Methodology Report

Coverage Estimation for the Swiss Population Census 2000

0

Statistische Grundlagen und Übersichten
Bases statistiques et produits généraux
Basi statistiche e presentazioni generali
Basic statistical data and overviews

Estimation Methodology and Results

Office fédéral de la statistique
Bundesamt für Statistik
Ufficio federale di statistica
Uffizi federal da statistica
Swiss Federal Statistical Office

OFS BFS UST

Neuchâtel, 2004

Statistik der Schweiz

Die vom Bundesamt für Statistik (BFS) herausgegebene Reihe «Statistik der Schweiz» gliedert sich in folgende Fachbereiche:

0 Statistische Grundlagen und Übersichten
1 Bevölkerung
2 Raum und Umwelt
3 Arbeit und Erwerb
4 Volkswirtschaft
5 Preise
6 Industrie und Dienstleistungen
7 Land- und Forstwirtschaft
8 Energie
9 Bau- und Wohnungswesen
10 Tourismus
11 Verkehr und Nachrichtenwesen
12 Geld, Banken, Versicherungen
13 Soziale Sicherheit
14 Gesundheit
15 Bildung und Wissenschaft
16 Kultur, Medien, Zeitverwendung
17 Politik
18 Öffentliche Verwaltung und Finanzen
19 Rechtspflege
20 Einkommen und Lebensqualität der Bevölkerung
21 Nachhaltige Entwicklung und regionale Disparitäten

Swiss Statistics

The «Swiss Statistics» series published by the Swiss Federal Statistical Office (SFSO) covers the following fields:

0 Basic statistical data and overviews
1 Population
2 Territory and environment
3 Employment and income from employment
4 National economy
5 Prices
6 Industry and services
7 Agriculture and forestry
8 Energy
9 Construction and housing
10 Tourism
11 Transport and communications
12 Money, banks, insurance companies
13 Social security
14 Health
15 Education and science
16 Culture, media, time use
17 Politics
18 Public administration and finance
19 Law and justice
20 Income and standard of living of the population
21 Sustainable development and regional disparities

BFS / SFSO Methodology Report – Coverage Estimation for the Swiss Population Census 2000

Swiss Statistics

Methodology Report

Coverage Estimation for the Swiss Population Census 2000

Estimation Methodology and Results

Author **Anne Renaud**
Swiss Federal Statistical Office

Publisher **Swiss Federal Statistical Office**

Office fédéral de la statistique
Bundesamt für Statistik
Ufficio federale di statistica
Uffizi federal da statistica
Swiss Federal Statistical Office

OFS BFS UST

Neuchâtel, 2004

Preamble et Thanks

The «Coverage Estimation for the Swiss Census 2000» project seeks to assess one of the aspects of census quality. This report presents the general estimation methodology and the final results for overcoverage, undercoverage and resulting net coverage. The work was carried out by Anne Renaud from the Statistical Methods Unit (METH) of the Swiss Federal Statistical Office (SFSO).

A special word of thanks to Rajendra Singh and his team at the Decennial Statistical Studies Division of the U.S. Census Bureau for the general assistance, discussions about the methodology and review of this report. The contacts per e-mail and meetings at the Joint Statistical Meeting in Atlanta in August 2001, in Neuchâtel in March 2003 and in Washington in March 2004 were of great help. Special thanks also to Philippe Eichenberger (METH) for the helpful discussions throughout the project. Thanks also to census staff members who carried out matchings, performed various clerical checks, and furnished considerable information about the census data. Thanks also to Randall Jones from the Languages Services (LING) for reviewing the English in this report.

Summary

Coverage of the Swiss population census is estimated for the first time for the census 2000. Both undercoverage and overcoverage are analyzed apart and then combined by using the dual system methodology. The estimates are based on two samples: the Enumeration sample (E-sample) and the Population sample (P-sample) in order to capture both the overcoverage and the undercoverage components.

Similar to results in other countries, we determined that 1.6% of the resident population were overlooked in the census (undercount) and that 0.4% were counted erroneously (overcount). The resulting overall rate of net undercoverage is 1.4% with larger values for some subgroups of the population such as 20-31 years-old people (2.8%) or foreigners (2.9-3.5%).

Other types of errors were analyzed such as error in the type of domicile, time delay between census day and effective data collection day for movers around the census day, or potential misclassification variables. The results and experience gained during the project can be used to improve the subsequent censuses.

Key Words

methodology report; population census; VZ2000; RFP2000; coverage; undercoverage; overcoverage; sampling; estimation estimations; couverture; plan d'échantillonnage; plusieurs niveaux; stratification; allocation; post-enumeration; coverage.

Published by:	Swiss Federal Statistical Office (SFSO)
Information:	Anne Renaud, Phone +41/ (0)32 713 62 65 Anne.Renaud@bfs.admin.ch
Realisation:	Statistical Methods Unit, SFSO
Obtainable from:	Swiss Federal Statistical Office CH-2010 Neuchâtel 'Phone +41/ (0)32 713 60 60 / Fax +41/ (0)32 713 60 61 Order@bfs.admin.ch
Internet:	http://www.statistik.admin.ch
Order number:	338-0027
Price:	free
Series:	Swiss Statistics
Field:	0 Basic statistical data and overviews
Original text:	English
Graphics/Layout:	SFSO
Copyright:	SFSO, Neuchâtel 2004 Reproduction with mention of source authorized (except for commercial purposes)
ISBN:	3-303-00305-X

Contents

Introduction

In every population census some people are overlooked and others are counted more than once. There is therefore undercoverage and overcoverage of the population. The combination of both components typically leads to net undercoverage of the population with values around 1-3%. However, net overcoverage has also been observed (*e.g.* in the US census 2000); see Table 1.

The net coverage may vary considerably between subgroups of the population; see Table 2. Some subgroups, such as 20-30 year-old males living in large cities, typically have larger undercount than 40-50 year-olds in rural regions. This is mostly due to higher mobility. Special omissions such as newborns or elderly people in retirement homes are also observed. It is therefore possible to have a net overcount for some subgroups but a net undercount for the population as a whole.

Coverage of the Swiss population census is estimated for the first time for the census 2000. Both undercoverage and overcoverage are analyzed apart and then combined by using the dual system methodology.

The dual system estimator (DSE) is based on the capture-recapture methodology. It combines the census counts with some estimators based on two samples: the *Enumeration* sample (*E-sample*) and the *Population* sample (*P-sample*). The E-sample is selected in the census data set and forms the basis for the estimation of overcoverage in the census. The P-sample is a subset of a post-census coverage survey, which is as independent as possible from the census, and forms the basis for the estimation of undercoverage in the census.

Estimates are expected for large demographic groups, for small and large municipalities and for the census methodologies CLASSIC and TRANSIT. See Appendix A for general information about the census 2000.

The purpose of the project is not to adjust the census counts but rather to gather information about the coverage quality in the census results. In other words, the aim is to gather information that can be used to plan and improve the quality of subsequent censuses.

The general methodology of the project was developed with the grateful assistance from Dr. Rajendra Singh and his team of the Decennial Statistical Studies Division of the U.S. Census Bureau.

Table 1: Census coverage estimates, with the standard error in parentheses [%]. Net undercoverage, overcoverage and undercoverage. References: Thibault (2003), StatCan (1996), Hogan (1993), Fenstermaker (2002), Brown et al. (1999) and ABS (1997, 1999, 2004). Results for the Census 2001 in Canada: state on April 2003.

	Census	net	over	under
Canada	1991	2.9	0.6	3.4 (0.12)
	1996	2.5 (0.10)	0.7 (0.04)	3.2 (0.09)
	2001	3.2 (0.14)	0.9 (0.04)	4.1 (0.13)
USA	1990	1.6 (0.20)	3.1	4.7
	2000	-0.5 (0.20)	not available	not available
Australia	1991	1.8 (0.10)	not available	not available
	1996	1.6 (0.10)	0.2	1.8
	2001	1.8 (0.10)	0.9	2.7
New Zealand	1996	1.2 (0.10)	1.4	0.2
UK	1991	2.2	not available	not available

Table 2: Census coverage estimates. Net coverage in subgroups [%] with standard error. Same references.

UK (1991)	>20% for young males in inner cities
Australia (1996)	1.1 - 3.1% depending on the State/Territory
	2.0% males and 1.1% females, 4.3% for males aged 20-24
USA (1990)	0.7% (0.22) White, 5.0% (0.82) Hispanic, 4.6% (0.55) Black
USA (2000)	-1.1% (0.20) White, 0.7% (0.44) Hispanic, 1.8% (0.43) Black

This report is a continuation of three methodology reports[1]: Renaud and Eichenberger (2002) (see also Renaud (2002) in English), Renaud (2003), and Renaud and Potterat (2004). It seeks to summarize the methodology and results from the coverage estimation for the Swiss population census 2000. Part I "Methodology" presents the general methodology for carrying out census coverage estimations as well as the methodology developed for the Swiss estimates. Part II "Data and Preliminaries" presents the data available for the estimations and the matching processes (searches). Part III "Results" presents the results of the analysis and a conclusion with remarks for future censuses and future coverage estimations. Some complementary information is gathered in the appendix.

[1]The methodology reports may be downloaded from the SFSO web site (pdf files in French).

Part I

METHODOLOGY

Chapter 1

Measuring the Coverage of a Population Census

Coverage error has been studied for several decades in the USA (since 1950) and Canada (since 1961) (Wolter, 1986). Coverage estimation has also become a standard practice since the 80s or 90s in many other countries such as UK, New Zealand, Australia, Germany, Italy, Estonia and Norway. Methods and results mainly from the USA, UK and Australia are described below. These countries have the advantage of having extensive experience with coverage estimation and available detailed documentation.

1.1 Reasons Why Estimating Coverage is Important

Census undercoverage has been estimated to be non negligible in many countries, especially in some subgroups of the population. Discrepancies between real population and census counts lead to an imprecise image of the population for planning and decisions.

The undercount was often observed to be larger for the census 1990 than for the census 1980. As a result, extensive research has been done for the censuses around 2000. On the one hand, special measures for improving the census process have been developed to improve coverage (advertising for targeted subgroups, etc). On the other hand, estimation methodology has been further studied.

The justification for more research is also that measuring coverage has become an important statistical issue. As a case in point, the U.S. Census Bureau was sued in federal court many times in 1980 and 1990 on the issue of completeness of the census.

The knowledge gained from these research efforts help the census agencies to improve the quality of future censuses.

1.2 Basics

Let a population U with size N, that is assumed fixed but unknown.

A census is conducted in order to enumerate each and every person in U at a particular point in time. For a variety of reasons, some individuals of U are overlooked, others are counted twice or erroneously (*e.g.* not born or abroad on census day). The enumerated list U_c with size C is therefore different from U; see Figure 1.1.

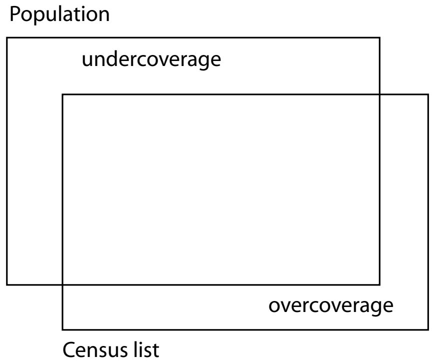

Figure 1.1: Illustration of overcoverage and undercoverage of the census as a comparison between the real population to enumerate and the census list.

The number of overlooked people is the *undercount* C_{under}. The number of people counted erroneously is the *overcount* C_{over}.

The *net error of coverage* is defined by $C_{net} = N - C = C_{under} - C_{over}$. It is positive if the population count is underestimated and negative if the count is overestimated.

We also define the *rate of net coverage* $R_{net} = C/N$ and the *rate of net undercoverage* $R_{net,under} = C_{net}/N = (N - C)/N$.

In practice, C is generally smaller than N (net undercoverage) but the opposite may also be observed if serious problems of double entries occur in the lists.

To produce a measure of census coverage, we need some auxiliary information alongside the census data set. The auxiliary information is usually demographic data or data from a sample survey; see Sections 1.3 and 1.4.

1.3 Demographic Data

Comparison of demographic data and census data has been used for numerous census coverage estimations. If high-quality demographic data are available, such comparisons are not expensive and can lead to interesting results for aggregated data such as global population, male-female, or large regions.

In practice, comparisons are limited by the information that can be derived from demographic data, the quality of data and the level of aggregation. Only net coverage can be estimated, without any information about the number of individuals overlooked or counted twice. Furthermore, results are not reliable if the definitions in the demographic data and in the census data do not match up.

As a case in point, demographic analysis for the census 1990 in the USA show results about male-female, black-non-blacks and six age groups (Robinson et al., 1993). The variance of the results was also estimated by assuming that the census data were error-free. The methodology was shown to give more reliable results for differences between subgroups then for net undercounts in subgroups.

It is interesting to note that demographic analysis may be a complement to other estimation methodologies (*e.g.* check for coherence). For example, the comparisons between census and demographic data for the US census 2000 detected an important error in the results of the Accuracy and Coverage Estimation (A.C.E. based on sample surveys, see below). The A.C.E. estimated an undercount of 1.18% (Davis, 2001) and demographic analysis estimated an overcount of 0.12%; see Robinson et al. (2002) for preliminary results and Robinson (2001) for revised results. The subsequent revision of the A.C.E. estimates produced a final value of 0.49% overcount (Fenstermaker, 2002; Mule, 2003a).

1.4 Sample Survey Data

Methods based on sample surveys have been developed in order to remedy the limitations of demographic analysis in coverage estimations.

The most common methodology is called dual system estimation and is based on the capture-recapture system developed by biologists for estimating the size of wildlife populations. For this methodology, a sample survey is organized as independently as possible from the census and then matched with the census data set.

The survey is called *e.g.* post-enumeration survey, coverage survey or census validity survey. It is used to estimate coverage of the census and sometimes also measurement error in the census. Note that the use of data from coverage surveys - to be organized - needs more resources than use of existing demographic analysis.

1.4.1 Capture-Recapture

Various capture-recapture models may be used to estimate population coverage and all of them are based on log-linear models (Wolter, 86).

The capture-recapture methodology makes use of two independent lists. In our case: the census (capture) and the survey (recapture).

Let us assume that A is the list of individuals enumerated in the census and B is the list of individuals enumerated in the post-enumeration survey. For now, we assume that B is a complete enumeration of the population. The sample-related characteristics will be included in the particular case of the dual system estimation, see Section 1.4.2.

The general coverage error model is based on the following assumptions (see details in Wolter, 1986):

1. Closure assumption: the population U is closed and of fixed size N.

2. Multinomial assumption : the joint event that the individual i is in the list A or not and in list B or not is modelled by the multinomial distribution ξ_i with parameters $p_{i11}, p_{i12}, p_{i21}$ et p_{i22}:

	in B	out B	
in A	p_{i11}	p_{i12}	p_{i1+}
out A	p_{i21}	p_{i22}	p_{i2+}
	p_{i+1}	p_{i+2}	1

(1.1)

where p_{i11} is the probability for i to be captured in both lists, p_{i12} is the probability for i to be captured in A but not B and $p_{i1+} = p_{i11} + p_{i12}$ is the probability to be captured in the list A.

3. Autonomous independence: lists A and B are created as a result of N independent trials in U, using distributions $\xi_1, \xi_2, ..., \xi_N$. The number of individuals in each cell (in-out A crossed by in-out B) is:

	in B	out B	
in A	N_{11}	N_{12}	N_{1+}
out A	N_{21}	N_{22}	N_{2+}
	N_{+1}	N_{+2}	$N_{++} = N$

(1.2)

where N_{11} is the number of individuals in both lists and N_{1+} is the number of individuals in the census (list A). Note that N_{1+} is a random variable with mean $\mu_{1+} = \sum p_{i1+}$ and variance $\sigma_{1+} = \sum p_{i1+}p_{i2+}$ under the model.

4. Matching and quality assumptions: it is possible to match list B to list A exactly and without error. The possible nonrespondents from list B are so documented that we can match them to the census. The lists do not include double, out-of-scope or fictitious entries.

The general model, with $3N$ unknown parameters (3 for each individual, see assumption 2 above), is underidentified. Further assumptions are needed to estimate the true population size N. Various special cases of the general model may be considered, as well as combinations between them.

1.4.2 Dual System Model

The Peterson or dual system model is the most employed capture-recapture model in the framework of population censuses. Two assumptions complete the general model:

1. Causal independence: the event of being included in list A is independent of the event of being included in list B: $p_{i11}/p_{i12} = p_{i21}/p_{i22}$ for $i = 1, .., N$ or $p_{ijk} = p_{ij+}p_{i+k}$, for $i = 1, .., N, j = 1, 2, k = 1, 2$.
2. Fixed enumeration probability in A, resp. in B: the capture probabilities satisfies $p_{i1+} = p_{1+}$ and $p_{i+1} = p_{+1}$ for $i = 1, .., N$.

The likelihood associated with the dual system model is:

$$L(N, p_{1+}, p_{+1}) = \binom{N}{N_{11} N_{12} N_{21}} p_{1+}^{N_{1+}} \cdot (1 - p_{1+})^{N - N_{1+}} \cdot p_{+1}^{N_{+1}} \cdot (1 - p_{+1})^{N - N_{+1}} \tag{1.3}$$

The sufficient statistic is now (N_1, N_{1+}, N_{+1}), where $N_1 = N_{11} + N_{12} + N_{21}$ denote the total number of distinct captures, and the maximum likelihood estimators are:

$$\widehat{N} = N_{1+} \frac{N_{+1}}{N_{11}}, \widehat{p}_{1+} = \frac{N_{11}}{N_{+1}} \text{ and } \widehat{p}_{+1} = \frac{N_{11}}{N_{1+}} \tag{1.4}$$

In practice, we assume that all N members are exposed to capture in the list A but only a sample of the N members are exposed to possible inclusion in list B. Of the population quantities of (1.2), only the census total N_{1+} is known. The survey total N_{+1} as well as the totals N_{11}, N_{12} and N_{21} may be estimated based on the sample survey data and the result of the match between both lists. The quantities N_{22} and N are then estimated on the basis of the dual system model.

The dual system estimator of N is therefore:

$$\widehat{N} = N_{1+} \frac{\widehat{N}_{+1}}{\widehat{N}_{11}} \tag{1.5}$$

where N_{1+} is the census total, and $\widehat{N}_{+1}$ and $\widehat{N}_{11}$ are estimators based on survey enumeration B and the matching of B to list A.

1.4.3 Non Independent Sample Survey

An alternative to the dual system was tested in the USA during the 1995 and 1996 Census Tests. The methodology was called CensusPlus and was designed to resolve some of the problems due to unsatisfied assumptions such as statistical independence (causal independence) between capture in the census and recapture in the survey and optimal matching (Bell, 1994).

The CensusPlus methodology consists of two phases. During the first phase, people are enumerated in a sample of units selected independently from the census data set (coverage survey). During the second phase, data from the survey and the census are compared and possibly completed in a final list. This is done by resolving differences between two lists in the field to get one high-quality list for the sampled units. The assumption of complete coverage therefore replaces the independence assumption in the estimation. This methodology was not further studied as the quality of the final lists was not sufficient during the 1995 and 1996 tests to be considered complete.

1.5 Combination of Data Sources

Various data sets may be combined to get an estimate of census coverage. For example, the dual system may be extended to a triple-system which is potentially more precise than the dual system but more complex (*e.g.* census, survey and demographic data). However, the two additional lists required, which should be complete and of high quality, are rarely available for census purposes.

Combination of dual system and demographic analysis has been tested in various countries; see *e.g.* Bell (1993). The idea is to apply the dual system to get a coverage estimate for females and then to use the male-female ratio from demographic data to get a coverage estimate for males. This approach is based on the assumption that the bias of the dual system will be larger for males than females.

1.6 Dual System in Practice

The dual system methodology is the basis for coverage estimation in many countries; see the bibliography of Fienberg (1992).

A post-enumeration survey and a matching with the census are organized in order to get the survey data and the information about data in both the survey and the census.

In practice, the census data set is not perfect. It may include some erroneous entries such as double entries or other people that should not be counted in the census (*e.g.* born after the census day, abroad on census day, dead before census day). In that case, we have an overcount in the census list.

All countries include an estimation of undercoverage because this component is known to be the most effective component of coverage.

Some countries, such as the UK, have chosen not to estimate the overcoverage component of the census list because it is expected to be negligible. Other countries, such as the USA, include this component in the coverage estimator. As a consequence N_{1+} is no longer a fixed value but is estimated by $\widehat{N}_{1+}$.

Some other countries combine the results from different sources. In Canada, the undercoverage and overcoverage components are derived in 1996 and 2001 from four studies (Morel and Kleim, 2003; Clark and Tourigny, 1999)[1]: the vacancy check (undercoverage), the reverse record check (undercoverage and overcoverage), the automated match study (overcoverage) and the collective dwelling study (overcoverage). The reverse record check uses the capture-recapture methodology.

1.6.1 U.S. Census Bureau Estimator

The U.S. Census Bureau dual system estimator *DSE* $= \widehat{N}$ is based on the Equation (1.5) (Hogan, 1992, 1993, 2003):

$$\mathit{DSE} = \widehat{N} = [\widehat{N}_{1+}] \left[\frac{\widehat{N}_{+1}}{\widehat{N}_{11}}\right] = [\widehat{N}_{1+}] \left[\frac{\widehat{N}_p}{\widehat{M}}\right] = \left[(C - II)\frac{\widehat{CE}}{\widehat{N}_e}\right] \left[\frac{\widehat{N}_p}{\widehat{M}}\right] \tag{1.6}$$

The number of people correctly counted in the census N_{1+} is estimated by $\widehat{N}_{1+}$, which is in turn based on the census count C with a correction for the number of whole-person imputations II

[1] The "Census 2001 Technical report on coverage" is scheduled for release in December 2004.

and the rate of correct enumeration $\widehat{CE}/\widehat{N}_e$. The estimated number of correct enumerations in the census data set $\widehat{CE}$ and the estimated census count $\widehat{N}_e$ are weighted totals based on a sample selected in the census data set. The sample is called *E-sample* (*Enumeration*-sample). $\widehat{CE}$ is based on the result from a search for erroneous and correct enumerations in the E-sample.

The second part of the formula is the opposite of the rate of correct matches $\widehat{N}_{+1}/\widehat{N}_{11} = [\widehat{M}/\widehat{N}_p]^{-1}$. The estimated number of matches $\widehat{N}_{11} = \widehat{M}$ and the estimated census count $\widehat{N}_{+1} = \widehat{N}_p$ are weighted totals based on a sample independent from the census (recapture). The sample used for recapture is called *P-sample* (*Population*-sample). $\widehat{M}$ is estimated on the basis of the result from a search for matches in the census data set.

We note that the dual system estimator $\widehat{N}$ is the product of a fixed amount $(C - II)$ by two random ratios $\widehat{CE}/\widehat{N}_e$ and $\widehat{N}_p/\widehat{M}$.

The U.S. Census Bureau uses a quite complex method to estimate the various totals; see for instance the treatment of movers in Section 1.6.7, the decomposition of the initial A.C.E. estimator in Mule (2001) and the Target Extended Search plans of the A.C.E. 2000 (Navarro, 2000). The general methodology, with the revisions and an evaluation, is described in National Research Council (2004).

1.6.2 Note about Alternative Estimators

Alternative estimators to Equation (1.6) could be:

$$\widehat{N}^{(1)} = C - \widehat{\text{EE}} + \widehat{\text{UN}} = \widehat{CE} + (\widehat{N}_p - \widehat{M}) \tag{1.7}$$

$$\widehat{N}^{(2)} = C \left[\frac{\widehat{CE}}{C}\right] \left[\frac{\widehat{N}_p}{\widehat{M}}\right] = \widehat{CE} \left[\frac{\widehat{N}_p}{\widehat{M}}\right] = (C - \widehat{\text{EE}}) \left[\frac{\widehat{N}_p}{\widehat{M}}\right] \tag{1.8}$$

where $\widehat{\text{EE}} = C - \widehat{CE}$ is the estimated number of erroneous enumerations and $\widehat{\text{UN}} = \widehat{N}_p - \widehat{M}$ is the estimated number of overlooked enumerations.

The estimator $\widehat{N}^{(1)}$ is probably the most intuitive one. However, it is not applied in practice because the cell of people missed by both lists is omitted. Furthermore, the variance of the estimator is very high (sum of estimated totals).

The estimator $\widehat{N}^{(2)}$ is directly deduced from Equation (1.6), but the constant C is used instead of the estimator $\widehat{N}_e$ as the total in the overcoverage component of the estimator. This estimator was tested by the U.S. Census Bureau but not kept for application because of the larger variance. Note also that $\widehat{CE}/C$ may have a larger bias than $\widehat{CE}/\widehat{N}_e$.

1.6.3 Application of the Dual System

Most of the assumptions of the dual system model may be considered as satisfied if the post-enumeration survey and the data processing are achieved with care. However, the assumption of fixed enumeration probability in the census, and in the survey respectively, is clearly not satisfied. It is known, for instance, that young people have a lower enumeration probability than older people. Similarly people in urban regions usually have a lower probability of being counted than people in rural regions.

Two types of methodologies were developed to deal with this point: (1) construction of homogeneous groups (also called estimation cells or post-strata[2]) in which *DSE* is calculated and then recombined to get estimates for various subgroups of the population or (2) modelling of the capture probability, for instance with a logistic model.

Traditionally, the U.S. Census Bureau uses estimation cells (Hogan, 2003) and the UK uses models for their coverage results (Brown et al., 99). The U.S. Census Bureau is, however, considering the idea of using more models for 2010. Estimation cells are easy to deal with but offer fewer degrees of freedom than models. Note, however, that the choice of the set of cells is also a modelling task.

In the U.S. Census Bureau, estimation cells were defined equally for the P-sample and the E-sample in 1990 and for the first results of 2000. However, during evaluation, the behavior of the probability of correct enumeration was found different from the behavior of the probability of match. The revised version includes two sets of estimation cells in the estimations; one for the P-sample and one for the E-sample (Kostanich, 2003).

The choice of the set of estimation cells is a key point in the dual system estimation. For one thing, people with similar census capture probabilities should be grouped together without ending up with cells that are two small. Furthermore, cells must be definable in both the P-sample and the census data set and based on variables with a low misclassification error.

1.6.4 Direct and Synthetic Estimation

When using estimation cells, direct coverage estimates are available for both estimation cells and aggregates of estimation cells. For instance, if the cells are defined as a combination of both genders in three age groups, results are available for the whole population, for male, for female, for the age groups 1 to 3, as well as for males in age group 1, etc. However, an estimation for females in large cities requires another methodology.

Synthetic estimation is used to produce estimates for any subgroup of the population. The assumption is that a proportion measured at an aggregate level applies to all sub-groupings.

Let $\Lambda = \{1, .., \ell, .., L\}$ the set of L estimation cells. In each estimation cell, we define the census count C_ℓ and the DSE estimator $\widehat{N}_\ell$.

Let d be the domain for which we want to estimate the total N_d. The synthetic estimator $\widehat{N}_d^s$ is defined by:

$$\widehat{N}_d^s = \sum_{\ell \in \Lambda} C_{d\ell} \frac{\widehat{N}_\ell}{C_\ell} = \sum_{\ell \in \Lambda} C_{d\ell} \, \widehat{CCF}_\ell \tag{1.9}$$

where $C_{d\ell}$ is the census count in the intersection between domain d and estimation cell ℓ ($d \cap \ell$) and $\widehat{CCF}_\ell = \widehat{N}_\ell / C_\ell$ is the estimated coverage correction factor in the estimation cell ℓ.

If domain d is a set of estimation cells $J \subset \Lambda$, we have $C_{d\ell} = C_\ell$ for $\ell \in J$ et $C_{d\ell} = 0$ for $\ell \notin J$.

[2]The term "post-strata" is traditionally used in the U.S. Census Bureau to refer to the estimation cells of the DSE. It is not related to a post-stratification in the general meaning of the sampling techniques.

The synthetic estimator is reduced to the direct estimator $\widehat{N}_d^s = \sum_{\ell \in J} \widehat{N}_\ell$.

We note that synthetic estimation for small areas may have a low accuracy and possibly a large bias. Results are expected to be more accurate for larger subgroups of the population.

1.6.5 Variance Estimation

The U.S. Census Bureau traditionally uses the jackknife techniques to estimate the variance of the DSE estimator. For 2000, the jackknife was quite complex so that the stratified two-phase sampling could be taken into account (Kim et al., 2000 and Sand and Navarro, 2001). However, a comparison between the production variances and a simple jackknife showed that results where very similar in most of the cases (Schindler, 2002). Jackknife is available for direct as well as synthetic estimations. Some adjustments may be included such as grouping of estimation cells or of sampling strata in order to stabilize the variance estimator; see *e.g.* Sand and Navarro (2001).

The Office for National Statistics (ONS) in the UK tested the jackknife and the ultimate cluster variance estimator[3]. The final choice was to use the jackknife methodology (Brown et al., 1999 and ONS,2000).

The jackknife methodology has the advantage of being an all-purpose method which works in stratified multistage samples and serves as a consistent estimator of variance when the parameter θ is a smooth function of population totals (Lohr,1999). If a nonlinear statistic has a local linear quality, then, the jackknife method should produce reasonably good variance estimates (Wolter, 1985). Jackknife may also be more stable than direct/explicit estimation because it is less influenced by extreme values (Brewer, 2002). However, bias may be larger and jackknife is not an accurate means of estimating the variances of some statistics such as percentiles. It is important to note that little is known about how jackknife performs in unequal probability without replacement sampling designs in general; see also Brewer (2002). The main justification for the jackknife in nonlinear problems is that it works well and its properties are known in linear problems.

Some checks are usually applied in order to detect extremely influential units that may contribute disproportionately to variance. Generally, robust techniques as well as weight trimming deal with outliers. All methods entail trading possibly increased bias for reduced variance to reduce mean square error. Smoothing of the adjustment factors in the estimation cells as well as grouping of estimation cells are also applied to deal with this problem for synthetic estimation (Hogan, 1993).

[3]The ultimate estimator is:

$$\widehat{V}(\widehat{\theta}) = \frac{1}{n(n-1)} \sum_{g=1}^{n} (\widehat{\theta}_g - \widehat{\theta})^2 \tag{1.10}$$

where n is the number of PSUs, $\widehat{\theta}$ is the estimator based on the sample and $\widehat{\theta}_g$ is an estimator based only on the data from the PSU g.

1.6.6 Post-enumeration Survey

In each country, the lists available - or constructible - as sampling frames for the post-enumeration survey are different (blocks, postal areas, geographical areas, buildings, etc.). The sampling procedure therefore needs to be adapted on a case-per-case basis. The only common point is the need for a multistage sampling as no complete list of people is available.

Examples of post-enumeration survey organization and operation may be found for instance in Hogan (1992, 1993) for the US census 1990 and Hogan (2000, 2001, 2003) and ZuWallack et al. (2000) for the US census 2000. General information can be found for UK in Pereira (2002) and Brown et al. (1999). The documentation for the Canadian census 1996 may be found in StatCan (1999) and the information about the PES 1996 in Australia in ABS (1999).

Survey procedures depend on the country (hardcopy questionnaires, phone interviews, etc.). The questionnaire has to include the variables necessary for matching with the census data set and the variables useful for the coverage analysis. A special emphasis has to be put on the search for the sampled households or people (contacts) and on the response rate in order to limit the bias in the coverage estimation.

The timing of the post-enumeration survey must be chosen with care. It should not be conducted too early (to avoid overlap with the census) nor too late (since changes in the population may occur). The survey is usually organized after census day, although some operations such as address listing sometimes occur before census day.

1.6.7 Treatment of Movers

Coverage estimations have to deal with changes in the population between census day and survey day (movers, births, deaths).

Births and deaths are usually dealt with during the survey and data editing is handled in a pragmatic manner. Newborns are removed from the survey data set. Deceased people are treated as a non-response if they are listed in the sample and simply disregarded if they are not (small bias).

Movers are potentially more susceptible to omission than non-movers in the census. They have to be treated with care during both the survey and the matching with the census data set. We define two types of movers in relation to the sample survey. The *out-movers* that lived in a sampling unit on census day but moved out before survey day and the *in-movers* that moved into a sampling unit between census day and survey day.

The treatment of the movers is related to possible bias in the DSE estimation because of the heterogeneity of the movers (Griffin, 2000):

$$\mathrm{bias}(\mathit{DSE}) = -\frac{N\,d\,(1-c)(1-m)}{(1+d\,c\,m)(d+1)} \tag{1.11}$$

where N is the population total being estimated, $d =$number of movers / number of non-movers, $c =$ census coverage for movers / census coverage for non-movers and $m =$survey coverage for movers / survey coverage for non-movers. The aim of the procedure is to get $d = 0$, or $c = 1$ or $m = 1$.

Different procedures relating to the enumeration of the people in the sampling units have been tested in the U.S. Census Bureau:

A. Construction of the list at the time of the census: information about out-movers collected by proxy. No information collected about in-movers.

B. Construction of the list at the time of the survey. Respondents are asked to provide the address where they lived on census day.

C. Construction of two lists: the first, at the time of the census and the second, at the time of the survey. Demographic information from the in-movers and matching information from the out-movers (better match rates) is used.

All procedures give an estimate of the number and percent matched for non-movers. Procedures A and B also give an estimate of the number and percent matched for out-movers and in-movers, respectively. Procedure C gives an estimate of the number of out-movers and in-movers, as well as the percent matched for out-movers. The bias of DSE with procedure B is expected to be smaller than with procedure A, but similar to procedure C.

Procedure B was used in 1990 but the unresolved match rate for in-movers was high. Procedure A was tested during the 1995 et 1996 Census Tests to improve the rate of match among movers and to avoid problems related to the planned 2000 non-response follow-up procedure (sampling, no more 100%). Results had good matching ability but problems occurred in the collection of demographic data for out-movers. Procedure C was tested during the Dress Rehearsal of 1998 and chosen for 2000.

With procedure C, $\widehat{N}_p = \widehat{N}_n + \widehat{N}_i$ and $\widehat{M} = \widehat{M}_n + (\widehat{M}_o/\widehat{N}_o)\widehat{N}_i$, where $\widehat{N}_n$, $\widehat{N}_i$ and $\widehat{N}_o$ are the estimated totals of non-movers, in-movers and out-movers, respectively; and $\widehat{M}_n$ and $\widehat{M}_o$ are the estimated totals of non-mover and out-mover matches (Griffin, 2000). Note that the estimated total of matches among the movers $(\widehat{M}_o/\widehat{N}_o)\widehat{N}_i$ is based on the rate of match for out-movers and on the total of movers estimated with the in-movers.

A fourth procedure was considered in UK. This option (procedure D) is to collect no information on movers and assume that they are just non-responses missing at random in the survey (*i.e.* no different from the non-movers that the survey does enumerate). The specific features of the UK and the approach to the One Number Census resulted in a choice between procedures A and D. Procedure A was chosen, with Procedure D as a reserve for specific cases where proxy information is poor (ONS, 2001).

1.7 General Remarks and Deviation from DSE Assumptions

Some points are worth noting about dual system estimation; see National Research Council (2004, p. 162).

First, the DSE formula includes II if we have census enumerations that either lacked sufficient information or were added too late to be included in the matching.

Second, there is no assumption that the P-sample must be more complete than the census. It is expected that the P-sample will miss some people. What is important is that the information obtained in the P-sample be of high quality for matching and satisfy the assumption of independence.

Third, a key assumption is that the procedures used to define who is in and who is not in the census are balanced. Failure to apply the same criteria for the correct enumerations in the E-sample and the matches with correct enumeration in the census will create a balancing error.

Fourth, the DSE is sample based. Consequently it is important to estimate not only the DSE itself but also its variance due to sampling and other errors. In addition, the number of individual population groups for which reliable coverage estimates can be developed is limited.

Fifth, if DSE results are to be used for domains that are smaller than those used in the post-stratification (estimation cells), we assume that the match rate and the correct enumeration rate are also valid for the domain. This assumption - known as the synthetic assumption- is strong.

The key assumption underpinning the DSE methodology is the independence between the census and the survey. The bias due to dependance may be important but is expected to be small provided both the census and the survey have high response rates (Brown et al., 1999. Adjustment for dependance between census and survey was applied in the UK to get a more accurate estimate of the population (Abbott et al., 2003). This adjustment makes use of auxiliary information.

Other points may disturb DSE model assumptions. We have to deal with heterogeneity caused by the movers, with unit and item non-response in the survey, with matching errors, with changes in the population between census day and survey day, with measurement error (*e.g.* misclassification in estimation cells) and with heterogeneity of capture probability in the estimation cells.

Many studies have been conducted at the U.S. Census Bureau in order to analyze the various errors; see for instance the total error analysis of 1990 in Mulry and Spencer (1993). The list of errors contains the matching error, the imputation error and the survey operation errors. Synthetic error was also included for 2000.

Correlation bias, due to causal dependence and heterogeneity in capture probability (see the DSE assumptions), was corrected in the A.C.E. Revision II (Shores, 2002). For example, demographic data were used to adjust for correlation bias due to a lack of independence between the probability of being counted in the census and in the survey (Kostanich, 2003). Other corrections such as measurement errors in the residence status and adjustment for missing data were included in the final results (Kostanich, 2003). No demographic data were included in the official results for coverage of the 1990 census.

Information about the treatment of outliers in the framework of the US coverage estimation may be found in *e.g.* Zaslavsky et al. (2001) for a general overview, Mule (2000) for the plan and Mule (2003b) for the results.

1.8 Adjustment of Census Counts

When estimating census coverage it becomes clear that the census count is not perfect. Therefore, different population counts are sometimes available: *e.g.* census count, census count adjusted to demographic data, census count adjusted by using the DSE methodology. One count may be used to distribute seats in the government and another may be used to distribute social funds between regions.

The USA has a long history of political and legal controversy when it comes to census adjustment. The question for the 1980 census was about whether coverage estimations could or should be used to adjust the census results for undercounts (Freedman and Navidi, 1992 and Hogan, 1992). In 1989, litigation about the 1980 census culminated in an agreement between the U.S. Department of Commerce (of which the Census Bureau is part) and a coalition of states, cities, and organizations led by New York City (region with a rather high undercount). The Bureau did not adjust for 1980 but was to prepare for adjustment for the 1990 census. New controversy among statisticians and further litigation led to the decision in 1991 not to adjust the 1990 census. As an unresolved issue, new discussions took place for the census 2000 but the A.C.E. did not successfully measure the large number of duplicates in the census; see *e.g.* Whitford (2002). The distribution of the seats in the House of Representatives and redistricting are therefore still based on the unadjusted census counts. Interesting general literature about the decisions regarding the 2000 census may be found in National Research Council (1999 and 2004) and the document prepared for Congress (U.S. Census Bureau, 2001).

In the UK, the One Number Census project was developed to adjust the 2001 census counts for undercounts. The aim was to measure this level of underestimation in the most acceptable way to provide a much clearer link between the census counts and the population estimates in order to adjust all the census count for undercounts. This means that the individual level data base was adjusted by including imputed households and persons to reflect underestimation. All counts can therefore be added to "one number" (Brown et al., 1999 and Pereira, 2002). Because final numbers are estimates, confidence intervals are supplied with the totals; see the documents on the website `www.statistics.gov.uk`. To our knowledge, UK is the first country that adjusted the official census counts completely based on sampling methodology.

In Australia, the results of the census post-enumeration survey are used with other administrative sources in the calculation of the Estimated Resident Population (ERP) (ABS,1999). The ERP is the population for all official purposes such as financial distribution and distribution of seats in parliament (ABS, 2004). Special discussions occur for the estimates of Queensland Aboriginal and Torres Strait Islander communities because the values are expected to be underestimated (Evans et al.,1993, Taylor and Bell, 2003, Hunter and Dungey, 2003).

Chapter 2

General Methodology for the Swiss Estimation

The coverage estimation of the Swiss 2000 Census mainly bases itself on the methodologies developed in the U.S. Census Bureau, the UK's Office for National Statistics (ONS) and the Australian Bureau of Statistics (ABS).

2.1 Objectives

Our aim is to estimate the components of overcoverage, undercoverage as well as the combined net coverage of the Swiss population census 2000.

Results are expected for the whole census data set as well as subgroups such as male and female, age groups, small and large communes, urban and rural communes, Swiss and foreigners, and the census methodologies CLASSIC and TRANSIT; see Appendix A.

The *target population* for the coverage estimations is defined as the resident population at their economic domicile and living in a private household; see Appendix A for the definitions. We assume that the population at the economic domicile and assigned to a collecting household is also part of the target population.

