ORNL/TM--9421

DE91 004170

ENGINEERING PHYSICS AND MATHEMATICS DIVISION

NONRESIDENTIAL BUILDINGS ENERGY CONSUMPTION SURVEY (NBECS): STUDY TO DEVELOP REGRESSION MODELS TO IMPUTE MISSING ELECTRICITY AND NATURAL GAS CONSUMPTION VALUES

D. M. Flanagan, Project Manager
H. J. Tsao
R. L. Schmoyer, Jr.
J. M MacDonald

Date Prepared: January 1985
Date Published: October 1990

Research sponsored by
Office of Energy Markets and End Use,
Energy End Use Division, Energy Information Administration
U.S. Department of Energy
Washington, D.C.

Prepared by the
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831
operated by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
for the
U.S. DEPARTMENT OF ENERGY
under Contract No. DE-AC05-84OR21400

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF TABLES (continued)

ACKNOWLEDGMENTS

The authors wish to extend a very special thanks to Toby Mitchell and Terri Vineyard whose helpful comments in the areas of statistics and engineering, respectively, improved the quality of this document. A special thanks also goes to the Energy Division technical editors who made this report appealing to the eyes.

ABSTRACT

Imputation procedures were designed for the 1983 Nonresidential Buildings Energy Consumption Survey (NBECS) of the Energy Information Administration (EIA) using 1979 NBECS data. The study included methodology development, data analysis, regression analyses, empirical evaluations of the regression models, and imputation procedures. Models considered were engineering models, stepwise regression, weighted regression, nonlinear regression, and log transformation regression. A method for determining the appropriateness of the imputation model for a particular set of independent variables is recommended.

Although this study was completed in 1985, this final version of the report is being issued due to continuing requests for information.

EXECUTIVE SUMMARY

This report documents a study to review and design imputation procedures for the Energy Information Administration's (EIA) 1979 Nonresidential Buildings Energy Consumption Survey (NBECS). Procedures were designed for buildings using electricity and natural gas and which reported a consumption of less than 31 days. The methodology was to be used by the EIA for its 1983 NBECS, which uses an almost identical survey and sample population.

The study included methodology development, data analysis, regression model analyses, empirical studies of the regression model, and imputation procedures. Highlights from each area are presented.

Methodology

1. Reproduce and study residuals of regression models previously produced by the EIA.
2. Study data base to find more potential independent variables for regression models.
3. Let an engineering approach also suggest a set of predictor variables.
4. Based on the above, develop an initial functional form.
5. Assess predictive capability of models, addressing outliers, negative imputed values, appropriate ranges for imputation, and imputation error.
6. Define imputation procedures.

Data Analysis

7. Edits such as the filed, logic, and range edits can provide benefits such as: (1) detecting potential errors; (2) developing rules for contacting respondents and correcting errors; (3) automating processes and reducing manual time and human error, and (4) reducing costs.
8. Variables important to the regression equation should be screened carefully for missing values: these values (usually square footage and number of workers) should be obtained rather than imputed.
9. Use the regression model to detect erroneous values that need to be resolved.

10. A previous-value check of important regression variables can be made for the 1983 NBECS, especially if the 1979 value is to be substituted for a missing 1983 value.

Regression Model Analyses

11. Study of previous EIA regression models showed variables important for different building types, but also revealed negative imputed values and non-normal error terms. Residual analysis indicated that neither polynomial nor weighted least squares techniques were applicable.

12. All of section 5.2 provides information, tabled as much as possible, to show the structure of the NBECS data and to indicate important predictor variables.

13. Section 5.3 provides an engineering model and its justification, based on the ASHRAE Handbook.

14. The regression model development of section 5.4 recommends the log transform analysis because it provides a highly significant fit to the data in every building category and has the best residual plots. This model also provides a non-negative predicted value when back transformed to the original value.

Empirical Study of the Regression Model

15. The nonrespondents and respondents need to follow approximately the same response surface. The recommended ORNL models provided estimates that are 1.2% away from the actual consumption value; the weighing adjustment method is 7.6% away.

Imputation Procedures

16. A detailed flow chart is provided and discussed in section 7.1, showing how to use the ORNL models in the total imputation procedure. Recommended ad hoc procedures to be used when regression adjustment is inappropriate include sample weight adjustment (first choice) and lower bound estimation.

17. Section 7.2 notes that an imputation model should be used only for data that might be expected to have an imputed value within the range of data used to calculate the imputation model. A method for doing this is recommended and requires quadratic programming software.

1. INTRODUCTION

In 1979, the Energy Information Administration conducted the first Nonresidential Buildings Energy Consumption Survey (NBECS) to collect data about commercial energy consumption. That survey provided valuable first-time data about several aspects of energy use in commercial and other nonresidential buildings: energy consumption, conservation activities, energy end use, and heating and cooling characteristics. The 1979 sample of 6,776 buildings was drawn to represent all commercial buildings in the contiguous United States and is used primarily to calculate energy consumption totals for the EIA report, Nonresidential Buildings Energy Consumption Survey: 1979 Consumption and Expenditures (1983).

The two primary fuels used in nonresidential buildings are natural gas and electricity; however, some or all of the data for 1979 consumption of these fuels were not available for 38% of those buildings using electricity and 27% of those using natural gas. The ability to provide good imputed values became very important, but it was especially difficult to impute totally absent data. The Office of Energy Markets and End Use, within the Energy End Use Division of EIA, imputed consumption using a multiple linear regression model of consumption vs. selected building characteristics and climatic conditions. Although this complex, time-consuming procedure produced good results from some building types, it was not satisfactory for all of them.

This imputation study, conducted at Oak Ridge National Laboratory (ORNL), investigates alternative imputation procedures, including better regression models and, possibly, hot-decking methods. Simpler imputation methods, if available, are desirable. Imputation procedures are restricted to imputing electricity and natural gas consumption for those building respondents with a reported consumption period of at least 31 days. The procedures are based on results obtained from building respondents with reported consumption periods of at least 331 days. The methodology used to develop these models will be used by EIA to calculate imputation models for the 1983 NBECS data. The methodology should be applicable to subsequent data bases because most of the buildings included in the 1983 and other surveys were included in the 1979 survey, and because the NBECS questionnaire has had few changes.

This report documents the development of the study and presents some interesting results concerning the data base, data structure, and inference--as well as those results directly related to the recommended imputation procedures. Section 2, entitled "Background," summarizes previous studies and procedures developed by the EIA for the 1979 NBECS data base. Section 3, "Methodology," discusses the analytical approach and provides additional references. Section 4, "Data," describes the EIA data base, the data bases created by ORNL personnel, edits, and questionable observations. The results of the previous EIA models and the models developed by ORNL are presented and discussed in detail in Section 5. An evaluation of the regression models is documented in Section 6, and Section 7 describes the imputation procedure. Conclusions form the study are discussed in Section 8.

The many variable names in the data set are defined in Tables 5.7 and 5.10 and in subsections 5.2.1, 5.2.2, and 5.3.

2. BACKGROUND

The NBECS survey is the first attempt at collecting nonresidential buildings' characteristics and fuel consumption data from a national statistical sample. The NBECS publications are to be used by government agencies at the federal, state and local levels, as well as by representatives from the private sector who are concerned with buildings' energy consumption for forecasting, modeling, and analysis. Buildings that were primarily residential but showed evidence of commercial or industrial activities are also within the scope of this survey. The information was collected through personal interviews with building representatives between October 1979 and January 1980. A description of the survey design, the data collection procedures, and the techniques used to convert the sample data to national estimates is found in the EIA report, *Nonresidential Buildings Energy Consumption Survey: 1979 Consumption and Expenditures, Part 1: Natural Gas and Electricity* (1983).

The two most important fuels used in buildings across the country, natural gas and electricity, are the two fuel types described in this report. As in most national energy consumption surveys, the NBECS is subject to various sampling and nonsampling errors when the sample information is used to estimate the total consumption of a particular fuel type. One major concern of the nonsampling error is the "item nonresponse," where buildings in the NBECS sample provide some, but not all, of the requested information, or where the information provided for some items is not usable. Item nonresponses for selected building characteristics, such as square footage and number of employees, were treated by imputing data from a reservoir of responding cases, using a hot-deck procedure for one of a combination of two items. To impute the square footage and number of employees, the EIA used a simultaneous hot-deck procedure.

One of the major goals of the NBECS was to produce estimates of natural gas/electricity consumption in the nonresidential sector during calendar year 1979. To accomplish this, the EIA collected consumption data from electricity and natural gas suppliers. However, the item nonresponse problem on the consumption data amounts to 27% of the natural gas records and 38% of the electricity records. There were two main causes for the incompleteness of the building consumption data: (1) a waiver was not obtained from the building representative interviewed; or (2) the utility billing records were either missing or not usable for all or part(s) of 1979. There were other problems with respondent errors that caused incomplete fuel consumption values. These included the following:

1. Utility billing or meter reading data rarely started in January 1, 1979, and ended on December 31, 1979. Most cases of complete reporting of 1979 data included billing periods that overlapped into 1978 and 1980.

