Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

METHODOLOGY OF THE SWISS CENSUS 2000 COVERAGE SURVEY

Anne Renaud, Swiss Federal Statistical Office
Anne Renaud, SFSO, Statistical Methods Unit, CH-2010 Neuchâtel, Switzerland

Key Words: Census, Coverage Survey, Post-enumeration, Multistage Sampling Design.

1 Introduction

In every population census some persons are missed and others are enumerated more than once, which typically leads to an underestimation of the population count. Global net undercount was estimated to 1.6% in 1990 (Hogan, 1993) and 1.2% in 2000 (Davis, 2001) in the USA, 2.2% in 1991 in the UK (Brown *et al.*, 1999) and 1.6% in 1996 in Australia (ABS, 1996). However, larger undercounts were observed for subgroups of the population, *e.g.* 5% for Hispanics in 1990 in the USA, over 20% for young males in inner cities in 1991 in the UK and 4.3% for men aged 20 to 24 in 1996 in Australia (same references).

The undercount of the Swiss population census is estimated for the first time for the Census 2000. Estimations are expected for large demographic groups, for small and large municipalities and for the census methodologies CLASSIC and TRANSIT, see the Appendix. The estimations are based on the dual system, see Wolter (1986) for the method, Hogan (1992) for an application and Fienberg (1992) for a bibliography.

The post-enumeration required for the dual system is called the Swiss Coverage Survey (SCS). The SCS of 27,000 households is designed to obtain information about the undercount in the various groups of interest.

The purpose of the present paper is to discuss the SCS survey methodology and sampling design.

2 Survey Methodology

The SCS survey methodology may be summarized in four phases: (1) selection of a sample of buildings, (2) enumeration of the households in the sampled buildings, (3) subsampling of buildings to reduce the sample size of households, and (4) interviews with the sampled households.

The sample of buildings consists of two parts: all cantons except Ticino (noted below NORTH) and the canton of Ticino (noted below TICINO). This partition is due to differences in the structure of the data sets available in both areas. The sampling designs are discussed in Sections 3, 4 and 5.

The enumeration of the households in the sampled buildings requires a temporary list from the electronic phone book, correction and adaptation by post employees and capture of the corrections. Maps are made available to find the buildings without identifying postal addresses.

A subsampling is applied after the enumeration in order to get the required sample size of 27,000 households.

The interviews are conducted by phone if a phone number is known and face-to-face otherwise. They are computer assisted, with questions on the household, such as address and number of rooms in dwelling, and questions for each member of the household: first and last names, sex, date of birth, marital status, language, nationality, position in the household, working status, possible second place of residence and change of address between census and SCS days. The SCS questions allow the match with the census, including the identification of moving or second domicile, and the classification into various demographic groups.

The households enumeration took place from January to February 2001 and the interviews were organized from mid April to the end of May 2001, *i.e.* before the third callback of the census. The census response rate was about 95%.

3 Sampling Procedure

The sampling designs NORTH and TICINO are multistage designs with a stratification in the first stage, see Table 1. In both designs, we have an exhaustive selection of the households in the sampled buildings and information is asked about all the persons resident in the selected households. The SCS samples of households and persons are based on data sources independent of the census data, which ensures satisfaction of the inde-

I thank my colleagues Philippe Eichenberger, for the helpful discussions during the project, and Paul-André Salamin, for reviewing this paper.

pendence assumption for the dual system method.

Table 1: Stages of the sampling designs NORTH and TICINO.

Design NORTH	Design TICINO
(1) postal areas (PA)	(1) municipalities
(2) mail delivery areas (MD)	
(3) buildings	(2) buildings
(4) households	(3) households
(5) persons	(4) persons

The multistage selection reduces field costs but leads to a clustering effect. The stratification in the first stage, with postal areas and municipalities as primary sample unit (PSU), decreases the loss of precision due to clustering and controls the sample size in the subgroups of interest. It also gives the opportunity to under-sample or over-sample PSUs depending on the expected difficulties for the field work.

Note that the target population is the resident population of Switzerland, but the surveyed population consists in all residents living in a building of the available lists of buildings. In order to maximize the coverage of the target population, both the buildings coded as inhabited and not inhabited (ex. industrial buildings) are part of the frame.

