ESTIMATING NEW CONSTRUCTION LAG IN THE MASTER ADDRESS FILE: RESULTS FROM THE 1999 PERMIT ADDRESS MATCH STUDY[1]

Anthony Tersine, Clifford Loudermilk, and Randall Parmer, U.S. Census Bureau
Anthony Tersine, U.S. Census Bureau, DSMD 3784-3, Washington, DC 20233

Keywords: New construction, Master Address File, Lag time

1.0 Introduction

The Permit Address Matching Study (PAMS) was undertaken to determine the match rate between the United States Postal Service's Delivery Sequence File (DSF) and the household demographic surveys' permit frame. We also attempted to measure the time lag between when a unit is built and when it shows up on the Census Bureau's Master Address File (MAF), since the primary source of housing unit updates to the MAF is the DSF.

Our interest in coverage and lag of new construction is two-fold. First, we were interested in whether the lag in MAF updating of new construction is adequate for the American Community Survey or if not will we need to consider special updating procedures, such as are planned for the Community Address Updating System (CAUS). Second, we were interested in determining how the lag in MAF updating compares with the lag in the permit frame for the demographic surveys.

Our hypothesis had been that the DSF provides fairly good coverage of permit new construction. We suspected, however, that the lag in updating the MAF with new construction could be fairly substantial due to delays in updating the DSF or due to inherent delay caused by the MAF creation methodology. To study the issue, the PAMS was undertaken in July-November of 1999.

The permit addresses were obtained for this study from the demographic surveys' sample database. These permits are collected monthly from local building permit offices. We decided to study permit addresses from throughout the decade of the 90's. Since we had no way of determining what the lag might be, we decided this was the safest course of action. Since we had strong hope

that the lag was no more than 2 or 3 years, we decided to subsample permits older than that.

The study was limited to 17 counties in the 1999 American Community Survey with moderate to substantial permit activity plus an additional 6 counties with substantial permit activity. These counties were almost exclusively areas with house-number-street-name addresses, which we will call "city-style" addresses. The advantage to confining our study to city-style addresses is that the MAF can readily be updated with DSF information in these areas. In areas with Post Office (PO) Box or Rural Route addresses, the MAF cannot readily be updated with DSF information.

The first step in the process was to sample the permit files. A description of the files is provided in Section 2.0. The sampling procedure is described in Section 3.0. Next, computer matching was performed between the permit file and the MAF. This is described in Section 4.1. Clerical matching is described in Sections 4.2 and 4.4, with field follow-up described in Section 4.3. Undercoverage results are presented in Section 5.0; lag results are presented in Section 6.0. Conclusions are summarized in Section 7.0.

2.0 Description of Source Files

The two sources of data used in PAMS were:

- the permit addresses collected for current demographic surveys (permit years 1989-1998)

- the Master Address File (MAF) extracts used for sampling the American Community Survey for sample year 1999

2.1 Permit Addresses

A building permit is an authorization by a building permit office (covering a county, city, township, etc.) to build a

---

structure for a particular purpose on a given plot of land. Because a permit is authorized, however, does not guarantee that the structure will ever be completed or occupied, although most are completed. Since permits are required for new housing nearly everywhere in the nation, building permits are a good source of addresses for new construction.

Although the information available on building permits varies from office to office, there are a few essential data items that every permit should contain: the date the permit was issued, the size (i.e., the number of housing units) of the structure to be built, and the location or address of the structure. The permits also may have ZIP Codes, subdivision names, a physical description, and other data that can help in locating the structure later in the field.

Each month, field representatives from the Census Bureau visit selected permit offices to collect address and other information from the newly-issued permits for privately-owned, residential housing. The purpose of this is to update the sample of housing units for each of our household-based demographic surveys with new construction. Data collected during these permit office visits from the past decade are available on computer files and are the source for the permit addresses used in PAMS.

There were two reasons these permit addresses present an attractive opportunity for an evaluation of the DSF updating of the MAF:

- The permit addresses, readily available, represented exactly what we expected the DSF to be adding to the MAF over time—new construction housing that did not exist at the time of the last census. So each permit address was a ready-made data point, a potential "success" or "failure" for the DSF.

