CREATING A FRAME OF NEWLY CONSTRUCTED UNITS FOR HOUSEHOLD SURVEYS

Bridgett Bell, Leyla Mohadjer, Jill Montaquila, and Lou Rizzo, Westat
Bridgett Bell, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: New construction, area sampling, building permit, sampling frame, and coverage

1. Introduction

Since the 1940's, area sampling has been one of the primary methods used for household surveys in the United States. Area sampling is usually carried out in a number of stages. The first stage is the selection of primary sampling units (PSUs) which are usually counties or groups of counties. The second stage of selection is a subset of areas within the PSUs. Additional stages of sampling may include households and persons.

In general, the first stage of area sampling (PSU selection) requires relatively accurate county-level data. These data are usually available by year. This is not the case for the second stage of selection. Data are not readily available on subsets of areas within PSUs. The subset areas, known as segments, are formed by grouping census blocks. In order to retain a self-weighting sample, the probability of selecting a segment is proportionate to a measure of size. Typically, the measure of size is derived from decennial census data such as total population or total number of housing units. This sampling method provided reasonable survey estimates until the 1950's and 1960's.

Table 1. Change in population between decennial censuses

	Population		Housing units	
Year	Total	Percent change from last census	Total	Percent change from last census
1960	179,323,175	18.5	58,326,357	26.4
1970	203,302,031	13.4	68,704,315	17.8
1980	r226,542,199	11.4	88,410,627	28.7
1990	248,709,873	9.8	102,263,678	15.7

Source: U.S. Bureau of the Census. The 1990 Census of Population and Housing. "Population and Housing Counts: 1790-1990", Publication CPH-2-1, U.S. Government Printing Office, Washington, D.C.

During 1950's and 1960's, the geography of the country began to change. Rapid growth occurred outside the central cities. The number of suburban communities increased significantly. For example, Table 1 shows an 18.5 percent change in population and a 26.4 percent change in the number of housing units built after the 1950 decennial census. Consequently, the U.S. Census Bureau began to see higher variability in survey estimates. The primary concern was how to handle rapid growth between decennial censuses. Therefore, the U.S. Census Bureau began working on methods to reduce the variance that many large surveys had experienced at this time.

In 1959, Joseph Waksberg, then a U.S. Census Bureau statistician, began work on a new area sample design (Waksberg, 1998). He proposed supplementing the most recent census data with statistics on new housing starts. Data on permit issuance are collected from all local building permit offices in the country on a monthly or annual basis. In the early 1960's, this new form of list sampling was introduced in the Current Population Survey (CPS). As of 1970, the number of building permit offices covered approximately 85 percent of all new residential housing built in the country. The CPS continues to use the building permit office survey data as a frame. Other area surveys have used the dual sampling frame approach to offset variance as well, particularly area surveys conducted late in a the decade.

Section 2 describes the methods currently used at Westat to create a sampling frame using the building permit files. Section 3 briefly describes a method for sampling segments and Section 4 describes listing procedures. Sections 5 and 6 discuss the use of permits as a sampling unit and other sampling issues, respectively. Section 7 provides a conclusion.

2. Frame Development

A frame of new construction is developed based on data collected from the Building Permit Survey conducted by the U.S. Census Bureau. The data are available on CD-ROM and there are four report source files: Annual Summary, Annual Cumulative, Monthly, and Monthly Cumulative. The Annual Summary file contains data from building permit offices that report on a yearly basis. Usually, annual reports come from offices that issue a small number of permits per year (approximately 50 or less). The Annual Cumulative files contain monthly data for all months over a number of years. Monthly files contain data from the current month and year and the Monthly Cumulative files contain cumulative monthly data for the current year. For instance, the monthly cumulative file for March 1999 contains the number of units for which permits are

issued for January 1999, February 1999, and March 1999 separately.

Surveys have applied different approaches in using the permit sampling frame as a supplement to the census data. For example, Westat is currently conducting two area surveys using the building permit survey data to develop a supplemental frame: The National Health and Nutrition Examination Survey (NHANES) and the National Survey of Parents and Youth (NSPY).

NHANES is sponsored by the National Center for Health Statistics (NCHS). The survey is designed to assess the health and nutritional status of the civilian noninstitutionalized population of the U.S. In addition, NHANES provides national estimates of selected diseases and disease risk factors. Data are obtained from a nationally representative sample that includes an adult, youth, and family interview in selected households. Upon completion of the interview(s), all sampled persons (SPs) are given a standardized physical examination. The examinations take place in a specialized mobile examination center (MEC).