Note that the decision of making the estimations in the above-defined target population, especially including the type of domicile, led to some complexity in the estimation. This decision should be reviewed for a possible future coverage estimation; see Chapter 15.

Some comparisons between census and demographic data are shown in Appendix B. More reliable and detailed comparisons are expected with the dual system methodology.

2.2 Estimation Methodology

Overcoverage

The estimation of overcoverage is based on the *Enumeration-sample* or *E-sample*, that is selected in the census data, see Section 8.2, and on the results from the search for correct (CE) and erroneous enumerations (EE) in the E-sample, see Figure 2.1 and Section 10.2.

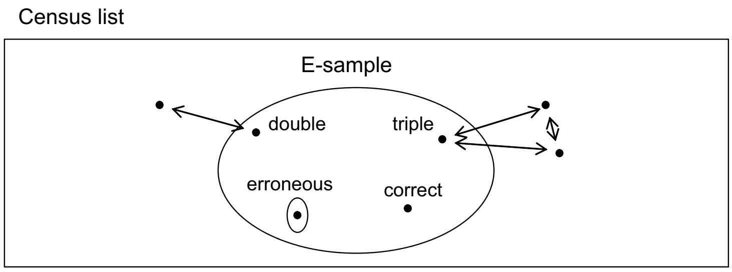

Figure 2.1: Illustration of the search for correct and erroneous enumerations in the E-sample.

The basic estimator of overcoverage is $1 - R_{ce}$ where $\widehat{R}_{ce} = \widehat{CE}/\widehat{N}_e$ is the estimated rate of correct enumeration based on the E-sample, with $\widehat{CE}$ the estimated number of correct enumerations in the census and $\widehat{N}_e$ the estimated census count; see the methodology in Chapter 3 and the results in Chapter 11.

Undercoverage

The estimation of undercoverage is based on the *Population sample* or *P-sample* and the results from the search for matches in the census, see Figure 2.2 and Section 10.1. The P-sample is a subset of the post-enumeration survey, called the Swiss Coverage Survey (SCS) and is as independent from the census as possible; see Section 8.1.

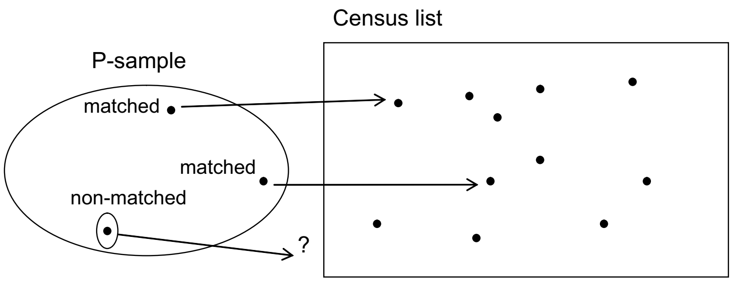

Figure 2.2: Illustration of the search for matches from the P-sample.

Information is collected during the SCS for non-movers and in-movers, but not for out-movers (procedure B for treating the movers). We have a 100% follow-up in the census and we expect a better rate of resolved matches than in the USA. Switzerland is much smaller and nearly final census data are available for matching.

The basic estimator of undercoverage is $1 - R_m$ where $\widehat{R}_m = \widehat{M}/\widehat{N}_p$ is the estimated rate of correct match based on the P-sample, with $\widehat{M}$ the estimated number of correct matches in the census and $\widehat{N}_p$ is the estimated census count; see the methodology in Chapter 4 and the results in Chapter 12.

Net coverage

The estimator of the net coverage is $\widehat{R}_{net} = C/\widehat{N}$, where C is the census count and $\widehat{N}$ is the dual system estimator of the population total. The corresponding rate of net undercoverage is $\widehat{R}_{under} = 1 - \widehat{R}_{net}$.

The dual system estimator of the U.S. Census Bureau with data from the E-sample and P-sample is the basis for the estimation of $\widehat{N}$; see Equation (1.6):

$$\widehat{N} = C\, \frac{\widehat{CE}}{\widehat{N}_e}\, \frac{\widehat{N}_p}{\widehat{M}} = C\, \frac{\widehat{R}_{ce}}{\widehat{R}_m} = C\, \widehat{CCF} \tag{2.1}$$

where $\widehat{CCF}$ is the estimated "coverage correction factor" and the other terms as above. Note that we do not include the term II in the formula. The reason is that we do not have whole person imputation in the Swiss census and very few entries are not data-defined (no names, imputation for most of the variables).

The dual system estimation methodology is applied in estimation cells (post-strata) $\ell = 1, .., L$ and then combined by using synthetic estimation to get estimates for the whole population as well as various sub-groups d of the population; see the methodology in Chapter 5, the construction of the estimation cells in Chapter 13 and the results in Chapter 14:

Variance

Variance estimation of the coverage estimators makes mainly use of the jackknife methodology, see the methodology in Chapter 6 and the results in Chapters 11, 12 and 14.

2.3 Data and Preliminaries

Three main data sets are used for the estimations: the census data set, the P-sample data set and the E-sample data set (subset of the census data set). Some complementary data sets are available for instance to define geographical areas and domains; see Chapters 7, 8 and 9.

Two important procedures are applied in order to get the basic information about the under- and overcoverage: (1) search for matches between the P-sample and the census data set, and (2) search for correct enumerations (CE) and erroneous enumerations (EE), in the E-sample; see Chapter 10.

The search for matches between the P-sample and the census is a complete search in the entire census data set. For matched entries, we have the information collected in the census and SCS questionnaires.

The search for correct enumeration in the E-sample is mostly restricted to a search for double entries in the census. For the E-sample people, we have only the data collected during the census. We have no complementary information about the real location, the real type of domicile and the real type of household of the E-sample people on census day. This point is further discussed in Chapters 3 and 15.

2.4 Checks before Estimations

Some checks are applied to the P-sample, E-sample and combined results before making the estimations. These checks aim at detecting possible units that are extremely influential on the estimates that may consequently contribute disproportionately to its variance.

Robust techniques are not broached because they would lead to even more complicated coverage estimations. If necessary, the intervention entails trimming of weights. The aim is, however, to modify weights only if we can expect a large improvement in the variance estimate while keeping a small bias.

2.5 Expected Results

The results for undercoverage are assumed to give us most of the information. We have detailed information about the P-sample matched and non-matched entries. Furthermore, comparisons between census and SCS characteristics of matched entries provide us with information about the potential measurement and misclassification errors.

The results for overcoverage are supposed to be less detailed and accurate than those for matches. The reason is the simplified procedure for the search for CE and EE. Only double entries and few fictitious entries are detected. We do not have any complementary information about the real enumeration status of the person (proper location? right population?).

Some choices and assumptions about the definition of correct matches and correct enumeration are necessary in order to combine the results from the search for matches and the search for CE in the DSE estimation methodology. For example, one of the challenges is the balancing error; see Section 5.

Chapter 3

Correct/Erroneous Enumeration and Overcoverage

The estimation of the overcoverage of the census data set is based on the results of the search in the E-sample for correct enumerations (CE) and erroneous enumerations (EE); see Sections 2.3 and 10.2.

The overcoverage may be due to various problems in the census process, such as a missing link between two domiciles, an enumeration at two places due to a move without the proper administrative notification, or scanning of non valid questionnaires.

The search for CE and EE is designed to detect only a part of the overcoverage component: multiple entries and fictitious entries. As a consequence we have only partial information about the CE and EE and we assume that all other types of overcoverage are negligible. The estimated overcoverage is a minimum value.

If an E-sample person matches a person with another entry in the census, we call it an *E-sample double*; see Figure 2.1 on page 24. The corresponding entry in the census is called a *doublet*. Similarly, we have an *E-sample triple* with the two corresponding *triplets*.

3.1 Rate of Correct Enumeration R_{ce} and Overcoverage

Analysis of the CE and EE in the E-sample leads to the estimation of R_{ce} the *rate of correct enumeration* or the proportion of CE in the census data set:

$$\widehat{R}_{ce} = \frac{\sum_{j \in s_e} w_{e,j} \, P_{ce,j}}{\sum_{j \in s_e} w_{e,j}} = \frac{\widehat{CE}}{\widehat{N}_e} \tag{3.1}$$

where $P_{ce,j} \in [0, 1]$ is the *status of correct enumeration* and $w_{e,j}$ is the weight of people j of the E-sample s_e, $\widehat{N}_e = \sum_j w_{e,j}$ is the estimated population total and $\widehat{CE} = \sum_j w_{e,j} \, P_{ce,j}$ is the estimated number of correct enumerations.

The complement $1 - \widehat{R}_{ce}$ is the estimator of the proportion of people that should not have been counted in the census. This is the *overcoverage* due to the erroneous enumerations. For example, a rate of correct enumeration of 99% means that 1% of the people in the census should not have been counted; *i.e.* there is an overcoverage of 1% in the census count.

Note that the estimator is defined as $\widehat{R}_{ce} = \widehat{CE}/\widehat{N}_e$ and not as $\widehat{CE}/C$ with the constant census count C; see Section 1.6.2.

Below, we develop the methodology for R_{ce}. The results for overcoverage can be directly deduced from those for R_{ce}.

The estimator for a specific subgroup or *domain* U_d of the population U is given by:

$$\widehat{R}_{ce,d} = \frac{\sum_{j \in s_e} w_{e,j}\, P_{ce,j}\, I_{jd}}{\sum_{j \in s_e} w_{e,j}\, I_{jd}} \text{ with } I_{jd} = \begin{cases} 1 & \text{if } j \in U_d \\ 0 & \text{otherwise, } j \in U \backslash U_d \end{cases} \tag{3.2}$$

I_{jd} is the domain indicator.

The choice of the status of correct enumeration $P_{ce,j}$ for each element $j = 1, .., n_e$ of the E-sample s_e is very important. Various definitions have been considered (see below).

Note that $P_{ce,j}$ is a characteristic for all elements of the census data set but measured only for the E-sample.

3.2 Definition of CE and EE in the Census

A general definition of correctness may be expressed as follows: an enumeration is assumed to be *correct* if it is complete, appropriate, unique, in the right population and properly located (see *e.g.* Hogan, 2003, with extension for the target population):

- Completeness means that the record is sufficient to identify a single person.
- Appropriateness means that the person should be included in the census.
- Uniqueness means that each person is enumerated only once.
- Right population means that the person is a member of the target population (private household and economic domicile).
- Proper geographical location means that the person is included where he/she should be included.

Various levels may be defined for the five criteria. These levels may be very strict or not.

We assume that completeness is satisfied in all cases because most of the population received a questionnaire with preprinted address, names and demographic information such as sex, date of birth, marital status and nationality. Furthermore, there is little imputation of these demographic variables.

We say that an E-sample person is appropriate if he/she is a real person (not fictitious). This definition is not very strict but a more constrained definition would not be verifiable for the E-sample.

Uniqueness was verified during the search for CE/EE. However, no information is available to determine which of the double/triple entries is the correct one, or even if one is correct among them. The status of multiple entries has to be estimated.

We do not use the criteria of right population and proper location in the definition of the correct enumeration because we do not have any information, besides the census, that would confirm or refute the membership in the target population and location.

An enumeration is *erroneous* if it is not correct.

None of the E-sample entries have an unresolved status of correct enumeration at the end of the search process. However, some assumptions have to be set; see below.

3.2.1 Simple Status of Correct Enumeration

We define the status of *simple* correct enumeration $P^{(s)}_{ce,j}$ for each element $j = 1, .., n_e$ of the E-sample:

$$P^{(s)}_{ce,j} = \begin{cases} 0 & \text{if } j \text{ is a fictitious enumeration} \\ 1 & \text{if } j \text{ is a match with the P-sample} \\ 1/2 & \text{if } j \text{ is a double} \\ 1/3 & \text{if } j \text{ is a triple} \\ 1 & \text{otherwise} \end{cases} \quad (3.3)$$

Fictitious elements are clearly not correct (not appropriate) and the status of correct enumeration is therefore $P_{ce,j} = 0$.

We assume that matches with the P-sample are correct because the existence of the person was confirmed by the SCS interviews ($P_{ce,j} = 1$); see the remarks in the Section 10.2.

Various approaches may be used to determine the status of correct enumeration for multiple entries. Without auxiliary information such as interviews, the status has to be estimated. We may consider two extreme situations: the E-sample entries are correct ($P_{ce,j} = 1$) and the doublets/triplets are erroneous, and the E-sample entries are erroneous ($P_{ce,j} = 0$) and the doublets/triplets are possibly correct. We choose the medium situation: $P_{ce,j} = 1/d$ with $d = 2$ for doubles entries and $d = 3$ for triple entries. The idea is that one of the two (of three) enumerations is correct (assumption) but we have no information about which is the correct one. Therefore, each double is considered as half correct and each triple is considered as one third correct.

E-sample entries not identified as fictitious or multiple entries are assumed to be correct ($P_{ce,j} = 1$). We do not have any other information that would enable us to better determine their correct enumeration status.

3.2.2 Alternative Statuses of Correct Enumeration

Some alternative correct enumeration statuses are defined for multiple entries. These alternatives use membership in the population and location of the doublets et triplets, as well as the partner if they have two domiciles.

Membership in the Population

The membership in the target population (private households, economic domicile) has a special role in the estimation. Actually, only people enumerated in the target population were eligible in the E-sample. The reason is that coverage estimations are expected for the target population.

In the definition of the simple status $P_{ce,j}^{(s)}$, doublets and triplets out of the target population are considered as real doublets and triplets. For the alternative status $P_{ce,j}^{(pop)}$, doublets and triplets are maintained only if they are members of the target population. The idea is that we have over-coverage of the target population only if the doublets and the triplets are in the target population too.

For multiple entries, we have $P_{ce,j}^{(pop)} = 1/d'$, with d' the number of doublets/triplets *in* the target population; see Figure 3.1.

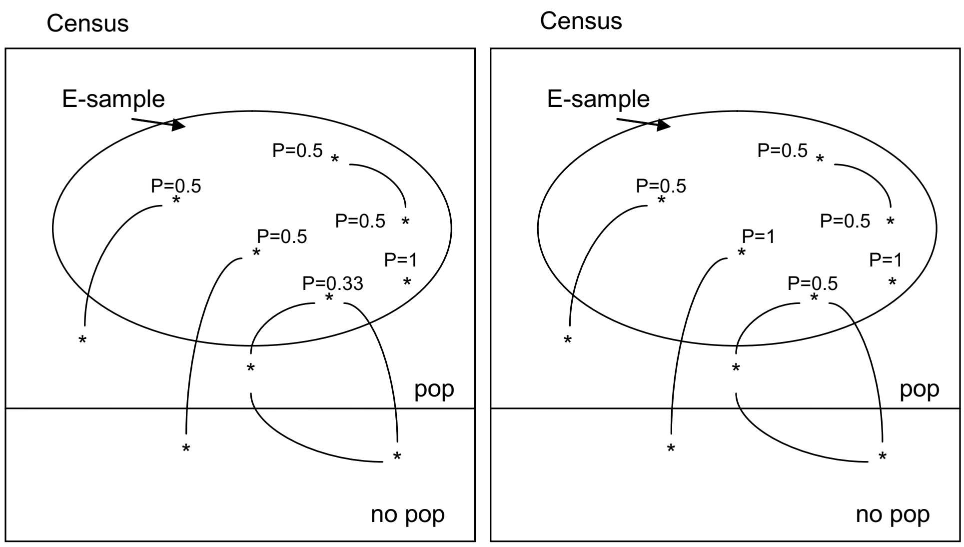

Figure 3.1: Status of correct enumeration $P_{ce,j}^{(s)}$ (left) and $P_{ce,j}^{(pop)}$ (right). The link between the points indicate the double and triple entries.

The resulting status $P_{ce,j}^{(pop)} \geq P_{ce,j}^{(s)}$ may be expressed by:

$$P_{ce,j}^{(pop)} = \begin{cases} 1/2 & \text{if } j \text{ is a double } \underline{\text{and}} \text{ the doublet is } \textit{in} \text{ the population} \\ 1 & \text{if } j \text{ is a double } \underline{\text{and}} \text{ the doublet is } \textit{out} \text{ of the population} \\ 1/3 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ both triplets are } \textit{in} \text{ the population} \\ 1/2 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ one and only one triplet is } \textit{in} \text{ the population} \\ 1 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ both triplets are } \textit{out} \text{ of the population} \\ P_{ce,j}^{(s)} & \text{otherwise} \end{cases} \tag{3.4}$$

Location

The location has less impact on the estimation methodology than the membership in the population. The reason is that, contrary to the restriction of the E-sample to the target population, we do not restrict the estimation to a particular place.

In the analysis of overcoverage, it is however interesting to look at the location of the doublets and triplets in comparison with the location of the E-sample entries. Therefore, we define the status of correct enumeration at the right location $P^{(loc)}_{ce,j}$, where the multiple entries are considered as real (partial) erroneous enumerations only if the doublets and triplets are enumerated around the address of the E-sample entry. Note that the accepted area around the address has to be defined.

For multiple entries, we have $P^{(loc)}_{ce,j} = 1/d'$, with d' the number of doublets/triplets *in* the area around the address.

The resulting status $P^{(loc)}_{ce,j} \geq P^{(s)}_{ce,j}$ is expressed with the same scheme as $P^{(pop)}_{ce,j}$ in Equation (3.4) but replacing "in the population" by "in the area" and "out of the population" by "out of the area".

Partner

When determining the status of correct enumeration of multiple entries, including the characteristics of the *partner enumeration* of the doublets and triplets can be quite revealing. Partner enumerations are actually available for the doublets and triplets that have civil addresses that differ from their economic addresses (additional record for the same person; two domiciles).

For example, consider one E-sample entry j with a doublet i living in a collective household and/or at a civil domicile, *i.e.* i is out of the target population, $P^{(s)}_{ce,j} = 1/2$ and $P^{(pop)}_{ce,j} = 1$. Suppose now that i has a partner k ($\neq j$) that is living in a private household and at the economic domicile, *i.e.* in the population. One can relax $P^{(pop)}_{ce,j}$ by assuming that k should also be considered as a doublet of j in the target population but was not found during the search, *i.e.* $P^{(popR)}_{ce,j} = 1/2$.

The relaxation of the status of correct enumeration $P^{(pop)}_{ce,j}$ for multiple entries is therefore: $P^{(popR)}_{ce,j} = 1/d'$, with d' the size of the set defined by the union of "doublets/triplets *in* the target population" with "doublets/triplets with a partner *in* the target population". We are now counting as doublets and triplets any partners of these non-population doublets and triplets which are in the population.

We define $P^{(popR)}_{ce,j} \leq P^{(pop)}_{ce,j}$:

$$P^{(popR)}_{ce,j} = \begin{cases} 1/2 & \text{if } j \text{ is a double } \underline{\text{and}} \text{ the doublet is } \textit{in} \text{ pop} \\ 1/2 & \text{if } j \text{ is a double } \underline{\text{and}} \text{ the doublet is } \textit{out} \text{ of pop} \\ & \quad \text{with partner } \textit{in} \text{ pop} \\ 1 & \text{if } j \text{ is a double } \underline{\text{and}} \text{ the doublet is } \textit{out} \text{ of pop} \\ & \quad \text{without partner or with partner } \textit{out} \text{ of pop} \\ 1/3 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ both triplets are } \textit{in} \text{ pop} \\ 1/3 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ one triplet is } \textit{in} \text{ pop} \\ & \quad \text{and one triplet is } \textit{out} \text{ of pop with partner } \textit{in} \text{ pop} \\ 1/3 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ both triplets are } \textit{out} \text{ of pop} \\ & \quad \text{with both partners } \textit{in} \text{ pop} \\ 1/2 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ one triplet is } \textit{in} \text{ pop and one triplet} \\ & \quad \text{is } \textit{out} \text{ of pop without partner or with partner } \textit{out} \text{ of pop} \\ 1/2 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ both triplets are } \textit{out} \text{ of pop with} \\ & \quad \text{only one partner } \textit{in} \text{ pop} \\ 1 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ both triplets are } \textit{out} \text{ of pop} \\ & \quad \text{without partner or with partner } \textit{out} \text{ of pop} \\ P^{(s)}_{ce,j} & \text{otherwise} \end{cases} \tag{3.5}$$

With the same scheme but replacing "in pop" by "in area" and "out of pop" by "out of the area", we can relax $P^{(loc)}_{ce,j}$ in order to define $P^{(locR)}_{ce,j}$.

Combined Effects

We can also define various statuses that combine the simple status with membership in the population, the location and relaxation for partners. To illustrate this, let $P^{(poploc)}_{ce,j}$ that depends on the population and location:

$$P^{(poploc)}_{ce,j} = \begin{cases} 1/2 & \text{if } j \text{ is a double } \underline{\text{and}} \text{ the doublet is } \textit{in } \Omega \\ 1 & \text{if } j \text{ is a double } \underline{\text{and}} \text{ the doublet is } \textit{out} \text{ of } \Omega \\ 1/3 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ both triplets are } \textit{in } \Omega \\ 1/2 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ one and only one triplet is } \textit{in } \Omega \\ 1 & \text{if } j \text{ is a triple } \underline{\text{and}} \text{ both triplets are } \textit{out} \text{ of } \Omega \\ P^{(s)}_{ce,j} & \text{otherwise} \end{cases} \tag{3.6}$$

with Ω the set of the elements "in the population and in the area".

Note that the membership and location effects increase the rate of correct enumeration (status increases) but the relaxation decreases the rate (status decrease).

3.3 General Comments

The most interesting point in the estimation is the rate of correct enumeration with the simple status $P^{(s)}_{ce,j}$ as well as the rate that includes population membership, especially with the relaxation for the partners (statuses: $P^{(pop)}_{ce,j}$ and $P^{(popR)}_{ce,j}$). The information about location has a lower impact on the analysis; see the results in Chapters 11.

Chapter 4

Matches and Undercoverage

The estimation of the undercoverage in the census data set makes use of the results of the search for matches between the SCS (P-sample) and the census; see Section 10.1.

The undercoverage may be due to various problems in the census process, such as people not listed in the registers, people not contacted by the enumerators or failure in the data processing (*e.g.* loss during scanning).

If a P-sample entry is matched during the search, we call it a *matched* entry. It is called a *non-matched* entry otherwise. The corresponding entry in the census is called the *match*.

Contrary to the treatment of correct enumerations, we have two data sets: SCS and census. Therefore, the correctness must first be defined and checked for the P-sample entries before defining the correct matches found in the census data set; also called "match with correct enumerations".

4.1 Correct P-sample Entries

The general definition of Section 3.2 is used to define the correct P-sample entries.

All P-sample people are assumed to be correct. We do not need to select any subsection of the sample for the estimations. The reason is (1) that P-sample people were contacted during the interviews (assumption: complete, appropriate), (2) that P-sample entries were checked for double entries during the SCS data editing (uniqueness), (3) that the determination of population membership during the SCS is assumed accurate (right population), (4) that the locations on census day and on SCS day collected during the SCS are assumed to be accurate because of the special effort made to collect all the addresses during the interviews.

However, we should note that errors may also occur in SCS data collection. For instance, measurement errors in demographic data (*misclassification*) as well as uncertainties in the type of domiciles (with possible inversion of locations as well) are possible but not checked.

4.2 Rate of Match $\widehat{R}_m$ and Undercoverage

Analysis of the matches between the P-sample and the census leads to the estimation of R_m the *rate of correct matches* or the proportion of P-sample people that have been correctly enumerated in the census:

$$\widehat{R}_m = \frac{\sum_{j \in s_p} w_{p,j}\ P_{m,j}}{\sum_{j \in s_p} w_{p,j}} = \frac{\widehat{M}}{\widehat{N}_p} \tag{4.1}$$

where $P_{m,j}$ is the *status of correct match* and $w_{p,j}$ is the weight of people j of the P-sample s_p, $\widehat{N}_p = \sum_j w_{p,j}$ is the estimated total population from the P-sample and $\widehat{M} = \sum_j w_{p,j}\ P_{m,j}$ is the estimated number of correct matches from the P-sample.

The complement $1 - \widehat{R}_m$ is the proportion of people that should have been but were not enumerated in the census. This is the *undercoverage* due to some errors in the census process. For example, a rate of match enumeration of 99% means that 1% of the people should have been counted in the census but were not; *i.e.* there is an undercoverage of 1% in the census count.

Below we develop the methodology for $\widehat{R}_m$. The results for the undercoverage can be directly deduced from those of $\widehat{R}_m$.

The estimation for a specific subgroup or *domain* U_d of the population U is given by:

$$\widehat{R}_{m,d} = \frac{\sum_{j \in s_p} w_{p,j}\ P_{m,j}\ I_{jd}}{\sum_{j \in s_p} w_{p,j}\ I_{jd}} \text{ with } I_{jd} = \begin{cases} 1 & \text{if } j \in U_d \\ 0 & \text{otherwise, } j \in U \backslash U_d \end{cases} \tag{4.2}$$

Similarly to the status of correct enumeration $P_{ce,j}$ in the E-sample, we need to define the status of correct match $P_{m,j}$ for each element $j = 1, .., n_p$ of the P-sample (match with a correct enumeration). Various definitions were considered (see below).

4.3 Definition of Correct Match

We assume that completeness is satisfied for all matches in the census data set; see Section 3.2.

We also assume that all matches are appropriate. With an accurate matching, none of the matches should be fictitious if matched to one P-sample entry (appropriate). This definition may be restricted by including the information about the demographic variables; see below the alternative statuses of correct match.

The uniqueness of the matches is assumed satisfied but was not checked. The assumption lays on the observation about the low number of multiple entries in the E-sample.

The criteria of right population and proper location may be checked by comparing with the information collected in the P-sample. More or less restrictive criteria may be defined; see the alternative statuses below.

Note that the right population and proper location criteria may be related. In the case of two domiciles, an inversion in location is possibly related to an inversion in the types of domiciles.

A match is *erroneous* if it is not correct.

4.3.1 Simple Status of Correct Match

We define the status of *simple* correct match $P_{m,j}^{(s)}$ for each element $j = 1, .., n_p$ of the P-sample:

$$P_{m,j}^{(s)} = \begin{cases} 0 & \text{if } j \text{ is a non-matched entry} \\ 1 & \text{if } j \text{ is a matched entry} \end{cases} \tag{4.3}$$

The definition of $P_{m,j}^{(s)}$ does not include any information about the classification of the matches (sex, origin, etc.), the membership in the population and location. If any entry is found somewhere in the entire census data set, it is a match.

4.3.2 Alternative Statuses of Correct Match

Alternative statuses of correct match are defined by including the classification (sex, origin, etc.), membership in the population and location. The information about partners is also included in a similar way as for the status of correct enumeration in the E-sample.

Classification

For matched entries of the P-sample, we can compare the information collected during the SCS and the census.

Suppose that the element j from the P-sample is in the domain d (*e.g.* male age 10-19) and that the match i in the census is out of the domain d (domain d', *e.g.* male age 80+); see Figure 4.1. We have a *misclassification error*.

		Census	
		d	d’
P-sample	d		$n_{dd'}$
	d’	$n_{d'd}$	

Figure 4.1: Distribution of the data between the P-sample and the census (matched entries). Classification of data in the non-overlapping domains d and d'.

A misclassification error may be seen as a special case of coverage error. If the SCS data collection is chosen as the reference value, there is an underestimation of the census count in domain d and an overestimation of the census count out of d.

We can define the status of correct match $P_{m,j}^{(d)}$ in a domain d:

$$P_{m,j}^{(d)} = \begin{cases} 0 & \text{if } j \text{ is a non-matched entry} \\ 0 & \text{if } j \text{ is a matched entry } \underline{\text{and}} \text{ the match is } \textit{out} \text{ of } d \\ 1 & \text{if } j \text{ is a matched entry } \underline{\text{and}} \text{ the match is } \textit{in } d \end{cases} \tag{4.4}$$

In the results of Chapter 12, we give preference to some general analysis of the misclassification error such as the detection of highly misclassified variables or the balancing between "in d in SCS and out of d in the census" and "out of d in SCS and in d in the census". In that context we define the asymmetry factor:

$$\phi_{d,d'} = \max\left(\frac{n_{dd'}}{n_{d'd}}, \frac{n_{d'd}}{n_{dd'}}\right) = \frac{\max(n_{dd'}, n_{d'd})}{\min(n_{dd'}, n_{d'd})} \tag{4.5}$$

This analysis gives important results about the potential misclassification of the people and is also useful as a preliminary step for the choice of variables that are eligible for the construction of estimation cells in the dual system estimation. We note that many reasons may lead to misclassification such as bad or unclear formulation of the question, or differences in the survey methodology (*e.g.* paper form versus computer assisted interviews).

Membership in the Population

The simple rate of correct match may be extended to take into account membership of the match to the target population. The idea of the status of correct match in the population $P^{(pop)}_{m,j}$ is that we have an undercoverage of the target population if the match is out of the population: $P^{(pop)}_{m,j} \leq P^{(s)}_{m,j}$:

$$P^{(pop)}_{m,j} = \begin{cases} 0 & \text{if } j \text{ is non-matched} \\ 0 & \text{if } j \text{ is matched } \underline{\text{and}} \text{ the match is } \textit{out} \text{ of the population} \\ 1 & \text{if } j \text{ is matched } \underline{\text{and}} \text{ the match is } \textit{in} \text{ the population} \end{cases} \tag{4.6}$$

The membership in the population cannot however be considered in the same way as a domain as in the classification above because only people in the target population have been selected in the P-sample (according to the SCS); see Figure 4.2. An undercoverage in the target population is related to an overcoverage out of this population but we do not have the balanced information about people out of the target population with matches in the population.

		Census	
		IN pop	OUT pop
P-sample	IN pop		
	OUT pop	?	?

Figure 4.2: Comparison of data in and out of the target population for the P-sample and the census (matched entries). We do not have any information about the cells in the lower part of the table.

Location

We also define a rate of correct match that depends on the location. The idea is that the match is correct if it is located near the location given by the SCS.

We define the status of correct matches in the proper location $P^{(loc)}_{m,j} \leq P^{(s)}_{m,j}$ with the same scheme as Equation (4.6) but replacing "population" by "area".

The analysis of location of the matches gives potentially interesting results about the location in the census, such as possible time delay for movers between census day and enumeration day, as well as errors in determination of the type of domicile.

Relaxation for the Partner

Determining the type of domicile is a known difficulty in the census and SCS data collection. Furthermore, the search for matches stopped after having found one eligible match. No further search was processed to determine if the partner would also be eligible.

The match, as well as the partner, may be in the right type of domicile or not, the right type of household or not, and in the proper location or not. If the match is in the right domicile and may have possible the false location A but the partner is in the false domicile and the proper location B, we may have possibly an exchange between the two types of domiciles; in the census and/or the SCS. It would then not be an underestimation of the target population (underestimation at the place A balanced by an overestimation at the place B) but an error in the type of domicile.

In order to take into account the weakness arising from the type of domicile and the search for matches (in the case of two domiciles), we define the relaxed rate of correct matches in the population to "correct population for the match or its partner (if it has one partner)" $P^{(popR)}_{m,j}$:

$$P^{(popR)}_{m,j} = \begin{cases} 0 & \text{if } j \text{ is non-matched} \\ 0 & \text{if } j \text{ is matched } \underline{\text{and}} \text{ the match is } \textit{out} \text{ of pop} \\ & \quad \text{without partner or with a partner } \textit{out} \text{ of pop} \\ 1 & \text{if } j \text{ is matched } \underline{\text{and}} \text{ the match is } \textit{out} \text{ of pop} \\ & \quad \text{with a partner } \textit{in} \text{ pop} \\ 1 & \text{if } j \text{ is matched } \underline{\text{and}} \text{ the match is } \textit{in} \text{ pop} \end{cases} \tag{4.7}$$

Similarly, we define $P^{(locR)}_{m,j}$ by replacing "pop" by "area".

Differences between $P^{(pop)}_{m,j}$ and $P^{(popR)}_{m,j}$, as well as between $P^{(loc)}_{m,j}$ and $P^{(locR)}_{m,j}$ can be observed only for matches with partners.

Combined Effects

Many combined statuses may be defined in the same way as used for the statuses of correct enumerations of Chapter 3.

To demonstrate this, we define the status $P^{(poploc)}_{m,j}$ that combines the membership in the population and the location:

$$P_{m,j}^{(poploc)} = P_{m,j}^{(pop)}\, P_{m,j}^{(loc)} = \begin{cases} 0 & \text{if } j \text{ is non-matched} \\ 0 & \text{if } j \text{ is matched \underline{and} the match is \textit{out} of pop or \textit{out} of area} \\ 1 & \text{if } j \text{ is matched \underline{and} the match is \textit{in} pop and \textit{in} area} \end{cases} \tag{4.8}$$

Note that, contrary to the rates of correct enumerations, the membership and location effects decrease the rate of correct matches but the relaxation to the partners increases the rate.

4.4 General Comments

Determining various statuses and corresponding rates of correct matches is of great interest when analyzing undercoverage, especially the simple rate and the effect of the population and location.

The analysis will also place emphasis on parallel analysis such as misclassification error and detailed location of the matches; see Chapter 12.

Chapter 5

DSE and Net Coverage

The estimation of the net coverage of census data set is based on the results from the search in the E-sample for correct enumerations (CE) and erroneous enumerations (EE) and from the search for matches between the SCS (P-sample).

The estimated rates of correct enumeration $\widehat{R}_{ce}$ and correct match $\widehat{R}_m$ are combined in the dual system estimator to get the rate of net coverage and the rate of net undercoverage. The combined results include all the effects coming from undercoverage and overcoverage components. The overcoverage may partially compensate for undercoverage with different effects depending on the considered subgroup of interest.

5.1 Rate of Net Coverage $\widehat{R}_{net}$ and Coverage Correction Factor *CCF*

The estimated *rate of net coverage* $\widehat{R}_{net}$ is given by:

$$\widehat{R}_{net} = \frac{C}{\widehat{N}} \tag{5.1}$$

where C is the census count and $\widehat{N}$ is the dual system estimator of the population count.

The complement $\widehat{R}_{under} = 1 - \widehat{R}_{net}$ is the estimated proportion of people that should have been counted in the census but were not. This is the estimated *net undercoverage* due to missing entries in the census (undercoverage) partially or fully compensated by erroneous enumerations (overcoverage).

The rate of net coverage $\widehat{R}_{net}$ may be larger than 1 with a negative rate of net undercoverage $\widehat{R}_{under}$. In this case, we have more overcoverage than undercoverage and the census count is overestimated.

We also define the estimated *coverage correction factor* $\widehat{CCF} = \widehat{N}/C = 1/\widehat{R}_{net}$ as the factor of correction by which the census count has to be multiplied to reach the dual system estimator.

For example, a rate of net coverage of 98.5% means that, overall, 1.5% of the population have not been counted in the census (*e.g.* 2% undercoverage compensated by 0.5% overcoverage). In that case, the net undercoverage of the census is 1.5% and the correction factor is 1.0152.

Results are mostly given below in the form of the net coverage $\widehat{R}_{net}$ and of the correction factor $\widehat{CCF}$.

5.2 Dual System Estimator

The dual system estimation is applied by estimation cells, or post-strata, in order to satisfy the assumptions of the methodology where possible; see Section 1.6.

Let $\Lambda = \{1, .., \ell, .., L\}$ the set of L estimation cells. The population total N is estimated by:

$$\widehat{N} = \sum_{\ell=1}^{L} \widehat{N_\ell} \tag{5.2}$$

where the dual system estimator $\widehat{N}_\ell$ of the population total N_ℓ in a estimation cell $\ell = 1, .., L$ is given by:

$$\widehat{N}_\ell = C_\ell \left[\frac{\widehat{CE}_\ell}{\widehat{N}_{e,\ell}} \right] \left[\frac{\widehat{M}_\ell}{\widehat{N}_{p,\ell}} \right]^{-1} = C_\ell \, \frac{\widehat{R}_{ce,\ell}}{\widehat{R}_{m,\ell}} = C_\ell \, \widehat{CCF}_\ell \tag{5.3}$$

with

- C_ℓ is the census count in the estimation cell ℓ;
- $\widehat{N}_{e,\ell}$ is the estimated total population in ℓ based on the E-sample;
- $\widehat{CE}_\ell$ is the estimated number of correct enumerations in ℓ in the census data set based on the E-sample and search for CE/EE;
- $\widehat{N}_{p,\ell}$ is the estimated total population in ℓ based on the P-sample;
- $\widehat{M}_\ell$ is the estimated number of people in ℓ matched to the census based on the P-sample and search for matches;

and:

- $\widehat{R}_{ce,\ell} = \widehat{CE}_\ell / \widehat{N}_{e,\ell}$ is the estimated rate of correct enumeration in ℓ; see Chapter 3;
- $\widehat{R}_{m,\ell} = \widehat{M}_\ell / \widehat{N}_{p,\ell}$ is the estimated rate of correct match in ℓ; see Chapter 4;
- $\widehat{CCF}_\ell = \widehat{R}_{ce,\ell} / \widehat{R}_{m,\ell}$ is the coverage correction factor in ℓ.

Note that C_ℓ is a constant. Only the ratio $\widehat{CCF}_\ell$, $\ell = 1, .., L$, that combine the results $\widehat{R}_{ce,\ell}$ and $\widehat{R}_{m,\ell}$ from both the P-sample and the E-sample are random values.

5.3 Estimation in Domains

The estimation in specific subgroups or domains is based on the *synthetic assumption*. We assume that the coverage correction factor, the inverse of the rate of net coverage, is constant in each estimation cell.

The estimated rate of net coverage $\widehat{R}_{d,net}$ and the coverage correction factor $\widehat{CCF}_d$ are given by:

$$\widehat{R}_{net,d} = \frac{C_d}{\widehat{N}_d} \quad \text{and} \quad \widehat{CCF}_d = \frac{\widehat{N}_d}{C_d} = (\widehat{R}_{net,d})^{-1} \tag{5.4}$$

with C_d the census count in domain d, $\widehat{N}_d$ the total dual system estimator in domain d, and

$$\widehat{N}_d = \sum_{\ell=1}^{L} C_{d\ell}\, \frac{\widehat{N}_\ell}{C_\ell} = \sum_{\ell=1}^{L} C_{d\ell}\, \frac{\widehat{R}_{ce,\ell}}{\widehat{R}_{m,\ell}} = \sum_{\ell=1}^{L} C_{d\ell}\, \widehat{CCF}_\ell \tag{5.5}$$

where the $C_{d\ell}$ is the total in the census of the elements in the intersection between the domain d and the estimation cell ℓ and $\widehat{CCF}_\ell$ is the ratio between $\widehat{R}_{ce,\ell}$ and $\widehat{R}_{m,\ell}$ in ℓ.

Note that, as in Equation (5.3), the only random quantities are $\widehat{CCF}_\ell$, $\ell = 1, .., L$, which combine the results $\widehat{R}_{ce,\ell}$ from the E-sample and $\widehat{R}_{m,\ell}$ from the P-sample at the estimation cell level.

5.4 Construction of Estimation Cells (Post-Strata)

The choice of the set of estimation cells, or post-strata, $\Lambda = \{1, .., \ell, .., L\}$ is a key point in the estimation.

The first objective is to group people with similar census capture probabilities in order to reduce bias in the DSE. The second objective is to group people with similar net undercount to get reliable synthetic estimations. The third objective is to group people in order to capture differences in important variables.

The construction also has to take into account some constraints. The estimation cells should not be too small in order to control variance and minimize ratio bias (DSE=ratio estimator). They must be definable in both the P-sample and the census data set. They should also be based on variables with a low misclassification error in order to avoid heterogeneity (where possible, classification should be in the same estimation cell for both the P-sample and the census).

Besides the SCS and the 2000 census data sets, we do not use any auxiliary data sets to construct estimation cells. Therefore, we are in a possible bias situation where data are used to determine the estimation cells and these estimation cells are then applied to the same data. However, we do not have any other available information such as variance estimates or demographic data.

A set of eligible variables is first selected as a basis to define the estimation cells. Selection makes use of general considerations about the important variables for coverage issues and analysis of misclassification errors of variables (comparison, for P-sample matched entries, between the SCS and census data).

Logistic and discrimination models are then used to extract a subset of variables for further work in the construction. The set of final variables is used to construct the preliminary estimation cells. New cross-classifications are then collapsed and integrated step by step in order to assure a minimum sample size in each final estimation cell; see Chapter 13 for the detailed procedure.

