THE QUALITY OF THE ELCAP ENGINEERING DATA SET

R. S. Crowder III
N. E. Miller

January 1987

Prepared for the
Bonneville Power Administration
under a Related Services Agreement
with the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

CONTENTS

THE QUALITY OF THE ELCAP DATA SET

INTRODUCTION

The purpose of this document is to describe the quality of the ELCAP data set for analysts. This document is a guide to the problem areas in the ELCAP data. In general, the quality of the ELCAP data set is extremely high, but as in any data set of this size and relatively young age, there are potential problems that the analysts should be aware of. Discussed are these problems along with suggested ways of identifying and correcting them. The motivation for this document's preparation is the belief that "blind use" of the data is dangerous and can easily be remedied by a judicious plan for data review.

Two data types contained in the ELCAP collections are engineering and characteristics data. This document emphasizes the engineering data, which is the record of the actual energy consumption and, in some cases, meteorological data for each installed channel in a given logger over a specified period of time. The second type of data, the characteristics data, is used to control and interpret the engineering data. Characteristics data used to control the engineering data includes, among other things: 1) what each channel represents, 2) the relations that describe which individual sensors should be combined into one end-use; 3) which channels represent redundant measurements that can be used for quality control checks, and 4) the gross verification status of the data. This subset of the characteristics data is referred to as control data in this document. Another type of characteristics data, which will not be addressed in any detail in this document, is the characteristics data that describes the physical and economic characteristics of the buildings and their occupants.

ENGINEERING DATA

This section will describe how the engineering data is measured, processed, stored, and verified from the perspective of how those steps could affect your analysis. We will separate these topics into a discussion of how data is verified, missing data values, how the time stamps are assigned, how

data is converted from digital signals to engineering units, and implications of the hardware design and installation process.

The basic engineering data set is stored as a huge collection of time series records. The data for each logger is stored in individual files containing a week's worth of data for all sensors in that logger. Within each file the data is stored in a two-dimensional array with each column representing the time series data for one sensor channel and each row corresponding to the readings for all sensors for a unique period of time. These files are stored using a special compressed format that allows storage of about six times as much information as would be possible using standard ASCII format. This special format still affords access times comparable to the standard ASCII format. The data records are recorded with an integration period of either 5, 15, or 60 minutes. The vast majority of the data is collected using a 60-minute integration period. The energy consumption data is presented in units of average watts, and the meteorological data is reported as the average of the appropriate metric system unit.

HOW THE LOGGER MEASURES POWER

Electrical power is the product of the voltage and the current. The ELCAP logger explicitly measures both of these quantities and multiplies them together for each monitored channel; the value reported represents the true power consumed by the device or devices being monitored. The current is measured by a device called a current transformer (CT), which produces a signal proportional to the current flowing through the wires being monitored. The voltage is measured with a sample voltage transformer, which produces a signal proportional to the line voltage. The current is measured separately for each sensor channel. The voltage is measured only once for each phase of the electrical service for the entire logger.

Since the logger can report only 256 different values, or counts, and we are interested in circuits from a single light bulb to the main power feed for a large commercial building, it is important that we be able to configure the instrumentation in such a way that each count will represent a different number of watts, depending on the types of circuits that are being monitored. There are three different ways to change the sensitivity of the individual channels in the ELCAP logger. The first is by choosing a CT to match the

rated capacity of the circuits being monitored. The second is to change the scaling resistor used to condition the signal produced by the CT. The third is to change the transformer used to sample the voltage of the building electrical service.

HOW THE LOGGER AVERAGES OVER TIME

Once the sensors have been installed, the logger will record a reading once every second and convert this reading into the proper number of counts. These 1-second readings are then added together for the duration of the integration period; for instance, 3600 readings are added together for an hourly reading. At the end of the integration period, the binary sum is truncated so that it is again less than eight digits long. This is the same as dividing by 2 to however many digits have been truncated. In the case of hourly records, 12 digits are truncated or the total of the 3600 readings is divided by 4096. This averaging by truncation does introduce a small amount of systematic error, but it never exceeds minus 2 percent.

It is this truncated number that is stored in the logger memory and subsequently transmitted to the laboratory. We then convert this digital count into engineering units by multiplying by a calibration factor that is the number of watts per count. It is very important to any analysis that the transmitted count is correctly converted into an engineering unit. The resolution of each channel depends on the size of the CT, scaling resistor, and sample voltage transformer, and on the ratio of the truncation to the number of seconds in the integration period. Figure 1 is a schematic of a typical sensor channel.

