. With each further child the chance of making this transition increases by about 17%. The above interpretation holds for both samples.

To determine whether the effects of covariates differ by mobility, we apply a Hausman-test for the differences in coefficients. According to Hausman-test there are no significant differences between two samples. The results from Table 3 suggest that controlling for socio-demographic characteristics removes any mobility effect in the transition from inactivity to employment. This holds also for longer time-spans. Thus, there is no trend in the non-coverage bias of this passage if standard demographic control variables are used.

Table 3: Logit analysis: Transition from Inactivity to Employment

Logistic Regression: Transition from Inactivity to employment

from 96 to 97		All	Immo	Δ
Intercept		**-2.5037**	**-2.5391**	0.0354
		<.0001	**<.0001**	0.5168
Age	< 30	**1.3229**	**1.3763**	0.0534
		<.0001	**<.0001**	0.3177
Sex	Male	**0.4625**	**0.395**	0.0675
		<.0001	**0.0025**	0.2021
Region	West	-0.16	-0.1168	0.0432
		0.1968	0.3936	0.4572
School	without exam	-0.0796	-0.1656	0.086
		0.6407	0.3766	0.2659
	grammar school (Abitur)	0.0529	0.0683	0.0154
		0.734	0.6903	0.83
Education	without vocational training	**-0.5074**	**-0.5478**	0.0404
		0.0002	**0.0002**	0.5039
	tertiary level	0.0297	-0.0538	0.0835
		0.8998	0.8391	0.4929
Number of children		**0.1737**	**0.167**	0.0067
		0.0011	**0.0041**	0.7804
Number of observations		3158	2826	
from 96 to 98				
Intercept		**-2.2304**	**-2.2353**	0.0049
		<.0001	**<.0001**	0.9401
Age	< 30	**1.3315**	**1.3292**	0.0023
		<.0001	**<.0001**	0.9742
Sex	Male	**0.458**	**0.3631**	0.0949
		<.0001	**0.0044**	0.1421
Region	West	-0.1917	**-0.2983**	0.1066
		0.0955	**0.0252**	0.1133
School	without exam	-0.0686	-0.1643	0.0957
		0.6787	0.3627	0.2847
	grammar school (Abitur)	0.2291	0.2443	0.0152
		0.1045	0.1421	0.8635
Education	without vocational training	**-0.3116**	-0.2225	0.0891
		0.0109	0.1191	0.2254
	tertiary level	-0.093	0.0024	0.0954
		0.6715	0.9925	0.4477
Number of children		**0.2085**	**0.2132**	0.0047
		<.0001	**<.0001**	0.8695
Number of observations		2982	2434	
from 96 to 99				
Intercept		**-1.9742**	**-1.9596**	0.0146
		<.0001	**<.0001**	0.8243
Age	< 30	**1.6323**	**1.6925**	0.0602
		<.0001	**<.0001**	0.4222
Sex	Male	**0.2965**	0.2196	0.0769
		0.006	0.0854	0.2583
Region	West	**-0.2621**	**-0.3334**	0.0713
		0.0198	**0.0122**	0.3149
School	without exam	0.1428	-0.019	0.1618
		0.3357	0.9148	0.0982
	grammar school (Abitur)	**0.3314**	**0.3812**	0.0498
		0.0168	**0.024**	0.6058
Education	without vocational training	**-0.3668**	**-0.3298**	0.0370
		0.0026	**0.0234**	0.6429
	tertiary level	0.1004	0.0205	0.0799
		0.6274	0.9346	0.5692
Number of children		**0.2542**	**0.2652**	0.011
		<.0001	**<.0001**	0.6854
Number of observations		2738	2124	

Table 4 contains logistic regression model of the transition from unemployment to employment. In general, the results for this transi-

tion are similar to those of the previous example in Table 3. However, some results in Table 4 are worth mentioning. In Table 4 we included a new variable, duration of unemployment, which is highly significant. As expected, the long-term unemployed are less likely to become employed. Contrary to the Table 3 results there is no significant effect of sex and school on transition from unemployment to employment. Also here we found a not significant differences in the coefficient estimates between the two samples. This indicate that controlling for socio-demographic characteristics the non-coverage bias for the transition from unemployment to employment can be removed.

