COVERAGE IMPROVEMENT IN THE ANNUAL HOUSING SURVEY

Irene C. Montie and Dennis J. Schwanz, U.S. Bureau of the Census

ACKNOWLEDGEMENT

The intent of this paper is to describe some methodologies used to reduce undercoverage bias in the Annual Housing Survey, National Sample.

Identification of deficiencies in the sampling frames and development of methodologies to close the gaps have been a joint effort within Statistical Methods Division. The work of the various statisticians are so commingled that individual acknowledgement would mask the synergistic process. Therefore, in this paper footnotes are limited to peripheral publications. The reader is directed to Census Bureau planning, procedural, and results memoranda from which this effort is derived. However, some recognition must be given to Dave Bateman and his sample design staff, Len Baer and his sampling systems staff, the computer programming staff under Dave Diskin, the staffs of sampling procedures branch and special surveys branch, and of course the Division Chief, Charlie Jones, for his relentless encouragement in carrying out the coverage improvement procedures. Special notice must also be given to the Field Division staff, who implemented the procedures, with precision and concern, and to Arthur F. Young Chief of Housing Division, Elmo Beach and his staff, and the staff of Demographic Surveys Division for their knowledgeable assistance, critique, and support of the entire undertaking.

I. SURVEY BACKGROUND

The Annual Housing Survey (AHS-National Sample) is a sample survey conducted annually by the Bureau of the Census for the Department of Housing and Urban Development to obtain national and regional estimates of the size and composition of the housing inventory in the United States. The series estimates year to year changes in the inventory due to losses and new construction (including mobile home placements), and provides characteristics of the total inventory.

The survey was first conducted in 1973. At that time approximately 59,300 sample units were contacted. The 1974 sample included 1,358 additional units to represent new construction built since the 1973 survey. This updating of new construction has been continued on an annual basis. In addition, in 1974 the sample in rural areas was doubled (an increase of 15,500 units) to provide for greater precision in measuring certain characteristics of housing in rural areas. Each year, demolished units and other types of nonexistent units have been deleted from the sample, thus partially offsetting the increase from new construction. At present the sample consists of 81,850 units.

II. PURPOSE AND SCOPE OF THIS REPORT

This report is principally concerned with nonsampling errors related to undercoverage in the Annual Housing Survey (AHS-National Sample). In comparing the first year results to independently derived estimates it became apparent that certain types of units, such as mobile homes, were not adequately represented in the sample. The types of omissions had been generally recognized, but their magnitude and impact on components of the inventory had not been fully recognized. In particular, for mobile homes the undercoverage was compounded by census misses and definitional differences in the basic sampling frame.

The purpose of this report is to describe the types of undercoverage, the methodology for representing undercovered units in the sample, and their effect on the undercoverage bias. These topics are discussed in sections IV - VI below; summary and conclusions appear in section VII.

III. AHS SAMPLE DESIGN AND ESTIMATION PROCEDURES

A. Sample Design

The AHS is a multi-stage cluster sample of about 82,000 units spread over 461 PSU's, comprising 923 counties and independent cities. Of the 461 PSU's, 156 were included in sample with certainty; these are referred to as self-representing. The remaining PSU's were grouped into strata and a sample of PSU's was selected from each stratum. This resulted in an additional 305 PSU's, which are referred to as non-self-representing.

Within each sample PSU, a sample of units from the 1970 Decennial Census listings was selected. This was accomplished in several stages. First, a sample of census enumeration districts (ED's) was selected. The next stage consisted of the formation and selection of clusters of housing units (HU's) within each sample ED, where the selection method was dependent on the type of ED. There are two types of sample ED's - Address and Area. Address ED's are those in which building permits are authorized for new construction and at least 90 percent of the 1970 census addresses were listed with house number and street address. In these ED's a compact cluster of an expected four units was selected from the 1970 census address listings.

Area ED's are those that do not meet one or both of the Address ED criteria. These ED's were divided into small land areas referred to as area segments. Each area segment selected for AHS was canvassed and all units (both 1970 census units and units built after the census) were listed. A systematic sample was then selected from this listing for AHS; this resulted in a noncompact cluster of an expected four units in each area segment.

In addition, a sample of new construction building permits was selected within each sample PSU to represent units built after the census. These are called permit segments. Finally, as a result of a 1970 census evaluation study, a sample of units missed in the census was also included; these are referred to as CEN-SUP segments.

B. Estimation Procedure

The estimation procedure, utilized for AHS in 1973-1975, employed three stages of ratio estimation. The first stage was employed for sample units from NSR PSU's only and was designed to reduce the between-PSU component of variance, due to the sampling of PSU's.

The second stage ratio estimation, which is very relevant to the undercoverage problem, was only employed for units built in April 1970, or later (new construction units). This procedure was designed to adjust the AHS sample estimates of new construction to independently derived new construction estimates that were considered to be the best estimates available. These estimates were derived from the Survey of Construction (SOC), a survey of building permits conducted monthly by the Census Bureau (for conventional new construction), and from mobile home shipments reported by the Mobile Home Manufacturers Association (for new mobile homes). This adjustment was necessary to correct for the undercoverage biases in AHS with respect to new construction.

The third stage ratio estimation was employed for all sample units. It was designed to adjust the AHS sample estimates to independently derived estimates for four types of vacant units and 24 residence-tenure-race of head-sex of head categories for occupied HU's. These estimates were derived from the Housing Vacancy Survey (HVS), a quarterly vacancy survey conducted by the Bureau, and the Current Population Survey (CPS), a monthly unemployment survey also conducted by the Bureau.

