END-USE LOAD AND CONSUMER ASSESSMENT PROGRAM

Residential Characteristics Data Base

. . . .

User-accessible SAS Residential Datasets

(USRD)

MAIN GUIDE

Version 2.2

June 1988

USRD Review and Development Group

END-USE LOAD AND CONSUMER ASSESSMENT PROGRAM

Residential Characteristics Data Base

. . . .

User-accessible SAS Residential Datasets

(USRD)

MAIN GUIDE

Version 2.2

June 1988

USRD Review and Development Group

CONTENTS

FIGURES

Acknowledgements

This document is a product of the Review and Development Group for the User-accessible SAS Residential Datasets: a completely unofficial, more-or-less group of analysts from Bonneville and the Portland office of the Battelle Pacific Northwest Laboratory. With some exceptions, the group has met regularly on Friday mornings beginning in August 1987 to discuss direction, plan strategy and make the necessary collective decisions. The "group" originated as meetings of William (Bill) Monroe (PNL) and Phillip Windell (BPA). As his participation in the project grew, Steve Paquin (PNL) joined the Friday meetings on a regular basis beginning in October 1987. Early in 1988, Roger Moore (BPA) joined the regular Friday morning RDG meetings.

A number of PNL and BPA staff have reviewed previous versions of this Guide and asked insightful questions, provided constructive comments and made useful suggestions. The members of the RDG are most grateful for the atmosphere of honesty and dedication to quality which has pervaded these discussions and reviews. This feedback is appreciated and we hope it will continue.

The USRD is intended to be "alive" in the sense that it evolves and grows over time. Similarly, it is anticipated that this Main Guide will be a "living document" in the sense that we expect it to change and grow as time passes. For example, the issues of "time windows" and "time stamping" are clearly not yet resolved with Version 2.1, and won't be without additional use of the data to address real-world problems.

ACCESSING THE USRD

This Chapter contains a brief outline of the steps necessary to access the User-accessable SAS Residential Datasets (USRD). Throughout this Section there are examples of what the user will see and what the user should enter from the keyboard. To keep things clear, the computer prompts will be denoted with single quotes (' '); the user responses will be denoted by the use of double quotes (" "). It is always assumed that the user terminates commands with a carriage return.

Access to ELCAP1

In order to access the USRD you must first have access to the ELCAP1 microVAX computer. If you already have an account on that system and are familiar with the login procedure you may proceed directly to the next Section, "Access to the USRD."

To obtain an account on ELCAP1 you must contact the ELCAP1 System Manager. PNL staff and others who already have an account on another PNL VAX may simply send a mail message to ELCAP1::SYSTEM requesting an account. Please be sure to ionclude your phone number in the message. Others are advised to first contact the Chief, End Use Analysis Section, Bonneville Power Administration (Mail Stop RPEE) for permission to access the ELCAP Computational Network.

The ELCAP1 System Manager will give you a username, password and project name necessary to gain access to ELCAP1. Connecting to the system is dependent on your location and your equipment. At the time you get your account information, you will be instructed on how to make the connection.

Once you have made the physical connection to ELCAP1, you will first be prompted for a 'Username: '. You then enter the username given to you by the System Manager. The system will then prompt you for a 'Password: '. You then respond by entering your assigned password. If all goes well, you will receive a '$' prompt indicating a successful login. If you make a mistake entering your username or password, you may have to repeat the above steps. If, after three tries, you are not successful, call the System Manager and describe the problem.

Following your successful login,I the next step is to move to the "project" within which you will be working. In some cases, the "project" may be equivalent to your personal account, in which case you need do nothing more. To move to a project, you will need to enter the following from your keyboard:

"SET PROJECT projectname"

where "projectname" is the name of the project under which you will be working and through which you shall have access to the USRD. You will then have access to the USRD.

Following is an example of a successful login sequence. Comments are enclosed in parentheses.

```
Username: XXXXXX        (where XXXXX is your username)
Password:               (your password will not be echoed)
$ SET PROJECT myproj    (where myproj is the project you work with
                          which has access privileges to the USRD)
```

When you are finished working on the computer and wish To log out, simply enter "LOGOUT" at the system prompt ($). The system will inform you that you have logged out and disconnect your terminal from the ELCAP1 computer. If you are using a modem, it will also hang up the phone, so there is no need force a hang up with your communications package.

Privilege to Access the USRD

If you opened a new account specifically to work with the USRD by following the instructions in the above section, you already have access to the USRD and should move directly to the next Section, "Using SAS." This Section is directed at those who already have an account on the ELCAP1 computer and simply wish to obtain privileges to access the USRD.

To obtain the required privileges you must notify the ELCAP1 System Manager. The best way to do this is by using the VMS mail utility. If you know how to use Mail, just send a message to ELCAP1::SYSTEM asking for access to the USRD. If you are not familiar with the Mail utility, the following paragraph provides sufficient information to enable you to send the appropriate message.

At the '$' prompt type "MAIL". You will then get the 'MAIL>' prompt. Type 'SEND' to send a message. The next prompt you will see is 'To:' to which you respond by typing "SYSTEM". The next prompt will be 'Subject:'. You can type anything here that relates the fact that you would like access to the USRD. Next you are allowed to type the text of your message on the screen. When you are done with your message type "^Z" (hold down the control key and enter "Z" at the same time). At that point, your message will be sent and you will get the 'MAIL' prompt again. This time type "EXIT" and you will return to the '$' prompt.

Following is an example of a VMS mail message as it will appear on your screen. As in the previous example, comments are included and enclosed parentheses.

```
$ MAIL
MAIL> SEND
To: SYSTEM
Subject: Give me USRD access
Please give me USRD access.^Z  (text of message - end with ^Z)
MAIL> EXIT
$                              (back to the original prompt)
```

Using SAS in the VMS Environment

If you are familiar with running SAS in a VMS environment you may wish to skip to the following Section , "Accessing the USRD." Those who are unfamiliar with SAS should consult "SAS User's Guide: Basics," SAS Institute Inc., Cary, NC. What follows here is a very brief introduction to the use of SAS in the VMS computing environment.

To use SAS on the ELCAP1 microVAX you simply enter "SAS" at the '$' prompt. This will put you into interactive SAS which allows you to enter any series of SAS statements that you wish. To execute a series of statements use the SAS command "RUN;" as the last statement in the series.

To use SAS in batch mode, you will first need to know how to use a text editor. If you do not know how to use any of the VMS text editors, consult with the System Manager to obtain documentation.

The first step in executing a SAS batch job is to create a file with the SAS commands you wish to execute. It is conventional to name the file with a ".SAS" extension although it is not required. Next you need to create a file with the following two lines in it (including the "$"):

```
$ SET DEFAULT sas_program_location
$ SAS sas_program
```

where "sas_program_location" is the VMS pathname to where your SAS program resides. If you do not know, use the "SHOW DEFAULT" command at the VMS prompt when you are in the directory where you created the SAS file. The results of this command can be copied exactly and put in place of "sas_program_location" in the above file. "Sas_program" is the name of the file where your SAS source code is located.

When these two files are created you use the command "SUBMIT/NOPRINT submit_program_name" at the VMS prompt, where "submit_program_name" is the name of the two line file created above. The results of the batch job will be contained in two files. They will have the same first name as your SAS program file with the extensions ".LOG" and ".LIS." The .LOG file contains the diagnostics from the program run. The .LIS file contains the results of the SAS program run.

Accessing Data in the USRD

Once permission has been granted, access to the USRD is simply a matter of incoporating the proper control statements in your SAS program. The following three statements must precede any "DATA" or "PROC" steps in your program:

```
LIBNAME libref1 'dataref';      /* data library */
LIBNAME libref2 'USRDFMT$';     /* format library */
LIBSEARCH libref2;              /* use the format library */
```

where "libref1" and "libref2" are names of your choice (8 characters or less) and "dataref" is either CURC$ or PCURC$. Using CURC$ gives you access to the CURC data and PCURC$ gives you access to the Pre-CURC. If you do not wish to use the format library, you can replace the last two statements with:

```
OPTIONS NOFMTERR;
```

As an example, suppose you wanted to see the contents of the CURC data set. The following program would do that:

```
LIBNAME INDATA 'CURC$';          /* reference the CURC data */
LIBNAME FMTLIB 'USRDFMT$';       /* reference the formats */
LIBSEARCH FMTLIB;                /* use the formats */
PROC CONTENTS DATA=INDATA.CURC; /* list the contents of CURC */
RUN;  /* execute the program - not necessary in batch */
```

Now suppose you wanted to get frequencies for all of the ELCAP87 survey variables in the Pre-CURC. The following program would accomplish this:

```
LIBNAME INDATA 'PCURC$';         /* reference the PCURC data */
LIBNAME FMTLIB 'USRDFMT$';       /* reference the formats */
LIBSEARCH FMTLIB;                /* use the formats */
PROC FREQ DATA=INDATA.ELCAP87; /* run frequencies for ELCAP87 */
/* no table statement so we get frequencies for all variables */
RUN;  /* execute the program - not necessary in batch */
```

To get a list of the available data sets in any library, use the DIRECTORY command at the VMS prompt ($). For example, to get a listing of the available Pre-CURC data sets use the following command:

```
$ DIRECTORY PCURC$
```

The data sets available within the USRD are documented more fully in the VMS on-line HELP facility and in hard copy form as "ELCAP RCDB: USRD - Complete Set of Residential Characteristics."