Finally, we repeated the same analysis for the case that employed persons stay employed (Table 5). In contrast to transitions from either inactive or unemployment to employment, we observe here a significant effect of region. West Germans are more likely to stay employed than their counterparts in East Germany. This effect is also stable over time. Older persons are less likely to become either unemployed or inactive. The higher the school/education level the more likely individuals are to stay employed. Also here the non-coverage bias is removed by the control variables.

Table 4: Logit analysis: Transition from Unemployment to Employment

Logistic Regression: Transition from Unemployment to Employment

from 96 to 97		All	Immo	Δ
Intercept		-0.2867	-0.2861	0.0006
		0.1487	0.1757	0.9936
Age	< 30	**0.8568**	**1.0118**	0.1550
		<.0001	**<.0001**	0.1241
Sex	Male	0.0734	-0.0083	0.0817
		0.704	0.9690	0.3815
Region	West	-0.1175	-0.1168	0.0007
		0.5877	0.6286	0.995
School	without exam	-0.1777	-0.1069	0.0708
		0.6331	0.7917	0.6562
	grammar school (Abitur)	0.1080	0.3396	0.2316
		0.7149	0.2964	0.0877
Education	without vocational training	-0.4148	**-0.5368**	0.1220
		0.0913	**0.0482**	0.2936
	tertiary level	0.8069	0.5941	0.2128
		0.0160	0.1116	0.1973
duration of unemployment	> 1 year	**-1.0612**	**-1.1035**	0.0423
		<.0001	**<.0001**	0.6436
Number of observations		533	452	
from 96 to 98				
Intercept		-0.1231	-0.1113	0.0118
		0.5652	0.6408	0.9108
Age	< 30	0.3380	0.3166	0.0214
		0.1062	0.1994	0.8706
Sex	Male	**0.5397**	**0.5272**	0.0125
		0.0067	**0.0260**	0.9218
Region	West	-0.3092	-0.3682	0.059
		0.1649	0.1591	0.6668
School	without exam	-0.3221	-0.4198	0.0977
		0.3978	0.3283	0.6220
	grammar school (Abitur)	-0.2836	-0.2372	0.0464
		0.3550	0.5062	0.7998
Education	without vocational traninig	0.2177	0.3219	0.1042
		0.3822	0.2703	0.4942
	tertiary level	**0.9858**	**0.9024**	0.0834
		0.0048	**0.0294**	0.7071
duration of unemployment	> 1 year	**-1.1598**	**-1.2401**	0.0803
		<.0001	**<.0001**	0.4892
Number of observations		497	377	
from 96 to 99				
Intercept		**0.5615**	**0.7188**	0.1573
		0.0237	**0.0128**	0.2855
Age	< 30	**0.4658**	**0.6346**	0.1688
		0.0407	**0.0237**	0.3036
Sex	Male	0.3532	0.2679	0.0853
		0.0955	0.2985	0.5608
Region	West	-0.3989	**-0.6869**	0.2880
		0.0978	**0.0169**	0.0669
School	without exam	-0.5235	-0.461	0.0625
		0.1945	0.3308	0.8018
	grammar school (Abitur)	-0.1365	0.0640	0.2005
		0.6782	0.8654	0.2795
Education	without vocational training	-0.2606	-0.2703	0.0097
		0.3284	0.3966	0.9558
	tertiary level	0.4885	0.3711	0.1174
		0.1832	0.3956	0.6201
duration of unemployment	> 1 year	**-1.2083**	**-1.2396**	0.0313
		<.0001	**<.0001**	0.8240
Number of observations		439	316	

