APPLYING PATTERN RECOGNITION TO VALIDATING TIME SERIES DATA FOR ELECTRIC UTILITY LOAD RESEARCH

Andrew E. Allen, Minimax Research Corporation
Tom L. Johnston, Minimax Research Corporation
Edward L. Tabakin, Minimax Research Corporation

1. INTRODUCTION

In 1981, PGandE's Residential Peak Load Reduction (RPLR) Program included the direct control by PGandE of over 38,000 central air conditioners and 600 electric water heaters, distributed over eleven test site areas. Since then, the program has grown to include over 68,000 air conditioners and 2,000 water heaters. Participants are given an incentive in the form of a monthly discount on their bills.

To provide load data for analysis of the program, PGandE installed pulse-initiating watthour recording equipment, including both magnetic tape and solid-state load profile recorders, at 500 participants' houses distributed over seven test sites. Figure 1 shows the configuration of the cycling and metering equipment. The receiver is coded to respond to signals for a certain "address" (radio or tone frequency). When an appropriate signal is received, the switch opens, turning off the air conditioner(s). The meter measures both energy usage of the air conditioners and of the total household, and sends a pulse to the recorder each time a predetermined number of watthours have been consumed. The recorder has four tracks: one records timing pulses generated within the recorder, and the other three are available for recording meter pulses. Once a month, the meter reader removes the entire record, which is taken to a microcomputer that translates the pulses into kWh usage averaged over 5- or 15-minute intervals and checks for certain types of recorder malfunctions. The translated data are then put into a SAS dataset for cleaning and analysis.

The conclusions that utility company analysts could draw from the load data are very sensitive to certain kinds of errors. For example, it is very important to know how much the verage load of a group of customers can be made to drop by curtailing their air conditioners during a peak usage period. If some customers' data were being reported "out-of-step" with that of others, and this condition were not recognized, then the group's average load drop would appear smaller than it really was.

Failures can occur in any link of the chain illustrated in Figure 1:

- The transmitter can fail to send the correct signals, or can send them at the wrong time.
- The signals may occasionally or consistently fail to reach the receiver.
- The receiver may fail to recognize signals sent to the correct address, and may even recognize signals sent to the wrong address.
- The switch may fail to open, either intermittently or consistently, upon receipt of the correct signal.
- The air conditioner may continue to function after the switch opens, if the switch is mis-installed.
- The meter may produce an incorrect number of pulses.
- The recorder may either add spurious pulses or drop valid pulses, on either the data tracks or the timing track.
- The cartridge may be incorrectly removed or inserted.
- The meter reader may write down the wrong cartridge removal or insertion time.
- The data may be mistranslated on the microcomputer, for example, by using an incorrect scale factor.

In addition, mis-installation may prevent any usage from being recorded on one or more channels.

