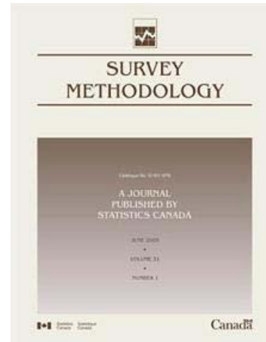
Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Estimation of the coverage of the 2000 census of population in Switzerland: Methods and results

by Anne Renaud



December, 2007

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2007
Vol. 33, No. 2, pp. 199-210
Statistics Canada, Catalogue no. 12-001-X

Estimation of the coverage of the 2000 census of population in Switzerland: Methods and results

Anne Renaud [1]

Abstract

Coverage deficiencies are estimated and analysed for the 2000 population census in Switzerland. For the undercoverage component, the estimation is based on a sample independent of the census and a match with the census. For the overcoverage component, the estimation is based on a sample drawn from the census list and a match with the rest of the census. The over- and undercoverage components are then combined to obtain an estimate of the resulting net coverage. This estimate is based on a capture-recapture model, named the dual system, combined with a synthetic model. The estimators are calculated for the full population and different subgroups, with a variance estimated by a stratified jackknife. The coverage analyses are supplemented by a study of matches between the independent sample and the census in order to determine potential errors of measurement and location in the census data.

Key Words: Census; Coverage errors; Dual system; Multi-stage sampling plan; Measurement errors.

1. Introduction

In any census, some people are not enumerated and should be, while others are counted twice or should not have been enumerated. There is both undercoverage and over-coverage, and quite often, the combined result is net under-coverage. For example, net undercoverage is estimated at 1.6% in the United States in 1990 (Hogan 1993), 2.2% in the United Kingdom in 1991 (Brown, Diamond, Chambers, Buckner and Teague 1999) and 3% in Canada in 2001 (Statistics Canada 2004). By contrast, in the United States in 2000, there is estimated to be net overcoverage of 0.5% (Hogan 2003). Coverage deficiencies may vary greatly between subgroups of the population. In the United States in 2000, blacks were found to have a net undercoverage of 1.8%, while whites had an overcoverage of 1.1%. Also, values often vary between age classes and regions, for example. These coverage deficiencies, and other errors such as measurement errors, result in a biased picture of the population. They are therefore studied in order to obtain information on the quality of the available data and to find ways to improve censuses of the population.

The 2000 population census in Switzerland gives a picture of the population on December 5, 2000. In this article, coverage deficiencies in a Swiss census are estimated for the first time. Undercoverage, overcoverage and net coverage resulting from the 2000 census are all analysed. Undercoverage is estimated from a sample of individuals S_P, independent of the census, on which a coverage survey was organized a few months after the census (collection took place in April and May 2001). The data from the survey are matched with data from the census to determine whether persons in S_P were enumerated. Overcoverage is estimated from a sample of individuals S_E drawn from census records. A search for duplicates and other erroneous records then serves to determine whether a given record corresponds to a real person to be enumerated. Net coverage is estimated on the basis of a capture-recapture model known as the dual system (Wolter 1986, Fienberg 1992). The dual estimator is applied in homogeneous cells, and the results are recombined using a synthetic model to obtain results for different domains of the population (Hogan 2003). The purpose of the project is not to adjust the census figures but rather to obtain information on the quality of the 2000 census and potential improvements for future censuses.

This article describes the different steps followed in obtaining estimates, then presents the results. Sections 2 and 3 describe the data sets and the coverage estimators. Section 4 provides the details on constructing the different statuses used in the estimators. Section 5 describes the approach used to compare the values collected in the census and in the survey for the matched persons from S_P. Sections 6 and 7 present the numerical results and the conclusion.

2. The three data sets

2.1 Census

The 2000 census was conducted under the auspices of the Federal Statistical Office, with the reference date of December 5, 2000. Information was collected for 7.3 million inhabitants, 3.1 million households, 3.8 million dwellings and 1.5 million buildings. The different levels were then linked by common identifiers when the data were processed.

1. Anne Renaud, Service de méthodes statistiques, Office fédéral de la statistique, Espace de l’Europe 10, CH-2010 Neuchâtel, Switzerland. E-mail: Anne.Renaud@bfs.admin.ch.

The collection of information on persons and households was the responsibility of Switzerland's 2,896 political communes. The latter had a choice between different methods of collection:

- TRADITIONAL: use of census agents;
- SEMI-TRADITIONAL: pre-printed questionnaires based on the communal register of inhabitants are mailed out and then collected by census agents;
- TRANSIT: pre-printed questionnaires are mailed out and mailed back;
- FUTURE: identical to TRANSIT except that links between households and dwellings are supplied by the commune;
- TICINO: similar to TRANSIT but limited to the canton of Tessin.

Most of the SEMI-TRADITIONAL, TRANSIT, FUTURE and TICINO communes also offered the option of completing questionnaires online. The 2,208 SEMI-TRADITIONAL, TRANSIT, FUTURE and TICINO communes that used the pre-printing of questionnaires based on communal registers of inhabitants account for nearly 96% of the population. For most of these communes, the tasks of mailing out questionnaires and controlling their return were organized at a national centre.

The data set for individuals contains 7,452,075 entries. One feature of this data set is that it contains two records for the same person if that person has two residences (2.3% of the population; for example, a student who both resides with his parents and has a residence close to his school). In the case of two residences, one is coded as the *economic residence* and the other as the *civil residence*. The economic residence is the place where the person spends the most time per week and the civil residence is where the person's official papers are kept (birth certificate for Swiss citizens, residence permit for foreigners). Where there is just one residence, it is both the economic and the civil residence. Switzerland is considered to have a *resident population* of 7,280,010 based, on the set of records showing the economic residence.

Households are classified as *private*, *collective* or *administrative*. Examples of private households are families, couples and persons living alone. Examples of collective households are groups of occupants of a home for the aged or a boarding school or the inmates of a prison. Administrative households group together people with no fixed residence, travellers and persons - by building or commune - who could not be assigned to private or collective households (2.4% of the resident population).

