DOCUMENT RESUME

ED 286 908 TM 870 540

AUTHOR Gates, William A.; Witt, Barbara
TITLE Creating a Database for Demographic Research: A Case Study.
SPONS AGENCY Wisconsin Univ., Madison. Center for Demography and Ecology.
REPORT NO CDE-WP-86-30
PUB DATE 86
GRANT HD-05876
NOTE 14p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Census Figures; Database Management Systems; *Databases; *Demography; Extended Family; *Family (Sociological Unit); *Family Characteristics; Family Relationship; Family Structure; Marital Status; National Surveys; *Residential Patterns; *Search Strategies; User Needs (Information)
IDENTIFIERS *Census 1980

ABSTRACT

The PUS801000 (Public Use Samples) database was created as a subsample of the Census Bureau's 1980 Public Use Microdata Sample (PUMS) for the purpose of meeting the needs of demographics research. PUMS needed to be reorganized along relational lines by identifying the variables most widely used by researchers and constructing proper relations containing these variables. PUMS contains 11 million records for individuals and 4 million records for households; in its original format, each household record is followed by the "person" records for that household. The PUS801000 database is a reduction of PUMS database to construct a 1/1,000 sample of U.S. households (94,201 households; 226,458 persons), in which each household is coded uniquely (by state/number) and each person in the household receives a meaningful number code. (All variables are integer format except for H=Household and P=Person.) Questions that could not be addressed by the basic relational operations of traditional projection and restriction, and the problem of record concatenation can be addressed by the QUEL programming language. Programming steps are described for creating married couples as a unit of analysis. This involves the complex process of identifying primary family couples and subfamily couples in households. A portion of the user directory table illustrates some of the variable fields selected and the storage saving achieved by coding "person" relations. (LPG)

