COVERAGE STUDY FOR HOUSING STARTS IN NONPERMIT AREAS

Jack Pullen, U.S. Bureau of the Census

1. INTRODUCTION

1.1 Nonpermit (NP) Survey Design

The NP Survey, part of the Surveys of Construction (SOC), measures monthly housing starts in nonpermit areas. The sampling units consist of all PSUs from the SOC sample containing nonpermit areas. Every month Census interviewers collect information about new housing from "Sources," who are familiar with construction activity in these areas. The Sources are chosen to cover the entire NP area in sample. Examples of Sources include the local postmaster or government official, a representative of a utility, a building supplier, the owner of a lumber yard.

Since the list of new residential construction supplied by the Sources will be incomplete, a sample of area segments was selected in sampled PSUs to measure the undercoverage. Segments are Enumeration Districts (EDs) defined for the most recent, in this case the 1970, Census of Population. After collecting Source reports, the Census interviewers canvass the segments and compare starts reported by the Sources to the starts picked up in the enumeration. In areas not covered by segments, interviewers are required to call the owner or builder of each project to determine the date of start and whether the project is in-scope. When neither the owner nor builder can be reached by telephone, interviewers may visit the construction site.

In brief, the survey's design is an example of dual-frames sampling, with Source-reports forming an incomplete list frame, relatively inexpensive to enumerate, and area segments forming a complete geographic frame, more expensive to enumerate.

1.2 Estimation in Nonpermit

This design makes possible two different estimates of housing starts. The first is the dual-frames estimate, consisting of the single-staged sample estimate from Sources, together with an estimate of undercoverage from the segments; the second is a two-staged sample estimate (PSUs and area segments) from segments alone. Letting these estimates be denoted by S_1 and S_2, respectively, their expressions are:

Notation: Throughout this paper, h refers to the PSU index.

$$S_1 = \sum_h \{X_{1h} \cdot W_h + P_h \cdot (X_{2h} + X_{3h})\}$$
$$S_2 = \sum_h W_h \cdot (X_{1h} + X_{3h})$$

X_{1h}: Starts not reported by a Source, but found in the canvass of segments

X_{2h}: Starts reported by a Source outside the segments

X_{3h}: Starts reported by a Source and also found in the canvass of the segments

P_h : Inverse of probability of selecting the h^{th} PSU

W_h : Inverse of (overall) probability of selecting a segment in the h^{th} PSU

The starts represented by X_{1h}, X_{2h}, and X_{3h} are referred to, in order, as class 1, 2, and 3 starts.

The lower sampling error, owing to its smaller second-stage component, makes S_1 preferable to S_2 as an estimator of housing starts. The second estimate, S_2, ignores any units from outside the segments, i.e., not within the two-staged sample. If S_1 and S_2 are unbiased, their difference should be within sampling error; or, equivalently, $E(S_1/S_2) = 1$. Since the beginning of this survey in the early 1960s, however, S_1 generally has been considerably greater than S_2. Some recent estimates are given in Table 1 (and the Appendix). References will be made to the ratio, S_1/S_2, rather than to the difference.

Table 1.

STARTS RATIOS

Year	Ratio	C.V.
1975	1.18	0.11
1976	1.14	0.11
1977	1.23	0.11
1978	1.36	0.11
1979	1.15	0.09
1980	1.14	0.11
1981	1.21	0.09
1982	1.15	0.10
1983	1.07	0.10
1984	1.12	0.09

1.3 Possible Sources of Bias

The bias in the starts ratio seems to be due to some combination of class reporting errors (class 1, 2, or 3), or to some systematic difference in the rate at which Sources choose to report in the segments compared to the areas outside the segments. The kinds of **Source-reporting** errors that could produce a positive bias in the ratio are the overestimation of class 2 starts and the underestimation of class 3 starts. Errors in counting class 1 starts in themselves are not detectable, but they may augment an existing bias brought on by Source-reporting errors. In the analysis of available data, no other reasons for the bias were implicated. Therefore, discussion will be limited to these three.

Experience with the NP canvass suggested that class 1 errors were rare, and normally would occur when special circumstances, such as poor weather, prevented a complete canvass. On the occasions when it happens under normal conditions, an experienced interviewer has enough opportunities that the chance of completely missing the project should be minimal. This is borne out by the following example. During an observation of pre-listing for the study, the completed frame of a two-story house was missed by the special interviewer, because it was nearly invisible from the road. The observer first noticed the case. Such a case ordinarily would have been discovered while canvassing one of the

other roads; in this instance, not even an entrance road could be found. Nevertheless, the interviewer managed to collect the necessary information from a man she talked to about another project.

