ORGANIZING THE ANNUAL HOUSING SURVEYS AS A VERY LARGE RELATIONALLY ORIENTED DATA BASE[1]

Andrew A. Beveridge,
Columbia University

Jennifer A. Norris,
Barnard College

Since 1973, the Department of Housing and Urban Development, through the Bureau of the Census, has conducted a yearly nationwide survey of housing. Data on a wide range of topics are collected during face to face interviews with over 190,000 individuals. Plainly, the Annual Housing Surveys represent one of the largest longitudinal general social and economic data collection efforts ever undertaken.

Due to changing policy and substantive interests, as well as government requirements, the interview schedules have changed significantly from year to year. Since the great potential for data from the Annual Housing Survey is in longitudinal analysis, it is necessary to have common variable definitions and consistent formats.

To accomplish this, we have developed and implemented a system which includes: 1) documentation of the variables, questionnaires, and files across all years and surveys; 2) files created using one homogeneously defined data structure; 3) a simple system to produce custom user files; 4) a method to easily produce routine custom analyses and tabulations using the data.

We have applied the relational model to create a small data base which documents the interview schedules, files and variable definitions. From this we produce up to date documentation and computer programs which are used to update the Annual Housing Survey data base, to handle custom file requests, and to perform analyses.

Since 1973, the Department of Housing and Urban Development (HUD) through the Bureau of the Census has conducted a yearly nation-wide survey of housing. This survey represents the largest longitudinal data collection effort on housing and related social and economic issues ever undertaken in this country. It was initiated by HUD not only as a tool for studies to be conducted within the department but also with the expectation that such a data set would be an invaluable resource to academic and private users around the nation. Until recently, however, there has been little usage outside HUD, and what studies have been undertaken have been largely restricted to cross-sectional analyses, in no way realizing the full potential of this data. It has been our primary function at the Annual Housing Survey Project (AHSP) to put the data into an easily documented, usable format. In order to accomplish this, we created one central data base which contains all the necessary information on the incoming files, documentation and final output and which controls the production of all the reformatted files. We developed a simple system for updating the current data base in order to accommodate each new year of the survey as it arrives.

The Annual Housing Survey

The Annual Housing Survey actually breaks into surveys of two distinct samples, together comprising some 190,000 units interviewed each year. There is the national survey and then the surveys of the selected standard metropolitan statistical

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

©1981 ACM 0-89791-056-7/81/0500/0009 $00.75

areas (SMSA's). Surveys are conducted annually with SMSA samples rotating, with the national sample interviewed each year. The SMSA samples were initially split into three groups of twenty SMSA's and interviewed on a rotating basis every three years. Currently, due to cost considerations, there are four groups of fifteen, and they are interviewed once every four years. For all samples, during face-to-face interviews, data are collected on a wide range of topics including those such as housing quality, the income of family members, energy usage, and family structure. There are approximately 75,000 units in the national sample which are interviewed over a three-to-four month period with October considered as the approximate date of contact. The sixty SMSA's include the largest and several of the smaller fast-growing areas in the country. Of these, the twelve largest have a sample size of approximately 15,000 housing units, while the size of the remaining forty-eight is approximately 5,000 units. The SMSA interviews are conducted throughout a twelve month period from April of one year to March of the following year. Since the samples are distinct, users can combine data from the national and SMSA surveys to increase their coverage of any particular area.[2]

Aside from the scope and magnitude of the Annual Housing Survey (AHS), its great potential lies in the fact that the same housing units are surveyed every year, with appropriate adjustments for lost housing and new construction. Every unit is assigned a unique control number which remains constant from year to year, enabling researchers to track changes at the individual level rather than merely in aggregate. In addition, since the housing unit is the basis for the survey rather than the individual, the AHS is ideally suited to analyzing the flow of households through the nation's housing stock. It is also highly useful for any research on the population in general, since it contains detailed demographic and economic information on household members of the 190,000 units surveyed each year.