IV. SOURCE AND TYPE OF UNDERCOVERAGE

As noted in the previous section, there are two types of Enumeration Districts (ED's); i.e., those for which permits are issued for new construction (permit-issuing areas) and those for which permits are not required for new construction (nonpermit areas). This paper is concerned with undercoverage in address segments, which are located in permit-issuing ED's, and in permit segments, which are used to represent new construction in these ED's.

The frames used for selecting the sample in address segment areas have certain deficiencies which, in total, represent something less than 2 percent of the universe (about 1,080,000 units, of which about 959,000 are eligible to be counted in the housing inventory). However, the undercoverage is disproportionately concentrated in certain types of units. These units are described below, along with estimates of their undercoverage.

One source of undercoverage bias is in units constructed since the census. For AHS, new construction is defined as units created on the site, including prefabricated housing, and occupied new mobile home placements. Prefabricated housing is represented in address segment ED's through permit segments. However, units completed after the census for which permits were issued before January 1, 1970, are not included in the sampling frame. These are referred to as permit lag units and are estimated at about 598,000 units.

The other type of new construction consists of occupied mobile home placements for which the undercoverage bias is estimated at 294,000 units.[1/] These units may be located in mobile home parks or on individual lots at large. Some of these parks have been created since the census; others existed prior to 1970 but were either missed in the census or unreported due to definitional differences.[2/] There is also undercoverage of mobile homes that were manufactured prior to 1970.

Another source of coverage loss is nonresidential units that have been converted to residential use since the census. The permit universe consists of permits for residential new construction only; it does not include permits for alterations to existing structures. Although these conversions are a small component of the housing inventory, they have unique characteristics that may not be fully represented in the independent estimates used in the third stage sample adjustment and therefore, contribute to biases in the characteristics of the total inventory.

Houses that have been moved into address segments since the census are also undercovered. They have no chance of selection at the census address nor at the new address, unless they replace housing that existed at the new address at the time of the census. The estimate of this undercoverage is 50,000 units.

Procedures have been developed to represent all of these types of units in the AHS National sample. These coverage improvement procedures are described in section V.

V. COVERAGE IMPROVEMENT SAMPLE DESIGN AND IMPLEMENTATION

Four coverage improvement procedures were developed to reduce undercoverage bias of the types of units described in section IV. The design and implementation of the samples are discussed in this section; survey results appear in section VI.

A. Permit Lag

The permit lag sample provides coverage of new construction for which permits were issued prior to January 1, 1970 but construction was completed after the census.

1. Sample Design

The permit lag sampling frame was created from the Survey of Construction (SOC), a survey of authorized building permits conducted monthly by the Bureau of the Census to determine the rate at which these authorized units are started and completed. Between 1964 and 1973, SOC was conducted in a 122-PSU design, which was a subset of the CPS 449-PSU design. Within each of these PSU's, a sample of permits authorized each month was selected from each of the sample permit-issuing places. A three stage sample selection was used which resulted in an overall probability of selection of 1-in-100 for each sample permit.

For each permit in SOC, the month construction started and the month it was completed were determined. From this a sampling frame was created which consisted of permits for residential structures that had been authorized before January 1970 but were completed after the 1970 census (i.e., April 1970). The AHS permit lag sample was selected from this frame. However, some of these sample permits were in PSU's which were not in the AHS sample design or in any other sample design. It was decided to drop these permits from the sampling frame since interviewing units in these PSU's would not be cost effective. To compensate for these units, the weights associated with the remaining sample units were adjusted by the following factor:

$$\frac{\text{Wt'ed. HU's in non-AHS PSU's} + \text{Wt'ed. HU's in AHS PSUs}}{\text{Weighted HU's in AHS PSU's}}$$

For cost efficiency reasons, it was decided that the ultimate sampling unit for the permit lag sample should be a compact cluster of about four units. Thus, the units for each permit in the frame were divided into clusters of about four units.

Each permit in the frame had a measure of size which was the weighted number of HU's represented by the particular permit. Prior to selecting a sample of the clusters, this permit measure of size was transformed into a cluster measure of size according to the following formula:

$$\text{Measure of size of cluster } j = \frac{M_i}{(K_i)N_{ij}}$$ 3/

where: M_i is the measure of size for the i^{th} permit.

K_i is the number of clusters associated with the i^{th} permit.

N_{ij} is the size of the j^{th} cluster in the i^{th} permit.

Prior to sample selection, the clusters of this sample frame were stratified according to the following variables:

1. Size of structure
   a. 1 unit
   b. 2-3 units
   c. 4-5 units
   d. 6-7 units
   e. 8-9 units
   f. 10-16 units
   g. 17-49 units
   h. 50-99 units
   i. 100-199 units
   j. 200 or more units
2. Region
3. SMSA/Non-SMSA
4. PSU Number
5. Permit Number
6. Cluster Number

This stratification was employed to insure a representative sample of these types of units by size of structure, region, SMSA/Non-SMSA, etc.