2. For a given building and in a given billing period, billing or meter reading data covered more than (or less than) the actual fuel consumption. In this case, the interviewer may or may not have known of the existence of the error.

Preliminary procedures for adjusting item nonresponses, respondent errors, and other types of nonsampling errors were developed and documented by the EIA. To facilitate the EIA investigation, utility records were classified into three groups: (1) records whose period of consumption covered less than or equal to 30 days in 1979; (2) records whose periods of consumption covered 31-330 days in 1979; and (3) records whose period of consumption covered 331 days or more in 1979. After that, a separate imputation procedure was devised to impute fuel consumption for each of the three groups in order to estimate the total 1979 fuel consumption. Simple statistical adjustments were made to impute incomplete records in Groups 2 and 3. Linear regression procedures were developed for imputing missing consumption records in Group 1 and to meet the data needs for the first NBECS publication. However, the EIA believed that the imputation methodology developed for the Group 1 records was preliminary, and they were not completely confident about the results obtained.

The field work for NBECS II (the second NBECS update)-which collected energy consumption data for 1982 and 1983 from the buildings' energy suppliers-was completed in 1984. In addition to the NBECS II, planning began for a new, more effective buildings survey to be fielded in 1986. Although the respondents may change because of growth, decline, personnel changes, or mergers, the item nonresponse problems on fuel consumption will remain at a certain level for the Group 1 records. Therefore, possible methods of improving the previous imputation procedures on Group 1 records for natural gas and electricity consumption will directly benefit the analyses of the two successive NBECS data reports. Therefore, as stated in the Introduction, one of the tasks of this ORNL-conducted imputation study, is to investigate alternative imputation procedures for the Group 1 records.

3. METHODOLOGY

The task description for this project specifies the development of regression models to impute missing electricity- and natural gas-consumption values for the 1979 NBECS data as a means of developing imputation methodology for the 1983 NBECS data. The observations to be imputed are those building respondents with a reported consumption period of less than 31 days. The observations to be used for modeling include all respondent data of buildings with a reported consumption period of at least 331 days. The methodology used to develop these models will be used to calculate imputation models for the 1983 NBECS data. The methodology should be applicable to the new data base because (1) most of the buildings included in the 1983 survey were included in the 1979 survey; and (2) the NBECS questionnaire has had few changes.

The NBECS sample data are used primarily to calculate energy consumption totals for the Energy Information Administration (EIA) Report, *Nonresidential Buildings Energy Consumption Survey: 1979 Consumption and Expenditures* (1983). Several statistical areas of study are of great interest and value with respect to the imputation problem and its impact on survey estimates from statistical samples - such as variance estimation, impacts on weights, total estimates, and cell and total biases. However, this study concentrates on developing regression models that are logically sensible (with respect to the variables used and functional form) and display normally distributed (or at least symmetric) error terms that are as small as reasonably possible. In addition, limited analyses of the models' imputation preformances and some recommended imputation procedures are discussed.

This section discusses the statistical methodology used to conduct the study. The general approach considers four areas. First, the regression models previously developed by the EIA were reproduced to study the residuals and to learn from that prior experience. Second, the variables in the NBECS data base were studied to include more factors (or independent variables) that would be of potential importance. Next, an engineering approach to the model development suggested a set of predictor variables. The regression diagnostics from the EIA models, the data structure study, and engineering modeling were used to develop an initial functional form, which was then tested and revised until a final form and algorithm emerged. Finally, the predictive ability of the models was assessed. This assessment addressed negative imputed values and other outliers, the appropriate range for imputation, and the model imputation error. Imputation procedures were also defined and assessed. Each of these areas is discussed in detail.

As discussed in Section 2, EIA personnel previously developed ordinary least squares (OLS) regression models for imputing missing electricity- and natural gas-consumption values for the 1979 NBECS data base. The analyses were programmed using the Statistical Analysis System (SAS), and printouts from the regression procedures (which included a list of the observed, predicted, and residual values) were available to ORNL personnel. The 26 building regression categories for the natural gas models and 18 for the electricity models and associated building activity class codes are listed in Appendix A. To build on and learn from these previous models and, in particular, to understand the model error terms, the EIA models were reproduced using the 1979 NBECS data base described in Sect. 4. Observations (buildings) designated as outliers by EIA are discussed in Sect. 4. Additional observations deleted as a result of the ORNL data screening effort are also discussed in Sect. 4. Regression diagnostics included checking the normality of the residuals (i.e., stem-and-leaf plot, skewness, and kurtosis); analyzing the residual plots described by Draper and Smith (1981); and checking for influence, multicollinearity, autocorrelation, etc. The results of the regression diagnostics for each model are summarized in Sect. 5.

The structure of the NBECS data set was studied to learn about data dependencies and to include more factors that could be important, especially from an engineering standpoint. The variables used previously are noted in the EIA report, *Nonresidential Buildings Energy Consumption Survey: 1979 Consumption and Expenditures* (1983). ORNL's Efficiency and Renewables Research Section of the Energy Division--which is dedicated to performing engineering studies concerning the energy consumption of residential and commercial buildings and appliances--suggested additional independent variables for regression equations, based on their own engineering studies and those of others in that field. The study also involved checking the distribution of the independent variables and calculating frequencies for conservation variables. These analyses are documented in Sect. 5.

The ORNL imputation team again drew on the expertise of the ORNL Efficiency and Renewables Research Section for the development of a regression model based on engineering principles. For example, the conceptual heating model can be represented as:

$$Q_{heat} = Q_C + Q_D - Q_I - Q_E + Q_P \ , \qquad (3.1)$$

where

- Q_{heat} = total fuel consumption for space heating;
- Q_C = heat lost from the building envelope by conduction and convection;
- Q_D = heat required to warm outside air that has infiltrated the building through doors, crevices, etc.;
- Q_I = heat from internal gains such as lighting, equipment, occupants, etc.;
- Q_E = heat from external gains such as sun and wind; and
- Q_P = heat required to protect the building's contents (e.g., water pipes) when empty.

More detailed engineering models for heating and cooling from which the initial regression models are derived follow.

The heating model is represented as:

$$Q_{heat} = \frac{24 \times HDD65}{E_H} \times \Bigg\{ \underbrace{(U^* \times Ae)}_{Q_C} \times [1 - \underbrace{(B_W \times \frac{G}{100})}_{Q_E}] \quad (3.2)$$

$$+ \underbrace{(AV_W \times \rho \times C_P \times A_W)}_{Q_D} + \underbrace{(\lambda \times S \times \mu \times A_W) - (P \times \mu)}_{Q_I} \Bigg\} ,$$

where

U^* = conductivity and convection,
Ae = surface area of envelope,
AV_W = air volume of outside winter air per square foot (ft^2) of heated floor area,

$\rho \times C_P$ = conversion from $\frac{ft^3}{hr - ft^2}$ to $\frac{Btu}{hr - {}^{\circ}F - ft^2}$,

A_W = heated floor area,
B_W = solar fraction heating reduction,
G = percent glass,
$\lambda \times S$ = estimate of internal loads other than people (i.e., lighting and equipment),
μ = occupied fraction = hours operation per week divided by 168 hours per week,
P = number of employees,
$HDD65$ = number of heating degree days using 65°F as the basis, and
E_H = efficiency of the heating equipment.

Q_p is part of the heating component as defined by Eq. 3.1 but is not included in Eq. 3.2 because of a lack of appropriate data.

The cooling model is represented as:

$$Q_{cool} = \frac{24 \times CDD65}{E_C} \times \Bigg\{ \underbrace{(U^* \times Ae)}_{Q_C} \times [1 + \underbrace{(B_S \times \frac{G}{100})}_{Q_E}] \quad (3.3)$$

$$+ \underbrace{(AV_S \times \rho \times C_P \times A_C \times \Delta T)}_{Q_D} + \underbrace{(\lambda \times S \times A_C) + P}_{Q_I} \Bigg\} ,$$

where

CDD65 = number of cooling degree days using 65°F as the basis,
B_S = solar fraction cooling increase,
E_C = cooling equipment efficiency,
AV_S = air volume of outside summer air per square foot of cooled floor area,
A_C = cooled floor area, and
ΔT = design temperature difference.

All other variables are as defined for heating in Eq. 3.2.

The total annual fuel consumption (Q_{TOTAL}) for electricity or natural gas can be represented by:

$$Q_{TOTAL} = Q_{heat} + Q_{cool} + Q_{INTERNAL} + Q_{EXTERNAL} \quad , \tag{3.4}$$

where

$Q_{INTERNAL}$ = Fuel needed for internal uses (such as cooking and manufacturing) that affect the internal heat component Q_I; and
$Q_{EXTERNAL}$ = Fuel required for external uses (such as water heating, electricity generation, outdoor lighting, and other uses) that do not affect the internal heat component Q_I.

Other details of the engineering regression models are explained in Sect. 5.

On the basis of the diagnostics for the EIA OLS models and the data base structure study, the ORNL study team determined whether OLS or other linear modeling methodology, such as transformed data or weighted least squares (WLS), was applicable, or whether nonlinear models were necessary. The need for WLS was assessed via residual analysis methods described by Draper and Smith (1981), Neter and Wasserman (1974), and others. Independent variables, if any, contributing to the WLS problem were identified via plots of the residuals vs that independent variable. Particular weighting schemes considered were based on variance-proportional-to-mean and variance-proportional-to-squared-mean models.