The Raw Materials

The difficulties which arise when using the data and documentation as they are available from the Bureau of the Census are many and varied, but they are most significant for someone wishing to take advantage of the possibility of longitudinal analysis provided by these surveys. Questions are decided on by ongoing discussions between HUD, Census and other government agencies. As a result, each year the survey instrument changes, sometimes with minor modifications, other times in a more significant fashion. Changes have included such things as a more detailed scale for responses, which occurred in the question concerning property value, and a marked change in the sense of a question about a particular topic, as in the battery of questions concerning housing quality. Although the main focus of the survey is housing, the AHS is a household survey, and so questions concerning many other directly and indirectly related areas are included. Also, other government departments occasionally sponsor a set of supplementary questions to be incorporated into the questionnaire as occurred in 1974 when the Department of Transportation paid for a series of questions about the journey to work or several surveys for which the Department of Energy sponsored questions on home fuel usage. More recently, a large supplement concerning the handicapped was added to the survey.

Since the Census treats each year's survey as a separate entity, until recently there had been no serious attempt to coordinate the documentation and file layouts of the surveys across years. Thus, it was up to the individual researcher to trace questions and responses across lists of documentation and file layouts. For anyone wishing to undertake a longitudinal study there was the further task of matching records from the same units across years. Clearly these are time consuming and tedious jobs, susceptible to errors at any stage. The AHSP was created in order to make the data more accessible to small- and large-scale users alike and also to develop a system so that data from future years could be easily incorporated within the larger framework.

System Design

We designed a system based on the relational data model to organize the Annual Housing Surveys' data and their documentation. We defined a number of unique keys and developed tables which map their relations to other elements in our system.[3] This system is used to produce documentation, order forms, data files, cross-tabulations, and retrievals. The various tables can be easily updated since each relation is defined once and only once, and therefore is "normalized."

In short, the two main elements needed for the system are a set of unique variable names and a set of unique control numbers, which identify the housing unit, both of which remain constant across years. Then tables indicating the relation of the variable name survey and year to incoming files, to the questionnaire, to the response codes, and to the master data base are defined. These relational tables are processed to generate the program statements, to update the file, to produce the codebook, and to provide various indices and the lists of variables used in the handbook and order form. Custom tabulations and retrievals are produced by processing a table submitted by a user.

Thus, the relational model provides a simple and elegant method to organize this massive data base and its accompanying documentation.

Implementation

Due to the variability from year to year of the tape locations and variable names assigned by Census and HUD, when we created the documentation data base, we decided upon unique variable names for each question and made these names the central organizing data item for all documentation either received or created by us.[4]

This simplification process logically broke down into two parts. The first was to define a simple, homogeneous file structure for the survey data itself and the other was to develop a documentation system which would incorporate all the extant sources of information and allow for easy cross reference. Once these two structures were defined in relation to one another, all that remained to be done was to write two series of programs: one that processed incoming files; and the other which would enable us to create subfiles of the large file for users with special interests.

The Documentation

The structure of the documentation data base became very simple since every entry was uniquely identified by the variable name (there are now 572) and a set of relational tables which flagged the type of information it contained. Our tables are of six main types (See Figure 1 for examples):

1) Variable name by incoming tape locations and field lengths for documenting the origin of all data on the homogeneous file (there are 11 extant files) as well as the value and universe codes.

2) Variable name by outgoing tape location which reduce to the final tape layout.

3) Value code by values which is merged with the file definition table to produce documents such as the codebook.

4) Variable name by the allocation variables and codes which enable us to preserve the allocation detail for each variable.

5) Universe code by universe.

6) Variable name by the question numbers and year of survey for indexing all variables back to the instruments that produced them (there are 17 through 1981).

Thus our central documentation files not only contain all the necessary information to trace each variable across every year but also can easily be broken down into a series of smaller lists for any of our file management programs.

The Design

This documentation system evolved in conjunction with the file structure we decided upon. Given the organization of the survey itself the simplest file structure was one in which each housing unit-year-

Figure 1a. Example Tables for AHS Data Base

1. File Definition Table

80N	VARNAME	START POSITION	LENGTH	UNIVERSE CODE	VALUE CODE

2. Value Record

*VALUE CODE
1 VALUE
2 VALUE
3 VALUE
*VALUE CODE

3. Allocation Code Table

ALLNN	VARNAME	VALUES	YEARS

4. Universe Code Table

UNIVERSE	CODE	VALUES

5. Questionnaire Item Number Table

QUESTIONNAIRE	QUESTION NUMBER	VARIABLE NAME

Figure 1b. File Structure for AHS Data Base

Control Number	Year	Survey	Var 1	Var 2	Var 3	.....	Var 572
0001	74	N					
0001	75	N					
0001	76	N					
0001	77	N					
0001	78	N					
0001	79	N					
0002	74	N					
0002	75	N					
.	.	N					
.	.	N					
.	.	N					
.	.	N					
.	.	N					
n	74	N					
n	75	N					
.	.	N					
.	.	N					
.	.	N					
n	79	N					

survey (i.e. national or SMSA) combination represented an individual record, with all the records for a particular unit grouped together and such that every record had exactly the same layout. Then any user would need only one file layout and the documentation of variable values compiled from the individual years in order to use the data. We simply made everything from the documentation to the survey data itself as "flat" as possible (see Figure 1 for file layout).