GLOSSARY

For the convenience of the reader, we present here the definitions of a small number of terms which are used in a somewhat idiosyncratic, but relatively precise manner in this document. The section also includes the full interpretation of all of the acronyms used in the document.

Terms

Characteristics Data: that portion of the ELCAP residential data which includes administrative records, information concerning the structural properties of the participating residences, physical and demographic properties of the occupants, as well as occupant reports of their behavior and opinions obtained through personal or telephone interview or mail surveys. [Compare with "Engineering Data," below.]

Data Element: a piece of information describing a characteristic or some portion of a characteristic of a specific site at a given point in time stored in a manner which corresponds as closely as possible to the form in which the data were originally collected; the product of the administration of a data collection instrument to a specific site. The components of the Pre-CURC are referred to as "Data Elements." [Compare with "Variable," below.]

Engineering Data: That portion of the LELCAP residential data which includes information collected at the logger board, including the amount of electricity consumed, by end-use, indoor temperature, and, in some cases information regarding the outdoor environment -- temperature, humidity, and wind. [Compare with "Characteristics Data," above.]

ELCAP Site: An electricity account serving a household occupying a structure a specific geographic location, primarily for residential purposes; a solid state monitoring system is attached to the electric service panel of participating residences.

Transformations: a series of arithmetic or logical operations which are carried out on one or more data elements in the USRD producing a new data element. As examples, both of the following are legitimate transformations:

1) $c = a + b$

2) if $a = x$ and $b > y$ then $c = z$

where "a" and "b" are members of the USRD so that "c," as a product of "a" and "b" is also a member of the USRD; and "x" and "y" are constants.

Variable: a collection of data elements which purport to measure the same underlying concept at various points in time. The CURC is composed of variables and sites. Thus, in most cases, several Pre-CURC data elements comprise a single CURC variable. [Compare with "Data Element," above.]

Windows: a portion of the ELCAP data, defined in terms of the sites and the beginning and ending dates; generally, greater emphasis is placed on the time dimension.

Acronyms

CSRC = Complete Set of Residential Characteristics

CSRC = ORSD + RSII + RCTC

CURC = Used Residential Characteristics

ELCAP = End-use Load and Consumer Assessment Program

ELCAP/R = ELCAP/Residential

ELCAP86 = 1986 survey of ELCAP residential participants

ELCAP87 = 1987 survey of ELCAP residential participants

Meta-CURC = Data file containing indicators of the availability of characteristics and engineering data for each ELCAP residential site

ORSD = Original Residential Survey Datasets

PNWRES-83 = 1983 Pacific Northwest Residential Energy Survey

Pre-CURC = An intermediate stage of the RCDB which is more comprehensive than the CURC, but which lacks the time-stamping and associated assumptions of the CURC

RCTC = Residential Characteristics Transformation Collection

RI = 1985-6 Residential Inspections

ROS85/M = 1985 Residential Occupant Survey - Mail

ROS85/T = 1985 Residential Occupant Survey - Telephone

RSII = Residential Site Identification Information

USRD = User-accessible SAS Residential Datasets

META-CURC

Introduction

It is unlikely that analysts will ever attempt to use the entire set of ELCAP data in a single analysis. Rather, the "Typical Analyst" (TA) will seek to select some subset of the available sites and some "window" or "period" of time which is most appropriate given the objectives of the study in which they are engaged. This selection process frequently involves a complex series of interactions between the analyst and the database. To date, the selection process has, required multiple iterations between the engineering data and the characteristics data.

The "Meta-CURC" is designed to facilitate the data selection process by placing in one location indicators of the availability of the various types of data. These indicators range from simple "present/not present" bivariates to relatively intricate patterns intended to convey several pieces of information.

Structure

The Meta-CURC database is best conceptualized as a simple flat file with fixed and equal length records. Functionally, the Meta-CURC operates as a SAS system file and is stored together with the other members of the USRD.

There is one and only one Meta-CURC record for each ELCAP/R site: the rows in Figure 1. The columns in Figure 1 may be interpreted as variables representing certain "characteristics" of the site. Apart from the site identification information, the "characteristics" represent indicators of "participation" in the ELCAP -- participation in characteristics data collections and rates of engineering data collection.

A list of the variables included in the Meta-CURC is contained in "ELCAP RCDB: USRD: Guide to the Meta-CURC," forthcoming.

Use of the Meta-CURC

The following two examples illustrate the way in which the Meta-CURC may be used to assist users in the selection of sites and appropriate time periods. The descriptions are intentionally brief and no attempt is made to describe the analytic process or to present the results of analytic procedures. Our attention is focused entirely on the process of site and window identification.

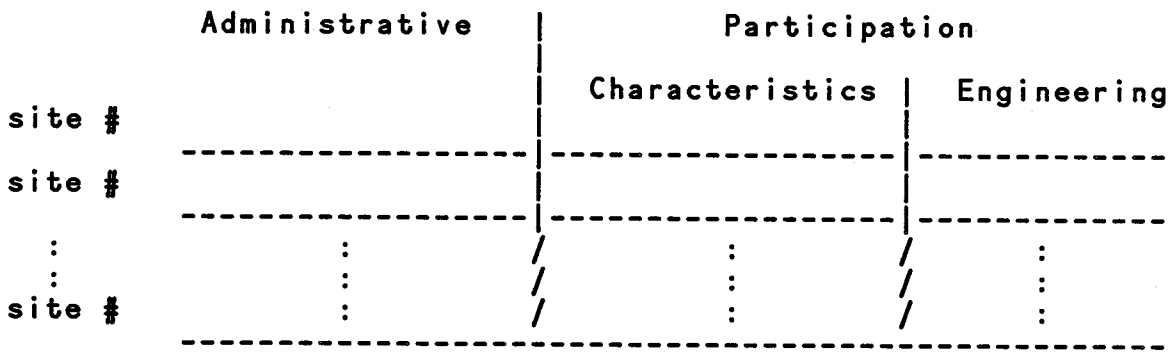


FIGURE 1

Structure of the Meta-CURC

Example 1: Determinants of PNW refrigerator usage

In this case, our typical analyst is exploring the data pertaining to refrigerators. There is no theory to provide a framework for the exploration. The only "hypothesis" to be tested is relatively primitive, as follows:

H(o): The estimate of refrigerator usage in the current end use forecasting models is an accurate representation of actual usage among PNW residences.

However, this "hypothesis" is not the sole objective of the analysis. The analyst also wishes to understand whether and to what extent usage varies over any of several demographic and behavioral characteristics -- income, household size, size and age of refrigerator, for example.

Although there are a variety of ways in which an analyst might approach this topic, the first step would always be to define a "window" (see glossary) of the available sites and period of time for the extraction of characteristics and engineering data. Because it is the most broadly representative of existing residential stock, the analyst may wish to select only sites from the Base Study. Unfortunately, in only a few of these sites is the refrigerator on a dedicated circuit. However, in many of the sites drawn from the Residential Standards Demonstration Program (RSDP), the refrigerator is on a dedicatd circuit. In any case, the analyst will likely wish to begin by subseting the sites by Study Code and on the basis of whether the refrigerator is on a dedicated or mixed circuit. This may be easily accomplished through the initial generation of a "PROC FREQ" table in SAS, to determine the number of available sites, and then through the

application of "IF" statements followed by a "PUT" statement to identify the site identification numbers. The analyst can then determine the availability of engineering data, by month, as well as various other characteristics (i.e., income and number of occupants) through the use of "PROC FREQ" tables or "PUT" statements in SAS.

Having identified the sites and the period of time in which s/he is interested, the analyst then extracts the relevant characteristics and engineering data, merges the two files to form a single analytic dataset, and proceeds to carry out the exploratory analysis.

Example 2: Estimated average space heat use for RSDP and control

In contrast to the first example, an explicit hypothesis is under examination here:

> H(o): Homes built under the RSDP use less electricity per square foot for purposes of heating than do homes built to current building codes -- the "control" homes.

