OPEN-ENDED SEGMENTS: VARIATION ON AREA SEGMENTING AND LIST FRAME SUPPLEMENTATION

Irene C. Montie and William G. MacKenzie, U.S. Bureau of the Census

I. INTRODUCTION

The intent of this paper is to describe the development of a sample survey coverage methodology which was originally used to reduce undercoverage bias in the Current Population Survey (CPS), but appears to have the potential for broader utilization in sample designs. The labels associated with this methodology are Successor Check Procedure and Open-Ended Segments. These two terms reflect a difference in utilization rather than method. The successor check is more closely associated with supplementing existing frames, whereas open-ended segments are an extension of the half-open interval concept for providing a separate source for obtaining sample units.[1]

The present Successor Check Procedure (called a structure successor check) works as follows. One selects n of the N structures from the sample frame. Each selected structure is called a sample reference structure. Using the sample reference structure as a starting point an enumerator in the field lists structures in a prescribed path of. travel order until k structures (not including the sample reference structure) on the sample frame have been listed. Any structures listed that are not on the list frame but are in the population of interest are called inscopes. In this manner n open-ended segments (successor check strings) of length k successors are generated. This successor check resembles in some ways an area segment in which all structures are listed within a relatively small land area with well-defined boundaries. Also, if one selects for interview all or a sample of inscopes identified by the successor check, then this procedure can be used as an alternative method of list frame supplementation.

It is not the intent of this paper to fully describe the sample design or particular procedures of the surveys referred to here. Extensive documentation is provided elsewhere.[2] However, an overview of Census Bureau experience with the successor check procedures is provided in Section II as background and linkage to the research efforts discussed in Section III.

II. EXPERIENCE IN THE USE OF SUCCESSOR CHECK PROCEDURES

A. Successor Check in the 1960 Surveys

During the 1960's the Census Bureau used a unit successor check procedure in B segments (addresses from list frame) for the Current Population Survey (CPS) and the Health Interview Survey (HIS). This procedure differed in two respects from the structure successor check described in the introduction. First of all, reinterview units were designated as reference points. Secondly, the enumerator listed units in a prescribed path of travel from the reference point until obtaining a unit that was in the sample frame (called the successor). Listed units between the reference point and the successor that met one of the following conditions were considered inscopes:

Units missed in the census
Mobile homes - in parks or at large - put in place since the census
Mobile homes excluded from the census by definition, but eligible for sample surveys
Houses moved to their present location since the census
Units converted from nonresidential use since the census
Special places missed or created since the census.

The results of the unit successor check were reasonably satisfactory in terms of reducing nonsampling errors. For CPS the check identified inscope units representing about 2.6 percent of the inventory. For HIS the yield was about 2 percent.[3] However, the complexity of this type of successor check, especially in terms of path of travel in multi-unit structures, resulted in high preparatory (including the preparation of detailed maps) and high field costs (field work was done by supervisory personnel).

B. Successor Check in 1970 Surveys

Based on experience during the 1960's with unit successor checks in CPS and HIS, the Census Bureau decided to implement structure successor checks for surveys in the 1970's. This type of structure check is defined in the introduction. It has been used for list frame supplementation in address type segments of the Annual Housing Survey (AHS)-SMSA and for coverage evaluation in the Survey of Income and Education (SIE). These are described below, along with a partial successor check used in the AHS-National sample.

1. Successor Check in AHS-SMSA Samples

To reduce undercoverage of census misses and other types of units not fully represented in the AHS-SMSA samples, a structure successor check was designed.[4] An intended string length of k=eight successor structures was used. In most cases this string length was attained. Where the string length was not attained an adjustment in the weighting of inscope units was necessary.

The use of a structure, rather than a unit successor check, reduced the complexity of the field operation. From the experience gained during the first year of implementation, field procedures, instructions and training have been revised. These revisions should result in greater efficiency in the field operation for the second year of implementation.