The required number of persons for acceptable results is fixed at 10,000 net and 11,800 gross, assuming 15% nonresponse. This value is based on experience abroad since no reference data exists for Switzerland yet. We note for instance in Australia an undercoverage of 1.2-1.8% with a standard deviation of 0.3% for groups of about 10,000 persons (ABS, 1996).

The total number of PSUs is set to 305 and the final sample has to reach about 26,000 households for the sampling design NORTH and 1,000 households for the sampling design TICINO.

4 Sampling Design NORTH

The sampling design NORTH aims at satisfying the required sample size in the subgroups of interest and at controlling the variability of the inclusion probabilities in the building selection stage.

4.1 Data

The Swiss directory of buildings is built from the previous census of housings and censuses of entreprises. It was checked by the municipalities prior to the 2000 Population Census. In order to maximize the coverage of the population, the subset used for the SCS contains both the buildings coded as inhabited and the buildings coded as not inhabited. Data about municipalities (estimation of the population size, official language and census methodology) and postal areas (PAs, 6 digits postal codes) are linked by using geocoded information. Data about mail delivery areas (MDs, 2nd stage), households (4th stage) and persons (5th stage) are collected during the SCS procedure.

4.2 Methodology

The sampling methodology uses proportional to size sampling without replacement (PPS) in the first and second stages (PAs and MDs) and simple random sampling without replacement (SRS) in the third stage (buildings). The measure of size in PPS sampling is related to the number of buildings in the sampling unit. In order to get an even distribution of workload, we define the Target Cluster Size TCS as the number of buildings to select in each PA. This amount is then equally distributed among the post employees (MDs). We also take into account the expected difficulties for the field work, such as missing postal addresses.

4.2.1 Selection Probabilities

The three types of sampling units are PAs, MDs and buildings.

The set of PAs, $U_{(PA)} = \{1, ..., i, ..., M\}$, is partitioned into H strata of size M_h, $h = 1, .., H$. The number of buildings in PA i from stratum h is denoted B_{ih}. Consider also $U_{(MD,ih)} = \{1, ..., j, ..., C_{ih}\}$ the set of MDs in PA i from stratum h, and $U_{(building,jih)} = \{1, ..., k, ..., B_{jih}\}$ the set of buildings in MD j from PA i in stratum h.

In the first stage, we select m_h PAs in each stratum h. The selection probability π_{ih} of PA i is

$$\pi_{ih} = m_h \frac{B^+_{ih}}{\sum_{i \in h} B^+_{ih}} \qquad (1)$$

where $B^+_{ih} = \max(B_{ih}, TCS)$ is the measure of size of PA i in stratum h. The total sample of PAs is denoted $s_{(PA)}$ and its size is $m = \sum m_h$.

In the second stage, we select c_{ih} MDs in $U_{(MD,ih)}$. The selection probability $\pi_{j|ih}$ of MD j, given PA i in $s_{(PA)}$, is

$$\pi_{j|ih} = c_{ih} \frac{B^+_{jih}}{\sum_{j \in (i,h)} B^+_{jih}} \qquad (2)$$

where $B^+_{jih} = \max(B_{jih}, TCS/c_{ih})$ is the measure of size of MD j from PA i and stratum h. The total sample of MDs is denoted $s_{(MD)}$ and its size is $c = \sum \sum c_{ih}$.

In the third stage, we select $b_{jih} = \min(B_{jih}, TCS/c_{ih})$ buildings in $U_{(building,jih)}$. The selection probability of building k, given MD j in $s_{(MD)}$, is

$$\pi_{k|jih} = \frac{b_{jih}}{B_{jih}} = \frac{TCS/c_{ih}}{B^+_{jih}} \qquad (3)$$

The total sample of buildings is denoted $s_{(building)}$ and its size is $b = \sum\sum\sum b_{jih}$.

The overall selection probability π_k of building k from MD j in PA i from stratum h reduces to

$$\pi_k = \pi_{k|jih}\pi_{j|ih}\pi_{ih} = m_h \frac{TCS}{\sum_{i\in h} B^+_{ih}} \rho_{ih} \qquad (4)$$

where $\rho_{ih} = B^+_{ih}/\sum_{j\in(i,h)} B^+_{jih} \leq 1$.