An example may be helpful in understanding this concept. Suppose there is a hot water heater attached to a two-phase, 50-amp circuit breaker. The ELCAP instrumentation would be installed using a 100-amp CT for each phase. The CTs would be attached to two channels, and both channels would be scaled so that they read full scale for 50 amps. The channel resolution for the above configuration would be 27 watts per count for hourly data. The resolution could be made finer by scaling the channels for a full-scale reading at 30 or 40 amps instead of 50 amps, because most circuit breakers are oversized. However, ELCAP practice has been to scale to the maximum

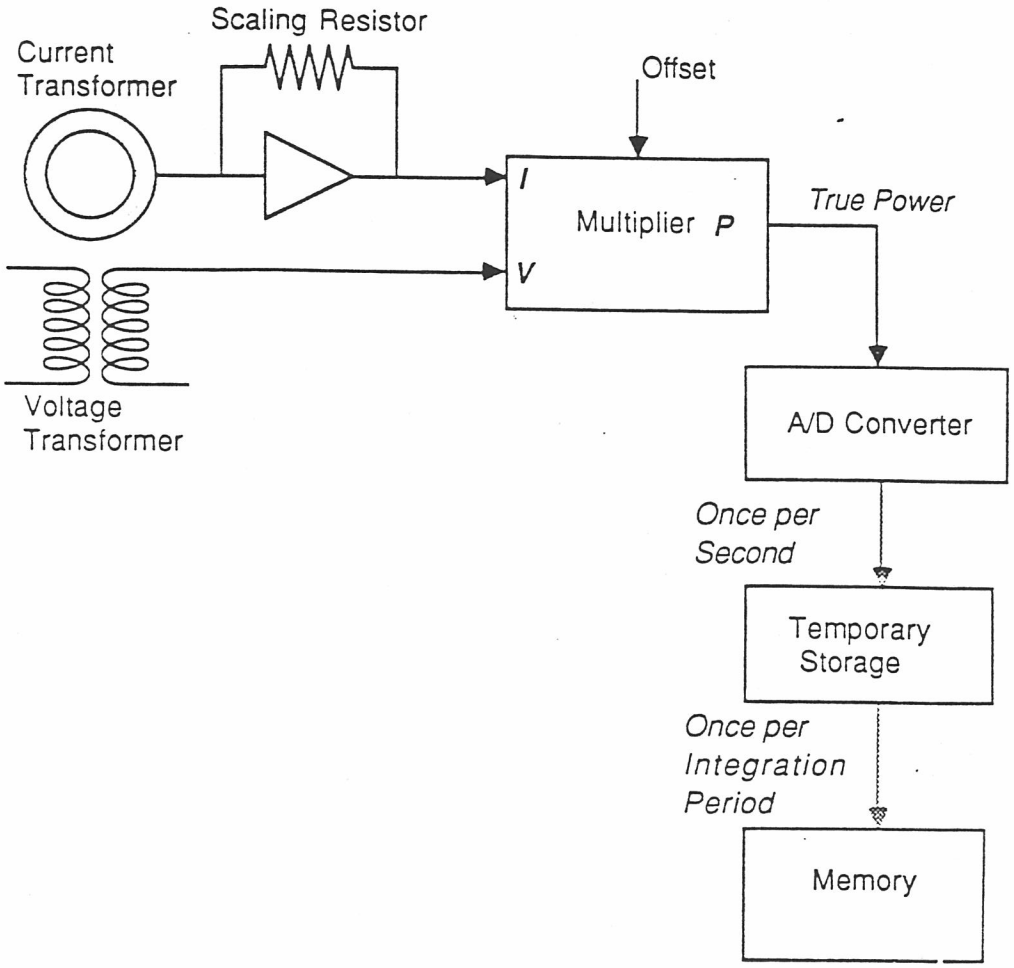

FIGURE 1. Schematic of a Typical Sensor Channel

circuit breaker rating, even though the resulting channel resolutions are larger than may be absolutely necessary.

One other important aspect of the logger that affects the quality of the engineering data is the offset that is applied to each channel within the logger. This offset is required so that the electronics will function properly. The offset is individually calibrated for each channel before the logger is sent to the field. We know that the offset can drift from laboratory calibration, and we will discuss the effect of this in a later section.

The meteorological data is also recorded as a digital signal within the logger and is accumulated and truncated to form the data record in the same way as the energy data. The item of most concern to the analysis is how the temperature data has been treated.

Before March 1986, temperature data was reported in integer degrees Kelvin. Data collected after March 1986 has been converted taking advantage of the full precision of the measurements. To continue storing the data in an integer format, the temperature data was converted so that it is stored as tenths of a degree K instead of degrees K as originally planned. This change has caused two problems. The first is that data collected before March 1986 is now reported with four significant digits but really has only three significant figures, because it was recomputed by simply multiplying the existing value by 10. Second, a few files in the data set still have three-digit temperatures; these missed the multiplication by 10. These files are steadily being corrected as they are found. If temperature values are observed to be three digits only on data records with time stamps before March 1986, this information should be provided on a Problem Identification Form so that the problem can be corrected.

