MANAGING THE DATA FROM THE 1979 INCOME SURVEY DEVELOPMENT PROGRAM

Pat Doyle, Mathematica Policy Research
Constance F. Citro, National Academy of Sciences

During 1979 and 1980 the Department of Health and Human Services and the Bureau of the Census, with support from other federal government agencies including the Food and Nutrition Service, USDA, administered a panel study of households representative of the civilian noninstitutionalized population in the United States called the 1979 Income Survey Development Program (ISDP) Research Panel. The survey was designed as the final pretest for the Survey of Income and Program Participation (SIPP) which had been under development since 1975 and was fully implemented in late 1983. The 1979 panel study was extremely complex due to the efforts put forth to improve the measurement of income, net worth, and program participation and to increase the information available on behavior, attitudes, expenses and disposable income of the population.

The complexity of the 1979 ISDP survey design led to the production of public use files which are cumbersome to use thus making it difficult to access the newly available data for research. The subject of this paper is to describe a project conducted by Mathematica Policy Research (MPR) under contract to the Food and Nutrition Service, USDA, to solve the data access problems through the use of data base management system technology. The DBMS chosen for this work was RAMIS II™ developed and distributed by Mathematica Products Group. The system developed by MPR is referred to as the ISDP/RAMIS II system.

It is important to note that a number of problems that were confronted in designing the access system described in this paper have been resolved in the release of the public use ISDP files (in fact, data from the ISDP/RAMIS II system were the source of some of these improvements). Furthermore, some, but by no means all, of these access problems have been explicitly taken care of in the design of the SIPP. Consequently, designing an access system for the new survey should be easier than for the ISDP. It is also true that the best design for a SIPP access system is likely not to be the design chosen for the ISDP system.

In the subsequent section, an overview of the panel study with emphasis on the contents and problems of the data files is provided. The report concludes with an overview of the newly created system with a summary of the data problems solved in the course of this work. For detailed information on the contents and use of the ISDP system, the reader is referred to Doyle and Citro (1984).

Overview of the ISDP and Its Applications

Figure 1 gives a graphic representation of the key features of the ISDP design. Briefly, note that:

-- There were 6 waves of interviewing providing 12 to 15 months of data for each household.

-- Interviewing was staggered; one-third of the sample was interviewed each month, with, thus, a different 3-month reference period for each rotation group.

-- This pattern was regular, except that the third rotation group, for various reasons, was skipped over Wave 4.

-- Each wave asked a core set of items, including monthly income and employment, plus one-time supplemental items.

The SIPP design for the first panel is very similar, including skipping one wave for part of the sample.

The ISDP, by virtue of gathering detailed month-by-month data over a span of at least a year, offered the potential for exciting research that simply could not be carried out before. But, to make it possible for the researchers at MPR to realize that potential, we had to design an access system that would do the following:

-- Generate reports and analysis files from individual waves, undoubtedly the easiest way of using the data

-- Generate reports and analysis files linking data across waves

-- Let researchers apply different rules to identify households and families across waves for longitudinal analysis

-- Link supplemental data collected in one wave to core data in other waves

-- Make it possible to carry out sophisticated statistical as well as tabular analysis of the data

-- Make it possible to use the ISDP data with data from other sources, for example, 1980 census summary data.

All of these access requirements apply equally well to the SIPP.

Problems for Access Posed by the ISDP

Various design features of the ISDP posed more or less serious problems for developing an access system that would satisfy the requirements just listed. These are summarized below.

- Staggered Interviewing. The use of a staggered interviewing schedule results in a situation where data from more than one interview must be accessed to study a common calendar period for the entire sample (except where the user can make do with the single calendar month that is common to all rotation groups within a wave).

- Skipping Wave 4. The alteration of the interviewing schedule to have the third rotation group skip over the Wave 4 interview means that, although two-thirds of the

sample cases have a full 15 months of data (from the five regular waves if they did not attrite), the other third has only 12 months. Moreover, the third rotation group does not have responses to any of the topical supplemental items asked at Wave 4.

- o Different reference periods for wave-specific information. For any one interview, there is a potential mismatch between the wave-specific data and the monthly data, given that monthly data for the month of an interview were actually asked at the subsequent interview.

- o Identifier problems. The Census Bureau encountered problems in uniquely identifying individuals across the survey waves, necessitating creation of a new unique person identifier, called the link index, as a separate file from the interview data files. It also turned out that the Bureau erroneously included some persons on the cross-section interview files who were not in fact present and vice versa.

- o One-time wave-specific supplemental data. The fact that important data were asked on a supplemental one-time basis creates problems for using these items together with the monthly and quarterly data.