Nonlinear models were analyzed by either the Marquardt or Gauss-Newton method in SAS and started with models of these forms,

$$(1) \quad \log(Y) = \log \Psi \, (\alpha + \beta_1 X_1 + \ . \ . \ . + \beta_P X_P) + \epsilon \ , \text{ and} \tag{3.5}$$

$$(2) \quad \sqrt{Y} = \sqrt{\Psi(\alpha + \beta_i X_1 + \cdots + \beta_p X_p)} + \varepsilon . \qquad (3.6)$$

where p is the number of independent variables and ψ is a function with continuous derivative and positive lower bound. The choice of lower bound is dependent on the fuel type.

Linearizations of the nonlinear models were also considered. A linearization for (3.5), for example, is based on the approximation

$$\log(\alpha + \sum_{j=1}^{p} \beta_j X_j) = \log \alpha + \frac{1}{\alpha} \sum_{j=1}^{p} \beta_j X_j . \qquad (3.7)$$

The nonlinear, WLS, and all other models were checked with the same residual diagnostics applied to the EIA OLS models. The final models were selected based on normality and size of the error term, good model diagnostics, simplicity of form, and quality of predicted values. The details of the development for each model are explained in the results of Section 5.

A performance assessment was made to test the regression models and is discussed in Section 6. A percentage of the modeling group (approximately equal to the percentage requiring imputation for the building category) was reserved, and the model was recalculated without these reserved observations. Imputed values were then estimated for the reserved observations via the recalculated models. The predictive performance was assessed by calculating the percentage difference between an estimated total energy consumption using the ORNL model and the actual total energy consumption of the test population.

The state-of-the-art efforts on treating incomplete data as a general survey problem were organized by the Panel on Incomplete Data established by the Committee on National Statistics within the Commission on Behavioral and Social Sciences and Education of the National Research Council in 1977. With funding provided by National Science Foundation, the Social Security Administration, and the U.S. Department of Energy, a report was published in three volumes and titled "Incomplete Data in Sample Surveys" (1983). The ORNL research team has reviewed portions of the report to draw upon experiences and case study results from this current research effort.

The imputation procedures define the rules for each building type and address the applicable range of each regression model and the handling of negative imputed values and other outliers. The rules for model usage are based on one of several methods, such as discriminant analysis, variance of predicted values, and check to see if imputation involves interpolation or extrapolation. These analyses will determine whether the independent variables for the imputation groups are within ranges similar to those for the modeled groups. In all cases, alternate procedures were developed when the population needing imputation was statistically different from the population modeled, in specific ways to be discussed, regardless of the fit of the model. Hot-decking procedures as described by Sande (1982) and by Madow, Olkin, and Rubin (1983) will not be appropriate for observations that are out of range.

Variance estimates for predicted (imputed) values in regression models are straightforward to compute. Usually large variances (e.g., with respect to the prediction itself) indicate the inability of the model to impute accurately, even if the model itself is correct.

Discriminant analysis methods--such as those described by Tatsuoka (1971)--are appropriate if the independent variables are normally distributed. Analyses summarized in Sect. 5 will verify or negate this assumption. The nonparametric nearest-neighbor method described by Cover and Hart (1967), Fix, Evalyn, and Hodges (1959), and Hand (1981) is inappropriate. This method requires two "known" parent populations (imputed and modeled) and two samples (imputed and modeled) to be classified on the basis of a nearest-neighbor match with the two parent populations. This procedure would be possible if more data were available, especially for the imputed group. Therefore, if the independent variables are not normally distributed, the discriminant analysis will be omitted.

On the other hand, the model may not be exactly correct, but it may serve as an adequate approximation over the range of the data used to estimate the model parameters. Imputation outside this range (i.e., extrapolation) should then be avoided. Extrapolation occurs when the vector of independent variables for the observation to be imputed lies outside the convex hull of the independent variables used to estimate the model parameters. Quadratic programming or other optimization techniques can be used to ascertain whether imputation involves extrapolation and, if so, by how much. Section 7 summarizes these problems and notes how differences between the modeled and imputation groups can affect the validity of the imputation procedure.

4. DATA

This section documents the relationships between the master file of the 1979 Nonresidential Buildings Energy Consumption Survey (NBECS) and all other data sets generated from the survey during the ORNL imputation study. The discussions start with a creation of the natural gas and the electricity working data sets in Sect. 4.1. These two working data sets were created separately in the early stages of the imputation study and were frequently modified. Each of the two working data files carries approximately half the data load (input/output) of the original master data file. Data records of these data files were coded in a format suitable for statistical analysis. New variables were created as a result of a data requirements review study that focused on a study of the independent variables required in developing the regression models for building energy consumption. The structure of the working data sets provided more programming flexibility, which allowed testing programs of one working data set to be used under another working data set without having to change the names of fuel-consumption-related variables. The structure also allows the use of more meaningful and identifiable variable names.

Additional imputation flag counts, which are the counts of the number of variables imputed, obtained from the EIA in July 1984, were included in the two working data sets. Because of the large number of buildings having imputed values of square footage or number of workers, frequency counts of imputation flags were examined in Sect. 4.2.

The two working data sets also identified outliers deleted by EIA personnel prior to the ORNL study. It is assumed that similar preliminary data edits on the 1983 NBECS working data files will be performed so that potentially erroneous records can be identified and deleted or corrected. Section 4.3 documents the data edits previously established by EIA personnel. Section 4.4 documents the discrepancy between the input data set previously used by the EIA and the input data set employed by ORNL to reproduce the EIA regression models.

To conduct a meaningful regression analysis based on the engineering relationships of building energy use, it is necessary to create a clean and frozen data set to support the modeling efforts. The initial input data set went through a set of special edits--in addition to the EIA edits --to complete the data-screening process. These additional data edits and their impact on the data sets are discussed in Sect. 4.5. To observe the impact of data edits, frequency counts were made before and after the inclusion of a particular set of data edits.

Section 4.6 suggests a total approach for creating initial input data sets. Studies of large buildings that were excluded from the modeling process are described in Sect. 4.7 with frequency counts by data groups (Group 1 and Group 3) and by building types. Section 4.8 describes analyses for building records with zero fuel consumption, and Sect. 4.9 discusses additional editing and outlier detection activities.

4.1 CREATION OF NATURAL GAS AND ELECTRICITY WORKING DATA SETS

The original Nonresidential Buildings Energy Consumption Survey (NBECS) master data file is CLASS.NBECS79, which was accessed using this job control language (JCL) statement:

```
//CLASS DD DSN=CN6616.RL2.GENE.NBECS79.TAPE2.OAKR.SASTEST3,DISP=SHR.
```

The large SAS data file, with 6222 records and 520 variables, is the master NBECS data set used by ORNL personnel throughout the imputation study period. The master data file was first converted into these two SAS data sets:

1. CONVERT.NATGAS, with 4129 records and 346 variables, was created to include only those building units that reported natural gas use. Information associated with other types of fuel use was excluded, except for the six end-use variables of all fuels: space heating, space cooling, water heating, electricity generation, cooking, and manufacturing.

2. CONVERT.ELECT, with 6145 records and 346 variables, was created to include only those building units that reported electricity use. Information associated with other types of fuel use was excluded, except for the six end-use variables.

Each of the two converted data files represents approximately half the data load (with respect to computer input/output) of the original master data file. Data records of these converted files are mostly coded directly from the 1979 NBECS form, as are the master data file records.

Because the data record files were not in a format suitable for statistical data analysis, it was necessary to alter the two converted data sets so that the corresponding working data sets could be constructed to meet the following needs for an efficient statistical data analysis:

1. Deleting records with contradictory variable values. These deletions include building records that reported natural gas (or electricity) use in more than one energy source field. Unless hard copies of these observations can be examined or follow-up phone calls can be made, the variables, "bill coverage days of natural gas consumption" and "the amount of natural gas consumed" cannot be clearly defined because of the multiple response fields.

2. Deleting variables that are not required for data analysis. For example, survey variables such as interview month, interview waiver time, number of tanks, etc., not used in the analysis of natural gas or electricity consumption are deleted to reduce the data load.

3. Recoding character survey variables to numerical variables for analysis or recoding continuous variables to categorical variables.

4. Transforming or combining existing variables into more meaningful variables.

5. Creating class variables that partition the data set into homogeneous subcategories for regression analysis.

6. Adding new variables that must be obtained from other existing data sets.

7. Minimizing the amount of Central Processing Unit (CPU) time required to load the data sets. (A SAS program that requires over 20 seconds of CPU time may require a waiting time of four hours on weekdays.)

8. Adding flags for those records that failed routine edit checks.

As a result, two working data sets, one for natural gas and one for electricity, were created for the analysis needs. Data set WORKING.NATGAS, the working data set for the natural gas-use building, contains 4127 observations and 313 variables. Data set WORKING.ELECT is the working data set for the electricity use buildings, and it contains 6143 observations and 306 variables. Table 4.1 lists the four building records that were excluded from the working data sets because of contradictory source fields.