IMPLICATIONS OF THE HARDWARE

The major effect of the hardware design on data quality is the fact that there is a soft zero. What this means is that when a particular circuit is consuming no power, and the data has been correctly converted, the logger may report a consumption (count) of +1, 0, or -1. We address this problem during the initial verification of the data. If the offset appears to have drifted from the laboratory calibration to a different value, we compute the new offset value and use that value to convert future data and to correct existing data. Over the course of the project we have observed that about 10 percent of the channels exhibit offset drift from the time of initial installation. Continuing analysis has shown that the offsets for most channels remain stable after this recalibration. However, we have discovered that 1-2 percent of the channels, over the entire database, do show drift from the original offset value. The amount of drift has not been observed to be more than 1 or 2 counts in either direction, but it does not seem to be readily predictable, so no action has yet been taken to try to reset the offset for the drifting channels.

The impact of this hardware limitation on analysis could be quite profound but can be easily minimized by designing your analysis so that it is not overly sensitive to the minimum value of the time series. Given this

limitation, an example of a bad analytic approach for looking at air conditioner usage would be to assume that any time the air conditioner end-use was greater than 0 the air conditioner is "on". In this approach, it would be possible to find that an air conditioner was "on" 20 percent of the time in the month of January for some loggers in the sample. Closer examination of the data would reveal that the reported load was at most only 10 watts, which is an indication not that the AC is "on" but rather that the channel was reading +1 count instead of 0 for 20 percent of the time.

Of course, because the calibrated channel-specific offset value is truly an offset that is applied to a channel all the time, if the channel's zero is off by 1 count, then all readings for that channel will be off by 1 count during the period of drift upward. The offset problem is pointed out in relation to 0 because 0 is such a convenient threshold to use for many types of analysis.

The engineering data itself is accurate to 2 to 4 percent of full scale (255 counts). This "lid" error figure integrates the accuracy of the electronics in the hardware, the truncation in the logger's firmware, and conversion to engineering units in the laboratory by the central computer. The amount of wattage the 2 to 4 percent figure represents varies from channel to channel, depending on the number of watts an individual count represents. In terms of precision, the average error between the main feed to an electrical panel and the sum of its feeder channels is closer to one half of 1 percent of full scale.

INTRODUCTION TO VERIFICATION OF THE ENGINEERING DATA

The verification procedures are designed to address most aspects of the engineering and control data systems. The emphasis is on proper equipment installation and appropriate conversion parameters for individually metered channels. These procedures have proven to be highly accurate and reliable but they do have limitations. It is important that analysts are aware of these limitations so they can set up appropriate procedures within their analysis to guard against incorrect data biasing results.

The primary tool used in verification is an internal comparison of redundantly metered data. This means that for each electrical panel or set of electrical panels, the main feed to the panel is explicitly metered as well

as all the individual circuits. Figure 2 shows a simple electrical panel instrumented using the ELCAP protocol. By taking a difference-between the reading for the main channel and the sum of the readings for all the feeder channels through a process called sum checking, a determination can be made as to whether the instrumentation has been installed properly. The difference between the main channel reading and the sum of all the feeder channels readings should be close to zero for a properly installed site. The difference is not exactly equal to zero because of the finite resolution of the digitization of the data from individually metered channels. The sum check equations that would be used for the example panel are:

$$|P1-(P3-P4)| < \text{tolerance\_A\_Phase}$$
$$|P2-(P5+P6+P7)| < \text{tolerance\_B\_Phase}$$

where Pn represents the power measured by CTn and tolerances A and B are determined by the resolution of channels 1 and 2, respectively.

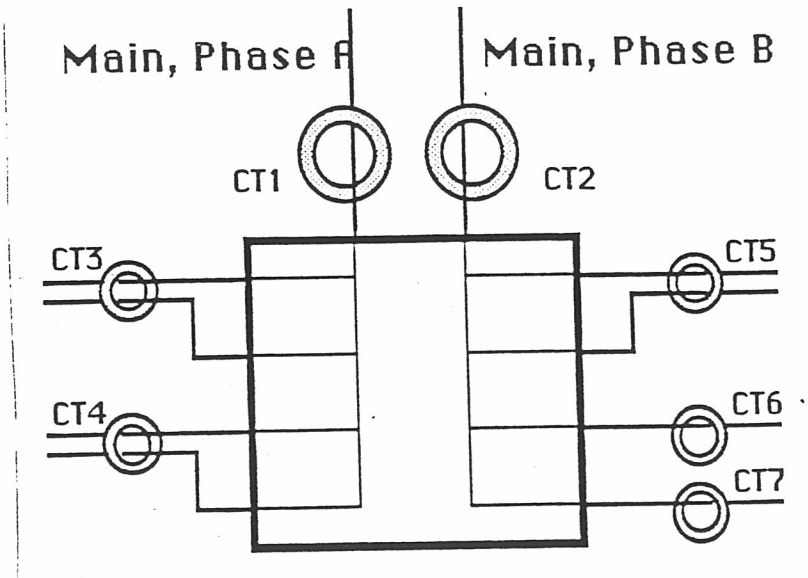

FIGURE 2. Simplified Two-Phase Electrical Panel Monitored to ELCAP Protocol

If the difference is not close to zero, the verification team attempts to identify the reason for the discrepancy. The cause of the error is then corrected, either in software or by a site visit, before the data is made available to the analysts. It should be noted that the ability of this process to identify errors in the data depends on the actual installation. Some review of the verification team's notes in the computer "project" INFORM is recommended for detailed analyses.