Since all of the Bureau's recurring surveys (i.e., CPS, AHS, the National Crime Survey [NCS], and the Health Interview Survey [HIS] fail to properly represent these permits, representative national samples of clusters necessary for the rest of the decade were selected for each of these surveys. This included one sample for AHS, thirteen samples for CPS, six samples for NCS, eight samples for HIS, and two samples to be held in reserve for future surveys. The clusters were selected with probability proportionate to the cluster's measure of size at a rate of 1-in-47. The selected clusters or hits were assigned to each of these samples according to the following scheme:

Hit 1 :	AHS
Hits 2-14 :	CPS (samples A36-A48)
Hits 15-16:	Reserve samples
Hits 17-19:	NCS (samples J03, 05, 07)
Hits 20-27:	HIS (samples Y77-Y84)
Hit 28 :	AHS
Hits 29-41:	CPS (samples A36-A48)
Hits 42-44:	NCS (samples J04, 06, 08)
Hits 45-46:	Reserve samples
Hits 47-54:	HIS (Y77-Y84)
Hit 55 :	AHS
Hits 56-58:	NCS (samples J03, 05, 07)
Hits 59-71:	CPS (samples C20-C32)
Hits 72-73:	Reserve samples
Hits 74-76:	NCS (J04, 06, 08)
Hits 77-84:	HIS (Y77-Y84)

The assignment order presented in the above scheme was repeated for every 84 hits, which means that 3 out of every 84 selected clusters were assigned to AHS.

2. Systems and Procedures

As indicated above, the permit lag universe was developed from a computer listing of 12,920 permits issued during the years 1967-1969 in the sample PSU's. The permit issuing date and the date construction was completed appeared on the list for each unit. Thus the universe was created by stripping off addresses of all structures that were completed after April 1, 1970. A sample of 1,386 units was selected for the AHS national sample.

The selected units were clustered by geographic location into 438 segments of size 1-5. A total of 1,386 units were assigned for interview during the regular AHS interview period (roughly September - November 1976).

Some overlap between the permit universe and census addresses was discovered at time of AHS interview. This occurred, in part, because the reported date of completion for multi-unit structures was the date when more than half of the units were completed. Thus some of the units were completed earlier and could have been reported in the census. In these situations the basic address, and all units at it were eligible for inclusion in the AHS sample. In the case of single-unit structures the census enumerator could have considered construction sufficiently complete to report such units as vacant. (Some subjectivity entered into the determination of vacancy status.)

Overlap could also occur between the permit lag universe and the regular permit universe or the CEN-SUP sample, for methodological reasons or due to permit issuing practices. For example, all units at a sampled permit address are listed, regardless of the number of structures involved. However, separate permits may have been issued for each structure and, depending upon the timing, subsequent permits might not be discovered.

In the case of overlap with CEN-SUP, that sample was developed after the census and may have included some permit lag units. Since CEN-SUP is a sample, not a universe, and the PSU's in the permit lag universe are a subset of the PSU's for which CEN-SUP was developed, a complete unduplication cannot be accomplished.

The overlap among the various universes is expected to be small. However, it is presently under investigation. In addition, some procedural controls are imposed to correct the overlap. For example, the interviewer is told the number of units for which the permit is issued. If more units are found than expected, a check is made to determine if this is the result of overlapping frames, or due to permit problems such as overbuilding or underreporting on the permit. Adjustments in the sample estimates are made as a result of duplication discovered through procedural controls.

B. Woodall Sample

This sample was selected from a universe of mobile home parks obtained from a commercial list. The list was updated each year through 1974, when the commercial operation was terminated. Thus the Woodall sample provides coverage of mobile homes located in parks created after the census and through calendar year 1974. Parks that were begun before 1970, but completed after the census, also were included. (Mobile home parks and other special places are not included in the Permit Lag sample since they are not sampled from permits.)

1. Sample Design

This sample was designed to provide coverage of mobile homes located in parks which were created after the 1970 census. Since the sample was limited to address ED's, it was necessary to unduplicate these places from area segment ED's. In addition a check needed to be made against the Census listings for places reported as created through 1972 in case any part of these places existed at the time of the Census. To do this it was necessary to determine (as described in paragraph 2) the ED in which each park was located. The unduplication and matching procedures were costly and time consuming. In order to reduce costs and preparatory time, it was decided to implement this procedure in the 266-PSU design (the representative national sample of PSU's which is a subset of the AHS design). The savings in cost and time were considered sufficiently important to outweigh any increase in the between PSU variance component resulting from this design. Therefore, the Woodall sampling frame consisted of the mobile home parks on the Woodall commercial listing which (1) were identified as having been created after the 1970 census and (2) were located in an address ED in the 266-PSU design.

Since it was decided to employ noncompact clusters of size four for this procedure, similar to what is done for other mobile home parks in AHS, the measure of size associated with each park in the Woodall sample frame was equal to the following:

$$\frac{\text{Number of sites in park}}{4}$$

Prior to sample selection, the mobile home parks were stratified according to the following variables:

1. Region 2. SMSA/Non-SMSA 3. SR/NSR

This stratified sampling frame of mobile home parks was them sampled with probability proportionate to the park's measure of size such that the overall probability of selection for each hit, or sample cluster, was 1-in-1366. This resulted in 30 sample mobile home parks from which 31 noncompact clusters of 4 mobile home sites were selected for the AHS Woodall samples. The procedure for selecting the sample units appears below.

2. Systems and Procedures

The full universe consists of 794 parks that were not available for listing at the time of the census. The universe was created by determining the geographic location of each park on the commercial list and allocating the parks to the appropriate census ED. Then the ED's were identified as area or address segment ED's, according to their permit issuing status and certain other criteria related to the adequacy of addresses in the ED.4/ Parks in area segment ED's were dropped from the universe because they had a chance of selection in the AHS sample through area segments.