Once these documentation data base and file structures had been determined, the system to process the incoming files became trivial. For each year-survey we received a questionnaire, tape and a file layout. The first stage, then, and by far the most tedious, was a manual comparison, item by item, of every question on the instrument and its possible values with the central documentation file. Once this had been done we entered the file layout as it had been received, a table referencing the census variable name to our variable name and a short file consisting of any changes that needed to be made to the central value documentation (see Figure 2).

Once this data has been entered the remainder of the process is automated. Essentially there are only four programs. The first (INFORM) takes in the entered data and updates the documentation data base renumbering variables as necessary. The second (REFORM) is concerned only with the tape layouts, and it takes in the incoming variable list and file layout and the current master file layout from the central data base. It then matches entries by name and prints them in the sequence defined in the documentation data base (see Figure 3). Next it compares field lengths and creates a new file layout incorporating all the variables from both sets of files. Once the new layout has been generated it then creates a third program (or set of programs) to shuffle the data on the incoming file and if necessary also on the current national and SMSA files, expanding the two layouts into one larger layout incorporating both.

Once both layouts have been expanded to the same format, the new file is sorted by control number and merged with the appropriate survey-years. That is, incoming national files are merged with that longitudinal master file, and the incoming SMSA's are merged by SMSA. As yet since we have only received SMSA data through 1977, only the 1974-77 SMSA wave has been longitudinally linked. After merging the files another program (OUTFORM) checks to see which units have been added to or deleted from the sample and blank records are created to fill out the records for each housing unit so that every unit has a record for each year. This whole process is extremely straightforward and requires a minimum of human intervention, thus reducing as much as possible the likelihood of errors. Once the files are merged, however, we run a series of tabulations and manual checks both to verify our procedure and to double-check the census documentation.

Retrieval System Design

Since our purpose was not only to link and document the existing files, but also to be the primary contact for users and to make it as simple as possible for people to use the data, a similar series of programs was designed for the production of user files. This series starts with the fully merged files and creates custom files.

Users can reduce their file size in three ways. The first is that they can request specific variables relating to their particular study. As yet, this is the only part of the custom file production procedure that is completely automated (see Figure 4). The program to perform this reduction (USERFORM) is very similar to the second program in the first series, REFORM; but unlike REFORM, USERFORM is keyed by variable numbers rather than names to save in staff time. Users fill out an order form, really part of the documentation, which merely requires them to check off the variables they are interested in. We then enter these in sequence into a small file which is used as input to USERFORM. USERFORM produces the user's file layout including a brief description of each variable and the variable's sequence number in the documentation for easy reference. It also produces part of the program that ultimately controls the data selection.

The other two ways of reducing the file size are to take a random sample of the larger file or to select specific housing units based on their household characteristics. Users can also combine these two forms of requests. For example, one order we filled requested all units with a head of household fifty-five years old and older and a 20% sample of all other units. In addition, since the file has been longitudinally linked, file requests can be made which select units because of changes in values between years. An example here is a file which we produced for someone who was studying racial transitions. He wanted all white to black and black to white housing transitions plus a 20% sample of the white-white and black-black transitions, plus 5% of all other units. As you can probably see, users may request any boolean retrieval and random samples on any strata they define.

Conclusion

Since the AHS is so large, our ability to quickly and simply produce small custom files has greatly increased the accessibility of the data. No single user could possibly treat the many different topics covered by the survey, and relatively few have the computer resources to process so much data. Currently, the national file alone, which has been longitudinally linked from 1974 to 1979 (although with just the core variables for 1978 and 1979), contains 800 million bytes of data. There is almost as much data from the SMSA surveys and as yet we have not received any SMSA data after 1977. Once all the survey data which already has been collected is added to the file, we expect there to be about 2.5 billion bytes of data.