In addition to making sure that the engineering data is internally consistent, the verification team also checks to make sure that equations used to produce the end-use data from the channel data are correct, that the labels assigned to the individual channels agree with the measurement plan, and that the parameters used to convert the digital data to engineering units are correct.

While this approach has proven to be an incredibly powerful tool, it does have some limitations. The most important limitation is that, in general, only the first 12 days of data are examined after the instrumentation has been changed, either by the initial installation or a maintenance site visit. The second limitation is that in some buildings the main power circuits cannot be monitored, so it is not possible to compare the feeder channels to any redundant measure. Sites where it is not possible to compare redundant measurements for all channels are denoted with a logger status of 3, instead of a 1 or 2, where a sum check was performed. About 8 percent of the loggers have one or more channels that cannot be sum checked. Despite these limitations, one group of residential data analysts, who were very careful in screening their data, were able to use over 95 percent of the data that the verification team had labeled as ready for analysis.

IMPLICATIONS OF THE INSTALLATION PROTOCOL

It is the built-in redundancy--monitoring both the mains and the feeders--in the ELCAP installation protocol that provides one of the project's unique features. For the bulk of the sites where such redundant measurements have been feasible, this sum checking facilitates initial data verification and holds the key to ongoing and automated data quality assessments for the energy channels.

None of the current engineering data has been routinely sum checked, record by record, past the initial verification. To do such requires that the database of time stamped sum check equations be completed to make the power comparisons possible for all data collected--past, present, and future. For those sites without the possibility of redundant measurement to be automatically scanned for blatant hardware or installation problems or owner building modifications that have occurred after initial verification, the time-stamped database of channel-specific information must be completed. Current plans call for this measure of data quality to be encoded into all data records as part of the data quality flag. We expect that this task will be completed and integrated into the data access tool, EASE, by March 1987. Until then, the tool HISTCHECKER is available to all analysts. Appendix HC gives detailed instructions for operation of HISTCHECKER and a discussion on interpretation of the output.

(Suggestions are made later in this document as to ways existing tools for sum checking data may be used by analysts.)

GENERATION OF TIME STAMPS ON THE ENGINEERING DATA RECORDS

Every engineering data record has a time stamp indicative of the period over which the data was taken. This time stamp records the end of the period the data values cover. At the beginning of the data processing chain, in the data logger, this time stamp is the number of seconds since midnight at precisely the moment the data record is written into the logger's memory. The logger clocks are set to Greenwich mean time (GMT). Each time the logger is interrogated, its clock is compared to Greenwich mean time. Should there be a discrepancy out of the range of 2.5 minutes, the logger clock gets automatically reset.

Consider an example. Suppose that the logger has parameters sent to it instructing that a record be collected every hour. The logger will at the end of each hour precisely, according to its own internal clock, write a record to memory. If this occurred at logger time 8:00 a.m., the time stamp on the record would be 28,800 seconds. The data record, however, represents the accumulation of data scans from logger time 7:00 a.m. to logger time 8:00 a.m. If the logger clock is keeping Greenwich mean time correctly, then the data record is synchronized to both its own internal clock and GMT.

SPIKES CAUSED BY INCOMPLETE INTEGRATION PERIODS

Another small but bothersome problem with the data is caused by wrong number of scans (one second readings) being included in a data record. This can only happen during the integration period when the logger clock is being resynchronized to Greenwich mean time. The result is that a record will be produced that contains too many or too few scans for the integration period. Since the logger software assumes that all integration periods have the correct number of scans, it will truncate the wrong number of digits. If too few scans are included, the record appears to have very low or negative readings, because the summed reading is less than the sum of offsets for the full integration period for the energy channels. If too many scans are used, all channels will report values that are considerably higher than the actual consumption for the guilty record.

In May 1986 a program to remove all the records suspected of incomplete integration periods from the data set was run. This process removed less than 0.02 percent of the total records from the data set; however, the spikes in the records with too many scans were large enough that they had been significantly biasing analysis for those analysts who had not screened their data. Also implemented were changes in the data processing software, so that suspect records are removed from the files as the data is converted from raw to engineering units. Data that was extracted before May 1986 should be re-extracted or examined to remove records contaminated by incorrect integration periods. Any data extracted after May 1986 will not have this problem.

EFFECTS OF POWER OUTAGES

In the case of a power outage when the logger comes back up (upon restoration of power), the logger clock has had both its day and seconds since midnight counters set to zero. If the logger was on hourly integration, then in precisely 1 hour after power restoration the first record would be written to memory. Although this record represents an integration of exactly 1 hour, it has a very small chance of representing a sum of scans going from one GMT hour to the next GMT hour. This could only happen if the power outage was exactly an integral number of hours long. Consequently, for data records written after a power outage, the number of

seconds since midnight will not be synchronized to the Greenwich mean time clock.