Since a mobile home park would have a chance of selection in the AHS sample if even one unit was occupied at the time of the census, an unduplication procedure was mounted. The address of each park located in an address segment ED and completed before January 1, 1973 was matched against the census listings. Any parks listed in the census, as either a regular address or a special place address, were dropped from the Woodall universe. The January 1973 cutoff was used because it was felt that a park, for which construction had begun before the census, would be completed by that date. Because vacant mobile home sites were not reported in the census, a review was made of parks that first appeared on the commercial list in 1969. These were processed as described above, and those not found in the census were included in the Woodall universe.

In order to avoid clustering, interviewers listed all sites (occupied or vacant) at the selected parks and a non-clustered sample of approximately 4 sites was selected from the listings. A total of 119 sites were assigned for AHS interview.

C. Windshield Sample

The Windshield sample was used to supplement the Woodall sample. It was originally conceived as a source for providing coverage of mobile home parks created after the termination of the Woodall operation; i.e., after January 1, 1975. However, some preliminary investigation indicated that the Windshield sample had the potential for improving undercoverage bias of parks missed or otherwise unreported in the census and in the Woodall sample. Thus, the scope of the Windshield sample was broadened to provide for this additional coverage.

1. Sample Design

The Windshield sample design was a two stage sample selection procedure implemented in the entire AHS 461 PSU design. The first stage consisted of selecting about 150 tracts within these PSU's. It was decided to select tracts5/ since they were small enough to be canvassed at relatively little cost and time, but were large enough to yield a significant payoff in terms of locating missing mobile home parks. One problem with using tracts as the area to be canvassed is that the sample was supposed to represent missing mobile home parks in address ED's only; but the sample tracts could contain some area ED's. This problem was resolved by eliminating all parks found to be in area ED's. The identification was made after the tracts had been canvassed, because it was more efficient than unduplicating the area ED's before canvassing the tracts. One-hundred and fifty tracts were selected because it was felt that this was the maximum number of tracts that could be canvassed, taking into consideration the time and cost constraints. Although this was not necessarily the optimum number of tracts, it was felt that canvassing this number of tracts would result in a relatively reliable estimate of mobile homes in missing mobile home parks.

The 150 tracts were selected from a file, created from the 1970 census fourth count tape, that contained a record for each tract in the 461 PSU design. A measure of size (M_i'), equal to the total number of mobile homes in the tract as reported in the 1970 census was assigned to each tract.[6] Even though the number of 1970 census mobile homes may not necessarily have been highly correlated with mobile homes in missing mobile home parks, it was felt that this measure of size was the best available for selecting the sample tracts. The measure of size was then adjusted by the inverse of the probability of selecting the PSU in which the tract was located, to reflect the sampling of NSR PSU's. The adjusted measure of size (M_i) was then used in the selection of sample tracts. This tract file was stratified, or sorted, by the following variables:

1. Region 2. SR/NSR PSU 3. M_i

The sample of tracts was then selected with probability proportionate to M_i using the following sampling fraction: $\frac{150}{M}$ (where M equals the sum of M_i's across all of the tracts in the 461 PSU's.)

The 150 selected tracts were then canvassed, as described in the next section. Mobile home parks identified in the canvassing operation that were found to be in area ED's, enumerated in the 1970 census, or duplicated on the Woodall list, were deleted from the Windshield sample.

The second stage procedure was the selection of noncompact clusters of size four (mobile home sites) within the remaining mobile home parks. Prior to this sample selection, the parks were sorted into two types - census misses (parks in existence in April 1970 which were not enumerated in the census) and Woodall misses (parks built after April 1970 which were not on the Woodall list). The second stage selection was implemented independently within each type of park. The noncompact clusters of size four were then sampled with equal probability within each park using the following sampling fraction for each sample tract:

$$\frac{1}{1366} \times \frac{M}{150\ M_i}$$

This within-tract sampling fraction was employed so that each noncompact cluster of four would have the same overall probability of selection, 1-in-1366, as the other AHS sample units (i.e., this sampling fraction was used to preserve, as much as possible, the self-weighting aspects of the AHS sample design).

2. Systems and Procedures

Census interviewers canvassed each of the 150 tracts selected in the Windshield sample. In order to reduce costs, all major roads were physically canvassed, as were any areas where signs indicated the location of a mobile home park, but inquiry was made in areas where parks were not likely to be located; e.g., in high cost housing projects.

A form was filled for each park discovered. This provided identification information and the size of the park. Through a matching operation the parks were unduplicated from the Woodall universe and from the census. This resulted in 85 parks, which were subsampled at a rate computed separately for each tract. A sample of 24 parks was selected, from which 29 segments were created. (Double hits occurred in some large parks.) In order to avoid clustering, a sample of units was selected across each park. A total of 118 units were assigned for AHS interview.

D. Successor Check

The successor check provides coverage of three types of units that would not have been reported in the census at their present location.

The first type is mobile homes at large (not located in parks) that were either placed on the present site since the census or were vacant at the time of the census. (Vacant mobile homes were not reported in the census even when they were affixed to a permanent foundation.) The second type is houses that were moved to the present site since the census. Finally, the successor check provides coverage of units in structures that were converted to residential use since the census. These three types of units are referred to as inscope successors.

1. Sample Design

Unlike the Permit Lag or Woodall coverage improvements, a universe (or sample-based) listing of these types of units, from which a representative sample could be selected, did not exist. Thus, it was decided to use a successor check procedure. This is a listing procedure that has been used previously by the Census Bureau (e.g., it was used in the spring of 1976 for the Survey of Income and Education and was used for CPS and HIS throughout the 1960 decade). The successor check procedure is described in more detail in the next section.

Briefly, it involves listing a string of k structures in a predetermined order. The string begins with an AHS sample unit and is bounded by the kth residential structure that existed in 1970. Inscope units are identified along the string, between these two structures.