2. Successor Check in the AHS-National Sample

CEN SUP (Census Supplement) segments are used in national Census Bureau samples to represent units missed in the Census. However, there was no procedure for reducing undercoverage of the other types of inscope units. For CPS and most of the recurring surveys done at the Census Bureau the impact of this undercoverage is considered trivial. This is not true for the AHS.[5] Therefore, in AHS-National (1976) a partial successor check (k=8) was undertaken to provide representation of these units.[6] The yield for mobile homes and houses moved in was reasonably good. However, the check was not that effective in eliminating undercoverage of structures converted from nonresidential use.

3. Successor Check in SIE

In order to meet precision requirements of the survey within the funds available it was decided to take unit samples in all areas. Since CEN SUP segments were not available for rural areas, a structure successor check (k=4) was done by crew leaders for coverage evaluation. About 6,000 sample reference structures were selected which meant that about 25,000 structure addresses had to be matched in rural areas six years after the Census. The matching was difficult and frequently required reconciliation in the field. In addition, there was some evidence of wrong units selected for interview, which resulted in the wrong reference structure for the successor check. Preliminary results also indicated a very high yield of inscope units. Thus, a number of validation and follow-up studies were conducted.[7] One of these, the Year Built Study, is briefly discussed in Section III.

III. RESEARCH RELATED TO SUCCESSOR CHECK/OPEN-ENDED SEGMENTS

A. Introduction

Until the early 1970's most surveys conducted by the Census Bureau were based on national samples. More recently, however, the Bureau has responded to the need for small area data, such as the AHS-SMSA sample, and conversely, for very large samples, such as those for SIE. These kinds of surveys require greater variety in sample design and methodology.

For example, there are special needs for unclustered samples in small area design, and for greater emphasis on cost analysis for samples of very large size.

A procedure that has the potential to meet special needs of both small areas and large areas is the structure successor check. Some Census Bureau studies relating to open-ended segments are described below.

B. Estimation and Variance

For many Bureau surveys the principal sample frame is the Census list of residential addresses (used for address segments). Units within a structure are listed sequentially. Conceptually, this type of frame contains N structures, each of which can be a reference structure for an open-ended segment of size k successor structures. If the N structures are in path of travel order on the frame, then one can obtain the exact probability of inclusion for any inscope unit identified by the specific sample of strings selected. However, these N structures usually do not appear on the frame in successor check path of travel order.

From the frame a sample of open-ended segments of length k may be chosen. For the AHS successor checks a systematic sample of n housing units was selected from the AHS sample units in address segments. The sample unit defined the reference structure. From each reference structure an open-ended segment of k length was formed. This procedure translates into a systematic ppes sample of successor check strings of length k because each string's probability of selection depends upon the number of units M_i in the string's sample reference structure (i.e. a string generated by a reference structure with M_i=6 has three times the chance of selection as a string generated by a reference structure with M_i=2).[8] Within these strings inscope units (structures) are identified. Estimation and variance related to inscope units (structures) are discussed in this section.

1. Present Estimation of Inscope Totals

In considering the estimation of inscope totals let y_{ui} be the number of inscope units found by the i^{th} sample string, k be the string length, n be the number of sample reference structures selected, N be the number of structures in the sample frame, and π_i be the probability of inclusion in the sample of the i^{th} selected reference structure. We do not assume that the N structures on the frame are in successor check path of travel order. Then it can be shown that

$$\hat{Y}_u = \sum_{i=1}^{n} \frac{y_{ui}}{\pi_i} \cdot \frac{1}{k} \qquad (1)$$

provides an unbiased estimate of Y_u, the total number of inscope units.[9]

If we define M_i to be the number of units in structure i on the sample frame and $M_o = \sum_{i=1}^{N} M_i$, then for the sample of open-ended segments defined in the introduction we have that $\pi_i = n\, z_i$, where $z_i = M_i/M_o \leq \frac{1}{n}$. Thus for our present systematic ppes sample of successor strings we have used the unbiased estimate of Y_u:

$$\hat{Y}_u = \sum_{i=1}^{n} \frac{y_{ui}}{n\, z_i} \cdot \frac{1}{k} \qquad (2)$$

We could similarly define an unbiased estimator $\hat{Y}_s$ for Y_s, the total number of inscope structures.