Census data contain no imputation at the record level, since communes sent basic information for non-respondents (unit non-response). However, values are imputed in the case of missing data or inconsistency in questionnaires (item non-response).

The *population of interest* for coverage estimates is the resident population (based on economic residence) in private and administrative households. Collective households, which account for 2.3% of the enumerated resident population, are excluded from the estimates.

2.2 S_P sample, coverage survey and matching (undercoverage)

The size objective for the S_P sample is set at approximately 50,000 people. In the absence of existing frames in Switzerland, this value was determined approximately, based on experiences in other countries. In particular, the Australian results for 1996 were used, since the sampling plan for Australia's coverage survey was similar to the one for Switzerland in 2000 (ABS 1997).

The S_P sample, which is independent of the census, is constructed in two parts: the canton of Tessin (TICINO) and the rest of Switzerland (NORD). Both parts use a multi-stage draw. The first stage consists in selecting 303 primary units - these are political communes for TICINO and postal codes for NORD - according to a stratified plan with a draw proportional to the number of buildings. The second stage consists in a simple random draw of a fixed number of 60 buildings per primary unit. In the NORD plan, these buildings are allocated to a maximum of three mail delivery routes, based on an intermediate sampling stage. The sampling is thus constructed so as to consolidate the field work while limiting the variability of the weights. For practical reasons and in light of available resources, postal codes that include a large proportion of buildings lacking complete postal addresses or coded as unoccupied are selected with a lower probability than other postal codes. These tend mainly to be postal codes in rural areas or industrial zones, which are unlikely to exhibit major coverage deficiencies. With the assistance of postal employees, complete lists of households are drawn up in the field within the sample of approximately 16,000 buildings. A sub-sample of buildings is then drawn so as to obtain a total of approximately 27,000 households. For more information on the sampling and survey procedure, see Renaud (2001) and, in greater detail, Renaud and Eichenberger (2002).

The coverage survey consists in contacting the 27,000 households - by telephone if a telephone number is found and in person if not. The variables collected are those that lend themselves to matching with the census and defining subgroups of interest for the coverage study (socio-demographic variables, addresses). The collection operation covers all members of all households in the selected buildings.

The final sample S_P contains $n_P = 49{,}883$ people in the population of interest (persons listed at their economic residence and residing in a private household). Of the households contacted, 88% were reached by telephone and 12% in person. The weighting depended on the sampling and an adjustment for non-response. The adjustment for non-response was based on a homogeneity model in cells constructed on the basis of the sampling strata and whether or not a telephone number was known to exist (interviews conducted by telephone or in person). It also incorporated an estimate of the proportion of true households among the households to be contacted, since a sizable portion of the households to be contacted actually consisted of vacant dwellings, stores or businesses. No calibration was applied, since the auxiliary data available were not independent of the census. There was no partial non-response. The weighting details are documented in Renaud and Potterat (2004).

Based on the questions asked in the survey and various plausibility controls, we hypothesize that the S_P data are correct and usable for matching with the census. The quality criteria used are as follows:

- *completeness*: the record is sufficient to identify the person;
- *appropriateness*: the person should have been enumerated;
- *uniqueness*: the person is listed only once;
- *belonging to population of interest*: the person is listed at his/her economic residence and in a private household;
- *correctness of location*: the person is listed at the correct address on Census Day.

The matching between the S_P sample and the census serves to determine the *matching status* P_j of each element j of S_P. Status P_j is equal to 1 if the element is matched in the census (enumerated person) and 0 if this is not the case (person not enumerated). In our case, the data collected in the coverage survey, the final census data and images of the census questionnaires are used for automatic matching, manual matching and controls. No supplementary interview took place in addition to the coverage survey. Persons who moved between Census Day and the day of the survey were sampled at their address on the day of the survey and then searched for on a priority basis at the address they had indicated for the day of the census. No case was unresolved by the end of the process.

2.3 S_E sample and search for erroneous records (overcoverage)

The size objective for the S_E sample was set at approximately 55,000 persons. This value, somewhat greater than that of S_P, had little influence on the processing of the data, since there was no field work or interview supplementary to the census.

The S_E sample was selected from the census data using a two-stage draw. Only elements included in the population of interest were eligible (records at the economic residence, without members of collective households). The primary units of S_E were identical to the primary units of S_P (postal codes and communes). However, the list of postal codes in the NORD plan that was used for S_P did not correspond exactly to the list of postal codes that were present in the census data. Census records found in postal codes that did not exist in the list used for S_P were therefore reallocated to existing codes, taking geographic location into account (this involved assigning fictitious postal codes for the sampling). In the second stage, records were drawn from the population of interest using a simple random plan, without intermediate stages. The allocation was done in such a way as to obtain constant weights in the sampling strata of the primary units. In the end, the sample contained $n_E = 55{,}375$ records (Renaud 2003).

We hypothesize that S_E records are sufficient to identify persons (completeness), since there is little imputation in the census data and most questionnaires were pre-printed based on registers of inhabitants. Appropriateness and uniqueness were determined in a matching between S_E and the rest of the census using a procedure similar to the matching between S_P and the census. In our case, this involves a search for duplicates or triplicates of elements of S_E, supplemented by an analysis of suspect cases in S_E. An element j is considered appropriate if it is not considered erroneous in the analysis of suspect cases (*e.g.*, a note on the questionnaire indicating that the person has gone abroad). An element j is considered unique if no duplicate or triplicate is detected in the census. There is no supplementary interview for S_E. There is therefore no information supplementary to the census for S_E persons (actual location? actual type of residence or household?). The search for duplicates/triplicates and suspect cases results in an *enumeration status* E_i for each element j of S_E. Status E_i is equal to 1 if the element should indeed have been enumerated in the census (default value) and 0 if it should not have been enumerated. In practice, it can take on values between 0 and 1 if the case is not determined precisely. Thus, duplicates and triplicates receive respectively the values 1/2 and 1/3 if there is no information allowing the correct record to be determined from among the records detected. These cases, which are rare, consist of persons who completed more than one questionnaire in the census without any link having been made between those questionnaires during the processing of the data.

