


# Preparing the Swiss public-use sample for generating a synthetic population of Switzerland

**Conference Paper****Author(s):**

Müller, Kirill; Axhausen, Kay W. 

**Publication date:**

2012-05

**Permanent link:**

<https://doi.org/10.3929/ethz-a-007340012>

**Rights / license:**

In Copyright - Non-Commercial Use Permitted

**Originally published in:**

Arbeitsberichte Verkehrs- und Raumplanung 791

---

# **Preparing the Swiss Public-Use Sample for generating a synthetic population of Switzerland**

**Kirill Müller**

**Kay W. Axhausen**

**STRC 2012**

**May 2012**

**STRC**

12<sup>th</sup> Swiss Transport Research Conference  
Monte Verità / Ascona, May 2 – 4, 2012

STRC 2012

## **Preparing the Swiss Public-Use Sample for generating a synthetic population of Switzerland**

Kirill Müller, Kay W. Axhausen  
IVT  
ETH Zürich  
8093 Zürich  
phone: +41-44-633 33 17  
fax: +41-44-633 10 57  
{mueller,axhausen}@ivt.baug.ethz.ch

May 2012

### **Abstract**

A microsimulation model for transportation planning or land use requires disaggregate information on the population of the study area. This kind of data is not often available and has to be synthesized. For Switzerland, the Swiss Public-Use Sample (PUS) seems to be a suitable data source for such a synthetic population.

This paper describes the process of preparing the PUS so that it can be used for generating a synthetic population.

### **Keywords**

Population synthesis, Microsimulation, Households, Disaggregation, Hierarchical, Data Preparation, Public-Use Sample

# 1 Introduction

A microsimulation model for transportation planning or land use requires disaggregate information on the population of the study area. If an up-to-date full sample of persons and households is not available, this population must be synthesized (Müller and Axhausen, 2011b). Recent techniques use a disaggregate household sample for synthesizing a population of households with disaggregate data for all persons in each household (Lee and Fu, 2011, Müller and Axhausen, 2011a, Ye *et al.*, 2009, Bar-Gera *et al.*, 2009).

For Switzerland, the Swiss Public-Use Sample (PUS, Swiss Federal Statistical Office (2000b)) seems to be a suitable data source for such a sample. However, the dataset is a person sample and must be converted to a household sample first. This paper describes the preprocessing of the PUS that is required before the conversion to a household sample can be carried out. The household generation will be similar to that presented by Pritchard and Miller (2012).

Bürgle (2007) has also used the Public-Use Sample to generate a synthetic population. However, she considered only the overall household structure and not the individual attributes of the persons within the household.

The remainder of this paper is organized as follows. The next section presents the attributes available in the PUS. In the two subsequent sections, the preparation of the attributes at the person and household levels is described. An outlook outlines further steps. The appendix presents attributes from the dataset that were not treated and technical information on the preparation process.

## 2 Attributes

This section outlines the format of the Swiss PUS and the steps necessary to prepare it for further use in a population synthesis routine. The PUS is a 5 percent sample of all persons from the full census sample for each census between 1970 and 2000, taken using simple random sampling without replacement. The sample represents the residential population (as opposed to citizens according to civil law). This is also the preferred choice for a population for a transport model, as trips are mostly generated where persons live and not where they are registered.

For each person (P), auxiliary information on household, housing, building and geographical location is available. In addition, for each household, a reference person (RP) has been established. Often, the reference person is the head of the household – the precise rules are defined in

(Swiss Federal Statistical Office, 2000a). For each person, complete person-level data is also provided for the reference person and, if appropriate, also for the partner of the reference person (PRP). All attributes are categorical, a few are ordered:

1. Person

- Demography
- Education
- Profession and social status
- Workplace location and journey
- Status within the household

2. Household

- Household type, total number of persons
- Existence of (grand)parents and non-relatives
- Number/existence of children in specific age groups
- Number of persons with specific attributes (in training, retired, unemployed, non-Swiss)

3. Housing

- Ownership, rent price, number of pieces, area

4. Building

- Type, age, number of housing units, heating

5. Geography

- Canton
- Agglomeration type (town, suburban area, rural area)

Although the PUS contains household attributes as well, it is not a sample of households. The selection probability is the same for each person, thus persons from larger households are overrepresented. In fact, usually not all persons of a multi-person household are sampled, and there are households where none of the constituent persons are part of the sample.