Clearly, the analyst's first step is to identify all the sites which satisfy the Study Code criteria -- either RSDP or Control. Next, the analyst will wish to identify the time period which contains the greatest volume of valid data for the space heating end use. This can be done by examining the Meta-CURC variables indicating the availability of the end-use engineering data on a monthly basis.

The analyst may wish to further subdivide the sites based on the residents' participation in the various occupant surveys in order to ensure the availability of information regarding the use of electricity for space heating (as opposed, for example, to wood). Alternatively, the analyst may wish to subdivide the sites based on the availability of meteorological data, either in full, or, as a minimum, outdoor temperature. All of this information is available in the Meta-CURC and may be extracted either through the use of "PROC FREQ" or the "PUT" statement in SAS.

Having identified the sites of interest, the analyst would then proceed to extract the relevant characteristics and engineering data and then merge the two data sets. The analyst is then ready to proceed with the analytic routines necessary to test the proferred hypothesis.

OVERVIEW

Background

The End-use Load and Consumer Assessment Program (ELCAP) is an electricity end-use monitoring study sponsored by the Bonneville Power Administration and conducted by the Pacific Northwest Laboratory operated for the U.S. Department of Energy by the Battelle Memorial Institute. The monitoring equipment and data collection and data verification procedures are described elsewhere (1). The conventions governing the organization of the end-use data are described in Halverson, Crowder and Bailey (1987). The purpose of this Guide is to describe certain conventions governing the storage of data pertaining to the characteristics of the participating residential structures and their occupants.

The residential portion of the ELCAP includes over 400 sites participating in eight different studies (see Windell (1987)). The data from each of these sites is conventionally divided into two large groups: the "engineering" data and the "characteristics" data (Figure 2). The engineering data includes the records of energy use, indoor temperature records and records of meteorological measurements. All of the site identification, energy data control parameters, building characteristics, equipment characteristics and occupant characteristics data are considered part of the characteristics data.

Characteristics data have been collected from the sites on several occasions using a variety of standard techniques (Figure 3). Until now, these data have been held solely in a relational data management system (Rdb) on the ELCAP computational network. Documentation for this database is contained in JD Schwab and PJ Cowley (1987).

Unfortunately, analysts have found it awkward to extract data from Rdb. In addition, the system is not particularly conducive to the execution of routine validity and consistency checks. The User-accessible SAS Residential Datasets (USRD) have been designed to improve access to the ELCAP residential characteristics data as well as to facilitate improvements in the quality of the data. The objective of this document is to describe the principles underlying the organization and stucture of the USRD.

---

(1) The equipment is described in SD Tomich & GJ Schuster (1985). Portions of the discussion in RS Crowder & NE Miller (1987) are also relevant. The latter also describes the data verification procedures. An early description of the procedures is contained in a paper by EW Pearson, GM Stokes and RS Crowder in RA Stokes et al. (1985).

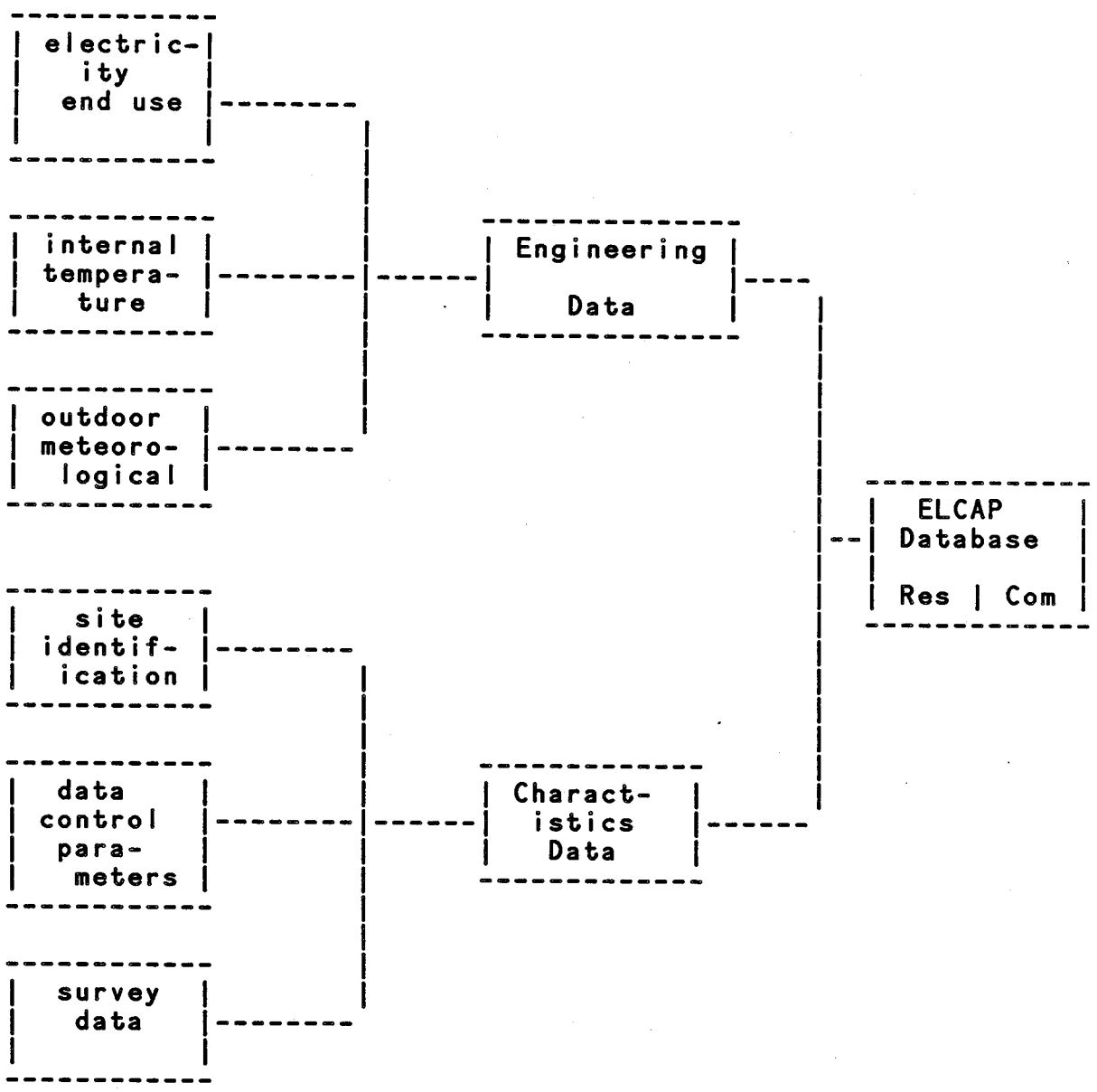


Figure 2

Composition of the ELCAP Database

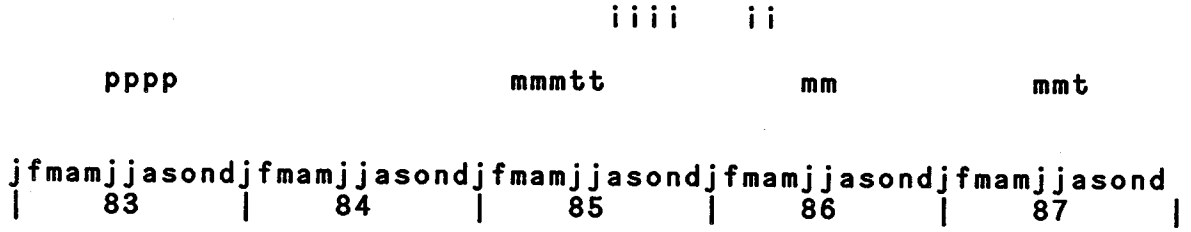


Figure 3

Dates of Characteristics Data Collections

Legend:

p = On-site interview
m = Mail surveys
t = Telephone surveys
i = On-site inspections

Scope

The ELCAP USRD contains all site characteristics data collected under the ELCAP and related projects which include ELCAP residential sites. The data are intended to characterize each of the residential sites, including the structure, the heating equipment, appliance portfolio and the occupants. Some of the data collections originally involved residences which are not participating in the ELCAP; the data associated with these sites have been excluded from the ELCAP USRD. In most cases, the complete survey dataset resides on the ELCAP computer network; however, for purposes of constructing the User-accessible SAS Residential Datasets, the data pertaining to the ELCAP sites have been extracted into separate datasets.

In general, the characteristics data are collected on an irregular basis with a periodicity approximating a year. For the most part, the data are collected from the sites using a common instrument which isadministered over some more-or-less brief window of time. Thus, we shall refer to the data collections as "surveys."