3. Coverage estimators

3.1 Undercoverage and overcoverage

The *undercoverage rate* is estimated by $\hat{R}_{\text{under}} = 1 - \hat{R}_m$, where $\hat{R}_m$ is the estimate of the *correct matches rate* based on the S_P sample. Similarly, the *overcoverage rate* is defined as $\hat{R}_{\text{over}} = 1 - \hat{R}_c$, where $\hat{R}_c$ is the estimate of the *correct records rate* based on the S_E sample. The correct matches rate and the correct records rate are estimated by the weighted means of matching status P_j and enumeration status E_j, as follows:

$$\hat{R}_m = \frac{\sum_{j \in S_P} w_{P,\,j} P_j}{\sum_{j \in S_P} w_{P,\,j}} \quad \text{and} \quad \hat{R}_c = \frac{\sum_{j \in S_E} w_{E,\,j} E_j}{\sum_{j \in S_E} w_{E,\,j}}, \quad (1)$$

where $w_{P,\,j}$ is the weight of element j of sample S_P and $w_{E,\,j}$ is the weight of element j of sample S_E. We note that the denominator of $\hat{R}_c$ is the sum of the weights $w_{E,\,j}$ of S_E and not the number C of known records in the census, so as to have a potentially less biased estimator.

The estimate of the undercoverage and overcoverage rates in a domain d is given by $\hat{R}_{\text{under},\,d} = 1 - \hat{R}_{m,\,d}$ and $\hat{R}_{\text{over},\,d} = 1 - \hat{R}_{c,\,d}$, with

$$\hat{R}_{m,\,d} = \frac{\sum_{j \in S_P} w_{P,\,j} P_j I_{jd}}{\sum_{j \in S_P} w_{P,\,j} I_{jd}} \text{ and } \hat{R}_{c,\,d} = \frac{\sum_{j \in S_E} w_{E,\,j} E_j J_{jd}}{\sum_{j \in S_E} w_{E,\,j} J_{jd}}. \quad (2)$$

Identifiers I_{jd} and J_{jd} take on the value 1 if element j, respectively of S_P and S_E, is found in domain d; otherwise their value is 0.

3.2 Net coverage

The *net undercoverage rate* is estimated by $\hat{R}_{\text{netunder}} = 1 - \hat{R}_{\text{net}}$ where $\hat{R}_{\text{net}} = C / \hat{N}$ is the estimate of the *net coverage rate*, C is the number enumerated in the population of interest and $\hat{N}$ is the estimate of the true total in the population of interest. If $\hat{R}_{\text{netunder}}$ is negative, there is net overcoverage.

The estimate of the true total $\hat{N}$ is based on the dual model (Wolter 1986). This model is built on the principle of capture (census) and recapture (coverage survey). It is applied in estimation cells $k = 1, \ldots, K$ in order to best satisfy the assumptions of the model; see discussion below. Thus, the estimate of the true total $\hat{N}$ is composed of the sum of the estimated true totals $\hat{N}_k$ in disjoint estimation cells covering the population of interest $k = 1, \ldots, K$:

$$\hat{N} = \sum_{k=1}^{K} \hat{N}_k. \quad (3)$$

The estimated totals $\hat{N}_k$ have the form given by the dual model:

$$\hat{N}_k = [N_{1+,\,k}] \left[\frac{N_{+1,\,k}}{N_{11,\,k}} \right], \quad (4)$$

where $N_{1+,\,k}$ is the total of records correctly counted in cell k during capture (census), $N_{+1,\,k}$ is the total in k during recapture (estimated from sample S_P) and $N_{11,\,k}$ is the number of records common to the two lists (estimated from matches between S_P and the census).

The different terms of equation (4) are estimated using undercoverage and overcoverage estimates. This is an extension of the model in Wolter (1986), similar to the one used by Hogan (2003). Thus, the total of the records correctly counted in the census $N_{1+,\,k}$ is estimated by the enumerated total C_k multiplied by the correct records rate $\hat{R}_{c,\,k}$ to take account of overcoverage. Also, the ratio between the total in the recapture $N_{+1,\,k}$ and the number of records common to the two lists $N_{11,\,k}$ is estimated by the inverse of the rate of matching $\hat{R}_{m,\,k}$ between the coverage survey and the census in order to take account of undercoverage. We obtain

$$\hat{N}_k = [C_k \hat{R}_{c,\,k}][\hat{R}_{m,\,k}^{-1}] = C_k [\hat{R}_{c,\,k} \hat{R}_{m,\,k}^{-1}] = C_k \hat{F}_k, \quad (5)$$

where $\hat{F}_k = \hat{R}_{c,\,k} \hat{R}_{m,\,k}^{-1}$ is the *coverage correction factor* in cell k. Factor $\hat{F}_k$ combines the effects of overcoverage and undercoverage of cell k estimated from samples S_P and S_E. We note that undercoverage in one domain may be offset by overcoverage in the same domain. Thus, nil net undercoverage in a domain does not mean that no coverage deficiency exists in it.

The proposed estimates are based on the assumptions of the dual model, the choice of estimation cells, and the choice of the statuses defining the estimators $\hat{R}_{c,\,k}$ and $\hat{R}_{m,\,k}$. The dual model is useful since it takes into account the fact that some persons are reached neither by the census (capture) nor by the coverage survey (recapture). However, a series of conditions must be met to avoid estimation biases. The coverage survey and the census must be totally independent. The matching must be of very high quality. The model must be applied in cells with persons who have the same probability of being enumerated in the census and the survey respectively; see Section 3.3. Lastly, the population must not change too much between Census Day and the day of the survey. As to the estimators $\hat{R}_{c,\,k}$ and $\hat{R}_{m,\,k}$, they are based on the quality of the matching and the search for erroneous records. Also, it is necessary to ensure that the definition of a correct match in S_P and the definition of a correct record in S_E are identical, *i.e.*, that there is a balance between overcoverage and undercoverage; see Section 4. All those elements are taken into consideration insofar as possible in the present estimates.