The data set contains four census years in one file, (every ten years between 1970 and 2000). In addition, each sample contains persons from regular as well as collective households. This work considers only persons from the year 2000 and from regular households, as persons from collective households (e.g., prisons and nursing homes) mostly have restricted activity patterns. Of course, students living in dormitories are a notable exception. This removed 14,609 out of 364,401 records (4 %) corresponding to persons from collective households.

In the remainder of this paper, attributes are typeset in *italics*, their categories are “quoted”, and their internal name is typeset in a non-proportional\_font.

### 3 Persons and geography

Table 1 lists all person-level attributes provided by the PUS. Most attributes are available for the person (P), the reference person (RP) and the partner of the reference person (PRP). The *is reference person or partner* and *position in household* attributes are useless for the (P)RP and

Table 1: Person-level attributes

Attribute	Internal name	Initial levels		“INAP” (%)			Missing (%)		
		P	(P)RP	P	RP	PRP	P	RP	PRP
<i>Age</i>	altj	17?	4?	0.14	0.09	24.77	0.03	0.08	0.07
<i>Sex</i>	gesl	2?	=	0.03	0.10	24.83	0.01	0.06	0.09
<i>Nationality</i>	nati	3?	2?	0.13	0.05	24.81	0.03	0.04	0.09
<i>Marital status</i>	zivl	4?	3?	0.03	0.02	24.75	0.00	0.01	0.04
<i>Denomination</i>	konf	4?	=	0.01	0.02	24.72	0.00	0.01	0.02
<i>Main language</i>	spra1	4?	=	0.01	0.07	24.79	0.01	0.06	0.07
<i>Highest level of education</i>	habgh	7?	=	17.52	0.21	24.93	6.58	7.07	5.10
<i>Work status</i>	ezux	9		0.00			0.00		
<i>Work status (RP/PRP)</i>	esch		7?	0.00	0.00	24.69	0.00	0.00	0.00
<i>Workload</i>	pens	4?	3?	50.66	20.44	62.52	2.14	2.54	1.94
<i>Work sector (1)</i>	pber1	4?	=	48.52	20.44	62.52	13.68	18.63	11.87
<i>Employment status</i>	sthb	6?	3?	48.62	20.50	62.63	4.35	5.52	4.39
<i>Work sector (2)</i>	wart1	4?	=	48.52	20.44	62.52	0.00	0.00	0.00
<i>Legal form of company</i>	refo	5?	=	55.41	29.58	68.46	6.90	9.14	5.94
<i>Location of work or school</i>	agde	4?	=	33.49	19.80	61.76	4.25	3.17	3.67
<i>Commute time</i>	wegz	6?	=	33.49	33.04	68.69	9.62	8.91	7.50
<i>Commute frequency</i>	wegh	3?	=	33.49	36.54	71.63	7.29	9.73	9.27
<i>Principal means of transport</i>	vemi	7?	=	33.49	27.13	66.33	6.12	4.99	5.12
<i>Socio-professional category</i>	sopk	10?	=	0.10	0.50	25.21	0.02	0.45	0.42
<i>Is reference person or partner</i>	rphh	4		0.00			0.00		
<i>Position in household</i>	sthh	6		0.00			0.00		
<i>Birthplace</i>	gortk	5		0.00			0.00		
<i>Residence 5 years ago</i>	wo5k	6?		5.49			5.49		
<i>Profession</i>	erlb1	5?		17.48			17.48		
<i>Work sector (ISCO)</i>	isco1	10?		48.58			48.58		
<i>Occupational prestige</i>	trei2	7?		61.68			61.68		

P – person

(P)RP – reference person or her partner

? – “INAP” values present

= – same value as in the column at the left

ISCO – International Standard Classification of Occupations

Table 2: Geographic attributes

Attribute	Internal name	Initial levels
<i>Canton</i>	kant	25?
<i>Large region of Switzerland</i>	grossreg	7
<i>Language areas</i>	spr93	3
<i>Agglomeration type (town, suburban, rural)</i>	agglo	3?
? – “INAP” values present		

hence provided only for P. Furthermore, the *birthplace*, *residence 5 years ago*, *profession*, *work sector (ISCO)* and *occupational prestige* attributes are available only for P.