5.5 Balancing Correct Enumerations and Correct Matches

One of the key subjects in the dual system estimation methodology is the definition of the status of correct enumeration $P_{ce,j}$ in the E-sample s_e and the status of correct match $P_{m,j}$ in the P-sample s_p. Actually, these statuses are the basis for calculation of $\widehat{R}_{ce,\ell}$ and $\widehat{R}_{m,\ell}$ in the estimation cells $\ell = 1,..,L$ in the dual system estimation:

$$\widehat{R}_{ce,\ell} = \frac{\sum_{j\in s_e} w_{e,j}\, P_{ce,j}\, I_{j\ell}}{\sum_{j\in s_e} w_{e,j}\, I_{j\ell}} \tag{5.6}$$

$$\widehat{R}_{m,\ell} = \frac{\sum_{j\in s_p} w_{p,j} P_{m,j}\, J_{j\ell}}{\sum_{j\in s_p} w_{p,j}\, J_{j\ell}} \tag{5.7}$$

with the weights $w_{e,j}$ in s_e and $w_{p,j}$ in s_p and the estimation cell indicator variables $I_{j\ell}$ for $j \in s_e$ and $J_{j\ell}$ for $j \in s_p$ that equal to 1 if $j \in \ell$ and 0 otherwise.

Generalities

The combination of overcoverage and undercoverage components needs *balancing*. In other words, the same definition of correctness must be used for enumerations in the E-sample and for matches with the P-sample. If balancing is not done, the dual system estimator will be biased.

Balancing has to be done for location: a match that is located far from the reference address collected during the SCS can be considered as erroneous (not correct) *only* if the search for correct enumeration also detects this match as erroneous because it is not in the proper location.

In addition to location, balancing has to be done for the target population. A match that is not in the target population may be excluded (considered as not correct) *only* if the search for correct enumeration also detects this match as erroneous because it is not in the target population.

As a result, we cannot combine any definition of the status $P_{ce,i}$ with any other definition of status $P_{m,i}$ without risk of bias in the estimate.

Preliminary Definition of Correctness

For balancing, the same definition of correctness has to be applied for correct enumerations and correct matches. We call the set of criteria: *dse-correctness*.

All P-sample and E-sample entries are assumed to be complete. Therefore, completeness is included in the definition of dse-correctness and assumed to be satisfied in all cases.

We assume that the P-sample entries are all appropriate and that the search for fictitious entries in the E-sample detected the non appropriate entries. Therefore, appropriateness is included in the definition of dse-correctness.

We assume that P-sample entries are unique, that the search for multiple entries in the E-sample is accurate to detect the non unique elements and that the estimation of the status of correct enumeration is also reliable for the multiple entries. Therefore, uniqueness is included in the definition of dse-correctness.

The location of P-sample matches is verifiable (according to the SCS data) but the location of the E-sample entries is not checked during the search for CE and EE. Therefore, we cannot include the location in the definition of dse-correctness.

Membership in the target population is a special case in the estimation. First, the P-sample is a set of people in the target population according to SCS data and the E-sample is a set of people in the target population according to census data. Second, some matches are found to be out of the target population according to the census but we do not have any complementary information to confirm or refute membership of the E-sample people to the target population[1]. The decision whether or not to include membership in the definition of dse-correctness is not straightforward. Some assumptions have to be set.

Membership in the Target Population

The dual system estimator $\widehat{N}_\ell$ in estimation cell ℓ is a function of C_ℓ, $\widehat{R}_{ce,\ell}$ and $\widehat{R}_{m,\ell}$.

The census count C_ℓ may include only people in the target population $C_\ell^{(pop)}$ or people also out of the target population (non private households or non economic domiciles) $C_\ell^{(s)}$. This term does not dictate any constraint in the choice of the treatment of membership in the population.

The rate $\widehat{R}_{m,\ell} = \widehat{M}_\ell / \widehat{N}_{p,\ell}$ is based on the P-sample selected in the target population according to the SCS. Therefore $\widehat{N}_{p,\ell}$ is the estimator of the population total in the target population. The search for matches give the information about membership of the matches to the target population. We can estimate the total number of matches $\widehat{M}_\ell^{(s)}$ (status $P_{m,i}^{(s)}$) as well as the total number of matches in the target population $\widehat{M}_\ell^{(pop)}$ (status $P_{m,i}^{(pop)}$ or $P_{m,i}^{(popR)}$). Note that the difference between $P_{m,i}^{(s)}$ and $P_{m,i}^{(pop)}$ is not large in our case; see Chapter 12.

The rate $\widehat{R}_{ce,\ell} = \widehat{CE}_\ell / \widehat{N}_{e,\ell}$ is based on the E-sample. According to the census, all E-sample entries are in the target population. Therefore $\widehat{N}_{e,\ell}$ is the estimator of the population total in the target population.

The search for CE in the E-sample leads to the detection of fictitious entries and multiple entries but not to entries that are enumerated by mistake in the target population. Based on the results from the search, we can estimate the number of correct enumeration $\widehat{CE}_\ell^{(s)}$:

$$\widehat{CE}_\ell^{(s)} = \sum_{j \in s_e} w_{e,j}\, P_{ce,j}\, I_{j\ell} \tag{5.8}$$

or choose one estimator that includes an estimated proportion φ_ℓ of E-sample entries that are out of the target population $\widehat{CE}_\ell^{(pop)}$:

$$\widehat{CE}_\ell^{(pop)} = \varphi_\ell \sum_{j \in s_e} w_{e,j}\, P_{ce,j}\, I_{j\ell} \tag{5.9}$$

[1]Note that matches that are in the E-sample are all in the target population according to the census and the SCS. This intersection cannot be used for the estimation of the proportion of E-sample people that are really in the target population.

In both forms, we have to define $P_{ce,j}$. The second form has the advantage of (partially) remedying some incomplete results from the search for CE and EE. However, we need to estimate the factor φ_ℓ.

We look at two forms of the dual system estimator:

$$\widehat{N}_\ell^{(s)} = C_\ell^{(pop)} \left[\frac{\widehat{CE}_\ell^{(s)}}{\widehat{N}_{e,\ell}}\right] \left[\frac{\widehat{M}_\ell^{(s)}}{\widehat{N}_{p,\ell}}\right]^{-1} \tag{5.10}$$

$$\widehat{N}_\ell^{(pop)} = C_\ell^{(pop)} \left[\frac{\widehat{CE}_\ell^{(pop)}}{\widehat{N}_{e,\ell}}\right] \left[\frac{\widehat{M}_\ell^{(pop)}}{\widehat{N}_{p,\ell}}\right]^{-1} \tag{5.11}$$

In both forms, we select the census counts in the target population $C_\ell^{(pop)}$ because of the general framework of the project: results expected for the target population and selection of the P-sample and the E-sample among this population.

The first form $\widehat{N}_\ell^{(s)}$ satisfies the needs for balancing because determination of correctness of the E-sample entries and matches does not depend on membership in the target population. The disadvantage of this form is that the effect of the target population is not fully included in all the estimated totals. A justification of this form is based on the assumption (not verifiable) that people who are mistakenly classified in the census as belonging to the target population are balanced by people who are mistakenly classified as belonging outside the target population.

The second form $\widehat{N}_\ell^{(pop)}$ is expected to satisfy balancing indirectly by the correction factor φ_ℓ. The definition of correctness is not the same for the status of correct enumeration and the status of match but the correction is designed to balance the estimate. The advantage of this form is that the effect of the target population is included at all stages but the difficulty is to estimate φ_ℓ in such a way that we reduce bias in the estimates.

If φ_ℓ is estimated by $\widehat{M}_\ell^{(pop)}/\widehat{M}_\ell^{(s)}$, we have $\widehat{N}_\ell^{(pop)} = \widehat{N}_\ell^{(s)}$. Here again, we assume that people who are mistakenly classified in the target population are balanced by people who are mistakenly classified outside the target population. The global estimator $\varphi_\ell = \widehat{M}^{(pop)}/\widehat{M}^{(s)}$ is probably more stable but may lead to a larger bias at the estimation cell level. Some other forms such as estimates in each first stage sampling stratum (variable `stradap`, see Section 8.1) are also conceivable.

For both $\widehat{N}_\ell^{(s)}$ and $\widehat{N}_\ell^{(pop)}$, the status of correct match $P_{m,i}$ is clearly defined. The status of correct enumeration of fictitious entries is clearly $P_{ce,i} = 0$ but the status of multiple entries still needs to be determined. Following the line of thought, we choose to define the status of correct enumeration on the basis of the doublets and triples in the target population (reference framework = census for the search). Obtaining a status that depends on the population but is also relaxed for the partners would be a wiser choice.

Final Dual System Estimator

For balancing purposes as well as to keep the estimators as simple as possible, we chose to use the first form $\widehat{N}_\ell^{(s)}$ for the dual system estimation in Equation (5.3).

The status of correct enumeration is set to $P_{ce,i} = P_{ce,i}^{(popR)}$ from Equation (3.5) in order to take into account the fictitious elements and the multiple entries in the target population, with relaxation for the partners.

The status of correct match is set to $P_{m,i} = P_{m,i}^{(s)}$ from Equation (4.3) in order to satisfy the need for balancing.

The final estimator is then:

$$\widehat{N}_\ell = \widehat{N}_\ell^{(s)} = C_\ell^{(pop)} \left[\frac{\widehat{CE}_\ell^{(s)}}{\widehat{N}_{e,\ell}}\right] \left[\frac{\widehat{M}_\ell^{(s)}}{\widehat{N}_{p,\ell}}\right]^{-1} = C_\ell^{(pop)} \frac{\widehat{R}_{ce,\ell}^{(popR)}}{\widehat{R}_{m,\ell}^{(s)}} \tag{5.12}$$

$$= C_\ell^{(pop)} \left[\frac{\sum_{j\in s_e} w_{e,j}\, P_{ce,j}^{(popR)}\, I_{j\ell}}{\sum_{j\in s_e} w_{e,j}\, I_{j\ell}}\right] \left[\frac{\sum_{j\in s_p} w_{p,j}\, P_{m,j}^{(s)}\, J_{j\ell}}{\sum_{j\in s_p} w_{p,j}\, J_{j\ell}}\right]^{-1} \tag{5.13}$$

The chosen estimator is somewhat conservative in the sense that it is not founded on the assumption that the determination of membership in the population is perfect for the P-sample. Actually, this determination, especially the type of domicile, is known to be difficult in any case.

We note that balancing is assumed to be met at the estimation cell level, which possibly limits the estimation bias in the various subgroups of interest.

Chapter 6

Variance Estimation

This chapter presents the methodology developed for estimation of the variance of overcoverage, undercoverage and net coverage estimators.

The variance of estimated overcoverage $1 - \widehat{R}_{ce}$ equals the variance of the rate of correct enumeration $\widehat{R}_{ce}$. It depends on the multi-stage sampling design of the E-sample, the final weights $w_{e,i}$ and the measured $P_{ce,j}$ in the E-sample, see Chapter 3. Similarly, the variance of the estimated undercoverage $1 - \widehat{R}_m$ equals the variance of the rate of correct matches $\widehat{R}_m$. It depends on the multi-stage sampling design of the P-sample, the final weights $w_{p,j}$ and the measured $P_{m,j}$ in the P-sample, see Chapter 4.

The variance of the estimated net undercoverage $1 - \widehat{R}_{net}$ equals the variance of the rate of net coverage $\widehat{R}_{net}$. It depends on the sampling designs of the E-sample and the P-sample, the final weights $w_{e,i}$ and $w_{p,j}$ and the measured $P_{ce,j}$ in the E-sample and $P_{m,j}$ in the P-sample, see Chapter 5.

The variance of $\widehat{R}_{ce}$ and $\widehat{R}_m$ may be treated in the same way. Both estimators are weighted means of P_{ce} and P_m, respectively and are based on one unique sample; *i.e.* the E-sample and P-sample, respectively. The variance of $\widehat{R}_{net}$ has to be treated in a special way as it is a combination of weighted totals from two different non-independent samples (E-sample and P-sample).

Note that we do not explicitly include influences such as the nonresponse model for the P-sample in the variance estimation. Furthermore, we do not develop the methodology to test differences between rates in various subgroups of the population. Comparisons of confidence intervals do,however, offer some indications.

6.1 Over- and Undercoverage: Variance of $\widehat{R}_{ce}$ and $\widehat{R}_m$

Various methodologies may be used to estimate the variance of $\widehat{R}_{ce}$ and $\widehat{R}_m$. In this report, we use three techniques: the Taylor expansion (PROC SURVEYMEANS from SAS), the classical jackknife and a stratified jackknife (implemented in SAS).

General information about the variance estimation, the Taylor expansion and the jackknife technique can be found in Wolter (1985), Särndal et al. (1992) or Rao (1997). More detailed information about jackknife may also be found in Shao and Tu (1995).

The variance estimators presented below mainly depend only on the first stage of the design (PSUs) without correction for the finite population. This approximation is valid if the first-stage sampling fraction is small or if the first-stage is drawn with replacement. In our case, the sampling fraction is small in most strata but some of them are quite high. Some of the estimated variances are therefore possibly overestimated or unstable.

Jackknife is usually known as an all-purpose method. In our case, we can have confidence in the method as $\widehat{R}_{ce}$ and $\widehat{R}_m$ are weighted means, in other words, smooth functions of population totals.

Strict adherence to random replicates would dictate that the adjustment of the basic weights be computed separately within each replicate, with inclusion of the imputation, nonresponse adjustment and possible other correction such as calibration to auxiliary data (Wolter, 1985). In practice, only the final weights are considered below in the weight adjustment of replicates. Although underestimation of the variance is possible, this problem does not seem to be a serious one.

Estimates are calculated for the overall data set as well as for various sub-groups of the population. Results from the various techniques are compared in Appendix E.

The numerical results given with the coverage estimates in Chapters 11, 12 and 14 are results from the stratified jackknife.

6.1.1 Notation

Let $h = 1, .., H$ be the stratum number at the first stage (`stradap`), $i = 1, .., m_h$ the cluster number in stratum h, and $j = 1, .., n_{hi}$ the unit number in cluster i from stratum h.

We define $\widehat{R}$ to be the estimated rate of interest based on the sample s (weighted mean of P_j):

$$\widehat{R} = \frac{\sum_h \sum_i \sum_j w_{hij} \, P_j}{\sum_h \sum_i \sum_j w_{hij}} \tag{6.1}$$

where w_{hij} is the weight of the unit j in cluster i of stratum h and P_j is the status of correctness of the unit j.

In the case of overcoverage, $s = s_e$, $R = R_{ce}$, $w_j = w_{e,j}$ and $P_j = P_{ce,j}$. In the case of undercoverage $s = s_p$, $R = R_m$, $w_j = w_{p,j}$ and $P_j = P_{m,j}$. The statuses may be simple ones $P^{(s)}_{ce,j}$ and $P^{(s)}_{m,j}$ or some other statuses defined in Chapters 3 and 4.

The variance estimators are described below for the global estimator $\widehat{R}$. The estimation in a domain d is calculated in the same way but replacing w_j by $w_j \cdot I_{jd}$ and P_j by $P_j \cdot I_{jd}$, with I_{jd} the indicator variable: $I_{jd} = 1$ if the observation j is in the domain d and $I_{jd} = 0$ otherwise.

6.1.2 Taylor Expansion Method

The Taylor expansion method (linearization) estimator of variance $V(\widehat{R})$ is defined by (SAS, 2004):

$$v_L(\widehat{R}) = \sum_{h=1}^{H} \frac{m_h(1-f_h)}{m_h - 1} \sum_{i=1}^{m_h} (e_{hi} - \bar{e}_h)^2 \tag{6.2}$$

with

$$e_{hi} = \frac{\sum_{j=1}^{n_{hi}} w_{hij}(P_j - \widehat{R})}{\sum_h \sum_i \sum_j w_{hij}} \tag{6.3}$$

$$\bar{e}_h = \frac{\sum_{i=1}^{m_h} e_{hi}}{m_h} \tag{6.4}$$

The finite population correction $(1 - f_h) = (1 - m_h/M_h)$, with M_h the total number of clusters in h in the study population, is used in the calculation only if the option TOTAL is included in PROC SURVEYMEANS.

6.1.3 Classical Jackknife Method

Let $\theta = R$ be the parameter of interest and its estimator $\widehat{\theta} = \widehat{R}$.

For jackknife purposes, the sample s is partitioned into $m = \sum_h m_h$ subsamples corresponding to people in the PSU $\alpha = 1, .., m$. We define the $\alpha-$th group as the set of people in the PSU α.

Note that different $\alpha-$groups could be defined such as aggregation or part of the PSUs. However, PSUs form a natural partition and reflect the structure of the sample.

Let $\widehat{\theta}_{(\alpha)}$ be the estimator of the same functional form as $\widehat{\theta}$, but computed from the reduced sample obtained by omitting the $\alpha-$th group, $\alpha = 1, .., m$ (*replicates*):

$$\widehat{\theta}_{(\alpha)} = \widehat{R}_{(\alpha)} = \frac{\sum_{j \in s \setminus \alpha} w_j \, P_j}{\sum_{j \in s \setminus \alpha} w_j} = \frac{\sum_{j \in s} w_j \, P_j \, I_{j\alpha}}{\sum_{j \in s} w_j \, I_{j\alpha}} \tag{6.5}$$

where $s \setminus \alpha$ defines the sample s with omission of the PSU α. The second form makes use of the indicator $I_{j\alpha}$=1 if $j \in \alpha$ and 0 otherwise. We do not need any correction of the weights as $\widehat{R}_{(\alpha)}$ is a ratio (correction terms can be simplified)

We define

- the *pseudo-values* $\widehat{\theta}_\alpha = m \, \widehat{\theta} - (m-1) \, \widehat{\theta}_{(\alpha)}$ for $\alpha = 1, .., m$,
- the *Quenouille's or jackknife estimator* $\widehat{\widehat{\theta}} = \sum_\alpha \widehat{\theta}_\alpha / m$, and
- $\widehat{\theta}_{(.)} = \sum_\alpha \widehat{\theta}_{(\alpha)} / m$ the mean of the $\widehat{\theta}_{(\alpha)}$.

Two alternative variance estimators may be used:

$$v_{JK1}(\widehat{\widehat{\theta}}) = \frac{1}{m(m-1)}\sum_{\alpha=1}^{m}(\widehat{\theta}_\alpha - \widehat{\widehat{\theta}})^2 = \frac{m-1}{m}\sum_{\alpha=1}^{m}(\widehat{\theta}_{(\alpha)} - \widehat{\theta}_{(.)})^2 \tag{6.6}$$

$$v_{JK2}(\widehat{\widehat{\theta}}) = \frac{1}{m(m-1)}\sum_{\alpha=1}^{k}(\widehat{\theta}_\alpha - \widehat{\theta})^2 = v_1(\widehat{\widehat{\theta}}) + \frac{(\widehat{\theta} - \widehat{\widehat{\theta}})^2}{m-1} \geq v_1(\widehat{\widehat{\theta}}) \tag{6.7}$$

In practice, we often assume that $v_{JK1}(\widehat{\theta}) = v_{JK1}(\widehat{\widehat{\theta}})$ and $v_{JK2}(\widehat{\theta}) = v_{JK2}(\widehat{\widehat{\theta}})$.

Both values $v_{JK1}(\widehat{\widehat{\theta}})$ and $v_{JK2}(\widehat{\widehat{\theta}})$ are identical for many estimators (*e.g.* linear). In our calculation, the difference is negligible in the whole data set and in various subgroups ($< 0.01\%$). Therefore, only results with $v_{JK}(\widehat{R}) = v_{JK1}(\widehat{\widehat{\theta}})$ are presented below.

6.1.4 Stratified Jackknife Method

When the jackknife method is applied to a stratified sample $h = 1, ..H$, one will commonly use other variance estimators than classical jackknife. Actually, one should be careful when applying classical jackknife to a stratified sample.

We assume that the sample has been selected by using a stratified sampling design with strata $h = 1, .., H$. Accordingly to the classical jackknife, let $\widehat{\theta}_{(h\alpha)}$ be the estimator of the same functional form as $\widehat{\theta}$, but computed from the reduced sample obtained by omitting the $\alpha-$th group of stratum h, $\alpha = 1, .., m_h$, $h = 1, .., H$.

We have:

$$\widehat{\theta}_{(h\alpha)} = \widehat{R}_{(h\alpha)} = \frac{\sum_h \sum_{i\in s_{e,h}} \sum_j w'_{hij}\, P_j}{\sum_h \sum_{i\in s_{e,h}} \sum_j w'_{hij}} \tag{6.8}$$

with the corrected weights

$$w'_{hij} = \begin{cases} 0 & \text{if } i = \alpha \\ w_{hij}\, \frac{m_h}{m_h-1} & \text{if } \alpha \in h \text{ and } i \neq \alpha \\ w_{hij} & \text{otherwise, i.e. } \alpha \notin h \end{cases} \tag{6.9}$$

We define:

- the *pseudo-values* $\widehat{\theta}_{h\alpha} = m_h\, \widehat{\theta} - (m_h - 1)\, \widehat{\theta}_{(h\alpha)}$ for $\alpha = 1, .., m_h$ and $h = 1, .., H$,
- the *Quenouille's or jackknife estimator* $\widehat{\widehat{\theta}} = \sum_h \sum_{\alpha=1}^{m_h} \widehat{\theta}_{h\alpha}/m$, and
- $\widehat{\theta}_{(h.)} = \sum_{\alpha\in h} \widehat{\theta}_{(h\alpha)}/m_h$ the mean of the $\widehat{\theta}_{(h\alpha)}$ in stratum $h = 1, .., H$.

The variance estimators are similar to the classical jackknife but applied within each stratum. Here, we define only the first form :

$$v_{JKS}(\widehat{\widehat{\theta}}) = \sum_{h=1}^{H} \frac{m_h - 1}{m_h} \sum_{\alpha=1}^{m_h}(\widehat{\theta}_{(h\alpha)} - \widehat{\theta}_{(h.)})^2 \tag{6.10}$$

Variations of the estimator are applied in practice. We may, for instance, replace $\widehat{\theta}_{(h.)}$ by $\widehat{\theta}_{(.)} = \sum_h \sum_{\alpha \in h} \widehat{\theta}_{(h\alpha)}/m$. This change has little effect on the performance of the jackknife since both are second order asymptotically equivalent (Shao and Tu, 1995).

6.1.5 More About Jackknife

The factor $m_h/(m_h - 1)$ has been used in the correction of weights in Equation (6.9). An alternative correction could be the ratio between the sum of the weights w_{hij} in the stratum and the sum of the weights w_{hij} in the stratum without the PSU α.

The alternative correction would better reflect the design but the first version is recommended in our case to account for the variability in the estimated population total for the stratum (Kostanich, 2004). Actually, we did not apply ratio estimation to control final weights within the stratum to a fixed value. In the opposite case of a fixed total value $\sum_i \sum_j w_{hij}$ in each stratum, the alternative correction would be recommended.

Some splitting of PSUs could be processed in order to get more replicates to improve the jackknife estimation (stability) in strata with few PSUs. However, a splitting of PSUs would partially destroy the dependence among units within the same PSU.

Alternative correction and splitting were tested for R_{ce} and R_m overall and in some subgroups of the population. In most of the cases, the results are pretty close but larger differences are observed in subgroups that include data only or mainly from Ticino; see Appendix E. The reason is that this region has only 2 to 6 PSUs selected out of the 3 to 75 PSUs in the six strata. For domains like Ticino, where we have few degrees of freedom, any variance estimate based strictly on the design will suffer from a high variance of the variance.

Another direction, not broached in the current report, could be an estimation that includes the within-PSU variance and finite population correction. For example to deal with Ticino, we could use the estimated ratio total/within variance from Switzerland except Ticino to inflate the within variance (more stable than "between") in Ticino (Kostanich, 2004).

6.2 Net Coverage: Variance of $\widehat{R}_{net}$

The estimator $\widehat{R}_{net}$ is a non-linear combination of estimated totals from two dependent samples: the E-sample and the P-sample. Both samples have the same PSUs but different following sampling stages. Therefore, they cannot be considered as independent samples. However, the common first stage may be used in the estimation of variance.

The jackknife techniques are typically useful for the complex estimator $\widehat{R}_{net}$. The linearization technique (Taylor expansion) is not applied because of the complexity of the estimator.

Jackknife techniques have been shown to be valuable for dual system coverage estimations in the USA. The first stage estimator has also been shown to be a good estimator of variance. The reason is that variability between people is larger than variability between sampling units such as buildings.

As with the variance estimation applied for $\widehat{R}_{ce}$ and $\widehat{R}_{ce}$, we do not strictly adhere to random replicates. Only the final weights are considered in the weight adjustment of replicates.

The variance estimators are described below for the global estimator $\widehat{R}_{net}$. The estimation in a domain d has the same form but replacing C by C_d and C_ℓ by $C_{d\ell}$, see Equation (5.5).

6.2.1 Simple Jackknife

Let $\theta = R_{net}$ the parameter of interest and its estimator $\widehat{\theta} = \widehat{R}_{net}$.

For jackknife purposes, the P-sample and the E-sample are both partitioned into m subsamples corresponding to people in the PSU $\alpha = 1, .., m$. We define the $\alpha-$th group of the P-sample, resp. E-sample, as the set of people in the P-sample, resp. E-sample, and in the PSU α.

Let $\widehat{\theta}_{(\alpha)}$ be the estimator of the same functional form as $\widehat{\theta}$, but computed from the reduced sample obtained by omitting the $\alpha-$th group, $\alpha = 1, .., m$ (replicates):

$$\widehat{\theta}_{(\alpha)} = \widehat{R}_{net,(\alpha)} = \frac{C}{\widehat{N}_{(\alpha)}} = C \left[\sum_{\ell=1}^{L} C_\ell \frac{\widehat{R}_{ce,\ell(\alpha)}}{\widehat{R}_{m,\ell(\alpha)}} \right]^{-1} \tag{6.11}$$

where $\ell = 1, .., L$ is the estimation cell, see the Equations (5.2), (5.3) and (5.1).

The rates are defined by:

$$\widehat{R}_{ce,\ell(\alpha)} = \frac{\sum_{j \in s_e \setminus \alpha} w_{e,j} \, P_{ce,j}}{\sum_{j \in s_e \setminus \alpha} w_{e,j}} \tag{6.12}$$

$$\widehat{R}_{m,\ell(\alpha)} = \frac{\sum_{j \in s_p \setminus \alpha} w_{p,j} \, P_{m,j}}{\sum_{j \in s_p \setminus \alpha} w_{p,j}} \tag{6.13}$$

where $s_e \setminus \alpha$ and $s_p \setminus \alpha$ define the sample s_e, resp. s_p with omission of the PSU α.

Note that C and C_l are not random elements and therefore do not need to be modified in the replicates.

We do not need any reweighting to define $\widehat{\theta}_{(\alpha)}$ as both $\widehat{R}_{ce,\ell(\alpha)}$ and $\widehat{R}_{m,\ell(\alpha)}$ are ratios (correction terms can be simplified). Note however that a correction would be necessary for instance for a ratio between totals from two different samples.

The variance estimators have the same form as for R_{ce} and R_m, see Section 6.1.3. The first form is used for estimations.

6.2.2 Stratified Jackknife

For the stratified jackknife estimator, we define $\widehat{\theta}_{(h\alpha)}$ as follows:

$$\widehat{\theta}_{(h\alpha)} = \widehat{R}_{net,(h\alpha)} = \frac{C}{\widehat{N}_{(h\alpha)}} = C \left[\sum_{\ell=1}^{L} C_\ell \frac{\widehat{R}_{ce,\ell(h\alpha)}}{\widehat{R}_{m,\ell(h\alpha)}} \right]^{-1} \tag{6.14}$$

where

$$\widehat{R}_{ce,(h\alpha)} = \frac{\sum_h \sum_i \sum_j w'_{e,hij}\, P_j}{\sum_h \sum_i \sum_j w'_{e,hij}} \tag{6.15}$$

$$\widehat{R}_{m,(h\alpha)} = \frac{\sum_h \sum_i \sum_j w'_{p,hij}\, P_j}{\sum_h \sum_i \sum_j w'_{p,hij}} \tag{6.16}$$

The weights of the E-sample and the P-sample are corrected in the same way: $w'_{e,hij} = w_{e,hij}\Psi_{h\alpha}$ and $w'_{p,hij} = w_{p,hij}\Psi_{h\alpha}$ with:

$$\Psi_{h\alpha} = \begin{cases} 0 & \text{if } i = \alpha \\ \frac{m_h}{m_h - 1} & \text{if } \alpha \in h \text{ and } i \neq \alpha \qquad and \\ 1 & \text{otherwise, i.e. } \alpha \notin h \end{cases} \tag{6.17}$$

The variance formula in Section 6.1.4 are also valid for $\widehat{\theta}_{(h\alpha)} = \widehat{R}_{net,(h\alpha)}$.

6.2.3 Computer Implementation

In the jackknife variance estimator, we note that the only random term is $CCF_{\ell(\alpha)} = \widehat{R}_{ce,\ell(\alpha)}/\widehat{R}_{m,\ell(\alpha)}$ for the classical jackknife and $CCF_{\ell(h\alpha)} = \widehat{R}_{ce,\ell(h\alpha)}/\widehat{R}_{m,\ell(h\alpha)}$ for the stratified jackknife, respectively. These replicates, corresponding to a matrix $L \times m$ (L=number of estimation cells, m=number of PSUs), are then used to estimate the overall rate of net coverage and applied to various domains of interest. Unlike the rate of rate of correct enumeration and the rate of correct match, whose replicates need to be calculated for each domain, jackknife estimators require replicates to be calculated only once with synthetic assumption.

Part II

DATA and PRELIMINARIES

Chapter 7

Census Data

Census data sets are available at the inhabitant level, at the household level, at the housing unit level (dwelling), at the building level and at the commune level. All levels are linked by common identifiers; see the general information and some definitions in Appendix A.

Provisional and final census data sets are used in the project. In this section, all numerical values are related to the final data set of September 2003.

7.1 Inhabitant

The census data set at the inhabitant level has 7,452,075 entries.

A person is repeated as many times as his/her number of domiciles. One person with a civil domicile different from the economic domicile is therefore counted twice in the data set. The category of domicile (`WKAT`) allows for the extraction of data sets without double entries; see Table 7.1.

The *resident population* (economic domicile) is defined by `WKAT in (1,3)` as being a total of 7,288,010 and the *civil population* is defined by `WKAT in (1,4)` as being a total of 7,287,357.

Table 7.1: Category of domicile `WKAT`.

`WKAT`		Nb	[%]
1	one single domicile	7,123,292	95.6%
3	economic domicile (not unique)	164,718	2.2%
4	civil domicile (not unique)	164,065	2.2%
Total		7,452,075	100%

About 2.3% of the resident population is enumerated in two different domiciles (civil and economic). Note that some people are coded as `WKAT=3` but the link with the partner enumeration does not exist (`PARTNR=-7`, 565 records in the entire data set).

Various characteristics are collected in the personal questionnaire; see Appendix A. In the current project, we use the names and addresses, date of birth, gender, marital status, nationality (residence permit), position in the household (reference: economic domicile) and occupation; see Table 7.2 for the distribution between categories (some of them grouped) for the resident population.

Note that the information about the "language you think in and know best" was collected during the census and is often used in analysis of the Swiss population. However, the data collected during the SCS would need some more clerical work before it can be used. As this work has not been done, we do not use the results of this variable in the coverage estimation.

The information about imputation is gathered into `FLAG` variables, with `FLAG=0` being used for the original and unmodified value, `FLAG=1` for an imputation for a blank (missing) and `FLAG=2` for an imputation related to a change in the original value (incoherent, not valid). Flag variables are not available for all the variables in the census data set. In Table 7.3, we use the flag about the year of birth `GEJAF` for the age and the flag `KAMSF` made up by Daniel Kilchmann (METH) for the occupation.

7.2 Households

The census data set at the households level has 3,204,914 entries; 3,181,568 when restricted to the resident population.

The households are distributed in *private households*, *collective households* and *administrative households* and a more detailed typology was developed in each of these groups.

A private household is defined as a group of people who live in the same housing unit (*e.g.* a family). A collective household is defined as a "non-private" household (*e.g.* jails, hospitals or retirement homes). All other cases are administrative households (homeless, travelling people, collecting households).

Collecting households are not real households. They are created during the census data processing when people could not be linked to a real household. The household composition and the housing unit are not known. In some cases, the building is also not known; see Table 7.4.

Households are formed by 1 to 6831 people (economic population). Family households have 2-16 people. Non family households have 2-17 people. Collective households have 1-695 people. Administrative households have 1-6831 people.

7.3 Housing Units and Buildings

The census data set at the level of the housing units has 3,758,939 entries and the census data set at the level buildings has 1,465,891 entries.

The list of housing units contains units that are inhabited or not, with private or collective households, with or without kitchen/kitchenette. Fictitious entries are also listed to get the link for people not assigned to a real housing unit. Extracts may be processed to obtain, for instance, the dwelling units or other groups. Note that the housing units in the list cannot be entirely identified in the field (*e.g.* we may have two 2-room housing units that are both situated on the 3rd floor). The housing units receive 0-17 people with a mean value around 1.8 and a median of 2.

Only buildings that can be used for housing are listed in the census data set. As with collecting households, *collecting buildings* are created to accommodate people not linked to an enumerated building (maximum 1 per commune, total of 2797).

The buildings receive up to 262 housing units with a mean around 2.4 and the median 1. Most of the inhabited buildings have less than 5 dwellings (87%). The larger number of people in one building is 10,116. Among the real buildings, the larger number of people is 720, with mean around 5 and median at 3.

Information about imputation is also available. Note that some values were taken from the 1990 census data set (`FLAG=3`). This is for instance the case for 2.6% of the variable "type of the building" (`GART`: only habitation, mainly habitation, provisory or mobile habitation, mainly for other use, collecting building).

Table 7.2: Distribution of the resident population according to age group `age=2000-GEJAF` (`GEBJA`=year of birth), gender `GESL`, marital status `ZIV`, permit `AUSW`, position in the household `STHHW` and occupation `KAMS`. Definition of new codes `sex,Cage2,ziv2,ausw2,stell` and `taet` (see the selection of categories `Cage2` in Chapter 13).

Variable	Class	Code	Nb	[%]
`GESL`		`sex`		
Male	1	1	3,567,567	49.0
Female	2	2	3,720,443	51.1
`age`		`Cage2`		
	1-9	1	814,293	11.2
	10-19	2	851,320	11.7
	20-31	3	1,145,557	15.7
	32-44	4	1,561,377	21.4
	45-59	5	1,445,163	19.8
	60-79	6	1,171,413	16.1
	80-	7	298,887	4.1
`ZIV`		`ziv2`		
Single	1	1	3,064,734	42.1
Married	2	2	3,400,396	46.7
Widowed	3	3	414,945	5.7
Divorced	4	3	407,935	5.6
`AUSW`		`ausw2`		
Swiss	-9	1	5,792,461	79.5
C permit	1	2	1,032,056	14.2
B permit	2	3	339,321	4.7
Other permits	3-9	3	124,172	1.7
`STHHW`		`stell`		
Living alone	111	1	1,120,878	15.4
Husband/wife	112	2	3,132,892	43.0
Common-law husband/wife	113	3	387,938	5.3
Single parent	114	4	162,321	2.2
Other head	115	5	81,581	1.1
Relative of the head	121-124	6	2,021,373	27.7
Other	131-134	7	85,828	1.2
Collecting household	210	8	123,235	1.7
Collective household	241-243, 300	9	171,964	2.4
`KAMS`		`taet`		
In employment	1	1	3,789,416	52.0
Unemployed	2	2	157,572	2.2
No occupation	3	3	2,096,362	28.8
Less than 15 years old	4	4	1,244,660	17.1
Total			7,288,010	100

Table 7.3: Information about imputation flags FLAG (7,452,075 observations). Distribution among the values 0 (no imputation), 1 (imputation for blank) and 2 (imputation for incoherence) [%]. [1]: Results for the resident population without information about collective and collecting households.

Variable	Flag	0	1	2	Total
age	GEJAF	99.89	0.05	0.06	100
GESL	GESLF	99.96	0.02	0.02	100
ZIV	ZIVLF	99.76	0.10	0.14	100
AUSW	AUSWF	99.45	0.22	0.33	100
STHHW[1]	STHHWF	83.51	6.90	9.59	100
KAMS	KAMSF	94.36	5.12	0.52	100

Table 7.4: Distribution of economic households and economic population according to type of household HHTPW: absolute values and proportions [%]. Details for administrative households.

Description	Codes	Nb HH	[%]	Nb Pers	[%]
One person	1000	1,120,878	35.2	1,120,878	15.4
Family	2111-2422	1,931,860	60.7	5,733,917	78.7
Not family	3110-3222	62,661	2.0	138,016	1.9
Collective	9111-9224	8,148	0.3	166,384	2.3
Administrative	9801-9804	58,021	1.8	128,815	1.8
Total		3181568	100	7,288,010	100

Administrative	Codes	Nb HH	[%]	Nb Pers	[%]
People physically not present	9801	0	0	0	0
Homeless, travelling people	9802	1,174	2.0	5,505	4.3
No link with a building	9803	2,409	4.2	47,899	37.2
No link with household but real building	9804	54,438	93.8	75,411	58.5
Total		58,021	100	128,815	100

7.4 Communes

The 2896 communes are grouped into 217 *districts*, 26 *cantons* and 7 *NUTS* regions (Nomenclature of Units for Territorial Statistics).

The *population count* (`POP`) is the census count of the resident population in the commune. It takes values between 22 (5102 Corippo) et 363,273 (261 Zürich) with half of the resident population in communes with less then 7,246 inhabitants; see Table 7.5. Note that about 35% of the communes have fewer than 500 inhabitants and about 4% have more than 10,000 inhabitants. For the analysis, we define `taipop2=1` for communes with fewer than 2000 residents, `taipop2=2` for communes with 2000-7999 residents and `taipop2=3` for communes with 8000 or more residents (see the selection of the categories in Chapter 13).

The *official language* of a commune (`LING`) is one of the four national languages (German, French, Italian or Romansh) and is used as the administrative language between the commune and its inhabitants. It is determined as the national language with the larger number of inhabitants; see Table 7.6. For analysis purposes, we also define `ling2` as the aggregation of Romansh and German.

Table 7.5: Distribution of the communes N_c and the resident population `POP` into classes of population counts `POP`.

Class	N_c	[%]	`POP`	[%]
22 - 99	155	5.4%	9,896	0.1%
100 - 499	856	29.6%	243,818	3.4%
500 - 999	563	19.4%	407,909	5.6%
1000 -4999	1023	35.3%	2,322,434	31.9%
5000 - 9999	180	6.2%	1,241,997	17.0%
10000 - 49999	111	3.8%	1,878,008	25.8%
50000 - 99999	3	0.1%	222,605	3.1%
100,000 - 363,273	5	0.2%	961,343	13.2%
Total	2896	100%	7,288,010	100%

Various commune typologies have been constructed on the basis of the census 2000's socio-demographic, socio-economic, territorial and geographic data; see Schuler and Joye (2004). We use the *urban-rural status* (`urbrur`) that is determined by using variables such as employment, building density, population size and structure. For the analysis, we define `urbrur2`, which groups the 5 isolated towns with the town centers.

Some information about census operations is also available at the commune level. We use the *census methodology* (`var`), see Appendix A and the new defined variable `var2` that groups TRANSIT and FUTURE, as well as information about *outsourcing* (`outsour`). Outsourcing was used by many communes for mail management and other tasks, such as management of transfers (moving, etc.). The communes of Ticino canton receive the code `outsour=0` because of the special procedure in that canton.

Table 7.6: Distribution of communes N_c and the resident population `POP` into NUTS regions `NUTS`, official language `LING`, urban-rural status `urbrur`, census methodology `var` and outsourcing information `outsour`. Definition of new codes `ling2`, `urbrur2` and `var2`.

	Class	Code	N_c	[%]	`POP`	[%]
`NUTS`						
Lake Geneva region	1		589	20.3%	1,326,729	18.2%
Espace Mittelland	2		913	31.5%	1,679,417	23.0%
Northwestern Switzerland	3		321	11.1%	994,946	13.7%
Zurich	4		171	5.9%	1,247,906	17.1%
Eastern Switzerland	5		471	16.3%	1,048,467	14.4%
Central Switzerland	6		186	6.4%	683,699	9.4%
Ticino	7		245	8.5%	306,846	4.2%
`LING`		`ling2`				
German	1	1	1669	57.6%	5,221,135	71.6%
French	2	2	892	30.8%	1,720,365	23.6%
Italian	3	3	269	9.3%	320,247	4.4%
Romansh	4	1	66	2.3%	26,263	0.4%
`urbrur`		`urbrur2`				
Town center	1	1	64	2.2%	2,078,003	28.5%
Agglomeration	2	2	910	31.4%	3,204,312	44.0%
Isolated town	3	1	5	0.2%	63,137	0.9%
Rural	4	4	1917	66.2%	1,942,558	26.7%
`var`		`var2`				
CLASSIC	1	1	688	23.8%	268,826	3.7%
SEMI-CLASSIC	2	2	224	7.7%	177,353	2.4%
TRANSIT	3	3	1718	59.3%	6,473,966	88.8%
FUTURE	4	3	21	0.7%	61,019	0.8%
TICINO	5	5	245	8.5%	306,846	4.2%
`outsour`						
No delegation	0		957	33.1%	722,217	9.9%
Global packet	1		1607	55.5%	6,233,509	85.5%
Only mail	2		332	11.5%	332,284	4.6%
Total			2896	100%	7,288,010	100%

Chapter 8

P-sample and E-sample Data

The P-sample and E-sample of people are selected by using a multi-stage sampling design. Both have identical primary sampling units (PSUs) but different following stages.