- The permit addresses were "time-stamped" (i.e., we know the date the permit for each was issued), with an adequate distribution over the years in which we had an interest. This gave us the ability to estimate the elapsed time between when a housing unit is authorized for construction and the time it appears on the DSF, an important issue for any survey that seeks to sample from the MAF.

After the 23 counties involved in PAMS were identified, all the permit addresses we had that were issued from 1989 to 1998 in those counties were gathered. Note that we had only a relatively small subset of all the new construction addresses that came into existence in these counties in 1989-1998. The demographic surveys collected information for sample permits only, not all the permits. This means that we could do a one-way match only: we would try to find a match on the MAF for each permit address, but we would not try to match each MAF record to the permit file.

We discarded a small number of permit address records that did not have good city-style addresses (i.e., the house number, the street name, or both were blank), since these could not be matched in any meaningful way to the MAF. The remaining records were sampled (details are provided in section 3.0). After sampling, the selected address records were matched against the Census Bureau's Topologically Integrated Geographic Encoding and Referencing (TIGER) files to try to assign to each a census tract and block code, which could be helpful in the computer matching. The resulting file was the input into the PAMS computer matching.

2.2 Master Address File (MAF) Extracts

The MAF extracts used for PAMS were those produced for the American Community Survey for sample year 1999. The most recent Delivery Sequence File (DSF) used to update these MAF extracts was the one delivered to the Census Bureau in September 1998. These MAF extracts were modified in two ways:

1) Only those MAF records with the DSF listed as a source were retained for matching in the 23 PAMS counties.

2) To be consistent with the permit records, the MAF records were collapsed into one record for each basic street address. That is, the different unit records within each multi-unit address were combined into a single record with a revised total number of units.

These MAF inputs, much larger than the corresponding permit files for each county, were then sent to computer matching.

3.0 Sampling

Due to the large volume of permit addresses obtained in our 23 counties, we needed to sample the permit addresses. We did this because we only had the resources to handle 3/4 of the permits we had available for this study. We designed our sample to provide a large enough workload in counties with a small number of permits. We used a stratified systematic sample with the following

strata: state/county and permit year. We set a sampling rate for each year, and then if the number of permits was not a predetermined minimum number for that year, we adjusted the sampling rate to obtain that minimum. If the number of permits for a state/county and permit year combination was less than the predetermined minimum, we sampled all permits for that combination. For permits issued in 1996-1998 all permits were in sample. The sampling rates and predetermined minimums are given in Table 1 below.

Table 1. Sampling Rates and Minimum Sample Sizes

Years	Rate	Minimum
1989-1991	1/4	40
1992-1993	1/3	50
1994-1995	2/3	60

In six counties there was a subsampling process of units that went to field follow-up due to time constraints. The subsampling rate was 50% in one county and 80% in the other counties.

4.0 Matching Process

The matching process involved 4 steps: computer matching of the address, clerical matching of non-matched addresses, field follow-up of any remaining non-matches, and a final resolution based on the field results.

4.1 Computer Matching

The addresses were first matched by a computer using probabilistic matching. Every address is assigned a probability of matching to an address on the other file. Those with high probabilities were considered matches, those with middle probabilities were considered possible matches, and those with low probabilities were considered non-matches. Also, any permit that matched to a non-residential address on the MAF was considered a possible match, and anytime the number of units at the address disagreed it also was considered a possible match. Every address on both files was assigned a match code when the computer matching was finished.

4.2 Clerical Matching

Clerks took the results from the computer matching and could modify the match codes based on the information available to them. In most cases, they changed possible matches to matches or non-matches to possible matches.

4.3 Field Follow-up

Any permit address that was coded as a possible match or non-match was sent to the field to obtain information on the address. The object of the field follow-up was to ascertain whether the unit existed or not. If it did exist, then we tried to obtain the correct (updated) address, the month and year built, and the month and year first occupied.

4.4 Final Resolution

After information was obtained in the field, the address was recoded using the information collected. This was the only operation during which a unit could be coded as being out-of-scope (i.e., the unit did not exist, it was found in a different county, it was not a housing unit, or it was built after the date of the DSF). All permit addresses had to have a final match code after this operation and there could be no possible matches.