Table 5: Logit analysis: Transition from Employment to Employment

Logistic Regression: Transition from Employment to Employment

from 96 to 97		All	Immo	Δ
Intercept		**1.5162**	**1.5605**	0.0443
		<.0001	**<.0001**	0.1670
Age	< 30	**-0.3672**	**-0.2587**	**0.1085**
		<.0001	**0.0006**	**0.0018**
Sex	Male	**0.2906**	**0.2794**	0.0112
		<.0001	**<.0001**	0.6970
Region	West	**0.1695**	**0.1764**	0.0069
		0.0208	**0.0284**	0.8362
School	without exam	-0.1413	-0.2159	0.0746
		0.3086	0.1476	0.1712
	grammar school (Abitur)	0.1620	**0.2771**	**0.1151**
		0.0969	**0.0129**	**0.0326**
Education	without vocational training	**-0.2862**	**-0.2962**	0.0100
		0.0003	**0.0006**	0.7854
	tertiary level	**0.5112**	**0.4709**	0.0403
		<.0001	**0.0003**	0.4717
Number of observations		7755	6869	
from 96 to 98				
Intercept		**1.2033**	**1.3129**	**0.1096**
		<.0001	**<.0001**	**0.0071**
Age	< 30	**-0.2184**	**-0.1901**	0.0283
		0.0004	**0.0122**	0.5183
Sex	Male	**0.2064**	**0.1469**	0.0595
		0.0004	**0.0299**	0.0915
Region	West	0.1155	**0.1910**	0.0755
		0.0860	**0.0150**	0.0629
School	without exam	-0.1363	-0.1348	0.0015
		0.2931	0.3607	0.9830
	grammar school (Abitur)	**0.2511**	**0.2486**	0.0025
		0.0044	**0.0193**	0.9655
Education	without vocational training	**-0.3538**	**-0.4777**	**0.1239**
		<.0001	**<.0001**	**0.0051**
	tertiary level	0.1773	**0.2547**	0.0774
		0.0776	**0.0336**	0.2365
Number of observations		7390	5850	
from 96 to 99				
Intercept		**1.1144**	**1.2242**	**0.1098**
		<.0001	**<.0001**	**0.0126**
Age	< 30	-0.0156	-0.0702	0.0546
		0.8136	0.3970	0.2745
Sex	Male	**0.2029**	0.1376	0.0653
		0.0007	0.0513	0.0830
Region	West	**0.1632**	**0.1630**	0.0002
		0.0191	**0.0474**	0.9950
School	without exam	-0.0399	0.0535	0.0934
		0.7682	0.7364	0.2642
	grammar school (Abitur)	0.1496	0.1393	0.0103
		0.0991	0.2011	0.8635
Education	without vocational training	**-0.6155**	**-0.6779**	0.0624
		<.0001	**<.0001**	0.2059
	tertiary level	**0.3517**	**0.3035**	0.0482
		0.0009	**0.0141**	0.4538
Number of observations		6795	5031	

5 Strategies to cope with the non-coverage of the residential mobility

In this section we will discuss some statistical tools that may correct the non-coverage bias of the residual mobility in the MC. At the present stage these tools have not been used empirically. So this section reflects our future intentions in making the MC a reliable instrument for longitudinal analysis. These tools use information from different sources.

The most valuable source is the SOEP, which covers residential mobility. However, its case numbers are low with respect to the MC, even if the MC is reduced to one rotation group in a four wave analysis. For the period 1996 to 1999 the SOEP amounts only one tenth of the MC longitudinal sample. Especially for rare longitudinal events the MC may be more informative as it delivers approximately ten times more events of interest.

We may dichotomize the population, and also the sample, into those without residential mobility during the reference period, called ”stayers”, and the rest with residential mobility, called ”movers”. For the stayers we know the distribution of the characteristic of interest at the start of the interval, at the end of the interval and all transitions in between. This holds not only for the SOEP but also for the MC. For the movers we know all three distributions for the SOEP. However, for the MC we know only the marginal distribution at start and at the end of the interval, but not the transitions in between. The knowledge of the marginal distribution at the end of interval is related to the fact that the MC sampling design includes the persons that move into the dwellings that have been left by the residential movers. Therefore the MC is representative for the population at the end of the reference period[2].

Finally for a reduced set of variables in the labour market there is a different source from the files of the labour force administration.

[2]These move-in persons are also persons with residential mobility. Their inclusion in the longitudinal sample results in a 100 percent over-representation of residential movers

The so-called ”Historik-Datei” (HD) contains longitudinal information about all persons that underlay payments into the German social security system. This excludes certain persons: self-employed persons, civil servants and persons above certain income limits. However, these characteristics are known from the MC and therefore we may use this data source to calibrate longitudinal information from the MC with the register information. For example, one may adjust the MC results with respect to the number of persons switching from unemployment to employment and/or vice versa. As it is typical for administrative data, the HD-files contain only very limited information. So in this data base one cannot distinguish residential movers and stayers.