8.1 P-sample

For the Swiss Coverage Survey (SCS), a special survey methodology was developed to obtain data that would be independent from the census and reliable for a match with the census. The data also include identification of moving or second domicile and classification into groups of interest. The survey requires intensive fieldwork, with a listing of households, including maps to identify buildings, and interviews; see Renaud (2002) and Renaud and Eichenberger (2002) for details.

The SCS multistage design leads first to a selection of 303 PSUs: 283 postal areas PAs in 22 strata in NORTH (=Switzerland except Ticino Canton) and 20 communes in 6 strata in TICINO (strata= `stradap`).

After a second stage related to field operations, we selected 15,877 buildings with 27,398 households. Information is collected about all household members during an interview conducted by phone (CATI), if the phone number was known and face-to-face (CAPI) otherwise; see the questionnaire in Appendix C. About 25,000 households were contacted. A set of 21,350 households were respondents and part of the population. We do not have any item non-response.

The list of households in the selected buildings was established from 18 January to 9 February 2001 for NORTH and from 28 February to 16 March 2001 for TICINO. The SCS-interviews took place from 17 April to 29 May 2001.

According to the information collected during the interviews and some checks during the matching process, we selected SCS people that belonged to the target population. The final P-sample of n =49,883 people is therefore a subset of the SCS data set; see Renaud and Potterat (2004) for details.

The P-sample weights are based on the sampling weights, with an adjustment for nonresponse at the household level. This adjustment takes into account the estimated proportion of non-respondents who belong to the target population. The reason is that many firms or vacation housing units were on the list of households. The final weights are quite variable (values between 5 and 489; CVs between 0.5 and 28% within the 28 `stradap`).

The distribution of P-sample entries into the categories for the most important analysis variables is shown in the undercoverage results tables in Chapter 12.

It is worth noting that some bias may appear in the distribution of the P-sample. For example, foreigners, particular those holding a B permit or less, are known to be generally underrepresented in households surveys. In the context of the coverage estimation, we did not apply any calibration and therefore bias may remain in the data. Actually, the proportion of Swiss people is 83.2% (weighted) in the P-sample and 79.5% in the census. However, the difference has to be put into perspective as the proportion based on the P-sample has some variability.

8.2 E-sample

The E-sample of n =55,469 people was selected in the census data set by using a two-stage design with a stratification at the first level (strata: `stradap`); see Renaud (2003) for the details.

After a special treatment of the census data to get a sampling frame of PSUs identical to the frame of the SCS, the 303 PSUs of the SCS were selected at the first stage (same selection probability). At the second stage, people were directly selected in the PSUs without going through buildings or households.

According to the information collected during the census, the E-sample contains only people that are part of the target population.

The E-sample sampling weights are almost constant within each stratum `stradap`. However, the variability within the subgroups may be quite high (*e.g.* CVs of about 60% in the language regions). We do not have nonresponse (no interviews, final weight = sampling weight).

A group of 21 people where excluded from the E-sample because they were absent from the final census data. The provisional E-sample used for the search for CE/EE therefore contains n =55,448 observations. For the estimations, the final E-sample contains n =55,375 observations because 73 people are no longer part of the target population in the final census data set.

Only census data are available for the E-sample people. We do not have any complementary field work to collect the information about the type of domicile, type of household and location on census day.

The distribution of the E-sample entries into the categories of the most important analysis variables can be found in the tables of Chapter 11.

Chapter 9

Geographical Location and Analysis Areas

Various data are available to help pinpoint the location of people in the census (and E-sample) and in the P-sample. Some new variables also need to be constructed for the estimation needs.

9.1 Location

Postal Areas

Postal areas (PAs) and communes are two different partitions of the Swiss territory. There are about 5000 PAs with 6 digits, aggregated in 3500 PAs with 4 digits, and 2896 communes.

In the census data set, information about the commune is more reliable than information about the PA.

The postal address of a given person or household is the address of the building they live in (street name, number in the street and PA)[1]. In many communes, the buildings do not have any street name or building number. Such buildings may be identified in most cases by using the coordinates or the insurance numbers and maps. The possible street names and numbers in the street are not used below. The location is defined by the PA (4 and/or 6 digits), the commune or the building ID.

Collected Addresses

One or two addresses were collected for each person during census data collection: (1) economic domicile on census day, (2) civil (not economic) domicile on census day. The first address is the reference for the estimation (population at its economic domicile).

The addresses 1 and 2 in the census data set are defined by a PA with 6 digits and by a commune. The commune is more reliable than the PA, especially in special cases such as collecting

[1]For example of location, Paul Meyer lives at the postal address "Rue Haute 11, 1450 La Sagne". The PA with 4 digits is 1450 (with the name: La Sagne) and the building is identified in the PA by the street "Rue Haute" and the number in the street "11". The PA with 6 digits "145002" is used for postal purposes and mail delivery. The commune is "Ste-Croix" (commune ID: 5568) which is in the district "Grandson", in the canton "Vaud" and in the NUTS region "Lake Geneva".

buildings.

Up to four addresses were collected during SCS for the P-sample people: (1) economic domicile on SCS day, (2) civil (not economic) domicile on SCS day, (3) economic domicile on census day, (4) civil (not economic) domicile on census day. The first address is the contact address during the SCS (contact at the civil domicile is excluded from the P-sample because it is out of the target population). The third address is the reference address for the estimation (economic population on census day).

In the P-sample the address 1 is defined by a postal area (PA) with 6 digits in NORTH (ex. 100004) and by a commune in TICINO (ex. 5042). Addresses 2-4 are defined by a PA with 4 digits (ex. 1000).

The P-sample of n =49,883 people are distributed into 48,303 (96.8%) non-movers and 1580 movers (3.2%). The *non-movers* are defined as people having the same economic address on SCS and census day (address 1 = address 3). The *movers* are defined as people having a different economic address on SCS and census day (address 1 $\neq$ address 3).

Geocoded Information

Geocoded information is available for the communes (GEOSTAT data from SFSO). We use the 3 dimensional *central coordinates* (state: 31 December 2000) and the list of *adjacent communes*. The central coordinates are used to geographically identify the most important place in each commune such as the market square or church. Coordinates are given in meters in the Swiss system of coordinates. The list of adjacent communes give the list of the pairs of communes that have a common frontier/boundary.

Others

The list of PAs (La Poste, status: 1 July 2003) with corresponding post offices as well as the electronic phone book (Twixtel) and the related maps (Twixroute) are used to find an address as well as to have an idea about the location of a PA or a commune.

9.2 Analysis Areas

For analysis purposes, we define two geographical areas around the P-sample and E-sample addresses: the *basic area* and the *extended area*. Both areas are defined as sets of communes.

Some addresses such as in the P-sample are defined by PAs without the information about the commune. Therefore, we need a link between PAs and communes.

9.2.1 Link between Postal Areas and Communes

We define the link between PAs and communes by two reference lists of couples (PA with 6 digits,commune) and (PA with 4 digits,commune) - below (PA6,co) and (PA4,co).

A couple (PA,co) is included in the list if at least one real building is included in both the PA and the commune co. In that case we say that the PA *touches* the commune co; and vice-versa, the commune touches the PA. The reference list of real buildings contains inhabited as well as uninhabited buildings (business, etc.) but no collecting buildings.

As an illustration, the PA A touches commune 1 and the PA B touches communes 1, 2 and 3 in Figure 9.1. Therefore, the couples (A,1), (B,1), (B,2) and (B,3) belong to the list (PA,co).

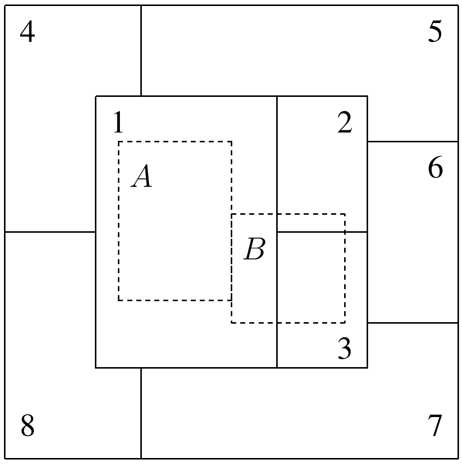

Figure 9.1: Illustration of the areas. Dashed boxes for postal areas (PAs A and B, 4 or 6 digits) and standard boxes for communes (1-8).

9.2.2 Main Commune

For each PA (4 or 6 digits) we also define the *main commune* as the commune touched by the PA with the more buildings. In Figure 9.1, the main commune of PA A is commune 1. If the PA B has 500 buildings distributed in 200 buildings in commune 1, 40 buildings in commune 2 and 260 buildings in commune 3, the main commune for B is commune 3.

9.2.3 Basic and Extended Areas

For P-sample non-movers in NORTH, the *basic area* is the set of communes being touched by the PA6 of the address 1 (address 1= address 3, 6 digits). For P-sample non-movers in TICINO, the *basic area* is the commune in the address 1 (= address 3). For P-sample movers in NORTH and TICINO, the *basic area* is the set of communes being touched by the PA4 of the address 3 (4 digits).

For balancing purposes, the basic area of each E-sample person is also defined by the set of the communes being touched by the PA6 of the building in NORTH and by the commune of the building in TICINO, respectively.

In all cases the *extended area* is defined by the set of communes adjacent to the basic area; including the basic area.

Some examples illustrated for the P-sample in Figure 9.1:

- a non-mover in NORTH with PA6 A on census day: the basic area is commune 1 and the extended area is the set of communes $\{1, 2, 3, 4, 5, 7, 8\}$;
- a non-mover in NORTH with PA6 B on census day: the basic area is the set of communes $\{1, 2, 3\}$, and the extended area is the set of communes $\{1, .., 8\}$;
- a non-mover in TICINO in commune 1 on census day: the basic area is commune 1, and the extended area is the set of communes $\{1, 2, 3, 4, 5, 7, 8\}$;
- a non-mover in TICINO in commune 2 on census day: the basic area is commune 2, and the extended area is the set of communes $\{1, 2, 3, 5, 6\}$.

9.3 Reference Commune

A *reference commune* is defined for each P-sample, E-sample and census person in order to determine the regional information such as the rural - urban status, the official language or the census collection methodology.

In the P-sample, the reference commune on census day is defined by the commune of the sampled building for non-movers (NORTH and TICINO) and by the main commune of the economic address on census day (PA with 4 digits) for movers. The reference commune of the 20 movers with missing PAs is set to the commune of the sampled building (assumption: move into a similar type of commune).

In the census and E-sample, the reference commune is the commune of economic domicile.

Chapter 10

Searches for Matches and Correct/Erroneous Enumerations

The coverage estimation involves two important processes before going into estimations: the search for matches between the P-sample and the census data set and the search for correct enumerations in the E-sample. These processes have to be of really good quality in order to avoid a bias in the coverage estimation; see Chapter 1.

Both searches make use of matching procedures: between the P-sample and the census, and between the E-sample and the census. The methodology is related to a record linkage (exact and probabilistic) and not to a statistical matching. The idea is not to find a similar person but the same person.

Both searches are applied with the entire census data set. We do not restrict it to any subpopulation such as resident population, private households or some search areas. At the moment of matching, all people are available in the census data set but some characteristics are not definitive (*e.g.* household type, domicile type, building, etc.).

We assume that all census data are eligible for matching. We do not need any special procedure for entries that would not be data-defined (such as having no names or a name but few non imputed data among the main demographic characteristics). No special treatment either for imputed characteristics.

The quality of the searches has not been checked. We do not have any information about the performance of the procedures such as false matches or missing matches and false correct enumeration or missing correct enumeration, respectively. The results of the search for matches and the search for correct enumerations are assumed to be accurate and are therefore not changed.

10.1 Search for Matches

The search for matches aims at determining which P-sample people are counted and which are not counted in the census data set.

The preliminary P-sample of 50,070 observations is used for matching; see Renaud and Potterat (2004).

If a P-sample person is matched somewhere in the census, we call it a *P-sample matched entry*. The corresponding entry in the census is called *a match* (regardless of the type of domicile, type of household and location). If a match is found in a place far from the address on census day or not in the target population, we look for another possible match in the correct area and in the right population. If we succeed, the new match replaces the old one. If we fail, the older match is kept.

A *P-sample non-matched entry* is a person who could not be found in the entire census data set.

The search is organized in various steps using computerized matching, computer-assisted clerical matching and non-assisted clerical matching.

Computer-assisted clerical matching and non-assisted clerical matching are checked clerically. A sample of the computerized matching is also checked clerically; but not all cases.

We do not have any follow-up of non-matched P-sample people to determine whether they are eligible for the census.

10.1.1 Matching Process

A step-by-step procedure is applied to the data[1]. If a unique match is found for person i at step a, then the search is no longer active for this person in step $a+1$. If ≥ 2 matches are found for person i at step a, and if clerical checks cannot determine the correct entry, the case is sent to step $a+1$. If the entry is not matched at step a, then search for a match at step $a+1$ begins.

Before any matching, the first names and surnames are standardized in order to avoid problems due to typing or scanning errors: capital letters, special characters deleted (*e.g.* /;:,*), modifications (*e.g.* É -> E, Ö -> OE). Note also that matching of names uses a phonetic procedure (*e.g.* John-Paul compared with John, Paul, John Paul and Paul-John, or Maria with Marie, Marie-Jeanne, Mary).

Steps in the ORACLE census database with SQL procedures:

1. computerized matching: first name, surname, date of birth;
2. computer-assisted: building ID code and date of birth;
3. computer-assisted: list of household members (if one member matched);
4. computer-assisted: commune, first-name and surname;
5. computer-assisted: commune, surname and date of birth.

Steps with SAS macros for non-movers:

1. computer-assisted: surname and year of birth;
2. computer-assisted: surname and date of birth;
3. clerical: building ID code, sex and marital status.

[1]The operations are processed by the census staff and only shortly described here.

Complementary steps for non-movers:

1. For matches that are not in the target population (collective household or civil domicile) or not in the PSU (postal area or commune): search for a better match by using the PSU, the surname and the year of birth and, for remaining cases: the PSU, the surname and the date of birth;
2. Clerical search for remaining non-matches: date of birth, building ID.

Step with SAS macros for movers (computer-assisted): commune, year of birth and surname.

Final process: in some cases, two people from the P-sample were matched to one person in the census. These cases were checked clerically in order to detect double entries in the P-sample, as well as eligible and non eligible matches.

The result of the matching operations with the 50,070 P-sample people is summarized in a list of 57 codes `match`, see Tables D.1 and D.2 in Appendix D. The codes can be aggregated into 4 groups: confirmed matched (49,238, 98.0%), confirmed non-matched (807, 1.6%), unresolved cases (217, 0.4%) and cases that have to be excluded from the P-sample (*e.g.* doubles, 4).

10.1.2 Final Matching Codes

New steps are applied to get the final matching codes. These steps include the information from special cases and supplementary checks that are summarized in the complementary codes `match_cont` and `match_cont2`; see Appendix D and the Table D.3.

After excluding 184 entries corresponding to people not in the P-sample (`matchG=0`) and detecting 3 people living abroad on census day, the final P-sample is a data set of 49,883 people.

The census data available for matching receive some adjustments before the final version. As a result, 75 matches are no longer available in the final census data set and the final result is: 49,107 matched entries (`matchG=10`) and 776 non-matched entries (`matchG=20`).

The multiple matches (one P-sample people matched with 2 or more census people) were rechecked at the end of the process in order to select potential matches (*e.g.* eligible as a replacement). The criteria do not include the type of domicile, the type of household and the location. The list of potential matches contains 54 entries.

10.1.3 Remarks

We do not use the status of "possible match". Therefore, the status of match is 0 (non-matched) or 1 (matched) for all P-sample units.

The matching process stops for unit i of the P-sample as soon as one unique match is found. In a complementary step for non-movers, the case is treated again if the match is not in the correct area or not in the target population. However, checks have shown that the definition of the area was not defined in the right way during the process. And the first match was not kept if the second match was considered better. Therefore we do not have information about all the potential matches and may have a small bias in the estimation; especially when comparing the location of the P-sample people and the matches in the census. This effect is however expected to be negligible.

Overall, the matching results for the undercoverage estimation could probably be improved by considering all the possible matches in the process and therefore creating the list of matches and a complementary list of potential matches. For example, the second domicile of a match could be checked to be possibly included in the list of potential matches.

Note also that the matching process is not documented in a detailed and systematic way. This shortcoming, especially in some of the phases such as SQL, means that we do not know the exact and effective steps of the process.

A measure of the quality of the matching process would also be of great interest.

10.2 Search for Correct/Erroneous Enumerations

The search for correct and erroneous enumerations (CE and EE) is mostly restricted to a search - based on the E-sample - for double entries in the census. Some few special cases have been determined to be fictitious enumerations.

The provisional E-sample of 55,469 people is used to match with the complete census data set.

During the search, if an E-sample person is matched with another entry in the census, we call it an *E-sample double*. The corresponding entry in the census is called a *doublet*. Similarly, we have an *E-sample triple* with the two corresponding *triplets*.

Partner enumerations of E-sample people with two domiciles are excluded from the search. If the *partner enumeration* of an E-sample entry is found during the search, we don't keep it as a doublet.

The data collected during the census are the only source of information in this process. We do not have any field operation or interviews of E-sample people to get a new source of information about these people.

The search is organized in various steps using computerized matching, computer-assisted clerical matching and non-assisted clerical matching.

10.2.1 Process

The searching process may be summarized in 3 steps[2].

First, E-sample people matched with P-sample people are coded as Correct Enumeration (CE).

Second, people not matched with P-sample people are processed in order to detect doublets in the census data set. The matching process makes use of the first-name, surname, date of birth, marital status, position in the household, etc. The type of domicile, type of household, and location are not matching variables.

Third, people not coded during the 1st and 2nd phases are checked in order to select people that may correspond to fictitious enumerations (*e.g.* names containing special symbols, people

[2]The operations are processed by the census staff and only shortly described here.

born before 1895). A group of unclear 120 entries were further checked by using the census database, the images made of census questionnaires and the phone book.

Finally, people not involved in phases 1-3 are considered as CE.

The result of the search is a set of lists that may be summarized in 7 codes `codeEE` grouped as follows:

- 6694 match with the P-sample `codeEE=11` (12.1%);
- 168 confirmed doubles `codeEE=15` (0.3%) and 1 possible double `codeEE=31`;
- 1 confirmed fictitious `codeEE=20` and 1 possible fictitious `codeEE=32` ;
- 48584 without any information `codeEE=-99` (87.6%) and 20 out of the census data set `codeEE=25`.

One entry received the code `codeEE=31` as a possible double. The first name and surname, date of birth, marital status and nationality are identical but the education and profession are different.

The confirmed fictitious enumeration is an entry with preprinted information in the database, although clear crossing out of the questionnaire on the scanned image ("Do not live in Switzerland"). The possible fictitious enumeration has no name and surname and is also crossed out with annotations (not readable).

10.2.2 Final E-sample Coding

The codes are completed in order to get the final information about the search for CE and EE:

1. Integration of checks about people with two domiciles at the time of the search but only one in the final census data set (63 cases);
2. Integration of a set of double/triple entries found during a new SQL search in the census data base (355 cases);
3. Exclusion from the E-sample of 21 people out of the final census data set and 73 people in collective households;
4. Integration of the final matches with the P-sample (6697 cases);
5. Exclusion of doublets corresponding to partner enumerations (74 cases).

The final correct enumeration status `codeEEA2` of the 55,375 E-sample people has the following distribution: 48,228 without information `codeEEA2=-9` (87.1%), 2 fictitious enumerations `codeEEA2=1`, 440 doubles `codeEEA2=2` (0.8%), 8 triples `codeEEA2=3` and 6697 matches with the P-sample `codeEEA2=4` (12.1%).

For the analysis, the possible double is combined with the confirmed double and the possible fictitious enumeration is combined with the confirmed fictitious one. The number of "possible" statuses is actually very small.

If we consider the people without information and matches as correct enumerations, a total of 54,925 out of 55,375 are considered as correct (99.2%).

10.2.3 Remarks

The search for multiple entries did not apply for matches with the P-sample. This simplification is a potential for bias in the search as these entries could also be a double or a triple in the census data set.

Most of the resources were allocated to the search for double and triple entries. The search for other erroneous entries such as fictitious or people not in the right population and location is more difficult without auxiliary information such as interviews.

The most important for the DSE estimation is probably the lack of information about the real location, real type of domicile and real type of household of the E-sample people on census day; see the discussion about balancing in Chapter 5. Data from interviews, at least from a sample of the E-sample, would be recommended for a future coverage estimation.

Part III

RESULTS

Chapter 11

Overcoverage

Based on the methodology described in Chapter 3, this chapter presents the results obtained from analysis of correct and erroneous enumerations and corresponding overcoverage in the census data set.

Chapters 7, 8 and 9 describe the census data, the basic and extended areas as well as the E-sample, which is used as the basis for all estimations in this chapter. Chapter 10 contains information about the search for correct and erroneous enumerations.

11.1 Checks before Estimation

Before going into numerical estimation, we note that none of the E-sample units has an extremely large influence on the point and variance estimates. Actually, the E-sample weights have a very low variability within each sampling strata `stradap`. Therefore, the E-sample and the weights defined in Renaud (2003) serve, without any changes, as the basis for estimation.

11.2 First Look at the Rate of Correct Enumeration $\widehat{R}_{ce}$

The simplest estimated rate of correct enumeration in the census $\widehat{R}_{ce}^{(s)}$ is based on the simple status of correct enumeration $P_{ce,j}^{(s)}$, see Section 3.2.1:

$$\widehat{R}_{ce}^{(s)} = \frac{\sum_{j \in s_e} w_{e,j} \, P_{ce,j}^{(s)}}{\sum_{j \in s_e} w_{e,j}} \tag{11.1}$$

where $w_{e,j}$ is the weight of people j in E-sample s_e.

The status of correct enumeration $P_{ce,j}^{(s)}$ is mostly equal to 1 (>99%) with 0.8% of the cases equal to 0.5 (double entries), 8 triple entries and only two fictitious enumerations; see Table 11.1.

With $P_{ce,j}^{(s)}$, we have $\widehat{R}_{ce}^{(s)} = 99.60\%$, with a standard error s.e. = 0.03%. Therefore, 0.4% of the census count are erroneous. This is the estimated overall overcoverage.

Table 11.1: Status of simple correct enumeration $P^{(s)}_{ce,j}$ in the E-sample, with n the number of people and $Pwei$ the weighted proportion.

$P^{(s)}_{ce,j}$	n	$Pwei$
0	2	0%
1/3	8	0%
1/2	440	0.8%
1	54,925	99.2%
Total	55,375	100%

Among the various rates defined in Chapter 3, $\widehat{R}^{(s)}_{ce}$ can be considered as a lower limit because all multiple entries j have a status $0 < P^{(s)}_{ce,j} < 1$, *i.e.* partially erroneous, none of them is fully correct.

Comparison between double entries

A comparison between the demographic characteristics of the 440 couples (double, doublet) shows that few values are different for the variables permit (`ausw`, 6 over 423 non-imputed values), marital status (`ziv`, 13 over 423 non-imputed values) and age (0 up to 4 years). Larger differences are observed for position in household (`STHHW`, 59 over 104 non-imputed values) and occupation (`KAMS`, 66 over 434 non-imputed values; with 57 shifts between "in employment" and "no occupation"). The size of household (economic definition) varies considerably between the two entries. The difference in size can be as high as 6800 or for private households as high as 7.

11.3 Alternative Rates of Correct Enumeration

The analysis of alternative rates enables detection of particular behaviors of multiple entries.

Membership in the Target Population

If we consider the doublets and triplets as real multiple entries only when they belong to the target population, 391 doubles and 2 triples remain multiple entries:

- 389 out of 440 doublets are in the target population and 51 are not in the population (5 civil domiciles and 46 collective households).
- 4 out of 8 triples have both triplets that are not in the population (civil domicile or collective household), 2 have one triplet that is not in the population (civil domicile) and 2 have both triplets in the population.

The total absolute number of EE enumerations decreases from $\sum_{j\in s_e}(1 - P^{(s)}_{ce,j}) = 2 + (8 * 2/3) + (440/2) = 227.33$ to $\sum_{j\in s_e}(1 - P^{(pop)}_{ce,j}) = 2 + (2 * 2/3) + (391/2) = 198.83$.

The rate of correct enumeration based on membership of multiple entries in the target population is $\widehat{R}_{ce}^{(pop)} = 99.65\%$ (s.e. = 0.03). Therefore, the rate of overcoverage of the target population 0.35% is slightly slower than $1 - \widehat{R}_{ce}^{(s)} = 0.4\%$.

Relaxing the Criterion of Membership in the Target Population

The criterion of membership in the target population may be eased (or relaxed) to include multiple entries only if the doublet/triplet or its partner are in the target population. The idea is that partners would also be potential records for multiple entries. The corresponding status of correct enumeration is $P_{ce,j}^{(popR)}$.

The relaxation of membership in the population has no effect on the status of correct enumeration in the E-sample; neither for doubles nor for triples:

- 12 out of 440 doublets have two domiciles and therefore a partner. However, 7 partners are not defined (`PARTNR=-7`) and can therefore not be considered in the easing adjustment. The resulting 5 partners are members of collective households (out of the target population); see Figure 11.1.
- 6 out of 8 triples have both triplets with a partner. However, they are all coded as partners of each other in the census (link between both triplets). The relaxation of this criterion, represented by the information for the partners, is therefore already "included" in the data.

As a result, $P_{ce,j}^{(popR)} = P_{ce,j}^{(pop)}$ and $\widehat{R}_{ce}^{(popR)} = \widehat{R}_{ce}^{(pop)} = 99.65\%$

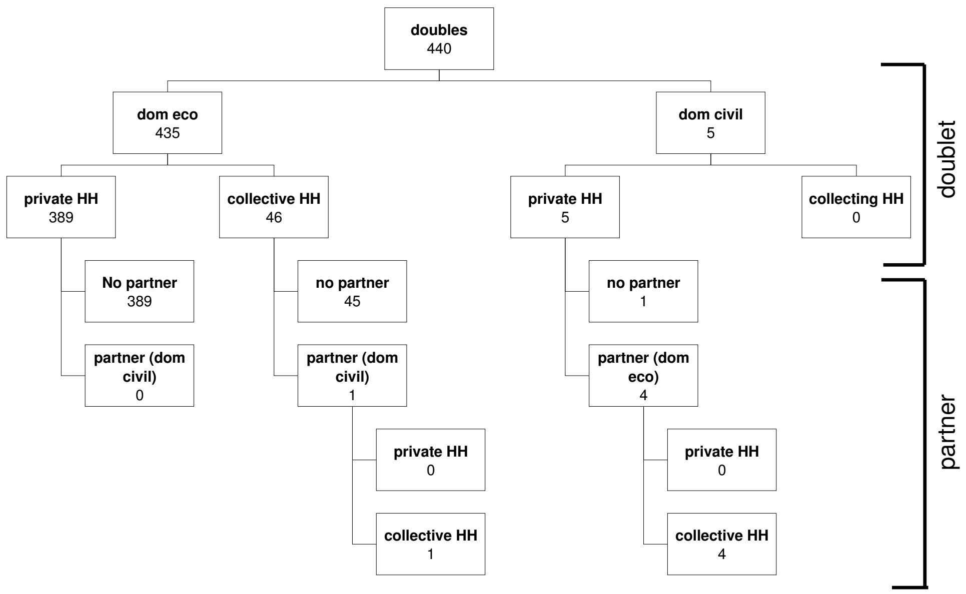

Figure 11.1: Breakdown of membership in the target population for doublets and their partners. *N.B.*: target population = economic domicile and private household.

It is remarkable that most of the partners of the doublets entries are not defined in the census data set (7 out of 12). One could ask if there is a missing link between the double and the doublet in the census data set. Actually, missing links are probable for 3 couples (double, doublet) since both have undefined partners. The remaining 4 doubles do not have two domiciles in the census data but such a possibility was possibly overlooked in the data processing stage.

Location

The doublets and triplets are found around the corresponding doubles and triples (basic or extended area) or farther (out of the extended area).

If a doublet or triplet is considered as a real multiple only if located in the basic or extended area, the number of doubles decreases to 258 and the number of triples decreases to 2:

- 218 out of 440 doublets are the basis area, 36 in the extended area and 186 out of the extended area (107 out of the canton).
- 4 out of 8 triples have one triplet in the basic area and the second triplet out of the extended, 2 have one triplet in the basic and one triplet in the extended, and 2 have 2 triplets out of the extended areas (out of the canton).

A group of 135 doublets are not only in the same basic area (218 doublets) but also in the same building as the E-sample entry. Therefore, a non negligible amount of multiple entries would be removed by a special process in the census data treatment (at the building level).

We note that the distance between the double and the doublet out of the extended area ranges between 2.3 km and 276 km, with an average of 59.5 km (distance between the central coordinates of the communes). This distance ranges from between 16 and 210 km for the triplets out of the extended area. Further study could inform about possible moves or lack of links between two addresses.

The total absolute number of EE enumerations decrease from $\sum_{j \in s_e}(1 - P_{ce,j}^{(s)}) = 227.33$ to $\sum_{j \in s_e}(1 - P_{ce,j}^{(loc)}) = 132.33$. The effect of location is larger than the effect of population.

The rate of correct enumeration that includes the information about location of the multiple entries is $\widehat{R}_{ce}^{(loc)} = 99.77\%$ (s.e. = 0.03). Therefore, the rate of overcoverage around the location is estimated to be small 0.23%.

Combination of Population and Location

If a doublet or triplet is considered as a real multiple only if it belongs to the population and is located in the basic or extended area, the number of doubles decreases to 229 and the number of triples to 0:

- 389 out of 440 doublets are in the population, with 199 in the basis area, 29 in the extended area, 161 out of the extended.
- 2 out of 8 triples have both triplets in the population but out of the extended area, 4 triples with both triplets out of the population, one triple with only one triplet in the population but out of the extended area and one triple with only one triplet in the population and in the basic area

The total absolute number of EE enumerations is only $\sum_{j\in s_e}(1-P_{ce,j}^{(poploc)}) = 116.5$ and the rate of correct enumeration is $\widehat{R}_{ce}^{(poploc)} = 99.80\%$ (s.e. = 0.03). Therefore, the rate of overcoverage in the population and around the location is estimated to be only 0.20%.

11.4 Results for some Domains

The rate of correct enumeration is high in the census data set but some variations are observed between sub-groups of the population; see Table 11.2 for the simple rate of correct enumeration $\widehat{R}_{ce}^{(s)}$ and the rate that depends on membership in the population $\widehat{R}_{ce}^{(popR)} = \widehat{R}_{ce}^{(pop)}$. The codes are defined in Chapter 7. Only data with a low potential misclassification error are presented here.

All the estimated rates of correct enumeration are larger than 99%. The lower rate is observed for the age group `Cage=3` (20 - 31 years old): 99% correct enumeration, *i.e.* 1% overcoverage. Young people are more mobile and may therefore be enumerated at two places without any link between the two addresses. We do not observe significant differences between categories for the other variables.

All estimated standard errors are smaller than 0.18%, which means that the rates are estimated with a very low coefficient of variation (note that the CV$\approx$ s.e. as $R_{ce} \approx 1$).

11.5 More about Overcoverage

We observe an overall overcoverage of 0.35 - 0.4% when considering the simple rate and the rate that depends on membership in the target population. These values are in the range of the estimates in other countries; see Table 1, on page 8.

Analysis of multiple entries allows detection of a non negligible number of doublets enumerated in the same building. This is a possible track for improvement in census data processing. Further improvement may also occur in the difficult task of linking people with two addresses.

The estimated low rates of overcoverage are possibly related to weaknesses of the search for correct and erroneous enumerations in the E-sample. The values may therefore be seen as minimum values. We do not have any information about the real situation of the E-sample people. Interviews, combined with an improved search for double entries, should be considered for future estimations.

Further analysis could be carried out to detect the most discriminant variables for overcoverage (logistic models etc.) and develop the results in other domains such as age groups by sex. At this point we stop the analysis of overcoverage because there were only a small number of erroneous enumerations in the census.

Table 11.2: Rates of CE for different domains. Number of elements n, number of erroneous enumerations $EE = \sum_{j \in s_e}(1-P_{ce,j})$, rates of correct enumeration $\widehat{R}_{ce}^{(s)}$ and $\widehat{R}_{ce}^{(popR)} = \widehat{R}_{ce}^{(pop)}$ [%] with the standard error s.e. [%].

Variable			n	EE	$\widehat{R}_{ce}^{(s)}$	s.e.	EE	$\widehat{R}_{ce}^{(pop)}$	s.e.
Overall			55375	227.3	99.60	0.03	198.8	99.65	0.03
`sex`	Male	1	27374	116.8	99.59	0.04	104.8	99.63	0.04
	Female	2	28001	110.5	99.62	0.03	94.0	99.67	0.03
`Cage2`	1-9	1	6449	16.7	99.74	0.05	16.5	99.74	0.05
	10-19	2	6689	23.2	99.66	0.06	19.0	99.73	0.05
	20-31	3	8652	85.5	99.03	0.09	82.2	99.07	0.09
	32-44	4	12090	47.7	99.65	0.05	44.7	99.67	0.05
	45-59	5	10902	29.0	99.74	0.04	25.0	99.78	0.04
	60-79	6	8802	12.5	99.88	0.03	9.5	99.90	0.03
	80+	7	1791	12.8	99.22	0.18	2.0	99.89	0.06
`ausw2`	Swiss	1	45550	177.3	99.61	0.03	153.8	99.67	0.03
	C permit	2	6851	28.0	99.63	0.06	25.0	99.67	0.06
	Other	3	2974	22.0	99.39	0.11	20.0	99.44	0.11
`ziv`	Single	1	23515	132.8	99.44	0.05	117.2	99.50	0.05
	Married	2	26040	71.7	99.76	0.04	69.2	99.77	0.04
	Widowed	3	2879	12.3	99.45	0.12	5.0	99.75	0.08
	Divorced	4	2941	10.5	99.65	0.09	7.5	99.76	0.08
`ling2`	German + R	1	36706	143.5	99.61	0.04	123.0	99.67	0.04
	French	2	16473	71.2	99.61	0.05	63.7	99.65	0.06
	Italian	3	2196	12.7	99.44	0.13	12.2	99.47	0.12
`NUTS`	Lake GE	1	10901	46.8	99.59	0.06	43.2	99.63	0.07
	Espace M.	2	16039	66.7	99.59	0.09	57.0	99.65	0.09
	Northwest	3	6592	20.5	99.73	0.04	16.0	99.82	0.04
	Zurich	4	8813	31.5	99.65	0.05	27.0	99.69	0.05
	East	5	7856	35.7	99.55	0.07	31.5	99.60	0.07
	Central	6	3478	15.5	99.59	0.08	14.0	99.64	0.06
	Ticino	7	1696	10.7	99.44	0.13	10.2	99.46	0.12
`taipop2`	Small	1	18668	79.0	99.63	0.06	71.3	99.66	0.05
	Middle	2	17013	75.2	99.54	0.07	67.0	99.59	0.07
	Large	3	19694	73.2	99.63	0.03	60.5	99.69	0.03
`urbrur2`	Town	1	12882	55.7	99.58	0.04	47.0	99.65	0.04
	Agglo	2	20733	85.0	99.60	0.06	76.7	99.64	0.06
	Rural	4	21760	86.7	99.63	0.04	75.2	99.68	0.04
`var2`	CLASSIC	1	11000	53.2	99.58	0.06	49.0	99.61	0.05
	SEMI-CLA	2	5298	20.2	99.60	0.08	18.7	99.63	0.08
	TRAN+FUT	3	37381	143.3	99.61	0.03	121.0	99.67	0.03
	TICINO	5	1696	10.7	99.44	0.13	10.2	99.46	0.12
`outsour`	No del	0	13548	68.3	99.49	0.07	63.7	99.51	0.07
	Global	1	35599	133.8	99.63	0.03	111.5	99.68	0.03
	Only mail	2	6228	25.2	99.45	0.17	23.7	99.46	0.17

Chapter 12

Undercoverage

Based on the methodology described in Chapter 4, this chapter presents the results obtained from analysis of the matches and related undercoverage in the census data set. Emphasis is placed on membership in the population and location in order to detect possible improvements in the census data processing such as time delay or difficulty in determining the type of domicile. Analysis is also enhanced by a comparison between characteristics collected during both the census and SCS. As with analysis of overcoverage, we present a choice of results and suggest ways to go further with the analysis.

Some P-sample people moved between the census day and the SCS day. Mobility is a known cause of coverage errors, making it more difficult to gather census data (*e.g.* questionnaire sent to the wrong address, etc.). Therefore, we often present the results for movers and non-movers alongside overall results.

Chapters 7, 8 and 9 describe census data, the basic and extended areas as well as the P-sample, which is used as the basis for all estimations in this chapter. Chapter 10 contains information about the search for matches.

12.1 Checks before Estimation

Various checks were applied to the P-sample in order to detect extremely influential elements. The weights are quite variable, especially in some sampling strata. For this reason, contrary to the E-sample, influential elements are likely to skew results based on the P-sample.

In our case, the weights $w_{p,j}$ vary between strata and clusters (PSUs) but also within the clusters. Variability within the strata is the most influential parameter for variance estimation. As status $P_{m,j}$ takes values 0 or 1, outliers are checked for $w_{p,j}$ and not $w_{p,j} \cdot P_{m,j}$.

Trimmed weights $w_{p,j}^{(t)}$ were defined with $w_{p,j}^{(t)} \neq w_{p,j}$ for 89 P-sample elements; see Appendix E. This trimming has a negligible impact on $\widehat{R}_m$ and its variance. More study could be devoted to the influential weights and trimming but this was not done in the current project. Therefore, the results below are based on the original weights $w_{p,j}$.

12.2 First Look at the Rate of Match $\widehat{R}_m$

The simplest estimated rate of correct match $\widehat{R}_m^{(s)}$ is based on the simple status of correct match $P_{m,j}^{(s)}$, see Section 4.3.1:

$$\widehat{R}_m^{(s)} = \frac{\sum_{j \in s_p} w_{p,j} \, P_{m,j}^{(s)}}{\sum_{j \in s_p} w_{p,j}} \tag{12.1}$$

where $w_{p,j}$ is the weight of people j in P-sample s_p.

The status of correct enumeration $P_{m,j}^{(s)}$ is equal to 1 for 49,107 elements and 0 for 776 elements; see Table 12.1.

We have an overall rate of correct match $\widehat{R}_m^{(s)} = 98.36\%$ with s.e. = 0.11%. Therefore, 1.64% of the population targeted during the SCS should have been, but were not, enumerated in the census (undercount). The undercount 4.5% (s.e. = 0.66) for movers is clearly larger than 1.5% (s.e. = 0.10) for non-movers.

Table 12.1: Simple matches, total and depending on moving status, with n number of people, and $Pwei$ weighted proportion $\widehat{R}_m$ and $1 - \widehat{R}_m$, respectively [%]. Standard error of $\widehat{R}_m$ in brackets [%].

	Overall		Non-movers		Movers	
$P_{m,j}^{(s)}$	n	$Pwei$	n	$Pwei$	n	$Pwei$
0 (no match)	776	1.64	708	1.54	68	4.52
1 (match)	49107	98.36 (0.11)	47595	98.46 (0.10)	1512	95.48 (0.66)
Total	49883	100	48303	100	1580	100%

12.3 Classification and Misclassification

In Chapter 4, we addressed the status of correct match in a particular domain such as correct sex or age group and the misclassification error of P-sample or census entries.

In this section, we first compare the characteristics of data collected during SCS and census. This comparison takes imputation flags into account. The second part gives some indication of possible influence of misclassification on coverage estimations.

When there is a difference between SCS and census data, we cannot initially determine which result is the right one. In some cases, a special approach such as looking at the image of the census questionnaire may help determine the right one. However, this is not possible in all cases.

Note that the date of birth is used as a matching variable during automatic steps. All other variables are used only during clerical checks.