Structural Principles

In a fundamental sense, the integrity of each data element in the USRD is rooted in the data collection from which it originates. As a result, one of the basic units of organization of the USRD is the survey, or segment thereof. That is, data elements are grouped into surveys, or segments of surveys, and the element labels are based on this initial survey grouping. Thus, the labels for the data collected under the 1983 PNWRES include the two letter prefix "PN" while those collected under the Residential Inspections begin with the two letter prefix "RI." The complete set of original survey data elements is referred to as the "Original Residential Survey Datasets" (ORSD).The continually growing set of data elements which have been derived from the ORSD elements through the application of some transformation logic are referred to as the "Residential Characteristics Transformation Collection," or RCTC. Together with the Residential Site Identification Information (RSII), these collections of data comprise the "Complete Set of Residential Characteristics," or "CSRC."

Just as the usage or "engineering" data may be divided into end-uses depending on the equipment or appliance actually using the electricity, so the characteristics data may be divided into categories based on the subject of the stimulus and/or the source of the information. Thus, data elements pertaining to the residential structure constitute one category; elements pertaining to the heating, ventilating and cooling equipment form a second category; and respondent perceptions of the structure constitute yet a third category. A complete list of the "Basic Data Categories" appears later in this document (see page 29).

A complete list of the data elements available in the USRD is accessible in machine-readable form on the ELCAP computational network both as part of the on-line HELP facility and as a separate document. The list is also available as a separate hardcopy document.

The characteristics data may be located within a two-dimensional matrix. The first dimension focuses on the surveys under which the elements were collected. As illustrated in Figure 4, the surveys may be temporally ordered, suggesting the time-series nature of the data. As the second dimension, the data elements may be grouped on the basis of the subjects to which they refer. For our purposes, these subjects have been divided into a limited number of Basic Data Categories (see below, page 35). As a result, each data element may be described in terms of its survey origins and its substantive category (see Figure 4). For example, the occupants were characterized in the 1983 PNWRES and again in the 1985 ROS. Alternatively, the 1983 PNWRES included elements intended to characterize the appliances as well as elements intended to characterize the occupants.

From an analytic perspective, certain data elements are more important than others. In addition, what is of interest and use to the analyst is often an index, scale, or other construct derived from the raw data elements, not the raw elements themselves. Finally, the data do form a time-series and it is necessary to place the data in time-series format in order to complete the linkage between the characteristics and the engineering data. To address these
needs, the USRD is comprised of three "levels." The final, most user-accessible level is known as the "Commonly Used Residential Characteristics" or "CURC." A slightly less accessible but more complete, intermediate level has also been established. For aesthetic reasons, we refer to it as the "Pre-CURC." The first, most basic level is the just the original surveys themselves.

Reference Survey	Data Categories					
	Immut	Mngmnt	Struct	Condt	Appl	HH	Bill
SCODES							
SFLAGS							
PNWRES83							
ROS85/M							
ROS85/T							
RES INSPCT							
ELCAP86							
ELCAP87							
EXT BILL							

Figure 4

Matrix of Data Reference Surveys by Data Categories

OBJECTIVES

The Main Guide

The purpose of this "Main Guide" is to convey the concepts underlying the User-accessible SAS Residential Datasets (USRD) and to present the basic axioms which form its skeleton. Listings of the available data sets and all the data elements of the USRD are included in the on-line VMS HELP facility, as separate machine-readable documents, as well as in hardcopy form as separate documents. Documentation of the transformation procedures is available in machine-readable form as part of the ELCAP INFORM as well as in hardcopy form as a separate document. Documentation for the reference and status matrices are also available in both machine-readable and hard-copy form as separate documents.

The User-accessible SAS Residential Datasets

The USRD is comprised of a set of SAS system files. Each file is analytic ready and some files are constructed in such a way as to explicitly facilitate linkage to the end-use engineering data. However, the real power of the USRD resides in the organization of the data. From this standpoint, the USRD is a set of conventions for the systematic and consistent organization of all ELCAP residential characteristics data.

The system is intended to,

- serve as a source of information concerning the availability and status of residential characteristics data;
- facilitate user access to the data; and
- contribute to improvements in the quality of the data.

As a result, less time and effort will be required for data preparation. In the process, the quality of the data is likely to improve, and our knowledge of the relative quality of the individual elements shall certainly increase.

The system offers users an opportunity to share the products of analytic routines through the permanent collection of available data transforms. For example, the "UA" values derived by PNL staff are

included in the USRD, as are fundamental derivations, such as "Permanent Electric Space Heating Equipment" (or "PESHE"), and a series of data elements derived from the 1985/6 Residential Inspections which are comparable to certain data elements collected in the 1983 PNWRES and the 1985 ROS. The collection includes both the results of the transformations as well as commented SAS code. It is described in detail in the ELCAP USRD: Guide to the Residential Characteristics Transformation Collection.

In the end, two factors are deemed critical to the USRD. First, it must make the data reasonably accessible to an interested analyst who is reasonably proficient in SAS but with little additional specialized training. Second, the system must be flexible and capable of growth.

COMPONENTS AND LEVELS

Overview

The occupants of each of the ELCAP sites have been asked to participate in a survey-like activity on approximately an annual basis since at least the initiation of the project (Figure 3). In addition, there are several datasets containing basic identification and administrative information pertaining to the ELCAP sites. As a result, the full set of residential characteristics data is relatively large and awkward in structure. Efficient storage criteria suggest that the data be held in files corresponding to the individual surveys or segments of the surveys. However, from even minimal experience, it seems clear that in this form the data are very awkward to use, at best, and, at worst, the structure poses significant danger of jeopardizing the integrity of at least some analyses.

In an effort to facilitate user access to the most current version of the residential characteristics data, three succeedingly more restrictive and highly refined datasets have been defined (see Figure 5). From the Complete Set of Residential Characteristics (CSRC), except for those elements deemed to be of interest only to highly specialized deep analyses, the data are standardized and brought forward to form the Pre-CURC. This includes the results of all transformations, aggregations, and derivations deemed to be of interest and use to energy analysts. Whereas portions of the Original Residential Survey Datasets (ORSD) may contain multiple records per site, this is not true of the Pre-CURC; the data have been organized such that for each site there is one and only one record per data item. Because it is still relatively large, the Pre-CURC is partitioned into several datasets and retains the survey-by-survey organization of the ORSD.

The Commonly Used Residential Characteristics (CURC) is a highly refined set of characteristics data which is easily accessible to the user in a single dataset. Only those data elements deemed to be of immediate interest based on the results of past research efforts are included in the CURC. To facilitate linkage with the electricity end-use data, the data are time-stamped indicating the period for which the data are applicable. Furthermore, much of the data have undergone at least simple transformations and reasonable steps have been taken to check for consistency and eliminate processing errors.

Each of these "levels" of residential characteristics data are described in greater detail in the following sections.