Unlike class 1 reporting, the quality of Source reporting was very much in question, chiefly because of inherent over-reporting. Since several Sources may report for an area, duplicates resulting from overlapping coverage present a constant problem. In addition, the same Source sometimes reports a project several times. For instance, in one area the postmaster commonly reported a project when it began, and again when the family moved in (probably corresponding to the initial request for mail service and the notice of change of address). A complication making detection of duplicates more difficult is that each category of Source -- building supplier, local official, etc. -- becomes aware of a project in a different phase of the planning or construction.

Along with duplicates will be a number of early reports, reports of abandoned projects, and various other kinds of out-of-scopes. These must be verified by the interviewer, possibly on the basis of inadequate information. Many reports contain little more than the site location, or give only the most general idea of the actual location of the project. In the event that the builder's name is given, the interviewer may be able to get information on the project by telephone. But with such incomplete descriptions, the wrong project could easily be picked up.

It was also thought that under-reporting of starts in the segments (class 3 errors) could contribute to the bias. Since the interviewer will canvass the segments, she may discourage reports, for instance, by requesting segment information last, or by not requesting details about reported construction. The interviewer also may tolerate poor Sources more readily in the segments and not seek replacements.

2. DESIGN OF THE NP STUDY

2.1 The Three Treatments

Attention was to be focused on Source reporting, since class 1 errors should not present a significant problem and their analysis would, in all likelihood, entail disrupting reporting and canvassing patterns. Accordingly, the NP Study was designed to detect differences in Sourcereporting rates between segment and nonsegment areas.

Nonpermit EDs residing in SOC sample PSUs were selected for the study. Each ED was assigned to one or more of the following three treatments:

(A) The ED is canvassed by Census interviewers using the procedures from the NP Survey for collection of Source-reports and canvassing. During the canvassing the interviewers match Source-reports to starts found on the ground. These EDs are called segments because they were represented by the NP Survey's area sample of segments from PSUs in the study.

(B) The ED is canvassed by specially trained, experienced interviewers. The special interviewers match Source-reports supplied by Census interviewers to starts found on the ground. In addition to this, reports are reviewed for errors in out-of-scoping and unduplication. These EDs measure the actual rate of Source-reporting for the NP Survey as accurately as possible. These are the Control EDs.

(C) The ED is canvassed for starts found on the ground, but no attempt is made to match starts to the Source-reports. Reports are allocated to the EDs beforehand by the special interviewers using only the site location listed on the report. These are called Source EDs because they represent conditions in the areas covered by only Source reports.

To reduce costs, treatments (B) and (C) were applied jointly to a single set of EDs, referred to as study segments.

2.2 Selection of Sample

Eighteen PSUs from the SOC Nonpermit sample of PSUs were selected for the NP Study. The NP PSUs were stratified into 3 regions, Northeast, Midwest, and South. The West was excluded for lack of NP activity. Twelve PSUs with high 1981 housing starts activity came into sample with certainty. Within each region, the remaining PSUs were divided according to whether their starts ratios were close to one. Then one PSU was sampled from each stratum with probability proportional to its 1981 housing starts activity.

Study segments were selected using random sampling from areas not in the existing sample of segments in study PSUs. Enough EDs were selected from a certainty PSU to ensure, on the average, at least one start per month over the six months of the study. In non-certainty PSUs the sample consisted of a single ED. All segments from the NP survey residing in study PSUs came into sample.

2.3 Timing of Recanvass

The NP Study was conducted between March 15 and November 15, 1983. The pre-listing of study segments began the latter part of March and was completed for all areas by May 1. Canvassing of study segments took place every two months, beginning in July. The interviewers recorded housing units started since the previous canvass (or pre-listing) and the start dates. Housing units started in the same month as the canvass were held for the next canvass. Study segments from the six non-certainty PSUs were originally scheduled for a single canvass in November. After the July canvass, we decided to begin normal bimonthly canvassing in these PSUs, also. They were canvassed in September and November.

2.4 Handling Late Reports

A Source-report with a start date after the date verified by the special interviewer was counted as a start in Source EDs for the date given on the Source-report. Control EDs, on the other hand, represent the actual rate of Source-reporting as best we can measure it. Thus, we must apply procedures similar to those used for segments. Census interviewers hold reports for projects that either do not have start dates, or for which the start dates are incorrect, until the owner or builder confirms the start. So, for Control EDs, a Source-report having an entry date

before the start date verified by the special interviewer was counted no matter what the recorded date of start was on the original Source-report.