5.0 Undercoverage Estimates

Undercoverage is when an existing residential address is missing from the MAF. This is possible when a permit address matches to a MAF address flagged as non-residential, a single unit permit address is missing from the MAF, or when a permit address has 30 units but the MAF has only 20 units, which yields an undercoverage of 10 units. The undercoverage estimates, by year, are shown in Table 2.

Table 2. Undercoverage Estimates, by Permit Year

Permit Year	Estimate	Standard Error	Total # of Permit Housing Units
1989	0.2%	3.9	3,739
1990	2.7%	0.7	5,656
1991	7.8%	1.2	4,716
1992	1.1%	0.4	3,425
1993	0.9%	0.4	6,226
1994	2.7%	0.6	6,597
1995	4.3%	0.1	7,523
1996	10.3%	0.4	7,240
1997	40.0%	0.3	5,847
1998	80.5%	0.2	1,678

6.0 Lag Estimates

We would like to have measured lag by looking at permits that were on the MAF extract and finding the date they first appeared on the MAF extract. This information was not available on the MAF extract at the time this study was done. In the future, the MAF extract will have the date of the DSF the unit came from. A limitation to this will be the date when the Census Bureau first started receiving DSFs. For this study, we measured lag by looking at the non-matches and seeing how long they existed until the date of the DSF. We will present three estimates of lag time. The first estimate is based on our best guess of when the unit was occupied, by using the month and year occupied if known; if not known, then

using month and year built, if known; and if not known, then using the permit month and year. The second estimate is similar to the first, but if the permit date is used, then we subtract six months from the lag time. Six months is the current delay time from when a permit is issued until it can show up in sample. The third estimate is based on the permit date. The estimate of lag time will be presented, in months, in Table 3.

Mean lag time has been measured for the three methods of estimation as 10.6, 9.6, and 16.5 months respectively. We believe method 2 is the most accurate reflection of the lag from the time the permit meets our definition of a housing unit, because we know there is a delay from when the permit is issued until the unit is completed.

Table 3. Lag Estimates (in months), by Permit Year

Permit Year	Best Guess (method 1)		Best Guess (method 2)		Permit Year (method 3)	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
1989	106.0	0.0	102.0	0.0	107.3	0.0
1990	100.3	0.2	95.0	0.2	102.2	0.2
1991	75.7	2.3	74.0	1.6	82.7	0.4
1992	68.3	0.8	67.0	0.8	74.3	0.2
1993	52.1	2.8	51.0	2.8	62.1	0.3
1994	40.0	8.9	38.9	9.0	50.4	0.4
1995	28.5	0.9	27.1	0.8	38.6	0.2
1996	13.0	0.5	12.6	0.5	24.8	0.4
1997	5.8	0.2	4.7	0.1	13.3	0.1
1998	2.3	0.2	1.5	0.1	3.9	0.1

7.0 Conclusions

Using method 2 of calculating lag time, and comparing Tables 2 and 3, we can see that for permit year 1998, the mean lag time is 1.5 months and the match (coverage) rate is around 20%. By permit year 1997, the mean lag time is 4.7 months and the coverage rate is about 60%. As we move back to permit year 1996, we see a 90% match rate with a lag time of 12.6 months. Permit year 1995 shows a 96% match rate with a 27.1 month lag. For years prior to 1995, the match rate appears to stabilize at around 98%. Year 1991 appears to be an anomaly that we cannot explain.

As mentioned in the introduction, one of the primary motivations for this paper was to determine the best

strategy for CAUS to use building permit records. As we see, the first year after a permit is issued seems to yield a rapid increase in the coverage rate of the DSF. After the first year, the coverage rate continues to increase, but at a much slower rate. Beyond the third year, we see a leveling of the coverage rate with no apparent increase in coverage. Therefore, we recommend that CAUS concentrate on using permits that are 1-3 years old. Note that this recommendation is based only on the 23 counties in this study. No inference can be made about the permit coverage in the country as a whole. Additionally, we would like to note that this study has purposefully excluded permit issuing areas with non-city-style addresses, so we cannot make any recommendations for CAUS in those types of areas.