12.3.1 Comparison of Data

The comparison of data collected during SCS and census is processed for the following variables; see Table 7.2 on page 54 for the definitions: gender `sex` (2 categories), age groups `Cage2` (7 categories), marital status `ziv2` (3 categories), permit `ausw2` (3 categories), position in the household `stell` (9 categories), occupation `taet` (4 categories) and size of household (relative to the economic domicile) `nbHH`.

Preliminary results

Among the 49,107 people for whom information from SCS and census was available, 47,213 people (96.0% weighted) show no difference in the variables `sex`, `Cage2`, `ziv2` and `ausw2`. Among the 48,707 non-imputed entries in the same four variables, 46,886 show no difference (96.1% weighted). We have 1754 people with 1 different value (mainly `ziv2` and `ausw2`), 62 with 2 differences, 4 with three differences and 1 with 4 differences.

Large differences between SCS and census values may be a sign of an erroneous match. Actually some selective controls showed that errors such as mix-ups between father and daughter occurred. We also observed errors in census scanning and possible typing errors in SCS. However, we consider all the matches as real matches. The assumption is not fully satisfied but complete controls are no longer possible.

Age

Differences between years of birth in SCS and census are observed for 1.65% of the matches (1.61% for non-imputed values). These values decrease to 0.54% and 0.51% if we consider differences only when age group `Cage2` is different.

The largest difference is 90 years. Among non-imputed cases, 53.7% (weighted) show a difference of 1 or 2 years, 15.7% a difference of 3 up to 5 years and 17.9% differences between 6 and 10. The remaining 118 entries with differences larger than 10 represent 12.7% of the differences.

Sex, marital status, permit and occupation

We have few differences for the gender variable `sex` (0.69% weighted, overall and for non-imputed values), marital status `ziv2` (1.67% weighted, 1.63% for non-imputed values) and permit `ausw2` (1.25% weighted, 1.21% for non-imputed values). We observe larger differences for the occupation `taet` (8.08% weighted, 7.95% for non-imputed values); see the details in Table 12.2.

The misclassification error of the gender variable `sex` is not completely symmetric as we have more people "male in SCS and female in census" (204) than "female in SCS and male in census" (166). The *asymmetry factor* is $\phi = 204/166 = 1.23$. Imputation is not the reason for asymmetry. We do not have any valuable explanation for this small asymmetry (randomness? systematic error?).

We also observe an asymmetry for marital status `ziv2` and permit `ausw2`. There are for instance more people "married in the SCS and single in the census" than vice versa ($\phi = 157/46 = 3.4$) as well as more people "Swiss in SCS and C permit in census" or "C permit in SCS and B permit or less in census" than vice versa ($\phi = 166/37 = 4.5$ and $\phi = 230/59 = 3.9$). Some asymmetries are also observed for the occupation `taet`. We have for instance more people "without occupation in SCS and unemployed in census" than vice versa ($\phi = 323/89 = 3.6$) as well as more people "without occupation in SCS and in employment in census" than vice versa ($\phi = 2007/901 = 2.2$).

Table 12.2: Comparison of the gender variable `sex`, marital status `ziv2`, permit `ausw2` and occupation `taet`. Total number of entries and number of imputed values in the census.

sex			Census out	overall 1	2	imputed 1	2	Total
SCS	Male	1	393	23967	204	6	0	24564
	Female	2	383	166	24770	0	13	25319
	Total		776	24133	24133	6	13	49883

ziv2			Census out	overall 1	2	3	imputed 1	2	3	Total
SCS	Single	1	414	20238	46	116	135	4	1	20814
	Married	2	284	157	23636	130	8	30	2	24206
	Other	3	78	24	293	4467	4	6	10	4862
	Total		776	20419	23975	4713	147	40	13	49883

ausw2			Census out	overall 1	2	3	imputed 1	2	3	Total
SCS	Swiss	1	526	41886	166	51	104	3	2	42629
	C permit	2	101	37	5075	230	11	8	0	5443
	Other	3	149	11	59	1592	1	5	27	1811
	Total		776	41934	5300	1873	116	16	29	49883

taet			Census out	overall 1	2	3	4	imputed 1	2	3	4	Total
SCS	In empl	1	424	23953	300	901	9	564	14	92	0	25587
	Unempl	2	23	188	221	89	0	22	8	15	0	521
	No occ	3	217	2007	323	12143	28	312	26	881	0	14718
	< 15	4	112	13	1	18	8913	6	1	5	0	9057
	Total		776	26161	845	13151	8950	904	49	993	0	49883

Differences may be explained by different interpretation of the question (*e.g.* a widow may say "married" instead of "widowed"), not really accurate questions, form of the interview (paper form or phone) and psychological reactions. Note also that one household member gives answers for all household members during SCS and that matching errors may occur.

As a result, we probably have few misclassification errors in the census for gender, marital status and permit but less accuracy for the occupation (subject to interpretation and psychological reactions).

Position in household

A group of 792 P-sample matched entries are removed from the comparison between positions in households `stell`. Actually, we have special cases in the census such as: not an economic domicile (359, `stell=0`), collecting household (384, `stell=8`), and collective household (49, `stell=9`). Such households are not private households and removed from the comparison.

Among the remaining entries, we observe a difference for 8.6% of the matches (weighted) with a high level of 12.4% imputation (13.7% weighted). The rate is about 5.1% for non-imputed values. All the possible combinations between codes in SCS and census are observed; see Table 12.3. For example, we see the large dispersion for the code `stell=5` ("other head") and `stell=7` ("other").

Table 12.3: Comparison of the position in the household `stell`, values without imputation and with imputation.

`stell`			Census (non-imputed)							
			1	2	3	4	5	6	7	Total
SCS	Alone	1	4984	85	53	19	10	85	16	5252
	Husband/wife	2	145	20650	140	58	5	25	6	21029
	Common law	3	191	21	1965	25	16	52	46	2316
	Single parent	4	45	31	34	419	3	11	4	547
	Other head	5	56	45	39	27	19	33	14	233
	Relative	6	80	355	38	89	30	12056	28	12676
	Other	7	43	5	48	1	9	41	133	280
Total			5544	21192	2317	638	92	12303	247	42333
			Census (imputed)							
			1	2	3	4	5	6	7	Total
SCS	Alone	1	559	20	44	31	54	75	26	809
	Husband/wife	2	191	1259	50	119	17	70	11	1717
	Common law	3	166	6	143	39	58	28	34	474
	Single parent	4	12	8	18	137	9	33	2	219
	Other head	5	28	10	28	31	48	66	18	229
	Relative	6	74	66	50	68	136	1865	71	2330
	Other	7	27	4	18	3	92	17	43	204
Total		1057	1373	351	428	414	2154	205	5982	

We observe many asymmetries. For example of non-imputed values and at least 100 observations in one of both cells: we have more people "common-law husband/wife in SCS and living alone in the census" than vice versa (`stell=3` and `stell=1`, $\phi = 191/53 = 3.6$). A similar but less impressive behavior is also observed for "married in SCS and living alone in the census" (`stell=1` and `stell=2`, $\phi = 145/85 = 1.7$). We also have many more people "relative of the head in SCS and married in the census" than vice versa (`stell=2 and 6`, $\phi = 355/25 = 14.2$).

The main reason for other asymmetries such as "head" versus "husband/wife" is probably difficulty interpreting the question; see questionnaire in Appendix A.

Size of the household

For comparisons between the sizes of households, we also restrict the data set to matches in private households (see above). Among the remaining entries, we observe a difference for 11.7% of the matches (weighted) with a maximum of 10 (one case). The difference is one person for 9.4% and two people for 1.6%.

It is interesting to note that we have more people "in households with 2 people in SCS and 1 person in census" then vice versa ($\phi = 587/383 = 1.5$). The reason is possibly the difficulty of grouping people in households during the census process. As a case in point, communes are not always aware of the fact that two people are living together. In that case, two household questionnaires had been sent, without grouping if people did not explicitly state that they live together.

We also have more people "in households with 2 people in SCS and 3 people in census" then vice versa ($\phi = 783/495 = 1.6$). The interpretation is not clear: overgrouping in the census? missing people in the SCS?

Table 12.4: Comparison of the size if the household `nbHH`.

`nbHH`		Census						
		1	2	3	4-6	7-10	>10	Total
SCS	1	5774	383	123	84	4	0	6368
	2	587	12450	783	226	12	0	14058
	3	133	495	6805	635	4	0	8072
	4-6	105	156	633	18104	155	0	19153
	7-10	2	5	5	78	548	7	645
	>10	0	1	1	7	10	0	19
Total		6601	13490	8350	19134	733	7	48315

A complementary analysis at the household level such as size and composition of households would be interesting but not dealt with here. Are SCS households matched with census entries that are also in the same household?

As an indication, we observe some cases where there are 3 people in the same household according to the P-sample with 1 matched with an entry in a 1-person household and the other 2 people matched with entries in a common 2-people household or in two separate 1-person households.

12.3.2 Misclassification and Coverage

Comparisons between data collected in SCS and census show that some variables have very similar values and can therefore be considered to have a low misclassification error (sex, age groups, nationality, permit and marital status). Some other variables are less reliable and should be used with caution (position in the household, occupation, size of household).

Misclassification errors may be due to an error in SCS, to an error in the census or also to an error in the matching process. The exact origin of the error is not known but some indications may allow for detection of the error in particular cases. A false match would for instance inflate the difference (whereas a false non-match would inflate the bias).

Misclassification between domains that have different behavior in term of coverage bring heterogeneity to the estimation cells and therefore to the coverage estimates.

The asymmetry observed in the results have an impact on the distribution of variables and on the coverage estimation. Undercoverage in some domain d may be due to a real undercoverage of people in that domain or to a shift to another category. The domain we can define with data does not exactly correspond to the expected domain we'd like to study. Therefore, only domains based on variables with a low misclassification potential can be selected for accurate coverage estimations.

12.4 Population Membership and Domicile Errors

As with analysis of correct enumeration, we can estimate the rate of correct match $R_m^{(pop)}$ based on the status of correct match $P_{m,j}^{(pop)}$ in order to take membership of the matches in the target population into account. The rate can also be relaxed for the partners to get $R_m^{(popR)}$ based on $P_{m,j}^{(popR)}$.

Among the 49,107 simple matched entries of the P-sample, we have 359 matches at the civil domicile and 44 at the economic domicile but in collective households; see Figure 12.1. Therefore, the number of non-matched entries is 776 + 403 = 1179 when considering membership in the population.

A group of 285 matches out of the population (civil domicile) have a partner in the population (284 + 1, see Figure 12.1). Therefore, easing the criterion (relaxation) leads to 776 + 44 + (359-285) = 776 + 44 + 74 = 894 non-matched entries in the population, distributed into 5 types:

1. 776 simple non-matched;
2. 74 matches in civil domiciles and private households and partner in economic domiciles and collective households;
3. 32 matches in economic domicile but collective households and no partner;
4. 11 matches in economic domicile but collective households and partners in civil domiciles and private households;
5. 1 match in economic domicile but collective households and partner in civil domiciles and collective household.

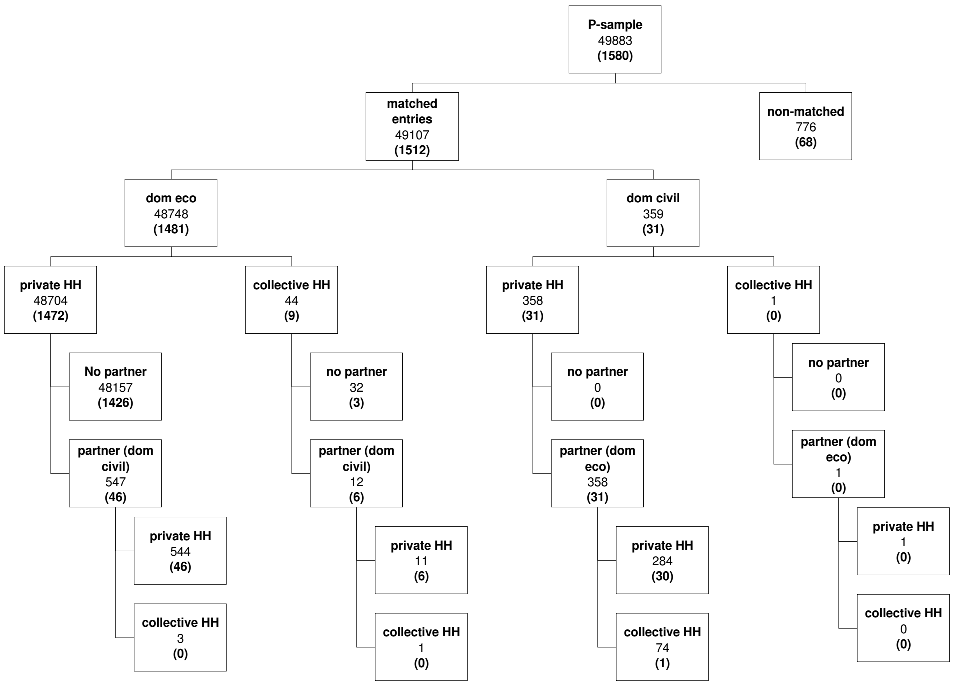

Figure 12.1: Decomposition of membership in the target population for the matches and their partners. Values for movers in parenthesis. In memory: target population = economic domicile and private household. The number of simple non-matched entries is 776.

The first type is clearly composed of non-matched entries. Types 3 and 5 are non-matched entries when taking into account the population (33 entries). Types 2 and 4 may be seen as possible errors in type of domicile (85 entries). If we reversed both types of domiciles, the entries would be accepted as matches in the population.

As a result, if we consider membership in the population but relaxed for partners and some errors in the type of domicile we get 776 + 33 = 809. This can be considered as a population-dependent lower threshold for non-matched entries. We call the corresponding status of correct match $P_{m,j}^{(popR2)}$ and rate of match $R_m^{(popR2)}$.

Note that unbalanced errors for the type of domicile would lead to problems in census counts in the target population (*e.g.* more cases with economic instead of civil than vice versa). This cannot be checked in our case.

Note also that the list of the 54 potential matches does not bring interesting new matches when already considering the partners.

The effect of membership in the population on the rate of correct match is not very large on the whole when we relax the criterion for partners and allow errors in the type of domicile (98.30% versus 98.36%); see Table 12.5. These two steps eliminate most of the difference of including membership in the population. The extension of the matching process to a check of partners may improve the estimation for future applications and for instance better identify errors in the type of domicile.

Table 12.5: Results for statuses of matches that depend on the population, total and depending on the moving status, with n the number of people, and $Pwei$ the weighted proportion $\widehat{R}_m$ and $1 - \widehat{R}_m$, respectively [%]. Standard error in parenthesis [%].

		Overall		Non-movers		Movers	
		n	$Pwei$	n	$Pwei$	n	$Pwei$
$P_{m,j}^{(pop)}$	0	1179	2.39	1071	2.24	108	6.89
	1	48704	97.61 (0.12)	47232	97.76 (0.10)	1472	93.11 (0.66)
$P_{m,j}^{(popR)}$	0	894	1.86	816	1.74	78	5.27
	1	48989	98.14 (0.11)	47487	98.26 (0.11)	1502	94.73 (0.77)
$P_{m,j}^{(popR2)}$	0	809	1.70	738	1.59	71	4.76
	1	49074	98.30 (0.11)	47565	98.41 (0.10)	1509	95.24 (0.66)
$P_{m,j}^{(s)}$	0	776	1.64	708	1.54	68	4.52
	1	49107	98.36 (0.11)	47595	98.46 (0.10)	1512	95.48 (0.66)
Total		49883	100	48303	100	1580	100%

12.5 Location and Time Delay

Most of the matches are found around the address on census day collected during SCS (97.7% in the same basic area, *i.e.* PA in NORTH or commune in TICINO). Few of them are found in the extended but not in the basic area (0.6%) and 1.7% are found at a farther address; see Table 12.6. We note that a high proportion of movers are matched at an address far from the SCS census day address (11.8%).

A special effort was made to collect addresses in the SCS. Therefore, enumeration in the census at a different location than the one collected in the SCS is most likely due to an error in the census.

Only 6 partners of matches are found around the address on census day (basic area) whereas the matches are out of the extended area (movers). Thus, relaxing the criterion for partners does not have much impact on the results and is not further studied in relation to location.

Table 12.6: Location of the matches. The address on census day is missing for 16 movers. "Out" for out of the extended area, "extended (no basic)" for extended but not basic area and "basic" for the basic area.

	Matched		Non-movers		Movers	
Area	n	$prop$	n	$prop$	n	$prop$
Missing	16	0.0%	0	0%	16	1.1%
Out extended	827	1.7%	648	1.4%	179	11.8%
Extended (no basic)	307	0.6%	258	0.5%	49	3.2%
Basic	47957	97.7%	46689	98.1%	1268	83.9%
Total	49107	100%	47595	100%	1512	100%

The rate of match when accepting matches in the proper extended area is $\widehat{R}_m^{(loc)} = 96.62\%$

(s.e. = 0.23), with 97.03% (s.e. = 0.22) for non-movers and 84.56% (s.e. = 1.37) for movers. As a result, undercoverage of movers around the location on census day reaches the high value of 15.4%.

More about non-movers

Most of the non-movers are found in the basic area around the address on census day collected during SCS (97.9%, weighted).

A set of 90.6% (weighted) people in the NORTH in same basic area are also in the same building. Therefore, about 9.4% of non-movers are found near but not exactly in the same building. This observation has to do with precise localization of people. It cannot be checked in TICINO because building IDs in SCS are not linked to IDs in the census.

During the survey, we noted that it was not an easy matter to identify buildings in the field and establish lists of households in the sampled buildings. Likewise, the information regarding who was assigned to which building is also potentially flawed in the census data process. Fine localization errors can therefore come from the SCS and/or the census.

It is important to note that fine localization error has a negligible impact on counts at the commune level. For example, only 51 out of 46,689 entries matched in same basic area are in a different commune.

More about movers

Some delay may occur in census data collection. The address on census day collected during SCS - assumed to be the real address on that day - may therefore differ from the address collected during the census.

Some interesting findings are observed when looking at the location of movers not only around the address on census day but also around the address on SCS day; see Table 12.7.

Table 12.7: Localization of matches for movers. Status of location around the address on census day and around the address on SCS day.

		SCS day			
		Out	Ext	Basic	Total
Census day	Missing	0	1	15	16
	Out	28	6	145	179
	Ext	3	4	42	49
	Basic	277	69	922	1268
	Total	308	80	1124	1512

A group of 277+3=280 matches are located around the address on census day but not on SCS day. These movers moved to a distant location. They are probably enumerated at the right address.

A set of 145+6=151 matches are in extended or basic area around the address on SCS day but not on census day. In NORTH, 128 among 147 movers are even found in the same building on SCS day (14 in the same basic area and 5 in the extended area). This means that a proportion of at least 8.7% (weighted) of the mover matches are probably enumerated in the census in the wrong location (time delay in the census process).

Many movers move between very nearby locations. A set of 922 mover matches are in the basic area of both addresses (63.9% weighted; same commune or commune that touches their postal area). The proportion is even higher when considering the extended area (1037, 71.8% weighted). Some of the cases are possibly also enumerated in the census at a wrong location. This is clearly the case for people found in the building on SCS day although they had stated that they were moving (in NORTH; 688 among 907 are found in the building on SCS day).

As a result, at least 688+151=839 out of 1512 matched movers seem to be enumerated at the wrong address (55%).

Relaxing the Criterion of Location

The rate of match depending on location but relaxed for movers to extended areas around both addresses on census and SCS days is $\widehat{R}_m^{(locR)}$ =96.93% (s.e. = 0.22%) with 97.03% (s.e. = 0.22%) for non-movers and 93.88% (s.e. = 0.71%) for movers.

When the criterion of location is relaxed, the rate sharply increases for movers from 84.6% to 93.9% due to the adjustment for the 151+15+1=167 time delay errors.

Note that we expect balancing between location errors due to time delays. Actually, the same census methodology is used for most of the movers around Switzerland (transfers, re-assignments). However, an unbalanced behavior should not be very important for overall estimates as only 3.3% of the P-sample are movers.

A summary of all location cases for non-movers and movers is found in Figure 12.2.

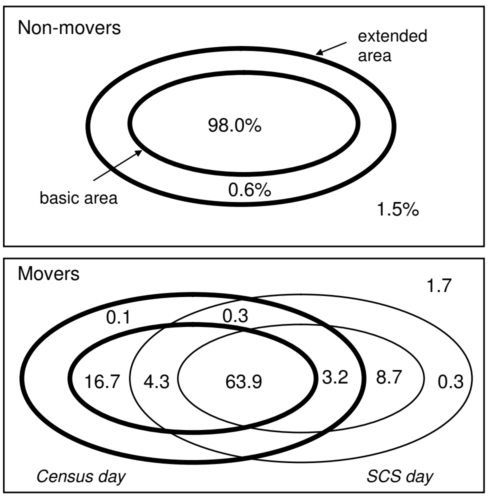

Figure 12.2: Distribution of non-movers and movers into basic and extended areas around the addresses on census (bold) and SCS (single) days (weighted proportions, [%]).

12.6 Combining Population and Location

Various combinations of results from membership in the population and location may be further analyzed.

Overall, relaxing the criterion for partners has more impact than relaxing the movers for location; see Table 12.8. Such results argue in favor of future matching processes including a special check to match status of partners.

Relaxing the criterion of addresses has the most noticeable effect on the rate of match for movers. Location errors (time delay) is a non negligible problem in the census. If the errors are not geographically balanced, this problem can also lead to coverage errors.

Table 12.8: Rates of correct match that depend on population and location, total and depending on the moving status [%]. Standard error in brackets [%].

	Overall	Non-movers	Movers
$\widehat{R}_m^{(poploc)}$	95.90% (0.23)	96.36% (0.22)	82.47% (1.42)
$\widehat{R}_m^{(popRloc)}$	96.42% (0.23)	96.85% (0.22)	83.88% (1.41)
$\widehat{R}_m^{(popR2loc)}$	96.58% (0.23)	97.00% (0.22)	83.32% (1.37)
$\widehat{R}_m^{(popR2locR)}$	96.89% (0.22)	97.00% (0.22)	93.64% (0.72)

with

$\widehat{R}_m^{(poploc)}$: right population (economic domicile and private household) and proper location (extended area around address on census day);

$\widehat{R}_m^{(popRloc)}$: relaxation of (1.) by including the partner of the match in the determination of population membership;

$\widehat{R}_m^{(popR2loc)}$: relaxation of (2.) by including the exchange between domiciles;

$\widehat{R}_m^{(popR2locR)}$: relaxation of (3.) by including the location around the addresses on SCS day for movers.

12.7 Results for Some Domains

The rate of match varies between subgroups of the population; see Table 12.9 for the results for the simple rate of match $\widehat{R}_m^{(s)}$, the restrictive rate that depends on population and location $\widehat{R}_m^{(poploc)}$ and the rate that also depends on population and location but with relaxed constraints (partners, error in domicile and error in location) $\widehat{R}_m^{(popR2locR)}$.

The data collected during the SCS are used as a reference for classification into categories. Only data with a low misclassification error are presented in the table.

We observe an overall undercoverage of 1-8% when considering the simple status of match. The lower value ca. 1% is observed for people in age group 60-79 (`Cage2=6`). The larger value is observed for foreigners holding a "B permit or lower" (`ausw2=3`).

Differences are observed between age groups with a larger undercount for the age group 20-31

(Cage2=3). The permit is also a discriminant variable but the difference between Swiss and C permit is very low. Differences are also observed in marital status, with a larger undercount for singles (also many young people).

We note that the variability of the estimates increases as the size decreases. For example, the confidence interval of the rate of match is 98.09% $\pm$ 0.52 for TICINO (var2=5). The difference between TICINO and CLASSIC (var2=1) or TRANSIT (var2=3), although large, is therefore not significant.

The rate decreases when we consider population and location. The smallest impact is observed in age group "80+" (-1%) and the largest impact in age group 20-31 (-5.5%). The impact is noticeable too for categories related to Ticino (-3.5% for ling2=3, NUTS=7, var2=5), the single (-3.5% for ziv2=1) and the French speaking part (-3% for ling2=2, -3.5% for NUTS=1).

Relaxing the criteria (partners, errors in domicile, location of movers) leads to an increase in the rate of correct match. The age group 20-31 is the most influenced (3%) and the age group 60-79 is the least influenced (0.3%) when criteria are relaxed. The larger difference between simple and relaxed rate that depends on population and location is observed for NUTS=1.

12.8 More about Undercoverage

Analysis of matches shows how difficult it is to isolate the various impacts. Undercoverage problems are mixed with misclassification errors, errors in type of domicile and errors in location.

Taking Switzerland as a whole, the various errors that coexist with coverage errors are not expected to lead to coverage problems if the effects are balanced. For example, location errors are not problematic if the number of people enumerated by mistake in region A instead of region B is the same as the number of people enumerated by mistake in region B instead of region A. Unbalanced mistakes may induce coverage errors at a lower level (*e.g.* region or age group): undercoverage on one side and overcoverage on the other. Time delay, for instance, will result in undercounting in the urban part and overcounting in the rural part if moves occur mainly from urban to rural communes.

The overall simple undercoverage of 1.6% is rather low and variation among subgroups are in the range of estimates in other countries; see Tables 1 and 2 on page 8. We also observe the effect of age group, origin, and some regional effects (Ticino and "Lake Geneva Region"). The difference between subgroups is, however, rather smooth compared with other countries. Especially between males and females.

Further analysis can be carried out to detect the variables that have the most influence (see the construction of estimation cells in Chapter 13) and other simple or combined effects such as characteristics of buildings and households or mover status within age groups. Some statistical tests can also be applied to determine significant differences between categories of variables (chi-square). New variables about the census process (transfers, re-assigment, origin of data) would also be of interest for estimations (not available).

Table 12.9: Rates of correct match for different domains. Number of elements n, number of non matched entries NM $= \sum_{j \in s_p}(1 - P_{m,j})$, rates of correct match $\widehat{R}_m^{(s)}$, $\widehat{R}_m^{(poploc)}$ and $\widehat{R}_m^{(R2R)} = \widehat{R}_m^{(popR2locR)}$ [%] with standard error s.e. [%].

Variable			n	NM	$\widehat{R}_m^{(s)}$	s.e.	NM	$\widehat{R}_m^{(poploc)}$	s.e.	NM	$\widehat{R}_m^{(R2R)}$	s.e.
Overall			49883	776	98.36	0.11	1998	95.90	0.23	1477	96.89	0.22
`sex`	Male	1	24564	393	98.27	0.13	1019	95.71	0.27	764	96.70	0.26
	Female	2	25319	383	98.45	0.10	979	96.09	0.23	713	97.07	0.22
`Cage2`	1-9	1	5957	82	98.54	0.21	158	97.21	0.34	134	97.60	0.33
	10-19	2	6189	86	98.70	0.19	285	95.94	0.35	167	97.44	0.29
	20-31	3	7339	247	96.50	0.34	646	91.03	0.57	410	94.13	0.50
	32-44	4	10826	164	98.35	0.16	369	96.41	0.28	310	96.93	0.26
	45-59	5	10303	104	98.82	0.14	316	96.77	0.28	268	97.25	0.25
	60-79	6	7879	75	99.10	0.13	194	97.47	0.36	165	97.82	0.35
	80+	7	1390	18	98.80	0.31	30	97.73	0.50	23	98.33	0.42
`ausw2`	Swiss	1	42629	526	98.72	0.09	1615	96.14	0.23	1144	97.19	0.23
	C permit	2	5443	101	98.15	0.29	185	96.60	0.41	156	97.08	0.38
	Other	3	1811	149	91.98	0.85	198	89.39	1.01	177	90.47	0.94
`ziv`	Single	1	20814	414	97.93	0.18	1145	94.47	0.31	744	96.27	0.29
	Married	2	24207	284	98.73	0.11	692	96.96	0.23	597	97.34	0.23
	Widowed	3	2486	34	98.77	0.26	57	97.95	0.37	49	98.20	0.36
	Divorced	4	2376	44	98.05	0.35	104	95.80	0.64	87	96.44	0.57
`ling2`	German + R	1	33724	467	98.50	0.11	1211	96.36	0.23	863	97.29	0.22
	French	2	14177	258	98.11	0.25	668	95.00	0.58	521	96.02	0.56
	Italian	3	1982	51	97.65	0.49	119	94.23	0.62	93	95.75	0.70
`NUTS`	Lake GE	1	9486	186	97.81	0.38	498	94.31	0.82	397	95.37	0.77
	Espace M.	2	13870	198	98.61	0.15	498	96.37	0.28	358	97.36	0.27
	Northwest	3	6056	78	98.50	0.27	177	96.85	0.35	130	97.59	0.36
	Zurich	4	8835	124	98.42	0.19	304	96.41	0.34	229	97.19	0.34
	East	5	6935	97	98.71	0.23	287	96.01	0.90	209	96.93	0.86
	Central	6	3150	50	98.43	0.25	135	96.54	0.36	75	98.00	0.27
	Ticino	7	1551	43	97.62	0.52	99	94.18	0.65	79	95.70	0.72
`taipop2`	Small	1	18306	246	98.50	0.15	713	95.83	0.29	507	96.91	0.27
	Middle	2	15845	216	98.68	0.16	607	96.11	0.48	446	97.13	0.47
	Large	3	15732	314	97.99	0.19	678	95.76	0.31	524	96.65	0.30
`urbrur2`	Town	1	10295	207	98.04	0.17	448	95.70	0.35	349	96.61	0.33
	Agglo	2	18295	262	98.51	0.19	673	96.27	0.39	496	97.23	0.39
	Rural	4	21293	307	98.44	0.17	877	95.63	0.41	632	96.69	0.38
`var2`	CLASSIC	1	8694	139	98.09	0.28	321	95.63	0.37	254	96.59	0.34
	SEMI-CLA	2	4940	51	98.93	0.24	183	96.16	0.61	108	97.82	0.39
	TRAN+FUT	3	34698	543	98.38	0.11	1395	95.98	0.25	1036	96.92	0.24
	TICINO	5	1551	43	97.62	0.52	99	94.18	0.65	79	95.70	0.72
`outsour`	No del	0	10487	185	97.93	0.28	427	95.09	0.48	336	96.36	0.43
	Global	1	33784	523	98.39	0.11	1357	95.98	0.26	1011	96.90	0.25
	Only mail	2	5612	68	98.46	0.42	214	95.92	0.45	130	97.51	0.43

Chapter 13

Estimation Cells (Post-Strata)

The choice of estimation cells, also called post-strata, is a key point in the dual-system estimation. A special effort is therefore made for this construction. The detailed procedure is described below; see Section 5.4 for the general methodology.

For practical reasons, the estimation cells are initially defined on the basis of the P-sample and simple status of match $P^{(s)}_{m,j}$. The results are then compared with the situation for the E-sample and the CE. Note that mover and CATI-CAPI statuses, both causes of heterogeneity, cannot be included in the definition of the estimation cells because they are available only on the P-sample side.

13.1 Eligible Variables

Three groups of variables are eligible to define estimation cells: demographic variables, variables about the reference commune (regional as well as socio-economical), and census data collection variables.

The demographic data are: sex (`sex`), age (`age`), marital status (`ziv`), nationality (`natio`, 1: Swiss, 2: other) and type of residence permit (`ausw`). These variables are supposed to have a low misclassification error; see Section 12.3. Position in household, occupation and size of household are not retained because of the greater likelihood of measurement error.

The variables about the reference commune are:

- Resident population 2000 (`pop`): continuous variable;
- Official language (`ling`): 1: German, 2: French, 3: Italian, 4: Romansh ;
- Urban-rural status (`urbrur`): 1: town center, 2: agglomeration, 3: isolated town; 4: rural;
- Nomenclature of Units for Territorial Statistics (`NUTS`): 7 regions in Switzerland.

The variables about the census data collection/process are:

- Census methodology (`var`): 1: CLASSIC, 2: SEMI-CLASSIC, 3: TRANSIT, 4: FUTURE, 5: TICINO;
- Outsourcing (`outsour`): 0: no delegation of tasks, 1: global packet, 2: only mail management.

The variable about the reference commune may be disturbed by some problems of location. However, the problems should not be very important (*e.g.* moves often occur between similar types of communes).

The two continuous variables (age and resident population of reference commune) are distributed into classes. The size of the group as well as the homogeneity of the match rate are used as criteria for selection of the classes. Result:

- Age groups (7 classes, `Cage2`): 1: 0-9, 2: 10-19, 3: 20-31, 4: 32-44, 5: 45-59, 6: 60-79, 7: 80 and older.
- Size of reference commune (3 classes, `taipop2` based on `pop`): 1: 0-1999, 2: 2000-7999, 3: 8000 or larger.

Some classes of categorical variables are quite small in the P-sample (and E-sample). Grouping is therefore applied after a comparison of sizes and match rates:

- Type of residence permit (`ausw2`): 1: Swiss, 2: permanent residence (C permit), 3: annual residence (B permit) and others;
- Marital status (`ziv2`): 1: single, 2: married, 3: widowed and divorced;
- Official language of reference commune (`ling2`): 1: German and Romansh, 2: French, 3: Italian;
- Urban-rural status of reference commune (`urbrur2`): 1: town-center and isolated town, 2: agglomeration, 4: rural;
- Census methodology of reference commune (`var2`): 1: CLASSIC, 2: SEMI-CLASSIC, 3: TRANSIT and FUTURE, 5: TICINO.

To sum up, the 11 variables that may be used to define estimation cells are: sex (2 classes, `sex`), age group (7 classes, `Cage2`), marital status (3 classes, `ziv2`), nationality (2 classes, `natio`), type of residence permit (3 classes, `ausw2`), size of commune (4 classes, `taipop2`), official language of commune (3 classes, `ling2`), urban-rural status (3 classes, `urbrur2`), Nomenclature of Units for Territorial Statistics (7 classes, `NUTS`), census methodology (4 classes, `var2`) and outsourcing (3 classes, `outsour`).

13.2 Selection of Variables

The choice of the final set of variables used to define estimation cells is based on two methodologies: logistic regression and discrimination. Note that this step corresponds to the search for an optimal model to explain the match - non match result (binary variable), and for a subset of variables that best reveals differences among classes, respectively.

We note that the correlation coefficient is quite high between some variables (*e.g.* 0.7 for age group - marital status, 0.55 for census methodology - size of the commune, -0.73 for urban-rural status - size of the commune).

Results from logistic regression analysis (PROC LOGISTIC; with weights but no sampling design): very significant variables (`ausw2, Cage2, ziv2`), significant variables (`taipop2,` NUTS, urbrur2), not very significant variable (ling2) and not significant variable (natio, well correlated with ausw2). The variables var2 and outsour are significant when combined with taipop2 (var2*taipop2 and outsour*taipop2). The variable sex is not significant as a simple effect but as a multiple effect with the age group (sex * Cage2).

Results from the discrimination methodology (PROC STEPDISC; with weights): the highly discriminant variables are ausw2 and ziv2. The variables ling2,taipop2 and urbrur2 are also significant.

Summary of the results from both analysis: variables significant for both (ausw2, ziv2, taipop2, urbrur2), significant for one of the methods (Cage2, ling2, NUTS), significant when combined with a significant variable (sex, var2, outsour), variable not significant (natio).

We decide to decrease the number of variables by excluding urbrur2, NUTS, var2, outsourc, and natio because of high correlations with the remaining variables and/or lower significance in the models.

The following 6 variables are kept for definition of the estimation cells: ausw2, ziv2, taipop2, ling2, Cage2 and sex, *i.e.* 4 demographic variables and 2 variables about the reference commune.

13.3 Construction of Cells

The minimum accepted size of estimation cells is 150 elements in the P-sample and 150 in the E-sample. This limit, which is expected to lead to stable variance estimates, is quite arbitrary. For example, the limit was set to 100 for the A.C.E. 2000 in the U.S (Davis, 2001). The influence of the limit is not quite clear but is expected to be negligible in our case, provided that the value is high enough. In this section, only P-sample sizes are considered.

We first aggregate some categories in order to avoid having estimation cells that are too small:

- Type of residence permit (ausw3, same as nationality natio): 1: Swiss, 2: others;
- Marital status (ziv3): 1: single, 2: married, widowed and divorced;
- Official language of the reference commune (ling3): 1: German, 2: French, Italian and Romansh.

Step 1: Combinations ausw3×ziv3×taipop2 (2×2×3=12 cells).

Step 2: Collapsing of taipop2 $\in \{1,2\}$ for foreigners (ausw3=2) (2×3 + 2×2=10 cells).

Step 3: Integration of the language ling3 within each of the 10 existing cells (2×3×2 + 2×2×2=20 cells).

Step 4: Integration of the Cage2×sex combinations within each existing cell (20×7×2 = 280 potential cells, 245 cells with elements, 157 with 1 to 150 elements).

Step 5: Collapsing in order to get a minimum of 150 elements of the P-sample in each estimation cell. Collapsing is based on analysis of the discriminant abilities of sex, resp. age group. For instance, within the data defined by Cage2 $\in \{1,2\}$ ×sex$\in \{1,2\}$ for the singles (ziv3=1), the sex has to be collapsed before the age group.

Step 5.1: Collapsing of `sex` if one of the 2 cells has less than 150 elements. Result: 171 cells (65 with less than 150 elements)

Step 5.2: Collapsing of age groups, with a different treatment for singles and the others (limit: 150): (1) `ziv3=1` (singles): `Cage2`$\in$ $\{1,2\}$, `Cage2`$\in$ $\{4,5\}$, `Cage2`$\in$ $\{6,7\}$; and (2) `ziv3=2` (others): `Cage2`$\in$ $\{1,2,3\}$, `Cage2`$\in$ $\{4,5\}$, `Cage2`$\in$ $\{6,7\}$. No collapsing for singles in the age group `Cage2=3` (20-31 years old). If a collapsing of the `sex` already occurred in any of the concerned cells, then collapsing of the `sex` took place for all the cells. Result: 136 cells (21 with fewer than 150 elements).

Step 5.3: Collapsing of languages (limit: 150). Result: 123 cells (4 with fewer than 150 elements).

Step 5.4: Collapsing of age groups `Cage2`$\in$ $\{4,5\}$ with $\{6,7\}$ in two remaining low numbers of elements. Result: 121 cells. We keep the unique cell with fewer than 150 elements (134 elements). No more collapsing.

Step 6: Small adjustment of the definition of estimation cells in order to include all the combinations of the 6 variables present in the E-sample and the census.

The resulting 121 estimation cells illustrated in Table 13.1 have between 134 and 1045 elements in the P-sample and between 151 and 1127 elements in the E-sample.

Note that we did not take into account the variability of the weights $w_{p,j}$ and $w_{e,i}$ within estimation cells during the selection of estimation cells (*e.g.* weights less variable within the census methodology than within the demographic data).

13.4 More about Estimation Cells

The rates $\widehat{R}_m^{(s)}$, $\widehat{R}_{ce}^{(popR)}$, $\widehat{R}_{net} = \widehat{CCF}^{-1}$ and $\widehat{R}_{under} = 1 - \widehat{R}_{net}$ clearly vary between estimation cells; see Appendix F. This variation confirms the need for considering heterogeneity in the probability of being counted in the census. A direct overall estimate would ignore the important variation that is modelled in the estimation cells.

The largest rate of net undercoverage $\widehat{R}_{under}$ is 7.6% in the cell `"A2Z1T12L12C 3S12"` (single foreigners in small or medium-sized communes, 20-31 years-old). This high value is due to the low rate of correct match 91.7%. The rate of net undercoverage $\widehat{R}_{under}$ also takes negative values (overcoverage) but none of them are significantly smaller than 0.

The smallest rate of correct enumeration $\widehat{R}_{ce}^{(popR)}$ =97.9% is observed for `"A1Z1T 1L 2C 3S 1"` (single Swiss males 20-31 in small communes where Romance languages are spoken). It is balanced by the rather small rate of match $\widehat{R}_m^{(s)}$ =96.9% with a resulting net undercoverage $\widehat{R}_{under}$ =1.1%.

The cell `"A1Z1T2L2C2S1"` (single Swiss males 10-19 in medium-sized communes where Romance languages are spoken) has exactly the same rate of correct enumeration $\widehat{R}_{ce}^{(popR)}$ and correct match $\widehat{R}_m^{(s)}$. The result is 100% net coverage.

The standard error of $\widehat{R}_{net}$ and $\widehat{R}_{under}$ ranges between 0.05% and 2.4% with an average value of about 1%. As a result, variability is quite high at the cell level because of the small sample sizes.