Since the two working data sets were created separately, each fuel consumption or end-use-related variable can carry the same name in both working data sets. This structure provides more programming flexibility, which allowed testing programs of one working data set to be used under another working data set without having to change the names of fuel-consumption-related variables. The separation into two data sets also allowed more meaningful, identifiable variable names.

Source statements for the creation of the four data sets--CONVERT. NATGAS, CONVERT.ELECT, WORKING.NATGAS, and WORKING.ELECT--are documented in Appendices B, C, D, and E, respectively.

For each of the two fuels, records were divided into three groups, as was previously done by the EIA:

Group 1. Records whose periods of reported consumption covered 30 or fewer days in 1979 (DAYCLASS=1);
Group 2. Records whose periods of reported consumption covered 31-330 days in 1979 (DAYCLASS=2); and
Group 3. Records whose periods of reported consumption covered 331 days or more (DAYCLASS=3).

Since the fuel consumption data in Group 1 were considered completely missing, their consumption value records are to be imputed using information from the building records in Group 3 which will serve as a reservoir of potential input records for the imputation of Group 1 records. Table 4.2 gives the distribution of Group 1 and Group 3 natural gas-use buildings before any data edits. Table 4.3 gives the distribution of Group 1 and Group 3 electricity-use buildings before any data edits.

Table 4.1. Four questionable records in the NBECS79[a] data file

		Energy source field Number 1			Energy source field Number 2			Energy source field Number 3			
Building ID[b]	Fuel	Covered days of consumption[c]	Total consumption (Btu[d])	Number of suppliers	Covered days of consumption	Total consumption (Btu[d])	Number of suppliers	Covered days of consumption	Total consumption[c] (Btu[d])	Number of suppliers	Adjusted
730	Natural gas	NA			0	99,008,078	1	365	123,691,985	2	289.1309
735	Natural gas	NA			365	469,048,757	1	98	68,318,855	1	203.9631
3348	Electricity	365	76,920,128	2	357	133,201,068	2		NA		201.8503
3380	Electricity	365	189,159,481,876	2	365	189,408,015,368	2		NA		1

[a]NBECS79 = The original Nonresidential Buildings Energy Consumption Survey master data file.
[b]ID = Identification.
[c]NA = Not applicable.
[d]Btu = British thermal unit.

Table 4.2 **Distribution of Group 1 and Group 3 natural gas-use buildings before any data screening edits**

Building type (BCWM1)	Regression Category (RECCAT)	0-30 days target records (DAYCLASS = 1) FREQUENCY	331-365 days potential input records (DAYCLASS = 3) FREQUENCY
Agriculture	260	4	3
Assembly	10	40	239
Education	20	40	387
Food sales services	30	19	160
Health care	40	19	62
Health care	50	18	74
Industrial	60	36	155
Industrial	70	27	119
Industrial	80	17	92
Retail/services	90	38	96
Retail/services	100	10	78
Retail/services	110	22	158
Retail/services	120	11	100
Retail/services	130	4	43
Office	150	19	87
Office	160	31	147
Office	170	21	174
Office	180	11	100
Office	190	22	118
Residential	200	16	58
Residential	210	23	141
Lodging	220	8	80
Lodging	230	5	55
Warehouse/storage	240	18	120
Warehouse/storage	250	16	125
Other	260	41	130
Vacant	260	24	23
Automobile sales	140	16	115
Total		576	3239

Source: Statistical Analysis System data set WORKING.NATGAS.

Table 4.3. Distribution of Group 1 and Group 3 electricity-use buildings before any data screening edits

Building type (BCWM1)	Regression Category (REGCAT)	0-30 days target records (DAYCLASS = 1) FREQUENCY	331-365 days potential input records (DAYCLASS = 3) FREQUENCY
Agricultural	440	5	12
Assembly	270	45	356
Educational	280	83	492
Food sales/services	290	26	254
Health care	300	52	137
Industrial	310	33	191
Industrial	320	22	148
Industrial	330	22	123
Retail services	340	94	627
Office	360	33	126
Office	370	64	492
Office	380	13	159
Office	390	23	153
Residential	400	21	93
Residential	410	19	175
Lodging	420	36	176
Warehouse/storage	430	71	378
Other	440	54	211
Vacant	440	40	64
Automobile sales	350	23	189
Total		779	4556

Source: Statistical Analysis System data set WORKING.ELECT.

4.2 ADDITIONAL IMPUTATION FLAG COUNTS

Imputation flag counts imputed by the EIA for the building square footage (SQFT1) or the number of workers (NWKER1) variables were obtained from the EIA subsequent to starting the ORNL analysis. In the new data file, there are 1555 additional EIA imputation flags for the variable SQFT1 and 664 additional imputation flags for the variable NWKER1. ORNL personnel then added these imputation flags to the NBECS master data set and analyzed this information regarding the degree of imputation associated with each record. Analyses of additional imputation flag counts are presented in this section.

4.2.1 The Data

The four data sets involved in this study are: CLASS.NBECS79, NATGAS, ELECT, and IMPUTED.UPDATE. The data sets and variables are described in detail below.

4.2.1.1 CLASS.NBECS79

The NBECS 1979 master data set, with 6,222 records. The imputations were made by the WESTAT Company. The relevant imputation variables are:

Variable IMPSF1	-	Gives imputation flags for imputed square footage values;
Variable IMPSFC1	-	Gives imputation flags for imputed square footage category values;
Variable IMPNW1	-	Gives imputation flags for imputed number of workers values; and
Variable IMPNWC1	-	Gives imputation flags for imputed number of workers category values.

Distribution of the imputation flag counts in the original CLASS.NBECS79 data set is given in Table 4.4.

4.2.1.2 NATGAS

The working natural gas data set with data screening edits provided by the EIA is also a subset of the CLASS.NBECS79 data set.

With all the outliers deleted and regression categories identified (by the variable REGCAT), the data set structure is very close to the input data set originally used by the EIA to develop the nonresidential buildings natural gas-use models. NATGAS has 3014 records. The overall distribution of imputation flag counts is given in Table 4.5.

Table 4.4 Distribution of original imputation flag counts from the NBECS[a] master file

IMPSF1[b]	IMPSFC1[c]	FREQ	IMPNW1[d]	IMPNWC1[e]	FREQ
Yes	Yes	191	Yes	Yes	63
No	No	6031	Yes	No	2
Total		6222	No	Yes	6
			No	No	6151
				Total	6222

[a]NBECS – Nonresidential Buildings Energy Consumption Survey;
[b]IMPSF1 – Gives imputation fla s for imputed square footage values;
[c]IMPSFC1 – Gives imputation flags for imputed square footage, category values;
[d]IMPNW1 – Gives imputation flags for imputed number of workers values; and
[e]IMPNWC1 – Gives imputation flags for imputed number of workers category values.

Table 4.5. Distribution of imputation flag counts for the natural gas data file

IMPSF1[a]	IMPSFC1[b]	FREQUENCY	IMPNW1[c]	IMPNWC1[d]	FREQUENCY
Yes	Yes	110	Yes	Yes	21
No	No	2904	No	Yes	3
			No	No	2990
	Total	3014			
				Total	3014

[a] IMPSF1 – Gives imputation flags for imputed square footage value;
[b] IMPSFC1 – Gives imputation flags for imputed square footage category values;
[c] IMPNW1 – Gives imputation flags for imputed number of workers values; and
[d] IMPNWC1 – Gives imputation flags for imputed number of workers category values.

4.2.1.3 **ELECT**

The working electricity data set with data screening edits provided by EIS is also a subset of the CLASS.NBECS79 data set.

With all the outliers deleted and regression categories identified (by the variable REGCAT), the data set structure is very close to the input data set originally used by the EIA to develop the nonresidential buildings electricity-use models. ELECT has 4,222 records. The overall distribution of imputation flag counts is given in Table 4.6.

Table 4.6. Distribution of imputation flag counts for the electricity data file

IMPSF1	IMPSFC1	FREQ	IMPNW1	IMPNWC1	FREQ
Yes	Yes	149	Yes	Yes	38
No	No	4073	Yes	No	2
Total		4222	No	Yes	1
			No	No	4181
				Total	4222

4.2.1.4 **IMPUTED.UPDATE**

This data set includes an additional set of imputation flag counts with variables:

IMPSFX - Identifies building square footage values imputed by EIA personnel; and
IMPNWX - Identifies number of workers values imputed by EIA personnel.

This data file was obtained from the EIA on July 5, 1984. The new data file has 2,030 records, with a few questionable records that include:

One blank record (the last data line);

One complete duplicate (BLDGID1 - 3650, IMPNWX - Yes, IMPSFX - Yes);

Two records not found in the CLASS.NBECS79:

(1) BLDGID1 - 6732, IMPNWX - Yes, IMPSFX - No; and
(2) BLDGIDI - 7550, IMPNWX - No, IMPSFX - Yes

An overall distribution of these additional imputation flag counts is given in Table 4.7.