- o School lunch data problems. The ISDP files include valid data only for the last child in a family, and these data were erroneously written into the records for all other children.

- o Lack of editing on Wave 6. In the case of Waves 1-5, the Census Bureau performed edits on demographic variables and also edited income recipiency flags. No editing was performed on the Wave 6 data, which were collected in an entirely different format.

- o Asset income reporting experiment. This experiment creates practical problems of associating asset income data with other data for each month of the panel.

- o Incomplete determination of monthly unit composition. The design of the cross-section files, coupled with a high level of noise in the data on arrival and departure dates, made it very difficult to assemble a stream of monthly unit composition indicators consistent with reported monthly economic data.

- o Absence of longitudinal weights and imputations for missing data. The cross-section interview files contain weights and also imputations for missing income and employment data that were constructed strictly on a cross-section basis which are not suitable longitudinal studies.

- o Absence of longitudinal editing. With the exception of editing age and sex in the construction of the unique identifiers, no longitudinal edits were performed on the demographic variables.

These characteristics of the ISDP survey make retrieval of the information for analysis cumbersome and expensive. This is particularly true for longitudinal applications of the data such as the study of turnover in the Food Stamp Program.

The difficulty in using the ISDP for research was compounded by the structure of the available data files. At the time this project was carried out, the most suitable input file was a concatenation of cross-section files from all five waves. The format for each cross-section was similar to the public working files currently available (NTIS, 1982) except that the family level had not been fully developed. The records from all five waves were grouped by PSUSERIAL and a level 1 record was created which recorded information common to each group such as rotation.[1] In addition to inserting the level 1 record, the Bureau also merged the link index (constructed unique person identifier) and longitudinal edited values of age and sex to this file. However, the Bureau deleted from this file the results of the cross-sectional imputations for income and employment data. The rationale for this omission was the unsuitability of these imputations for longitudinal analysis, the purpose of the concatenated file.

This file was extremely cumbersome to access due to the lack of a true hierarchical structure, the large number of different record types (data from each topical module were recorded on a separate record with a distinct record length and layout) and the fact that some of the newly created person identifiers were erroneous.

Overview of the ISDP/RAMIS II System

The objective of this data base development effort, as noted above, was to take the information available on the series of cross-section files described above and array it in a manner that would facilitate longitudinal as well as cross-sectional analysis. The results of this effort were two RAMIS II data bases, one called SIPPMASTER and one called MH for monthly households. SIPPMASTER is the main file in that all of the data collected during each wave are stored there. This file is used for all cross-section applications as well as longitudinal applications which do not involve the formation of longitudinal households or other groupings of individuals. The MH file is the data base designed to support the construction of longitudinal units. It essentially provides information on monthly household, family, and food stamp unit composition. The data in MH are arrayed to permit a user to develop a definition of longitudinality and apply that in the construction of a longitudinal unit file. Once the longitudinal unit itself is determined, the user can employ the data stored in SIPPMASTER to derive variables like total household monthly income which reflect the longitudinal unit characteristics.

The remainder of this section provides an overview of the contents of the ISDP/RAMIS II system. A detailed discussion of the motivation for choosing this file design and the procedures

required to develop this data base is described in Doyle and Citro (1982).

SIPPMASTER. Figure 2 displays the logical organization of SIPPMASTER. It has a hierarchical structure with fifteen levels, five of which are real and ten of which are virtual.[2] The five real levels are wave, household, family, person and month. Some relevant comments on each of these levels follow:

Wave. (Level 1) Indicators for Waves 1 through 5 are contained in SIPPMASTER on level 1. The data from Wave 6 are treated as supplemental and are therefore stored in the virtual level PM (level 7). SIPPMASTER is physically separated into 5 data bases, one for each wave. They are linked together with RAMIS II USE commands to logically form one data base.

Household. (Level 2) This reflects household composition at the time of the interview. The household identifier (HHID) uniquely identifies households within wave. It cannot be used to identify households longitudinally. Non-interview households in each wave have entries at this level, however data for all other levels are zero. The contents of the household level consist of the data found in the household record in the cross-section files prepared by the Census Bureau.

Family. (Level 4) The family level simply identifies family units within households as they existed at the time of each interview. Primary individuals, secondary individuals, and outmovers are treated as one person families.