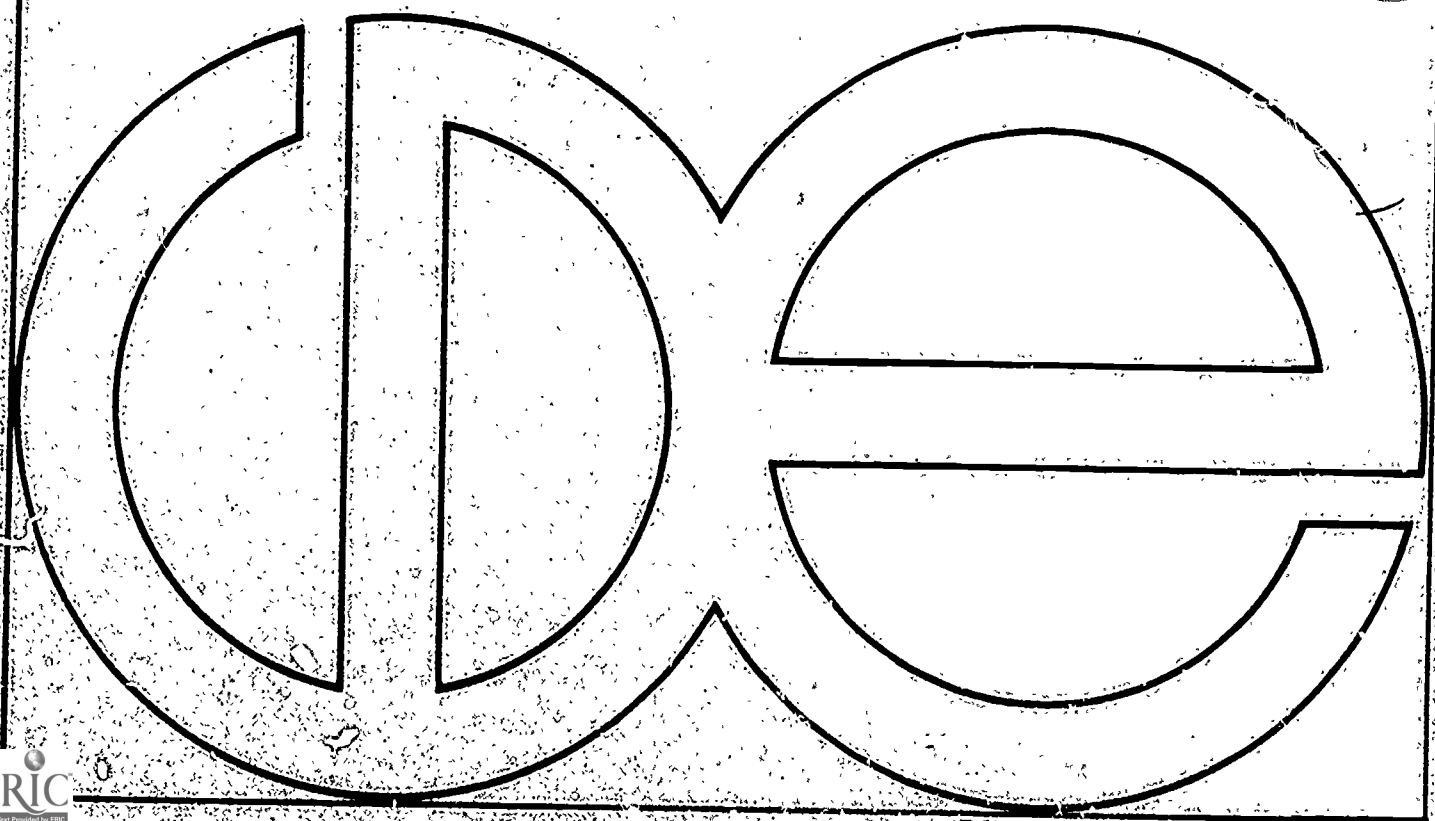
5-1-81

Center For Demography And Ecology
University of Wisconsin-Madison

CREATING A DATABASE FOR DEMOGRAPHIC RESEARCH: A CASE STUDY

William A. Gates
Barbara Witt

CDE Working Paper 86-30



CREATING A DATABASE FOR DEMOGRAPHIC RESEARCH: A CASE STUDY

William A. Gates
and
Barbara Witt
Center for Demography and Ecology
University of Wisconsin, Madison

This work was supported by the Center for Demography and Ecology (Grant No. HD-05876).

CREATING A DATABASE FOR DEMOGRAPHIC RESEARCH: A CASE STUDY

William A. Gates and Barbara Witt
Center for Demography and Ecology
University of Wisconsin-Madison

The Center for Demograhpy and Ecology at the University of Wisconsin-Madison acquired INGRES three years ago to begin exploring ways in which relational database concepts could prove useful in supporting research. We report here specifically on the development of the PUS801000 database derived from the Census Bureau's 1980 Public Use Microdata Sample. The original dataset consists of over 11 million records for individuals and 4 million records for households. It is produced by the U.S. Census Bureau from the 1980 Census of Population and Housing.

The experiment continues but many practical lessons have been learned. This case study documents a variety of strategies and tools for minimizing the resources required to support the PUS801000 database. Other issues addressed include how such databases can be extended to be of much greater use to researchers. Also discussed are the special needs of researchers and the problems they encounter in being direct users of INGRES QUEL.

Creating A Database for Demographic Research: A Case Study

What were once termed large files, sometimes modestly termed datasets, in the 1960s and 1970s, are now in the 1980s called databases. Does this really reflect basic change? In the best of circumstances the justification for this change in terminology seems to be a new cognizance of the importance of the relationships, organization and structure among and between the data elements themselves. Fortunately, dealing with such weighty questions as these is beyond the scope of this paper and instead we present our experience with transforming a "large file" into a more relational database.

This paper further describes our environment and presents the rationale that has led to the development of the PUS801000 database from the 1980 Public Use Microdata Sample (PUMS). This database, while still evolving, provides researchers with a readily available, reliable subsample from the 1980 PUMS. For some, the dataset provides a large enough sample for completion of analysis; for others, it is a place to begin, to test ideas without a great deal of energy and processing effort. Rarely is data in a form that can be used for analysis or browsed and explored. With a limited knowledge of relational database languages such as QUEL a researcher can create a set of observations specific to his needs which can then be exported to the statistical packages. Because the data is stored on disk, it is always available. There is no need to wait for tape queues or even acquire the knowledge necessary to process tape files. The data can be easily shared and re-creating datasets because of lack of knowledge about their existence can be decreased. Because of the static existence of the dataset, reproducing results can be achieved.