The estimate of net undercoverage in a domain d has the form $\hat{R}_{\text{netunder},\,d} = 1 - \hat{R}_{\text{net},\,d} = 1 - C_d / \hat{N}_d$, where C_d is the enumerated number in the domain and $\hat{N}_d$ is the estimate of the true total. The estimate of the true total $\hat{N}_d$ is based on a *synthetic* model that assumes that the correction factor is fixed in each cell $k = 1, \ldots, K$:

$$\hat{N}_d = \sum_{k=1}^{K} \hat{N}_{k,d} = \sum_{k=1}^{K} C_{k,d} \hat{F}_k. \quad (6)$$

$C_{k,d}$ is the number enumerated in the population of interest in the intersection between cell k and domain d, and $\hat{F}_k$ is the correction factor for the coverage in cell k. The hypothesis of the synthetic model is satisfied if the behaviour of any subset in the cell is identical to that of the entire cell. This homogeneity is best controlled by the choice of cells. Here we are using the homogeneous cells defined by the dual model.

3.3 Estimation cells

The estimation cells $k = 1, \ldots, K$ are constructed in such a way as to group together elements that have homogeneous probabilities of enumeration in the census and the survey respectively (dual hypothesis) and homogeneous net coverage rates (synthetic hypothesis). We want a minimum of 100 persons per cell in S_E and S_P in order to control the variance and limit the estimation bias. The variables defining the cells are selected using a logistic regression model and a discrimination method applied to the data from S_P (binary variable: P_j). The three most influential variables are cross-tabulated: nationality in two categories, marital status in two categories and size of commune in three categories. The other variables are then successively integrated. Groupings are created when the cell sizes are too small (official language of commune in two categories, age class in seven categories and sex in two categories). In the end, 121 estimation cells are obtained; see Renaud (2004) for more details.

3.4 Variance of coverage estimators

The variance of the estimators is estimated by a stratified jackknife applied to the (identical) primary units of S_P and S_E. We note that the variance of the estimated under-coverage $\hat{R}_{\text{under}} = 1 - \hat{R}_m$ is equal to the variance of the estimated matching rate $\hat{R}_m$. Similarly, the variance of the overcoverage $\hat{R}_{\text{over}} = 1 - \hat{R}_c$ is equal to that of the correct record rate $\hat{R}_c$, and the variance of the net undercoverage $\hat{R}_{\text{netunder}} = 1 - \hat{R}_{\text{net}}$ is equal to that of the net coverage rate $\hat{R}_{\text{net}}$.

Let θ be the parameter of interest taking the form of a weighted mean of statuses in the case of undercoverage and overcoverage, and the form of a linear function of quotients between two weighted means in the case of net under-coverage. Its estimator is $\hat{\theta}$.

Let $h = 1, \ldots, H$ be the stratum used in the first stage of sampling, $i = 1, \ldots, m_h$ the number of the primary unit in stratum h (postal code for NORD or commune for TICINO), and $j = 1, \ldots, n_{hi}$ the number of the person in primary unit i of h. For the needs of the jackknife method, samples S_P and S_E are partitioned, in each stratum h, in m_h subsets corresponding to the persons in primary units $\alpha = 1, \ldots, m_h$.

Let $\hat{\theta}_{(h\alpha)}$ be the estimator having the same form as $\hat{\theta}$ but calculated on the sample from which primary unit α of stratum h. has been removed. We note that estimators $\hat{R}_{m(h\alpha),k}$ and $\hat{R}_{c(h\alpha),k}$, $k = 1, \ldots, K$ are combined to form $\hat{R}_{\text{net}(h\alpha)}$:

$$\hat{R}_{\text{net}(h\alpha)} = C \left[\sum_{k=1}^{K} C_k \frac{\hat{R}_{c(h\alpha),k}}{\hat{R}_{m(h\alpha),k}} \right]^{-1}. \quad (7)$$

The corrected weights w'_{hij} used to calculate values $\hat{R}_{m(h\alpha)}$ and $\hat{R}_{c(h\alpha)}$ have the following form:

$$w'_{hij} = \begin{cases} 0 & \text{if} \quad i = \alpha \\ w_{hij} \dfrac{m_h}{m_h - 1} & \text{if} \quad \alpha \in h \text{ and } i \neq \alpha \\ w_{hij} & \text{if} \quad \alpha \notin h. \end{cases} \quad (8)$$

This form of correction is preferred to the quotient between the sum of the weights of the elements in the stratum and the sum of the weights without primary unit α since it allows us to take account of the variability due to the unknown number of elements in the stratum.

The jackknife estimator becomes:

$$\hat{\theta}_{JK} = \frac{\sum_h \sum_{\alpha=1}^{m_h} \hat{\theta}_{h\alpha}}{\sum_h m_h}, \quad (9)$$

with pseudo values $\hat{\theta}_{h\alpha} = m_h \hat{\theta} - (m_h - 1)\hat{\theta}_{(h\alpha)}$. The estimator of its variance can take different forms; see the example of Shao and Tu (1995). We apply the following form:

$$v(\hat{\theta}_{JK}) = \sum_h \frac{m_h - 1}{m_h} \sum_{\alpha=1}^{m_h} (\hat{\theta}_{(h\alpha)} - \hat{\theta}_{(h.)})^2, \quad (10)$$

with $\hat{\theta}_{(h.)} = \sum_{\alpha=1}^{m_h} \hat{\theta}_{(h\alpha)} / m_h$. Lastly, we use $v(\hat{\theta}_{JK})$ as an estimator of the variance of $\hat{\theta}$. The estimates in the subgroups use the same form of estimator with integration of a domain indicator in the construction of $\hat{\theta}_{(h\alpha)}$. No correction for the finite population is applied in the estimates. Also, other variabilities are not taken into account, such as the variability induced by the weighting model for non-response in S_P.

Problems, such as the lack of stability of estimation in strata with few primary units, appeared in the course of applying this approach. However, tests on the sharing of some primary units and a comparison with the Taylor linearization or a simple jackknife suggest that the estimators of variance by stratified jackknife that are presented in this document are fairly conservative.

4. Choice of correct matching statuses and enumerations

A key element of coverage estimates is the definition of the *correct matching status* for the elements of S_P and the *correct enumeration status* for the elements of S_E. These correct statuses are defined on the basis of frames P_j and E_j determined during the matchings.