Table 4.7. Additional EIA imputation flag counts

IMPSFX		FREQ	IMPNWX		FREQ
Yes		1555	Yes		664
No		475	No		1366
	Total	2030		Total	2030

4.2.2 Updating the Imputation Flag Counts

Data set IMPUTED.UPDATE was merged into data set CLASS.NBECS79 by building identification numbers (BLDGID1). The resulting frequency counts for the merged data set are given in Table 4.8.

Table 4.8. Imputation flag counts from merged NBECS master file and imputed update file

IMPSF1	IMPSFC1	IMPSFX	FREQ	IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	191	Yes	Yes	No	63
No	No	No	4478	Yes	No	No	2
No	No	Yes	1553	No	Yes	Yes	6
	Total		6222	No	No	No	5495
	Total Imputed		1744	No	No	Yes	656
						Total	6222
					Total Imputed		727

Other analyses include the distribution of update imputation flag counts in the NATGAS data set as given in Table 4.9. The distribution of updated imputation flag counts in the ELECT data set is given in Table 4.10.

Table 4.9. Updated imputation flag counts for the NATGAS data set

IMPS[illegible]	IMPSFC1	IMPSFX	FREQ	IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	110	Yes	Yes	No	21
No	No	No	2139	No	Yes	Yes	3
No	No	Yes	765	No	No	No	2684
	Total		3014	No	No	Yes	306
	Total Imputed		875		Total		3014
					Total Imputed		350

Table 4.10. Updated imputation flag counts for the ELECT data set

IMPSF1	IMPSFC1	IMPSFX	FREQ	IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	149	Yes	Yes	No	38
No	No	No	3010	Yes	No	No	2
No	No	Yes	1063	No	Yes	Yes	1
	Total		4222	No	No	No	3768
	Total Imputed		1212	No	No	Yes	413
					Total		4222
					Total Imputed		454

The tables in Appendix M give a further breakdown of imputation flag counts by the 26 regression categories of the NATGAS data set and the 18 regression categories of the ELECT data set.

4.2.3 Findings

Records with square footage categories imputed by the WESTAT Company (IMPSFC1 = '1') or records with number of workers imputed by the WESTAT Company (IMPNWC1 = '1') can, in general, be deleted from their respective building categories during the model-building process without seriously affecting the number of available input records. Only 262 records were deleted from a total of 6,222 records.

The deletion of records with additional square footage or number of workers values imputed by the EIA may affect the validity of the resulting model because these deletions represent a maximum of 2,030 additional observations. These records provide good square footage category values because none of these records has an imputed square footage category value.

4.3 IDENTIFICATION OF RECORDS THAT FAILED THE EIA EDITS

Prior to the ORNL research effort, the EIA developed regression equations, within the 44 regression data categories, to help impute missing fuel-consumption values. Within each of the 44 regression categories--and before the regression model development stage--outliers were identified by the EIA, and potentially erroneous records were deleted from the input data set. These deleted input records were identified by the variable 'REGEDIT' in the ORNL working data sets.

Basically, two types of data screening were performed by the EIA. Building records with extremely high/low average fuel cost per unit consumption were identified, and survey forms were reexamined to see if these records had unreasonable consumption values. Records were selected for reexamination if the natural gas cost was less than \$0.10/100 ft^3 or greater than \$1.0/100 ft^3 or if the electricity cost was less than \$0.01/kw or greater than \$0.20/kw. Other "illegal" records were also identified as outliers with regression diagnostics. Within each of the two working data sets, the variable REGEDIT identifies the EIA-deleted outliers by regression category. Separate regression models are made for each building category (class number): so building category is the same as regression category. For example, a record in regression Category 10 (assembly buildings in the natural gas working data set) is an EIA-deleted outlier if REGEDIT = 10.

4.4 DIFFERENCES BETWEEN THE EIA AND ORNL INPUT DATA SETS

Two data sets will be discussed in this section. The OLD data set refers to the set of input records (after data screening) which the EIA used to develop its 1979 imputation models for the 44 regression categories. The NEW data set refers to the set of input records which ORNL employed to reproduce the imputation models developed by the EIA personnel.

The two data sets are nearly identical; however, some discrepancies still exist despite all efforts to match the two data sets. These differences will not affect the 1983 NBECS imputation activities. Observed differences are documented in Table 4.11.

4.5 CREATION OF THE INPUT DATA SETS TO SUPPORT ORNL REGRESSION ANALYSES

To conduct a meaningful regression analysis of the building energy use, it is necessary to create a clean, frozen, and archived input data set. While the input data set cannot be "error free," appropriate data screening procedures will provide an input data set with more reliable information.

Data screening was to be carried out in each of the 44 regression categories: 26 categories from the natural gas working data set and 18 categories from the electricity working data set. Within each regression category, data records were evaluated for their appropriateness in entering the input data set. Evaluation criteria were based on the degree of imputation associated with the records as well as the validity of the reported consumption values. Records considered not suitable for the regression analysis were set aside and excluded from the modeling input data set.

This portion of the report will describe the ORNL evaluation criteria used in the final data screening and the effect of this screening and will also examine the distributions of the large building groups.

Table 4.11. Differences in the input 1979 NBECS data sets

REGCAT	Observations in the NEW[a] data set but not found in the OLD[b] data set: BLDGID1[c]	Observations in the NEW[a] data set but not found in the OLD[b] data set: CNSUNIT[d]	Observations in the OLD[b] data set but not found in the NEW[a] data set: CNSUNIT
10	5074	577,266	0
	3057	616,953	0
20	2912		
70	423	602,322	6,023,222
	4669	5,253,408	115,719,500
	1982	44,799,225	
	5539	788,092,308	
90	5839	738,924	
	5193	2,732,211	
	5877	3,515,223	
	2099	8,231,288	
140			121,315
170	4388	143	
	3188	3,042	
	6752	98,982	
	0740	397,779	
	2994	481,774	
	2960	674,806	
	2863	688,030	
	2789	1,059,485	
	5865	1,090,292	
	4849	1,098,710	
	6063	1,165,424	
	4431	1,483,423	
	0898	2,052,704	
	3015	2,095,892	
	4227	2,617,867	
	0075	2,892,993	
	1989	4,415,574	
	5486	4,575,546	
	4329	5,891,928	
	6586	9,885,418	
	5574	12,499,922	
	5628	68,461,943	
	1253	4,105	
200	1351	220,381	0
270			55,439,473
			55,512,314
280			22,544
			39,039

[a]The data set created by ORNL to match OLD data set.
[b]The EIA input data set.
[c]BLDGID1 - Building identification number.
[d]CNSUNIT - Consumption of natural gas (electricity).

The following SAS data steps will create the input master data set for modeling natural gas-use buildings.

```
DATA MASTER;
SET WORKING.NATGAS;
IF A = 1;
IF DAYCLASS = 3;
IF SQFT1 <1000000;
IF NFLOOR1 < 50;
IF IMPSFC1 NE '1';
IF IMPNWC1 NE '1';
IF IMPSF1 NE '1';
IF IMPNW1 NE '1';
IF IMPSFX NE '1';
IF IMPNWX NE '1';
IF IMPNOF1 NE '1';
IF IMPPG1 NE '1';
IF IMPPGC1 NE '1';
IF IMPYRC1 NE '1';
IF BLCOVX NE '2';
IF CNSUNIT GE 0;
IF REGCAT NE REGEDIT;
IF REGCAT NE 001;
```

The input master data set for modeling electricity-use buildings is created by these SAS data steps:

```
DATA MASTER;
SET WORKING.ELECT;
IF A = 1;
IF DAYCLASS = 3;
IF SQFT1 <1000000;
IF NFLOOR1 < 50;
IF IMPSFC1 NE '1';
IF IMPNWC1 NE '1';
IF IMPSF1 NE '1';
IF IMPNW1 NE '1';
IF IMPSFX NE '1';
IF IMPNWX NE '1';
IF IMPNOF1 NE '1';
IF IMPPG1 NE '1';
IF IMPPGC1 NE '1';
IF IMPYRC1 NE '1';
IF BLCOVX NE '2';
IF CONSUNIT GE 0;
IF REGCAT NE REGEDIT;
IF REGCAT NE 002;
```

A detailed description of these evaluation criteria is given below.

Condition 1. IF A = 1;

A = 1 means that the fuel is one of the first three primary fuels used in the building. If the fuel (natural gas/electricity) is not reported as one of the first three primary fuels (A = 0), then the building fuel consumption is not likely to be influenced by the building characteristics under consideration. None of the 1979 sample buildings failed this edit.

Condition 2. IF DAYCLASS = 3;

Buildings with DAYCLASS = 3 include records whose periods of reported consumption covered 331 days or more in 1979. This group of records forms the basis of all the potential input records.

Condition 3. IF SQFT1 < 1,000,000;

Buildings with areas smaller than 1 million ft^2 were included. Buildings with areas larger than 1 million ft^2 were previously given the weighted average square footage of all sample buildings with areas larger than 1 million ft^2 in the same census region.

Condition 4. IF NFLOOR1 < 50;

Buildings with fewer than 50 floors were included. Buildings with more than 50 floors were previously assigned a truncated value of 50 floors.