In Table 2 all geographic attributes are shown. The *canton* categories representing “Appenzell-Innerrhoden” and “Appenzell-Ausserrhoden” are collapsed in the original data. Apart from the *canton*, only the *agglomeration type (town, suburban, rural)* attribute contains missing values.

Almost all attributes contain the code “INAP” (inappropriate) for some records. Furthermore, some attributes have different levels for P and (P)RP. The remainder of this section describes the matching between R and (P)RP attributes, the distinction between “inappropriate” and “missing”, an imputation procedure for the “missing” values, and the treatment of the differences in categorization.

### 3.1 Matching attributes between P and (P)RP

For almost all attributes there were cases with “INAP” values. However, sometimes P is identical to RP or PRP, and a value has been deleted only in P but not in (P)RP, or vice versa. For example there are cases where P is indeed reference person of the household and the *marital status* attribute has been deleted in P but not in RP. In such cases, the missing P attribute can be inferred from the same (P)RP attribute, and vice versa. This important step towards a consistent dataset has been carried out for 56,242 RP and 24,427 PRP values. For 6 values of the *location of work or school* attribute, the P and (P)RP values are different but not “INAP”; the (P)RP values have been substituted by the P value here.

### 3.2 Disambiguation

The “INAP” category is ambiguous in some cases.

- For households where the reference person has no partner, all attributes that refer to the partner of the reference person are set to “INAP”.
- For all persons aged 14 or younger, the *highest level of education* attribute is “INAP”.
- The attributes related to the work place (*workload*, *work sector (1)*, *employment status*, *work sector (2)* and *legal form of company*) are always “INAP” if the person is non-working.
- If all commute-related attributes *location of work or school*, *commute time*, *commute frequency* and *principal means of transport* are “INAP” then it is assumed that the person is not commuting at all.

For all of the above attributes, the distinction has been inferred from other attributes. The six rightmost columns of Table 1 show the share of cases with “INAP” and missing information.

### 3.3 Missing values

All remaining missing values were substituted by values drawn from the empirical one-dimensional distribution of the corresponding variable. For example, if the *nationality* of a person has been deleted, the value has been drawn randomly from the overall nationality distribution observed in the sample. Notable exceptions are *age* and *main language*, and *canton* and *agglomeration type (town, suburban, rural)*: Here, for the imputation, the sample is stratified by *work status*, *language areas* and *large region of Switzerland*, respectively. This avoids ridiculous results such as using the age group “5–9” for a person in “upper management”, or assigning the canton of “Geneva” to a person from “central Switzerland”. In total, 1,120,585 “INAP” values have been substituted.

### 3.4 Aligning classification between P and (P)RP

For some attributes, the categorization detail for (P)RP attributes is coarser than that for P attributes. In order to simplify the further analysis, all attributes were harmonized, while keeping the original values for further reference. Table 3 provides an overview of the information not available for the reference person and her partner (cf. the third and fourth column of Table 1):

Another option would be to treat the finer categorization as an unobserved latent variable for the reference person and her partner and to infer the missing information (Little and Rubin, 2002). The application of more sophisticated imputation procedures for both substitution of missing values and refinement of the categorization is an open task.