2. Present Variance Estimation

Due to the nature of the sample frame it is very difficult to calculate the joint inclusion probability of any two selected strings. Thus we don't use the familiar Yates and Grundy estimate of the variance of $\hat{Y}_u$. Instead we note that if a pps with replacement sample of n sample reference structures were selected, then $\hat{Y}_u$ is once again an unbiased estimator of Y_u and furthermore (see Chapter 9 of Cochran's Sampling Techniques)

$$\hat{V}(\hat{Y}_u) = \frac{1}{k^2} \sum_{i=1}^{n} \left(\frac{y_{ui}}{z_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{y_{ui}}{z_i} \right)^2 \Big/ n(n-1), \qquad (3)$$

is an unbiased estimator of $V(\hat{Y}_u)$. Now for the AHS National and SMSA successor check samples $\frac{n}{N}$ and π_i are small. Since the variance for a without replacement systematic pps sample estimator is less than the variance for a with replacement pps sample estimator, we use equation (3) as our estimator of $V(\hat{Y}_u)$.[10]

Utilizing equation (3), estimated coefficients of variation $[\hat{CV}(\hat{Y}_u)]$ of $\hat{Y}_u$ have been calculated for 13 AHS-SMSA successor checks of length K=8. In general, the $\hat{CV}(\hat{Y}_u)$'s were less than 0.4 with $\frac{n}{N}$ less than 0.003. Estimated design effects (DEF's) from these data are about 2.2.[11]

Inscope units usually represent between 1 and 3 percent of the housing units in a jurisdiction. Thus, when open-ended segments are used to supplement the existing sample frame for inscope units, they usually have little effect on the variance of the overall sample estimate, and yet improve coverage.

C. Successor Check Modeling

It is of interest to study the distribution of inscope structures by strings. If the number of inscope structures per string can be reasonably well approximated by a known distribution, then we could derive better estimates of design effects, workload, and costs.

1. First Attempts At Modeling

The proportion of inscope structures in the population is small, about two percent nationally. Thus, for moderate size k the percentage of strings with inscope structures should also be small. At first we speculated that t, the number of inscope structures in a string, might follow a Poisson distribution with parameter λ or a negative binomial distribution with parameters k, p. By making use of various AHS-National, AHS-SMSA, and SIE successor check data sets, unbiased estimates of λ and p were obtained. However, in each case the Poisson and negative binomial models provided poor fits to the observed distributions of inscope structures per string.[12]

2. Development of Markov Dependent Bernoulli Trials (MDBT) Model

The first two models had the advantage of being simple. However, observation of the actual distributions of inscope structures in strings shows that the inscopes often cluster together.

This significant observation of the clustering of inscope structures leads us to consider a model where each string is assumed to constitute a series of Markov Dependent Bernoulli Trials. We define a one-step transition matrix T and two states: state 1 being a census structure and state 2 being an inscope structure, where

$$T = \begin{pmatrix} 1-p & p \\ \alpha & 1-\alpha \end{pmatrix} \quad \text{where}$$

1-p = conditional probability of a census structure following a census structure
p = conditional probability of an inscope structure following a census structure
α = conditional probability of a census structure following an inscope structure
1-α = conditional probability of an inscope structure following an inscope structure

The only data readily available for estimating T are from AHS-SMSA successor checks (k=8). For each of these data sets we define the following:

V = (number of complete strings) x 8
U = number of times an inscope structure follows a census structure
W = number of inscope structures
X = number of times an inscope structure follows an inscope structure

Then let $\hat{p}$ = U/V and $\hat{\alpha}$ = (W-X)/W. We then let $\hat{T}$ estimate T where

$$\hat{T} = \begin{pmatrix} 1-\hat{p} & \hat{p} \\ \hat{\alpha} & 1-\hat{\alpha} \end{pmatrix}$$