3. DATA ANALYSIS

To restate, our objective was to compare rates of reporting in segments and in Source EDs to the Control EDs. A t-test was performed on the paired PSU observations, comparing the average of the differences, D_1 or D_2, to zero, where,

$$D_1 = \sum_h d_{1h}/11$$
$$D_2 = \sum_h d_{2h}/11$$

with,

d_{1h}: The difference in rate of Source-reporting between segments and Control EDs in the h^{th} PSU

d_{2h}: The difference in rate of Source-reporting between Source and Control EDs in the h^{th} PSU

Notice that 11 PSUs were averaged out of a possible 18. Only PSUs selected into the study with certainty were included in these tests. There are 12 certainty PSUs; however, PSU 665 did not generate any Source-reports during the study, and was eliminated for this reason from the analysis.

3.1 Summary of Results

The following table summarizes the start rates of reporting for PSUs in the study. The start rate for a treatment is the quotient of the number of verified Source-reported starts to the number picked up in the canvass.

Table 2.

STARTS RATES DURING STUDY PERIOD

PSU	Source	Control	Segment
248	.68	.68	.79
425	.41	.29	.50
586	.50	.50	.65
591	.71	.71	.80
626	.80	.70	.50
640	1.33	1.00	ZERO
657	.75	.75	.49
665	1.00	1.00	.09
669	.69	.54	.80
677	.59	.35	.22
679	1.13	.63	.39
699	.33	ZERO	1.00

No difference in average" effects was detected between rates for segments and the Controls at the 5 percent level of significance (t = .01). But it should be noted that the absolute deviations are quite large. This probably reflects the small sample size, but may indicate a difference in the quality or volume of reporting in the two kinds of areas. The average difference was so close to zero that the net bias expected to come from segments is probably small, but they may contribute a large portion of the bias observed in a particular month.

The test did detect very significant over-reporting in Source-EDs compared to the Controls (t = 3.14). Twenty percent of the reports in certainty PSUs were over-reports. The types of errors recorded by the special interviewers were:

Type	Units
Outside segment or not found	4
Late report	7
Duplicate	3
Out-of-scope	3

3.2 Discussion

According to instructions, the special interviewers were supposed to investigate all unmatched Source-reports to see if the projects could be found outside of the segments. Only projects with addresses placing them inside the study segments were to be considered. So many such cases were listed, though, that we decided to count as outside/not found only the ones which were thoroughly annotated on the listing sheets. In all, 27 projects were recorded in this category. But only four confirmed cases were accepted. An example of a "not found" is the report listing a possible segment address, the owner, and giving a church as a landmark, for which neither the owner, nor church could be located. This case was verified by an observer.

Late reports included several projects reported as started at the time additional work was done: A September report for a modular home, for which the foundation was laid in July, and the home set on the foundation in September; a June report for a house started in the summer of 1982 and moved into in June, 1983. Another report, actually counted as a duplicate in our tally, was a duplicate of a late report for a previously matched project.

Out-of-scope reports fell in the category "nonresidential reported as residential." For instance, a garage was mistakenly counted on one report. The most interesting out-of-scoping error noted was made by one of the **special** interviewers. A house had been built on the foundation of one recently gutted by fire. The project was reported by a Source, and the Census interviewer correctly classified the report out-of-scope. After relating that the house had been built on an existing foundation, and so on, the special interviewer casually reclassified the project to in-scope.

Several types of under-reports were discovered by the study. An interesting under-report that occurred too early to be counted in the results was a project started in April, 1983 and reported by a Source. The Census interviewer failed to contact the owner to confirm the start. The house was moved to a new site in September, but was not counted in either month. We also found unexpected under-reporting in a PSU where a permit office acts as Source. When a previously nonpermit-issuing area becomes permit-issuing, the area is canvassed for about three months. If no errors are detected in that time, the permit office becomes a Source. We expect it to be a very good Source. In this case we happened to

find a very poor one.

A look at the examples makes clear that most of the errors should have been avoided had interviewers followed existing SOC procedures. Telephone followups and site visits should have identified most of the over-reports, including late reports, as well as the under-reports caused by the failure to get a start date.

The potentially serious problem of under-reporting by the permit offices, deserves further investigation.