Condition 5. IF IMPSFC1 NE '1';

Buildings with imputed square footage category were excluded.

Condition 6. IF IMPNWC1 NE '1';

Buildings with imputed number of workers category were excluded.

Condition 7. IF IMPSF1 NE '1';
IF IMPNW1 NE '1';
IF IMPSFX NE '1';
IF IMPNWX NE '1';

Buildings with imputed square footage or imputed number of workers were excluded.

Condition 8. IF IMPNOF1 NE '1';

Buildings with imputed number of floors were excluded.

Condition 9. IF IMIPG1 NE '1';
IF IMPPGC1 NE '1';

Buildings with imputed percent of glass were excluded.

Condition 10. IF IMPYRC1 NE '1';

Buildings with imputed year-built category were excluded.

Condition 11. IF BLCOVX NE '2';

Buildings with BLCOVX = '2' include cases where reported fuel consumption covered more than the respondent building. These buildings were subject to disaggregation by the EIA, and they were excluded from the input data set.

Condition 12. IF CNSUNIT > 0;

This condition is to be combined with condition 2. Buildings with zero consumption and over a 330-day period of bill coverage were considered outliers. Section 4.8 discusses this condition in detail.

Condition 13. IF REGCAT NE REGEDIT;
If REGEDIT NE 001;

001 was used in the natural gas data set, and 002 was used in the electricity data set. The variable REGCAT is a group identifier that assigns a building in the respondent sample to one of the 44 regression categories. The variable REGEDIT identifies outliers (as defined in Sect. 4.3) of the corresponding regression category.

Example: Suppose a building is selected from REGCAT = 060. If its value of REGEDIT = 060, then this building is an outlier that is to be excluded from the particular regression category 060. If its value of REGEDIT = 001 (for the natural gas data set), then this building is a general outlier of the natural gas working data set. This record will also be deleted.

Tables 4.12 and 4.13 give the distributions of the two input data sets by building type and by regression category. These input building records satisfy all edits under Conditions 1-13.

Table 4.12. Distribution of Group 3 (DAYCLASS = 3) natural gas-use buildings by building type and by regression category: after all data screening edits

Building type (BCWM1)	Regression category (REGCAT)	FREQ	Building type (BCWM1)	Regression category (REGCAT)	FREQ
Agriculture	260	2	Office	160	76
Assembly	10	84	Office	170	102
Educational	20	201	Office	180	58
Food sales/services	30	79	Office	190	64
Health care	40	32	Residential	200	12
Health care	50	34	Residential	210	45
Industrial	60	100	Lodging	220	24
Industrial	70	81	Lodging	230	26
Industrial	80	53	Warehouse/storage	240	81
Retail/services	90	43	Warehouse/storage	250	95
Retail/services	100	41	Other	260	70
Retail/services	110	102	Vacant	260	15
Retail/services	120	62	Automobile sales	140	61
Retail/services	130	21			
Office	150	54		Total	1718

Table 4.13. Distribution of Group 3 (DAYCLASS = 3) electricity-use buildings by building type and by regression category: after all data screening edits

Building type (BCWM1)	Regression category (REGCAT)	FREQ	Building type (BCWM1)	Regression category (REGCAT)	FREQ
Agricultural	440	7	Office	380	99
Assembly	270	121	Office	390	80
Educational	280	228	Residential	400	19
Food sales/services	290	123	Residential	410	68
Health care	300	66	Lodging	420	53
Industrial	310	123	Warehouse/storage	430	233
Industrial	320	90	Other	440	96
Industrial	330	63	Vacant	440	37
Retail/services	340	337	Automobile sales	350	86
Office	360	74			
Office	370	266		Total	2269

Observe that the resulting natural gas (electricity) input data set has a total of 1718 (2269) records, which is approximately 47% (50.2%) of a reservoir of 3239 (4556) potential input records in record Group 3 (see Tables 4.2 and 4.3). This large reduction is caused mainly by the large number of buildings that failed Condition 7 or Condition 11 edits. For Condition 11 alone, a total of 368 Group 3 records of the natural gas data set failed the edit, and a total of 759 Group 3 records of the electricity data set failed the edit. Table 4.14 shows the individual as well as the joint influence of Conditions 7 and 11 on the potential input data set.

Table 4.14. The influence of data screening condition 7 and condition 11 on the initial input data set for the Group 3 (DAYCLASS = 3) records

Working data set	No edits	All edits except conditions 7 & 11	All edits except condition 7	All edits except condition 11	All edits
WORKING.NATGAS	3239 (100%)	2916 (90%)	2587 (80%)	1956 (60%)	1718 (53%)
WORKING.ELECT	4556 (100%)	4155 (91%)	3443 (76%)	2787 (61%)	2269 (50%)

NOTE: Percentages of the original potential data set with no edits are calculated and displayed inside the parentheses.

4.6 SUGGESTED APPROACH TO CREATE THE INPUT DATA SETS

A total approach, which time does not allow for this study, to create initial input data sets to support the imputation study would include the following steps:

1. Create the two working data sets as suggested in Sect. 4.1.
2. Complete all EIA edits as was described in Sect. 4.3 and identify these records.
3. More outlier detection efforts can be made exclusively for each of the two working data sets as discussed in Sect. 4.9.
4. For the natural gas (electricity) working data set, run PROC FREQ (SAS) over all the character variables and run PROC UNIVARIATE over all the numeric variables. The analysis should also be done for each regression category.
5. Conduct other analyses of frequency counts.
6. Develop evaluation criteria for the initial data screening process based on information obtained from 2-5.

7. Study the impact of data screening efforts as suggested in Sect. 4.5. The analysis may suggest some necessary changes in the modeling strategy such as collapsing a few regression categories or perhaps excluding some data screening edits.
8. Create an initial input data set for regression analysis. Adjust this data set as the modeling process progresses.

4.7 DISTRIBUTION OF LARGE BUILDINGS

The category of large buildings includes buildings that have square footage values greater than or equal to 1 million ft^2, or have 50 or more floors. As was discussed with data screening (Sect. 4.5), these large buildings were excluded from the ORNL regression modeling efforts. They were excluded through Conditions 3 and 4 because they carry mean imputed square footage values or truncated number of floors values. Tables 4.15 through Table 4.18 give joint distributions of these large buildings by SQFT1, by NFLOOR1, and by building type. Records under DAYCLASS = 1 are Group 1 records that are treated as buildings with missing consumption values. Records under DAYCLASS = 3 are Group 3 input records. These tables give frequency counts before and after the data screening process.

Table 4.15. Joint distribution of large natural gas use, building groups by building type: before the data screening edits

	SQFT1 > = 1,000,000 NFLOOR 1 > = 50		SQFT1 > = 1,000,000 only		NFLOOR1 > = 50 only	
Building type (BEWM1)	DAYCLASS = 1	3	DAYCLASS = 1	3	DAYCLASS = 1	3
Assembly			0	1		
Educational			0	2		
Health care			1	2		
Industrial			5	10		
Retail/services			7	19		
Office	6	7	6	15	0	3
Residential	1	0				
Lodging			0	2		
Warehouse			1	3		
Other			1	5		
Vacant			1	0		
Automobile sales						
Total	7	7	22	59	0	3

Table 4.16. Joint distribution of large natural gas-use building groups by building type: after the data screening edits

Building type (BCWM1)	SQFT1 > = 1,000,000 NFLOOR 1 > = 50 DAYCLASS = 1	SQFT1 > = 1,000,000 NFLOOR 1 > = 50 DAYCLASS = 3	SQFT1 > = 1,000,000 only DAYCLASS = 1	SQFT1 > = 1,000,000 only DAYCLASS = 3	NFLOOR1 > = 50 only DAYCLASS = 1	NFLOOR1 > = 50 only DAYCLASS = 3
Assembly						
Educational						
Health care			1	2		
Industrial			4	1		
Retail/services			3	9		
Office	5	2	6	8	0	2
Residential						
Lodging			0	1		
Warehouse			1	0		
Other			1	2		
Vacant			1	0		
Automobile sales						
Total	5	2	17	23	0	2

Table 4.17. Joint distribution of large electricity-use building groups by building type: before the data screening edits

Building type (BCWM1)	SQFT1 >= 1,000,000 NFLOOR1 > = 50 DAYCLASS = 1	SQFT1 >= 1,000,000 NFLOOR1 > = 50 DAYCLASS = 3	SQFT1 >= 1,000,000 only DAYCLASS = 1	SQFT1 >= 1,000,000 only DAYCLASS = 3	NFLOOR1 >= 50 only DAYCLASS = 1	NFLOOR1 >= 50 only DAYCLASS = 3
Assembly			0	1		
Educational			0	2		
Health care			1	2		
Industrial			7	12		
Retail/services			7	20		
Office	8	9	7	29	0	2
Residential			1	0		
Lodging			0	2		
Warehouse			2	5		
Other			3	3		
Vacant			1	0		
Automobile sales						
Total	8	9	29	76	0	2

Table 4.18. Joint distribution of large electricity-use building groups by building type: after the data screening edits

Building type (BCWM1)	SQFT1 >= 1,000,000 NFLOOR1 >= 50		SQFT1 >= 1,000,000 only		NFLOOR1 >= 50 only	
	DAYCLASS = 1	DAYCLASS = 3	DAYCLASS = 1	DAYCLASS = 3	DAYCLASS = 1	DAYCLASS = 3
Assembly						
Educational			0	1		
Health care			1	1		
Industrial			6	3		
Retail/services			3	10	0	1
Office	7	6	5	23		
Residential						
Lodging			0	1		
Warehouse			2	2		
Other			3	2		
Vacant			1	0		
Automobile sales						
Total	7	6	21	43	0	1

Group 3 records that failed data screening edits are records with inadequate information on either the fuel consumption values or important explanatory variable values of building fuel consumption. Group 1 records that failed data edits are records with inadequate information on important explanatory variable values of the building fuel consumption. Records that survived the data screening edits (Tables 4.16 and 4.18) are records with more usable information. Within a particular building type, if there are enough input Group 3 records that survive the data edits, then it is possible to impute the missing consumption values of the corresponding Group 1 records with some confidence.