HOW TIME STAMPS ARE CONSTRUCTED

When the raw data is brought into the laboratory the time stamp is:

1. converted to Pacific Standard time
2. used to construct a minute block representing the nearest number of minutes since Pacific Standard Time midnight
3. a data quality flag is encoded with information regarding whether this record is after a power outage and also how far the record was off from the Greenwich mean time synchronization.

In all versions of SELECT and EASE, the minute block, modified in accordance to the analyst's time zone choice, is part of the output file produced. The seconds since Greenwich midnight have never been part of the output file and are not planned to be part of EASE 2.0 output either. The data quality flags have similarly never been part of the output file. EASE 2.0 or a future version will take the data quality flag into account when deciding whether a record is suitable for aggregation. The specifics of the algorithm and timetable have not yet been worked out.

MISSING DATA

Missing values are an inevitable result of any data collection effort the size and complexity of ELCAP. To understand how missing values will affect analysis, it is important to understand the cause and frequency of the different types of missing data. It is also important to understand how the various extraction tools treat missing values.

The three different types of missing values in the ELCAP data set are:

1. Isolated records with missing values. These are usually caused by clock resets, random communication failures, or local power outages. These blocks of missing data are usually only one or two records long, and the blocks appear to be distributed fairly randomly over time.
2. Portions of months ranging from 20 to 200 continuous hours. These blocks of missing data are usually the result of the loss of one data dump from

the logger's memory. The data can be lost because of intermittent modem failures, loss of logger parameters, or bad telephone lines. These types of problems affect less than 0.5 percent of the data dumps, but they tend to affect some loggers more than others.

3. Extended periods of missing data ranging from 200 hours to several months. Bad modems or other hardware failures have been the principal causes of these more extensive blocks of missing data.

The discussion of missing values refers only to entire records, or time periods, as missing. This is because ELCAP data is collected in such a way that readings from all active channels must be present or the entire record is deleted. Therefore, there are not any cases where for a particular collection period there are values for seven channels and no values for six channels. There are, however, changes in the number of sensors that are installed at the site. See Appendix A for instructions on how to determine when sensors are changed.

To get an idea of what kind of missing data values you will need to be concerned with, you can use the logger status information that is provided by the new ELCAP data management tool EASE. EASE allows you to look at graphical representation of the amount of data available for a particular logger over time. Figure 3 is an example of the type of information that is provided through the EASE "status" option. (Note: This status information is also available outside of EASE. The files may be printed by using the command $PRINT ss$STATUS where ss is the study name.) The next matter of concern is how the extraction tool will treat missing values. Extraction simply produces a data set that is a complete time series, with each row representing a measurement at some unique point in time. An extraction might consist of five-minute, hourly, monthly, or even yearly data.

The simplest extraction is just a record-by-record copy from the original database. If the data of interest has a lower time resolution (i.e., 5-minute resolution instead of 60-minute resolution) than that used to collect the data, it will be necessary to combine more than one data record together to produce the requested data record.