We note that the measures of size B^+_{ih} and B^+_{jih} lead to π_k which are independent from MD j and slightly dependent on PA i through the factor ρ_{ih}. For comparison, the overall probability with the measures of size B_{ih} and B_{jih} is given by $\tilde{\pi}_k = m_h TCS/(\sum_{i\in h} B_{ih})\tilde{\rho}_{jih}$ where $\tilde{\rho}_{jih} = \min(c_{ih}B_{jih}/TCS, 1) \leq 1$. The factor $\tilde{\rho}_{jih}$ has two disadvantages. It depends on MD j and varies much more than ρ_{ih} since PAs often contain very small MDs together with bigger ones. Note also that PPS sampling with the chosen measures of size may be seen as an hybrid selection PPS-SRS. Actually, an increase of TCS leads to the selection of more and more units with equal probability.

4.2.2 Stratification and Allocation of PAs

The selection probabilities π_k depend mainly on the sampling design at the PAs level, see (4). Therefore, special attention is to be paid to the stratification and allocation of PAs, while taking into account the expected difficulties for the field work.

Suppose we have a stratification of the population $U_{(PA)}$ such that the strata $h = 1, ..., H$ may be combined into three disjoint groups: easy strata G_1, medium strata G_2, and difficult strata G_3. The allocation of m PAs is done in two phases.

The first phase allocates m PAs to the groups G_1, G_2 and G_3, resulting in m_{G1}, m_{G2} and m_{G3} PAs respectively:

$$\begin{aligned} m_{G_1} &= m(B^+_{G_1} + \varphi)/B^+ \\ m_{G_2} &= mB^+_{G_2}/B^+ \\ m_{G_3} &= m(B^+_{G_3} - \varphi)/B^+ \end{aligned} \qquad (5)$$

where $B^+_{G_r} = \sum_{h\in G_r} B^+_h$, $r = 1,2,3$, $B^+_h = \sum_{i\in h} B^+_{ih}$, and $B^+ = \sum_r B^+_{G_r}$. The parameter φ may be chosen in the interval $[0, B^+_{G_3}[$. It allows a transfer of $m\varphi/B^+$ PAs from the difficult group G_3 into the easy group G_1.

The second phase allocates m_{Gr} PAs to the strata within G_r, $r = 1, 2, 3$, with

$$m_h = m_{Gr}\frac{B^+_h}{B^+_{Gr}} \qquad (6)$$

From (4), (5) and (6), we derive the overall selection probability of building k from MD j in PA i from stratum h:

$$\pi_k = \begin{cases} m\frac{TCS}{B^+}\rho_{ih}\left[1 + \frac{\varphi}{B^+_{G_1}}\right] & \text{if } k \in G_1 \\ m\frac{TCS}{B^+}\rho_{ih} & \text{if } k \in G_2 \\ m\frac{TCS}{B^+}\rho_{ih}\left[1 - \frac{\varphi}{B^+_{G_3}}\right] & \text{if } k \in G_3 \end{cases} \qquad (7)$$

where $\rho_{ih} = B^+_{ih}/\sum_{j\in(i,h)} B^+_{jih} \leq 1$.

In the SCS, the factor ρ_{ih} equals or is very close to 1, which means that the π_k are mainly determined by the group of strata and the parameter φ.

Note that the parameter φ is related to the more intuitive parameter $\gamma \geq 1$, which is defined as the ratio of the maximum (group G_1) to the minimum (group G_3) selection probabilities. If ρ_{ih} is set to 1, the parameter φ is given by $B^+_{G_1}B^+_{G_3}(\gamma - 1)/(B^+_{G_3} + \gamma B^+_{G_1})$.

4.3 Implementation

The Target Cluster Size TCS of buildings in each PA is set to 60 to get an even distribution of workload and to reach the expected total sample size of households.

4.3.1 Sample of Postal Areas

The sampling frame U_{PA} of $M = 3,699$ PAs is stratified into 21 strata depending on the census methodology, the size of the municipality to which the PA belongs and the expected difficulty for the field work.