Since the check was related to the AHS sample, the only sample design questions that needed to be resolved for this coverage improvement procedure were the size of the string (k) and how many strings should be listed (i.e., the number of AHS sample addresses from which a listing should be started).

The 1970 Components of Inventory Change Survey (CINCH) showed that between April 1960 and October 1970 there were about 743,000 of these types of units added to the inventory. This represented about 1 percent of the inventory in a 10 3/4 year time period; therefore, it was assumed that these units added since April 1970 represented about .6 percent of the total inventory. Since these units represented such a small fraction of the total inventory, it was assumed to be unlikely that more than one inscope structure would be found in a string. Therefore, the intraclass correlation between inscope or missed structures would not depend on the string length which implied, in terms of variance constraints, that the string size should be as large as reasonable. This was also true to a certain extent, in terms of cost considerations. The cost per inscope unit decreases as the size of the string increases since the expected number of inscope structures listed per string also increases. However, the Bureau's field personnel felt that after a certain string size there would be a large incremental cost increase due to added complexity, travel, more supervisory referrals, etc. Although they did not know the exact size at which this increased cost would be incurred, it was speculated that this would happen for a string size of 12 or more. Even though it was not optimal, a string size of 8 was selected as a compromise, to minimize the risk of incurring this additional cost increase since the coverage improvement budget was very tight and would not allow for this additional cost.

Given the string size of eight, the number of strings was determined by equating this to an optimal allocation determination for a stratified sample involving two strata. The first stratum was the universe represented by the successor check units and the other stratum was the universe represented by the rest of the AHS sample units. This optimal allocation formula is given as follows:

$$\frac{n_{SC}}{n} = \frac{N_{SC}\, S_{SC} / \sqrt{C_{SC}}}{N_{AHS}\, S_{AHS} / \sqrt{C_{AHS}} + N_{SC}\, S_{SC} / \sqrt{C_{SC}}}$$

where:

N_{SC} = the number of units in the successor check universe.

N_{AHS} = the number of units in the regular AHS universe.

C_{SC} = cost per unit for the successor check universe (for a string of 8, this cost equalled \$306.25).

C_{AHS} = cost per unit for the regular AHS universe (C_{AHS} = \$24.50).

S^2_{SC} = the unit variance for the successor check universe.

S^2_{AHS} = the unit variance for the regular AHS universe.

We know that:

$$\left.\begin{aligned} N_{AHS} &= (1-P_{SC})\,N \\ N_{SC} &= P_{SC}N \end{aligned}\right\} \text{ where } P_{SC} = \text{proportion of the total universe represented by the successor check universe.}$$

$$n = n_{SC} + n_{AHS}$$
$$= n_{SC} + 1.57\ n_{AHS(oc)}$$

(where $n_{AHS(oc)}$ is the AHS sample size for units from address segments.)

Inverting the allocation formula and making the above substitution produced the following result:

$$\frac{n_{AHS(oc)}}{n_{SC}} = \frac{(1-P_{SC})}{(1.57)P_{SC}}\ \frac{S_{AHS}}{S_{SC}}\ \sqrt{C_{SC}/C_{AHS}}$$

Since the number of units in each address segment is two and the string size is eight, then:

$n_{AHS(oc)}$ = 2 (Number of address segments)

n_{SC} = $\frac{8\ Psc}{1-Psc}$ (Number of successor check strings)

Substituting the above into the allocation formula produces the following:

$$\frac{\text{Number of address segments}}{\text{Number of successor check strings}} = \frac{8}{2(1.57)}\ \frac{S_{AHS}\sqrt{C_{SC}/C_{AHS}}}{S_{SC}}$$

Since $\frac{S_{AHS}}{S_{SC}}$ did not vary greatly, the optimal ratio of address segments to the number of successor check strings for a string size of eight was determined by the square root of the ratio of the costs. This optimum subsampling rate was about twelve. In other words, one-twelfth of the AHS address segments (about 1,500) would be used as the starting points for the successor check strings. Since the AHS sample had been divided into six panels, each of which was a representative national sample, a systematic half sample of the address segments in one panel was selected for the successor check. The first address in each of these segments was used as the address from which the string of eight was determined.

The details of sample selection for the successor check appear below.

2. Systems and Procedures

The successor check was conducted at time of interview for 1,500 selected AHS units. For each of these units the interviewer listed a string of 8 structures[7] in a path of travel bearing to the right from the sample unit. The structures along the route were listed and a sketch drawn to show their location.

The year of construction was determined in order to identify regular structures built before April 1, 1970. These were called successor structures and were used to bound the string. By design, each string was to consist of eight successors and any intervening structures.

The string could cover one or more blocks in urbanized areas or a distance up to 10 miles in rural areas. In general, the path of travel was expected to proceed around the block in which the sample unit was located. In order to preserve probabilities, the string was terminated in the sample unit block when the northwest-most corner was reached. From this point the interviewer would continue an incomplete string, starting at the northwest corner of the next block to the right. This procedure would be continued until the string was completed.

No procedure was required for matching against the census address registers because the operation was not designed to identify census misses. (In address ED's units missed or otherwise not reported in the census are represented through the CEN-SUP sample.)

Interviewers recorded the number of units in each inscope structure listed. The regional office clerk reviewed the listing sheets and performed various quality checks. Units in inscope structures were assigned for AHS interview. Large multi-unit inscope structures were subsampled.