Person. (Level 5) This contains interview specific data for each individual and retrospective data that were not collected for specific calendar months such as total weeks unemployed. The identifier for level 5 is the link index (called PERID in RAMIS II) so that each person sampled is identified in the same way across all waves. The data for the person level were derived from record type 5 of the cross-section files. Some relevant points: outmovers in a given wave are included for that wave but have 0 in the weight fields; the weights are cross-sectional; all person identifiers with values exceeding 200000 should be deleted for longitudinal analysis but not for cross sectional analysis; corrected age (CORAGE) should be used instead of edited age (AGEED) except that corrected age is 0 on Wave 2: income recipiency flags on level 5 are not to be used to determine item nonresponse as they were retained here for other reasons (for example, if the interest flag in Wave 1 is 1 on level 5 but there is no entry for that income type in the WU or MU associated files, then the person was reported to have had an interest producing asset but did not actually receive interest income during the Wave 1 reference period).

Month. (Level 12) This represents the reference period for each wave. All months in the survey have been numbered longitudinally so that, for example, the 3 months pertaining to Wave 2 are 4, 5, and 6. Aside from identifying the longitudinal reference months, this level contains numerous fillers intended to support the construction of longitudinal household (or other aggregate unit) files.

The remaining data available through SIPPMASTER are stored in associated files which can be accessed directly if desired. A summary of the contents of each can be found in Doyle and Citro (1984).

MH. Figure 3 describes the logical organization of MH. It is a relatively simple hierarchical file with five real levels and one virtual level. This file reflects the outcome of a complicated procedure designed to determine monthly household and food stamp unit composition from the data collected in the 1979 ISDP Research Panel. Documentation on the methodology employed in the development of this file is included in (Doyle and Citro, 1984). The contents of this file are described below followed by a section summarizing how it is used to develop longitudinal units.

Unlike SIPPMASTER, MH contains a limited number of variables. It is comprised mostly of pointers detailing who lived with whom during each month covered by the first five waves of the survey. The remaining variables provide descriptive characteristics such as age and relationship to reference person which are necessary to effectively determine longitudinal units. Each of the levels of MH is described below.

PSUSERIAL. (Level 1) This level contains the scrambled values PSUSERIAL as well as the rotation group identifier. For the ISDP all persons who ever resided together have common values of PSUSERIAL, so this level was created to increase the efficiency of data retrieval and to minimize storage costs.

MONTH. (Level 2) This level simply identifies the month. Longitudinal reference months as described for SIPPMASTER were used. For rotation groups 1 and 2, the months range from 1 to 16 and for rotation group 3 they range from 1 to 13. Note that household composition can be described for one more month than is covered by the retrospective data collected in the ISDP. This extra time period is the month of the final interview.

Household. (Level 3) This level describes who lived with whom during each month and the Food Stamp Program participation and benefits of that group. The contents are the monthly household identifier and food stamp recipiency and amount variables for up to two food stamp units.

Family. (Level 4) This is an indication of family groupings within monthly households. The contents are family identifier, family type, and family kind.

Person. (Level 5) This level contains an entry for each person for every month he or she was present in the sample. The key to this level is PERID, the same identifier used in SIPPMASTER. The other variables stored in this level are age, relationship to reference person, marital status, food stamp unit membership and variables necessary to link to SIPPMASTER.

PD. (Level 6) This is a virtual level in MH. The associated file is called PD and it is the same PD file accessed through level 6 of SIPPMASTER. It contains presence in sample indicators as well as constant demographic data such as sex.

The intended use of the MH data base is to determine longitudinal units. In developing the ISDP/RAMIS II system, one objective was to allow researchers flexibility in the development of the definition of what constitutes the same unit when viewed over time. For some applications it may be appropriate to define a unit as being the same from one month to the next if all adults remain the same. For another application it may be sufficient to only require that the reference person (household head) be the same. More complicated definitions may be required in other situations. An example might be that units are the same if the composition changes from one month to the next are restricted to birth of a child, loss of a peripheral adult, e.g., an older daughter leaves for college, or a death of one spouse in a husband-wife primary family.

Each of these three definitions can be specified with the ISDP/RAMIS II system as can many others. The procedure is as follows. Using the preferred definition, an algorithm for uniquely identifying each unit each month is developed. In the second example above, this would simply involve assigning the PERID of the reference person to the monthly unit as the identifier. Next, a comparison across months within PSUSERIAL groups is made. All monthly units with common values of the newly created identifier constitute one longitudinal unit. Finally, an extract is created which records the available information organized by the longitudinal unit identifier.

The available data from MH are primarily demographic, the exception being Food Stamp Program characteristics. The user will of course also desire economic data to support the analysis of the longitudinal units. This can be achieved through the extraction of data from SIPPMASTER.