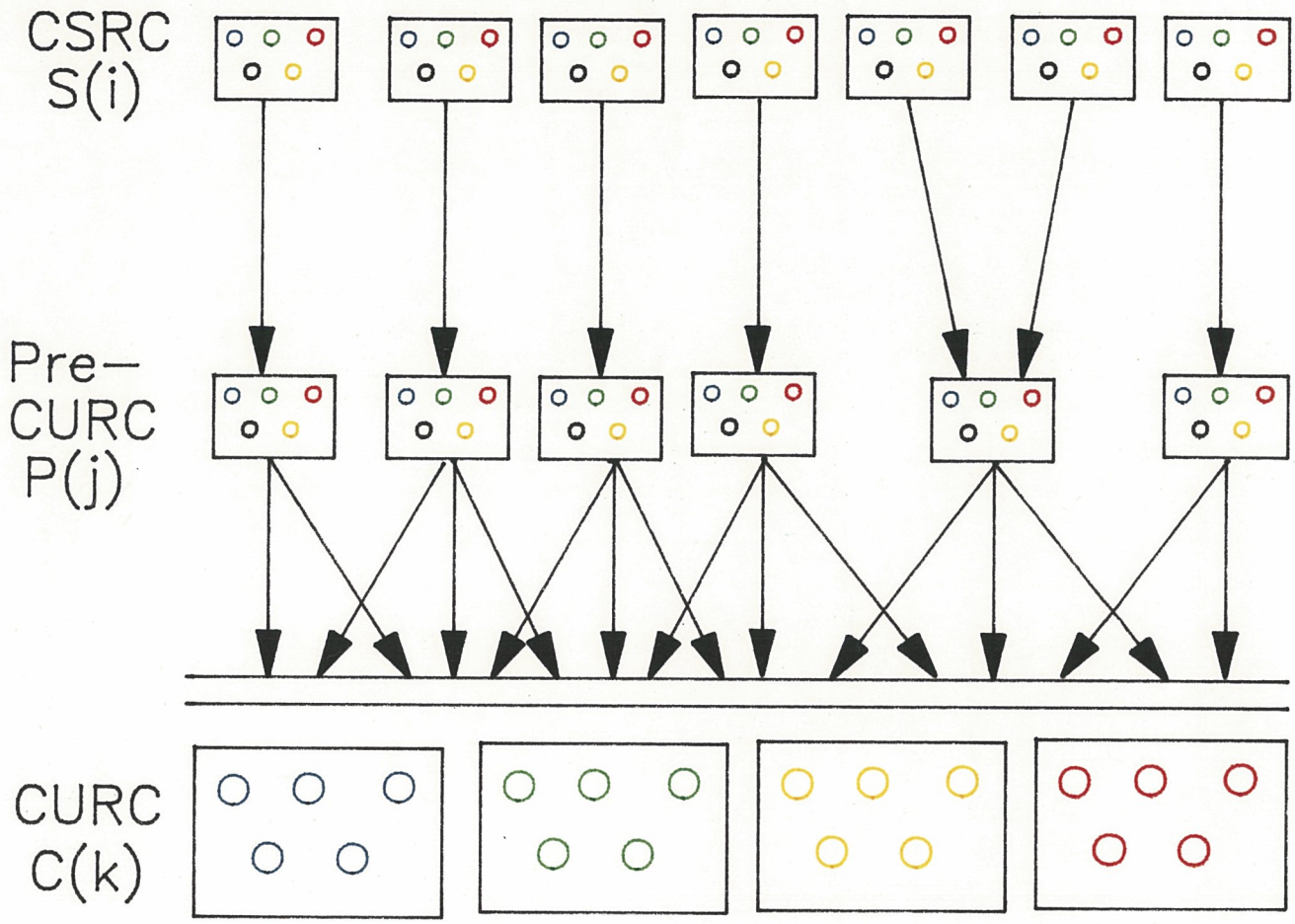


FIGURE 5
Basic Organization of the
ELCAP User–accessible SAS Residential Datasets

Complete Set of Residential Characteristics

The CSRC is comprised of three logical subsets of data. The Residential Site Identification Information (RSII), the Original Residential Survey Datasets (ORSD), and the Residential Characteristics Transformation Collection (RCTC). The RSII is comprised of all the site identification information which is assigned by administrative fiat. The ORSD is comprised of all the raw characteristics data collected from the ELCAP residential sites. The products of all transformations derived from the raw data comprise the RCTC.

In some cases, the original data collection may have involved residences which are not participating in the ELCAP. For purposes of the USRD, the data pertaining to the ELCAP sites have been extracted into a separate file. However, the basic survey-by-survey organization of the data has been preserved. Thus, each data element is stored in a file which corresponds to the survey, or logical segment of the survey in which it was collected. In most cases, the products of transformations are physically stored in the same SAS system file as the component elements. Thus, for example, all of the data collected from the ELCAP sites under the 1985 mail Residential Occupant Survey (ROS85/M) are stored in a single SAS system file together with the results of all transformations based on the ROS85/M data.

For the convenience of interested analysts, the complete list of data elements, by survey, are available in machine-readable form as part of the on-line VMS HELP facility as well as in hardcopy form under separate cover. To facilitate the ready identification of the source of each data element, the initial few characters of the element label denote the "reference survey." For example, the labels for the data elements collected under the 1983 Pacific Northwest Residential Energy Survey (PNWRES-83) all begin with the prefix "PN." The products of transformation operations involving elements drawn from a survey are denoted by adding a "Z" as a prefix to the label. Thus, transforms using data drawn from the PNWRES-83 have a label which begins with the prefix "ZPN."

With few exceptions, all of the data elements collected under each of the surveys are included in the ELCAP USRD. The results of transformations and derivations may also be included in the basic data file. The transformations are clearly identified as such and the procedures employed in the transformations are documented and stored both in hard copy form and as part of the ELCAP computational network.

We may represent the complete set of residential characteristics data as a matrix, say "C." Then,

$$c(i)(j) < C^{*}$$

---

* The symbol "<" is intended to denote "membership."

such that "c" is a data element in the matrix "C," representing the "j" characteristic for the "i" site. For example, c(HHI)(#167) may represent the response received from site #167 when we inquired about their total household income as part of the 1985 mail Residential Occupant Survey.

Then, let "S" represent a matrix such that,

s(i)(j) < S

and,

c(i)(j) <---> s(i)(j)

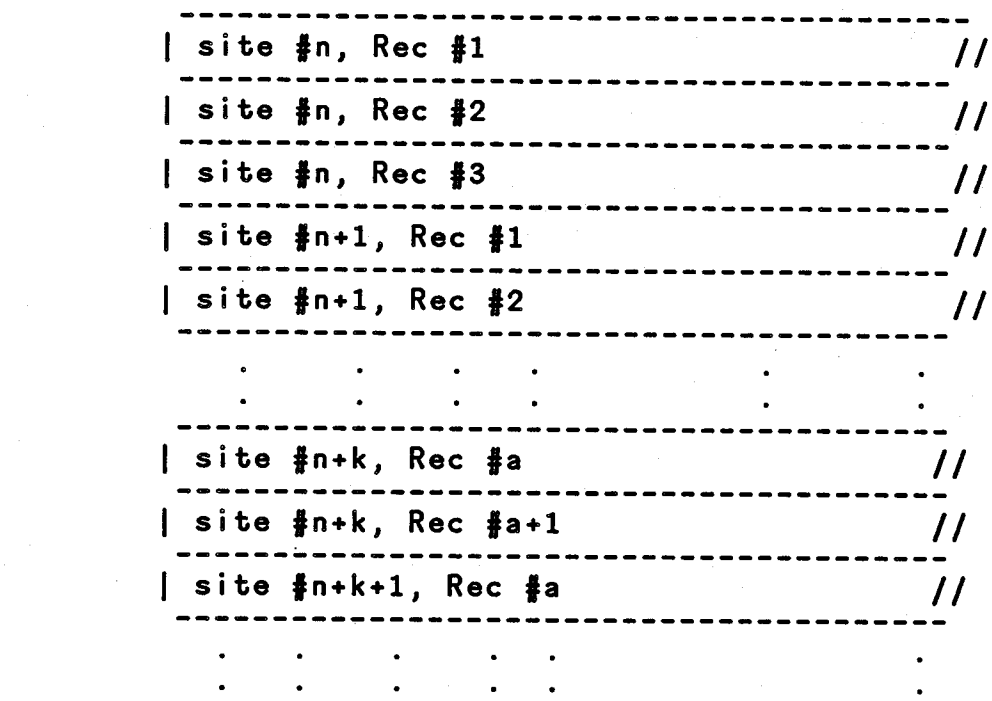
That is, "S" is a matrix of elements denoted by "s(i)(j)" and there is one and only one "s(i)(j)" for every "c(i)(j)" in the matrix "C."

The elements of "S" are intended to represent the "status" of the elements in "C." For our purposes, "status" includes three distinct criteria. First, the values in "S" will distinguish between values of "C" which were obtained directly from the survey and those which were not -- they were "missing" with regard to the survey. Second, the values in "S" may distinguish instances in which alternatives to certain types of survey missing values are available. Thus, for example, "S" may identify instances in which a vector of imputed values is available to replace an item-missing value in "C." It may also identify instances in which a "corrected" value is available, based on other available information. Third, in the case of values obtained from transwformations, the value in "S" will attempt to convey what might be termed the "quality" of the product. The meaning of "quality" and thus the interpretation of the "s(i)(j)" values will likely vary, depending on the data element. Among other considerations, "quality" might refer to the proportion of measured as opposed to defaulted data items used in the calculation of a result. In other instances, the value in "S" might differentiate between results based on field measurements and those based on respondent reports. The conventions governing the Status Matrix are described in greater detail in the ELCAP USRD: Guide to the Residential Characteristics Data Status Indicators.

Pre-CURC

The data included in the Pre-CURC are expected to satisfy the vast majority of the interests of members of the analytic community. The data are "analytic-ready" in the sense that there is one record per site, as depicted in Figure 6. In selecting elements for inclusion in the Pre-CURC, every attempt as been made to retain data which are of potential analytical interest.