The Center for Demography and Ecology (CDE) has operated and managed its own computing facility since the early 1970s. This has made it possible for demographers at the UW-Madison to specialize and more effectively process any of the numerous machine-readable large files that have become commonplace in the last fifteen years. In the last three years CDE has been successful in greatly expanding its VAX 11/780 computing environment to include disk systems sufficient to contemplate shifting tape-oriented large file processing to disk-oriented processing. The greater access that this provides has more than any other factor increased the demand for computing cycles; the increase in productivity has been worth it.

We also acquired INGRES three years ago and began exploring ways in which relational database concepts could further improve support for demographic research. With support for concurrent or shared access, we hoped to be able to gain back some of the computing cycles by lessening the amount of redundant tape-processing performed by researchers using the same major large files.

The Public Use Census files, as they may be called generically, are typical of the large files that see repeated and frequent processing at CDE. They now span the decades from 1940 to 1980 and as released after each U.S. Census represent the most widely analyzed dataset ever produced. Although we have dealt with these and many other large files we feel that our experience with the 1980 Public Use Microdata Sample (PUMS) typifies much of what can be anticipated in way of problems and decisions while pointing to certain realities that will persist for some time, even as frontiers move on.

The Large File

The 1980 Public Use Microdata Sample is a stratified 5% sample of housing units and the persons in those housing units as enumerated in the April 1980 U.S. Census and contains over 16 million records. This data is available from the Census Bureau in a 5% sample stored in 51 files (one file for each state and the District of Columbia). Given that each record has been written as 193 bytes we have a file of approximately 3.2 gigabytes. The 193 bytes on two basic record types, HOUSEHOLD and PERSON, contain nearly all the detail of census long form questionnaires. Let us ignore for the time being the conceptual problem of how households map to housing units. For each household, there is a HOUSEHOLD record containing geographic and housing items followed by a PERSON record for each person found in the household. The PERSON records contain information specific to an individual such as age, relationship to householder, sex and race. Re-read the last sentence. It should reveal one very fundamental problem; being hierarchical does not assure even minimal relational tests can be met.

The Conceptual Beginnings

It is implicit that the Public Use Samples need to be reorganized along relational lines, accepting some compromises toward the end of being able to process the data with greater efficiency while gaining a great deal in the ability to simplify the process of attaining the desired unit of analysis. We had the simplifying advantage of a static database in the sense that once created no additional data would be added to basic tables. We recognized from the beginning, given years of experience with traditional approaches, that not all of the data in the PUMS is relevant to even the broadest research questions in the social sciences, let alone demography. We also realized that CDE is quite successful at taking the more traditional approach and utilizing programmers and research assistants to deal with the logical difficulties of PUMS.

Therefore we took a very pragmatic approach to demonstrate the advantages of relational database concepts without consuming existing resources or displacing research at hand. Thus our approach has been to initially identify the most widely used attributes (variables) and construct proper relations containing them. We have given attention to being able to expand relations as necessary but have not as yet found it to be an urgent requirement. We also reduced the size of the problem by sampling the original PUMS

to construct a 1/1000 sample of households. Once the database has been properly conceived, a test we think can only truly be met by use and experience, the full sample becomes approachable, especially as storage technologies continue to evolve.

One additional aspiration is to gradually augment the basic tables to essentially create a statistical abstract. In a very modest sense this has begun with our extending the database to include documentation of basic tables traditionally found in code books or some data dictionaries. Again from a pragmatic viewpoint this is unlikely to occur until there is consensus and thus sufficient organizational motivation to move ahead.

The Traditional Problem

In the next section of this paper we will provide an example to motivate what demographers and other social scientists want to be able to do with these data. In all but the simplest of analysis, the basic organization of the data elements and data records present difficult logical problems. Of course first steps to make the "large file" more tractable actually parallel the basic relational operations of projection and restriction. This step is frequently preceded by one that makes the large file even larger or, i.e., take 3.2 gigabytes and make it into a 12 gigabyte file by appending the HOUSEHOLD record to every PERSON record. If the projection and restriction are performed first the "largeness" is not quite so bad and the computing budget may survive.

The selection of attributes (projection) and the selection of cases (restriction) does not carry one very far toward the end of analysis. The concatenation of record types furthermore does not allow the researcher to address the substantive questions of interest. One must now resort to "programmed" solutions to create the proper unit of analysis.