Now for the 16 small SMSA's, at most 208 strings of length 8 successors were listed. These sample sizes were quite small and thus no attempt at data fitting was made. However, for the four large SMSA's sample sizes were considerably larger. Tables 1 & 2 show the result of fitting the MDBT, Poisson, and negative binomial models to two of these four large SMSA successor check data sets. For the two large SMSA's shown in Tables 1 & 2 and for those two large SMSA's whose data are not shown here the MDBT model provides fits that are superior to those of the other two models.[13] It should be noted that goodness of fit tests for the MDBT model indicate that it fits the data reasonably well.

3. Use of MDBT Model to Estimate Successor Check Design Effects

If one believes that the MDBT model with transition matrix T provides a reasonably good approximation to Z, the distribution of inscope structures per string, then it can be used to estimate design effects. S_z can be calculated from the MDBT model. Furthermore, if r is the proportion of inscope structures for the population, then the successor check's design effect for its estimator of total structures equals the following:

$$DEF_k = \left(\frac{S_z^2}{k^2}\right) \Big/ \left(\frac{(r)(1-r)}{k+\bar{t}}\right) ,$$

where $\bar{t}$ is the MDBT model's expected number of inscope structures per string. For strings with k=8, $DEF_8 \doteq 1.8$ which yields an intraclass correlation coefficient δ of about .11. This result is in close agreement with more standard estimates of δ.

4. Use of MDBT Model to Estimate Successor Check Workload and Costs

Using the MDBT model, a distribution of string lengths can be predicted. Thus enumerator workloads are known approximately, which means total cost for the successor check can be estimated.

D. Year Built Study

It has become increasingly more difficult to obtain a representative sample of new construction from building permits. The problem is most severe for surveys that are introduced in later years of the decade (and require permits back to 1970). However, even current programs are affected. For example, some permit offices interfile addresses and discard the permits before they can be sampled.

Open-ended segments might be used to obtain new construction. However, in analyzing the results of SIE successor checks there was some concern about the accuracy of the data on new construction. An evaluation of this accuracy was done in the Year Built Study.[14] The preliminary results of this study, which analyzed the data from the 1976 AHS-National successor check sample of approximately 1,500 strings of size eight successors, indicate that year built data from open-ended segments are no less accurate than other year built data obtained by census enumeration or by area segmenting operations. Thus it appears that open-ended segments could replace building permits as the source of obtaining new construction in sample surveys.

E. Application to New York City Vacancy Survey

1. Related Background

The estimator $\hat{Y}_u$ defined by equation (1) has the additional advantage of making for easy processing of sample data from open-ended segments. Each inscope unit in an open-ended segment has the same weight, i.e. $M_0/(M_i \cdot n \cdot k)$. Since the weights are a function of M_i, they have large variability, especially in urban areas. Methods to reduce this variability were first considered during the design of a possible successor check sample to supplement the basic sample frame of the 1978 New York City Vacancy Survey. Open-ended segments were not used in the survey because of a limited budget; however, the methods proposed are of interest, and thus are discussed briefly in this section.

2. Differential Subsampling Method

M_i, the number of units in a sample reference structure, would vary considerably in New York City. From the AHS-SMSA New York successor check about 100 strings (k=8) had at least one inscope unit. For these strings the correlation between M_i and y_{ui}, the number of inscope units in the ith sample string, was about .58. Thus, as one might expect, y_{ui}/M_i was fairly stable around the value of 1. These observations indicated that for the New York City Vacancy Survey interviews, a differential subsample of inscope units could be selected, since the characteristics of these units probably don't vary much within a structure. The extreme case would be to take a 1 in M_i subsample of inscope units for interview, with a minimum of one interview from a string with inscope unit(s). Some type of subsampling certainly makes sense operationally; e.g. for an inscope structure containing 400 units it would not be reasonable to interview all of the inscope units.