Table 3: Differences in categorization between P and (P)RP

Attribute	Collapsed categories in (P)RP
<i>Age</i>	Age groups of 15–20 years for (P)RP vs. 5 years for P
<i>Nationality</i>	“European” and “non-European”
<i>Marital status</i>	“Widowed” and “divorced”
<i>Work status</i>	“Working in the own household” and “other non-earning person”
<i>Workload</i>	No distinction between amount of working hours for part-time workers
<i>Employment status</i>	“Working in the own household”, “apprentices” and “upper management”

## 4 Households

Most household attributes represent censored counts, denoting the number of household members with specific attributes bounded by an attribute-specific maximum value. Table 4 lists all count attributes along with their initial range in the third column. The level “–1” for the *total persons* attribute indicates presence of missing values. In order to simplify further computation, a new count attribute has been added that reflects the number of persons who are *reference person or partner*. The derivation of the *non-earning* attribute will be described in Section 4.2. Besides the count values, the only household information given is the *household type*; this attribute has 3 levels and no missing data.

Note that the count values do not correspond to a disjoint or complete classification. For instance, the *persons aged 65 and over* count also contains all *persons aged 80 and over*, and all *school-age children* and *apprentices* are counted as *people in training (even if employed)*. This is also reflected in the data.

In the following, the alleviation of the censoring that has been applied to the count values is described. Furthermore, a mapping between selected household-level count attributes and person attributes is described, which later will be necessary to estimate the full household structure.

### 4.1 Censoring

At the household level, only the *Total persons* attribute is missing in 94 out of 349,792 cases (0.027 %). No other attributes have missing values. However, all count attributes are right-censored, e.g., the exact number of *children aged 10–14* is not known for households with two or more of them. Some classifications like *parents (in law) and other relatives* and *unemployed*,

Table 4: Censored household-level count attributes

Attribute	Internal	Initial levels	Final levels	Uncertain
<i>Total persons</i>	aper	-1, 1-7	1-12	1.48 %
<i>Reference person or partner</i>	sth1		1-2	
<i>Children (in law)</i>	sth2	0-4	0-10	3.74 %
<i>Parents (in law) and other relatives</i>	sth3_4	0-1	0-6	3.84 %
<i>Non-relatives</i>	sth5	0-1	0-6	3.13 %
<i>Children aged 0-4</i>	a100_04	0-2	0-5	4.82 %
<i>Children aged 5-9</i>	a105_09	0-2	0-5	6.01 %
<i>Children aged 10-14</i>	a110_14	0-2	0-5	6.20 %
<i>Children aged 15-19</i>	a115_19	0-2	0-5	4.99 %
<i>Unmarried children aged 20-24</i>	a120	0-2	0-5	1.81 %
<i>Unmarried children aged 25-29</i>	a125	0-1	0-5	3.38 %
<i>Unmarried children aged 30 and over</i>	a130	0-1	0-5	2.32 %
<i>Persons aged 65 and over</i>	ap65	0-3	=	0.15 %
<i>Persons aged 80 and over</i>	ap80	0-2	=	0.80 %
<i>School-age children</i>	ia14	0-3	0-5	3.64 %
<i>Apprentices</i>	lehr	0-2	=	1.31 %
<i>People in training (even if employed)</i>	iaus	0-4	=	2.03 %
<i>Employed</i>	etto	0-4	=	3.86 %
<i>Unemployed</i>	elos	0-1	0-3	5.67 %
<i>Pensioners (not working)</i>	rent	0-2	=	8.09 %
<i>Foreigners</i>	aus1	0-5	0-11	3.67 %
<i>Non-earning</i>	nerw		0-7	

= – same value as in the column at the left

denote only presence. Hence, the count values will be biased towards smaller counts on average. Fortunately, in some cases we can derive a larger lower bound for some count attributes from the value of others. In addition to the improved consistency in the data set, bias is reduced.

The following *consistency criteria* must be met in the end:

- The *school-age children* value equals the sum of *children aged 5-9* and *children aged 10-14*
- The *children (in law)* value equals the sum of all attributes counting children
- The *total persons* value counts all household members of any kind: *employment status*, *position in household*, *reference person or partner*, *children (in law)*, *parents (in law)* and *other relatives* and *non-relatives*

These criteria are not fulfilled initially. In order to achieve consistency according to the above