Conclusion

This paper describes a system to access data from a complicated longitudinal survey of households when the survey itself was in its development stages. It represents a successful attempt to apply modern DBMS technology to solve access problems posed by complex social science data collection efforts. Some of its features are:

- Good report generation for easy specification of tables and extracts
- Procedural language interface to allow the use of FORTRAN or PL/1 to conduct complex applications
- SAS interface to permit more sophisticated statistical analysis.

The system is, of course, not without drawbacks. For example, the hierarchy imposed in the primary file, SIPPMASTER, is cumbersome and, with recent developments in relational data base technology, unnecessary. This structure could easily be simplified today. Furthermore, the system is on-line and therefore require large amounts of disk storage. As the cost of mass storage goes down with improved hardware now being developed, this will become less of a problem.

In spite of these imperfections the ISDP/RAMIS II system works. It represents the first truly integrated ISDP data base available to the public for research. With this system users can and indeed have carried out analyses that truly exploit the longitudinal nature of the data.

FOOTNOTES

[1]On the publicly available data bases, PSUSERIAL is a nine character field which uniquely identifies all households in Wave 1. Together with person number it was originally intended to uniquely identify persons followed in the panel.

[2]A virtual level is a level for which the data are not physically stored in the file. Instead there is an internal record of the location of another file which contains the information. With a DBMS, this other (or associated) file is accessed automatically when data from it are requested.

REFERENCES

Carr, Timothy; Doyle, Pat; and Lubitz, Irene. "An Analysis of Turnover in the Food Stamp Program." Draft Report. Washington, D.C.: Mathematica Policy Research, 1983.

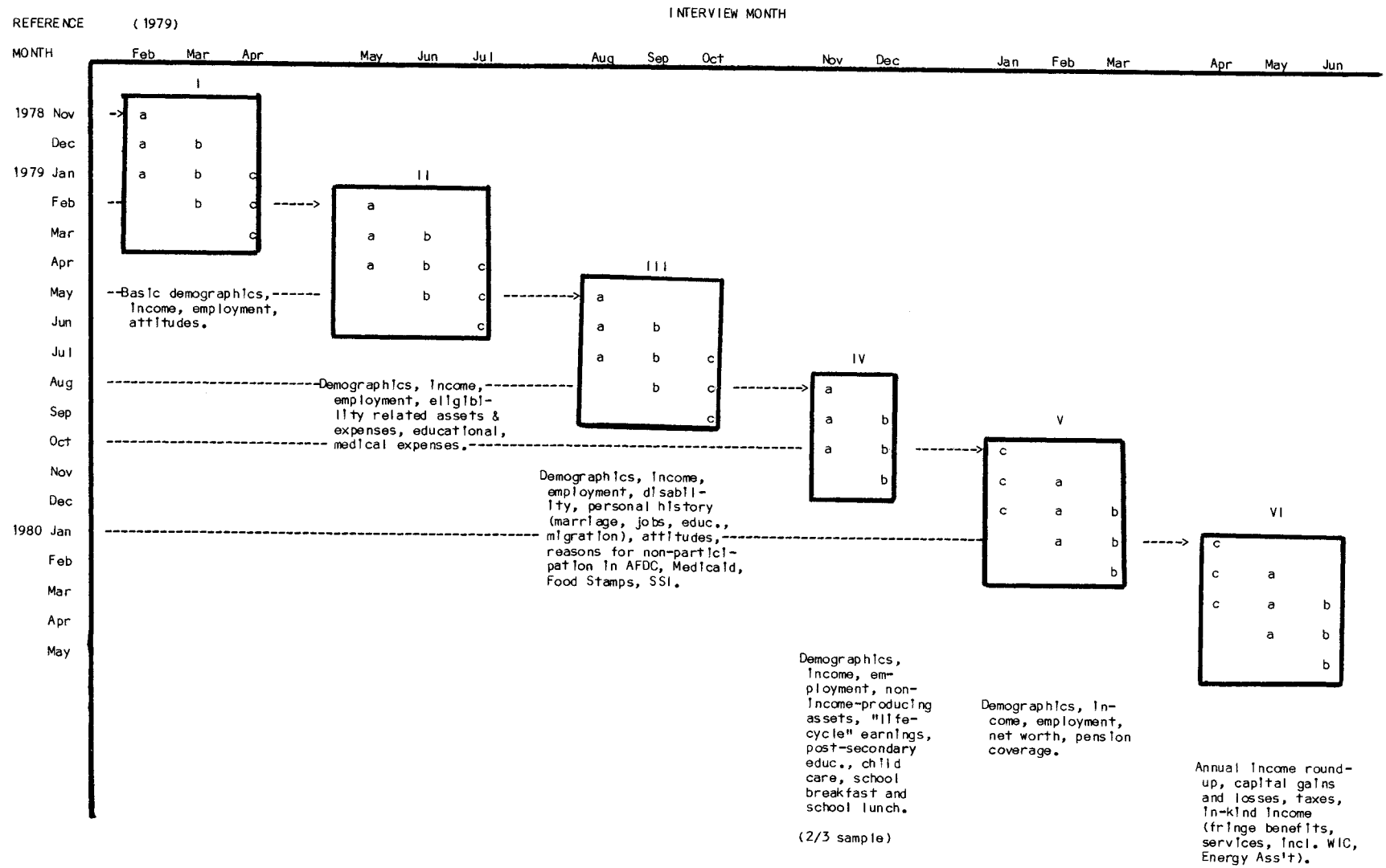
Doyle, Pat and Citro, Constance F. "The ISDP/RAMIS II System and Its Development." Draft Report. Washington, D.C.: Mathematica Policy Research, 1984.

