SAS'S VITAL ROLE IN A STATEWIDE FUEL CONVERSION AND ENERGY CONSERVATION STUDY

E. N. Caldwell

West Virginia Network for Educational Telecomputing
Associate Professor, West Virginia University

ABSTRACT


SAS was used extensively for the data management and analysis of the four separate databases in the first comprehensive coal conversion study for residential, commercial, and institutional buildings in the United States. SAS was very instrumental in the success of the project despite formidable handicaps.

INTRODUCTION

The severe winter of 1976-1977 led to substantial fuel curtailments within the State of West Virginia and reduced the availability of fuel oil. These conditions prompted the State Legislature of West Virginia, an abundant coal-producing state, to commission a feasibility study by West Virginia University's College of Engineering on the conversion of coal as the primary heating fuel for the public buildings throughout the State. "For the purpose of this survey, a public building is defined as any structure with a roof, built by tax money, and not owned by the Federal Government."1

A short term, but intensive investigation and analysis of the State's 6,000 public buildings was performed by two interdisciplinary teams established under the direction of the Engineering Experiment Station. The data team with five members was concerned with collection and processing. The major areas addressed by the eleven team members of evaluation and analysis included Regents buildings, individual buildings, classification, economics, conservation (thermodynamical), and conversion which was further broken down into hardware, auxiliaries, coal availability and transportation, and environmental impact.

1Study sponsored by the West Virginia University Engineering Experiment Station on behalf of the State of West Virginia Legislature.

DATA ORGANIZATION

Early in the study the decision was made to organize the information into four separate, but interrelated, databases, rather than one. These four were: Administrative, which was the Master; Monongalia County, which was the Pilot Study; Campus Buildings, which included university and college buildings; and Schools, which included the secondary schools. Their respective attributes are summarized below in Table I.

Table I. Table of Data Bases and Properties

<table>
<thead>
<tr>
<th>BASE</th>
<th>ADMN</th>
<th>MON. CO.</th>
<th>CAMPUS</th>
<th>SCHOOLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROPERTY</td>
<td>Master</td>
<td>Pilot</td>
<td>Colleges</td>
<td>Schools</td>
</tr>
<tr>
<td>Contents</td>
<td>Buildings</td>
<td>6,000</td>
<td>682</td>
<td>1,310</td>
</tr>
<tr>
<td>Record Size*</td>
<td>146</td>
<td>240</td>
<td>560</td>
<td>240</td>
</tr>
<tr>
<td>Base Size+</td>
<td>76,000</td>
<td>10,080</td>
<td>383,040</td>
<td>314,400</td>
</tr>
<tr>
<td>Variables</td>
<td>35</td>
<td>60</td>
<td>120</td>
<td>60</td>
</tr>
<tr>
<td>Storage</td>
<td>disk cards</td>
<td>cards hypertape</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heated Area</td>
<td>100%</td>
<td>--</td>
<td>20%</td>
<td>46%</td>
</tr>
<tr>
<td>Collection</td>
<td>on-site</td>
<td>on-site</td>
<td>question</td>
<td></td>
</tr>
<tr>
<td>Primary</td>
<td>existing</td>
<td>survey</td>
<td>survey</td>
<td>hairs</td>
</tr>
</tbody>
</table>

*in bytes
+in square feet

The Administrative database provided a building census containing a record for every public building in the State, whether or not it was actively studied. The 6,000 records in this base, encoded geographically, administratively, and functionally, included, but was not limited to, information such as age, square footage, fuel type, and consumption, and maintenance supervisor. The developmental time frame ran from the beginning of April until the end of September. The basic information, taken from existing records, was supplemented by mail questionnaires and phone calls.

The Monongalia County database contained more detailed technical information on the 42 buildings examined initially in the pilot study. The items collected were largely those ultimately collected for the public schools and were patterned after the PSECS2 format. The developmental time frame for this base ran from mid-May until the end of June.


3PSECS, Public Schools Energy Conservation Service.
The Campus buildings database, containing the 684 state college and university buildings, was the most detailed and voluminous. The information gathered in on-site surveys by specially trained teams, included, but was not limited to, energy consumption, heating system, lighting, ventilation, insulation, water treatment, building layout, and use. The statewide computer network proved useful here as survey teams transmitted data and reports to and from various locations around the State. The developmental time frame ran from mid-June until mid-August.

Figure 1 demonstrates the interaction between the bases and their respective programs.

SAS'S ROLE

SAS was responsible for the success of the data team on this study, probably the most complete study of its kind in the country, despite some very formidable handicaps.

First was the extremely short time frame, only six months from inception to report, which did not allow adequate design or development time. Answers and reports were frequently needed within a few hours of the requests.

As the developmental time frame on each base grew progressively shorter the portion of the system done using SAS increased dramatically. Figure 2 demonstrates the relationship between the development time available and the percentage of work done with SAS for each base. The ADMN base spanned a period of six months and SAS and PL/I were used about equally. The PL/I programs were designed at the beginning of the study and included a major edit program and programs to translate and reformat existing records. However, this assignment alone tied up the programmers involved for about two months which was over one-third of the total project time, and infringed on developmental time for the next base.