Table 5: Number of uncertain count attributes per household size

		<i>Total persons</i>									
		<b>-1</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>Sum</b>	<b>%</b>
<b># uncertain</b>	<b>0</b>	23	56,385	91,831	32,056	30,269	8,644	2,141	794	222,143	63.51
	<b>1</b>	36		6,955	27,226	42,050	18,069	5,074	1,851	101,261	28.95
	<b>2</b>	26			1,014	9,529	8,917	3,437	1,728	24,651	7.05
	<b>3</b>	9				21	247	640	707	1,624	0.46
	<b>4</b>							4	106	110	0.03
	<b>5</b>								3	3	
<b>Sum</b>		94	56,385	98,786	60,296	81,869	35,877	11,296	5,189	349,792	100.00

criteria, all attributes that denote the presence of *parents (in law) and other relatives* or *non-relatives* or the number or presence of children in a specific age group are processed in a round-robin fashion in the following order:

1. *Parents (in law) and other relatives*
2. *Non-relatives*
3. *Unmarried children aged 25–29*
4. *Unmarried children aged 30 and over*
5. *Children aged 0–4*
6. *Children aged 5–9*
7. *Children aged 10–14*
8. *Children aged 15–19*
9. *Unmarried children aged 20–24*

The count is increased by one if the following two *count increase conditions* hold:

- (A) The count value is *uncertain*, i.e., it is already at the rightmost boundary of the range so that it is possible that censoring is in effect here. (The rightmost column of Table 4 shows the share of uncertain cases for each attribute.)
- (B) There is a *violation* of at least one of the conditions involving the current attribute, and increasing the attribute would resolve or at least relax the violation.

This is repeated until no uncertain values exist anymore.

The processing order of the attributes is somewhat arbitrary. It matters only for those cases where at least two of the listed attributes are uncertain. Table 5 shows the distribution of uncertainty vs. household size. In total, only 7.54 % of the records contain two or more uncertain values,

so that this issue can be considered of only minor importance. A better approach would be to add all possible combinations for those records with two or more uncertain values with suitable weights. Note also that only 5,189 out of 349,792 total cases (1.5 %) have 7 or more *total persons*, so the effect of uncertainty also can be neglected for the household size.

Unfortunately, in 3,977 cases (in 2,545 records), at least one of the consistency conditions was violated even after all records in count increase condition have been processed. (Example: A record where the corresponding household has five *total persons*, two of which are *reference person or partner*, with one as count for both *children aged 0–4* and *children aged 5–9*, will never be in count increase condition.) This data error has been fixed by repeating the whole procedure with the uncertainty condition (A) dropped – only the violation condition (B) is in effect for this data-correcting pass.

Finally, the attributes *total persons*, *children (in law)* and *school-age children* are adjusted to reflect the potentially increased individual counts. After this adjustment, the consistency criteria for the treated attributes are met.

A better estimate for the *unemployed* attribute (which originally only shows presence but not count) can be obtained by analyzing the *work status* at the individual level. In total, for 860 cases the *unemployed* count has been increased. Perhaps a similar treatment could be applied for the other attributes, but only the *unemployed* are important for the disjoint classification established in Section 4.2.

From the data it can be inferred that the *unemployed* do not include *children aged 0–14*. Hence, the number of young children plus *employed*, *unemployed* and *pensioners (not working)* must not be less than the household size. This has been corrected in 996 cases by increasing the value of the *total persons* attribute regardless of uncertainty. This converts 81 one-person households to multi-person households, thus rendering the *household type* attribute inconsistent for these households; it has been set to “other” for these households.

Table 6 shows the cross-classification of the *total persons* and *foreigners* attributes. There are few if any cases where in a large household there is a minority of non-foreigners. Hence, it is safe to assume that if “5 or more” household members are foreigners, then all are. In addition, an obvious data error where there are “5 or more” *foreigners* but less than 5 *total persons* is corrected – 3 cases were affected here.