Figure 2. Documentation Update Procedure

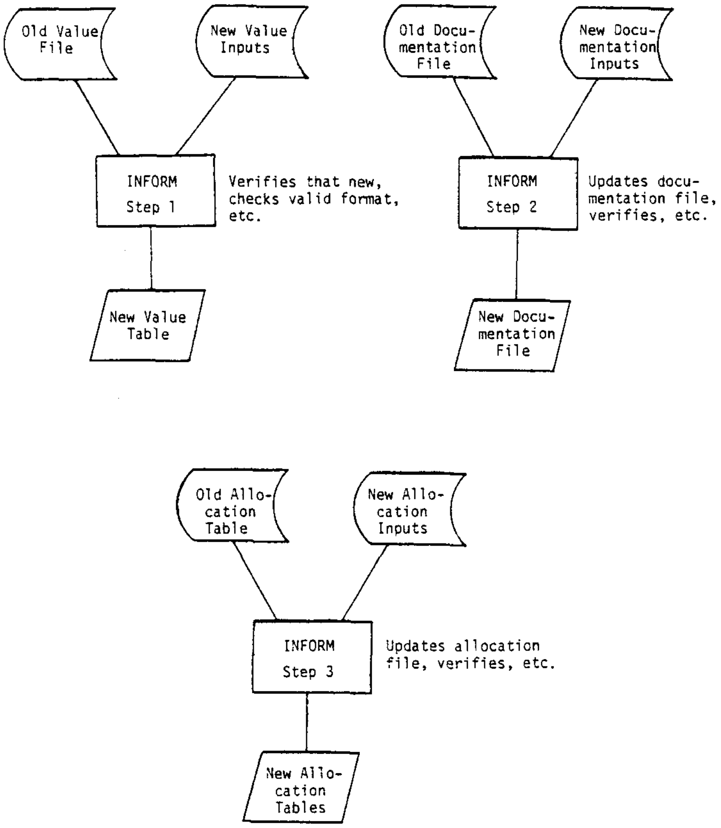


Figure 3.

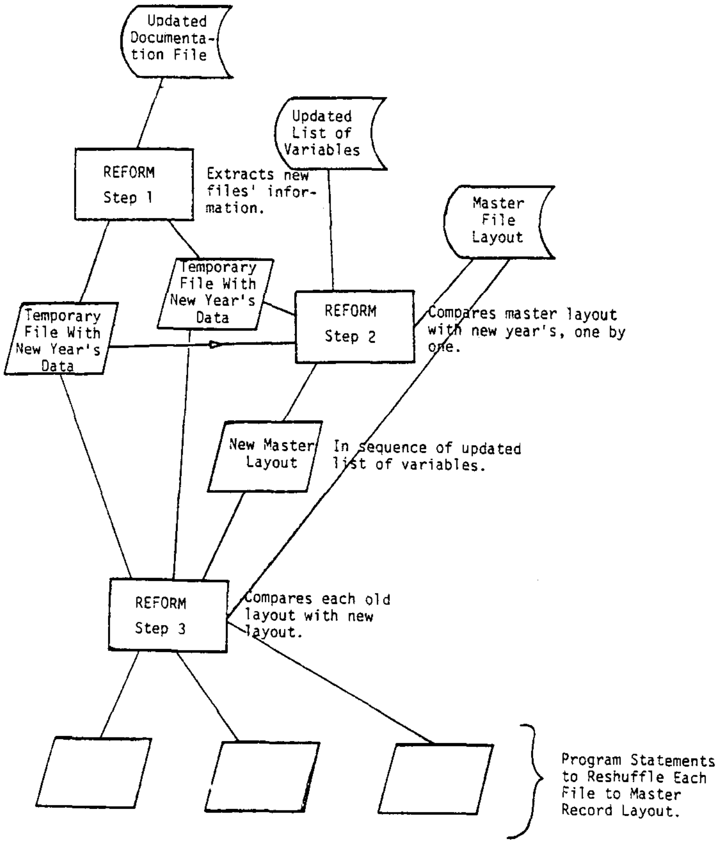
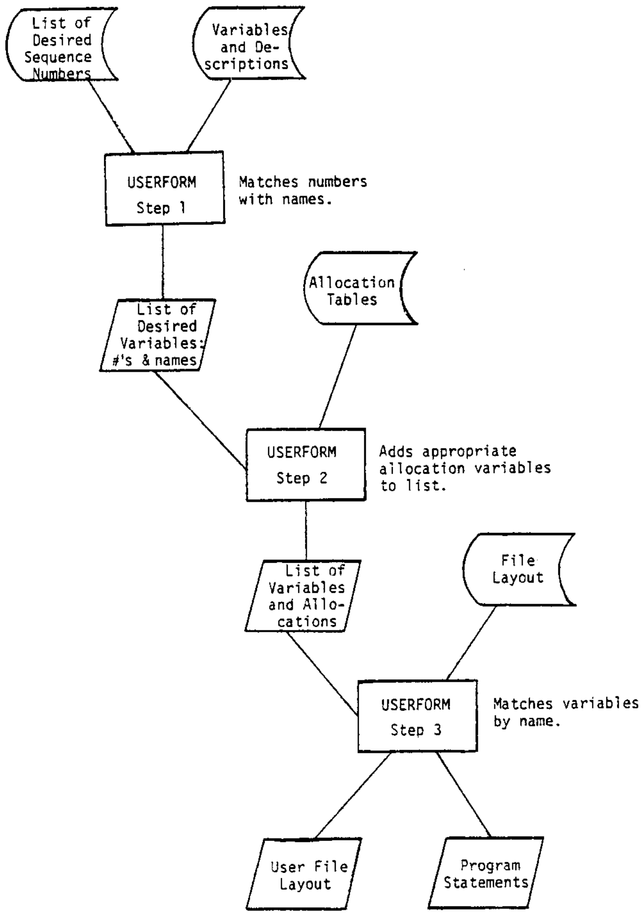