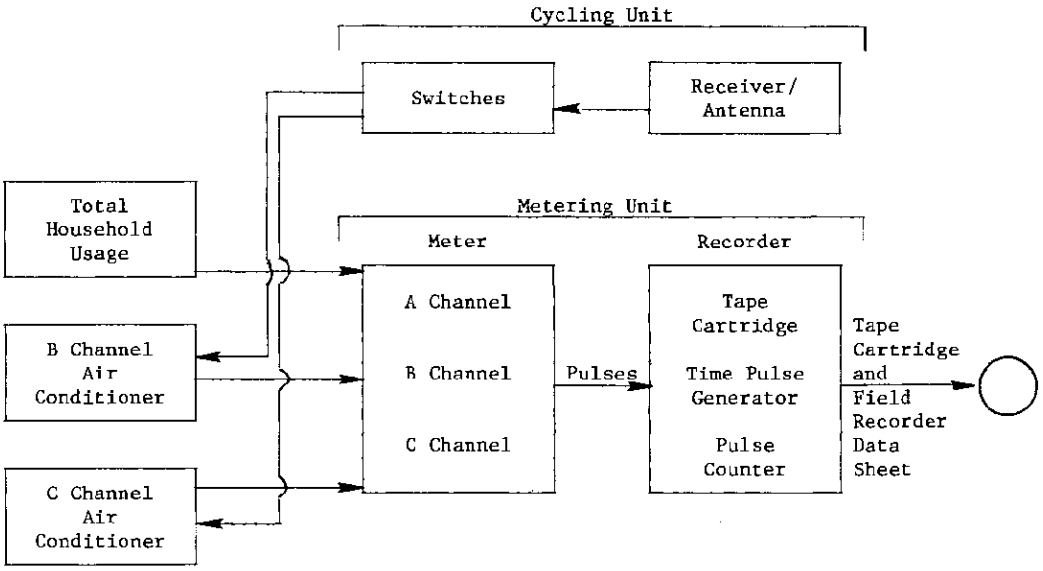
2. DISTINGUISHABLE DATA PROBLEMS

Although some types of failure are indistinguishable on the basis of their effects, this is not a problem for purposes of load impact analysis.[1] If the air conditioner fails to turn off when signaled, for example, it is irrelevant to the analysis whether that was due to mis-installation or to component failure in the switch or the receiver, since these problems are functionally equivalent. We have, therefore, divided the problems we have observed into functionally equivalent groups.

The types of problems that are mutually distinguishable using only their characteristic effects on the load data are as follows:

- consistent failure to receive or respond to signals;
- intermittent failure to receive or respond to signals;
- transmitter failure (affecting an entire site);
- asynchroneity (a valid record misplaced in time);
- persistent insertion of spurious time or data pulses;
- misaddressed receiver;
- impossible relationships between household and air conditioner loads;
- spikes; and
- certain types of mis-installation or component failure.

Figure 1: Configuration of LMG Cycling and Metering Equipment



The latter includes failure to record household usage, failure to record air conditioner usage, or recording a water heater instead of an air conditioner.

Most of these problems are fairly straightforward to detect. Among the more difficult to detect are the following:

- o asynchroneity and/or insertion of spurious time pulses;
- o misaddressed receiver; and
- o intermittent failure to respond to signals.

All of the above problems require the ability to perform pattern recognition, by matching the observed behavior against the expected behavior. Minimax developed an innovative method using SAS to recognize these patterns using only the load data and a few manually prepared parameters. This method was so successful that we were even able to identify occasions when transmitters departed slightly from the normal operational pattern, where this had either not been known or had not been recorded.

3. PATTERN RECOGNITION

3.1 Converting Five-Minute Load Data to Time/Duration Format

At the heart of our pattern recognition system is a process for converting the SAS data base of load data from 5-minute format to time/duration format. Unfortunately, load data averaged over 15-minute intervals has insufficient resolution for this conversion process. This type of load data comprised only 6% of the data base.

An example of 5-minute format is shown below. Each observation contains 288 values, corresponding to the average demand from midnight to 00:05, 00:05 to 00:10, . . . , and 23:55 to midnight.

Example of 5-Minute Data Format

DATE	DMD 1	DMD ...144	DMD 145	DMD 146	DMD 147	DMD 148...288	DMD
17JUN81	0.00	0.00	1.06	2.00	1.20	0.00	0.00

The corresponding time/duration format data are shown below. Each observation contains the time, a flag indicating whether the air conditioner was on or off, and the duration of that event.

Example of Time/Duration Data Format

DATE	TIME	FLAG	DURATION
17JUN81	0:00:00	Off	12:02:39
17JUN81	12:02:39	On	0:10:21
17JUN81	12:13:00	Off	11:47:00

Converting from 5-minute to time/duration format requires only one external parameter: the kW rating of the device (channel) whose load record is being converted. This we derived in a unique way, from the load record itself. Using PROC PCTL, we took the kW rating to be the 99th percentile of all 5-minute demand values on that channel during selected hot spells. This is equivalent to using the maximum value, except that, if the channel has any small spikes that were not caught during the cleaning process, they will not be selected as the kW rating.

Using the 99th percentile value requires one to assume that the air conditioner ran for an entire 5-minute demand measurement interval at

least 1% of the time during the selected hot spells. Our analysis has shown that this is a safe assumption to make. The cases where the 99th percentile value is too small to be credible as the kW rating are cases where the air conditioner was rarely if ever used, so no kW rating is needed and the conversion to time/duration format may be skipped. Observations of test (cycling) days showing no air conditioner use during the curtailment period are set aside by the pattern-recognition process, as described below in Section 3.3.

The conversion process takes each 5-minute demand value and converts it to number-of-minutes-on, by dividing by KW (the kW rating). We begin by starting at midnight, and assume the air conditioner is "off." It stays "off" until a 5-minute interval is encountered that has a value greater than the threshold (see below). It is assumed to change state at most once[2] during the interval, and remains "on" until a 5-minute interval is encountered whose value is less than KW.