Unit of Analysis Dilemma: The Couples Example

One use of Public Use Samples is for the study of relationships between persons in a household. Consider the problem of identifying and creating as a unit of analysis married couples from the 1980 Public Use Microdata Sample. Beginning with the data as already structured couples seem to be easy to identify. PERSON records are sorted on the relationship to the household head (RELAT1) code within a household. The head of the household has a code of 00 for that variable so his record is always first. His spouse has a code of 01 for the relationship variable and the record would be sequenced directly after the household head's record. This couple is called the primary family couple of the household. The problems start when one has to identify the other couples in the household. A household head may have his parents living with him. His spouse's parents may be present too. If we continue with this happy family, we might find the head has some married children and their spouses also living in the same household. These couples, where at least one person is related to the head of the household, are called subfamilies. Each subfamily in a household is identified by using two fields in

the data. The codes for SUBFAM1 identify the subfamily relationship with a code of 1 identifying a husband-wife subfamily. SUBFAM2 is the subfamily number. Since it was just illustrated that there may be more than one subfamily present in the household, each subfamily is assigned a number. There is a possibility of four subfamilies in each household.

Compounding the problem, one cannot assume that person records for subfamily members are positioned one after the other within the household. For example, a daughter of the household head is assigned a value of 02 (child of head) for the code relationship to head. If she is married and her husband is also present in the household, his value for the same code would be 05, the value given for an other relative of head. Recall again that the value for this code (RELAT1) establishes the sorting sequence for person records within a household. If other children of the head were also present, or if the head had a brother or sister present (code 03 on RELAT1), or a parent present (code 04), the person records for this couple would not be located next to each other on the file. Finding these couples is a difficult task compounded by the necessity of traditional reliance on the data being in "order". The simple concepts of projection, restriction or record concatenation do not address the problem adequately.

Within a relational database, primary family couples as well as subfamily couples can be easily identified using QUEL. First a relation is made with a universe of all married persons. This table contains all the variables needed to identify couples. These are sex, used to identify husband or wife, relationship to head, used to identify a primary family couple, and the two subfamily variables which are used to identify subfamily couples. In the 1980 PUS, a code of 0 on marital status identifies married spouse present adults as well as married spouse absent (from the household) adults. Therefore, not all the persons in the married relation will have a spouse present.

Next, a RETRIEVE INTO statement is used to create the couples relation. All married males (sex = 0) are put into this relation along with a variable which identifies the wife's person number (initialized to 0).

```
retrieve into cpls(married.all,wpersonum=int1(0)) where married.sex = 0
```

Two replace statements are then executed, one for primary family couples and one for subfamily couples. For primary families, a code of 0 represents the head of the household and a code of 1 represents his spouse. A code of 1 on sex identifies females.

```
range of m is married
replace c(wpersonum=m.personum) where (m.stserial = c.stserial) and
(c.relat1 < 2) and (m.relat1 < 2) and m.sex = 1
```

(stserial identifies a household)

For subfamily couples a code of 1 on SUBFAM1 identifies a husband-wife subfamily. SUBFAM2 identifies the subfamily number. Since there can be more than one subfamily in a household, the subfamily number for the husband and wife must be checked to make sure they are equal.

```
replace c(wpersonum=m.personum) where (m.stserial=c.stserial) and
c.subfam1=1 and m.subfam1 = 1 and m.sex = 1 and
c.subfam2 = m.subfam2
```

One now has obtained the person number within a household for husband-wife pairs. These two numbers as well as the stserial code can be stored alone in a relation and used to retrieve characteristics of the married spouse present couples found on the PUS801000.

The PUS801000 Database

The data for the PUS801000 database was produced using a subsample number contained in each HOUSEHOLD record on the PUMS sample which preserves the stratification. We used this field to create a one-in-one-thousand sample (2% of the original dataset). Thus, information for 94201 households and 226458 persons is contained in the PUS801000 database arrayed in various tables as appropriate to relational representation.