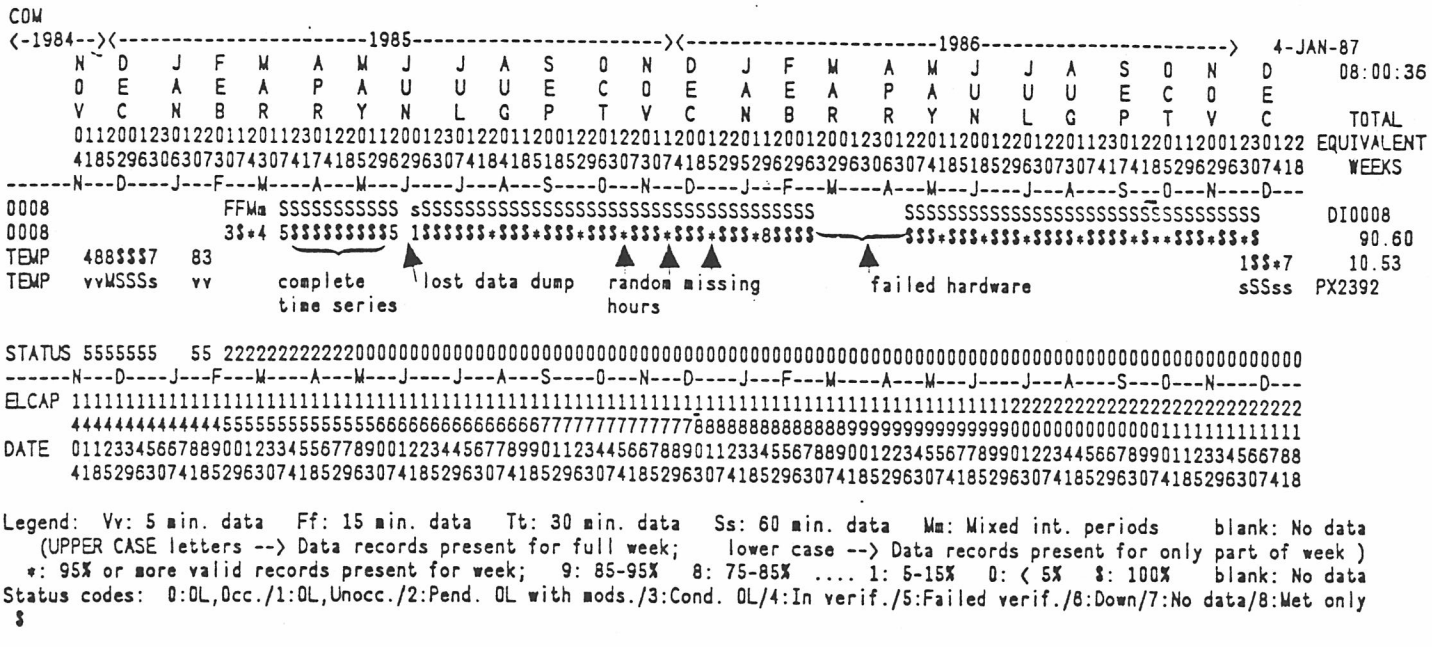

FIGURE 3. An Example of the EASE Status Option Which Shows the Different Types of Missing Data Found in the ELCAP Collection

ELCAP engineering data can be extracted two ways with respect to metering level. The two methods of extraction capture data either at the channel level or end-use level. The channel-level data is the most disaggregated form available. Within each data record, a data point corresponds to a specific sensor channel within the data logger itself. The channels often monitor a single phase of a device or group of similar devices. On the other hand, end-use-level data is derived by summing the readings from all channels that monitor devices with a similar end-use.

As an example of the channel versus end-use level of extraction, suppose a building has four three-phase heat pumps, each dedicated to one fourth of the building floor area. The measurement plan may show that channels 25, 26, and 27 monitor A, B, and C phases of the southeast and southwest heat pumps respectively, and the channels 28, 29, and 30 monitor the A, B, and C phases of the northeast and northbest heat pumps. Extraction of channel-level data would allow comparisons of loading between A, B, and C phases of the pairs of

heat pumps. Extraction at the end-use level would result in a HVAC value that would equal the sum of the readings on channels 25 through 30.

The extraction tool also allows substantial flexibility in tailoring the desired output files. For example, one may 1) sum records across buildings, decreasing the time resolution, if desired, and 2) produce typical profiles (or folded data sets) such as the average day over a given month, etc. Complete details are available in the EASE User's Guide. The analyst needs to be concerned with the way missing values are treated in extracted data sets 1) with lower time resolution and 2) that are folded over time.

How SELECT Treated Missing Values

SELECT was the original extraction tool provided for ELCAP data. SELECT made no attempt to fill in missing values for any of the extracted data sets. For averaged or folded data sets, SELECT reported that a row was missing if any of the component rows contained a missing value.

EASE Release 1.0

EASE 1.0 allows the user to specify a threshold missing percentage that will be used to determine when a row is reported missing for averaged or folded data sets. Refer to the EASE document for details on how EASE fills in for missing values for aggregated data sets when the missing percentage has been set to greater than zero. Note that the type of fill-in used depends on the statistic that the user has requested. EASE 1.0 still does not attempt to fill in missing values for extracted data sets.

EASE Release 2.0

EASE 2.0 will add the feature that the user will be able to specify whether or not missing values should be filled in for extracted data sets. The methods used to determined what value should be used to fill in a missing data value have not yet been finalized, but there will probably two or three options available. The availability of fill-in values for missing records will also give the user three additional options for filling in missing data for aggregated or folded data sets.

METEOROLOGICAL DATA

The meteorological data collected as part of the ELCAP study cannot be sum checked. Although it is conceivable that reasonableness checks specific to end-use could be performed on the meteorological data, these are not planned in the near future. The integrity of the time-stamped aggregation equation data base is crucial to extracting the correct channels, not just for the energy channel and end-use data, but for the meteorological data also. Any analyst not incorporating reasonableness checks on the meteorological data is running an unacceptable risk.

SELECT provided no easy way of performing reasonableness checks on the meteorological data. Before EASE 1.0, the reasonableness checks were largely done within customized analysis tools. This was often as simple as a data review macro used within the analytical processing itself. With the advent of EASE 1.0, some very preliminary snapshot information became available for the channels and end-uses extracted. The maximum, minimum, means, and standard deviations for each of the extracted columns are output along with the data itself in a separate file called the codebook. Depending upon the requirements of the analytic task at hand, automated or manual inspection of the codebook may be enough for meteorological review; often, however, it will not, and the development or adoption of a data review macro is strongly recommended.