3.3 Comparison of Starts Ratios

A starts ratio was computed for study segments to measure the effect of the class 2 overestimate. All 18 PSUs were included in the estimate. Let R'_{st} be the starts ratio for the study segments. Then

$$R'_{st} = \frac{\sum_h W_h (X_{1h} + X'_{3h})}{\sum_h W_h (X_{1h} + X_{3h})}$$

where,

X_{1h} : Number of starts reported only through canvass in study segments in the h^{th} PSU

X_{3h} : Number of starts reported by Sources and verified by special interviewers to be in study segments in the h^{th} PSU

X'_{3h}: Total number of starts reported by Sources and allocated to study segments by special interviewers in the h^{th} PSU

W_h : Inverse of probability of selecting study segments in the h^{th} PSU

The ratio for the study was $R'_{st} = 1.05$. This was the same as the starts ratio for the survey as a whole during the period. Unfortunately, the survey's ratio was not significantly different from one at 5 percent. Hence, there was no hard evidence of a bias. At the same time, Source-reporting errors confirmed by the study were sufficient to produce a 5 percent bias in R'_{st}. A fair conclusion seems to be that a large portion of the apparent bias in the survey's ratio during the period may be due to Source-reporting errors of the type confirmed by the study, instead of random fluctuations in the rate at which Sources reported in the various areas.

4. SUMMARY

In our study of the NP Survey significant over-reporting (class 2 errors) was detected in Source-only areas. The argument is far from ironclad, but it appears that most of the observed bias in the NP Survey may be the result of this over-reporting. The SOC procedures designed to handle the reporting errors discovered are reasonable, but for some reason in many cases are not being followed.

The possibility that under-reporting may contribute to the bias in the segments was not completely laid to rest. The reporting error in the segments was nearly zero, but there were wide swings between PSUs. If Source over-reporting were corrected this could lead to net under-reporting in the segments with an attendant bias.

APPENDIX

Starts Ratios

1975	1	2	3	4	5	6	7	8	9	10	11	12	ANN
Ratio	1.30	1.60	1.24	1.32	1.22	1.06	1.12	1.18	0.93	1.16	1.30	1.26	1.18
C.V.	0.19	0.12	0.12	0.14	0.14	0.17	0.18	0.12	0.27	0.15	0.18	0.18	0.11
1976													
Ratio	1.15	1.19	1.64	1.02	1.30	1.50	1.12	0.90	1.04	0.92	1.25	1.26	1.14
C.V.	0.18	0.13	0.13	0.29	0.24	0.13	0.17	0.10	0.17	0.14	0.15	0.18	0.11
1977													
Ratio	1.67	1.21	1.43	1.14	1.26	1.17	1.04	1.11	1.12	1.45	1.52	1.22	1.23
C.V.	0.21	0.15	0.16	0.20	0.13	0.21	0.17	0.20	0.13	0.11	0.13	0.21	0.11
1978													
Ratio	1.41	1.58	1.71	1.51	1.32	1.29	1.16	1.29	1.31	1.53	1.28	**	1.36
C.V.	0.15	0.18	0.18	0.12	0.17	0.14	0.21	0.25	0.19	0.18	0.18	**	0.11
1979													
Ratio	**	**	**	1.27	1.09	1.35	1.08	1.09	1.27	1.13	1.12	0.98	1.15
C.V.	**	**	**	0.15	0.14	0.09	0.12	0.08	0.11	0.10	0.19	0.19	0.09
1980													
Ratio	1.06	1.31	1.47	1.10	1.05	1.06	1.17	1.17	1.23	1.15	1.05	1.03	1.14
C.V.	0.11	0.18	0.29	0.21	0.15	0.11	0.16	0.16	0.13	0.16	0.13	0.11	0.11
1981													
Ratio	1.40	1.54	1.08	1.14	1.27	1.01	1.25	1.39	1.22	1.21	1.29	1.05	1.21
C.V.	0.16	0.17	0.14	0.11	0.15	0.11	0.14	0.22	0.11	0.14	0.16	0.20	0.09
1982													
Ratio	1.00	1.29	1.51	1.11	1.24	1.07	0.96	1.11	1.02	1.36	1.36	1.11	1.15
C.V.	0.15	0.21	0.18	0.11	0.19	0.11	0.15	0.13	0.14	0.15	0.16	0.24	0.10
1983													
Ratio	1.01	1.33	1.17	0.95	0.99	0.93	1.17	1.03	1.09	1.14	1.11	1.29	1.07
C.V.	0.13	0.16	0.17	0.13	0.13	0.15	0.17	0.10	0.14	0.16	0.15	0.22	0.10
1984													
Ratio	1.21	1.22	1.03	1.13	1.11	1.03	1.14	1.07	1.16				1.12
C.V.	0.17	0.12	0.11	0.15	0.14	0.13	0.15	0.13	0.13				0.09

**Phase in period of redesign.