Doyle, Pat and Citro, Constance F. "Proposed Design Strategy for Storing and Accessing Data from the 1979 Income Survey Development Program Research Panel." Draft Report. Washington, D.C.: Mathematica Policy Research, 1982.

National Technical Information Service. Income Survey Development Program: 1979 Research Panel Documentation. Springfield, V.A.: U.S. Department of Commerce, 1982.

FIGURE 1

Survey Waves -- 1979 ISDP Research Panel



"a" = households in first panel

"b" = households in second panel

"c" = households in third panel

FIGURE 2

FILE STRUCTURE FOR SIPPMASTER

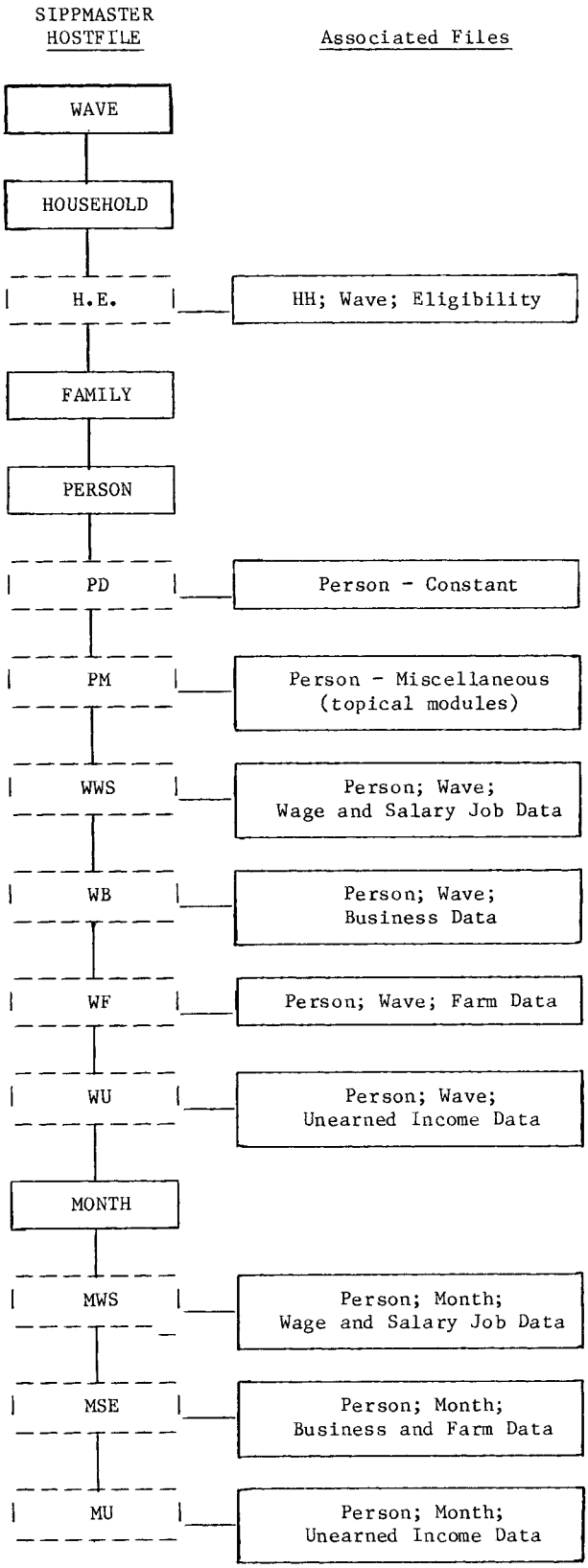


FIGURE 3

RAMIS II FILE REFLECTING MONTHLY UNIT COMPOSITION (MH)