CHARACTERISTICS DATA

This section will discuss the characteristics data that is used to control extraction of the engineering data. Another type of characteristics data, which will not be addressed in any detail in this document, is the characteristics data that describes the physical and economic characteristics of the buildings and their occupants. However, the physical and economic characteristic data is subject to two types of errors: data entry and incorrect information on the survey instrument itself. Any problems with the non-control characteristics data may be reported using the PIF process. (Any anomalous information noted on the survey instrument at the time of digitization by the data entry staff has been noted in ANALNOTES subdirectory in the project INFORM.)

TYPES OF CONTROL DATA

Four types of data are in this collection. First is the data describing what each energy channel is measuring. This relation, CHANNEL, includes the label of the channel, the parameters used to convert the channel from digital counts to engineering units, and a field stating whether or not the channel has been verified. The second type of data is contained in the ENDUSE relation. This relation gives the information about which individual channels should be added together to produce a specific end-use. The third relation, VERIFY, contains the sum check equations that are used to verify the data in initial verification. The final relation, LOGGER-HISTORY, contains information about the verification status of the logger and the time resolution of the data.

This data is used to control processing and extraction of the engineering data. It is possible that incorrect results may be generated at extraction time, not because the underlying engineering data is flawed, but because the characteristics data used to control the engineering data extraction is wrong. Discussed below are some of the problems with the control data and ways to spot the problems. If any problems are found in the control data, please file a PIF so that the problems can be corrected. The PIF should be accompanied by any logs or output files that illustrate the effects of the error.

TIME-STAMPING CHARACTERISTICS DATA

Originally all control information was contained in non-time-stamped files that were keyed to particular loggers. However, it soon became obvious that the information would change over time as building owners modified their buildings, instrumentation errors were corrected, or the sensor complements at the building were changed. Therefore the information was moved into the more flexible form of a relational data base. This change has both good and bad implications. The most important change is that all pertinent parameter sets are available at all times, so that it is possible to analyze data available for a logger even though there may be two or more different parameter sets involved. Unfortunately, we did not start out with the ability to time stamp the data, and the historical data must be backfilled

for sites with more than one parameter set, which affects most all sites. This backfilling process is an ongoing effort.

If the time stamps on the control data are incorrect, one may extract data that is incorrect or be denied data access altogether for the affected site. This is especially problematic when end-use assignments change. For instance, suppose that the AC end-use should be c25 + c26 from May_1 into the future, but should be c30 + c31 for any data before May 1. If the control information is not entered for the earlier parameter set, and the data is extracted for an entire calendar year, incorrect values will be returned for the dates from January through May.

These issues are addressed by the verification and data processing staff, but mistakes do slip through. An awareness of these types of problems should help guard against them when possible. It is also important to notify the data processing staff of any errors found so that corrections can be made for future analytic work. The PIF process is the recommended vehicle for documenting such problems.

THE CURRENT STATE OF THE CONTROL RELATIONS

The four relations are in various states of completion and verification based primarily on the amount of use each has received so far. The relations in the order of completeness are as follows:

1. The LOGGER-HISTORY relation is completely filled in.
2. The ENDUSE relation has been completely filled in for about 70 percent of the on-line sites, and every on-line site has at least the most recent parameter set complete and verified. The remaining sites will be backfilled as resources become available.
3. The VERIFY relation is in the same condition as the end-use relation. Because these two relations require the same information, they are typically updated at the same time by the data processing staff.
4. The CHANNEL relation has been completely loaded with historical data and has not been verified to any extent. The information stored on the data interrogation computers was used recently to load the information into CHANNEL. The relation is used only by the newest piece of verification software and by the newly released EASE extraction tool; hence it has not

been exercised to the extent that the other three relations have been exercised.

HOW THE EXTRACTION TOOLS USE THE CONTROL DATA

SELECT used only the time-stamped ENDUSE relation. Because of this it was very dangerous to use SELECT to examine data at the channel level, as SELECT would not have been able to detect changes in parameter sets at the channel level. SELECT did use the ENDUSE relation, and in many cases provided good results for data at the end-use level if simple extractions were required. (If the change in parameters sets was too complicated, SELECT would sometimes produce results incompatible with the actual engineering data.) EASE uses all four of the control relations previously discussed. Data is correctly extracted at the channel or end-use level using these time-stamped relations.

DEALING WITH A DATABASE UNDER DEVELOPMENT

USING THE EXTRACTION TOOLS TO ASSIST IN DATA REVIEW

Although no routine sum checking has taken place for data after initial verification, the mutual exclusiveness of several of the residential and commercial end-uses can be used to help validate: 1) that the end-use equations are formulated correctly in the characteristics data base, and 2) if 1) is correct, that the power on the main is comparable to that on the feeders.

This has been done by some of the analysts working within Battelle, for example, by those working on the residential thermal analysis projects. In the residential case the three end-uses: HVAC, HOT water, and OTHer mix are all mutually exclusive and should sum to TOTal within a certain error tolerance. The actual tolerance band is a function of both the channel resolutions for the individual site and the number of records in the aggregation. The greater the number of records in the aggregation, the greater the agreement that can be expected. In the commercial case all the end-uses other than TOE (total) should sum to the collection of all non-TOE (non-total) end-uses.