Is a match with a census element that is part of a collective household accepted as a correct match for an element of S_P, or is this a case of undercoverage of the population of interest? Is a duplicate outside the population of interest for an element of S_E really considered a duplicate, and hence an instance of overcoverage, or should it be excluded? A clear definition is needed. Also, the statuses used in estimates of net undercoverage must be chosen in such a way as to satisfy the balance between over- and undercoverage; see the concept of “balancing,” as, for example, in Hogan (2003). A match $(P_j = 1)$ with an element outside the population of interest may, for example, be rejected as a correct match (correct match status $= 0$, no undercoverage) only if the search for correct records would also detect this element as incorrect because it is out of scope (correct enumeration status $= 0$, no overcoverage).

The criteria for defining correct statuses are constructed using information available for elements of S_P and S_E. As regards S_P, we start with the assumption that census records that were matched with elements of S_P serve to identify persons (completeness) and these persons should indeed have been enumerated (appropriateness). We also consider that they are unique, since uniqueness, while controlled by matches, is achieved in the great majority of cases controlled in S_E. The criteria of belonging to the population and correctness of location are controlled by comparison with the information collected in the coverage survey, considered as reference information. No supplementary data collection was organized to resolve ambiguous cases. As regards S_E, we have the criterion of completeness considered as having been met in the census data and the results concerning uniqueness and appropriateness obtained in the matching with the rest of the census. For duplicates and triplicates, we define $E_j = 1/d'$, with $d' =$ number of duplicates/triplicates in the population of interest according to the census. The criteria of belonging to the population of interest and correctness of location for the elements of S_E cannot be controlled, since we have no reference data supplementing the census.

For estimates of net undercoverage, it is important to meet the balancing requirement. The criteria used in defining correct statuses are thus completeness, appropriateness and uniqueness. The criteria of belonging to the population of interest and correctness of location cannot be considered, since they are not usable in defining the correct enumeration status. The criteria of completeness, appropriateness and uniqueness are already integrated into the construction of the basic statuses P_j and E_j. Thus, estimates are made with basic statuses P_j and E_j.

For estimates not using the dual system and the need for balancing, it is possible to use other criteria to define correct statuses. Other types of correct match statuses are used in the analysis of potential measurement errors in Section 5 and the more detailed analyses of matches and enumerations presented in Renaud (2004).

5. Comparison of matches

5.1 Potential measurement errors

Measurement errors or classification errors are related to coverage errors. A person who is classified in domain d according to the census (*e.g.*, a person between 10 and 19 years of age) but who in reality is outside the domain (*e.g.*, a person 60 years of age) would end up as a case of overcoverage in domain d and an undercoverage case outside that domain. This misclassification does not cause a coverage error at the overall level, but it causes an error at the level of subgroups of the population.

The reasons for differences between the values collected in two surveys such as the census and the coverage survey may be quite varied and difficult to dissociate. It is to be expected that there will be matching errors, differences resulting from collection methods (paper questionnaire or telephone/face-to-face interview) and data processing methods, or real differences due to the time lag between the collection periods (December 2000 and April-May 2001). Also, it is difficult to determine the correct response if there are two different values. What is the correct choice - the census? the survey? another value not collected?

Potential measurement errors with respect to census data are analysed on the basis of a set of matches between the independent sample S_P and the census. We choose to determine which variables show respectively few or many potential classification problems, without making a judgment on the quality of either data collection. One use of this information is to evaluate the choice of estimation cells for the dual system and select subgroups for which the estimates of coverage deficiencies are soundest.

For the category variable X, we define the matching rate in the good domain $\hat{R}_X$ as follows:

$$\hat{R}_X = \frac{\sum_{j \in S_P \text{ match}} w_{P,\,j} P_{X,\,j}}{\sum_{j \in S_P \text{ match}} w_{P,\,j}}, \quad (11)$$

where $w_{P,\,j}$ is the weight of element j of matched sample S_P (S_P match) and the *classification status* $P_{X,\,j}$ is equal to 1 if element j appears in the same class in the census and the survey, and 0 otherwise. The value of $\hat{R}_X$ is estimated with the set of matched elements and with the subgroup of elements without imputation in the census.

We also define a measure of asymmetry $\varphi_X(d,\ d')$ for classes d and d' of variable X:

$$\varphi_X(d,\ d') = \frac{\sum_{j \in S_P \text{ match}} w_{P,\,j}\, I_j(d,\ d')}{\sum_{j \in S_P \text{ match}} w_{P,\,j}\, I_j(d',\ d)}, \quad (12)$$

where $I_j(d,\ d') = 1$ if element j appears in domain d according to the survey and in domain d' according to the census, and 0 otherwise. The factor $\varphi_X(d,\ d')$ is equal to 1 if there is a balance in the classification errors - in other words, if the number of elements in d according to the survey and d' according to the census is equal to the number in d' according to the survey and in d according to the census. The further the factor lies from 1, the less balance there is.

5.2 Potential location errors

Comparisons between the census and the survey can also be used to study people's geographic location. In the census data, we have a unique address if the person has a single residence and two addresses - principal and secondary - if the person has two residences. In the survey data, we have one or two addresses on Census Day, one or two addresses on the day of the survey and information on a possible move between the two dates. If a person has a single residence and has not moved, that person's principle address on Census Day and his/her principle address on the day of the survey are identical. The person does not have secondary addresses.

Different measures of distance are considered in order to determine potential location errors in the census. For practical reasons, including the data available, we define geographic areas around the person's principle address collected in the survey for Census Day (*reference address*). The areas are sets of political communes. They are defined on the bases of postal codes identified in the survey. The person's *basic area* is defined by the set of communes that have buildings within the postal code of the person's reference address. The definition of this area uses data from the Swiss building register, since the latter has information on buildings' postal address and the commune within which they are located. The *extended area* includes the communes within the basic area and the set of communes adjacent to them; see Renaud (2004) for examples.