The treatment of censoring creates an artifact – there is one household with 13 members but no household with 12 members. In this one household, the count of *children aged 0–4* is decreased by one in order to achieve a tight set of values for all count attributes. The final attribute ranges, which are extended sometimes substantially compared to the original ranges, are shown in the

Table 6: Cross-classification of the *total persons* and *foreigners* attributes

		<i>Total persons</i>											
		1	2	3	4	5	6	7	8	9	10	11	13
<i>Foreigners</i>	0	47,431	79,682	42,026	58,683	25,414	7,223	2,179	388	132	23	12	1
	1	8,877	8,618	5,471	5,581	2,026	611	200	57	8	9	1	
	2		10,350	1,791	920	370	172	60	19	10			
	3			10,922	1,029	274	78	44	13	6			
	4				15,685	439	82	36	7	1			
	≥ 5		1	1	1	7,452	3,152	1,647	399	145	28	5	

Table 7: Number of uncategorized persons per household size

		<i>Total persons</i>											
		1	2	3	4	5	6	7	8	9	10	11	12
<i># uncategorized</i>	0	50,453	79,548	37,531	42,627	14,099	3,461	1,105	412	187	20	3	
	1	5,855	16,201	19,871	31,806	16,523	4,930	1,542	206	32	15	3	
	2		2,902	2,663	6,440	4,141	2,249	974	166	40	17	8	
	3			146	993	1,059	551	435	73	33	6	2	
	4				33	144	106	78	23	7	2	1	1
	5					9	20	23	1	3			
	6						1	7	2			1	
	7							2					

second-rightmost column of Table 4.

## 4.2 Disjoint classification of person counts

For the anticipated synthesis of the full household structure, it seems helpful to have a set of count attributes so that the *total persons* attribute equals the sum of these count attributes. The following set of attributes almost defines such a disjoint classification: *Children aged 0–14*, *employed*, *unemployed* and *pensioners (not working)*. This is akin to the *work status* attribute at person level, although somewhat coarser; only for the “non-earning” category no count is provided. Table 7 shows a cross-classification of *total persons* vs. the number of uncategorized persons that does not fall into one of the above categories.

It seems reasonable to assume that the uncategorized persons are exactly the *non-earning*: Larger households tend to have at least one uncategorized member, often more. Hence, a new count

Table 8: Work status for one-person households per classification in the household data

	Classification			
	<i>Employed</i>	<i>Unemployed</i>	<i>Pensioners (not working)</i>	<i>Non-earning</i>
<i>Work status</i>				
“One or several part time jobs”	5,387			
“Full-time position”	25,542			
“Unemployed”		1,258		
“Other earning”	1,761			
“Working in the home”			7,183	1,322
“In education (school, college)”			178	852
“Preschool age”				
“Pensioner”			8,062	1,130
“Other non-earning”	335		747	2,551

attribute is created and initialized with the number of formerly uncategorized persons. This classification can be verified against the data by considering one-person households only: Here, all counts are either zero or one. Table 8 shows the correspondence between the *work status* attribute and the classification implied by the count attributes. This will be used to construct a bijection between count classification and *work status* attributes in order to estimate the full household structure. Note that Table 8 uses the original value for the *work status* attribute before harmonization (cf. Section 3.4), so that the distinction between “working in the home” and “other non-earning” is retained.

From Table 8 it follows that in all one-person households counting one *unemployed* person, the reference person also has *work status* “unemployed”. Similarly, *employed* corresponds mostly to a *work status* of “one or several part time jobs”, “full-time position” or “other earning”. The 335 out of 33,025 cases (1 %) with *work status* “other non-earning” are the only inconsistency observed, the treatment of this issue is detailed below. The distinction between *pensioners (not working)* and *non-earning* is also challenging. Both count categories are mapped to the non-earning *work status* categories “working in the home”, “in education (school, college)”, “pensioner” and “other non-earning” – there even exist persons counted as *pensioners (not working)* who are “in education (school, college)” according to their *work status*. A large share of persons is (somewhat contradictory) classified as both *pensioners (not working)* and “working in the home”; however, no further analysis has been carried out so far. Rather, as a first

approximation, the counts in both *pensioners (not working)* and *non-earning* are mapped to the above listed non-earning *work status* categories. Finally, there are no one-person households where the person's *work status* is "preschool age". This is not surprising as small children do not usually run a single-person household (except perhaps the strongest girl in the world, as described by Astrid Lindgren).