Consideration had been given to conducting interviews at inscope structures at the time they were identified in the successor check. This had some advantages in relation to cost and time constraints. However, it was felt that this might introduce interviewer bias. If interviews had to be obtained at each inscope structure, interviewers might be less scrupulous about identifying such structures.

A total of 44 inscope successor units were assigned for interview.

VI. RESULTS OF COVERAGE IMPROVEMENT PROCEDURES

The undercoverage bias affected both the total new construction estimates and the estimates of characteristics of the total housing inventory. For each year until 1976, a ratio estimation procedure was employed to adjust the AHS sample estimates of new construction units to independently derived current estimates.[8/] This procedure was used to correct for known deficiencies in the three categories of new construction represented in the sample.[9/] Although the independent estimates were considered the best available, their accuracy had become a matter of growing concern. In addition, the ratio estimation procedure may have had no effect on the bias in housing characteristics due to the undercoverage of certain types of units. The coverage improvement procedures addressed both of these issues. If the procedures could correct frame deficiencies so that all housing units had a known non-zero probability of selection in the survey, this would eliminate the bias and, in addition, valid unbiased estimates of total could be derived from the survey data itself. Another possible option relates to the third stage ratio estimation procedure. It is designed to adjust the AHS total inventory estimates to current independent housing estimates. These latter are derived from the CPS and the HVS. These two surveys have the same frame deficiencies as the AHS. Better estimates of the total housing inventory might be obtained by correcting the frame deficiencies in the CPS and HVS and then retaining the third-stage ratio estimation procedure.

In order to evaluate these options it is first necessary to examine the results of the coverage improvement procedures in terms of their effect on the undercoverage bias in the AHS sample.

The four coverage improvement procedures yielded a total of 1,667 unweighted units, of which 1,538 would be weighted to represent omissions in the housing inventory. The distribution by source and an analysis of their contribution to the sample appears below.

A. Permit Lag Sample Results

There were 1,386 units selected for the permit lag sample, representing 598,000 units which had no other chance for selection in AHS. This amounts to 0.88 percent of the total 1970 housing inventory and is all new construction.

The basic weight assigned to the permit lag sample units, during the AHS weighting procedure, was equal to the inverse of the probability of selecting a sample unit for the AHS permit lag survey.

The weight assigned to each of the sample units in the i^{th} cluster of the j^{th} permit was euqal to $\frac{1316}{N_{ij}}$.

Originally, the AHS permit lag sample design was to produce a self-weighting sample with each sample unit having a weight of 1316. However, during the AHS permit lag sampling operation the clusters were assigned the measure of size $\frac{N_j}{K}$ rather than $\frac{N_j}{K_j N_{ij}}$, which produced the actual non-self-weighting sample.

The permit lag coverage improvement procedure was probably the most successful in terms of eliminating the undercoverage bias associated with AHS. Since the permit lag sampling frame was based on a representative national 1-in-100 sample of all permits authorized before 1970, it should also be a representative sample and produce approximately unbiased estimates of any subset of these permits. Thus, one would expect that the permit lag sampling frame was a representative sample of units for which construction was authorized before 1970 but was completed after April 1970, and that a sample selected from this frame would produce unbiased estimates of characteristics of such units. The problem was that two possible sources of bias were introduced into the sampling operation. One resulted from eliminating units in non-AHS PSU's from the permit lag sampling frame and the other was the result of noninterviews in selected clusters. However, any bias in the sample estimates from these sources are likely to be quite small. In the first instance the number of units represented is about 16,000 and the weights on the remaining units in the permit lag sample frame were increased to represent these units. The second source of bias resulted from the fact that 21 of the 479 clusters selected for the AHS permit lag sample could not be visited because the corresponding SOC questionnaire, which contained the address, could not be located. Once again, the weights for the sample units that were visited were increases to reflect these 21 clusters. It is fairly safe to assume that these were approximately random misses and thus most of the bias associated with this problem was eliminated by the adjustment.

Although estimates from the permit lag sample are subject to sampling error, the magnitude of the sampling variability is probably lower than it would have been if these units were represented in the original AHS sampling operations. The decrease in variance was due to the larger-than-planned size of the AHS permit lag sample.

This gain was offset slightly by the fact that the permit lag sample frame was based on the 122 PSU design, which is therefore subject to more between-

PSU variance than the AHS 461 PSU design. Also, since the overlap between the permit lag universe and the address segment universe for multi-unit structures[10] was resolved at the sample unit level rather than the universe level, there may have been an increase in the variances associated with the sample estimates from the permit lag universe.

B. Woodall and Windshield Sample Results

A total of 237 mobile home sites located in parks were selected for sample from these two sources. This represented 342,000 mobile home units that had no other chance for selection in AHS. Approximately 50 percent of these mobile home units were in parks that existed in 1970 but were not reported in the census; the remainder were in parks created since the census.

The basic weights assigned to each Woodall or Windshield sample unit during the AHS weighting procedure was equal to 1,366. Thus both the Woodall and the Windshield samples were self-weighting sample designs.

The combination of the Windshield and Woodall coverage improvement procedures was successful in terms of eliminating the mobile home undercoverage bias in AHS. This was due, in part, to the supplemental effect of the Windshield Sample. The Woodall sampling frame consisted of what was purported to be a complete listing of new mobile home parks that were created between April 1970 and December 1974. Thus, the sample selected from this listing should be a representative sample and produce approximately unbiased estmates of that universe. However, there was evidence of undercoverage in the Woodall frame which was improved by the Windshield procedure. This latter procedure was able to represent not only mobile home parks built after 1974 and mobile home parks missed by the census but also mobile home parks that should have been on the Woodall list but were not. Twenty-three of these parks were picked up in the Windshield sample. Thus, the Windshield procedure attempted to eliminate the undercoverage bias in the Woodall procedure due to the deficiencies in the Woodall sampling frame.