Table 13.1: Illustration of the 121 estimation cells, as combinations of the variables `ausw3`, `ziv3`, `taipop3`, `ling3`, `sex` and `Cage2`. Each "*" denote an estimation cell.

					Cage2						
ausw3	ziv3	taipop3	ling3	sex	1	2	3	4	5	6	7
1	1	1	1	1	*	*	*	*		*	
				2	*	*	*				
			2	1	*	*	*	*			
				2	*	*	*				
		2	1	1	*	*	*	*		*	
				2	*	*	*				
			2	1	*	*	*	*			
				2	*	*	*				
		3	1	1	*	*	*	*	*	*	
				2	*	*	*	*			
			2	1	*	*	*	*			
				2	*	*	*				
	2	1	1	1	*			*	*	*	*
				2				*	*	*	
			2	1		*		*	*	*	*
				2				*	*	*	
		2	1	1	*			*	*	*	*
				2				*	*	*	
			2	1	*			*	*	*	
				2				*	*		
		3	1	1	*			*	*	*	*
				2				*	*	*	
			2	1				*	*	*	
				2				*	*		
2	1	1,2	1	1	*	*	*	*			
				2	*						
			2	1	*	*					
				2							
		3	1	1	*	*	*	*			
				2							
			2	1	*	*					
				2							
	2	1,2	1	1	*			*	*	*	
				2				*			
			2	1	*			*	*		
				2				*			
		3	1	1	*			*	*	*	
				2				*	*		
			2	1				*	*		
				2				*			

Chapter 14

Net Coverage

Results from the E-sample and rate of correct enumeration as well as the P-sample and rate of correct match are combined by using the dual system technique and the synthetic assumption. The results are net coverage $\widehat{R}_{net}$ and the corresponding net undercoverage $\widehat{R}_{under}$ developed in Chapter 5.

14.1 Checks before Estimation

No E-sample and P-sample unit has an extremely large influence on the point and variance estimates; see Chapters 11 and 12. The weights are therefore left unchanged.

Some checks are applied for the DSE estimates to detect influential PSUs or estimation cells. The test statistic is the *net error*, defined as $Z = |(\widehat{N}_p - \widehat{M}) - (\widehat{N}_e - \widehat{CE})|$, with $\widehat{N}_p$, the weighted population total from P-sample, $\widehat{M}$ the weighted number of matches, $\widehat{N}_e$ the weighted population total from E-sample and $\widehat{CE}$ the weighted total of correct enumerations in the census[1]. Some PSUs and cells have net errors that emerge from the group (*e.g.* 6% of the census count for `"A1Z2T2L2C123S12"` and `"A2Z1T12L12C3S12"`). However, none of them seems to be very influential on the estimates.

The number of PSUs in estimation cells ranges between 15 and 109 in the P-sample and between 16 and 136 in the E-sample (at least 1 and maximum 135 people in combinations PSU x cell). Therefore, variance estimates in each cell, and subgroup, should be reliable. For the variance replicates, we do not regroup the cells or smooth the coverage correction factors. As a result, estimation of variance is expected to be rather conservative.

[1]The net error can also be used during the search for EE/CE and matches in order to determine PSUs that need further checking. No further checking was done in the current project but is recommended for future applications.

14.2 First Look at Net Coverage $\widehat{R}_{net}$

The estimator $\widehat{R}_{net}$ is based on the rate of correct enumeration $\widehat{R}_{ce}^{(popR)}$ and the rate of correct match $\widehat{R}_{m}^{(s)}$:

$$\widehat{R}_{net} = C^{(pop)} \left[\sum_{\ell=1}^{L} C_{\ell}^{(pop)} \, \frac{\widehat{R}_{ce,\ell}^{(popR)}}{\widehat{R}_{m,\ell}^{(s)}} \right]^{-1} \tag{14.1}$$

where $C^{(pop)}$ is the overall census count in the target population and ℓ is the identifier of estimation cell; see Chapter 5.

As a result, the overall rate of net coverage of the census (target population) is $\widehat{R}_{net} = 98.59\%$ (s.e. = 0.12%) with the corresponding rate of net undercoverage $\widehat{R}_{under} = 1 - \widehat{R}_{net}$ =1.41% (s.e. = 0.12%).

The undercount $1 - \widehat{R}_{m}^{(s)} = 1.6\%$ is therefore, partially compensated by the overcount $1 - \widehat{R}_{ce}^{(popR)} = 0.4\%$ to get the net undercount $\widehat{R}_{under} = 1.4\%$.

14.3 Results for some Domains

The larger rates of net undercoverage $\widehat{R}_{under}$ are observed for people aged 20-31 (2.84% with s.e. = 0.36%; `Cage2=3`) and foreigners (2.89 and 3.48% with s.e. = 0.32 and 0.39%; `ausw2=2 and 3`); see Table 14.1. The smaller rates are observed for people aged 60-79 (0.82% with s.e. = 0.12%; `Cage2=6`) and widows (0.79% with s.e. = 0.13%; `ziv=3`).

The 20-31 age group (`Cage2=3`) has a significantly larger net undercoverage than other age groups; see also Figure 14.1. The low rate of match $\widehat{R}_{m}$ =96.5% is combined with the lowest rate of correct enumeration $\widehat{R}_{ce}$=99.1%. Compared to other domains, the 20-31 year-old group contains many multiple entries and overlooked people, with a positive outcome for the overlooked people.

Foreigners have a significantly larger net undercoverage than Swiss citizens. Holders of "C permit" (`ausw2=2`) and "B permit or less" (`ausw2=3`) do not have the same behavior. The difference is not significant for $\widehat{R}_{ce}$ but $\widehat{R}_{m}$ is much smaller for "B permit or less". The final $\widehat{R}_{under}$ is therefore larger for "B permit or less". However, confidence intervals overlap due to the rather high variability. In Figure 14.1, we note that the net undercount of people holding a "C permit" is larger than the rate of undercoverage, and that the net undercount of people holding a "B permit or less" is much smaller than the rate of undercoverage. Such a behavior may be due to the choice of the estimation cells; which does not split foreigners into "C permit" and "B permit or less". The synthetic estimate $\widehat{R}_{under}$ is a smoothed estimate. It does not include the detailed behavior of "C permit" and "B permit or less", respectively. Estimations for foreigners are bias. The confidence intervals overlap. If separate estimation cells had been used, the difference would be significant.

Differences are observed between marital statuses. Widows (`ziv=3`) and divorced people (`ziv=4`) have a lower net undercoverage than single (`ziv=1`) and married people (`ziv=2`). The confidence intervals for single and married people overlap, but single people have lower $\widehat{R}_{ce}$ and $\widehat{R}_{m}$ than married people.

Rural regions (`urbrur2=4`) have a lower $\widehat{R}_{under}$ than town centers (`urbrur2=1`). Both confidence intervals overlap the agglomeration rate (`urbrur2=2`). Similarly, communes with less than 8000 inhabitants (`taipop2 in (1,2)`) have a lower $\widehat{R}_{under}$ than larger communes (`taipop2=3`).

Confidence intervals overlap for all the other variables. Due to variability, we do not observe any difference between males and females, or between languages. For example, $\widehat{R}_{under}$ is much smaller for Ticino (`var2=5`) than other methodologies but the sample is two small to detect a significant difference; see Figure 14.1.

We note that the difference between the overall rate of undercoverage and the overall rate of overcoverage is not equal to the overall rate of net undercoverage. This is explained by the choice of estimator. The estimated total of people $\widehat{N}$ is based on the rates $\widehat{R}_{ce}$ and $\widehat{R}_m$ but not on totals such as $\widehat{CE}$ and $\widehat{\text{UN}} = \widehat{N}_p - \widehat{M}$; see Section 1.6.2.

We also note differences between the final estimates $\widehat{R}_{under}$ in Table 14.1 and the estimate $\widehat{R}^{(0)}_{under} = 1 - \widehat{R}^{(s)}_m / \widehat{R}^{(popR)}_{ce}$ which does not make use of estimation cells, *i.e.* does not take into account the different probability of being correctly enumerated[2]. The overall estimate $\widehat{R}^{(0)}_{under} = 1.30\%$ is smaller than $\widehat{R}_{under} = 1.41\%$ but still in the confidence interval. Seven $\widehat{R}^{(0)}_{under,d}$ out of the 40 domains d are significantly different from $\widehat{R}_{under,d}$. The larger difference, observed for foreigners holding a "B permit or less" (3.5% versus 7.5%), is combined with the significative difference for foreigners holding a "C permit" (2.9% versus 1.5%). This feature confirm the uncertainty when it comes to results for foreigners. The set of estimation cells probably smooths the results for these domains to a major extent. The estimation for foreigners needs to be further analyzed for future developments.

14.4 More about Net Undercoverage

The overall net undercoverage of 1.4% is rather low and the variation among subgroups is in the range of estimates in other countries; see Tables 1 and 2, on page 8. We also observe the effect of age group and origin. The difference between subgroups is however rather smooth compared with other countries. The higher net undercoverage is mostly due to a larger number of overlooked people. Differences between some groups cannot be detected due to variability of the estimates.

While determining sample designs and basing ourselves on the Australian results for 1996, we aimed to get a standard error of 0.3% for groups of about 10,000 sampled people and a rate of net undercoverage of 1.2-1.8% (Renaud, 2002). All in all, the results are better than expected. For example, the E-sample and P-sample have 10,000-12,000 people in the age groups `Cage2=4` and `Cage2=5` with a standard error of 0.14 and 0.12%, respectively. The results are similar to those obtained in other countries; see Table 1 on page 8.

Estimates for large demographic groups, small and large communes and census methodologies are also available. However, some potential differences cannot be confirmed due to the small sample size. Further analysis could be carried out to better understand the various effects of undercoverage, overcoverage and combined net coverage in the data. Statistical tests would also be of great interest to get better information about differences between categories.

[2]Estimations of $\widehat{R}^{(s)}_m$ and $\widehat{R}^{(popR)}_{ce}$ may also be based on synthetic assumption but such an assumption in not necessary and therefore avoided.

Table 14.1: Rates of correct enumeration $\widehat{R}_{ce}^{(pop)}$, correct match $\widehat{R}_{m}^{(s)}$, coverage correction factor *CCF*, net coverage $\widehat{R}_{net}$ and net undercoverage $\widehat{R}_{under}$ with corresponding standard errors s.e. [%].

Variable			C	$\widehat{R}_{ce}^{(pop)}$	s.e.	$\widehat{R}_{m}^{(s)}$	s.e.	CCF	s.e.	$\widehat{R}_{net}$	$\widehat{R}_{under}$	s.e.
Overall			7121626	99.65	0.03	98.36	0.11	1.0143	0.0012	98.59	1.41	0.12
`sex`	Male	1	3497940	99.63	0.04	98.27	0.13	1.0148	0.0013	98.54	1.46	0.13
	Female	2	3623686	99.67	0.03	98.45	0.10	1.0139	0.0013	98.63	1.37	0.13
`Cage2`	1-9	1	810373	99.74	0.05	98.54	0.21	1.0136	0.0027	98.66	1.34	0.26
	10-19	2	833185	99.73	0.05	98.70	0.19	1.0105	0.0022	98.96	1.04	0.22
	20-31	3	1115804	99.07	0.09	96.50	0.34	1.0292	0.0038	97.16	2.84	0.36
	32-44	4	1544721	99.67	0.05	98.35	0.16	1.0146	0.0019	98.57	1.43	0.19
	45-59	5	1431771	99.78	0.04	98.82	0.14	1.0105	0.0015	98.96	1.04	0.14
	60-79	6	1146709	99.90	0.03	99.10	0.13	1.0083	0.0013	99.18	0.82	0.12
	80+	7	239063	99.89	0.06	98.80	0.31	1.0104	0.0028	98.97	1.03	0.27
`ausw2`	Swiss	1	5674266	99.67	0.03	98.72	0.09	1.0099	0.0010	99.02	0.98	0.10
	C permit	2	1020242	99.67	0.06	98.15	0.29	1.0298	0.0034	97.11	2.89	0.32
	Other	3	427118	99.44	0.11	91.98	0.85	1.0361	0.0042	96.52	3.48	0.39
`ziv`	Single	1	2975643	99.50	0.05	97.93	0.18	1.0175	0.0020	98.28	1.72	0.19
	Married	2	3377223	99.77	0.04	98.73	0.11	1.0126	0.0012	98.75	1.25	0.12
	Widowed	3	369339	99.75	0.08	98.77	0.26	1.0079	0.0013	99.21	0.79	0.13
	Divorced	4	399421	99.76	0.08	98.05	0.35	1.0103	0.0010	98.98	1.02	0.10
`ling2`	German + R	1	5128353	99.67	0.04	98.50	0.11	1.0129	0.0012	98.72	1.28	0.12
	French	2	1680062	99.65	0.06	98.11	0.25	1.0182	0.0028	98.21	1.79	0.27
	Italian	3	313211	99.47	0.12	97.65	0.49	1.0158	0.0020	98.44	1.56	0.19
`NUTS`	Lake GE	1	1296464	99.63	0.07	97.81	0.38	1.0187	0.0029	98.16	1.84	0.28
	Espace M.	2	1640489	99.65	0.09	98.61	0.15	1.0127	0.0010	98.75	1.25	0.10
	Northwest	3	976699	99.82	0.04	98.50	0.27	1.0133	0.0012	98.68	1.32	0.12
	Zurich	4	1221014	99.69	0.05	98.42	0.19	1.0148	0.0013	98.54	1.46	0.13
	East	5	1020897	99.60	0.07	98.71	0.23	1.0126	0.0012	98.76	1.24	0.12
	Central	6	665904	99.64	0.06	98.43	0.25	1.0121	0.0012	98.81	1.19	0.12
	Ticino	7	300159	99.46	0.12	97.62	0.52	1.0160	0.0020	98.43	1.57	0.19
`taipop2`	Small	1	1372958	99.66	0.05	98.50	0.15	1.0113	0.0014	98.88	1.12	0.14
	Middle	2	2398256	99.59	0.07	98.68	0.16	1.0108	0.0019	98.93	1.07	0.19
	Large	3	3350412	99.69	0.03	97.99	0.19	1.0180	0.0020	98.23	1.77	0.19
`urbrur2`	Town	1	2078780	99.65	0.04	98.04	0.17	1.0186	0.0021	98.18	1.82	0.20
	Agglo	2	3145541	99.64	0.06	98.51	0.19	1.0136	0.0012	98.66	1.34	0.12
	Rural	4	1897305	99.68	0.04	98.44	0.17	1.0108	0.0012	98.93	1.07	0.12
`var2`	CLASSIC	1	265607	99.61	0.05	98.09	0.28	1.0108	0.0012	98.93	1.07	0.12
	SEMI-CLA	2	174501	99.63	0.08	98.93	0.24	1.0117	0.0013	98.84	1.16	0.13
	TRAN+FUT	3	6381359	99.67	0.03	98.38	0.11	1.0144	0.0012	98.58	1.42	0.12
	TICINO	5	300159	99.46	0.12	97.62	0.52	1.0160	0.0020	98.43	1.57	0.19
`outsour`	No del	0	707126	99.51	0.07	97.93	0.28	1.0138	0.0012	98.64	1.36	0.12
	Global	1	6087937	99.68	0.03	98.39	0.11	1.0145	0.0013	98.57	1.43	0.12
	Only mail	2	326563	99.46	0.17	98.46	0.42	1.0113	0.0012	98.88	1.12	0.12

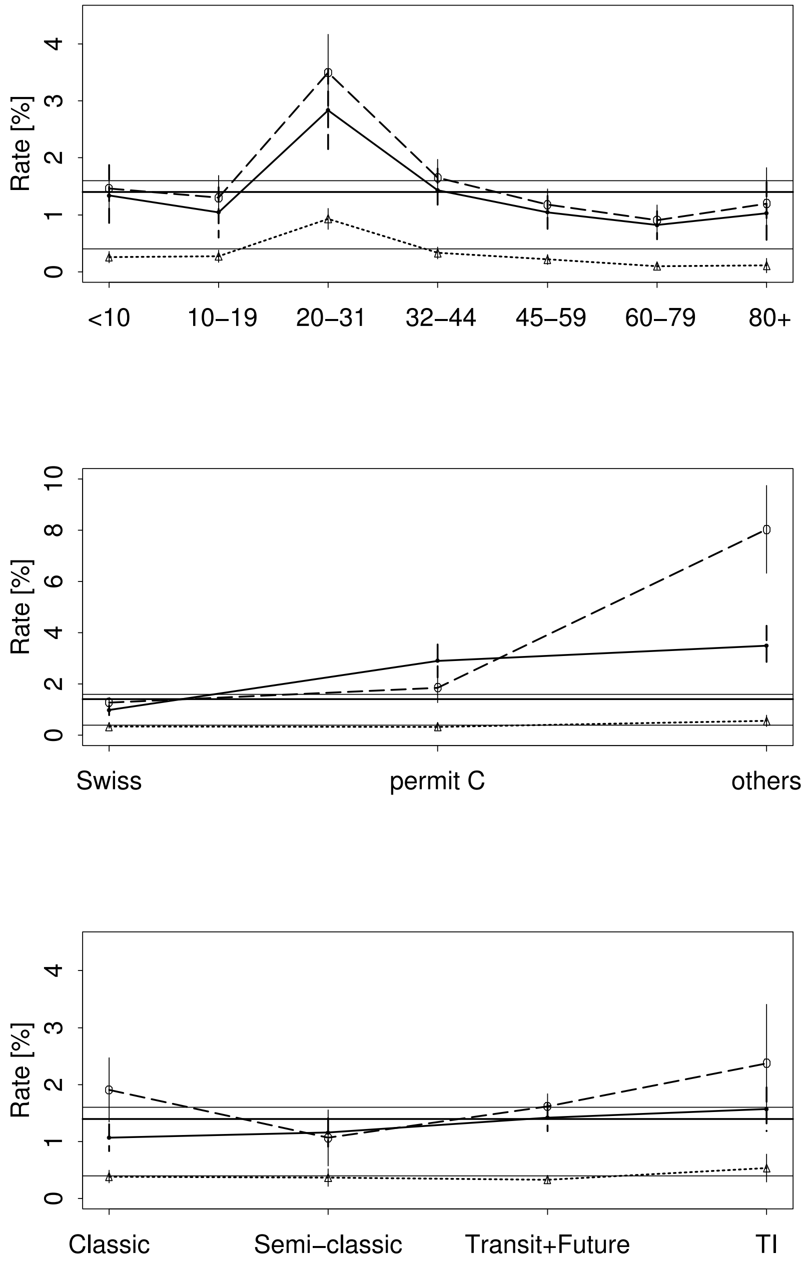

Figure 14.1: Detailed results for age groups (upper), permit (middle) and census methodology (lower). Rate of overcoverage $1 - \widehat{R}_{ce}^{(popR)}$ (triangles and dotted line), rate of undercoverage $1 - \widehat{R}_{m}^{(s)}$ (circles and dashed line) and rate of net undercoverage $\widehat{R}_{under}$ (plain bold line); with confidence intervals. Horizontal lines illustrate the overall rates.

Chapter 15

Conclusion

The conclusion summarizes results and remarks about the census data and the coverage estimation project.

Main Results about Census Data

The census overlooked 1.6% (s.e. = 0.11%) of the population and erroneously counted 0.4% (s.e. = 0.03%) entries. The resulting rate of net undercoverage is 1.4% (s.e. = 0.12%) with larger values for some subgroups of the population such as 20-31 years-old people (2.8%) or foreigners (2.9-3.5%). The results are in the range of the results in other countries.

We observe few overcoverage problems. However, a special search in the census data set for double entries at the building level would decrease the number of multiple entries even more. Furthermore, we suspect some missing links between both entries for people with two addresses.

Undercoverage is not negligible, especially in some subgroups of the population. During the analysis, we detect a mixture between totally overlooked people, misclassification errors, errors in type of domicile and errors in location.

Potential misclassification errors are detected for some variables collected during the census. For example, the data collected during census and SCS have large differences for position in the household, occupation and size of household. Therefore, census variables relating to the labor market and households may be somewhat flawed. Misclassification errors may also have an impact on coverage in the corresponding subgroups.

We observe a time delay between the census day and the effective data collection day. More than half of the movers seems to be enumerated at the address on SCS day but not at the address on census day.

SCS, P-sample and E-sample

The SCS operations worked relatively well but can also be improved. Having one common sampling frame for NORTH and TICINO would be one improvement. Avoiding some stages such as the mail delivery area in the sample selection would be another (weight variability, complexity). The list of households in the sampled buildings drawn up with the help of post office employees also needs some more checks and precise instructions to avoid including a number of non private households such as firms. Identification of the building in the field also needs some improvement (fine localization). In addition to the addresses on census and SCS day, we should also collect the moving date (time delay) and review the encoding of the results to make easier comparisons with the census data set.

One point to think about is the link between P-sample and E-sample. In the current project, we used the same PSUs but different further stages (P-sample stages not available in the census data set that is used as the frame for the E-sample). This choice made it easy to apply the jack-knife methodology for variance estimation but should be studied in comparison with alternative approaches such as two completely independent samples or identical units.

A special problem in the Swiss coverage estimation project is the lack of information about the E-sample. Having additional information about the situation of E-sample people on census day would be quite useful. A consequence of interviewing the E-sample is likely to be more erroneous enumerations which lowers the rate of correct enumerations and the net rate of coverage. Interviews with a subsample would already be quite useful, but interviewing the complete E-sample would clearly be even more enlightening. However, the practical consequences of such interviews call for further consideration (*e.g.* resources, time delay and availability of the census data set).

Searches for Matches and CE

Among difficulties that may occur in the dual system estimation, one of the most common is the lack of precision in the match. Accuracy such as number of false matches and number of false non-matches was not analyzed in the project. For further improvements, information may be found in Fellegi and Sunter (1969) and Ding and Fienberg (1994). For example, see also Cella et al. (2004) for the Italian agriculture census 2002.

The matching process can be improved in a future application. A search for multiple entries as well as a search for matches should be extended to include a check of partner entries. Continuing checks such as using net error at the PSU level would also improve control over the matching process. Finally, detailed documentation of all steps would also be of great help.

Estimation Methodology

One important point to work on for future coverage estimation is the choice of target population. The decision to restrict the P-sample and E-sample to people in private households and, especially, to people at economic domicile brings some complexity and uncertainties to estimations. We could, for instance, exclude collective households to avoid the complex procedure in SCS but include all types of domiciles. The type of domicile would then be suitably treated as a domain in the estimation.

Dual system estimation and synthetic assumptions work in a satisfactory manner. However, one could consider using more modelling instead of estimation cells and further study special effects such as the observed smoothing of the results for foreigners.

Some further analysis could be done to detect bias in coverage estimates. For example, the P-sample non-response adjustment is known to skip the Swiss-foreigner effect (as a consequence, net undercoverage is probably underestimated for foreigners). Comparisons with auxiliary data should be studied. One could also consider using the foreigner/Swiss ratio to correct estimates in a similar way as the sex ratio in some countries.

An additional area of study is the use of alternative, more refined, variance estimation methodologies. The effect of misclassification errors and matching errors may also be further studied to better split the various influencing factors.

Specific analysis of particular cases (non-matches, inversion of type of domicile, etc.) and special searches in the census database would also be of interest especially to detect possible improvements in the way that census data were processed. Did overlooked people receive a questionnaire, or did they not return the questionnaire, or did the commune fail to provide administrative data for missing responses, or did they disappear during census processing? Information about census processes such as transfers (*e.g.* moving, grouping of people in households) and origin of data (*e.g.* questionnaire, Internet, complementary phone interviews, commune) would be of great interest for future estimations.

Choices had to be made for the current report. Some points have been described in detail and others are only presented in a general way or not even mentioned. Various points could clearly be covered in more detail in future estimations. The 2000 results can be used to define objectives, to plan the sampling design and to estimate the expected variability of results.

Organization and Documentation

The organization of the project was split into two pools: census staff and Statistical Methods Unit (METH). As it turned out, METH ended up taking care of the sampling and estimation methodology as well as many general tasks having to do with design and organization. For future projects, the organization needs general review, re-evaluation of the required resources and greater integration of the coverage estimation project in the census framework, while keeping the operations strictly separate from the census process. There also needs to be more discussion about the process during operations and better documentation of what has been done.

We should point out that working with census data is quite a complex job. Improved documentation about the data and processes applied to the data would help having a better overview of the possible effects relating to the observed results.

General

The coverage estimation of a population census is quite a challenging project. The results of this first estimation for a Swiss census proved to be instructive. In many cases, we detected similar coverage behavior as that found in other countries. We also quantified errors and collected information about various possible improvements relating to census data. The experience we gained can be used to improve the future censuses as well as the methodology and organization of future similar projects of coverage estimation.

It is not yet known how future censuses in Switzerland will be conducted, whether they will be mostly based on administrative registers or similar to the Census 2000. What is clear, however, is that coverage estimations using data from independent surveys proved to be of great interest since they provided us information about the quality of the census data and possible improvements. We therefore feel that coverage estimations should be an integrated part of future censuses.

APPENDICES

Appendix A

Population Census 2000

The Swiss Population and Housing Census took place in 2000 with census day on 5 December 2000 (4 December at midnight). Information is collected for all 7.28 million inhabitants, 3.12 million households, 3.76 million housing units (dwellings) and 1.47 million buildings.

Census data are used for various purposes such as distributing the 200 seats of the national council among the 26 cantons, determining upon the official language of the communes, distributing of funds and subsidies and making other political decisions. Results may be found in OFS (2002b, 2003a, 2003b).

A.1 General Information

The Swiss Federal Statistical Office (SFSO) is responsible for conducting decennial censuses, but data collection is the responsibility of the communes. Communes had to make a choice between various *census methodologies*:

CLASSIC: enumerators visit households to bring and take back the questionnaires;

SEMI-CLASSIC: preprinting of questionnaires using the register of inhabitants (one in each commune), dispatching by mail and visit of enumerators to take back the questionnaires;

TRANSIT: preprinting the questionnaires using the register of inhabitants, dispatching and return by mail;

FUTURE: same as TRANSIT with link between households and dwellings in the register of inhabitants.

Ticino canton organized the census on its territory by using a methodology similar to TRANSIT.

There are three types of questionnaires: the household questionnaire, the personal questionnaire (see below) and the building questionnaire. The preprinting of questionnaires, the mail dispatch and the check of mail return was centralized for (almost) all Switzerland but Ticino. Most of the people in SEMI-CLASSIC, TRANSIT and FUTURE communes had the opportunity to fill in the personal and household questionnaires either by Internet or on paper. The people in CLASSIC communes had to fill in the questionnaires on paper.

Communes filled in the forms with their administrative information in case of non-response from the people (*e.g.* clear refusal or people not found but known to exist). Therefore, we do not apply any correction for non-response or whole person imputation in the census.

A.2 Processing and Definitions

The people are enumerated at one domicile (*e.g.* a person living in a single place) or at two domiciles (*e.g.* a student with one domicile with his family and the other at the place of study).

In the case of two domiciles, one of the enumerations is coded as the civil domicile and the other as the economic domicile. The *civil domicile* is the place where the official papers are registered ("acte d'origine" as well as taxes for Swiss people and residence permit for foreigners). The *economic domicile* is the place where the person mainly resides (4 or more days a week). For instance, the student has his civil domicile with his family and the economic domicile at the place of study.

Everybody has a civil domicile and an economic domicile, except people who have a civil domicile in Switzerland and an economic domicile abroad. In most cases, both the civil and the economic domiciles are the same (97.7% of the resident population), *i.e.* most people have one single domicile.

People with a civil domicile that differs from the economic domicile are linked by using a matching processing to avoid double counts (DD people). The type of domicile is mainly determined on the basis of the information collected in the questionnaire (question: "Where do you mostly reside (4 or more days a week)?").

People in vacation housing units that are not their civil or economic domicile (*e.g.* chalets, holiday dwellings) are not counted at this address in the census.

The census count is available for the civil domicile and the economic domicile. The *resident* population of a given commune is defined as the people who have their economic domicile in that commune.

There are two types of households. A *private* household is defined as a group of people that live in the same housing unit (*e.g.* a family). A *collective* household is defined as a "non-private" household (*e.g.* jails, hospitals or retirement homes).

During the census data processing, people are linked to the households, which are linked to the dwellings, which are then linked to the buildings. About 1.8% people are linked to *collecting households* (max 1 in each building) if the household could not be defined. Similarly, *collecting buildings* were created to accommodate 2.9% people not linked to an enumerated building (max 1 in each commune). The reason for being linked to a collecting building is mostly technical and not related to any "homeless" situation but to the fact that collected information does not allow for a confident link with a real building.

Automatic deterministic corrections as well as statistical imputation were applied to the census data. The statistical imputation is based on the New Imputation Methodology (NIM) developed by Statistics Canada (Kilchmann and Eichenberger, 2005). The final census data set contains very few missing items in the demographic variables. We do not have any imputation of whole persons in the census count.

A.3 Personal Questionnaire

Federal Population Census 2000

Data protection and statistical secrecy are regulated in arts. 4 and 5 of the Federal Act on Population Censuses. Answers on the first page may be used to update your commune's Population Register.

Personal Questionnaire

5 December 2000

To be completed by the commune

Commune: – SFSO No.: Register No.:

Building No.: – Dom.: Commune of registration:

Census District No.: Household No. 1: – Household No. 2: –

1 2 3

Please use a black or blue felt-tip or ball-point pen and not a pencil. Also please check whether the pre-printed details are correct and rectify any mistakes. Thank you!

Please complete in block capitals: A B C D E F
Where you have a choice of answers, please put a cross in the appropriate field(s):

A. Name and address

Residence **A**

Name:

First name(s):

(If subtenant) landlord/lady: C / O

Floor: Street: No.:

Postcode: Locality:

B. Do you have a second place of residence?

No (just residence A)

Yes (specify):

Residence **B**

(If subtenant) landlord/lady: C / O

Floor: Street: No.:

Postcode: Locality:

Canton: **or** foreign country:

Where do you mainly reside (4 or more days a week)? Residence **A** Residence **B**

English

1. Date of birth

Day: Month: Year:

2. Gender

Female Male

3. Marital status

Married persons should state the year when they married their present partner.

Legally separated persons should mark «married».

Single

Married → since: (year)

Widowed → since: (year)

Divorced → since: (year)

4. Nationality

Dual nationals of Switzerland and some other country should mark «Swiss» and name their second nationality.

Swiss

a) How long have you had Swiss nationality? from birth or since: (year)

b) Do you have another nationality besides Swiss nationality?

no yes → of which country?

Persons with several nationalities should indicate the country which last granted them citizenship. Stateless persons and refugees should indicate their country of origin.

Foreigner

a) Of what country are you a national?

Italy France Portugal Turkey Croatia

Germany Austria Spain Rep. Yugoslavia Rep. Macedonia

Of another country, namely:

The letter indicating the type of permit (A, B, C, F, L, N, S) appears in capitals on the permit.

b) Type of foreigner's residence permit/residence status

Permanent residence permit (C permit)

Applicant for asylum (N permit)

Short-stay permit (L permit)

Annual residence permit (B permit)

Person in need of protection (S permit)

Swiss Federal Department of Foreign Affairs permit

Seasonal permit (A permit)

Temporarily admitted foreigner (F permit)

Other status

5. Place of residence 5 years ago: where were you living on 5 December 1995?

- At the same address as now (residence **A**)
- In the same commune (as residence **A**) but at another address
- In another commune (specify):

Postcode: Locality: Canton:

- Abroad → Country:

6. Commune of residence at time of birth: where was your mother resident when you were born?

- In the same commune as residence **A**
- In another commune (specify): Canton:
- Abroad → Country:

7. To what church or religious community do you belong?

- Roman Catholic Church
- Protestant (Reformed) Church
- Old Catholic Church (altkatholisch)
- A Jewish community
- No affiliation
- A Muslim community
- An Orthodox community (Russian, Greek, Serb)
- Other church or religious community, namely:

8. Language

For infants who cannot speak yet, indicate the mother's language.
Persons speaking Friaulian or Ladin should not indicate «Italian» but «Rhaeto-Romansch».

a) What language do you think in and know best? (select just one reply)

- German
- French
- Italian
- Rhaeto-Romansch
- Other language, namely:

b) What language(s) do you speak regularly (several answers possible)

Schoolchildren and students should not list the languages they are studying but only those they speak regularly at school.

	Swiss German dialect	High German	Swiss French patois	French	Ticino or Grisons Italian dialect	Italian	Rhaeto-Romansch	English	Other language(s)
at school, at work									
at home, with your family									

9. Are you the father or mother of one or more children?

Including adult or deceased children

- no
- yes → **a) How many children?**
  → **b) Year of birth of your child/children?** Child 1: Child 2: Child 3: Child 4:

If you have more than 4 children, please add the year of birth of your youngest child:

10. What is your position in the household? (select only one reply)

All persons living in the same dwelling make up a single household.
Heads of households are persons socially and economically responsible for the household.
In households consisting of a couple (with or without children), both partners are regarded as heads of households.

Head of household

- Living alone
- Husband/wife
- Common-law husband/wife
- Single parent
- Other head of household

Relative of a head of household

- Son, daughter, stepson, stepdaughter, son-in-law, daughter-in-law
- Father, mother, stepfather, stepmother, father-in-law, mother-in-law
- Brother, sister
- Other relative of a head of household

Other position in household

- Flat-mate/commune member, non-related co-dweller
- Domestic employee, au pair
- Lodger, subtenant
- Other member of household (eg foster child, boarder)

11. Education

Tick all education/training you have completed in column a) and your present education/training in column b).

a) Completed education/training (mark all completed courses)	b) Ongoing education/training (select only one reply)	
○	○	None
○	○	Compulsory education (primary, junior secondary/high, assessment school, preparatory senior-secondary school, special school)
○	○	Certificated college (up to 2 years), administrative/transport college, social work, introductory course for nursing professions (1 or 2 years), preparatory vocational courses, basic vocational training (with contract)
○	○	Apprenticeship or full-time vocational college (eg commercial college, training in manual skills)
○	○	High-school certificate college, vocational high-school diploma, certificated college (3 years)
○	○	Teacher-training college (eg nursery, primary school), music, gymnastics and sports
○	○	Advanced technical and professional training (eg federal certificate of proficiency, diploma, master-craftsman certificate, higher commercial management college [HKG], technical college)
○	○	Higher college of technology (eg HTL, HWV, HFG, HFS) with full-time education lasting a minimum of 3 years (including post-graduate degree)
○	○	Specialized university (including post-graduate degree)
○	○	University, institute of technology (including post-graduate degree)

Questions 12, 13 and 14 are intended for those aged 15 and over

12. Profession studied, highest qualification obtained

Eg «CLERK», «ELECTRICAL MECHANIC», «NURSE (SRN)», «LL.B», «MD»

If possible, the official designation of the qualification/degree obtained should be entered.

13. Occupation: present situation

Please tick everything that applies.

Average number of hours a week

- ○ In (full-time) employment → hours
- ○ In (part-time) employment (at least one hour a week) → hours
- ○ Several (part-time) jobs → hours
- ○ Unemployed
- ○ Not in employment but seeking a job
- ○ Not in employment but future job guaranteed
- ○ Not employed and not looking for a job
- ○ Undergoing training (school, studies, apprenticeship) → hours
- ○ Retired, pension beneficiary (old-age, disability, etc.)

«In employment» means persons:
- who work **one hour or more a week** against payment
- who work in a **family business** without payment
- who are currently ill, on paid maternity leave or military service but are otherwise in employment.

Casual jobs should also be counted.

Apprentices should select both «In employment» and «Undergoing training». The appropriate number of hours must be given for both categories.

14. Work in home/family, voluntary work

(several replies possible)

Average number of hours a week

Including child care, nursing relatives and disabled persons in the same household

- ○ Work in own household → hours
- ○ Voluntary work → hours
- ○ No such activity

«Voluntary» means unpaid or only partly reimbursed work, eg:
- Caring for/nursing persons outside one's own household
- With charitable or church organizations, youth and environmental-protection organizations, interest groups, sport or cultural clubs, political parties, public office, etc.

Questions 15 and 16 are intended for people in employment and apprentices

15. What is your current professional status?

Indicate your main job (select just one reply)

- ○ Self-employed **without** employees (own business, free-lance)
- ○ Self-employed **with** employees (own business, free-lance)
- ○ Relative employed in family business

Employed as
- ○ apprentice (indentured or not)
- ○ employee in own corporation (eg stock corporation, plc)
- ○ manager, executive employee, senior civil servant
- ○ middle or junior level, eg office manager, section head, branch manager, group manager, workshop foreman, foreman
- ○ white-collar worker, blue-collar worker, trainee

- ○ Other position, namely:

English

16. What is your present occupation?

Wherever possible, indicate the exact title of your job. Your reply should clearly indicate the precise nature of your work, eg «METAL GRINDER» (instead of just «GRINDER»), «SHOP ASSISTANT, SHOES» (instead of «SHOP ASSISTANT»), «CLERK» (instead of «EMPLOYEE»), «MANAGER, FINANCIAL SERVICES» (instead of «MANAGER»), «CLERK OF COURT» (instead of «LL.B»), «ARTIST/PAINTER» (instead of «PAINTER/DECORATOR»), «PRIMARY TEACHER» (instead of «TEACHING»)

Indicate your main occupation (select just one reply).

Questions 17 to 21 are intended for employed persons, apprentices, schoolchildren and students

If you are both employed and in education/training, answer both columns

Employed persons

- If you work in several places, mention your main job base.
- If you move around in your job (eg driver, railway employee or construction-site worker), indicate where you usually start work.
- If you work from home, give your employer's address.

Schoolchildren, students

17. Where do you work, where do you normally start work? Where do you normally go to school?

Employed persons

State your place of work with the exact address:

Name of company:

Street (or usual designation): No.:

Postcode:

Locality (even if in neighbouring foreign country):

Canton: If abroad, indicate country:

Travelling (no fixed place of work)

Commercial travellers should enter «travelling» as well as their employer's address.

Schoolchildren, students

State your place of education with the exact address:

Name of school:

Street (or usual designation): No.:

Postcode:

Locality (even if in neighbouring foreign country):

Canton: If abroad, indicate country:

18. From which address do you normally leave for work/school?

Employed persons

- Residence **A** (as given on page 1)
- Residence **B** (as given on page 1)

Schoolchildren, students

- Residence **A** (as given on page 1)
- Residence **B** (as given on page 1)

19. How long does the trip to work/school usually take? (door-to-door)

Employed persons

- I work in the building I live in

Hours Minutes

Schoolchildren, students

- I live in the school building

Hours Minutes

20. How often do you commute to work/school (round trip)?

Employed persons

a) A day:
- once
- twice
- more than twice, namely times

b) On how many days a week? days

Schoolchildren, students

a) A day:
- once
- twice
- more than twice, namely times

b) On how many days a week? days

21. What means of transport do you usually use to go to work/school?

Mention all means of transport used on the same day for this journey.

Employed persons

- None, I walk all the way
- Bicycle
- Moped
- Motorcycle, scooter
- Car (driver)
- Car (passenger)
- Factory bus
- Train (SFR, private railway)
- Tram, municipal bus, trolley bus
- Postbus, coach
- Other (eg boat, cable railway)

Schoolchildren, students

- None, I walk all the way
- Bicycle
- Moped
- Motorcycle, scooter
- Car (driver)
- Car (passenger)
- School bus
- Train (SFR, private railway)
- Tram, municipal bus, trolley bus
- Postbus, coach
- Other (eg boat, cable railway)

Does your home have a telephone? Permanently installed / Mobile (Natel) / No telephone

Contact for queries

Home phone number: / Business phone number: /

Many thanks for your cooperation!

Appendix B

Demographic Estimations and Census Counts

Demographic estimates of the population are available for Switzerland on the 31st of December of each year. Estimates for the *permanent resident population* as well as for the *resident population* are published (OFS, 2002a, 2003c). Revisions occur after each decennial census. The adjustment due to the census 2000 corresponds to a small decrease of about 0.1% see *e.g.* OFS (2004).

The permanent resident population includes all people who maintain their civil domicile in Switzerland for at least one year. People with seasonal and short-stay permits (A and L permits) are not included. Similarly, people temporarily admitted (F permit) and applicants for asylum (N permit) are not included, except children 0-4 years old.

The resident population includes all people who maintain their civil domicile in Switzerland at a given time. The definition is similar to the definition used in the census, except the type of domicile: civil population instead of economic population.

Comparisons between demographic and census data are not very reliable since the definition of the population and the reference day are not identical. The demographic resident population is however more usable than the demographic permanent resident population. The difference between civil and economic population is not large at the national level.

A rough comparison shows that demographic counts are on the whole slightly larger than census counts; see Table B.1. Demographic counts are smaller for Swiss citizens and larger for foreigners. Non-negligible differences are observed for foreigners, with larger values for males than females.