The different sources of information are displayed in the following scheme. In each of the fields it is denoted whether there exists information from the three sources MC, SOEP and HD.

Figure 3: Different sources of information to cope with non-coverage problem

<table>
<tr><td>Distribution
of the characteristic</td><td>Marginal distribution at end:
MC, SOEP, HD</td></tr>
<tr><td rowspan="4">Marginal distribution at start:

MC, SOEP, HD</td><td>Transitions</td></tr>
<tr><td>Stayers: MC, SOEP</td></tr>
<tr><td>Movers: SOEP</td></tr>
<tr><td>Total: HD</td></tr>
</table>

In the following subsection we propose some strategies to cope with the non-coverage of residential mobility.

5.1 Mixture of MC and SOEP

This very simple strategy deletes all residential movers (in and out) from the MC and includes the mover group from the SOEP instead.

If the survey weights are given by the inverse of the selection probabilities, then valid population estimates are obtained by the weighted mixed MC/SOEP sample. Therefore it can be easily handled like every survey with weights.

However, the number of the SOEP-movers is quite small and the information of the MC-movers for the marginal distributions at the start and the end of the period is not used. Thus the use of information is inefficient. Furthermore, mixtures of surveys are often disliked because of minor deviations in item definitions which result in hard to control biases.

5.2 Use of weights

Here we estimate propensity scores for the residential mobility by adequate covariates. From the reciprocal propensity scores we compute weights for the stayers. The movers receive a zero weight which is equivalent to remove all movers (in and out) from the sample. This approach is very similar to the non-response treatment by response homogeneity groups, where propensity scores for non-response are estimated for different covariate combinations, see for example Särndal et al. (1992). The use of these weights for the stayers is therefore very similar to a traditional approach in non-response treatment.

However, this approach rests on implicit assumptions. It assumes that within the homogeneity groups, where equal weights are given to all its members, the distribution of the characteristic of interest is equal to movers and stayers. This assumption can be coined into a more formal expression. If R is the variable indicating whether a unit changes its residence or not and if Y is the outcome variable of interest and if the control variables for residual mobility are given by the vector X, then we assume conditional independence of R and Y given X. Of course, this relationship may vary for different outcome variables Y and therefore it may be desirable to choose variables in

the covariate set X that are closely related to Y. Note, however, that the covariate vector X has be known for movers and stayers. If the covariate set depends on Y, one ends up with item specific weights. From the point of design-based survey weights, which can be used for all outcome variables, item specific weights are a clear drawback.

In the weighting approach we use partly the information of the out-movers. However, the information of the in-movers is not used at all. Also for the out-movers the starting distribution of Y does not enter the estimation directly (only if it enters via the weights). This introduces some inefficiency in the use of available information.

Also the SOEP-information does not enter here. One might use the SOEP information to check whether the conditional independence assumption holds. This is not possible with the MC-data alone as the Y variable is not observed for R=Mover.

5.3 Calibration

Calibration is a different method that produces also weights. The general approach is to search for weights with a minimum distance to the design weights which guarantee the resulting estimates for certain variables coincide with known population totals, see Deville/Särndal 1992. Calibration is a general tool in the treatment of nonresponse, see Lundström/Särndal (2002). The fitting of two dimensional tables to given marginal distributions by the iterative proportional fitting algorithm is a special case of the calibration approach where the distance function is given by the entropy measure, see Deville/Särndal (1992).

There are two possibilities for calibration. First, we observe transitions for the movers on the basis of the SOEP. From the transitions we can calculate the marginal distributions at the start and at the end of the reference period. These distributions can be calibrated with the marginal distribution that arise from the MC move-out persons and the MC move-in persons. Note, that the procedure leaves the

distribution and weights of the stayers unchanged. The method uses efficiently the information for the movers that is found in the MC.