As a result, the major source of bias associated with the Woodall sample estimates, i.e., an incomplete universe, may have been eliminated, depending on the bias associated with the Windshield procedures. Although the Woodall sample was selected at the same rate as regular AHS, the Woodall estimates are probably subject to more sampling error than if they had been sampled with regular AHS since the Woodall sample was confined to the 266-PSU design and therefore is subject to more between-PSU variance.

One source of nonsampling error associated with the Windshield estimates is the completeness of the canvassing and of the matching operations. Additionally, since tracts were used as the areas to be canvassed for Windshield, this procedure only represents missing mobile homes in address ED's in tracts. The magnitude of this bias, obviously, depends on the proportion of missing mobile homes in non-tracted address ED's, which are approximately 9 percent of all address ED's. This bias would also impact on the effectiveness of the Windshield in terms of eliminating the undercoverage bias in the Woodall sampling frame. Even though the Windshield sample units were selected at the same rate and in the same PSU sample design as regular AHS, the resultant estimates are probably subject to more sampling error than if these units had been sampled with regular AHS. One source of this additional variance is the fact that the existence of parks in area ED's within tracts could not be corrected for until after canvassing, rather than before selecting the sample of tracts. Thus, the measure of size used in the selection of tracts included the effect of mobile homes in area ED's. The other major source of additional variance is the degree of effectiveness of the measure of size, assigned to the tracts during the selection of tracts, with respect to estimating missing mobile homes.

Missing parks were found in tracts with measures of size ranging from as low as 48 to as high as 2,435, whereas some tracts with measures of size as high did not contain missing parks. Thus, comparing the measures of size for tracts with and without missing parks does not uncover any obvious patterns. Nonetheless, the estimated correlation coefficient, based on these 150 tracts between the measure of size for a tract and the number of sites in missed parks found in the tract, is .67. Thus, based on the magnitude of this estimated correlation coefficient, it appears that the measure used in the selection of tracts was fairly effective in terms of the characteristic of interest.

C. Successor Check

The three types of inscope units discovered through the successor check produced a total of 44 sample units distributed as follows:

28 units - mobile homes at large, of which 24 represented omissions in the housing inventory

11 units - houses moved into the sample area

5 units - converted from nonresidential use

These represented roughly 124,000 units, which have unusual characteristics that were not adequately reflected in the original sample. (The total weighted figure would be 140,000 but 16,000 would not be considered part of the housing inventory.) The basic weight assigned to each successor check sample unit, during the AHS weighting procedure, was equal to the inverse of the probability of selecting the unit. The probability of selecting a sample unit was equal to the probability of selecting a successor check structure. As was mentioned before, the successor check sample design involved the listing of a string of addresses starting from the first address (referred to as "the sample address") in half of the address segments in one panel of AHS (panel 3). The string included exactly eight census addresses which had a prior chance of being selected for AHS (referred to as "the successor addresses") and any intervening new construction, mobile home parks, other types of special places, and inscope structures.

All units in an inscope structure (referred to as "the successor check sample units") were interviewed for AHS unless there was an excess number of successor check sample units in an inscope structure or the string. In that case, a subsample of the successor check sample units was selected for interview.

From this design, every successor address had a chance to be a sample address and vice versa. As a result of listing eight successor addresses in each string, any inscope structure could have been brought into the sample because of one of eight possible AHS sample addresses. If the eight preceeding sample addresses (or equivalently, successor addresses) for an inscope structure are denoted by $a_1, a_2, \ldots, a_8$ and the probability that sample address a_k was selected for the successor check is denoted by $P_r\ [a_k]$, then the probability that an inscope structure came into sample is $\sum_{k=1}^{8} P_r\ [a_k]$. However, information was obtained such that $P_r\ [a_k]$ could be calculated for only those successor addresses that preceeded the inscope address in the string. Therefore, $\sum_{k=1}^{8} P_r\ [a_k]$ could not be calculated from the information available. Nonetheless, the conditional probability of inclusion of the inscope address, given that the sample address is a_k, did provide an unbiased weight.

This conditional probability of inclusion is $8\ P_r\ [a_k]$ and was estimated as follows:

Let q = the number of address segments with some or all of their units in the sample address a_k.

m_i = the number of units in the i^{th} address segment at the sample address.

n_i = the total number of units in the i^{th} address segment.

Therefore:

$$8\ P_r\ [a_k] = 8 \sum_{i=1}^{q} P_r \begin{bmatrix}\text{Panel 3 was selected}\\ \text{for successor check}\end{bmatrix}$$

$$\times\ P_r \begin{bmatrix}\text{The half-panel}\\ \text{was selected for}\\ \text{successor check}\end{bmatrix}$$

$$\times\ P_r\ [i^{th}\ \text{segment was selected for AHS}]$$

$$\times\ P_r\ [\text{sample unit falls in } a_k]\}$$

$$= 8 \left\{ \sum_{i=1}^{q} \frac{1}{6} \times \frac{1}{2} \times \frac{2}{1366} \times \frac{m_i}{n_j} \right\}$$

$$= \frac{4}{3(1366)} \sum_{i=1}^{q} \frac{m_i}{n_i}$$

Thus, the basic weight for the units in an AHS successor check inscope structure was equal to:

$$\frac{3}{4} \left(\frac{1}{\sum_{i=1}^{q} \frac{m_i}{n_i}} \right) 1366$$

Thus, the successor check basic weight was

$\frac{3}{4} \left(\frac{1}{\sum_{i=1}^{q} \frac{m_i}{n_i}} \right)$ times as large as the regular base weight for AHS sample units from address segments. The most common successor check basic weight was 3.0 times the regular AHS base weight.