Unfortunately, no clear classification for the "other non-earning" could be found in the data. For a person with *work status* "other non-earning" it seems to be impossible to predict whether she belongs to the *employed*, *pensioners (not working)* or *non-earning*. It appears that 335 out of 3,633 cases (9.2 %) have a *work status* of "other non-earning" at the person level but are classified as *employed*. As the "other non-earning" persons do not have information on the commute anyway, they can be safely treated as "working in the home" for the purpose of a transportation model. (In fact, both categories are collapsed anyway, cf. Section 3.4). It remains to adjust the counts of the *employed* in those cases where all household members are present in the record (as P, RP or PRP) and the *employed* count is larger than the number of employed persons according to the individual data. In 10,695 cases, the *employed* count was decreased and the *non-earning* count increased. (Here, also the *total persons* had to be increased in one case.)

Ultimately, the following mapping is proposed.

- The "unemployed" *work status* is mapped to the classification with the same name, and vice versa.
- The *work status* categories "one or several part time jobs", "full-time position" and "other earning" are collapsed to a new category "employed". This new category is mapped to the classification with the same name, and vice versa.
- The classifications *pensioners (not working)* and *non-earning* and the *work statuses* "working in the home", "in education (school, college)", "pensioner" and "other non-earning" are collapsed and mapped to each other.

## 5 Conclusion and outlook

In this paper the preparation of the Swiss Public-Use Sample (PUS) has been described. The dataset is a 5 percent person sample of the population in 2000, providing detailed person-level information for the sampled individuals. Moreover, information on household, housing, building and location is attached to each person record.

The preparation considered mainly person and household attributes. At person level, attribute levels were disambiguated, harmonized, and missing values were imputed. For household attributes, the values of several right-censored count attributes have been increased where appropriate in order to achieve consistency between the different count values. Furthermore, a bijection between counts of persons in a household and person-level attributes has been established. In a few cases, geographic information has been imputed. Housing and building attributes have not been treated.

The resulting dataset contains no missing information in the attributes of interest; however, the imputation still can be improved. Categorization detail has been lost in the harmonization of attributes. The treatment of censored count attributes seems reasonable but has to be validated. Furthermore, the mapping between count and person-level attributes is rather imprecise. Collective households (including student dormitories) have been ignored altogether.

Synthetic populations for Switzerland, including household membership, will be generated based on the prepared data. A synthetic population for the year 2000 will then be compared against the full census dataset. The validation will finally show the impact of the various assumptions made during the preparation; that given, a validation is not considered necessary at this stage.

## 6 Acknowledgements

The authors gratefully acknowledge the financial support of the Swiss National Science Foundation (project no. 138270). The authors also would like to thank Christof Zöllig for his valuable remarks on the presentation.

## 7 References

- Bar-Gera, H., K. Konduri, B. Sana, X. Ye and R. M. Pendyala (2009) Estimating survey weights with multiple constraints using entropy optimization methods, paper presented at the *88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.
- Bürgle, M. (2007) Synthese von Haushaltsdaten für den Kanton Zürich, <http://e-collection.library.ethz.ch/view/eth:29494>, accessed on 28/03/2012.
- Lee, D.-H. and Y. Fu (2011) Cross-entropy optimization model for population synthesis in activity-based microsimulation models, *Transportation Research Record*, **2255**, 20–27.



- Little, R. J. A. and D. B. Rubin (2002) *Statistical analysis with missing data*, John Wiley & Sons, Hoboken.
- Müller, K. and K. W. Axhausen (2011a) Hierarchical IPF: Generating a synthetic population for Switzerland, paper presented at the *51st Congress of the European Regional Science Association*, Barcelona, September 2011.
- Müller, K. and K. W. Axhausen (2011b) Population synthesis for microsimulation: State of the art, paper presented at the *90th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2011.
- Pritchard, D. R. and E. J. Miller (2012) Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously, *Transportation*, **39** (3) 685–704, May 2012.
- Swiss Federal Statistical Office (2000a) Public use samples (PUS): Beschreibung der Variablen, [http://www.portal-stat.admin.ch/pus/files/var\\_d.html](http://www.portal-stat.admin.ch/pus/files/var_d.html).
- Swiss Federal Statistical Office (2000b) Public use samples (PUS): Excerpts for general use from the Swiss federal population censuses 1970-2000, [http://www.portal-stat.admin.ch/pus/files/index\\_e.html](http://www.portal-stat.admin.ch/pus/files/index_e.html).
- Ye, X., K. Konduri, R. M. Pendyala, B. Sana and P. A. Waddell (2009) A methodology to match distributions of both household and person attributes in the generation of synthetic populations, paper presented at the *88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.