The census methodologies are: CLASSIC (737 PAs), SEMI-CLASSIC (266 PAs), TRANSIT (2,670 PAs) and FUTURE (26 PAs). The sizes of the municipalities are defined as follows: small (less than 500 inhabitants; 944 PAs), middle (between 500 and 10,000 inhabitants; 2,406 PAs) and large (at least 10,000 inhabitants; 349 PAs). Using as criteria the proportion of buildings without postal address and the proportion of buildings declared as not inhabited, we define four types of expected difficulties: easy (1,669 PAs), medium (1,201 PAs), difficult 1 (366 PAs) and difficult 2 (463 PAs). The groups of strata are G_1 for easy, G_2 for medium, and G_3 for difficult 1 and 2. Note that $B^+_{ih} = TCS$ for PAs in difficult 2 (small PAs) and $B^+_{ih} = B_{ih}$ otherwise. The number of PAs per stratum ranges between 25 and 1,126.

The allocation of PAs is based on simulations carried out independently for each census methodology. The expected number of households and persons, as well as the π_k, are analyzed for varying φ and sample sizes m. The final values are set as follows: 80 PAs for CLASSIC, 37 PAs for SEMI-CLASSIC, 166 PAs for TRANSIT, and 4 PAs for FUTURE. The parameter φ is chosen in order to constrain the parameter γ, *i.e.* the ratio of the maximum to the minimum selection probabilities at the building stage, in the range 2.1-2.3. No groups are defined for the FUTURE methodology.

Note that one stratum has been divided into 2 strata before selection since inclusion probabilities of some PAs were larger than 1 in PPS sampling. The final sampling plan has 22 strata in which 2 to 60 PAs were selected. The total sample of PAs $s_{(PA)}$ has size $m = 285$.

4.3.2 Sample of Mail Delivery Areas

The sampling frame of 1,489 MDs is built for the PAs in the first stage sample $s_{(PA)}$. The MDs corresponding to real mail delivery areas are used in priority, but unique artificial MDs have been created in PAs with less than 3 real areas or where less than 80% of the buildings could be assigned to a real area. Note that ρ_{ih} ranges between 0.97 and 1 over all the sampled PAs, which confirms the small influence of this factor.

We allocate $c_{ih} = 1$ MD in PA i with $C_{ih} = 1$ (157 PAs) and $c_{ih} = 3$ MDs in PA i with $C_{ih} \geq 3$ (128 PAs). Departure from PPS sampling was unavoidable in some PAs, *e.g.* when $c_{ih} = C_{ih} = 3$. Special attention is then paid to the allocation of buildings to compensate part of this departure, see Section 4.3.3. The total sample of MDs $s_{(MD)}$ has size $c = 541$.

4.3.3 Sample of Buildings

The sampling frame of 123,276 buildings is constructed for the MDs in the second stage sample $s_{(MD)}$.

The allocation is defined by $b_{jih} = \min(B_{jih}, TCS/c_{ih})$ in all MDs, except in the PAs with $c_{ih} = C_{ih} = 3$. In those special 16 PAs, we allocate $TCS = 60$ buildings proportional to the number of buildings in the MD. The number of buildings b_{jih} may however not exceed 25 and the possible remaining buildings are distributed among the smaller MDs. This special procedure partly compensates for the departure from PPS sampling in the MDs stage. The total sample of buildings $s_{(building)}$ has size $b = 16,792$.

4.3.4 Subsampling of buildings

Due to non-response in the PA stage, the list of households was not created in 2 of the sampled PAs. The number of households in the remaining PAs is about 12% larger than the required 26,000. As information is gathered about all the households in the selected buildings, we choose a subsample of the buildings using SRS. A substratification is introduced in 5 touristic PAs with many holiday homes, which are then selected with a smaller probability. The subsample $s^{(2)}_{(building)}$ has $b^{(2)} = 14,697$ buildings. It contains 26,369 households which form the sample of households for the sampling design NORTH.

5 Sampling Design TICINO

The sampling design TICINO aims at satisfying the required sample sizes in the canton of Ticino and at limiting the variability of the inclusion probabilities in the building selection stage.

5.1 Data

The list of buildings is based on an exhaustive survey among the property-owners. Two versions of the list have been made available (temporary and final). Population size estimates are available for the 245 municipalities (MUs). Data about households (3rd stage) and persons (4th stage) are collected during the SCS procedure.

5.2 Methodology

The sampling methodology uses PPS sampling in the first stage (MUs) and SRS sampling in the building selection stage. The measure of size is the number of buildings in the sampling unit and TCS is the target number of buildings to select in each MU.