NSPY is sponsored by the National Institute of Drug Abuse (NIDA) in conjunction with the National Youth Anti-drug Media Campaign (NYAMC). The survey is a nationally representative sample of 9 to 17 year olds in the U.S., with in-person interviews of the sampled youths and their parents or caregivers. The sample design is an area probability sample with three stages: 90 primary sampling units (PSUs), area and new construction segments within PSUs, and households within segments.

Both surveys use all four reporting files to develop sampling frames. The frame development process is primarily the same for NHANES and NSPY. Note that timing of sampling of new construction can vary depending on the survey. For example, new construction sampling is done one PSU at a time for NHANES because of the complexity of the survey. However, for NSPY, new construction sampling was done at the same time for all PSUs.

Frame construction begins with an editing process. Each file is subset to the counties of interest For example, NHANES creates about 15 files per year. Next, a file source variable is created on each file. This variable is used to eliminate duplicate data in the final step of the frame creation. After creating the source variable, the four files are combined into one file. It is important to note that the U.S. Census Bureau changed many place name identifiers after 1991. To avoid any confusion later, the place name identifier from the most recent survey date is retained for all years (during the process of combining the files). In addition, all place descriptions such as "town" or "village" are removed from the file.

The next step in the process is to review the data on the combined file, particularly the number of housing units for 1990. The 1990 data is only available as annual figures; occasionally the number of units reported in a year is quite high. As a means to reduce the burden on field staff (or listers) and create more uniform sampling units in terms of size, the 1990 counts are reallocated across the 12-month period. That is, monthly data are imputed for 1990. The first step is to determine a cut-off for imputation. For example, NHANES uses a cut-off of 500. Note the cut-off depends highly on the survey (e.g., average segment size, screening effort). Next, decide how to impute the data. One approach is to use monthly data from 1991 through 1993 to impute monthly data for 1990. A mean proportion of units is calculated for each month. Then the 1990 annual number of units is allocated according to these proportions. New records are created on the file to retain the counts by month; a flag is assigned to the new records to signify an imputed value.

Some surveys may find that even a count of 500 is too high; therefore, they may decide on a lower cut-off. This was the case for NSPY. Imputations at the monthly level were developed for 1990 annual records with more than 90 units of activity, and for any other annual records with more than 90 units of activity. These monthly imputations were based on monthly percentages from complete monthly records from other offices in the same Census region (Northeast, South, Midwest, and West), from the same year or from neighboring years. It is important to impute from offices in the same region for monthly imputations, as there are significant regional differences due to variable weather conditions by region.

The reference time period for NSPY was 1990 through 1998, but the Census records were only current to September 1998 at the time of sampling, so the months October through December 1998 were imputed for all offices, using quarterly percentages from previous years. Many of the NSPY imputations were only carried out for the sampled building permit offices (see Section 4), as they were only needed for segment sampling within sampled offices.

The final step of frame building is the deletion of duplicate information. Whenever monthly data are available, annual data are deleted from the file because the monthly data provides a better distribution in terms of sampling units.

3. Segment Sampling

The next step is to construct and sample segments. Again, there are different methods for constructing segments. For example in NHANES, a regular area frame is constructed based on data from the last decennial census and the new construction segments are defined by sets of residential permits issued during a specific time period (e.g., June 1998 through September 1998). The two frames, area and permit, are combined and segments are constructed based on measures of size. Regular area segments are assigned measures of size based on number of persons in a block or group of blocks. The new construction segments are assigned measures of size based the number of persons expected to reside in the new units covered by a permit during a month or combination of months; some assumptions are made about the number of persons expected to reside in the new units. For example, NHANES uses the national average to approximate how many persons will live in each new unit. For details about the development of measures of size, see Montaquila et al. (1999). After segments are selected, the permit offices are identified.

Sometimes cost and/or time influences frame construction. Issues such as the number of building permit offices and the amount of building activity must be addressed if cost and/or time is a factor. For example, NSPY sampled segments from 90 NSPY primary sampling units (PSUs) in a two-stage selection process, given the large number of offices in the NSPY PSUs (making a one-stage sample of segments from the segment frame too costly to implement because too many offices would need to be visited). In the first stage of sampling, building permit offices were sampled in a systematic probability proportionate to size sample, with measures of size within each PSU proportional to the amount of activity for the office in the 1990-1998 period. Offices with less than 30 units of activity over this entire period were excluded from the frame, and small offices (offices with 30 to 240 units of activity) were combined into permit office groups for this sampling process. A total of 130 building permit offices or groups were sampled for NSPY across the 90 NSPY PSUs. Very large offices were allowed to be hit multiple times (using a with-replacement rather than a without-replacement sampling scheme).