Table B.1: Comparison between census (5 December 2000) and demographic estimates (31 December 2000, not revised) of the resident population. Census C and demographic D counts, as well as the relative differences $dr_{eco} = (D - C_{eco})/D$ and $dr_{civ} = (D - C_{civ})/D$ [%].

	D	C_{eco}	C_{civ}	dr_{eco}	dr_{eco}
Total	7,304,109	7,288,010	7,287,357	0.22%	0.23%
male	3,583,886	3,567,567	3,567,327	0.46%	0.46%
female	3,720,222	3,720,443	3,720,030	-0.01%	0.01%
Total Swiss	5,779,685	5,792,461	5,791,768	-0.22%	-0.21%
male	2,762,579	2,766,020	2,765,737	-0.12%	-0.11%
female	3,017,106	3,026,441	3,026,031	-0.31%	-0.30%
Total foreigners	1,524,424	1,495,549	1,495,589	1.89%	1.89%
male	821,307	801,547	801,590	2.41%	2.40%
female	703,116	694,002	693,999	1.30%	1.30%

Appendix C

Swiss Coverage Survey

Questionnaire used for the CATI and CAPI data collection. In French with some technical comments in German.

Questionnaires CATI et CAPI
Version française

A **Bonjour, mon nom est........ et je travaille pour l'Institut d'études de marché IHA-GfM à Hergiswil.**

Nous menons actuellement une enquête post-recensement 2000 pour le compte de l'Office fédéral de la statistique. Nous vous avons envoyé récemment un courrier d'information à ce sujet. J'aimerais à présent vous poser quelques questions dans le cadre de cette enquête de couverture. Auriez-vous un petit moment à nous consacrer?

F01. Vous êtes bien Madame / Monsieur................?

oui, nom et prénom sont corrects...... 1 ❑ → Question 4
oui, petite erreur - rectification.............. 2 ❑ → Question 4
non, autre personne............................. 3 ❑ → Question 2

EDV:	Bitte Name und Vorname einblenden.

F02. Notre questionnaire s'adresse à toutes les personnes qui font partie du même ménage (c'est-à-dire à toutes les personnes qui vivent dans le même logement, avec ou sans lien de parenté). Est-ce que vous habitez dans le même logement que Madame / Monsieur ....?

oui, 15 ans et plus ....1 ? → Question 4
oui, moins de 15 ans....................2 ? → Question 3
non..3 ? → Question 3

EDV:	Bitte Name und Vorname einblenden.

F03. Est-ce que je pourrais parler à Madame / Monsieur .... ou à quelqu'un qui habite le même logement que Madame / Monsieur...............?

oui...1 ? → Texte A
oui, à un autre moment..................2 ? → Rendez-vous
non...3 ? → Rendez-vous

INT:	Si „non“ fixer un rendez-vous pour un moment où il sera possible de joindre une personne qui pourra répondre aux questions
	ATTENTION! SI RENDEZ-VOUS STOPPER L'INTERVIEW!
EDV:	Bitte Name und Vorname einblenden.

F04. Vous habitez bien.....?

Rue, numéro: ___________________________
Appellation du bâtiment: __________________
NPA, localité: ___________________________
Nom du/de la logeur/se: ___________________

oui ....1 ? → Question 6
oui, petite erreur - rectification…….2 ? → Question 5
non...3 ? → Question 5

EDV:	Adresse 1 einblenden.

PF05. Un petit moment s'il vous plaît, je dois corriger votre adresse:

Rue, numéro: ___________________________
Appellation du bâtiment: ___________________
NPA, localité: ___________________________
Nom du/de la logeur/se:____________________ → Question 6

EDV:	Adresse 1 einblenden.

=

F06. A quel étage se situe votre logement?

Etage ....1 ?
Maison individuelle..................2 ?
Rez-de-chaussée....................3 ?
Entresol..................................4 ?
1er sous-sol.............................5 ?
2e sous-sol.........................6 ? → Question 7

F07. Combien de pièces compte votre logement?

Nombre de pièces ………….. → Question 8

INT:	Nombre de pièces = nombre de pièces habitables (sans les demi-pièces, la cuisine, le couloir, la salle de bain et autres pièces semblables, mais y compris les pièces d'habitation indépendantes)
EDV:	Max. 99 Zimmer

F08. Combien de personnes occupent ce logement, vous inclus/e**?**

Nombre de personnes…………→ Question 9

INT:	Egalement les sous-locataires, les navetteurs hebdomadaires, les colocataires, les enfants placés, etc.
EDV:	Max. 99 Personen

F9txt Pourriez-vous m'indiquer les noms et prénoms de toutes les personnes qui occupent le même logement que vous? IF (F08>1 AND FF01<3)

F9txt **Pourriez-vous m'indiquer les noms et prénoms de toutes les personnes qui occupent le logement, à commencer par vous?** IF (F08>1 AND FF01=3)

1re personne: Nom; prénom
2e personne: Nom; prénom
etc. → Question 10

INT:	Egalement les sous-locataires, les navetteurs hebdomadaires, les colocataires, les enfants placés, etc. (F09N, F09NA): Ne saisir d'abord que le NOM (Meyer) comme commentaire! (F09V, F09VA): Et saisir le PRENOM (Hans) comme commentaire!
EDV:	Plausib.: Anzahl der Personen muss mit Anzahl von Frage 8 übereinstimmen. Name und Vorname aus Adresse 1 einblenden, wenn Frage 1 mit ja beantwortet wurde.

F10. Avez-vous (ou un autre membre du ménage) un deuxième domicile? Une maison de vacances ou un appartement de vacances ne sont pas considérés comme deuxième domicile.

oui ....1 ? → Question 11
non...2 ? → Question 14

F11. Quels sont les membres du ménage concernés par un deuxième domicile?

Tous ....1 ?
1re personne: Nom; prénom..............2 ?
2e personne: Nom; prénom ....3 ? → Question 12
etc.

INT:	Sélectionner les personnes ayant un deuxième domicile
EDV:	Alle Personen der Liste aus Frage 9 einblenden. Werden alle Personen ausgewählt, ist dies gleichbedeutend wie der Code 1 „alle“.

PF12. Quelle est l'adresse de ce deuxième domicile de....... ?= ADRESSE 2

Nom: (apparaît à l'écran)
Prénom: (apparaît à l'écran)
Rue, numéro: ______________________________
Appellation du bâtiment: ______________________
NPA, localité: ______________________________
Nom du/de la logeur/se: ______________________
Etage: ____________________________________ → Question 13

INT:	(F12O) Localité! ATTENTION!! Si localité à l'ETRANGER ne saisir QUE LE PAYS.
EDV:	Loop: Frage 12 und 13 kommen in der Folge für alle Personen mit einem zweiten Wohnort, die bei Frage 11 ausgewählt wurden. Der Name und Vorname der betreffenden Person wird eingeblendet. Die Adresse 2 wird, sobald sie einmal notiert ist bei den folgenden Personen auch eingeblendet. Wird bei Frage 11 der Code 1 „Alle" gewählt, so wird anstatt der einzelnen Namen das Wort „allen" eingeblendet. Die TelefonistInnen müssen dann nur einmal die Adresse 2 notieren.

F13. A quel domicile .... vit-il/elle la plupart du temps, c'est-à-dire 4 jours ou plus par semaine?

EDV:	Adresse 1 und 2 einblenden und nacheinander alle diejenigen Personen, die bei Frage 11 ausgewählt wurden.

F14. Le dernier recensement de la population a eu lieu le 5 décembre 2000. Habitiez-vous (et les autres membres du ménage) déjà à la même adresse qu'aujourd'hui ?

oui, tous......................................1 ? → Question 17
non..2 ? → Question 15

EDV:	Bitte Adresse 1 einblenden.

F15.Qui habitait le 5 décembre 2000 à une autre adresse?

Tous ....1 ?
1re personne: Nom; prénom..............2 ?
2e personne: Nom; prénom ....3 ? → Question 16
etc.

EDV:	Alle Personen der Liste aus Frage 8 einblenden.

PF16. Quelle était l'adresse le 5 décembre 2000 de .... ? =

Nom: (apparaît à l'écran)
Prénom: (apparaît à l'écran)
Rue, numéro: ______________________________
Appellation du bâtiment: ______________________
NPA, localité: ______________________________
Nom du/de la logeur/se: ______________________
Etage: ____________________________________ → Question 17

INT:	(F16O) Localité! ATTENTION!! Si localité à l'ETRANGER ne saisir QUE LE PAYS.
EDV:	Loop: Frage 16 kommt in der Folge für alle Personen, die bei Frage 15 ausgewählt wurden. Der Name und Vorname der betreffenden Person wird eingeblendet. Als erste mögliche Adresse wird die Adresse 2 zur Auswahl eingeblendet. Die Adresse 3 wird, sobald sie einmal notiert ist bei den folgenden Personen einblendet. Wird bei Frage 15 der Code 1 „Alle" gewählt, so wird anstatt der einzelnen Namen das Wort „allen" eingeblendet. Die TelefonistInnen müssen dann nur einmal die Adresse 3 notieren.

F17. Aviez-vous (ou les autres membres du ménage) un deuxième domicile le 5 décembre 2000 ?

oui..1 ? → Question 18
non, personne.................................2 ? → Question 21

F18. Quels sont les membres du ménage concernés par un deuxième domicile?

Tous ....1 ?
1re personne: Nom; prénom..............2 ?
2e personne: Nom; prénom ....3 ? → Question 19
etc.

INT: Sélectionner les personnes qui avaient un deuxième domicile le 5 décembre 2000.

EDV: Alle Personen der Liste aus Frage 9 einblenden. Werden alle Personen ausgewählt, ist dies gleichbedeutend wie der Code 1 „alle".

PF19. Quelle était l'adresse le 5 décembre 2000 du deuxième domicile de .... ? **= ADRESSE 4**

Nom: (apparaît à l'écran)
Prénom: (apparaît à l'écran)
Rue, numéro: ________________________________
Appellation du bâtiment: ________________________
NPA, localité: ________________________________
Nom du/de la logeur/se: ________________________
Etage: ______________________________________→ Question 20

INT: (F19O) Localité! ATTENTION!! Si localité à l'ETRANGER ne saisir QUE LE PAYS.

EDV: Loop: Frage 19 und 20 kommt in der Folge für alle Personen, die bei Frage 18 ausgewählt wurden (Personen mit einem zweiten Wohnsitz am 5. Dez. 2000). Der Name und Vorname der betreffenden Personen wird eingeblendet. Als Voreinstellung die Adresse 1, 2 und 3 einblenden. Die Adresse 4 wird, sobald sie einmal notiert ist bei den folgenden Personen einblendet. Wird der Code 1 „Alle" gewählt, so wird anstatt der einzelnen Namen das Wort „allen" eingeblendet. Die TelefonistInnen müssen dann nur einmal die Adresse 4 notieren.

TF20. A quel domicile cette personne vivait-elle la plupart du temps le 5 décembre, c'est-à-dire 4 jours ou plus par semaine?

EDV: Adresse 3 und 4 einblenden und nacheinander diejenigen Personen, die bei Frage 18 ausgewählt wurden.

PF21. Je vais maintenant vous poser des questions qui vont se répéter pour chaque membre du ménage. Je vous prie de répondre tout d'abord pour vous, puis pour les autres membres du ménage.

F21. (indiquer le sexe)

féminin..1 ? → Question 22
masculin...2 ? → Question 22

INT: Ne pas citer!

EDV: Loop: Die Fragen 21 bis 31 werden nun für alle Personen im Haushalt gestellt, dazu nacheinander den Vornamen und den Namen der Hausbewohner einblenden (Liste der Frage 9).

KF22. Quelle est votre date de naissance?

___jour ___ mois ___ année → Question 23

F23. Quel est votre état civil?

célibataire 1 ?
marié/e // séparé/e 2 ?
veuf/ve 3 ?
divorcé/e 4 ? → Question 24

INT: Ne pas citer!

EDV: Getrennt wird gleich codiert wie verheiratet.

F24. Quelle est votre nationalité?

Suisse // Suisse avec double nationalité 1 ? → Question 27
Etranger/ère // Réfugié/e // Apatride // Plusieurs pays étrangers 2 ? → Question 25

F25. (Nationalité)

INT: NE PAS CITER!
Les personnes qui sont ressortissantes de plusieurs pays étrangers indiquent l'Etat dont elles ont obtenu la nationalité en dernier. Les apatrides et les réfugiés indiquent leur pays d'origine.
EDV: Frage 25 nur einblenden, wenn bei Frage 24 Code 2 angegeben wurde

Allemagne	1 ?
France	2 ?
Italie	3 ?
Croatie	4 ?
Autriche	5 ?
Portugal	6 ?
République de Yougoslavie	7 ?
République de Macédoine	8 ?
Espagne	9 ?
Turquie	10 ?
Autres:________________________	11 ? → Question 26

F26. Quel type d'autorisation de séjour / de statut avez-vous?

INT: Le type d'autorisation est désigné par une grande lettre sur le document d'autorisation.

EDV: Frage 26 nur einblenden, wenn bei Frage 24 Code 2 eingegeben wurde.

C = autorisation d'établissement	1 ?
B = autorisation de séjour annuel	2 ?
A = autorisation saisonnière	3 ?
N = requérant/e d'asile	4 ?
S = personne à protéger	5 ?
F = étranger/ère admis/e provisoirement	6 ?
L = autorisation de séjour de courte durée	7 ?
Autorisation du DFAE Département fédéral des affaires étrangères	8 ?
Autre statut:______________________	9 ? → Question 27

F27. Quelle est la langue dans laquelle vous pensez et que vous savez le mieux?

Allemand	1 ?
Français	2 ?
Italien	3 ?
Romanche	4 ?
Anglais	5 ?
Espagnol	6 ?
Portugais	7 ?
Turc	8 ?
Albanais	9 ?
Serbo-croate	10 ?
Autres:_________________________	11 ? → Question 28

TF28. Quelle est votre situation dans le ménage?

INT: Lire si nécessaire!
Ménage = toutes les personnes qui vivent dans un même logement
Chef de ménage = personne qui est responsable économiquement et socialement du ménage
Les membres d'un couple (avec ou sans enfants) sont tous deux chefs de ménage

LF28

Chef/fe de ménage	
Personne vivant seul/e	1 ?
Epoux/ épouse	2 ?
Personne vivant en union libre	3 ?
Personne élevant seule son/ses enfant/s	4 ?
Autre chef/fe de ménage	5 ?
Apparenté/e au chef de ménage	
Frère, soeur	6 ?
Fils, fille, beau-fils, belle-fille	7 ?
Père, mère, beau-père, belle-mère	8 ?
Autre/s parent/s du chef de ménage	9 ?
Autre situation dans le ménage	
Membre d'une communauté d'habitation / personne non apparentée partageant le logement	10 ?
Employé/e, garçon/fille au pair	11 ?
Locataire de chambre, sous locataire	12 ?
Autre/s personne/e vivant dans le ménage (p. ex. enfant placé, pensionnaire)	13 ?

→ Question 29

EDV: Filter: Frage 29 nur bei Personen über 15 Jahre einblenden (Prüfung mit Geburtsdatum Frage 22).

TF29. Ma prochaine question concerne votre situation ou vie active actuelle. Je vais d'abord vous définir ce que l'on entend par vie active et ensuite, j'aimerais que vous me disiez si l'un des exemples mentionnés correspond à votre situation ou non.

Les personnes actives sont:

- **les personnes qui travaillent une heure ou plus par semaine contre rémunération ou**
- **les personnes qui travaillent dans l'entreprise familiale sans rémunération ou**
- **les personnes qui sont actuellement en congé maladie, en congé maternité payé ou au service militaire, mais qui sont habituellement actives.**

Les personnes qui font des petits jobs occasionnels sont également considérées comme actives.

F29 Est-ce que l'un de ces exemples de personnes actives correspond à votre situation?

oui..1 ? → Question 30
non...2 ? → Question 31

F30. Avez-vous...

INT: LIRE LES POSSIBILITES!

Une activité professionnelle à plein temps (une seule occupation) 1 ?
Une activité professionnelle à temps partiel (au min. 1 heure par semaine, une seule occupation) 2 ?
Plusieurs activités professionnelles à temps partiel (plusieurs occupations) 3 ?
Un poste d'apprenti/e 4 ? → Question 32

EDV: Weiter zur nächsten Person oder Interview beenden mit Frage 32

F31. Etes-vous...

INT: LIRE LES POSSIBILITES!

LF31

Au chômage...1 ❑
Non occupé/e, mais en quête d'un emploi.........................2 ❑
Non occupé/e, mais futur emploi garanti ....3 ❑
Ni occupé/e, ni en quête d'emploi ....4 ❑
Femme / homme au foyer (travaux dans son propre ménage)...5 ❑
En formation (écolier/ière, étudiant/e)...............................6 ❑
Retraité/e (rente de vieillesse, d'invalidité, etc.) .... 7 ❑

EDV: Weiter zur nächsten Person oder Interview beenden mit Frage 32

TABSCH Nous sommes parvenus à la fin de notre interview. Je vous remercie de votre précieuse collaboration.

BEMER Bemerkungen?

ja,_______________________...1 ❑
nein...2 ❑

Appendix D

More about Matching

For additional information about matching between the P-sample and the census; see Section 10.1.

Matching Process

The match code `match` is the result of the matching phase processed by the census staff; see Tables D.1 for non-movers and D.2 for movers.

The data set contains multiple entries for unresolved multiple matches (indicated as "provisional" in the tables). Therefore, the total size of 50,266 is larger than the total P-sample sample size of 50,070. The phases in the census data based are noted SQL and the phases in SAS are noted SAS1 to SAS4.

The codes `match` can be aggregated into following groups:

- confirmed match (`match in (101-116, 121-145, 157, 161, 163, 201-203, 211, 221-233)`): 49,238 (98.0%)
- confirmed non-match (`match in (117, 158-159, 162, 164, 212-213, 234)`): 807 (1.6%)
- unresolved cases (`match in (152-154, 156, 204-206, 999)`): 217 (0.4%);
- cases that have to be excluded from the P-sample (*e.g.* doubles) (`match in (151, 155, 160)`): 4.

Table D.1: Match code `match` for non-movers.

`match`	Description	Groups	Nb
SQL in the census data base			
101	match SQL	match	44,519
102	match SQL	match	198
103	match SQL	match	104
104	match SQL, checked	match	212
105	match SQL, multiple	match	86
SAS for non match in SQL			
111	match SAS1	match	73
112	match SAS2	match	60
113	match SAS3	match	54
114	match SAS4	match	69
115	match SAS4	match	2
116	match SAS4	match	41
117	non match	non match	618
Matched in SQL but not right area			
121	match SAS1 confirmed	match	18
122	match SAS1 refused, SQL ok	match	87
123	match SAS2 confirmed	match	2
124	match SAS2 refused, SQL ok	match	1
125	non match in SAS, SQL ok	match	727
Matched in SQL in area but not PSU and not target population			
131	match SAS1 confirmed	match	86
132	match SAS1 refused, SQL ok	match	3
133	same match SAS2 and SQL	match	16
134	non match in SAS, SQL ok	match	15
Matched in SQL in PSU but not target population			
141	match SAS1 confirmed	match	1000
142	match SAS1 refused, SQL ok	match	27
143	match SAS2 confirmed	match	121
144	match SAS1 refused, SQL ok	match	1
145	non match in SAS, SQL ok	match	191
Special cases			
151	double entry in P-sample	excluded	1
152	multiple match	provisional	120
153	match but out of population	provisional	29
154	multiple match	provisional	18
155	double entry in P-sample	excluded	2
156	non match	provisional	8
157	match SAS1	match	9
158	match SAS1 refused	non match	7
159	non match, multiple	non match	69
160	double entry in P-sample	excluded	1
161	match SAS	match	19
162	match SAS refused	non match	3
163	same clerical match and SQL	match	4
164	non match	non match	38
Total			48,659

Table D.2: Match code `match` for movers.

`match`	Description	Groups	Nb
SQL in the census data base			
201	match SQL	match	1152
202	match SQL confirmed	match	1
203	match SQL confirmed	match	2
204	match SQL, multiple	provisional	36
205	match SQL, multiple	provisional	1
206	match SQL, multiple	provisional	3
999	match SQL, multiple	provisional	2
SAS for non match in SQL			
211	match SAS1	match	14
212	match SAS1 refused	non match	4
213	non match SAS	non match	67
Matched in SQL but not commune			
221	match SAS1	match	55
222	match SAS1 refused, SQL ok	match	28
223	non match in SAS, SQL ok	match	191
Matched in SQL in commune but not target population			
231	match SAS1	match	37
232	match SAS1 refused, SQL ok	match	2
233	non match in SAS, SQL ok	match	11
Special cases			
234	non match in SAS, SQL refused	non match	1
Total			1607

Final Matching Codes

New steps are applied to get the final matching codes. These steps include information from special cases and supplementary checks:

1. two P-sample people matched to one census entry (`match_cont in (1001-1004)`);
2. some special changes (`match_cont in (1005-1007)`);
3. one P-sample person matched with two or three census entries (multiple matches, (`match_cont in (1020-1023)`);
4. matches refused during the clerical checks and processed in a second phase (`match_cont in (1030-1033)`);
5. non-matched entries checked in order to detect people that should be excluded from the P-sample (`match_cont= 1040` and `match_cont2 in (2040-2042)`).

The match codes `match_cont` and `match_cont2` are complementary codes that have priority over the code `match`; see Table D.3.

P-sample people with codes `match in (151, 155, 160)` or `match_cont in (1002, 1005)` or `match_cont2 in (2041, 2042)` are excluded from the provisional P-sample (184 cases).

The final status of match `matchG` is equal to 0 for P-sample people out of the population, 10 for matches and 20 for non-matches; see Table D.4.

The identifier of the match `Vzid` in the census data set is `vz_pers_id` for SQL matches, `vz_pers_id_mac_p1 - p3` for SAS matches phase 1-3 and `vz_pers_cont` for the complementary matches.

Table D.3: Complementary match codes `match_cont` and `match_cont2`.

`match_cont`	Description	Number
Two P-sample entries with one unique match		
1001	match	195
1002	P-sample people is double	170
1003	match with another entry	15
Special cases		
1004	non-match	1
1005	P-sample people is double or born on 5 Dec. 2000	6
1006	match with another entry	4
1007	non match	4
Multiple matches		
1020	match	21
1021	non match, double in census	6
1022	non match, partner in census	13
1023	non match, other	2
Matched refused and treated again		
1031	non match but in P-sample	15
1032	match	110
1033	non match, double or partner in census	77
Non-matched entries to be checked for existence		
1040	non match	692
Total		1331

`match_cont2`	Description	Number
2040	in P-sample	689
2041	not in P-sample (non economic domicile)	3
2042	not in P-sample (collective household)	1
Total		693

Table D.4: Final match status `matchG` as a function of `match`, `match_cont` and `match_cont2`. The identifier of the match in the census data set is given in `Vzid`.

Status	`matchG`	`match_cont`	`match_cont2`	`match`	`Vzid`	Number
out pop	0	.	.	151, 155, 160		4
	0	1002	.	...		170
	0	1005	.	...		6
	0	1040	2041 - 2042	...		4
match	10	.	.	101-105, 114-116, 122,		
		.	.	124-125, 132-134, 142,		
				144-145, 163, 201-203,		
				222-223, 232-233	`vz_pers_id`	47,310
	10	.	.	111, 211, 221, 231	`vz_pers_id_mac_p1`	177
	10	.	.	112, 121, 123, 131,		
		.	.	141, 143, 157, 161	`vz_pers_id_mac_p2`	1299
	10	.	.	113	`vz_pers_id_mac_p3`	51
	10	1001, 1003, 1006,	.	...		
		1020, 1032	.	...	`vz_pers_cont`	345
non-match	20	1004	.	101		1
	20	1031	.	...		15
	20	1040	2040	...		688
Total						50,070

Appendix E

More about Variance Estimation

First Comparisons

Tables E.1 and E.2 show a comparison between variance estimation methodologies for the simple rate of correct enumeration $\widehat{R}_{ce}^{(s)}$ and simple rate of correct match $\widehat{R}_{m}^{(s)}$; see Section 6.1.

The notation is:

- n: number of persons;
- $\widehat{R}_{ce}$: estimated rate of correct enumeration;
- $\widehat{R}_{m}$: estimated rate of match;
- std_L: standard error for Taylor expansion without finite population correction (*fpc*);
- $std_L(tot)$: standard error for Taylor expansion with finite population correction;
- std_{JK1}: standard error for classical jackknife;
- std_{JKS}: standard error for stratified jackknife;
- $D_x = (std_x - std_{JKS})/std_{JKS}$, with $x \in \{L, L(tot), JK1\}$ [%]: difference relative to stratified jackknife.

Taylor expansion estimates are smaller than stratified jackknife ($D_L < 0$ and $D_{L(tot)} < 0$). Larger differences are observed when including the finite population correction (fpc) ($|D_{L(tot)}| > |D_L|$). However, the estimate with fpc $std_L(tot)$ is probably unstable because (1) it takes into account only the first sampling stage and (2) some strata have a large fpc. The large difference for Ticino (`NUTS=7`, `ling2=3` and `var2=5`) has to do with the small number of PSUs in the sampling strata.

The stratification of jackknife has a non-negligible effect on standard error, especially in some subgroups (*e.g.* `outsour=2`). The result is sometimes larger and sometimes smaller than the classical jackknife. We observe that the differences are generally larger for $\widehat{R}_{ce}$ than $\widehat{R}_{m}$ with an extreme value corresponding to a relative difference of -9%.

Table E.1: Variance estimates of $\widehat{R}_{ce}^{(s)}$, with n the number of persons. The text contains a definition of standard errors and differences in relation to the stratified jackknife.

Variable		n	$\widehat{R}_{ce}^{(s)}$	std_L	$std_L(tot)$	std_{JK1}	std_{JKS}	D_L	$D_{L(tot)}$	D_{JK1}
Overall		55375	0.99601	0.000299	0.000285	0.000303	0.000299	-0.16	-4.83	1.19
`sex`	1	27374	0.99586	0.000393	0.000374	0.000394	0.000394	-0.19	-5.01	0.10
	2	28001	0.99617	0.000324	0.000307	0.000334	0.000324	-0.11	-5.35	2.86
`Cage2`	1	6449	0.99742	0.000503	0.000476	0.000503	0.000504	-0.12	-5.48	-0.21
	2	6689	0.99659	0.000579	0.000551	0.000575	0.000579	-0.05	-4.88	-0.78
	3	8652	0.99028	0.000929	0.000876	0.000952	0.000930	-0.14	-5.84	2.38
	4	12090	0.99651	0.000515	0.000493	0.000519	0.000516	-0.15	-4.42	0.64
	5	10902	0.99745	0.000449	0.000424	0.000449	0.000450	-0.22	-5.77	-0.20
	6	8802	0.9988	0.000286	0.00027	0.000286	0.000287	-0.18	-5.76	-0.01
	7	1791	0.99218	0.001826	0.00174	0.001817	0.001830	-0.23	-4.93	-0.72
`ausw2`	1	45550	0.99613	0.000316	0.000302	0.000321	0.000317	-0.21	-4.63	1.45
	2	6851	0.99629	0.000642	0.000603	0.000641	0.000644	-0.27	-6.32	-0.47
	3	2974	0.99385	0.001114	0.001055	0.001105	0.001123	-0.83	-6.08	-1.60
`ziv2`	1	23515	0.99441	0.000468	0.000443	0.000467	0.000469	-0.18	-5.51	-0.36
	2	26040	0.99761	0.000401	0.000387	0.000405	0.000402	-0.16	-3.64	0.86
	3	5820	0.99556	0.000696	0.000657	0.000721	0.000697	-0.10	-5.70	3.49
`ling2`	1	36706	0.99611	0.00038	0.000366	0.000384	0.000380	-0.11	-3.79	0.96
	2	16473	0.99612	0.000527	0.000501	0.000539	0.000535	-1.57	-6.43	0.63
	3	2196	0.9944	0.001256	0.001076	0.001283	0.001262	-0.44	-14.71	1.67
`NUTS`	1	10901	0.99591	0.000633	0.000599	0.000660	0.000647	-2.14	-7.39	2.06
	2	16039	0.99592	0.000864	0.000837	0.000884	0.000874	-1.20	-4.28	1.06
	3	6592	0.99732	0.000399	0.000378	0.000411	0.000413	-3.43	-8.51	-0.51
	4	8813	0.99646	0.000518	0.000489	0.000529	0.000530	-2.29	-7.76	-0.24
	5	7856	0.99546	0.000694	0.000672	0.000718	0.000717	-3.14	-6.22	0.16
	6	3478	0.99587	0.000704	0.000671	0.000754	0.000767	-8.20	-12.50	-1.62
	7	1696	0.99438	0.001288	0.001103	0.001331	0.001293	-0.39	-14.70	2.96
`taipop2`	1	18668	0.99632	0.000557	0.00054	0.000566	0.000563	-1.06	-4.08	0.54
	2	17013	0.99541	0.000682	0.000663	0.000704	0.000686	-0.59	-3.36	2.69
	3	19694	0.99635	0.000312	0.000279	0.000311	0.000313	-0.25	-10.80	-0.62
`urbrur2`	1	12882	0.9958	0.000415	0.000366	0.000411	0.000420	-1.14	-12.82	-2.15
	2	20733	0.99598	0.000585	0.000566	0.000601	0.000589	-0.69	-3.92	2.10
	4	21760	0.99629	0.000434	0.000422	0.000437	0.000436	-0.45	-3.21	0.21
`var2`	1	11000	0.99578	0.000553	0.000508	0.000568	0.000561	-1.47	-9.49	1.25
	2	5298	0.99598	0.000779	0.00074	0.000784	0.000781	-0.22	-5.22	0.46
	3	37381	0.99613	0.000324	0.000311	0.000332	0.000324	-0.04	-4.05	2.39
	5	1696	0.99438	0.001288	0.001103	0.001331	0.001293	-0.39	-14.70	2.96
`outsour`	0	13548	0.9949	0.000688	0.000596	0.000632	0.000696	-1.12	-14.34	-9.15
	1	35599	0.99626	0.000333	0.000319	0.000334	0.000333	-0.01	-4.22	0.22
	2	6228	0.99448	0.00154	0.00151	0.001797	0.001680	-8.33	-10.12	6.95

Table E.2: Variance estimates of $\widehat{R}_m^{(s)}$, with n the number of persons. The text contains a definition of standard errors and differences in relation to the stratified jackknife.

Variable		n	$\widehat{R}_m^{(s)}$	std_L	$std_L(tot)$	std_{JK1}	std_{JKS}	D_L	$D_{L(tot)}$	D_{JK1}
Overall		49883	0.98359	0.00105	0.00098	0.001071	0.001051	-0.08	-6.74	1.94
`sex`	1	25319	0.9845	0.00103	0.000962	0.001047	0.001030	-0.05	-6.65	1.62
	2	24564	0.98265	0.001324	0.001238	0.001344	0.001326	-0.19	-6.67	1.34
`Cage2`	1	5957	0.98537	0.00209	0.001965	0.002087	0.002096	-0.27	-6.24	-0.42
	2	6189	0.987	0.001935	0.001817	0.001929	0.001936	-0.07	-6.17	-0.37
	3	7339	0.96504	0.00335	0.003153	0.003328	0.003371	-0.62	-6.47	-1.28
	4	10826	0.98349	0.001611	0.001492	0.001669	0.001612	-0.07	-7.45	3.55
	5	10303	0.98821	0.001381	0.001312	0.001363	0.001382	-0.11	-5.10	-1.41
	6	7879	0.99095	0.001349	0.001296	0.001406	0.001349	-0.03	-3.95	4.19
	7	1390	0.98803	0.003139	0.002962	0.003124	0.003148	-0.27	-5.89	-0.74
`ausw2`	1	42629	0.98721	0.000942	0.000891	0.000964	0.000943	-0.11	-5.52	2.17
	2	5443	0.98147	0.002921	0.002714	0.003046	0.002937	-0.53	-7.58	3.72
	3	1811	0.91975	0.008485	0.007949	0.008508	0.008526	-0.49	-6.77	-0.22
`ziv2`	1	20814	0.97926	0.001757	0.001648	0.001789	0.001761	-0.24	-6.43	1.55
	2	24207	0.98733	0.001062	0.000993	0.001054	0.001063	-0.14	-6.63	-0.85
	3	4862	0.98401	0.002008	0.001906	0.002000	0.002010	-0.10	-5.17	-0.48
`ling2`	1	33724	0.98496	0.001104	0.001053	0.001158	0.001107	-0.24	-4.85	4.68
	2	14177	0.98113	0.002443	0.00225	0.002536	0.002505	-2.46	-10.17	1.26
	3	1982	0.9765	0.004964	0.003704	0.005137	0.004945	0.39	-25.09	3.88
`NUTS`	1	9486	0.97814	0.003699	0.003469	0.003778	0.003812	-2.97	-9.01	-0.89
	2	13870	0.98606	0.001426	0.001351	0.001453	0.001453	-1.88	-7.04	-0.05
	3	6056	0.98496	0.002594	0.002429	0.002683	0.002674	-3.00	-9.17	0.32
	4	8835	0.98423	0.001877	0.001769	0.002022	0.001922	-2.36	-7.98	5.18
	5	6935	0.98706	0.002293	0.002197	0.002358	0.002346	-2.28	-6.37	0.49
	6	3150	0.98433	0.002363	0.002208	0.002620	0.002522	-6.29	-12.44	3.92
	7	1551	0.97624	0.00518	0.003863	0.005395	0.005158	0.42	-25.11	4.58
`taipop2`	1	18306	0.985	0.001488	0.001437	0.001498	0.001502	-0.93	-4.32	-0.24
	2	15845	0.98676	0.001634	0.00159	0.001668	0.001642	-0.49	-3.17	1.59
	3	15732	0.97994	0.001875	0.001693	0.001883	0.001884	-0.46	-10.13	-0.02
`urbrur2`	1	10295	0.98036	0.001624	0.001438	0.001663	0.001660	-2.18	-13.38	0.20
	2	18295	0.98508	0.001868	0.001731	0.001912	0.001890	-1.19	-8.43	1.16
	4	21293	0.98441	0.001678	0.001632	0.001722	0.001682	-0.26	-3.00	2.36
`var2`	1	8694	0.98093	0.002815	0.002548	0.002775	0.002815	-0.01	-9.50	-1.43
	2	4940	0.98934	0.002294	0.002146	0.002597	0.002440	-6.00	-12.06	6.43
	3	34698	0.98381	0.001114	0.001047	0.001138	0.001115	-0.11	-6.12	2.05
	5	1551	0.97624	0.00518	0.003863	0.005395	0.005158	0.42	-25.11	4.58
`outsour`	0	10487	0.97933	0.002788	0.002216	0.002754	0.002794	-0.21	-20.68	-1.42
	1	33784	0.98392	0.001121	0.001051	0.001157	0.001122	-0.11	-6.35	3.12
	2	5612	0.98462	0.003905	0.003783	0.004341	0.004209	-7.22	-10.12	3.14

Splitting of PSUs

Splitting of PSUs is tested for the P-sample and E-sample.

PSUs are selected in two phases. First, we select the set of strata `stradap` with a PSU sampling rate larger than 20%. Second, we select the PSUs from this set with at least 300 elements. The selection leads to 43 PSUs in the P-sample and 45 PSUs in the E-sample; all in the same 5 strata. Splitting is applied randomly into two fictitious equal-sized PSUs. This gives us 303+43=346 PSUs for the P-sample and 303+45=348 PSUs for the E-sample.

Splitting has a slight impact for most of the subgroups (between -1% and 1%). However, splitting clearly increases the standard error of categories related the Ticino. The effect is larger for classical jackknife (4%) than stratified jackknife (1.4%).

Stratified jackknife is expected to be more stable than classical jackknife in subgroups of the population observed in only few PSUs.

Alternative Correction of the Weights in Stratified Jackknife

Stratified jackknife with a weight correction that depends on the weights (and not only on the number of PSUs in the stratum) leads to a decrease in estimated variance.

The relative difference to reference stratified jackknife ranges between -0.1% and -23.7% with an overall value of -2.3%. Larger differences are observed for the categories relating mainly to Ticino and `outsour=0`.

The reference stratified jackknife estimate is more conservative than the estimate with the alternative correction. This is due to the added variability of the stratum size.

Checking P-sample Weights

The P-sample weights $w_{p,j}$ are checked for extremely influential elements; see Section 12.1.

The minimum $\min$, the maximum $\max$, the range $R = \max - \min$, the coefficient of variation CV, the first quartile $Q1$, the third quartile $Q3$ and the interquartile $IQR = Q3 - Q1$ are used for the tests. Note however that the statistics based on the quartiles or the CV are not reliable in many strata or PSUs because the weights have only few different values (*e.g.* 87 units with weight 10, 134 units with weight 34 and 3 units with weight 198).

The results of the checks on the P-sample may be summarized by:

- Overall, $w_{p,j}$ varies between $\min = 5.2$ and $\max = 489.2$ with a coefficient of variation of 56% and the interquartile $IQR = 155.7$. The 290 weights larger than $Q3 + 1.5\,IQR$ are grouped in `stradap=19`, that contains weights between 310.3 and 489.2. Therefore, these overall extreme values are not extreme at the stratum level.
- A group of three PSUs have large variability between the weights: `PSU=N461300` in `stradap=17`, `PSU=N541300` in `stradap=15` and `PSU=N811200` in `stradap=17`. Excepted `PSU=N461300`, the extreme values within the PSUs are not extreme at the stratum level.

- We detect 4 strata with possibly very influential elements. They are grouped in one particular PSU in each stratum: `PSU=N507400` for `stradap=9` (10 cases in 130), `PSU=N192300` for `stradap=11` (3 cases in 31), `PSU=N461300` for `stradap=17` (73 cases in 112) and `PSU=T5150` for `stradap=102` (3 cases in 40).

Based on the checks, we define the trimmed weights `weiPo`= $w_{p,j}^{(t)}$. With $w_{p,j}^{(t)} \neq w_{p,j}$ for 89 P-sample elements `weiPo=weiP;`

```
if stradap=11 and PSU='N192300' and weiP>80 then weiPo=50;
else if stradap=9 and PSU='N507400' and weiP<6 then weiPo=30;
else if stradap=17 and PSU='N461300' and weiP>370 then weiPo=300;
else if stradap=102 and PSU='T5150' and weiP>300 then weiPo=160;.
```

The effect of weight correction is negligible for the $\widehat{R}_m$ and standard error estimates std_L, std_{JK1} and std_{JKS} (relative absolute differences smaller than 0.4%).

Although weight variability was reduced, we unexpectedly observed a small increase in variance as a whole and in most of the subgroups (maximum 0.35%).

As a result, correction of the most influential weights has no positive impact on the estimates.

General Remarks about $\widehat{R}_{ce}$ and $\widehat{R}_m$

Based on the results from trimming weights, alternative weight correction and splitting PSUs, we can consider the reference stratified jackknife as a somewhat conservative and reliable estimate. This methodology is used for the results presented in Chapters 11 and 12.

Remark about $\widehat{R}_{net}$

The classical jackknife estimated standard errors of $\widehat{R}_{net}$ are 2-7% larger than the given stratified jackknife. The larger difference is observed for `ziv=4` (7.2%). We also have differences that exceed 6% for the French speaking region `NUTS=1` and corresponding language `ling2=2` and for the Italian speaking region `NUTS=7` and corresponding language `ling2=3` and census methodology `var2=5`.

Appendix F

Detailed Results for Estimation Cells

List of the 121 estimation cells (post-strata) with:

- Post-stratum ID: *e.g.* `A1Z1T 3L12C 67S12` for `ausw3=1` (Swiss), `ziv3=1` (single), `taipop3=3`, `ling3=1` or 2 (all languages), `Cage2=6 or 7` (60 and older), `sex=1 or 2` (male and female);
- `C`$= C^{(pop)}$: census count (target population);
- `Np`$= n_p$: number of elements in the P-sample;
- `M`$= \sum P^{(s)}_{m,j}$: number of simple matches;
- `CVp`: variation coefficient of the weights $w_{p,j}$ (P-sample) [%];
- `rateM`$= \widehat{R}^{(s)}_{m}$: weighted rate of simple match [%];
- `Ne`$= n_e$: number of elements in the E-sample;
- `CE`$= \sum P^{(pop)}_{ce,j}$: number of correct enumerations (multiple entries in the population);
- `CVe`: variation coefficient of the weight $w_{e,i}$ (E-sample) [%];
- `rateCE`$= \widehat{R}^{(popR)}_{ce}$: weighted rate of correct enumeration [%];
- `Rnet=rateM/rateCE`$= \widehat{R}_{net}$: rate of net coverage [%];
- `Run=1-Rnet`$= \widehat{R}_{under}$: rate of net undercoverage [%];
- `Rnetse`: standard error of `Rnet`$= \widehat{R}_{net}$ and `Run`$= \widehat{R}_{under}$ [%].