The point within the interval at which the air conditioner changes states is determined by whether it is switching from "off" to "on" or vice versa. For example, a 5-minute demand value of 1.20, with KW = 2.00, indicates that the air conditioner must have run for either the first three or the last three minutes. If the previous state was "off," then we assume it stayed off for the first two minutes, turned on in the middle of the interval, and stayed on for the last three minutes; otherwise, the previous state was "on," so we assume it stayed on for the first three minutes, and was off for the last two.

These strings of "on" and "off" states are then aggregated so that all consecutive elements in the same state are reduced to a single element. This can be seen in the above example: all of the elements up to 12:02:39 are condensed into a single "off," the next three form a continuous "on" of 10 minutes 21 seconds duration, and then it stays off the rest of the day.

Only one exception condition can arise, and it is easily handled: If demand is greater than KW, we have a spike that was not removed during data cleaning; just assume demand equals KW.

Some fine-tuning was required to make the process work. For example, on many channels, a demand value equivalent to as many as three pulses could appear within a 5-minute period when the air conditioner was off. These would usually appear as a value of .06, .12, or .18 kW, respectively, because a pulse usually represents 5 watt-hours on the recorders that were installed. They were either spuriously produced by the meter, or represented the demand from a 60- or 120-watt crankcase heater. In either case, they were not part of the curtailable air conditioner demand. Consequently, we set the "threshold," or minimum demand value, to .19 kW, and treated any smaller values as zero.

Conversely, a 5-minute demand value could equal KW minus the equivalent of one or two pulses; in this case, we would assume the air conditioner was "on" for the full five minutes. Examination of the data indicated that, in most cases, the air conditioner had in fact stayed on, and a pulse or two had been "dropped." Very rarely, of course, the air conditioner will turn off just before the end, or start just after the beginning, of the five-minute period. In these cases, no harm was done by assuming it stayed on the full five minutes.

3.2 Defining the Cycling Patterns

During the summer of 1981, the LMG experienced eight basic curtailment strategies. These were prototypical patterns of enforced off-periods during the afternoon and early evening hours. However, there were a few aberrations in the strategies on certain days, bringing the total number of patterns to 13. Examples of a basic strategy and a variant are shown in Figure 2. The aberration distinguishing this variant occurs in the first three transitions of the pattern. Clearly, air conditioners following the variant pattern will not satisfy the basic pattern.

Examining the data to establish the set of cycling patterns was not very difficult, once we had converted the data to time/duration format. An example of a correctly operating and correctly recorded air conditioner is shown below in Figure 3. When the pattern-recognition program (WEEDOUT) would reject all observations in a given site on a given date, we would review the data for that site and date until we could establish the variations in the pattern. The review process rarely took very long, once we knew where to look. Of the 98 site-days when curtailment took place, only 13 exhibited variant patterns; these would generally occur in several sites on the same date.

Once we had defined the cycling patterns, we embodied this information in two SAS datasets: one containing the patterns, and one cross-referencing the patterns to the cycling days, by test site. These files were then merged with the time/duration format load data, and input to WEEDOUT.

Two iterations were required to enable us to identify all of the data errors. First, we defined the basic patterns. Then we ran WEEDOUT and used PROC FREQ to examine the number of rejected observations by site and date. Whenever an unusually large number was rejected, we would examine the time/duration format load records to identify the anomalies in the pattern. When we had defined all of the variant patterns, we reran WEEDOUT to identify the data errors on those dates.

3.3 Performing the Pattern Recognition

The pattern-recognition process in WEEDOUT compares the load data to the cycling pattern to see if any of the curtailment periods were violated. First, however, it checks whether the air conditioner was on too long during the cycling period to have curtailed properly. For the Moderate Load Following[3] strategy, for example, the

maximum duration of an "on" state between 2:00 and 8:30 p.m. is 21 minutes; any channel with a longer "on" is automatically rejected. By preprocessing the patterns and the observations to calculate the maximum "on," we were able to save on computer time during the pattern-recognition phase. Figure 4 shows an example of an observation that would be rejected for being continuously on from 16:30 to 19:00. The reference curve shos the average air conditioner load for all participants on the Moderate Load Following strategy.

WEEDOUT also checks whether the air conditioner was on at all between an hour before and an hour after the cycling period. If not, the observation is rejected into the "free rider" category: those getting a "free ride" by not contributing to curtailed load. Figure 5 gives a graphic illustration of a free rider. Clearly, there is no point in trying to compare a free-rider observation to a cycling pattern.