4.8 BUILDING RECORDS WITH ZERO-FUEL CONSUMPTION VALUES

In the ORNL modeling approach, building records that failed Condition 12 of Sect. 4.5 were deleted (the buildings with zero-fuel consumption and over 330 days of utility bill coverage). After all other edits, only three records of the electricity data set were considered as outliers because of the zero consumption values. Table 4.19 lists selected building characteristics of the three Group 3 records that did not pass the zero-consumption edit.

These outliers were deleted from the input data set because the limited survey information did not provide evidence to justify a possible zero

Table 4.19. Some building characteristics of the three Group 3 records that passed all other edits but failed the zero-consumption edit

Variable	Description[a]	Variable Value		
BLDGID1	Building ID[b]	4510	5529	6474
PORVAC	Portion vacant	No	No	No
AVGNHR1	Average weekly open hours	30	168	124
BCLASS1	Building class	Senior High	Parking garage	Industrial
NWKER1	Number of employees	166	6	500
SQFT1	Square footage	298,116	108,233	609,982
NFLOOR1	Number of floors	4	14	5
BOILR1	Boiler present	Yes	No	Yes
BOILRX	Boiler powered by elect.	No	No	No
NGUSED1	Natural gas used	Yes	No	No
FOCH1-FOCC1	Fuel conversion (fuel switched)	No	No	No
DAYS	Days of consumption	365	365	365
CNSUNIT	Electricity consumption	0	0	0
CSTDX	Total annual electricity cost	$47	$12	$87
BASEWT1	Basic sampling weight	11.999	100	15.1735
HDD651	Heating degree days (65°F)	7123	6503	4694
CDD751	Cooling degree days (75°F)	6	37	73
CDD651	Cooling degree days (65°F)	429	685	768
REGION1	Census region	North central	North central	South
ENDUSE1	Elect. for space heating	No	No	No
ENDUSE2	Elect. for cooling	Yes	No	No
ENDUSE3	Elect. for water heating	No	No	Yes
ENDUSE4	Elect. for elect. gen.[c]	No	No	No
ENDUSE5	Elect. for manufacturing	No	No	No
ENDUSE6	Elect. for cooking	Yes	No	No

[a] Elect. = electricity.
[b] ID = identification
[c] gen. = generation.

electricity consumption value. First, these buildings have used electricity as the first primary source of energy, and they did not switch from one fuel to another in 1979. Second, these buildings were not partially vacant for the most part of 1979. Third, these building activities (senior high, parking garage, industrial) require a large amount of electricity under normal operating conditions. A parking garage of 14 floors needs an electrically-powered elevator to operate over the reported 168 hours per week. Fourth, electricity consumption of these buildings should have exceeded the use of an 100-watt electric bulb for eight hours a day and five days per week over the entire year of 1979. A senior high school of 108,233 square feet, which had 166 workers and used electricity for cooking and cooling, should have used a large amount of electricity.

If the survey questionnaire collected information on lighting fixtures and the source of energy for lighting, then a lower bound for the energy might be established. The best approach, however, is not to delete these records immediately, but to contact the building representatives for other possible factors that might support the zero consumption or the observed close-to-zero costs of electricity in these buildings.

4.9 SOME ADDITIONAL SUGGESTED OUTLIER DETECTION (EDIT) ACTIVITIES

"Outliers" in this section refer to any potentially erroneous input values, whether due to incorrect information, transcribing, or data entry errors. Different types of errors may occur in a data set, and at least five types of edits can be useful: field, logic, range, regression, and previous value edits.

Field edits identify errors that most commonly occur in transcribing and data entry (e.g., alphanumeric values in numeric fields), negative values, multiple response instead of a single response, any other inapplicable multiple-choice-type response, or no response.

The following variables for future NBECS should be screened carefully for missing values (a field edit): square footage, percent heated, percent cooled, number of workers, percent vacant, hours of operation, number of floors, percent glass, the six end-use variables (space heating, space cooling, water heating, manufacturing, electricity generation, and cooking), cooling degree days based of 65°F (CDD65), and heating degree days based on 65°F (HDD65). These variables are important parts of the engineering regression model discussed in Sect. 5.3. As discussed in Sect. 4.2, the number of workers and square footage values are the most likely to be missing and are very important variables. Efforts should be made to call the respondent to obtain this information or, better yet, to emphasize with the interviewer that this information must be obtained.

Logic edits try to detect illogical patterns in response. For example, if question 14 is "yes," should question 15 be "no"? An NBECS example of this problem was a natural gas user in the lodging category who responded that natural gas was used for manufacturing. A more logical response would have been cooking or water heating.

The field edits are easily developed, but the logic edits take more effort. A good time to develop the logic edits is *before* the form is approved so that unclear questions or unclear response paths can be rethought. A good example of field and logic edits used successfully is the EIA-764 distillate fuel oil survey conducted by ORNL. The actual detailed field and logic edits are documented in *EIA-764 Survey Processing Specifications for Automated Systems and Form Edits*, S. Anderson et al. (1983). The field and logic edits were applied immediately after the data were entered. A list of the edits that had failed and an appropriate source of error were printed for each observation. Some problems were data entry errors or other obvious errors that could be resolved easily. Other problems required telephone calls to the respondents.

The EIA-764 edits were utilized to *detect inconsistencies* and, depending on the *error*, to develop *rules* for calling the respondents. These steps helped to reduce the contractor and respondent effort, reduce the number of calls, reduce costs, and create a better data base for further analysis.

The field and logic edits alone cannot provide enough confidence in the validity of a data base: range edits are the next step. The range edit is what the name implies: an edit which verifies that the reported value is within a reasonable range. A reasonable range is often difficult to define but is worth the attempt. Univariate edits (e.g., $0 < x_1 < 40$) are ideal but multivariate edits (e.g., $x_1 + x_2 = 50$) should also be evaluated. An entire technical literature is devoted to the problems of error localization (i.e., determining which variable is the most likely to be in error) for multivariate edit problems. One particular example of a multivariate edit, the cost/ consumption ratio, was useful for the NBECS and is discussed in Appendix F. The ORNL study team found it to be a useful tool.

The regression models can also be used to detect possible outliers (a type of multivariate edit). The entire respondent record for any building with a large studentized residual should be studied to determine whether an input error might have occurred or whether another factor needs to be added to the model. Therefore, this is important for a clean data base and proper model development.

The final type of edit is the previous value check. For the 1983 NBECS data base, the 1979 nonimputed data can be used to verify data or substitute for missing values. These checks (substitutions) should be made carefully. For example, if the building consumption for 1983 is larger/ smaller by a sizeable amount and no change in percent vacant occurred, the 1979 square footage value is expected to be larger/smaller than the 1983 square footage value. Rather than compare all variable values, only the subset of important variables should be compared initially. If considerable differences arise with respect to these important variables, then a more detailed study can be made. The purpose of the NBECS imputation is to provide electricity- and natural gas-consumption values; however, good values for the independent variables must be available to (1) create imputation models and (2) impute. The analyst will need to make a decision as to how much checking is necessary for a particular problem. It is also important to remember that the 1979 and 1983 data need not match exactly. Some reporting errors can be expected between the two time periods even though no real changes have occurred. The analyst, again, will have to determine the magnitude of a reasonable change.

In summary, these points are appropriate:

1. Edits are valuable tools for checking the quality of a data base and, therefore, provide more confidence in analytical results derived from the data.
2. For the NBECS data, edits such as the field, logic, and range edits can provide benefits with respect to "cleaning up" the data base, such as: (1) detecting potential errors; (2) developing rules for contacting respondents and correcting errors; (3) automating processes and, thus reducing manual time and human error; and (4) reducing costs.
3. Variables important to the regression equation should be screened carefully for missing values. These values should be obtained rather than imputed. In the past, the values most likely to be missing were square footage and number of workers.
4. The regression model can also be used to detect erroneous values, and potential outliers should be resolved.
5. A previous value check of important regression variables can be made for the 1983 NBECS and should be made especially if the 1979 value is to be substituted for a missing 1983 value.