Within the sampled building permit offices, a total of 4 segments were drawn from the segment frame for that office (4 times the number of hits for large offices). The sample of four segments was a probability proportionate to size sample, with measure of size proportional to segment activity.

4. Listing Procedures

After identifying segment permit offices included in the sample, Westat field staff (referred to as listers) visit the building permit offices. Both NHANES and NSPY use listers to collect information on permits issued for the time period of interest.

In most building permit offices, the information is available in electronic format or computer printouts; the sort order of the file or listings differ from place to place. The type of information available differs as well. Occasionally, the permit information is retained on index cards or in hand-written logs. Although rare, sometimes the only source of information is the actual permit because the office does not produce any type of summary report. A sample information sheet is given to the listers as a guide. The sheet includes the time period of interest; the number of units reported to Census for the period of interest; and sampling information (e.g., random start, sampling rate). A pre-programmed calculator is used to select the sample (as described below).

Second, the lister records the selected permit number and number of units for which the permit covers. In addition to recording the permit number, the lister assigns a serial number to each unit; if the permit covers multiple units, a series of serial numbers is assigned to the permit.

The third step in the process is to select a sample or "chunk" of serial numbers listed on the sampling form. Listers have been given a pre-programmed calculator to select the chunk. The listers enter the random start, sample size, actual number of units, and expected number of units from the sample information sheet as input. Then the calculator displays a range of serial numbers. The range determines the starting and ending point of the sampled chunk.

Listers use another form to record unit or building addresses and builder information (e.g., name, telephone number, and address) for the chunk of serial numbers. Separate building addresses are recorded for multi-dwelling units that are covered by one permit.

The next step is to record the specific unit address including specific mailing address for multi-dwelling units (i.e., apartment number) of sampled units on a listing sheet. This sheet serves as the frame for further stages of sampling.

Depending on the nature of the survey an additional step may be taken to reduce the effect of clustering in sampled multi-dwelling units. This procedure involves re-sampling in multi-dwelling buildings that come into the sample through the process

described above. The NSPY survey uses this approach. The lister must complete one additional form. Whenever the chunk includes a portion of a multi-dwelling unit, the lister must record the address of all units in the building. If possible, the lister makes a site visit to the building and records the unit address of the physical location. Rather than including the cluster of units from the original chunk, a set of units is selected systematically from the listed units in the building.

During the data collection period, interviewers screen the households to determine whether the structure was built after the census. Structures built after the census are declared out of scope for the area sample because they will have a chance of coming into the sample through the new construction segments (segments created using the building permit file).

5. Identification of Units (Permit vs. Dwelling Units)

In area surveys, the dwelling unit is used as a sampling unit. For new construction, the permit serves as a proxy-sampling unit. It is important to note that the housing unit may never be built. During the 1980's, approximately 2 percent of all new housing units authorized by permits were never built. In recent years, the number of units never built has decreased; approximately 1 percent of all new housing units authorized by permits are never completed. Thus, the permit serves as a good proxy for the actual physical unit. Other sampling issues are discussed in the next section.

6. Sampling Issues to Consider

As mentioned above, with the method of permit sampling, dwelling units constructed prior to the census have a chance of coming into the sample only through area segments. On the other hand, dwelling units constructed after the census have a chance of coming into the sample through the permit segments only. Thus, the coverage of newly constructed units is a critical point for surveys that use permit files to supplement the area sample. There are a number of issues that need to be considered when permit files are used for supplementation purposes.

The building permit frame is not a complete file of all new housing units built in the U.S. This is due to the fact that:

- About 5 percent of housing units that are built in the U.S. are not included in the frame (because they are built in areas that do not require permits for new construction);
- Mobile homes do not require a building permit in most places, and thus are not included in the permit file; and
- About 1 percent of the housing units authorized by permits were never built, however, most constructions start the same month as the permit is issued or within three months of issuance.

An investigation of the counties included in a recent large survey showed that 99.6 percent of the counties have permit issuing offices. However, the geographic coverage of the permit offices varies from county to county. Approximately 63.9 percent of the counties have permit issuing offices that cover the entire county and 35.6 percent have partial coverage of the county. Places that have partial coverage tend to have permit issuing offices that cover towns or cities (excluding the surrounding areas). Since the difference is small, coverage can be addressed by using a poststratification adjustment in order to calibrate the base weights to Census estimates of the proportion of persons living in newly constructed homes.

Although some places require a building permit for new mobile homes, the Building Permit Survey excludes mobile homes from the survey. Since mobile homes tend to be geographically clustered (in mobile home parks), new mobile homes may be quite prevalent in some areas and non-existent in other areas. Furthermore, the distributions of some characteristics (health or education) may differ between persons living in mobile homes and persons not living in mobile homes. The exclusion of new mobile homes from the sampling frame may result in significant increase in variance. Several alternatives were examined to offset the increase in variance associated with new mobile homes.