The observations passing the two above tests are then compared against the patterns. The pattern-recognition process checks each individual curtailment period for violations. This catches the asynchronous observations (see Figures 6 and 7) and the intermittent switch failures or radio interference (see Figure 8).

Some fine-tuning was required to permit us to use data from switches that did not quite stay open the required amount of time. Only about 50% of the valid observations adhered strictly to the pattern, 30% came back on half a minute early, and 20% came back on a full minute early. There was a sharp drop-off at this point: increasing the grace period by another half minute would only have increased the sample size by about 1%.

4. SUMMARY

Minimax developed an innovative pattern-recognition process that would detect a wide range of unanticipated data anomalies. Both the development and the subsequent implementation required combining SAS's capabilities for efficiently manipulating data strings with its capabilities for statistical analysis. Depending on SAS for these fundamental tasks allowed us to concentrate on the analytical issues which were central to our problem. This pattern-recognition process can be used in a variety of applications where accurate representations of usage patterns in time series data are critical.

The translation of the load data from 5-minute data format to time/duration format and the subsequent cleaning of the data by WEEDOUT were crucial steps in the data cleaning task. The pattern-recognition process partitioned the load data into two groups: a dataset of properly cycled and recorded air conditioner loads, and a data set of observations showing failures as described in Section 1. This permitted the preliminary analysis of the program to be undertaken concurrently with the examination of the problem load data.

Figure 2: Examples of Basic and Variant Cycling Patterns

Basic			Variant		
Time of Day	On/ Off	Duration	Time of Day	On/ Off	Duration
12:00	Off	15	12:02	Off	8
12:15	On	15	12:10	On	10
12:30	Off	15	12:20	Off	25
12:45	On	15	12:45	On	15
13:00	Off	15	13:00	Off	15
(Repeats every 15 minutes until 17:45)					
17:45	On	--	17:45	On	--

Figure 3: Example of Correct Curtailment (Low Load Following Strategy)

Time of Day	Duration (Min.)	On/ Off
13:07.5	12.5	ON
13:20.0	6.5	OFF
13:26.5	11.5	ON
13:38.0	7.0	OFF
13:45.0	13.5	ON
13:58.5	14.0	OFF
14:12.5	6.0	ON
14:18.5	6.5	OFF
14:25.0	5.0	ON
14:30.0	12.0	OFF
14:42.0	18.0	ON
15:00.0	15.0	OFF
15:15.0	15.0	ON
15:30.0	15.0	OFF
15:45.0	15.0	ON
16:00.0	15.0	OFF
16:15.0	15.0	ON
16:30.0	15.0	OFF
16:45.0	15.0	ON
17:00.0	15.0	OFF
17:15.0	15.0	ON
17:30.0	15.0	OFF
17:45.0	15.0	ON
18:00.0	15.0	OFF
18:15.0	15.0	ON
18:30.0	11.5	OFF
18:41.5	18.5	ON
19:00.0	12.0	OFF
19:12.0	18.0	ON
19:30.0	10.0	OFF
19:40.0	20.0	ON
20:00.0	10.0	OFF
20:10.0	20.0	ON

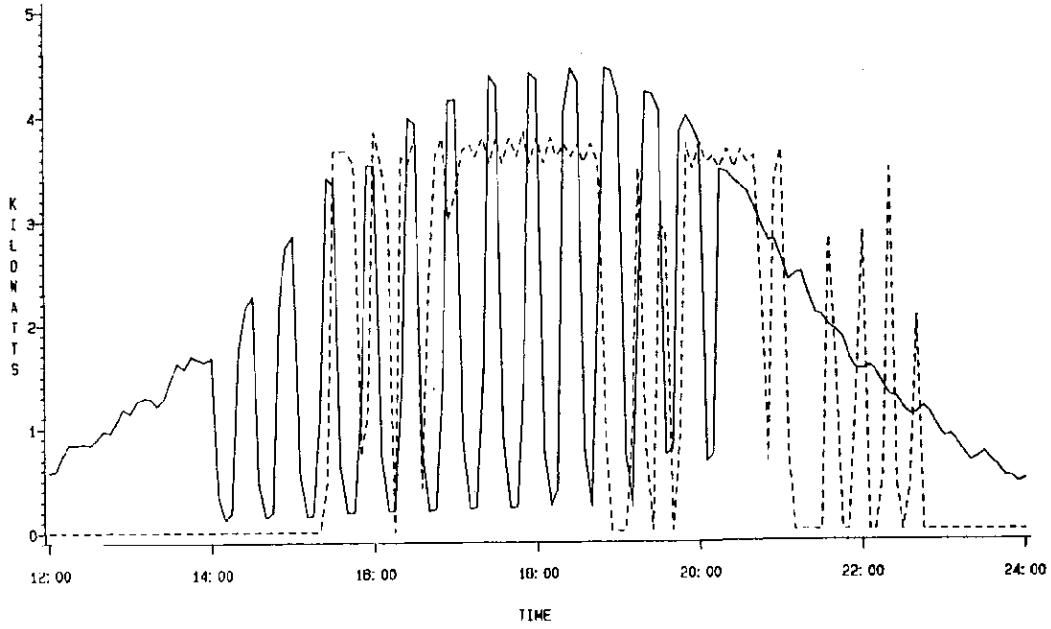
Notes:

[1] It is, however, an important problem for those responsible for installing and maintaining the equipment.

[2] Air conditioners and water heaters do not turn rapidly on and off. In fact, air conditioners can be damaged by being turned on within five minutes of being turned off.

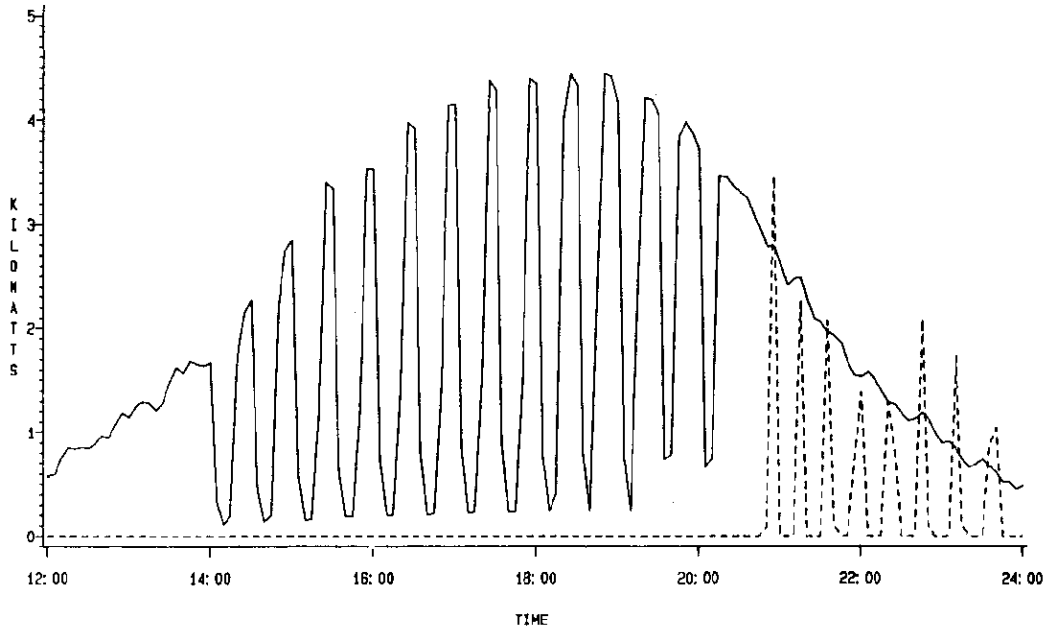
[3] The Moderate Load Following strategy is a particular pattern of off-periods varying from 9 to 18 minutes off per half-hour.

Figure 4 ILLUSTRATION OF AN UNCYCLED AIR CONDITIONER LOAD CURVE COMPARED TO MODERATE LOAD FOLLOWING STRATEGY



SOLID LINE=CYCLED, BROKEN LINE=UNCYCLED

Figure 5 ILLUSTRATION OF A FREE-RIDER AIR CONDITIONER LOAD CURVE COMPARED TO MODERATE LOAD FOLLOWING STRATEGY



SOLID LINE=PROPERLY CYCLED, BROKEN LINE=FREE-RIDER

Figure 6: Example of Asynchronous Operation

Time of Day	Duration (Min.)	On/ Off
11:00.0	56.5	ON
11:56.5	8.5	OFF
12:05.0	10.0	ON
12:15.0	25.0	OFF
12:40.0	15.0	ON
12:55.0	15.0	OFF
13:10.0	15.0	ON
13:25.0	15.0	OFF
13:40.0	15.0	ON
13:55.0	15.0	OFF
14:10.0	15.0	ON
14:25.0	15.0	OFF
14:40.0	15.0	ON
14:55.0	15.0	OFF
15:10.0	15.0	ON
15:25.0	15.0	OFF
15:40.0	15.0	ON
15:55.0	15.0	OFF
16:10.0	15.0	ON
16:25.0	15.0	OFF
16:40.0	15.0	ON
16:55.0	15.0	OFF
17:10.0	15.0	ON
17:25.0	15.0	OFF
17:40.0	80.0	ON