Obs	poststra		C	Np	M	CVp	rateM	Ne	CE	CVe	rateCE	Rnet	Run	Rnetse
1	A1Z1T 1L 1C	1S 1	49572	720	712	96.18	98.93	662	661.5	95.88	99.81	99.12	0.88	0.60
2	A1Z1T 1L 1C	1S 2	47662	655	649	96.87	98.38	663	662.5	93.14	99.98	98.40	1.60	0.79
3	A1Z1T 1L 1C	2S 1	54334	774	769	97.20	99.46	824	823.0	97.83	99.82	99.64	0.36	0.39
4	A1Z1T 1L 1C	2S 2	50819	698	692	100.7	99.68	695	694.0	96.47	99.81	99.88	0.12	0.23
5	A1Z1T 1L 1C	3S 1	40543	535	521	92.82	97.67	529	524.5	94.22	99.17	98.49	1.51	0.88
6	A1Z1T 1L 1C	3S 2	29122	389	379	98.17	97.70	435	429.0	97.31	98.85	98.84	1.16	1.37
7	A1Z1T 1L 1C	45S12	40867	503	497	93.52	98.63	533	528.0	96.54	99.30	99.32	0.68	0.76
8	A1Z1T 1L 2C	1S 1	31677	461	456	97.66	98.63	445	443.0	115.3	99.87	98.76	1.24	0.98
9	A1Z1T 1L 2C	1S 2	30072	401	400	94.68	99.93	424	423.0	114.2	99.92	100.0	-0.01	0.05
10	A1Z1T 1L 2C	2S 1	30265	455	447	95.00	98.68	400	398.0	108.6	99.64	99.04	0.96	0.69
11	A1Z1T 1L 2C	2S 2	28927	408	401	97.16	98.15	396	395.0	116.7	99.92	98.23	1.77	0.88
12	A1Z1T 1L 2C	3S 1	23693	293	285	98.72	96.85	374	367.8	111.8	97.88	98.95	1.05	1.93
13	A1Z1T 1L 2C	3S 2	17917	214	207	94.96	96.32	211	208.5	103.4	99.20	97.10	2.90	1.85
14	A1Z1T 1L 2C	45S12	25073	296	286	93.22	97.40	369	367.5	108.1	99.60	97.79	2.21	1.32
15	A1Z1T 1L12C	67S12	17960	205	201	97.42	97.83	233	231.5	103.4	99.41	98.40	1.60	1.46
16	A1Z1T 2L 1C	1S 1	91105	595	588	46.09	98.88	571	567.5	40.03	99.49	99.38	0.62	0.46
17	A1Z1T 2L 1C	1S 2	86778	556	553	39.71	99.57	526	524.5	40.19	99.67	99.90	0.10	0.35
18	A1Z1T 2L 1C	2S 1	102328	600	595	40.54	99.48	620	617.0	39.14	99.55	99.93	0.07	0.39
19	A1Z1T 2L 1C	2S 2	96688	575	570	39.72	99.22	551	549.5	45.97	99.67	99.55	0.45	0.51
20	A1Z1T 2L 1C	3S 1	87591	546	534	43.30	97.36	569	562.5	39.50	98.78	98.56	1.44	1.18
21	A1Z1T 2L 1C	3S 2	69504	419	411	42.02	97.90	408	403.0	41.29	98.67	99.22	0.78	1.31
22	A1Z1T 2L 1C	45S12	90529	525	515	42.56	97.90	560	558.5	38.23	99.71	98.19	1.81	0.78
23	A1Z1T 2L 2C	1S 1	26310	257	254	49.15	98.63	273	271.5	55.20	99.51	99.11	0.89	0.88
24	A1Z1T 2L 2C	1S 2	24731	250	247	50.74	99.28	245	245.0	61.20	100.0	99.28	0.72	0.55
25	A1Z1T 2L 2C	2S 1	25886	285	284	53.04	99.57	301	300.0	55.88	99.57	100.0	0.00	0.55
26	A1Z1T 2L 2C	2S 2	25352	250	247	48.79	99.31	314	313.5	55.62	99.82	99.49	0.51	0.53
27	A1Z1T 2L 2C	3S 1	23543	232	225	57.16	96.78	272	268.5	57.98	98.77	97.99	2.01	1.90
28	A1Z1T 2L 2C	3S 2	19528	188	183	50.85	97.53	215	212.0	57.17	98.23	99.29	0.71	1.41
29	A1Z1T 2L 2C	45S12	23916	179	174	50.77	97.19	245	243.0	59.17	98.83	98.34	1.66	1.52

30	A1Z1T 2L12C	67S12	23680	134	133	50.13	97.93	185	185.0	54.75	100.0	97.93	2.07	2.08
31	A1Z1T 3L 1C	1S 1	84547	368	359	13.99	97.79	458	456.0	12.23	99.55	98.23	1.77	0.70
32	A1Z1T 3L 1C	1S 2	80335	379	374	14.26	98.56	470	469.0	12.06	99.79	98.77	1.23	0.68
33	A1Z1T 3L 1C	2S 1	93261	474	470	17.22	99.24	483	481.5	10.56	99.74	99.50	0.50	0.42
34	A1Z1T 3L 1C	2S 2	91029	450	444	15.48	98.85	480	479.5	9.33	99.89	98.96	1.04	0.52
35	A1Z1T 3L 1C	3S 1	127403	610	590	16.36	96.60	765	760.5	11.28	99.40	97.18	2.82	0.90
36	A1Z1T 3L 1C	3S 2	112746	603	584	14.58	97.08	691	685.5	9.63	99.24	97.82	2.18	0.71
37	A1Z1T 3L 1C	4S 1	68752	348	345	14.91	99.13	455	453.0	9.20	99.56	99.57	0.43	0.50
38	A1Z1T 3L 1C	4S 2	50750	269	265	11.76	98.40	332	330.5	11.48	99.54	98.86	1.14	1.04
39	A1Z1T 3L 1C	5S12	50572	255	249	7.13	97.66	336	335.0	9.79	99.70	97.96	2.04	1.13
40	A1Z1T 3L 2C	1S 1	29131	186	185	12.81	99.45	151	151.0	18.97	100.0	99.45	0.55	0.59
41	A1Z1T 3L 2C	1S 2	28267	166	164	19.31	98.68	178	177.5	17.79	99.56	99.12	0.88	1.19
42	A1Z1T 3L 2C	2S 1	28604	166	162	14.42	97.37	226	226.0	16.71	100.0	97.37	2.63	1.69
43	A1Z1T 3L 2C	2S 2	28573	172	169	16.98	98.41	224	223.5	18.80	99.80	98.61	1.39	1.05
44	A1Z1T 3L 2C	3S 1	39157	184	175	20.50	95.32	257	253.0	17.05	98.31	96.96	3.04	1.47
45	A1Z1T 3L 2C	3S 2	36997	190	179	21.75	94.78	246	243.5	18.29	99.07	95.67	4.33	2.24
46	A1Z1T 3L 2C	45S12	43793	233	227	22.18	97.61	278	277.0	22.17	99.60	98.00	2.00	1.13
47	A1Z1T 3L12C	67S12	48429	226	222	17.23	97.92	330	329.5	16.43	99.85	98.07	1.93	1.03
48	A1Z2T 1L 1C	4S 1	63284	916	912	91.70	99.23	891	887.5	95.87	99.85	99.38	0.62	0.42
49	A1Z2T 1L 1C	4S 2	72605	1045	1038	93.78	98.87	1052	1050.0	95.57	99.83	99.04	0.96	0.52
50	A1Z2T 1L 1C	5S 1	71513	1035	1029	97.55	99.09	1018	1017.5	95.06	99.99	99.10	0.90	0.42
51	A1Z2T 1L 1C	5S 2	69582	1042	1038	98.03	99.51	1073	1071.5	95.73	99.96	99.55	0.45	0.29
52	A1Z2T 1L 1C	6S 1	51077	738	730	94.24	98.91	705	705.0	97.99	100.0	98.91	1.09	0.50
53	A1Z2T 1L 1C	6S 2	58882	828	824	93.94	99.38	834	834.0	92.96	100.0	99.38	0.62	0.37
54	A1Z2T 1L 1C	7S12	21957	266	261	98.39	97.12	292	292.0	95.50	100.0	97.12	2.88	1.49
55	A1Z2T 1L 1C	123S12	26693	368	363	98.93	98.99	374	371.5	90.30	99.56	99.42	0.58	0.76
56	A1Z2T 1L 2C	4S 1	37605	522	519	96.27	99.84	512	509.8	114.6	99.81	100.0	-0.03	0.16
57	A1Z2T 1L 2C	4S 2	43688	587	584	94.83	99.56	615	613.5	110.6	99.60	99.96	0.04	0.52
58	A1Z2T 1L 2C	5S 1	44236	604	599	96.44	99.16	594	592.5	118.9	99.71	99.44	0.56	0.41
59	A1Z2T 1L 2C	5S 2	46522	626	623	95.41	99.69	586	585.0	109.2	99.78	99.91	0.09	0.29
60	A1Z2T 1L 2C	6S 1	33793	470	464	93.90	97.59	448	447.0	118.3	99.94	97.65	2.35	1.17
61	A1Z2T 1L 2C	6S 2	40054	561	550	90.74	98.41	542	542.0	113.7	100.0	98.41	1.59	0.84
62	A1Z2T 1L 2C	7S12	16170	164	163	96.43	99.60	207	206.5	109.1	99.38	100.2	-0.22	0.76
63	A1Z2T 1L 2C	23S12	19317	242	235	99.99	97.41	249	246.0	123.1	98.73	98.66	1.34	1.84
64	A1Z2T 2L 1C	4S 1	119927	752	748	43.47	99.42	751	747.5	41.25	99.59	99.82	0.18	0.36
65	A1Z2T 2L 1C	4S 2	140070	855	850	43.10	99.28	833	829.5	38.58	99.58	99.69	0.31	0.41
66	A1Z2T 2L 1C	5S 1	147892	884	879	40.53	99.45	869	866.5	38.30	99.63	99.81	0.19	0.30
67	A1Z2T 2L 1C	5S 2	152611	895	891	37.11	99.38	826	825.0	40.38	99.81	99.57	0.43	0.35
68	A1Z2T 2L 1C	6S 1	108093	658	655	46.05	99.50	680	678.5	41.44	99.81	99.69	0.31	0.31
69	A1Z2T 2L 1C	6S 2	129011	794	789	44.48	99.56	750	749.0	37.55	99.92	99.64	0.36	0.28
70	A1Z2T 2L 1C	7S12	46311	196	193	37.17	99.27	296	295.5	41.63	99.81	99.46	0.54	0.59
71	A1Z2T 2L 1C	123S12	50154	320	315	41.69	98.54	307	303.0	39.62	98.87	99.66	0.34	1.27
72	A1Z2T 2L 2C	4S 1	31296	328	327	54.84	99.95	330	330.0	58.94	100.0	99.95	0.05	0.05
73	A1Z2T 2L 2C	4S 2	38128	372	371	49.12	99.70	378	378.0	56.57	100.0	99.70	0.30	0.31
74	A1Z2T 2L 2C	5S 1	39942	380	373	55.27	98.06	407	407.0	61.84	100.0	98.06	1.94	0.81
75	A1Z2T 2L 2C	5S 2	45807	450	446	56.89	99.01	440	439.0	59.10	99.73	99.28	0.72	0.65
76	A1Z2T 2L 2C	67S12	91304	813	801	49.57	98.63	928	927.0	55.55	99.87	98.75	1.25	0.54
77	A1Z2T 2L 2C	123S12	17123	194	186	53.62	95.71	200	200.0	64.65	100.0	95.71	4.29	2.24
78	A1Z2T 3L 1C	4S 1	116734	524	516	14.41	98.49	609	607.5	11.54	99.80	98.69	1.31	0.63
79	A1Z2T 3L 1C	4S 2	140138	666	656	14.53	98.63	750	749.0	12.49	99.90	98.72	1.28	0.43
80	A1Z2T 3L 1C	5S 1	159999	816	809	11.75	99.11	844	842.5	9.93	99.82	99.29	0.71	0.33
81	A1Z2T 3L 1C	5S 2	183494	905	898	14.60	99.20	1021	1020.5	9.09	99.95	99.24	0.76	0.29
82	A1Z2T 3L 1C	6S 1	150812	685	681	15.92	99.39	812	811.5	10.45	99.94	99.45	0.55	0.30
83	A1Z2T 3L 1C	6S 2	199380	892	889	14.56	99.65	1127	1127.0	11.69	100.0	99.65	0.35	0.20
84	A1Z2T 3L 1C	7S12	82104	337	334	12.51	99.09	450	450.0	12.12	100.0	99.09	0.91	0.51
85	A1Z2T 3L 2C	4S 1	33901	169	163	18.57	96.71	247	246.5	19.27	99.81	96.89	3.11	1.10
86	A1Z2T 3L 2C	4S 2	42969	255	251	17.52	98.48	275	274.5	18.01	99.82	98.66	1.34	1.13
87	A1Z2T 3L 2C	5S 1	45525	270	267	20.09	99.03	317	315.5	17.63	99.54	99.49	0.51	0.55
88	A1Z2T 3L 2C	5S 2	58767	312	306	18.30	98.07	374	372.0	20.33	99.45	98.61	1.39	0.81
89	A1Z2T 3L 2C	67S12	135081	676	670	17.43	99.22	937	934.0	21.57	99.70	99.52	0.48	0.39
90	A1Z2T 3L12C	123S12	72397	340	329	17.01	96.77	392	389.5	14.55	99.40	97.35	2.65	1.02
91	A2Z1T 3L 1C	1S12	78878	260	253	14.22	97.49	366	365.0	8.93	99.72	97.76	2.24	1.29
92	A2Z1T 3L 1C	2S12	67459	256	252	11.20	98.53	346	345.5	9.15	99.85	98.68	1.32	1.06
93	A2Z1T 3L 2C	1S12	40908	183	176	18.66	96.02	334	334.0	19.67	100.0	96.02	3.98	2.39
94	A2Z1T 3L 2C	2S12	34727	150	144	20.43	95.38	240	239.0	20.49	99.60	95.76	4.24	2.00
95	A2Z1T 3L12C	3S12	91775	380	362	17.59	95.08	600	595.0	17.38	99.13	95.92	4.08	1.60
96	A2Z1T 3L12C	4567S12	57235	294	279	11.72	94.63	347	345.5	20.15	99.50	95.10	4.90	1.81
97	A2Z1T12L 1C	1S 1	27846	156	153	54.67	99.07	197	197.0	52.33	100.0	99.07	0.93	0.77
98	A2Z1T12L 1C	1S 2	26071	160	154	54.04	97.07	202	201.5	59.09	99.69	97.37	2.63	1.71
99	A2Z1T12L 1C	2S12	49207	260	251	57.72	96.69	329	327.0	57.78	99.53	97.15	2.85	1.45
100	A2Z1T12L 2C	1S12	26483	204	198	56.55	96.12	284	283.0	67.21	99.53	96.57	3.43	1.45
101	A2Z1T12L 2C	2S12	23096	194	188	49.31	98.01	235	234.0	70.94	99.94	98.07	1.93	0.87
102	A2Z1T12L12C	3S12	46592	284	262	57.02	91.65	405	402.5	61.06	99.18	92.40	7.60	2.13
103	A2Z1T12L12C	4567S12	27048	156	150	68.49	95.40	222	219.5	68.78	98.83	96.52	3.48	2.32
104	A2Z2T 3L 1C	4S 1	69014	244	237	13.16	97.07	350	348.5	9.15	99.56	97.50	2.50	1.34
105	A2Z2T 3L 1C	4S 2	59427	213	206	13.89	96.54	305	304.5	11.92	99.83	96.70	3.30	1.34
106	A2Z2T 3L 1C	5S 1	52665	224	217	11.91	96.59	267	267.0	13.96	100.0	96.59	3.41	1.52
107	A2Z2T 3L 1C	5S 2	37080	166	163	10.68	98.18	206	206.0	12.86	100.0	98.18	1.82	1.04
108	A2Z2T 3L 2C	4S 1	37272	177	172	19.76	97.54	242	242.0	19.61	100.0	97.54	2.46	0.99
109	A2Z2T 3L 2C	4S 2	33437	151	145	20.41	96.36	224	222.5	20.41	99.37	96.97	3.03	1.86
110	A2Z2T 3L 2C	5S12	51476	241	235	18.74	97.24	319	318.0	23.36	99.70	97.53	2.47	1.01
111	A2Z2T 3L12C	67S12	73661	277	275	15.44	99.24	432	432.0	21.41	100.0	99.24	0.76	0.52
112	A2Z2T 3L12C	123S12	101751	390	370	16.46	94.47	601	597.5	16.86	99.47	94.97	5.03	1.16
113	A2Z2T12L 1C	4S 1	45658	270	265	59.26	97.34	336	335.0	61.72	99.59	97.74	2.26	1.24
114	A2Z2T12L 1C	4S 2	39773	259	246	62.08	95.61	348	346.5	62.08	99.76	95.84	4.16	1.54
115	A2Z2T12L 1C	5S12	54331	314	308	60.76	97.78	367	364.0	59.12	99.55	98.22	1.78	1.18
116	A2Z2T12L 1C	123S12	44795	272	255	54.31	94.20	344	339.5	62.49	98.60	95.54	4.46	1.72
117	A2Z2T12L 2C	4S 1	26736	209	203	57.99	96.95	278	275.5	71.72	99.50	97.43	2.57	1.48
118	A2Z2T12L 2C	4S 2	22213	180	175	56.83	98.86	220	220.0	64.62	100.0	98.86	1.14	0.82
119	A2Z2T12L 2C	5S12	35143	265	261	57.61	98.86	327	323.5	65.31	99.65	99.21	0.79	0.71
120	A2Z2T12L 2C	123S12	20093	168	160	56.84	95.12	233	231.0	65.54	99.04	96.04	3.96	1.98
121	A2Z2T12L12C	67S12	45510	297	289	49.21	98.00	319	318.5	60.23	99.96	98.04	1.96	0.86

Appendix G

List of SAS Programs

1. Data:
   - Census data sets: people, households and buildings:
     `...\DSE\ProgSAS\VZsurUnix\lecVZpers.sas`
     `...\DSE\ProgSAS\VZsurUnix\lecVZhh.sas`
     `...\DSE\ProgSAS\VZsurUnix\lecVZbat.sas`
   - Imputation flags from the census data set:
     `...\DSE\ProgSAS\Matchs\matchFlag.sas`
   - Geographical data and construction of the analysis area:
     `...\DSE\ProgSAS\VZsurUnix\domaines.sas`
2. Search for matches:
   - Matching results from the census staff
     `...\Estimations\sysVZ\reprisematch\base.sas`
   - Correction of matching codes:
     `...\Estimations\sysVZ\reprisematch\adapt.sas`
   - Final matching codes:
     `...\DSE\ProgSAS\Matchs\repriseM.sas`
3. Search for CE/EE:
   - CE/EE results from the census staff and additional steps:
     `...\Estimations\sysVZ\repriseEE\baseE.sas`
   - Final CE/EE codes:
     `...\DSE\ProgSAS\EE\repriseE.sas`
4. Results about CE and EE:
   - Simple status and population:
     `...\DSE\ProgSAS\EE\repriseE.sas`
   - Location:
     `...\DSE\ProgSAS\EE\EEArea.sas`

- Population, location and partners:
  ...\DSE\ProgSAS\EE\EEDomArea.sas
- Variance:
  ...\DSE\ProgSAS\EE\EEvar.sas

5. Matching results:
   - Simple status and population:
     ...\DSE\ProgSAS\Matchs\repriseM.sas
   - Comparison of characteristics:
     ...\DSE\ProgSAS\Matchs\matchDemo.sas
     ...\DSE\ProgSAS\Matchs\matchDemoComp.sas
   - Location:
     ...\DSE\ProgSAS\Matchs\matchArea.sas
   - Population, location and partners:
     ...\DSE\ProgSAS\Matchs\matchDom.sas
     ...\DSE\ProgSAS\Matchs\matchDomArea.sas
   - Variance:
     ...\DSE\ProgSAS\Matchs\matchVar.sas

6. Estimation cells:
   - Selection of variables:
     ...\DSE\ProgSAS\poststraDSE\poststraChoix.sas
   - Construction of cells:
     ...\DSE\ProgSAS\poststraDSE\poststraConstrSex.sas

7. DSE estimation:
   - Overall results:
     ...\DSE\ProgSAS\poststraDSE\poststraConstrSex.sas
   - Results with variance:
     ...\DSE\ProgSAS\Var\DSEVar.sas

8. Others:
   - Various calculations in the census data sets:
     ...\DSE\ProgSAS\divers.sas
   - Checks and outliers:
     ...\DSE\ProgSAS\Outlier\checks.sas

Bibliography

Abbott, O., Brown, J., Diamond, I. (2003). Census 2001. Dependence within the One Number Census (draf form). National Statistics website.

ABS (1997). The 1996 census of population and housing. Annual report 1996-97, Australian Bureau of Statistics.

ABS (1999). Measuring census undercount in Australia and New Zealand. Demography working paper 1999/4, Australian Bureau of Statistics.

ABS (2004). Results for the census 2001. Personal communication, email from Paul Williams, Australian Bureau of Statistics.

Bell, R. M. (1994). Sampling and statistical estimation in the decennial census. *Proceedings of the section on survey research methods, American Statistical Association*, pages 71–79.

Bell, W. R. (1993). Using information from demographic analysis in post-enumeration survey estimation. *Journal of the American Statistical Association*, 88(423):1106–1166.

Brewer, K. R. W. (2002). *Combined survey sampling inference : Weighing Basu's elephants*. Arnold, London.

Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J., Teague, A. D. (1999). A methodological strategy for a one-number census in the UK. *J. R. Statist. Soc. A*, 162(2):247–267.

Cella, P., C. N. and Tuoto, T. (2004). The coverage rate estimation for the fifth agriculture census in the presence of matching error. Q2004, European meeting on quality and methodology in official statistics.

Clark, C. and Tourigny, J. (1999). Designing coverage studies for the 2001 Canadian census. 1999 FCSM Conference Papers.

Davis, P. (2001). Accuracy and coverage evaluation: Dual system estimation results. DSSD Census 2000 Procedures and Operations Memorandum Series B-9*, U.S. Census Bureau.

Ding, Y. and Fienberg, S. E. (1994). Dual system estimation of census undercount in the presence of matching error. *Survey methodology*, 20(2):149–158.

Evans, J., Kahles, D. and C. Bate (1993). 1991 Census data quality: Aboriginal and Torres Strait Islander counts. Demography working paper 1993/6, Australian Bureau of Statistics.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Fenstermaker, D. (2002). A.C.E. Revision II: summary of estimated net coverage. DSSD A.C.E. Revision II Memorandum Series PP-54, U.S. Census Bureau.

Fienberg, S. E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey methodology*, 18(1):143–154.

Freedman, D. A. and Navidi, W. C. (1992). Should we have adjusted the U.S. census of 1980 ? *Survey methodology*, 18(1):3–24.

Griffin, R. (2000). Accuracy and coverage evaluation: Dual system estimation. DSSD Census 2000 Procedures and Operation Memorandum Series #Q-20, U.S. Census Bureau.

Hogan, H. (1992). The 1990 post-enumeration survey: An overview. *The American Statistician*, 46(4):261–269.

Hogan, H. (1993). The post-enumeration survey: Operations and results. *Journal of the American Statistical Association*, 88(423):1047–1060.

Hogan, H. (2000). Accuracy and Coverage Evaluation: Theory and application. 2000 DSE Workshop of the Academy of Science Panel to Review the 2000 Census, U.S. Census Bureau.

Hogan, H. (2001). Accuracy and Coverage Evaluation: Data and analysis to inform the ESCAP Report. DSSD Census 2000 Procedures and Operation Memorandum Series B-1*, U.S. Census Bureau.

Hogan, H. (2003). The Accuracy and Coverage Evaluation: Theory and design. *Survey Methodology*, 29(2):129–138.

Hunter, B.H. and Dungey, M.H. (2003). Creating a sense of 'closure': Providing confidence intervals on some recent estimates of Indigenous populations. CAEPR Discussion Paper 244, Center for Aboriginal Economic Policy Research.

Kilchmann, D. und P. Eichenberger (2005). Einsetzungsverfahren in der Volkszählung 2000. Methodenbericht, in preparation, Bundesamt für Statistik.

Kim, J.-K., Navarro, A. and Fuller, W. A. (2000). Variance estimation for 2000 census coverage estimates. In *ASA, Proceedings of the Section on Survey Research*. American Statistical Association.

Kostanich, D. (2003). A.C.E. Revision II: summary of methodology. DSSD A.C.E. Revision II Memorandum Series #PP-35.

Kostanich, D. (2004). Response to "coverage of the Swiss population census 2000: questions about variance estimation". Personal communication with remarks from R. Fay, D. Olson and E. Schindler.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Morel, J. et Kleim, G. (2003). Recensement de la population 2001. Etudes de couverture. Rapport sur le surdénombrement. Revue de la méthodologie. Rapport de mars 2003, Statistique Canada.

Mule, T. (2000). Accuracy and Coverage Evaluation Survey: Weight trimming plan. DDSD Census 2000 Procedure and Operations Memorandum Series #Q-26, U.S. Census Bureau.

Mule, T. (2001). Accuracy and Coverage Evaluation: decomposition of dual system estimate components. DDSD Census 2000 Procedure and Operations Memorandum Series B-8*, U.S. Census Bureau.

Mule, T. (2003a). A.C.E. Revision II Results: change in estimated net undercount. DSSD A.C.E. Revision II Memorandum Series PP-58, U.S. Census Bureau.

Mule, T. (2003b). Accuracy and Coverage Evaluation Survey: Summary of results for weight trimming. DDSD Census 2000 Procedure and Operations Memorandum Series Q-87, U.S. Census Bureau.

Mulry, M. H. and Spencer, B. D. (1993). Accuracy of the 1990 census and undercount adjustments. *Journal of the American Statistical Association*, 88(423):1080–1091.

National Research Council (1999). *Measuring a Changing Nation. Modern Methods for the 2000 Census*. National Academy Press, Washington, D.C. Panel on alternative census methodologies.

National Research Council (2004). *The 2000 Census. Counting under adversity*. National Academy Press, Washington, D.C. Panel to review the 2000 census.

Navarro, A. (2000). Accuracy and coverage evaluation: Targeted extended search plans. DSSD Census 2000 Procedures and Operation Memorandum Series #Q-18, U.S. Census Bureau.

OFS (2002a). Portrait démographique de la Suisse. Edition 2002. Numéro de commande: 480-0200, Office fédéral de la statistique.

OFS (2002b). Recensement fédéral de la population 2000. Evolution de la population des communes 1850-2000. Numéro de commande: 001-0015, Office fédéral de la statistique.

OFS (2003a). La dynamique spatiale et structurelle de la population de la Suisse de 1990 à 2000. Numéro de commande: 494-0000, Office fédéral de la statistique.

OFS (2003b). La population étrangère en Suisse. Edition 2003. Numéro de commande: 276-0300, Office fédéral de la statistique.

OFS (2003c). Recensement fédéral de la population 2000. Structure de la population, langue principale et religion. Numéro de commande: 001-0019, Office fédéral de la statistique.

OFS (2004). Statistique de l'état annuel de la population (ESPOP) 2003. Résultats définitifs, après l'adaptation au RFP 2000. Actualités OFS. Numéro de commande: 341-0302-05, Office fédéral de la statistique.

ONS (2000). One Number Census estimation update. One Number Census Census Steering Committe Paper **00/16**.

ONS (2001). The treatment of movers in the 2001 CCS. One Number Census Census Steering Committe Paper **01/04**.

Pereira, R. (2002). The Census Coverage Survey: the key element of a One Number Census. National Statitsics. Population Trends, Summer 2002, No 108, pp. 16-31.

Rao, J.N.K. (1997). *Resampling methods for complex survey data. In: Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth*, pages 149–156. Birkhäuser Verlag Basel, Edited by C. Malaguerra, S. Morgenthatler and E. Ronchetti edition.

Renaud, A. (2001). Methodology of the Swiss Census 2000 Coverage Survey. In *ASA, 2001 Proceedings of the Section on Survey Research [CD-ROM]*. American Statistical Association.

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample). Rapport de méthodes 338-0019, Office fédéral de la statistique.

Renaud, A. et Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Rapport de méthodes 338-0023, Office fédéral de la statistique.

Renaud, A. et P. Eichenberger (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Rapport de méthodes 338-0009, Office fédéral de la statistique.

Robinson, J. G. (2001). ESCAP II: demographic analysis results. Executive Steering Committee for A.C.E. Policy II. Report 1, October 13,2001, U.S. Census Bureau .

Robinson, J. G., Adlakha A. and West K. K. (2002). Coverage of population in census 2000: results from demographic analysis. Annual Meeting of the Population Association of America, Atlanta, Georgia, May 8-11.

Robinson, J. G., Ahmed, B., Das Gupta, P., Woodrow, K. A. (1993). Estimation of population coverage in the 1990 united states census based on demographic analysis (with discussion). *Journal of the American Statistical Association*, 88(423):1061–1071.

Sands, R. D. and Navarro, A. (2001). 2000 Census Accuracy and Coverage Evaluation survey variance estimates. In *ASA, Proceedings of the Section on Survey Research [CD-ROM]*. American Statistical Association.

Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New-York.

SAS Institute Inc. (2004). *SAS/STAT© 9.1 User's Guide*. SAS Institute Inc.

Schindler, E. L. (2002). Simplified variance estimation for complex surveys. In *ASA, 2002 Proceedings of the Section on Survey Research [CD-ROM]*. American Statistical Association.

Schuler, M. and Joye, D. (2004). Typologie des communes suisses : de 1980 à 2000. Unpublished report, 30 juillet 2004.

Shao, J. and Tu, D. (1995). *The jackknife and bootstrap*. Springer Series in Statistics.

Shores, R. (2002). A.C.E. Revision II: adjustment for correlation bias. DSSD A.C.E. Revision II Memorandum Series #PP-53, U.S. Census Bureau.

StatCan (1999). Coverage. 1996 Census technical reports 92-370-XIE, Statistics Canada.

Statistique Canada (1999). Couverture. Rapports techniques du recensement de 1996. 92-370-XIF, Statistique Canada.

Taylor, J. and Bell, M. (2003). Options for benchmarking ABS population estimates for Indigenous communities in Queensland. CAEPR Discussion Paper 243, Center for Aboriginal Economic Policy Research.

Thibault, C. (2003). Recensement de la population 2001. Rapport d'évaluation des études de couverture. Rapport du 10 avril 2003, Statistique Canada.

U.S. Census Bureau Monitoring Board (2001). A guide to statistical adjustment: How it really works. Report to Congress, U.S. Census Bureau.

Whitford, D. (2002). Chronologic overview of the Census 2000 adjustment decision. In *ASA, Proceedings of the Section on Survey Research [CD-ROM]*. American Statistical Association.

Wolter, K. M. (1985). *Introduction to variance estimation*. Springer Series in Statistics.

Wolter, K. M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81(394):338–346.

Zaslavsky, A. M., N. Schenker and T. R. Belin (2001). Downweighting influential cluster in surveys: Application to the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, 96(455):858–869.

ZuWallack, R., M. Salganik, R. Cromar and V. Th. Mule Jr. (2000). Final Sample Design for the Census 2000 Accuracy and Coverage Evaluation. In *ASA, Proceedings of the Section on Survey Research*. American Statistical Association.

Methodenberichte des Dienstes Statistische Methoden des BFS
Rapports de méthodes du Service de méthodes statistiques de l'OFS
Methodological reports of the Statistical Methods Unit of SFSO

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Order number: 338-0027

Kilchmann, D. (2004). Revision des Schweizerischen Lohnindex. Schätzmethoden der Lohnindices und deren Varianzschätzer. Bestellnummer: 338-0026

Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. Numéro de commande: 338-0025

Renaud, A. (2004). Analyse de données d'enquêtes. Quelques méthodes et illustration avec des données de l'OFS. Numéro de commande 338-0024

Renaud, A., Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Numéro de commande: 338-0023

Graf, M. (2004). Fusion de données. Etude de faisabilité. Numéro de commande: 338-0022

Potterat, J. (2003). Mietpreis-Strukturerhebung 2003. Entwicklung des Stichprobenplans und Ziehung der Stichprobe. Bestellnummer: 338-0021

Potterat, J. (2003). Landwirtschaftliche Betriebszählung 2003. Stichprobenplan der Zusatzerhebung. Bestellnummer: 338-0020

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample). Numéro de commande: 338-0019

Hulliger, B. (2003). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2000. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Bestellnummer: 338-0018

Renfer, J.-P. (2003). Enquête 2000 sur la recherche et le développement dans l'économie privée en Suisse. Plan d'échantillonnage. Numéro de commande: 338-0017

Potterat, J. (2003). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe. Schätzverfahren. Bestellnummer: 338-0016

Graf, M., Matei, A. (2003). Stratégie de choix des modèles de désaisonnalisation. Application aux séries de l'emploi total. Numéro de commande : 338-0015

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. Bestellnummer: 338-0014

Renaud, A. (2002). Programme international pour le suivi des acquis des élèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro de commande: 338-0013

Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d'échantillonnage. Numéro de commande: 338-0012

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. Bestellnummer: 338-0011

Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro de commande: 338-0010

Renaud, A., Eichenberger P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Numéro de commande: 338-0009

Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001. Bestellnummer: 338-0008

Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro de commande: 338-0007

Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro de commande: 338-0006

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande: 338-0005

Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer: 338-0004

Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande: 338-0003

Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer: 338-0002

Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande: 338-0001

Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen: Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer: 338-0000

de Rossi, F.-X. (1998). Méthodes statistiques pour le compte routier suisse.

Hulliger, B., Kassab, M. (1998). Evaluation of Estimation Methods for the Survey on Environment Protection Expenditures of Swiss Communes.

Salamin, P.A. (1998). Etablissement d'une clef de passage pondérée entre l'ancienne (NGAE 85) et la nouvelle nomenclature (NOGA 95) générale des activités économiques.

Peters, R. (1998). Extrapolation des données de l'enquête de structure sur les loyers.

Bender, A., Hulliger, B. (1997). Enquête suisse sur la population active: rapport de pondération pour 1996.

Salamin, P.A. (1997). Evaluation de la Statistique de l'emploi.

Peters, R. (1997). Etablissement du plan d'échantillonnage pour l'enquête 1996 sur la recherche et le développement dans l'économie privée en Suisse.

Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse: établissement du plan d'échantillonnage.

Peters, R. (1996). Pondération des données de l'enquête sur la famille en Suisse.

Comment, T., Hulliger, B., Ries, A. (1996). Gewichtungsverfahren für die Schweizerische Arbeitskräfteerhebung (1991-1995).

Hulliger, B. (1996). Haushalterhebung Familie 1994: Stichprobenplan, Stichprobenziehung und Reservestichproben.

Peters, R., Hulliger, B. (1996). Schätzverfahren für die Lohnstruktur-Erhebung 1994 / Procédure d'estimation pour l'enquête de 1994 sur la structure des salaires.

Peters, R. (1996). Schéma de pondération des indices PAUL.

Hulliger, B., Peters, R. (1996). Enquête sur le comportement de la population suisse en matière de transport en 1994: plan d'échantillonnage et pondération.

Hulliger, B. (1996). Gütertransportstatistik 1993: Schätzverfahren mit Kompensation der Antwortausfälle.

Salamin, P.A. (1995). Estimation des flux pour le module II des comptes globaux du marché de travail.

Peters, R. (1995). Enquête de structure sur les loyers: établissement d'un plan d'échantillonnage stratifié.

Hulliger, B. (1995). Konjunkturelle Mietpreiserhebung: Stichprobenplan und Schätzverfahren.

Schwendener, P. (1995). Verbrauchserhebung 1990 - Vertrauensintervalle.

Peters, R., Hulliger, B. (1994). La technique de pondération des données: application à l'enquête suisse sur la santé.

Hulliger, B., Peters, R. (1994). Enquête sur la structure des salaires en Suisse: stratégie d'échantillonnage pour le secteur privé.

Publikationsprogramm BFS

Das Bundesamt für Statistik (BFS) hat – als zentrale Statistikstelle des Bundes – die Aufgabe, statiastische Informationen breiten Benutzerkreisen zur Verfügung zu stellen.

Die Verbreitung der statistischen Information geschieht gegliedert nach Fachbereichen (vgl. Umschlagseite 2) und mit verschiedenen Mitteln

SFSO Publications

The Swiss Federal Statistical Office (SFSO) is the central and official purveyor of statistical information to the Swiss Government. It is officially mandated to supply this information to a wide range of users.

This statistical data is organized and disseminated on the basis of a subject-matter classification (see inside cover page).

Diffusionsmittel	*Kontakt Phone*	*Distribution medium*
Individuelle Auskünfte	032 713 60 11 info@bfs.admin.ch	Individual information
Das BFS im Internet	www.statistik.admin.ch	The SFSO on the Internet
Medienmitteilungen zur raschen Information der Öffentlichkeit über die neusten Ergebnisse	www.news-stat.admin.ch	Press releases: fast access to the latest results
Publikationen zur vertieften Information (zum Teil auch als Diskette/CD-Rom)	032 713 60 60 order@bfs.admin.ch	Purchase of printed and electronic publications and CD-Rom
Online-Datenbank	032 713 60 86 www.statweb.admin.ch	On-line Database

Nähere Angaben zu den verschiedenen Diffusionsmitteln liefert das laufend nachgeführte Publikationsverzeichnis im Internet unter der Adresse www.statistik.admin.ch→Aktuell→Publikationen.

The SFSO’s detailed online publications catalogue is updated daily (www.statistics.admin.ch→News→Publications).

Methodenberichte des Dienstes Statistische Methoden
Rapports de méthodes du Service de méthodes statistiques
Methodology Report of the Statistical Methods Unit

Die Methodenberichte beschreiben die mathematischen und statistischen Methoden, die den Resultaten und Analysen der öffentlichen Statistik zu Grunde liegen. Sie enthalten ausserdem die Evaluation und Entwicklung von neuen Methoden im Hinblick auf eine zukünftige Anwendung. Diese Publikationen sollen einerseits die verwendeten Methoden dokumentieren, um Transparenz und Wissenschaftlichkeit sicher zu stellen, und sie sollen andererseits die Zusammenarbeit mit den Hochschulen und der Wissenschaft fördern.

Zur Illustration der beschriebenen mathematischen Konzepte, werden im Bericht numerische Resultate aufgeführt. Diese sind allerdings nicht als offizielle Resultate der betreffenden Erhebungen zu verstehen. Ebenfalls können die tatsächlich angewendeten Methoden leicht von den hier beschriebenen abweichen.

Die Methodenberichte sind auf der Internetseite des BFS in elektronischer Form verfügbar.

Methodology reports describe the mathematical and statistical methods used to produce findings and carry out analysis in relation to official statistics. These reports also contain assessments and descriptions of new methodological approaches that could be implemented in the future. The aim is to ensure clarity and scientific rigor by keeping a record of the methods used and to encourage a closer working relationship with the scientific community and academic circles.

Numerical findings are presented in methodology reports for the purpose of clarifying the mathematical concepts described and should therefore not be taken as official findings for the surveys in question. Likewise, the methods actually used may differ slightly from those described in the reports.

The electronic version of methodology reports can be downloaded directly from the SFSO Web site.

Coverage of the Swiss population census is estimated for the first time for the census 2000. Both undercoverage and overcoverage are analyzed apart and then combined by using the dual system methodology. The estimates are based on two samples: the Enumeration sample (E-sample) and the Population sample (P-sample) in order to capture both the overcoverage and the undercoverage components.

Similar to results in other countries, we determined that 1.6% of the resident population were overlooked in the census (undercount) and that 0.4% were counted erroneously (overcount). The resulting overall rate of net undercoverage is 1.4% with larger values for some subgroups of the population such as 20-31 years-old people (2.8%) or foreigners (2.9-3.5%).

Other types of errors were analyzed such as error in the type of domicile, time delay between census day and effective data collection day for movers around the census day, or potential misclassification variables. The results and experience gained during the project can be used to improve the subsequent censuses.

Order Number:
338-0027

Orders:
032 713 60 60 Fax: 032 713 60 61 E-Mail: order@bfs.admin.ch

Price:
free

ISBN 3-303-00305-X