Like classification errors, location errors do not cause coverage errors at the overall level but they cause errors at the level of subgroups such as regions or types of communes. Different rates may be defined. We will retain the basic location rate and the extended location rate, both weighted by w_{Pj}, the weight of element j of matched sample S_P. The location status takes on the value 1 if the element lies within the basic area or the extended area, as the case may be; otherwise it equals 0. In particular, we will study the correctness of location of persons who have moved, in order to detect possible problems relating to the time lag between Census Day and the actual day of collection of census data.

6. Results

6.1 Estimates of coverage deficiencies

The overall net undercoverage rate is estimated at 1.41% with a standard deviation of 0.12%. The overcoverage rate is 0.35% (standard deviation = 0.03%) and the undercoverage rate is 1.64% (standard deviation = 0.11%). These results are of the same order of magnitude as those of other countries, although they are in the lower range; see Table 1.

Overcoverage is minor in the great majority of the domains studied. The highest rate is observed for persons between 20 and 31 years of age (0.93% with a standard deviation of 0.09%); see Table 2. However, undercoverage is high in several domains. For example, a rate of 8.03% (standard deviation = 0.85%) is observed for foreigners with temporary settlement permits ("other permits") and a rate of 3.50% (standard deviation = 0.50%) is observed for 20-31-year-olds. Also, an undercoverage rate of 2.4% is observed in the Italian-speaking region of the country (language of commune: Italian; NUTS region: Ticino, and collection method: TICINO). However, the results are related to relatively great variability (standard deviation of approx. 0.5%), since samples S_P and S_E include only 1,500 and 1,700 persons respectively in this region.

Net undercoverage is positive in all the domains studied. There is therefore no net overcoverage. The highest values are observed for foreigners with permanent or temporary permits (2.89% and 3.48%, standard deviations = 0.32 % and 0.39%) as well as for 20-31-year-olds (2.84%, standard deviation = 0.36%). No significant difference is observed between males and females, between languages or between NUTS regions. Because of the small size of the sample with the collection variant TICINO, this method cannot be differentiated from the others used in the country. On the other hand, significant differences are observed between marital statuses, as well as between types and sizes of communes.

Table 1
International comparison of overall results. Estimated rates of overcoverage $\hat{R}_{\text{over}}$, undercoverage $\hat{R}_{\text{under}}$ and net undercoverage $\hat{R}_{\text{netunder}}$, with corresponding estimated standard deviations. References: Statistics Canada (1999, 2004), Hogan (1993, 2003), McLennan (1997) and Trewin (2003)

		Overcoverage [%]	Undercoverage [%]	Net undercoverage [%]
Switzerland	2000	0.3 (0.03)	1.64 (0.11)	1.41 (0.12)
Canada	1996	0.74 (0.04)	3.18 (0.09)	2.45 (0.10)
	2001	0.96 (0.05)	3.95 (0.13)	2.99 (0.14)
United States	1990	3.1	4.7	1.6 (0.10)
	2000	-	-	-0.5 (0.21)
Australia	1996	0.2	1.8	1.6 (0.10)
	2001	0.9	2.7	1.8 (0.10)

We note that the net undercoverage rate is greater than the undercoverage rate in the case of permanent settlement permits. This effect, which is unrealistic, is due to the choice of estimation cells and the resulting smoothing. The construction of the cells made it necessary to group foreigners with permanent and temporary permits into a single category for aggregates so as to obtain the minimum size of 100 persons per cell. By making this grouping, we are treating foreigners as a homogeneous group, whereas this is not the case. This shows the limitations of the method and the difficulty of satisfying the assumptions of the models used in applying the approach. In the case of foreigners, we note, however, that the confidence intervals of the undercoverage and net undercoverage rates overlap. The consequences of the weaknesses of the application are therefore limited.

It should also be noted that the results are presented in domains defined by variables for which low levels of potential measurement errors were observed. The fact is that results for groups as defined by household or labour market characteristics would not be very reliable; see Section 6.2.

The precision of the results obtained is generally better than the objective set at the beginning of the project. That objective was to have a standard deviation of 0.3% for subgroups of 10,000 individuals in S_P. In the case of, for example, age classes 32-44 and 45-59, which have between 10,000 and 12,000 persons, the standard deviations are 0.19% and 0.14 %.

6.2 Potential measurement and classification errors

Of the 49,107 elements matched between the coverage survey and the census, 96% exhibit no difference in sex, the seven age classes, the three marital status classes and the three settlement permit classes (Swiss, permanent, temporary). The matching rate in the good domain $\hat{R}_X$ is 99.3 % for sex (with and without imputations), 98.3% for marital status (98.4% for non-imputed values) and 98.7% for settlement permits (98.8% for non-imputed values). The $\hat{R}_X$ rate is 99.5% for age classes (with or without imputations). However, it should be noted that date of birth, along with surname and given name, was one of the main variables in the matching. Age differences are therefore possible only in the case of a non-automatic (computer-assisted or manual) matching. Three variables exhibit a matching rate in the good domain that is markedly lower than that observed for sex, age, permit and marital status. These are the variables for labour market status (in the labour market, unemployed, not in the labour market), position in household (alone, spouse, common-law union, person with child or children, other head of household, related to head of household, other; results limited to private households), and size of the person's household (according to economic residence and in private households). The $\hat{R}_X$ rate is 90.4% for labour force status (91.1% for non-imputed values), 91.4% for position in household (94.9% for non-imputed values) and 88.3% for household size.