The first alternative is to retain new mobile homes in the sample. Under this alternative, a new mobile home is not screened out, unlike new permanent residential units. Due to the clustered nature of new mobile homes, however, this alternative may result in a need for subsampling in areas with large numbers of new mobile homes. This is true particularly if controlling the sample size is critical. Subsampling may necessitate truncation or trimming to reduce the variability in the weights. Thus, with this alternative, it is likely that an underestimate of the number of new mobile homes would result.

The second alternative involves adding "new mobile home" segments to the area component of the survey. The lister would look for mobile home parks within the area segment, and determine whether the mobile home park has new mobile homes. If so, the mobile home park will be excluded from the area

segment and a new segment would be created. This alternative is attractive since creating the new segments for new mobile homes in mobile home parks would maintain the correct measure of size for the area segment, and would assign correct measures of size to the new mobile home segments as well.

The last alternative is to conduct research on creating a new mobile home sampling frame. For example, the Yellow Pages could be used as a resource to identify mobile home parks. The managers of listed mobile home parks would be contacted and asked to provide counts of the number of new mobile homes in the mobile home park. However, this alternative does not account for unlisted mobile home parks. Coverage of new mobile homes not located in parks is a problem as well. This alternative is not as attractive as the first two; furthermore, it would require a substantial amount of research and testing.

An additional source of undercoverage in new construction permits is the possibility of incomplete listings of permits provided by permit offices. Based on Westat's experience during the late 1980's, in some small numbers of occasions, the permit office could not locate the permits; reported no new construction permit for a given period, although the U.S. Census Bureau had reported a number of permits for the associated time; or refused to spend the resources to compile old data. NSYP experience the latter. A total of 516 segments were fielded altogether in NSPY in the 130 offices. Of these segments, 486 (94.2 percent) were successfully listed, and 30 (5.8 percent) were not listed. The 30 unlisted segments were generally segments from early in the decade, and the offices were unwilling to expend the resources to recover these segments (in some cases, the permit information was destroyed). This 5.8 percent non-completion is a contribution to the non-response rate for NSPY, as the households corresponding to these units have no other chance of responding to the survey.

Also, note that the boundary between the regular area sample and new construction sample is not perfectly defined. The screening out of the new construction in regular segments relies on the respondents' knowledge of when their residential home was constructed. Furthermore, persons residing in households with permits issued before 1990 but built after 1990 will be excluded from the sample and thus will not be represented. The prevalence of this occurrence is expected to be very small.

7. Conclusion

Since most area surveys rely on census data to construct sampling frames, it is important to address the accuracy of measures of size or the "up-to-dateness" of the frame particularly for surveys conducted late in the decade. New construction or building permit sampling is cost effective and relatively easy to implement; however, additional procedures are required to reduce variance associated with new mobile homes and to address coverage issues related to new homes built in areas that do not require building permits. Never the less, permit sampling remains a plausible approach to maintain the efficiency of the sample and thus reduce variances in area segments.

In this paper, we have presented the history behind building permit sampling; a procedure for creating a sampling frame; and a procedure for selecting new homes from new construction permits. In addition, we have examined coverage issues that still exist. Particularly, we have discussed homes built in places that do not require permits; new mobile homes; and prevalence of new homes never built.

Currently, the Building Permit data represents approximately 95 percent of newly built homes in the country. We believe that the remaining 5 percent can be accounted for through weighting. Sampling issues associated with new mobile homes is an area that should be examined more carefully. We have described three methods (or alternatives) for accounting for new mobile homes built after the census and reducing variance associated with them. Finally, we have concluded that the number of homes that are never built is small; thus, permits provide an accurate estimate of measure of size for homes built after the census.

8. References

Montaquila, J., Bell, B., Mohadjer, L., and Rizzo, L. (1999). A method for sampling households late in a decade. To appear in *Proceedings of the Survey Research Methods Section.* Washington, D.C: American Statistical Association.

U.S. Bureau of the Census (1994). *The Current Construction Report. Housing Units Authorized by Building Permits: Annual 1994,* Publication C-40/94-A, U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census (1978). *The Current Population Survey. Design and Methodology*, Technical Paper No. 40, U.S. Government Printing Office, Washington, D.C.

Waksberg, J. (1998). Waksberg: The Hansen Era: Statistical Research and its Implementation at the U.S. Census Bureau, 1940-1970. *Journal of Official Statistics*, Vol. 14, No. 2, p. 115-118.

Acknowledgments

We would like to thank Joe Waksberg for his support and insight in the development of this paper.