SURVEY S(i), CSRC



SURVEY P(j), Pre-CURC

where P(j) <--- S(i)

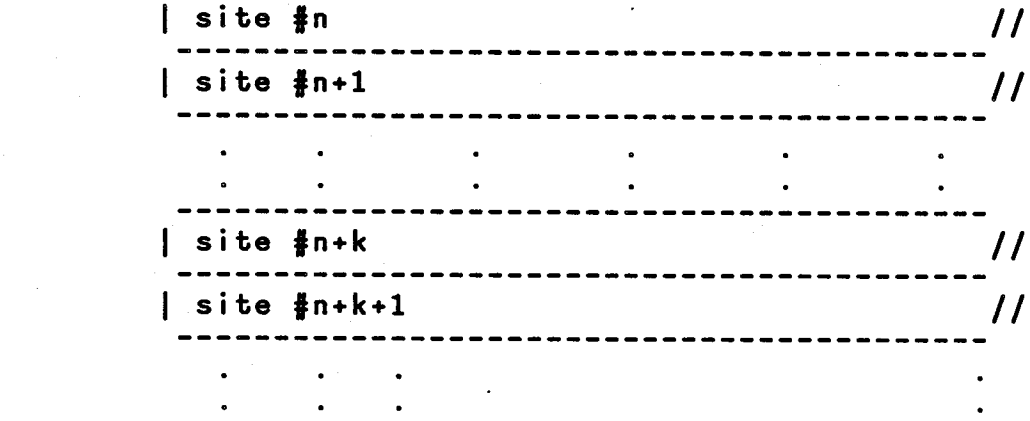


Figure 6

Comparative Structure of CSRC and Pre-CURC

The axioms governing the Pre-CURC are as follows:

1. Data elements in the CSRC are not brought forward to the Pre-CURC if they are deemed to be of interest only to highly specialized and very deep analyses.

2. Where possible, information which requires multiple records in the CSRC is summarized in such a manner that it can be efficiently expressed in a single record in the Pre-CURC.

3. In situations where summarization according to Rule 2 is deemed not feasible, the data elements contained in multiple records are simply transposed so that they can be contained in a single Pre-CURC record.

4. All aggregations, indexes, scales and other transformations which are intended for the CURC are to be included in the Pre-CURC. Unless it is otherwise deemed not feasible, the component elements of such transformations are also included in the Pre-CURC. To the extent possible, the components and the product are stored together in the same SAS system file.

5. The basic survey organization of the data will be retained at the Pre-CURC level. Segments of surveys, especially the Residential Inspections, may be combined into larger physical files in order to make more efficient use of the space. Unless otherwise noted, the products of transformations shall be stored in the survey file containing the component data elements.

6. All data elements in the Pre-CURC will retain the label which was assigned under the Complete Set of Residential Characteristics (CSRC).

Most of the information collected through interview-based surveys is already organized on a "one-record-per-case" basis. It is the information collected in the 1985-86 on-site inspections which have been stored in multiple-records-per-case fashion. Thus, Rules 2 & 3 above are directed primarily at the summarization of these several datasets.

The information concerning vertical components (windows, doors and walls) is summarized on the basis of the "orientation" of the component, based on an eight-point compass. That is, the characteristics of all windows facing a given direction, say southwest, are summarized -- gross square footage, net square footage, average UA. Similarly for doors and walls.

The characteristics of "horizontal components," -- that is, ceilings/-roofs, floors and foundations -- are summarized based on "type" classifications. Thus, all "attic/finished ceiling" areas are combined; all "no attic/finished ceiling" areas are combined; and so forth.

Under this version of the Pre-CURC, the entire dwelling is treated as a single unit. It is possible to break the dwelling into "space heat zones" based on information contained in the Residential Inspections regarding the space heating systems and dwelling architectural characteristics.

Commonly Used Residential Characteristics (CURC)

The data included in the CURC constitute a strict subset of the items included in the Pre-CURC; no new elements are introduced. However, the data are organized on the basis of different principles and they have been explicitly time-stamped and time-sequenced. The data are expected to satisfy most of the day-to-day analytic interests of the casual user and the user responding to time-sensitive management requests.

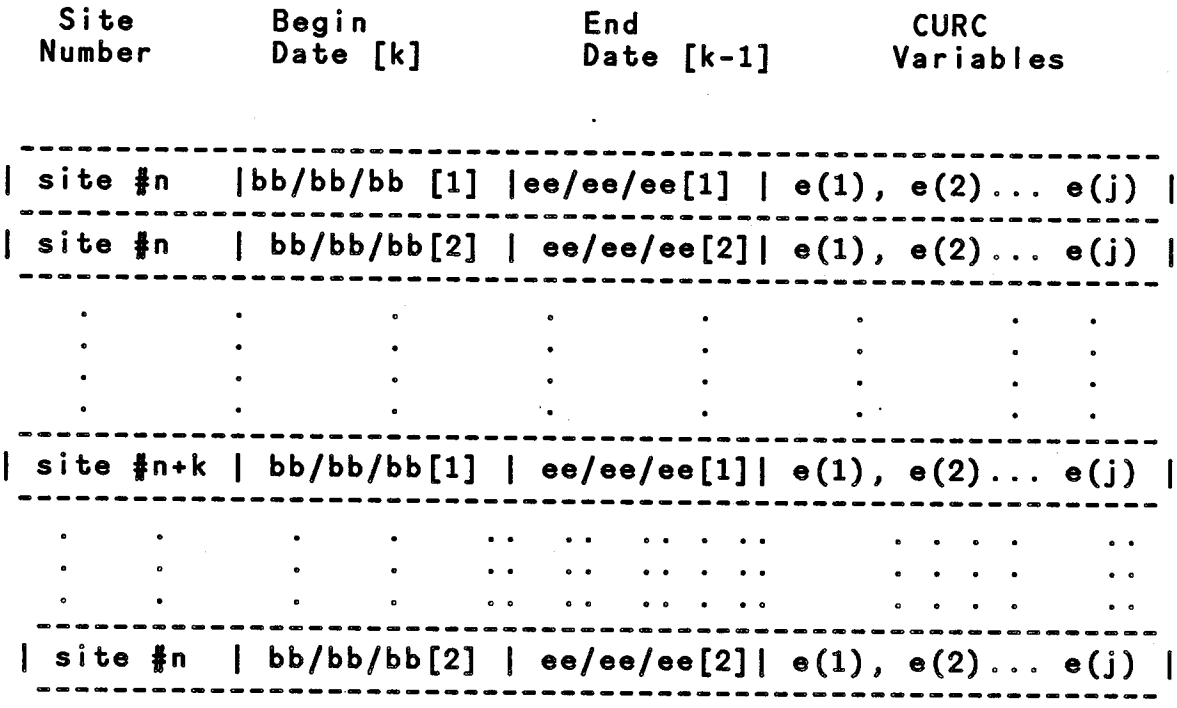
Although no new elements are introduced, the organization of the data differs from that employed in the Pre-CURC. First, the data are organized on the basis of the substantive categories to which they pertain, rather than the reference surveys under which they were collected. In order to avoid confusion, we shall use the term "variable" to refer to particular members of a data category. Thus, a CURC "variable," say "annual income," may include two or more CSRC or Pre-CURC "data elements," i.e., "1984 annual income," "1985 annual income," and so forth.

Second, and as a result, the labeling scheme for the variables is altered so that each variable in the CURC is labeled on the basis of the category into which it falls. The categories and associated neumonics are listed in a subsequent section of this document (see page 19).

The axioms which define the basic structure of the CURC are as follows:

1. Data elements are brought forward to the CURC if they are deemed to be of interest to a large number of users on a regular basis or are of methodological import to a large number of analyses which are regularly conducted. In the case of transformations, (i.e., a = b+c), with few exceptions, the product (a) shall be brought forward, but not the components (b, c).

2. The data are organized according to the "Basic Data Categories" as defined elsewhere in this document, rather than by survey.



where bb/bb/bb = begin date
ee/ee/ee = end date

and bb/bb/bb[k] = (ee/ee/ee)[k-1] + 1

Figure 7

Internal Organization of the CURC:

Implementation of Time Stamping

3. Except for the "Immutable" data elements, each data element is explicitly "time stamped" in the sense that it is associated with a specific time period which is encoded into the database together with the value of the element itself. Time stamping is to be accomplished in the following manner (see Figure 7):

    a. For each physical data file, each record is encoded with a "begin date" and an "end date," denoting the time period to which the data elements pertain. In the record applicable to the current time period, the end date is coded as "missing." When a new record is added, the "end date" of the previous record shall precede by one day the "begin date" of the new record. All changes are assumed to occur at midnight. To facilitate the integration of characteristics and engineering data within the current software environment, minute blocks are included as part of the "begin" and "end" dates. The minute block for "end" dates will always be set to "1440." The comparable block for "begin" dates is always set to "1."

    b. A record is updated each time there is a reported change in any one or more values in the record. Unchanged values are simply repeated in the next record.

4. For a specified time period, the data are organized so that there is one record per case. However, to the extent that the values for a specified site change over time, each change will be denoted with a new record for the site.