The measure of asymmetry $\varphi_X(d,\ d')$ takes on the value 1.33 for sex ($d =$ male and $d' =$ female). There are more persons coded as males according to the survey who are coded as females in the census than there are females according to the survey who are males according to the census. The proportion of males is slightly higher in the survey. However, these results must be interpreted with caution, since they based on very few cases; see Table 3. A McNemar test is just significant at the 5% level without taking the design into account, but it is no longer significant at that level if the design is factored in. On the other hand, quite substantial asymmetries are observed for marital status. There are fewer single persons in the survey who are married in the census than the reverse (factor 0.33 for $d =$ single and $d' =$ married). Similarly, there are fewer married persons in the survey who are widowed in the census than the reverse (factor 0.42 for $d =$ married and $d' =$ other). Asymmetry is also observed for the settlement permit variable. The tendency is to have more Swiss persons in the survey who are described as foreigners in the census than the reverse, and to have more permanent permits in the survey and temporary permits in the census than the reverse (factors 5.22 for $d =$ Swiss and $d' =$ foreigner with permanent permit and 3.83 for $d =$ foreigner with permanent permit and $d' =$ foreigner with temporary permit). The factors calculated are based on few cases. However, they give an insight into the potential differences between data collection via the census questionnaire and a survey conducted mainly by telephone. The labour force status variable includes more divergent cases; see Table 4. Thus, for example, we observe fewer persons employed force in the survey and fewer persons not in the labour force census than the reverse (factor of 0.46 for $d =$ in labour force and $d' =$ not in labour force). There are also fewer unemployed persons in the survey and persons not in the labour force in the census than the reverse (factor of 0.26 for $d =$ unemployed and $d' =$ not in labour force). The position-in-household variable also exhibits asymmetries, but these are based on few elements, since the dispersion of the elements in the boxes (d, d') is sizable. The census variables at the household level (position in household and size of household) are influenced by the complex process of household formation. They are less reliable than those concerning persons. The values at the household level are more reliable in the survey.

Table 2 Enumerated number C and estimated rates of overcoverage $\hat{R}_{\text{over}}$, undercoverage $\hat{R}_{\text{under}}$ and net undercoverage $\hat{R}_{\text{netunder}}$ for different domains [%], with corresponding estimated standard deviations (SDs)

Variable	Category	C	$\hat{R}_{\text{over}}$	SD	$\hat{R}_{\text{under}}$	SD	$\hat{R}_{\text{netunder}}$	SD
Overall		7,121,626	0.35	0.03	1.64	0.11	1.41	0.12
Sex	Male	3,497,940	0.37	0.04	1.74	0.13	1.46	0.13
	Female	3,623,686	0.33	0.03	1.55	0.10	1.37	0.13
Age class	≤ 9	810,373	0.26	0.05	1.46	0.21	1.34	0.26
	10-19	833,185	0.27	0.05	1.30	0.19	1.04	0.22
	20-31	1,115,804	0.93	0.09	3.50	0.34	2.84	0.36
	32-44	1,544,721	0.33	0.05	1.65	0.16	1.43	0.19
	45-59	1,431,771	0.22	0.04	1.18	0.14	1.04	0.14
	60-79	1,146,709	0.10	0.03	0.91	0.13	0.82	0.12
	≥ 80	239,063	0.11	0.06	1.20	0.31	1.03	0.27
Settlement permit	Swiss	5,674,266	0.33	0.03	1.28	0.09	0.98	0.10
	Foreigner, permanent	1,020,242	0.33	0.06	1.85	0.29	2.89	0.32
	Foreigner, temporary	427,118	0.56	0.11	8.03	0.85	3.48	0.39
Marital status	Single	2,975,643	0.50	0.05	2.07	0.18	1.72	0.19
	Married	3,377,223	0.23	0.04	1.27	0.11	1.25	0.12
	Widowed	369,339	0.25	0.08	1.23	0.26	0.79	0.13
	Divorced	399,421	0.24	0.08	1.95	0.35	1.02	0.10
Commune language	German + Romansh	5,128,353	0.33	0.04	1.50	0.11	1.28	0.12
	French	1,680,062	0.35	0.06	1.89	0.25	1.79	0.27
	Italian	313,211	0.53	0.12	2.35	0.49	1.56	0.19
NUTS region	Région lémanique	1,296,464	0.37	0.07	2.19	0.38	1.84	0.28
	Espace Mittelland	1,640,489	0.35	0.09	1.39	0.15	1.25	0.10
	Nordwestschweiz	976,699	0.18	0.04	1.50	0.27	1.32	0.12
	Zurich	1,221,014	0.31	0.05	1.58	0.19	1.46	0.13
	Ostschweiz	1,020,897	0.40	0.07	1.29	0.23	1.24	0.12
	Zentralschweiz	665,904	0.36	0.06	1.57	0.25	1.19	0.12
	Ticino	300,159	0.54	0.12	2.38	0.52	1.57	0.19
Commune size	Small	1,372,958	0.34	0.05	1.50	0.15	1.12	0.14
	Medium	2,398,256	0.41	0.07	1.32	0.16	1.07	0.19
	Large	3,350,412	0.31	0.03	2.01	0.19	1.77	0.19
Type	City/town	2,078,780	0.35	0.04	1.96	0.17	1.82	0.20
	Agglomeration	3,145,541	0.36	0.06	1.49	0.19	1.34	0.12
	Rural	1,897,305	0.32	0.04	1.56	0.17	1.07	0.12
Collection method	TRADITIONAL	265,607	0.39	0.05	1.91	0.28	1.07	0.12
	SEMI-TRADITIONAL	174,501	0.37	0.08	1.07	0.24	1.16	0.13
	TRANSIT + FUTURE	6,381,359	0.33	0.03	1.62	0.11	1.42	0.12
	TICINO	300,159	0.54	0.12	2.38	0.52	1.57	0.19

Table 3 Comparison of values collected in the survey and the census for the sex variable

Sex			Survey		
			Male	**Female**	**Total**
Census	Not matched	Total	393	383	776
	Matched	Total	24,171	24,936	49,107
	Matched	Male	23,967	166	24,133
		Female	204	24,770	24,974
	Matched (imputed value)	Male	6	0	6
		Female	0	13	13
	Total		24,564	25,319	49,883

Table 4 Comparison of values collected in the survey and the census for the labour force status variable

Labour force status			Survey				
			Employed	**Unemployed**	**Not in labour force**	**≤ 15 years of age**	**Total**
Census	Not matched	Total	424	23	217	112	776
	Matched	Total	25,163	498	14,501	8,945	49,107
	Matched	Employed	23,953	188	2,007	13	26,161
		Unemployed	300	221	323	1	845
		Not in labour force	901	89	12,143	18	13,151
		≤ 15 years of age	9	0	28	8,913	8,950
	Matched (imputed variable	Employed	564	22	312	6	904
		Unemployed	14	8	26	1	49
		Not in labour force	92	15	881	5	993
		≤ 15 years of age	0	0	0	0	0
	Total		25,587	521	14,718	9,057	49,883

6.3 Potential location and time lag errors

Of the 49,107 elements matched between the coverage survey and the census, 97.7% are found within the basic area around the reference address collected in the survey. The corresponding value is 98.1% for persons who did not indicate any move between Census Day and the day of the survey. It is 83.9% for those who indicated a move (1,512 persons); see absolute numbers in Table 5.