## A Housing and building

Table 9 shows the housing and building attributes for each record. All attributes have cases with missing values. Especially the *monthly rent* attribute will be important later as a proxy for the household income.

Table 9: Housing and building attributes

Attribute	Internal name	Initial levels
<i>Residence type (renting, owning)</i>	wbtyp	6?
<i>Monthly rent</i>	wmiet	27?
<i>Number of persons in apartment</i>	wapto	8?
<i>Number of living rooms per person in apartment</i>	wapra	5?
<i>Area per person in apartment</i>	wapfl	6?
<i>Type of building</i>	ggart	4?
<i>Construction period</i>	gbaup	7?
<i>Number of housing units in the building</i>	gazwt	9?
<i>Heating type</i>	gheiz	4?
? – “INAP” values present		

## B Technical information

The entire analysis has been carried out following the *literate programming* paradigm. Both R R Development Core Team (2012) code and L<sup>A</sup>T<sub>E</sub>X markup is contained in the code side by side. The sources are processed with knitr (Xie, 2012), a successor of Sweave (Leisch, 2002), to simultaneously produce this document and the results. The source files are available upon request.

- R version 2.15.0 (2012-03-30), x86\_64-pc-linux-gnu
- Base packages: base, datasets, graphics, grDevices, methods, splines, stats, utils
- Other packages: foreign 0.8-50, Hmisc 3.9-3, knitr 0.6.15, plyr 1.7.1, reshape 0.8.4, scales 0.2.1, survival 2.36-14, xtable 1.7-0
- Loaded via a namespace (and not attached): cluster 1.14.2, colorspace 1.1-1, dichromat 1.2-4, digest 0.5.2, evaluate 0.4.2, formatR 0.6, grid 2.15.0, labeling 0.1, lattice 0.20-6, munsell 0.3, RColorBrewer 1.0-5, stringr 0.6, tools 2.15.0

## C References to R packages

- R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/>. ISBN 3-900051-07-0.
- Xie, Y. (2012) *knitr: A general-purpose package for dynamic report generation in R*, <http://yihui.name/knitr/>. R package version 0.6.15.
- Leisch, F. (2002) Sweave: Dynamic generation of statistical reports using literate data analysis, paper presented at the *Compstat 2002 — Proceedings in Computational Statistics*, 575–580.
- R Core Team (2012) *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ...*, <http://CRAN.R-project.org/package=foreign>. R package version 0.8-50.
- Jr, F. E. H. and with contributions from many other users. (2012) *Hmisc: Harrell Miscellaneous*, <http://CRAN.R-project.org/package=Hmisc>. R package version 3.9-3.
- Wickham, H. (2012a) *plyr: Tools for splitting, applying and combining data*, <http://CRAN.R-project.org/package=plyr>. R package version 1.7.1.
- Wickham, H. (2011) *reshape: Flexibly reshape data.*, <http://CRAN.R-project.org/package=reshape>. R package version 0.8.4.
- Wickham, H. (2012b) *scales: Scale functions for graphics.*, <http://CRAN.R-project.org/package=scales>. R package version 0.2.1.
- Therneau, T. (2012) *survival: Survival analysis, including penalised likelihood.*, <http://CRAN.R-project.org/package=survival>. R package version 2.36-14.
- Dahl, D. B. (2012) *xtable: Export tables to LaTeX or HTML*, <http://CRAN.R-project.org/package=xtable>. R package version 1.7-0.