5.2.1 Selection Probabilities

The two types of sampling units are MUs and buildings.

The population of MUs, $U_{(MU)} = \{1, ..., i, ..., M\}$, is partitioned into H strata of size M_h, $h = 1, ..., H$. The temporary and final numbers of buildings in MU i from stratum h are denoted by B_{ih} and B^*_{ih} respectively, and $U_{(building,ih)} = \{1, ..., k, ..., B^*_{ih}\}$ is the final set of buildings in MU i from stratum h.

In the first stage, we select m_h MUs in each stratum h. The selection probability π_{ih} of MU i is

$$\pi_{ih} = m_h \frac{B_{ih}}{\sum_{i \in h} B_{ih}} \tag{8}$$

The total sample of MUs is denoted $s_{(MU)}$ and its size is $m = \sum m_h$.

In the second stage, we select $b_{ih} = \min(TCS, B^*_{ih})$ buildings in $U_{(building,ih)}$. The selection probability $\pi_{k|ih}$ of building k, given MU i in $s_{(MU)}$, is

$$\pi_{k|ih} = \frac{b_{ih}}{B^*_{ih}} = \min(TCS/B^*_{ih}, 1) \qquad (9)$$

The total sample of buildings is denoted $s_{(building,TI)}$ and its size is $b_{TI} = \sum\sum b_{ih}$.

The overall selection probability $\pi_{k,TI}$ of building k in MU i from stratum h reduces to

$$\pi_{k,TI} = m_h \frac{b_{ih}}{\sum_{i \in h} B_{ih}} \frac{B_{ih}}{B^*_{ih}} \qquad (10)$$

In the SCS, few MUs have less than TCS buildings. Therefore, b_{ih} equals to TCS in most cases and the variability of the selection probability $\pi_{k,TI}$ within a given stratum is mainly due to the ratio B_{ih}/B^*_{ih} between the temporary and final sizes of the lists of buildings.

5.2.2 Stratification and Allocation

The stratification and allocation in the first stage take into account the expected difficulties in the survey procedure. We do not use a special method of allocation as the variability in the selection probability is mainly due to the unknown ratio B_{ih}/B^*_{ih}.

5.3 Implementation

The target cluster size TCS of buildings is set to 60 in each MU.

The sampling frame $U_{(MU)}$ of $M = 250$ MUs (municipalities or land register parts) is partitioned into 6 strata depending on the population size and the availability of postal addresses. The sizes are: small (less than 1,000 inhabitants; 168 MUs), middle (between 1,000 and 8,000 inhabitants; 79 MUs) and large (more than 8,000 inhabitants; 3 MUs). We have three types of address problems for small MUs, 2 types for middle MUs and 1 type for large MUs. The sample of MUs $s_{(MU)}$ has size $m = 20$ and holds a temporary list of $b = 16,127$ buildings.

The final set of buildings has 24,728 buildings in $s_{(MU)}$. The ratio between final and temporary sizes ranges between 1.2 and 2.56 depending on the MU. The allocation of buildings in each sampled MU is given by $b_{ih} = TCS$ since all of them have more than TCS buildings. The final sample of buildings $s_{(building,TI)}$ has size $b = 1,200$.

A subsampling of buildings is applied to 5 touristic municipalities to decrease the number of buildings declared as temporarily inhabited and without phone number. The final sample $s^{(2)}_{(building,TI)}$ of $b' = 1,180$ buildings contains 962 households which form the final sample of households for the sampling design TICINO.

6 Overall Sample

The SCS sample of 27,331 households is the union of the samples for the designs NORTH and TICINO. It is distributed over all of Switzerland, see Figure 1. The linguistic distribution is 67% German, 28% French and 5% Italian. A proportion of 88% households can be contacted by phone and 12% have to be contacted in face-to-face interviews.

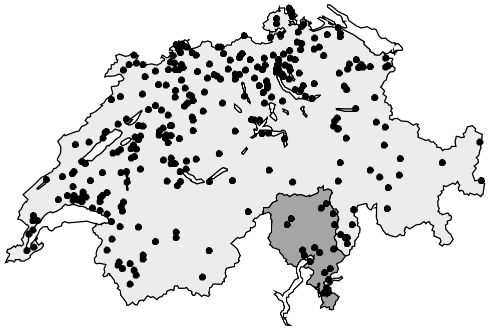

Figure 1: Geographical distribution of SCS PSUs; NORTH (light grey) and TICINO (dark grey). Source: Basemap ©SFSO GEOSTAT / S+T.