There is evidence that the successor check coverage improvement was not very effective in terms of eliminating the undercoverage bias in AHS for the types of units involved. As was mentioned before, the 1970 CINCH Survey estimated that additions from other sources, which are comparable to the units represented by successor check, between April 1960 and December 1970 amounted to 1 percent of the 1960 census inventory. Extrapolating this rate to the time period April 1970 - October 1976 indicates that additions from other sources in this time period should be about .6 percent of the 1970 census inventory or 400,000 units. (This is probably an underestimate of the actual rate for a 6½ year period because the CINCH estimate which covers a 10 3/4 year, would not represent units added by other sources after 1960 that were removed from the inventory by December 1970.) Since the successor check was designed to represent these types of units in address segments only, this number should be adjusted by the percent of the old construction sample represented by address segments (75 percent). This produces a figure of about 300,000 units which, even though it is probably an underestimate of the actual figure, is substantially larger than the estimate from the AHS successor check sample. Most of this difference was probably due to the successor check's coverage of the units converted from nonresidential to residential use. For these types of units, the successor check sample produced an estimate of about 16,000 units, which seems extremely low for a 6½ year period.

In evaluating the results of the successor check two important matters of resource must be considered. First, better estimates would have been obtained by selecting a sample from a representative list of these units, such as was done for the permit lag and the Woodall sample. The problem was that no such list existed or could be compiled with available resources. A second approach would have been to use the CINCH (area block listing) method. However, this would have been a costly operation and would have taken more time to finalize than was available for AHS. As a result it seemed best to develop a procedure that could be integrated into the basic AHS sample. The successor check was a reasonable choice for identifying mobile homes at large and even homes moved in. However, some type of stratification, or a very large sample might be required to provide an adequate sample and control excessive variability for nonresidential conversions. For example, to alleviate the deficiency, the successor check sample could have included a disproportionate number of AHS sample units in nonresidential areas. Alternately, a block sample approach could have been used for nonresidential conversions, in which a sample of blocks within a sample of tracts, which were highly nonresidential in 1970, could have been canvassed to identify such units. However, either of these methods would have added to the costs and funding was a serious problem. In any case, further consideration will have to be given to this matter before the successor check can be introduced into other surveys.

VII. SUMMARY AND CONCLUSIONS

A number of known deficiencies existed in the sampling frames for the AHS National sample. These resulted in under-representation in the survey of mobile homes, new construction, housing converted from nonresidential structures and houses that were moved from their original sites. The total estimate of this undercoverage was about 959,000 units. Although the survey data were adjusted to compensate for these omissions, biases may still have existed in the characteristics of the housing inventory. In addition, the independent estimates employed in the estimation procedures were not entirely satisfactory, especially for new construction mobile home estimates. Therefore, supplementary coverage procedures were introduced into the 1976 survey to provide more adequate coverage in the survey itself. The results of these procedures in terms of the undercoverage bias were the subject of this paper.

Since the coverage improvement procedures virtually eliminated the undercoverage bias for new construction units, it was decided to eliminate the second-stage ratio estimation procedure for most categories in the 1976 AHS estimation process. However, undercoverage bias in the AHS sample still exists for units converted from nonresidential use since 1970 in address ED's and for units in area ED's, so some concern still existed about the estimation of the total inventory. Therefore, it was decided to continue using the third-stage ratio estimation procedure for the 1976 AHS, even though it was felt that the independent estimates were overstated. This is a conservative approach and is subject to change in later years, as more experience is gained from the use of the coverage improvement procedures.

[1] Some additional sample units (representing 121,000 units) that were picked up by the coverage improvement procedures are not included in this figure, although they could become part of the housing inventory. These include vacant mobile homes and unoccupied sites in mobile home parks that may be occupied. in the next AHS interview period and therefore included in the housing inventory at that time.

[2] For example, vacant mobile homes or sites in parks were not recorded in the census but are included in AHS since the units might be occupied when the annual survey is conducted.

[3] Inadvertently, during the actual sampling operation, the N_{ij}'s were not divided out. This meant that the probability of selection for sample units was N_{ij} times as large as had been intended to produce a self-weighting sample. This produced a larger-than-expected sample for this coverage improvement procedure which necessitated a special adjustment in the estimation procedure for these units.

[4] Details of this process can be found in the 1970 redesign documentation, which is mostly internal Census Bureau memoranda. They will also appear in the Bureau's Technical Paper No. 7 which is currently being revised.

[5] This refers to Census tracts, which are geographic areas containing two or more ED's.

[6] Tracts in which no mobile homes were reported in the 1970 census were assigned a measure of size of 1 to insure that these tracts had a chance of selection.

[7] The term structure, as used here, includes mobile homes at large on permanent foundations if occupied by persons with no usual residence elsewhere, as well as regular residential structures.

[8] For more detail on the estimation procedure see Current Housing Reports Annual Housing Survey: 1973 United States and Regions, Part A, Series H-150-73A, Appendix B, "Sources and Reliability of the Estimates," pp. app. 32-3.

[9] This included categories for conventional new construction units and for new mobile home placements.

[10] The situation is described in section V.A.2. of this paper.

NOTE: This paper is an abstraction which omits details of the sample design and estimation procedures, related research, and references that appear in the original paper.