The second possibility uses the information in the HD file for selected variables. Here we can calibrate the estimated transitions from the MC to the known population totals from the HD-file. Note, that we calibrate here the longitudinal characteristics, like changes between employment status, for a subpopulation of all employees. In the previous approach we did calibrate the cross-sectional information at the start and at the end of the reference period. The second approach needs a starting value for the transitions including stayers and movers. Here we can use the estimates from the first calibration step. This exploits the information at hand in an exhaustive way.

This calibration strategy delivers weights that vary across variables as the marginal distribution depends on the variable of interest. Such a feature is uncommon among survey statisticians, who like to have only one set of weights that is used for all tabulations. In principle it is possible to use a weighting scheme that is related to a special set of outcome variables. However, there is no generally agreed set of such calibration variables as in the cross-sectional case, where gender and age-group are used frequently. It is up to further research to decide whether such all-purpose weighting schemes are useful.

5.4 Estimation of statistical models

The above procedures relate to the design-based approach of sampling theory. This approach is often used to estimate totals and proportions in the population. In the context of regression analysis where conditional distributions are analysed, the use of statistical models is standard. Besides regression analysis which is predominant for metric outcome variables, Logit analysis and loglinear models are used. The non-coverage of residential movers is a missing data problem in this context. Here the taxonomy of Rubin

(1976) into Missing completely at random (MCAR), Missing at random (MAR) and Not Missing at random (NMAR) is essential, see Little/Rubin (2002).

The MCAR case does not apply here, as the probability to move is not independent from covariates. In the NMAR-case the probability of the move depends on variables that are not observed. These are variables relate to causes after the last interview. In general such a relationship cannot be checked because the data needed for verification are just the ones that are not observed. However, with the SOEP-data at hand one can directly test whether changes, for example in the labour force status, have an impact on residential mobility.

In the case of two-dimensional transition tables the MAR property is equivalent to the equality of the transition probabilities, for example between labour force states, across the movers and the stayers. The numerical examples in tables 2 and 3 indicate[3] that the MAR assumption will not hold in some cases.

If the MAR condition holds, the transition behaviour for the movers and stayers can be regarded as equal[4]. Hence, the transition parameters may be simply estimated from the stayers. In more complex models we can use the general likelihood approach, where the likelihood is based only on the observed data, see Little/Rubin (2002).

In the case of linear regression models for panels this leads to mixed models that may be estimated by standard software, for example the SAS procedere Proc Mixed, see Verbeke/Molenberg (2000).

In the NMAR case there is an important model inhomogeneity between movers and stayers. There are two possibilities: Either one includes this inhomogeneity into the model, for example by using different transition matrices for movers and stayers or by estimating a marginal transition matrix, which is the mixture of transition matrices of the movers and the stayers. The estimation of both variants depends on results from the SOEP for the movers.

[3] A formal test on equality was not performed there.

[4] However, the starting distribution for movers and stayers can be different.

On the basis of MC data alone, the NMAR case is difficult to treat. In the case of longitudinal loglinear models Fay (1986) has proposed causal models, which use causal restrictions to identify the model parameters. Due to the missing subpopulation of the movers not all cells of the contingency table spanned by the loglinear model are observed. Therefore one needs restrictions to identify the model parameters. The causal restriction imply that future observations have no impact on previous observations.

5.5 Imputation

Imputation is a standard approach in design-based analysis, although its foundation is a good statistical prediction model. Imputation can be also used in the estimation of statistical models, see Schafer (1997). In contrast to the weighting approach which bases on a prediction model for the occurrence of residential mobility, the imputation uses a prediction model for the outcome variable. In a panel, predictions for metric variables may be very powerful. However, for discrete variables one has to predict changes which appears to be more difficult. The statistical methodology strongly votes for multiple imputation, see Schafer (1997) and Rubin (1987). In the multiple imputation for every missing value n=5 estimates are generated. The analysis is performed for the 5 complete datasets. From the variance of the 5 estimates one can easily derive the variance of the maximum likelihood estimate with missing observations.

If we use only MC data we have to assume the MAR assumption, as we have to predict the missing transitions for the movers from the observed transitions for the stayers. With the SOEP data for the movers we could use a prediction model that bases on the SOEP movers. Here the MAR assumption is no longer needed.