The expected number of persons for acceptable results is reached for the CLASSIC and TRANSIT methodologies, for the medium and large municipalities (overall and TRANSIT), see Table 2, as well as for the German and French speaking areas (overall and TRANSIT).

Table 2: Estimated gross number of persons in the SCS as a function of the census methodology and size of the municipality (classification as in plan NORTH).

Census	Size of the municipality			
meth.	Small	Medium	Large	Total
C	5939	6531	-	12470
S	1165	5336	-	6501
T	302	23475	19723	43500
F	-	340	584	924
Ticino	320	1116	488	1924
Total	7726	36798	20795	65319

The sampling weights $w_k = 1/\pi_k$ and $w_{k,TI} = 1/\pi_{k,TI}$ vary from 4 to 328 depending on the census methodology, the group of strata, departure from PPS sampling and subsampling in touristic units. The coefficient of variation of the households sampling weights equals to 37% for CLASSIC ($16 \leq w_k \leq 54$), 15% for SEMI-CLASSIC ($4 \leq w_k \leq 61$), 45% for TRANSIT ($45 \leq w_k \leq 326$), 0% for FUTURE ($w_k = 71$), and 45% for TICINO ($68 \leq w_{k,TI} \leq 328$).

7 Conclusion

The Swiss Coverage Survey (SCS) is based on a multistage stratified sampling design with two disjoint sampling frames in the first stage: postal areas in NORTH and municipalities in TICINO. An enumeration of households is conducted at the building level, and information is asked of all persons in the selected households.

The sampling design NORTH is developped in order to control, for each census methodology, the variability of the inclusion probabilities in the household stage. The proposed sampling plan gives selection probabilities which depend on the group of strata and only slightly on the PA, but are independent of the stratum and of the MD. We note however that the implementation requires some adaptations, mainly due to the inapplicability of PPS sampling in some units at the PA and MD levels.

The sampling design TICINO has to deal with temporary and final lists of buildings. The main consequence is the variability of selection probabilities within strata.

The overall sample of 27,331 households achieves the required sample sizes in the various subgroups of the population.

Appendix: The Swiss 2000 Census

The Swiss population and housing census took place in 2000 with Census day on 5 December 2000. Information is collected for all 7.1 millions residents of Switzerland. The Swiss Federal Statistical Office is responsible for the decennial censuses but data collection is under the responsibility of municipalities. In 2000, municipalities had to make a choice between various census methodologies in order to help build up a coordinated register of inhabitants for Switzerland. The methodologies may be summarized as follows:

CLASSIC (C): enumerators visit the households to bring and take back the questionnaires (690 municipalities);

SEMI-CLASSIC (S): preprinting of the questionnaires using the register of inhabitants, dispatch by mail and visit of enumerators to take back the questionnaires (223 municipalities);

TRANSIT (T): preprinting of the questionnaires using the register of inhabitants, dispatch and return by mail (1717 municipalities);

FUTURE (F): same as TRANSIT with link between households and dwellings in the register of inhabitants (21 municipalities).

The preprinting of the questionnaires, the mail dispatch and the check of mail return was centralized for all Switzerland but Ticino. The canton Ticino (245 municipalities) organized the census on its territory by using a methodology similar to TRANSIT.

References

ABS (1996), Census Population and Housing: Data Quality - Undercount. Information paper 2940.0, Australian Bureau of Statistics

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, *Journal of the Royal Statistical Society A*, 162 (2), 247-267

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, Washington, D.C.

Fienberg, S. E. (1992) Bibliography on Capture-Recapture Modelling with Application to Census Undercount Adjustment, *Survey Methodology*, **18**, 1, 143-154

Hogan, H. (1992) The 1990 Post-Enumeration Survey: An Overview, *The American Statistician*, **46**, 4, 261-269

Hogan, H. (1993) The 1990 Post-Enumeration Survey: Operation and Results, *Journal of the American Statistical Association*, **88**, 423, 1047-1060

Wolter, K. M. (1986) Some Coverage Error Models for Census Data, *Journal of the American Statistical Association*, **81**, 394, 338-346