*[5. In some cases, the underlying characteristic may be expressed as a real number. As a result of problems associated with the measurement of the characteristic, however, a given site may have an ordinal measure assigned. In such cases, it is possible to assign a real number based on any one of several techniques. Here, we have chosen to interpolate using a multinomial approach (see ELCAP USRD: Guide to the Residential Characteristics Data Imputation Procedures, forthcoming).]

*[6. Based on current plans, analysts may be offered an opportunity to substitute imputed values for item nonrespondents for selected items. In general, replacement values will not be offered for cases which failed entirely to participate in the collection (so-called, "unit non-respondents"). Rather than a single value, analysts will be offered a vector of values for each instance of a missing value (see Rubin, (1987)). The particular imputation techniques employed vary over the data elements. Documentation of the imputation procedures is included in machine-readable form on the computational network and in hardcopy form as a separate document (see ELCAP USRD: Guide to the Residential Characteristics Data Imputation Procedures, forthcoming).]

7. If we let "C" represent the matrix of values for each site, variable, and time period then there is a second matrix, "R," in which the values represent the "Survey Reference" for each of the values in "C" (see Figure 8). Thus,

c(i)(j)(t) < C

r(i)(j)(t) < R

such that, i = site,
j = variable, and
t = period of time

and,

c(i)(j)(t) <----> r(i)(j)(t)

For example, "c(HHI)(#66)(84)" may represent the income for site # 66, for the year 1984. Then, "r(HHI)(#66)(84)" will represent the reference for that value, say, 1985 Residential Occupant Survey, Mail. For the convenience of the analyst, the two matrices are stored as separate SAS Datasets within the USRD. The "Reference Matrix" is described in greater detail in "ELCAP: RCDB - USRD, Guide to Data References and Sources," forthcoming.

---

* These axioms have not been implemented under the current version of the USRD.

8. In a similar manner, there is a matrix "S" comparable to that described previously for the CSRC and Pre-CURC, which portrays the status of each variable for each site, for each time period (see above, pages 20 ff and Figure 8). That is, each entry in "S" indicates whether the value in "C" has been obtained as a straightforward extraction from the original reference; whether a "corrected" value is available upon request; whether an imputation vector is available; or, if the value is the product of some transformation procedure, the "quality" of the result. A more detailed description of the "Status Matrix" is available under ELCAP, RCDB: USRD, Guide to the Status Matrix, forthcoming.

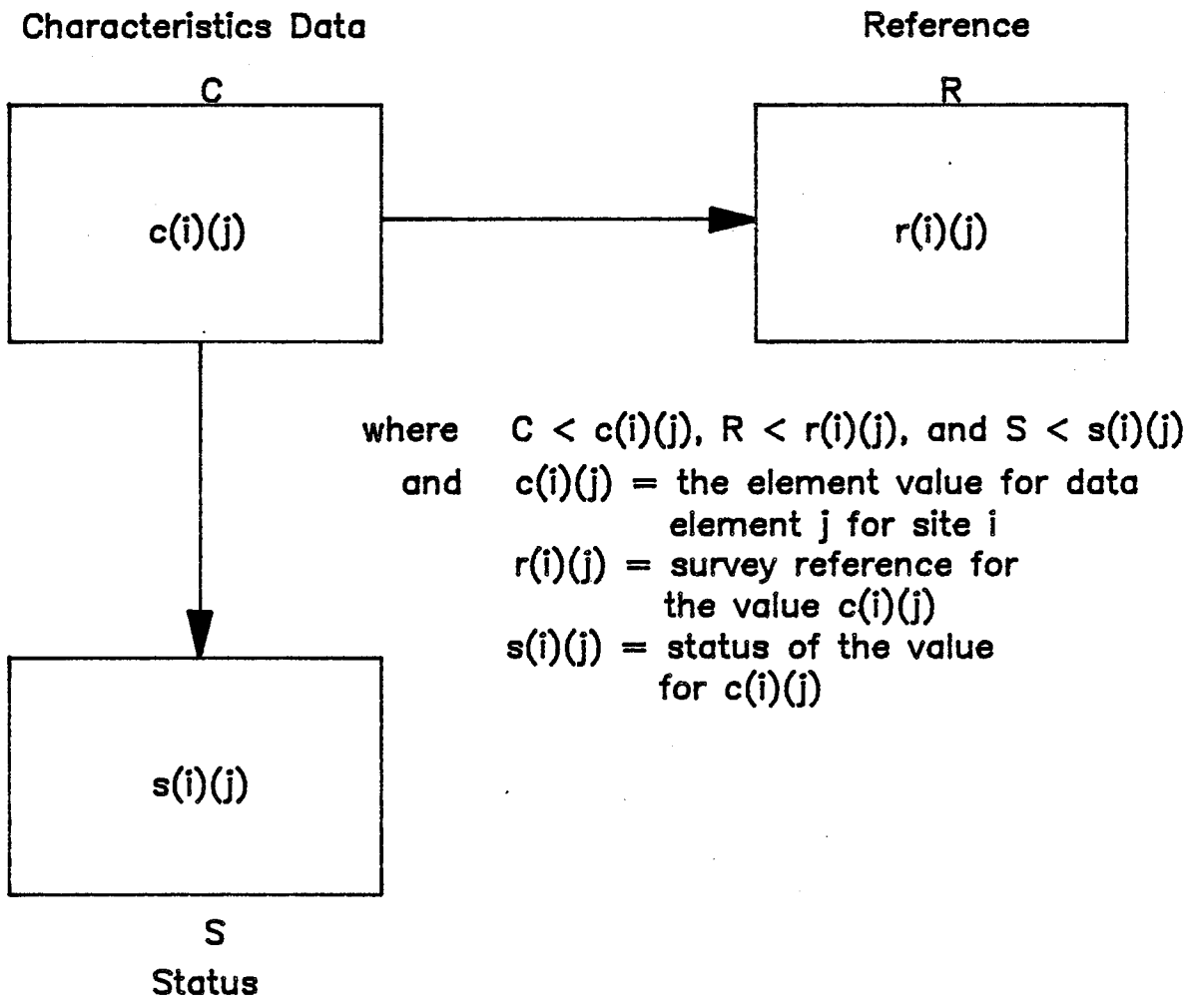


FIGURE 8

Implementation of Data Reference and Status Identification

CHANGES IN SITE CHARACTERISTICS AND DATA ELEMENT VALUES

Introduction

One of the primary purposes of developing the USRD and particularly the CURC is to facilitate the analysis of data with an explicit time dimension. Clearly, what is of interest is the way in which the various data elements change over time. In conducting such analyses, it is obviously important to first distinguish between valid and erroneous data points. The structure imposed on the data by the USRD facilitates such distinctions by explicitly juxtaposing the series of values for specified characteristics, by site. SAS procedures may then be executed to compare values between time periods as well as between elements - over time as well as cross-sectionally. The results may then be reviewed and judgements made regarding the validity of the available value.

The conclusions from this review will be based on our expectations regarding the way in which these elements change over time and on some more-or-less well-defined criteria of validity. Our review may result in suggestions for changes in the values of specific data elements for specific cases. Any such changes will be documentated within the ELCAP computational network as well as in hardcopy form.

These quality control checks are vital to the preservation of the integrity of the USRD. However, in the conduct of such checks, it is important to keep two things in mind. First, "validity" is not a simple dichotomous variable. It is a real-line variable, best conceptualized as ranging between "0" and "1." Second, in many instances, there is either no question regarding the validity of the assigned value or no way to question its validity. For example, there is rarely any question regarding the validity of the US Postal ZIP code which has been assigned; and there is no way to question the validity of an householder's opinion regarding recent changes in electricity prices. Only rarely is it crystal clear that the assigned value is not valid, and in most instances, these mis-assignments are quickly and easily resolved. Between these extremes exists a not-overly large, but at times very frustrating arena in which the validity of the assigned values is difficult to measure.

To facilitate our analysis, the data elements and their associated change rules may be divided into four categories. The category boundaries are based roughly on our expectations for change in the underlying characteristic or in the product of our measurements; and, again roughly, on the significance of the element for regular analyses. The latter is admittedly vague, but the system is flexible so that data

elements may shift categories over time. That is, at some point in the future, we may decide to accord certain characteristics greater significance; or to accord some other characteristics somewhat less significance.

Static Definitional Characteristics

The fundamental point of departure for this entire discussion is the definition of the term "Site." Within the ELCAP, a "Residential Site" is defined as an electricity account serving a more-or-less well defined structure located at a specific point in geographic space and which is used primarily for residential purposes. Accordingly, certain characteristics are by definition static: i.e., any change in one of these characteristics results in the termination of the site and the initiation of a new site. While there is some room for error, measuring each of these characteristics is relatively straightforward so that disagreements may be easily resolved.