6 Outlook

Although the German MC was originally designed for cross-sectional purposes it can be used for longitudinal analysis up to four annual measurements. There are various statistical tools to treat the non-coverage problem of the residential movers. The most powerful means seems to be the use of out-of-sample information, either from the SOEP, which covers residential mobility, or from register data for subpopulations. Up-to-now these tools have not been used empirically and it still deserves some considerations which out of the alternative approaches should be recommended. It appears that an answer can be only given on the basis of empirical results.

The need to adjust the MC results for non-coverage of residential mobility disappears if residential movers are followed. This would be the easiest way to solve the problem in some future. As most moves are short distance moves the additional field costs should be not so dramatic. However, the administrative effort may be high as the field work is organized by 16 separate field organisations of the German Federal States. Hence a move across federal State borders induces a change of corresponding field organisation. However, one has to bear in mind that the alternative, the non-coverage of the residential movers, also incur costs that arise from the maintenance of the above mentioned statistical tools to the new MC waves.

In future revisions one should also enforce the longitudinal information of the MC questionnaire. Up to now most of the time related questions ask for the status at the week of the interview. If this question is changed into ”Have there been changes since the last year and when did they take place?”, one would have information on the complete time interval of the last year. Such a questionnaire design is especially well designed for recording spells in event history analysis. With such a design one of the main advantages of the MC-panel, the cumulation of rare events by high case numbers, becomes more powerful.

References

[1] Deville, J.C.; Särndal, C.-E. (1992): Calibration estimators in survey sampling, J. American Statistical Assoc., 87, 376-382.

[2] Fay, R. (1986): Causal models for patterns of Nonresponse, J. American Statistical Assoc., 81, 354-365.

[3] 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, 251-299.

[4] Heidenreich, H.-J. (2002): Längsschnittdaten aus dem Mikrozensus. Basis für neue Analysemöglichkeiten. [Longitudinal Data on the Basis of the German Microcensus,] Allgemeines Statistisches Archiv, 86, 213-231

[5] Kasprzyk, D.; Duncan, G.; Kalton, G.; Singh, M. (eds) (1989): Panel Surveys, Wiley, New York.

[6] Little, R.; Rubin, D. (2002): Statistical Analysis with Missing Data. Second Edition. Wiley, New York.

[7] Lundström, S.; Särndal, C.-E. (2002): The estimation in the presence of nonresponse and frame imperfections, Statistics Sweden, Stockholm.

[8] Rendtel, Ulrich (2002): Attrition in Household Panels: A Survey, CHINTEX Working Paper #4.

[9] Rendtel, U.; Schimpl-Neimanns, B. (2001): Variance Estimation for the Scientific Use File of the German Micro Census, Paper presented at the International Conference on Quality in Official Statistics, Stockholm,

[10] Rubin, D. (1976): Inference and missing data, Biometrika, 63, 581-592.

[11] Rubin, D. (1987): Multiple imputation for Nonresponse in Surveys, Wiley, New York.

[12] Särndal, C.-E.; Swensson, B.; Wretman, J. (1992): Model assisted Survey Sampling, Springer Verlag, New York.

[13] Schafer, J. (1997): Analysis of incomplete missing Data, Chapman and Hall, London.

[14] SOEP Group (2001): The German Socio-Economic Panel (GSOEP) after more than 15 years - Overview. In: Elke Holst, Dean R. Lillard und Thomas A. DiPrete (eds.): Proceedings of the 2000 Fourth International Conference of German Socio-Economic Panel Study Users (GSOEP2000), Vierteljahrshefte zur Wirtschaftsforschung, Jg. 70, Nr. 1, S. 7-14.

[15] Verbeke,G.; Molenbergs, G. (2000): Linear Mixed Models for Longitudinal Data. Springer Verlag, New York.

[16] Zühlke,S. (2003): Systematische Ausfälle im Mikrozensus-Panel: Ausmaß und Auswirkungen auf die Qualität von Arbeitsmarktanalysen. [Systematic drop out in the Micro Census Panel: Extent and Effect on the Quality of Labour Market Analysis] Allgemeines Statistisches Archiv, 87, 39-58.