With SELECT, end-uses could be combined to facilitate such comparisons. Alternatively the comparison could be done within analysis software. EASE 2.0 will allow for algebraic combinations of channels and end-uses which, together with the codebook statistics, allows a nice end-use "sum checking" of all the data to be used in a given extraction.

THE PROJECT INFORM

The computer project INFORM on the AVAX was created to facilitate information exchange between analysts, verifiers, and site maintenance personnel. Stored in INFORM are tools with world READ and EXECUTE privileges that can be of great help in providing checks or deeper looks at both the control and engineering data for a specific site. Although no routine sum checking of the engineering data has been performed, there are tools residing in this project that can greatly assist data inspection.

The QUERY program can be used to retrieve all the end-use aggregation equations, all the channel information currently in the database, the sum check equations, and logger history records--all with their windows of applicability. The logger history records often give anecdotal information regarding the data processing steps for a given site as well as information similar to that found on the line graphs that are part of EASE 1.0 project status option. The HISTCHEKER program operates on the sum check equations in the characteristic database by providing an energy comparison statistic for each week of data that the analyst has access to. This program is particularly useful in identifying exact troublesome data windows or performing a quick check on the data quality for the energy channels.

A variety of other programs in the project INFORM make it easier to get at the information analysts and verifiers at Battelle have found useful. For example, one program finds the identification numbers for a given site based on input of only one of these numbers, for example the site identification. Another program makes the conversion back and forth between calendar and ELCAP days.

Also within the project INFORM are also all the verification summary notes for both the residential and commercial sample and the analysis notes for the RSDP and RES BASE thermal analysis work. The reader is referred to

the section in the ELCAP User's Guide that discusses INFORM's structure and content.

WORDS OF CAUTION

As the engineering database gets more usage, problems with data quality and data entry errors will certainly be uncovered and must be brought to the attention of the data processing and verification task. One of the more common problems encountered with SELECT, the first extraction utility, was an error of "bad equations". Giving the error log from such a run to the data processing staff assured that the problem would be fixed. The SELECT "bad equation" error usually meant that the end-use equation called out for a channel that did not exist in some part of the requested data window. This was a fatal error and prevented any data from being extracted at all. This was true even if only one equation was incorrect and all other equations in the requested window were correct.

This was annoying enough that such problems were brought to the attention of the data processing staff. With EASE 1.0, a single or set of bad equations will not necessarily stop the entire extraction. It actually only stops the affected extraction of end-uses or channels after the first bad string is encountered. A log, however, will still give the same explicit information that needs to be channeled to the data processing staff for correction. Most often, "bad equation" errors occur when the meteorological data channels are selected for extraction. The analyst needs to be wary of any data that was extracted with such an error message.

BUDGETING TIME AND MONEY FOR DATA REVIEW

Any analyst should understand the nature of the data with which they are working. This is true whether or not ELCAP data is the data under study. The ELCAP redundant measuring protocol has made the automated data quality checks for the bulk of the energy data possible, although it has not yet been routinely carried out. Currently some tools exist to facilitate this process for the analyst.

The ELCAP data consists of the engineering and characteristics data. The engineering data is made up of energy channel data and meteorological data. The characteristics data makes the interpretation and checking of the engineering data possible. The ELCAP analyst needs to take the time to:

1. understand the developmental state of each type of data they will be working with
2. understand the potential problems of each type of data
3. report any anomalies found.

For example, a characteristics database data entry error could give an analyst windspeed data when wood stove data was requested. Since data has not been historically sum checked, an installation problem or hardware failure may be apparent in current data. The point here is that the data review plan for any research project is just as important as the analysis plan itself. The extent of the data review required will be determined by the sensitivity of the particular analysis to data aberrations.

The initial verification process--bringing sites on line--has provided an amazing track record of predicting future data quality. During the RSDP and RES BASE thermal analysis, very few sites were removed from analysis for the degradation in data quality once they were brought on line. Few installation errors or hardware failures have been found to date that were not identified in the initial 1-1/2 week inspection comprising the initial verification process.

ELCAP data quality is quite high, even for its present stage of development. This, however, does not free the analyst from the responsibility of understanding the data and reviewing the data that his or her professional

reputation may rest on. The budget for data analysis should always include the budget for data review.

Distribution

No. of
Copies

OFFSITE

2 W. M. Warwick, Chief
Assessment and Evaluation Branch
Office of Conservation
Bonneville Power Administration
P.O. Box 3621 - KES
Portland, Oregon 97208

30 DOE Technical Information Center

ONSITE

DOE Richland Operations Office

J. J. Sutey

32 Pacific Northwest Laboratory

R. S. Crowder
L. O. Foley
S. G. Hauser
N. E. Miller
W. F. Sandusky
G. M. Stokes
R. A. Stokes
ELCAP Program Office (20)
Publishing Coordination (2)
Technical Report Files (5)