It is worth noting that 9.4% of the persons in NORD who did not move are found close to their reference address but not in exactly the same building. While these problems of exact location have a negligible effect on the census data, they show the difficulty of identifying the buildings sampled when constructing lists of households in the field during the survey, as well as the difficulty of assigning persons to buildings during the processing of the census data. However, a supplementary survey would be needed to evaluate the respective effects of these two difficulties.

Efforts to locate persons who moved indicate that $151 = 145 + 6$ persons were located near their address reported on the day of the survey and not near their Census Day address (9%, weighted). Also, a set of 688 persons in NORD, among the 922 located in the two basic areas, were actually found to be residing in the building on the day of the survey. During the coverage survey, special care was taken regarding questions on addresses on Census Day and on the day of the survey. We therefore believe that the addresses of persons who moved are of better quality in the survey data than in the census data. On this basis, we deduce that out of the 1,512 persons who moved, at least $151 + 688 = 839$ are enumerated in the census at an address that they did not have on the official day of data collection but at an address that they had some time after that date. The exact time lag is not known, since the moving date was not collected in the survey.

Table 5
Comparison of the location of matched persons. The areas are defined for the address on Census Day (according to information collected in the survey) and for the address on the day of the survey (also according to information collected in the survey). Presence in the basic area, the extended area (outside the basic area) or outside the extended area for persons who did not move (stayed) and persons who moved (moved) between the census and the survey

		Day of survey				
		Stayed	Moved			
		Basic area	Basic area	Extended area	Outside extended area	Total
Census	Basic area	46,689	922	69	277	1,268
Day	Extended area	258	42	4	3	49
	Outside extended area	648	145	6	28	179
	Missing	0	15	1	0	16
	Total	47,595	1,124	80	308	1,512

7. Conclusion

Overall coverage deficiencies in the 2000 census of population in Switzerland are of the same order of magnitude as those for the censuses of other countries. However, differences are noted for subgroups (*e.g.*, regions). Of the three components, undercoverage is of great interest, since it not only serves to detect groups of persons not well enumerated, but it also lends itself to analysing location and measurement errors. As to overcoverage estimates, these are limited by the lack of information supplementary to the census for S_E. In the future, they could be improved by collecting supplementary information on characteristics reported on Census Day in a survey of persons in that sample (*e.g.*, location and household type). Net undercoverage estimates are based on several assumptions. The results in large domains seem reliable, but certain risks, notably related to the choice of estimation cells, exist when domains are smaller. For future estimates, we propose to evaluate the model approach applied in the United Kingdom instead of the estimation cells traditionally used in the United States.

An important element to review for future estimates is the choice of the population of interest. The decision to limit that population to persons in private households and the economic residence led to a few problems in the estimates, since it was difficult to delimit that population precisely. In a future estimation, collective households could be excluded so as to avoid practical problems relating to collection but retain all types of residences. The set of records for the economic residence would then be treated as a domain.

Estimating the coverage deficiencies of a census is an ambitious project that has proved to be worthwhile. The results provide information on the quality of the data from the 2000 census and the different coverage problems. Upcoming censuses will essentially be based on registers. Coverage estimates will be based on the experience acquired in making the 2000 estimates, with probable adaptations to take account of the new data collection system.

Acknowledgements

I wish to thank Philippe Eichenberger of the statistical methods unit of the Office fédéral de la statistique for the productive discussions throughout the project. Warm thanks are also extended to Dr. Rajendra Singh and his colleagues in the Decennial Statistical Studies Division of the U.S. Census Bureau for their assistance in developing methods and estimates. I also wish to thank to all census personnel who performed tasks and provided information needed to carry out the project, and to Paul-André Salamin of the statistical methods unit for his careful rereading of the article.

References

ABS (1997). The 1996 census of population and housing. Annual report 1996-97, Australian Bureau of Statistics.

Brown, J.J., Diamond, I.D., Chambers, R.L., Buckner, L.J. and Teague, A.D. (1999). A methodological strategy for a one-number census in the UK, *Journal of the Royal Statistical Society*, Series A, 162(2), 247-267.

Fienberg, S.E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, 18(1), 143-154.

Hogan, H. (1993). The post enumeration survey: Operation and results. *Journal of the American Statistical Association*, 88(423), 1047-1060.

Hogan, H. (2003). The accuracy and coverage evaluation: Theory and design. *Survey Methodology*, 29(2), 129-138.

McLennan, W. (1997). Census of Population and Housing, Data Quality - Undercount. Autralia 1996. Information paper 2940.0. Australian Bureau of Statistics.

Renaud, A. (2001). Methodology of the Swiss Census 2000 Coverage Survey. *Proceedings of the Survey Research Methods Section* [CD-ROM], American Statistical Association.

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Échantillon pour l'estimation de la sur-couverture (E-sample). Methodology Report 338-0019, Federal Statistical Office.

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Methodology Report 338-0027, Swiss Federal Statistical Office.

Renaud, A., and 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. Methodology Report 338-0009, Federal Statistical Office.

Renaud, A., and 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. Methodology Report 338-0023, Federal Statistical Office.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics.

Statistics Canada (1999). Couverture. Rapport technique du recensement de 1996, 92-370-XIF, Statistics Canada.

Statistics Canada (2004). Couverture. Rapport technique du recensement de 2001, 92-394-XIF, Statistics Canada.

Trewin, D. (2003). Census of Population and Housing, Data Quality - Undercount. Autralia 2001. Information paper 2940.0. Australian Bureau of Statistics.

Wolter, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81(394), 338-346.