Updated Documentation File
REFORM Step 1
Extracts new files' information.
Updated List of Variables
Master File Layout
Temporary File With New Year's Data
Temporary File With New Year's Data
REFORM Step 2
Compares master layout with new year's, one by one.
New Master Layout
In sequence of updated list of variables.
REFORM Step 3
Compares each old layout with new layout.
Program Statements to Reshuffle Each File to Master Record Layout.

Figure 4. Procedure for User Requests


List of Desired Sequence Numbers
Variables and Descriptions
USERFORM Step 1
Matches numbers with names.
List of Desired Variables: #'s & names
Allocation Tables
USERFORM Step 2
Adds appropriate allocation variables to list.
List of Variables and Allocations
File Layout
USERFORM Step 3
Matches variables by name.
User File Layout
Program Statements

Because we can provide custom files, computer related costs including programmer time for all researchers are greatly reduced and small-scale academic users with limited computer facilities are able to take advantage of the survey's great potential. Prior to the development of the AHSP, virtually no longitudinal studies had been undertaken and outside HUD there even was very little cross-sectional work. At present, we have filled requests from over 30 users. Thus the AHSP has fulfilled a major part of its goal.

Clearly the AHS is one of the most valuable and underused data sources on social and economic issues in the country. We have developed a system to manage the data and make it publicly available which is straight forward and relationally oriented and which thus makes the data much easier for researchers to use. If anyone is interested we have handbooks and brochures from the AHSP, as well as a few copies of this paper.

Footnotes

[1] The authors gratefully acknowledge the help and support of the following: Duane McGough, Kathryn P. Nelson, and especially, Paul Burke, at the Office of Policy Development and Research, U.S. Department of Housing and Urban Development. Mr. Burke contributed mightily to the effort described here and much of this work was inspired by and builds on his earlier efforts. Jonathan Cole, Burton Singer, and Harold Watts at Columbia have also aided us. Juliette Levin, Marcelo Nacht, George Sweeting, Paul Sweeting, and John Susskind all participated in building our data base. The work that provided the basis for this paper was supported by a Co-operative Agreement (HA-5289-CA) with U.S. Department of Housing and Urban Development. The substance and findings of that work were dedicated to the public. The authors are solely responsible for the accuracy of the statements contained herein. Such interpretations do not necessarily reflect the views of the government.

[2] For more information on the Annual Housing Survey, anyone may request the Annual Housing Survey Project Handbook, available from Annual Housing Survey Project, 811 IAB, 420 West 118th Street, New York, NY 10027. There are several volumes of statistical tabulations published each year by HUD and the Census.

[3] For a description of the relational data model see C.J. Date, An Introduction to Data Base Systems, Reading, Massachusetts, 1977.

[4] We began with a list of variable names which had been assigned by John C. Sneed. These became another piece of raw material for the system. Paul Burke at HUD collaborated in this aspect of the work.

* * * * * * * * * * * * * *