There was never any question that anything but the SAS PROC UPDATE would be used as the update program. It was necessary to be able to pick up any combination of variables from one or two cards of input and update the corresponding fields in the database without rekeying existing valid information. To have accomplished this by the target date in another language would have been impossible for us, because at this time the developmental effort was being devoted to the edit and transformation program. All subsequent reporting, analyses, and statistical profiles done on the bases were in SAS.

A very valuable program was the preview program which printed pre-processed record images for verification before actual updating. This proved to be very useful because results from the update run took overnight for printing. Any error necessitated another update run for correction, putting us into the second day. The preview program, since it printed out only the updated records, received much shorter turnaround time.

SAS also was used in the ADMN base with GEORZP to generate the latitude, longitude, and county code from the existing zip code.

A special SAS macro library5 (distinct from the system macro library) and corresponding catalogued procedure were established and proved very valuable. The macros included defined the input, translated county code numbers into names, made classifications, and gave degree days for specific geographical locations.

Some preliminary work on the Public Schools base was attempted in PL/I. A major effort here involved transforming and reformatting the "raw" school data from four card format into the PSECS three card format. After running long past the deadline and producing no usable results, a five week PL/I effort was abandoned. In a few days a SAS program was coded which served the same purpose and even did preliminary screening.

The lack of data processing professionals proved to be a major handicap. Students were used to do the programming and typically took a couple of months to develop a PL/I program. Of far greater impact were weaknesses in the design specified by the investigator themselves. The largest problem was the establishment of four separate databases, each in a different format, with different variables, coding conventions, and categories of classification. Thus, information had to be run through a preprocessor as a preliminary step to common analysis programs, such as for heat loss. This also unnecessarily complicated the interchange of information between bases for updating.

Another design problem was the fact that numeric fields were coded with dashes under certain circumstances and alphabetic fields also were coded with dashes under different circumstances. The PL/I preupdate translate programs required additional programming to handle this. The Fortran programs abandoned.

Unfortunately, since the number of buildings for the Campus Buildings base was small, it was felt that the time necessary to develop edit programs was not warranted; rather editing and sorting would be done manually. The price, of course, was very high. The amount of time consumed by rewriting analysis programs and/or writing preprocessors programs to compensate for

4Borbash, Steven R., West Virginia Map Construction and Geolocational Analyses with SAS and a Georzip File, West Virginia University, Morgantown, WV, June, 1976, 27 pp.

5Changes to the SASSAMACRO catalogued procedure, Academic Services Instruction #42.74.5, West Virginia Network for Educational Teleprocessing, Morgantown, WV, August 1, 1977, 2 pp.
the bad data was probably three times more than it would have taken to develop the edit program. (And still there was a very high level of dissatisfaction with the data.) An extensive preprocessor was written, almost exclusively with SAS, as a preliminary step between the Campus Buildings base and the heat loss analysis program. After the data was filtered (in PL/I) the up to seven cards per observation were sorted, tested for lost cards and critical missing values, transformed, and screened. Only a clean, usable set of data was then passed to the sensitive heat loss analysis program.

Another major handicap was, of course, the ubiquitous poor and incomplete data. Ultimately, only twelve percent of the records were complete enough to run through the key Fortran heat loss analysis program and even less through the PSECS program. SAS played a key role in preprocessor programs to screen incomplete records which abended Fortran programs. A decision had been made to run the standard PSECS program, as opposed to modifying it to handle less than perfect, complete cases. The flexibility that SAS gave us here was invaluable. The results of PROC FREQ were also used to detect invalid data values.

SAS was used in each base to generate investigative and final profiles, descriptive statistics, regression equations, and to plot variables. All of these functions were done on the whole bases as well as various subsets which were being continually defined and redefined as the study progressed. SAS's flexibility in subsetting on variables and combination of variables allowed the investigators much more latitude than they had supposed was possible. Results from SAS programs, for example, PROC FREQ, were passed to Fortran coded CalComp programs to generate plots.

The obstacles to using SAS for this study included its reputation (among some engineers) as a statistical package. Some investigators felt that it could not handle the engineering type of problems that Fortran could; that it had a reputation as being good merely for "quick and dirty" requests. Some felt that their work warranted specially designed programs and output. Another obstacle was the cost — our final computer costs were more than three times the initial estimate. Many factors (some of which have been described in this paper) contributed to this; SAS's overhead was certainly one. Yet without it, even the skeptics agree, the completion date could not have been met without significantly more manpower than was available for the task.

In conclusion, I believe that this project would have taken three to four times as long if SAS had not been used and much flexibility would have been sacrificed. In many cases where other kinds of programs were developed, SAS could have done the job, often better, and with far less development time.
Figure 1. System interaction between databases
Figure 2. Relationship between length of developmental time and use of SAS for each database.