3. Post Stratification Method

Rather than relying totally on subsampling to reduce the wide variation in weights for interviews arising from the successor check, we have considered the use of a post-stratification technique to reduce the variability.[15] The method here is to select the sample reference structures by the usual method and then to group the sample reference structures by size into strata. Within each stratum select a 1 in C_{strata} of sample reference structures for field generation of open-ended segments, where C_{strata} is proportional to some function of M_i, $f(M_i)$. The question arises as to what $f(M_i)$ should be. Certainly C_{strata} should increase as $\overline{M}_i$ for the strata increase. Thus, before any subsampling the weight of an inscope unit would be $(M_0 \cdot C_{strata})/(k \cdot n \cdot M_i)$. The method would also reduce the number of strings that would have to be field listed, and would still allow for some subsampling as described in section 2 above (although to a lesser extent).[16]

IV. SUMMARY AND RELATED CONSIDERATIONS

A. Summary

In summary, the successor check or open-ended segment methodology has been used with varying degrees of success. It has the potential for providing representation of new construction, for replacing area segments in certain situations and for reducing the nonsampling errors due to undercoverage of certain types of units.

Recent cost studies within the Bureau indicate that for the same reliability on estimates of new construction the successor check costs are about the same as the present method of sampling for new construction in address segments. Similarly, for the same reliability on estimates in area segments the successor check costs are about the same as present methods of area segmenting. We are quite satisfied with the successor check as a method to reduce undercoverage bias. More investigations into the feasibility of using open-ended segments as a replacement for permit segments and area segments are presently being conducted at the Census Bureau.

B. Related Considerations

In undertaking this paper the authors communicated with Leslie Kish, Frank Yates, and W. Edwards Deming to discuss the development and operation of the half-open interval, as well as applications proposed by the Bureau. The half-open interval has been used by private firms to create segments without prelisting and to capture new construction and other unlisted units in a previously listed area.

REFERENCES AND EXPLANATION

1. See Deming, W. Edwards: Sample Design in Business Research; (New York: John Wiley and Sons, Inc., 1960.) Chapter 12.

2. See The Current Population Survey: Design and Methodology, Technical Paper 40; U.S. Bureau of the Census; January 1978. Specifically, for a detailed description of various types of segments see pp. 16-18.

3. See internal Census Bureau memorandum from Joseph Waksberg to William Hurwitz, "Two Memoranda Concerned with the Successor Check Operation," dated 1/5/66; and attachment dated 12/22/65.

4. For a complete description of the coverage improvement procedures introduced into the Annual Housing Survey samples, see Coverage Improvement in the Annual Housing Survey, Irene C. Montie and Dennis J. Schwanz; Proceedings of the Social Science Section of the American Statistical Association Annual Meetings, Chicago, Illinois, 1977.

5. See Montie and Schwanz, Coverage Improvement, pp. 7 and 8.

6. For a partial structure successor check census misses are not considered inscope, since they are already represented by CEN SUP segments.

7. See Harold Nisselson and Eli Marks, Problems of Nonsampling Error in the Survey of Income and Education: Coverage Evaluation, paper presented at the American Statistical Association Meetings, Chicago, Illinois, 1977.

8. For a discussion of systematic pps sampling see Hartley, H. O. and Rao, J. N. K. (1962), "Sampling With Unequal Probabilities and Without Replacement," Annals of Mathematical Statistics, 33, 350-374. For a discussion of ppes sampling see Cochrane, William G., Sampling Techniques, New York: John Wiley & Sons, Inc., 1963, pp. 252+.

9. The original proof is due to William Smith III, mathematical statistician with the Census Bureau. If one expresses

$\hat{Y}_u = \frac{1}{k} \sum_{i=1}^{N} t_i \frac{y_{ui}}{\pi_i}$ where $P(t_i=1)=\pi_i$ and

$P(t_i=0)=1-\pi_i$; then it is easy to show that

$E(\hat{Y}_u)=Y_u$.