In order to document the database, four relations have been added. The DIRECTORY table provides information about each of the data relations. Each row contains a name of a relation, the number of rows it contains and a description of the universe and attributes it contains. This table was designed to fit on a monitor screen so a user can obtain an overview of the contents and setup of the database. All of the attribute names used in PUS801000 are the same as those found in the Technical Documentation published by the Census Bureau. The VARTABLE relation can tell a user if the database contains a variable he wishes to look at and the relation in which it is stored. Since the variable names are limited to eight characters, the VARDESCR table contains a text description of each variable name. This was also taken directly from the Census Bureau documentation. The VARCODE relation provides the user with the numeric codes for each variable and a description of each code. Retrievals from this relation can provide a user with a codebook for a new dataset he creates. One attribute yet to be added to this table is the frequency distribution of each variable within the database.

The organization of the database can best be understood by examining the directory relation, one of the documentation tables in the database. This relation lists the tables in the database, the number of rows for each, and a brief description of it.

directory table

tabname	nsize	descript
per80all	226458	all persons info about age sex race educ rela to hd pov st
relat280	5785	var RELAT2 for anyone coded 5 (oth rel) on RELAT1
subfam80	3472	variables SUBFAM1, SUBFAM2 for anyone in a subfamily
income280	8214	contains anyone with a code other than 0 on INCOME2
income380	2941	contains anyone with a code other than 0 on INCOME3
income480	43678	contains anyone with a code other than 0 on INCOME4
income580	28249	contains anyone with a code other than 0 on INCOME5
income680	7500	contains anyone with a code other than 0 on INCOME6
income780	21895	contains anyone with a code other than 0 on INCOME7
spanish80	14704	anyone coded 1-4(Mex, P.Rican, Cuban, Other) on SPANISH
veteran80	30008	contains VETERAN1 to VETERAN8 for anyone coded 0 on VETERAN
adult80	174726	persons 15+, marriage, fertility, labor force variable
migrsam80	113470	Migration/Place of Work/Travel Time vars for sample persons
citimig80	14995	Citizenship and Yr of Immig. for anyone not born in U.S.
house80	94201	hsehld record info on area,hsehold type and income

Information on thirteen variables from the household record is stored in the HOUSE80 relation. One can obtain the location (state,SMSA, division) of the household plus information on the type of household, the income of the household and family size. The other 14 relations contain information from the person data. PER80ALL is the largest relation in the database and contains a row for each person in the 1/1000 sample. Basic information about each person such as sex, race, age and education is stored here. These are data items that were answered by everyone in the Census questionnaire.

In the other 13 person relations, the universe is limited to those who have a non-zero code recorded in the data field. The value 0 is most often used in the data as the non-applicable code. Since the Census data is very detailed, many questions are asked that pertain to a limited population. Good examples of this are the income questions. There are eight income fields in the 1980 data. For each of them a code of 0 means either the person is under 16 or there is no income from the source mentioned. Two of the eight fields, INCOME1 (wage or salary income in 1979) and INCOME8 (total income from all sources in 1979) have enough non-zero codes that they have been stored with several other fields in the ADULT80 relation. The other six income fields are stored in separate relations. The directory table shows the size of these tables. The smallest, INCOME380 (farm self-employment income) contains 2941 rows—approximately 1% of the 1/1000 population. The largest, INCOME480 (interest, dividend or net rental income in 1979) contains 43678 rows which is still only 20% of the total. All of the income fields are stored in 4-byte integer (i4) fields. By carrying each of these in a separate relation, the storage saving is substantial.

More storage saving is realized in the other person relations. The ADULT80 relation contains information only asked of persons 15 years or older. Information on marriage, fertility and work histories is stored here. The RELAT280 relation is a detailed breakdown of relationship to head for anyone coded as an other relative of head on the main relationship to head variable. This is only 2% of the population but is still valuable when determining relationships among persons in a household. The same is true of the SUBFAM80 relation. This relation contains variables which identify anyone in a subfamily, a second husband-wife or parent-child family in the household. The SPANISH80 relation identifies the Spanish population in the census and could be used to retrieve person information about this population. The VETERAN80 relation details military service history for anyone who has served in the armed forces. While tabulating the Census questionnaires, a sample was selected for questions concerning migration history, place of work history and travel time to work. This sample comprises 65% of the adult population and variables are stored in the MIGRSAM80 relation. The last person relation listed is the CITIMIG80 relation. This relation stores information on the citizenship status and year of immigration for anyone not born in the U.S.