Closely associated with these definitional characteristics are a small number of characteristics which, once a site is identified, are assumed not to change. Some of these characteristics may be less easily measured so that disagreements may be less easily resolved. Nevertheless, there is an underlying "true value" for each site which does not change. Perhaps the best example is the year in which the structure was built.

Significant Characteristics

Of the remaining characteristics, some are easily accessible, from a measurement perspective, and of such significance that monitoring change is relatively straightforward and, in consequence, may be done on a relatively frequent basis. Determining whether the site is occupied by the same household is sometimes less straightforward than one might at first think; however, it is of such fundamental significance to the accurate characterization of the site as to sometimes warrant the investment of extraordinary efforts. Several other characteristics are more-or-less closely related to whether the site is occupied by the same household (i.e., year household moved in). Several others are either extremely significant (i.e., fuel and system used for space heating) or the measurement procedures are relatively straightforward (i.e., size of the dwelling).

Characteristics Which Change Systematically

If the site and the occupying household remain the same from one point in time to another, there are a few characteristics which change in more-or-less systematic and expected ways. For example, people get older, not younger; it is possible to complete more years of education, but presumably not possible to lose years of education. Some of these characteristics are relatively accessible to accurate measurement; others may be somewhat more problematic. In consequence, measurement error may present more serious difficulties for some of these characteristics than for others.

Other Characteristics

The remaining characteristics in the USRD change in unpredictable ways, the characteristics are themselves difficult to measure, any changes in the characteristics are difficult to measure, and/or the characteristics are not of such unique significance as to warrant extensive measurement efforts. For whatever reason, changes in any of these characteristics over time must simply be taken at face value. Apart from checking to ensure that it is within the range of legitimate values, there is no basis for concluding that either of two different values for a given characteristic are "incorrect." Indeed, in some cases, this is even true for simultaneous or nearly simultaneous differences in values. The 1985 ROSM and ROST were administered within approximately 6 months of one another and both contained a series of questions concerning the temperature at which the main living area of the home is kept. Although the mean difference is nearly zero, some of the responses differ by as much as 15 degrees Fahrenheit.

This last example may be considered somewhat unique, especially since we are monitoring the indoor temperature as a part of ELCAP so that it is possible to compare each of the responses with the monitored data to obtain some measure of respondent accuracy. However, similar logic may be applied to responses to our enquiries regarding household income. Apart from ensuring that the value stored in the machine compares favorably with that recorded on the collection instrument, we have no basis for concluding that any one of several different values is "correct" and the others "incorrect." As is always the case, it is well to question the validity of extreme cases (say, for example, a change from "Over $60,000" to "Less than $8,000" in annual household income). However, if the relevant collection instruments support the values shown, we must assume that they accurately reflect the respondent's answer to our question.

List of Characteristics by Quality Control Criteria

Static definitional characteristics

- Site number
- Participation begin & end dates
- U.S. Postal ZIP Code
- Time Zone
- Geographic region
- Climate Zone
- Primary & secondary met sites & outdoor temperature reference
- Original serving electric utility
- Original study assignments
- Dwelling type (Single/Multi-family; Attached/Detached)
- Year dwelling built

Significant characteristics

- Status of the site
- Street address*
- Occupant name*
- Year occupants moved in
- Size (square feet; number of rooms, by type)
- Current study eligibility
- Fuels used for space heating
- Fuels used for water heating
- Basic home appliances (stove top, oven, refrigerator, freezer, dishwasher, clothes washer and dryer)

Characteristics which change systematically

- Age of householders
- Education of householders

---

* Access restricted, in accordance with the provisions of the Privacy Act of 1974.

BASIC DATA CATEGORIES

Introduction

The ELCAP residential characteristics data may be divided into groups based on the object being characterized and/or the source of the characterizations. For example, we may distinguish between data elements pertaining to the structural characteristics of the dwelling and elements pertaining to the occupying household. Similarily, we may distinguish between dwelling characteristics obtained through an on-site inspection conducted by a trained professional and similar characteristics obtained through an interivew with an adult member of the occupying household.

By sub-dividing the data we intend to achieve several objectives. First, the scheme should facilitate user comprehension of the data and greatly reduce the time spent searching for relevant data items. Second, some of the groupings correspond to our expectations regarding the rate and direction of change in the underlying characteristics and the corresponding data elements, as discussed previously. The third objective is closely related to the second: facilitation of verification and quality control activities.

The following list of characteristics is extensive, but may not be exhaustive. Users interested in a complete list of the data elements included in the USRD are referred either to the on-line VMS HELP facility or to the machine-readable or hardcopy versions of the lists of the CSRC or any of three components -- the RSII, the ORSD and the RCTC.

The Data Categories

IMMUTABLE

The data elements in this category are integrally related to the definition of an ELCAP site. We therefore assume that the values of the elements never change.

Site Number
Participation {begin; end}
Location {time zone; climate zone; geo region; Zip}
Prefered Meteorological Data Source
Original Study Assignment
Serving Electric Utility at the time of selection
Year dwelling unit built

MANAGEMENT

The data elements in this category are closely related to those in the previous category, except that the values may change without a corresponding change in the site identity. Some of the elements are purely administrative, others are of some substantive import.

- Change in study eligibility
- Current study eligibility
- Change in electric utility service
- Current serving electric utility
- Linkages to other study sites (i.e., RSDP matched pairs)
- SAME
- Urban/rural indicator
- Type of structure
- Survey participation indicators {PNWRES-83, ROS-85/M, ROS-85/T, RI85/6, ELCAP86, ELCAP87}
- Maintenance visits
- Amount of good data; ratio of good to possible; "sporadicity'

STRUCTURE

All of the data elements pertaining to the structural characteristics of the dwelling are included in this category. For the most part, the elements in this category derive from on-site inspections conducted by trained professionals. However, elements based on occupant interviews may have a secondary categorization which denotes them as "structural."

- Basic {Type, attachment, etc}
- Size {Square feet, #stories, heated sf]
- Foundation
- Floors
- Walls
- Doors
- Windows
- Ceiling/roof

CONDITIONING (Space and Water)

This category includes the data elements which pertain to the characteristics of the space and water conditioning systems. The elements include references to the fuels, the source, and the distribution systems.

- Heating {Fuel, Source system, Distribution}
- Cooling
- Ventiliation {Room fans; air-to-air heat exchangers}
- Hot Water { Fuel, Tank capacity, External blanket wrap}

APPLIANCES

The characteristics of the various appliances, recreational, health and business equipment are included in this category. In many cases, the "characteristic" may be a simple "presence/absence" or count of the number of pieces of equipment.

- Food preparation
- Food storage
- Dish & utensil cleansing
- Clothes cleansing
- Water pumping
- Entertainment and recreation
- Business and employment
- Health & Medical

HOUSEHOLD

All of the data elements pertaining to the occupants of the dwelling are included in this category.

Demographic	{number, age, education, income, occupation}
Behavior	{temperature setting, schedule, room closure, appliance purchases}
Perception	{respondent descriptions of structure, appliances, hvac}
Attitude/Opinion	

BILLING

The data elements in this category derive from the energy consumption records obtained from the serving electric utility.

- Periodicity of billing cycle
- Utility charging rates
- Billing period consumption
- Billing period expenditures ($)
- Annual consumption & expenditures
- Consumption ratios (per day, per capita, per square foot)

REFERENCES

Crowder, R.S. and N.E. Miller, 1987, "The Quality of the ELCAP Engineering Data Set," in ELCAP User's Guide, Richland, WA: Pacific Northwest Laboratories

Halverson, M.A., R.S. Crowder and S.A. Bailey, 1987, "From Logger to Archive," in ELCAP User's Guide, Richland, WA: Pacific Northwest Laboratories

Pearson, E.W., G.M. Stokes and R.S. Crowder, 1985, "Data Verification in End-use Metering," in R.S. Stokes, et al., Papers Presented at the Electric Power Research Institute's Commercial End-Use Metering Workshop, Portland, OR: Bonneville Power Administration

Rubin, D.B., 1987, Multiple Imputation for Nonresponse in Surveys, New York: John Wiley & Sons

Schwab, J.D. and P.J. Cowley, 1987, "Characteristics Data Documentation," in ELCAP User's Guide, Richland, WA: Pacific Northwest Laboratories

Tomich, S.D. and G.J. Schuster, 1985, "End-use Load and Conservation Assessment Program: Description of the Field Data Acquisition System for Building Energy Monitoring," Portland, OR: Bonneville Power Administration

Windell, P.A., 1987, "End-use Load and Conservation Assessment Program: Residential Sector Sample Design," Portland, OR: Bonneville Power Administration