10. This is true assuming no systematic effects. See page 370 of Hartley and Rao (1962). Also note that equation (5.20) of that paper would have been preferred as an estimate of $V(\hat{Y}_u)$.

However, $\sum_{i=1}^{N} (M_i)^2$ was not readily available.

11. See internal Census Bureau memorandum from William MacKenzie to David Bateman entitled "Successor Check Research and Its Application to RAV Planning," December 13, 1977, pp. 10-12. Also note that:

$$\left(\text{Design Effect}\right) = \left(\frac{\text{True variance of an estimate under sampling scheme used}}{\text{True variance of an estimate under simple random sampling}}\right)$$

12. See tables 1-6 of internal Census Bureau memorandum from William MacKenzie to Henry Woltman entitled "Successor Check Modeling: Development of Markov Dependent Bernoulli Trials Model," March 7, 1978.

13. Further details can be found in internal Census Bureau memorandum from MacKenzie to Woltman entitled: "Successor Check Modeling: Further Development of Markov Dependent Bernoulli Trials Model," March 21, 1978.

14. See also internal Census Bureau memorandum (draft) from MacKenzie to Bateman entitled "Some Preliminary Results from the Year Built Study," October 11, 1977.

15. If a list of structures with the number of units in each were readily available, it would have been more efficient in terms of variance to have stratified by structure size before selecting sample reference structures (and then using different sampling intervals in the different strata).

16. See also internal Census Bureau memorandum entitled "Possible Coverage Improvement Designs for New York City Vacancy Survey," May 18, 1977 from MacKenzie to O.S. Cullimore.

Table 1. Comparison of Markov Dependent Bernoulli Trials With Poisson and Negative Binomial Models For the Houston AHS SMSA Successor Check, k=8 1/

Number of Inscope Structures Found in the String	Observed Number of Strings	Expected Number of Strings Under:		
		Markov Dependent Bernoulli Trials Model 2/	Poisson Model 3/	Negative Binomial Model 4/
0	339	336.3	329.2	329.5
1	21	24.5	36.7	36.2
2	6	5.6	2.0	2.2
3	1	1.2	.1	.1
4+	1 5/	.4	0	0
Total	368	368.0	368.0	368.0

1/ Not adjusted for fact that strings were not selected with equal probability.

2/ $\hat{T} = \begin{pmatrix} 1-\hat{p} & \hat{p} \\ \hat{\alpha} & 1-\hat{\alpha} \end{pmatrix} = \begin{pmatrix} .98879076 & .01120924 \\ .80487805 & .19512195 \end{pmatrix}$

3/ $\hat{\lambda} \doteq .1114$

4/ $\hat{p} \doteq .9863$

5/ Includes one string with five inscope structures.

Table 2. Comparison of Markov Dependent Bernoulli Trials With Poisson and Negative Binomial Models For the Seattle AHS SMSA Successor Check, k=8 1/

Number of Inscope Structures Found in the String	Observed Number of Strings	Expected Number of Strings Under:		
		Markov Dependent Bernoulli Trials Model 2/	Poisson Model 3/	Negative Binomial Model 4/
0	436	426.6	387.9	390.0
1	52	59.6	115.2	111.7
2	20	22.5	17.1	18.0
3	6	8.4	1.7	2.1
4	3	3.1	.1	.2
5+	55/	1.8	0	0
Total	522	522.0	522.0	522.0

1/ Not adjusted for fact that strings were not selected with equal probability.

2/ $\hat{T} = \begin{pmatrix} 1-\hat{p} & \hat{p} \\ \hat{\alpha} & 1-\hat{\alpha} \end{pmatrix} = \begin{pmatrix} .97509579 & .02490421 \\ .68387097 & .31612903 \end{pmatrix}$

3/ $\hat{\lambda} \doteq .2969$

4/ $\hat{p} \doteq .9642$

5/ Includes two strings with five inscope structures, one with six, one with eight, one with nine.