Methodolgy: Decisions and Problems

Many decisions needed to be made since this was a first attempt at putting Census data into relational form. Should all of the 143 data items available be stored in the database? Is there a unique key already present in the data records that can be used to prevent confusion? How can the person data be linked to its household data? Should the data be stored in its original character format or changed to integer format? What structure should be used for the tables? Answers to these questions are discussed below.

The PUS801000 database contains a subset of data items from both the household and person records. Through many years of working with Census data, it has been observed that many data items available are not used in demographic analysis. Extracts are usually made from the original dataset containing only the variables to be studied. The variables selected for the PUS801000 are based on the extracts that have been made in the past. Thirteen of the sixty-five data items from the household record were chosen. Examples of items selected from the household data are SMSA (Standard Metropolitan Statistical Area) code, type of group quarters, whether the unit is owner or renter occupied and household and family incomes. Data items not selected include number of bathrooms, whether the structure is air-conditioned, trucks and vans available and amount of second mortgage. Fifty-three of the seventy-six data items from the person record were selected. Demographic interest seems to be in person-related characteristics. These include such variables as age, sex, race, work history and fertility. Examples of variables not selected are ancestry-second entry, activity in 1975, attending college, and carpooling.

One other major category of variables not selected are the allocation fields. For each data item on the file there is a corresponding allocation field to relay whether the code is allocated or actually reported by the respondent. The Census data contains no missing data or blank fields because of the allocation process. These fields may become important when analyzing small subpopulations within the data.

It is possible to add any missing data items to the database. This can be accomplished by bringing in a single variable into a separate relation. Timing estimates have been run on this process and it has been decided that it is easier to add these variables as researchers require them than to use the diskspace required to store every data item available.

The data in its original form appears as a household record followed by the person records for that household. There is no other link between a household and its persons other than where they physically appear in the file. If the data was inserted into a database record for record as it originally exists there would be no way to link together a household with the persons in it or a person in a household with another person in the same household. An identifier needs to be added to each record before it can be used in a database. Each household record contains a variable called SERIAL NUMBER which is a unique number within a state. The state code is also found on each household record. By using a combination of these two variables uniqueness for each household record can be established. The eight-digit variable STSERIAL (two-digit state code + six-digit serial number) was created for this purpose. In order to link a household to the persons in it, the variable STSERIAL was also added to each person record. To give each person uniqueness within a household the variable PERSONUM (person number) has been created. The first person in the household is assigned a value of 01, the second person is assigned 02...with the last person assigned a value equal to the number of persons in the household.

The database PUS801000 is stored entirely in integer format. The PUMS 80 sample is a completely numeric character dataset except for one field identifying the record type (H=household, P=person) which is not needed in the database. Because all two character fields can be stored in a one byte integer field, and three and four character fields in two byte integer fields, the storage savings amounts to 35% in this case for integer format over character format. Integer format also makes working with INGRES QUEL much easier. Data selections can be made without worrying about the field length of attributes. New variables can easily be created within the bounds of integer arithmetic.

The relations are stored in ISAM structure. The STSERIAL code is the primary key with PERSONUM as the secondary key. Optimization of all the attributes is planned. This will most likely be done using a smaller representative sample of a relation and then changing the table names in the zopt relations to the full sample relation names.

Summary and Observations

This paper principally described how the PUS801000 database was developed from the 1980 Public Use Microdata Sample. As a process this development experience was intended to provide a model that could be applied to many other datasets like it of interest to demographers and other social scientists. As a process, it was intended to remain flexible to allow the database to continue to evolve and grow as more researchers with more diverse interests came to appreciate its utility.

The appeal of databases such as the PUS801000 is that of accessibility, ease of retrieval and data manipulation, and saving and sharing of resources. Nonetheless, creating and/or maintaining or extending the database requires an effort and interest beyond that of most researchers. It was observed that the basic ideas of projection and restriction have been well understood by social scientists for quite some time. They have been used with more traditional sequential files at least since the 1960s. What has been lacking, short of resorting to programming, are reasonable tools for the researcher to use directly in dealing with the complex structure of the datasets. Relational query languages such as QUEL provide reasonable tools for dealing with this complexity and in creating the correct unit of analysis. Albeit, anything can be improved.

Mailing Address:

Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive
Madison, Wisconsin 53706-1393
U.S.A.