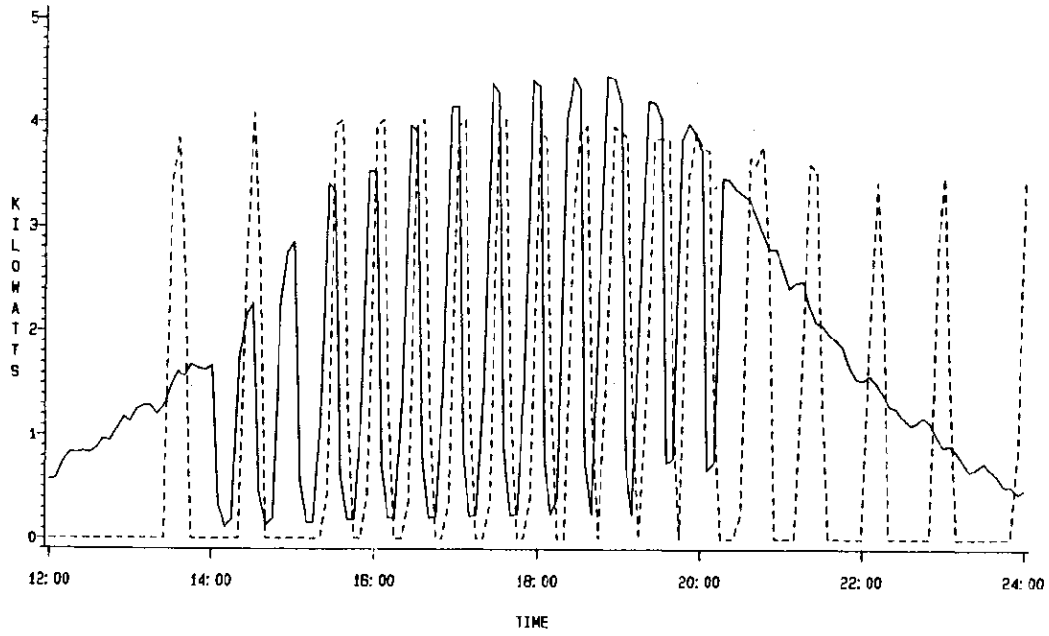
This channel appears to have lost a timing pulse. It follows Variation 1 of the 15-Minute pattern, but it "anticipates" the transmissions by 5 minutes.

Figure 8: Example of Radio Interference

Time of Day	Duration (Min.)	On/ Off
14:15.0	15.0	ON
14:30.0	15.0	OFF
14:45.0	15.0	ON
15:00.0	25.0	OFF
15:25.0	15.0	ON
15:40.0	15.0	OFF
15:55.0	15.0	ON
16:10.0	17.0	OFF
16:27.0	13.0	ON
16:40.0	17.0	OFF
16:57.0	18.0	ON
17:15.0	12.0	OFF
17:27.0	13.0	ON
17:40.0	17.0	OFF
17:57.0	13.0	ON
18:10.0	17.0	OFF
18:27.0	13.0	ON
18:40.0	17.0	OFF
18:57.0	13.0	ON
19:10.0	15.0	OFF
19:25.0	15.0	ON
19:40.0	10.0	OFF
19:50.0	20.0	ON
20:10.0	10.0	OFF
20:20.0	70.0	ON

This channel was being cycled according to Variation 2 of the Moderate Load Following pattern. At 16:57, the channel should have been allowed to come "on" for only 12 minutes. It missed the "off" signal, however, and so stayed "on" for 18 minutes. Thereafter, it followed the pattern.

Figure 7 ILLUSTRATION OF AN ASYNCHRONOUS AIR CONDITIONER LOAD CURVE COMPARED TO MODERATE LOAD FOLLOWING STRATEGY



SOLID LINE=PROPERLY CYCLED, BROKEN LINE=ASYNCHRONOUS