5. RESULTS OF REGRESSION ANALYSIS

This section presents the results of all regression analyses and data structure studies. The models previously developed by the EIA were reproduced, and diagnostics were calculated. These activities are summarized in Sect. 5.1. The calculations for the ORNL data structure study are summarized in Sect. 5.2. The ORNL detailed engineering regression model is described in Sect. 5.3. The results of the regression engineering analysis are presented in Sect. 5.4.

5.1 STUDY OF PREVIOUS EIA MODELS

Section 2 describes the background for the EIA models that are listed in Tables 5.1 and 5.2. Of the 44 regression building categories, 26 are for the natural gas models and 18 for the electricity models. (The model for electricity consumption in health care facilities was not available, so only 17 electricity models are summarized.) The regression categories and associated building activity class codes are listed in Appendix A. Table 5.3 contains more information about the regression equation variables from Tables 5.1 and 5.2. It was important to reproduce these models and to provide and interpret diagnostics to learn from the time-consuming work already expended on regression and data analysis. Diagnostics helped to point out the weaknesses of the models and to suggest some potential avenues for improvement.

Tables 5.1 and 5.2 provide the square root of the model mean squared error ($\sqrt{MSE}$), number of observations (N), R^2 statistic, coefficient of variation (CV), average consumption value (Avg Y), and coefficients $\beta_0 \ldots \beta_p$ (where p is the number of independent variables in the model). The variable name is listed below each coefficient. The models are the result of stepwise regression procedures (forward and backward) available in the Statistical Analysis System (SAS) software package. The R^2 statistic is included for completeness although it is not really helpful in this analysis. The R^2 statistic, in general, is used by many to measure how well a model predicts and is calculated as: regression sum of squares divided by the total sum of squares. The inadequacy of the R^2 statistics is discussed further in Sect. 5.4.

As noted in Table 5.3, the CLIMZONi variable is based on a 40-year average of heating degree days (HDDs) and cooling degree days (CDDs). This variable is not available in the current NBECS master file, and it was replaced by the variable Weather Zone i in all equations except mixed retail/wholesale (natural gas), which already included both the CLIMZON4

Table 5.1. Previous regression models developed by the Energy Information Administration for electricity consumption imputation

Electricity

Category	$\sqrt{MSE}$	N	R^2	CV	Avg Y	B_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}	B_{12}
(1) Assembly Bldgs	1762321	338	.92	158	1112646	232232*	-233002 #Floors	19 Sq ft	13154 #Employees	-16 Sq ft Cool	-3129396 Enduse 5	1584706 BLCL2						
(2) Education Bldgs	611704	450	.68	78	780511	-833161	83 HDD60	135171 Bldg Age Cat	1.5 Sq ft	-45 Sq ft Res†	3900 #Employees	-1 Sq ft Heat	5 Sq ft Cool	2990 Hrs Op	207029 Weather Zone 1	193552 Region 1	173213 BLCL1	
(3) Food Sales Bldgs	280030	232	.91	88	317759	-294758	-12 HDD60	47 Sq ft Res†	164956 #Emp Cat	72852 Sq ft Cat	22 Sq ft Heat	-116976 CLIMZON 4	231580 BLCL1					
(5) Assembly Plants	5401752	169	.77	125	4305737	1059306	16 Sq ft	3171 #Employees	25 Sq ft Cool	-1561573 % Glass								
(6) Raw Goods Industrial	10175950	141	.59	165	6149923	650186	20 Sq ft	25926 #Employees	-67 Sq ft Cool	5813457 Induse 3	-5234553 CLIMZON 4							
(7) Other Industrial Bldgs	7259869	117	.82	109	6706596	-1756525	32 Sq ft	34 Sq ft Cool	79365 Hrs Op	-6458039 Enduse 2								
(8) Retail Sales/Service	2615693	530	.78	166	1570390	-48514	-142705 #Floors	2 Sq ft	3939 #Employees	4 Sq ft Heat	1330309 Enduse 6	1073217 CLIMZON 3	1605570 BLCL1					
(9) Auto Sales/Service	74117	171	.76	113	65334	-66194	84341 #Emp Cat	-5 Sq ft Heat	7 Sq ft Cool	68430 Enduse 4								
(10) General Office Bldgs	4110761	120	.68	124	3314974	-1901637	14 Sq ft	-11960 CDD75	2092 #Employees	-38 Sq ft Vacant	27839 Hrs Op	10485354 Weather Zone 6						
(11) Professional Office Bldgs	4634308	447	.72	118	3909877	727747	12 Sq ft	900 #Employees	-6 Sq ft Heat	2 Sq ft Cool	-11 Sq ft Vacant	-1510565 Region 4						
(12) Financial Office Bldgs	3166605	153	.85	99	3195842	-2158110	189758 #Floors	-461 Sq ft Res†	2331 #Employees	-6 Sq ft Heat	11 Sq ft Cool	28207 Hrs Op	2193418 Weather Zone 6					
(13) Mixed Use Office Bldgs	5188291	144	.59	229	2262574	1062120	-477015 #Floors	12 Sq ft	7797 #Employees									

†Residential

Table 5.1. (Continued)

Category	√MSE	N	R^2	CV	Avg Y	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
(14) Residential Bldgs	216075	81	.86	86	249634	-358117	-155 CDD64	53119 YREST	149921 #Emp Cat	4 Sq ft Heat	8 Sq ft Cool	-14 Sq ft Vacant	81874 % Glass	555436 BLCL1
(15) Mixed-Use Residential	364791	158	.85	206	176730	-19587	14690 #Employees	10 Sq ft Heat	-10 Sq ft Cool	61 Sq ft Vacant				
(16) Lodging Bldgs	1046279	[illegible]	.93	70	1502965	-801990	96 HDD60	101814 #Floors	4 Sq ft	6393 #Employees	-18 Sq ft Vacant	200634 % Glass	602655 CLIMZON 5	5959162 BLCL1
(17) Warehouse Storage Bldgs	2591614	357	.56	230	1124550	-1378273	557 HDD60	15 Sq ft	-12 Sq ft Heat	14032 Hrs Op	-1259840 Enduse 1	1058360 Region 3	2499131 BLCL1	
(19) "Other" Bldgs	1703631	267	.90	123	1384161	366933	15 HDD80	-160189 #Floors	7 Sq ft	2517 #Employees	11 Sq ft Heat	12 Sq ft Cool	6873417 BLCL1	-1258552 BLCL2

Table 5.2. Previous regression models developed by the Energy Information Administration for natural gas consumption imputation

Natural Gas

Category	$\sqrt{MSE}$	N	R^2	CV	Avg Y	B_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8
(19) Assembly Bldgs	8420910	225	.86	237	3547588	-1091693	-212974 CD080	898982 #Floors	178 Sq ft Cool	261 Sq ft Vacant	19425118 Enduse 5			
(20) Education Bldgs	6717425	353	.73	93	7227655	-1951660	409 HDD60	-138765 CD080	17670 #Employees	5 Sq ft Heat	30 Sq Ft Cool	2169726 CLIMZON4		
(21) Food sales Bldgs	1431983	152	.58	131	1095462	-929892	59 Sq ft Heat	16281 Hrs Op						
(22) Health Care < 350,000 sq ft	19180721	69	.57	89	21444202	-4488086	1533249 CD080	114 Sq ft Heat	75288556 Enduse 5	13938634 Enduse 6	19780686 Weather Zone 1			
(23) Health Care ≥ 350,000 sq ft	56363236	61	.52	80	70717089	4269622	111 Sq ft Heat	153 Sq ft Cool						
(24) Assembly Plants	27224909	136	.56	152	17954601	-13148574	94 Sq ft Heat	258355 Hrs Op						
(25) Raw Goods Indust	69571526	109	.68	171	40618326	-9455862	336438 #Employees	5451 Sq ft Vacant	-40998112 Enduse 6					
(26) Other Industrial Bldgs	48685765	76	.71	122	39765477	-28886858	177 Sq ft	-83910 #Employees	286 Sq ft Heat	370859 Hrs Op	39789617 Enduse 2			
(27) Shopping Centers	9150223	85	.58	102	8972849	-15283287	2300204 Sq ft Cat	175 Sq ft Vacant	35699807 Enduse 5	5314035 Weather Zone 2				
(28) Retail Sales < 3 Floors	523622	149	.78	74	702215	-36423	312 CDD60	214577 #Emp Cat	23 Sq ft Heat	-40 Sq ft Cool	-4436995 Enduse 4	-298427 Enduse 6	-303964 CLIMZON5	260159 Weather Zone 2
(29) Retail Sales ≥ 3 Floors	964229	73	.97	57	1690364	291263	3293 #Employees	25 Sq ft Heat	-1259121 CLIMZON5					
(30) Personal Services Bldgs	1690287	97	.73	108	1558516	-372627	11 Sq ft	680863 #Emp Cat	-14 Sq ft Heat	2860224 Enduse 5	-1269408 Region 3	1852312 BLCL1		
(45) Other Bldgs	14990871	149	.80	154	9722814	-1796812	100 Sq ft Heat	69 Sq ft Cool	16846580 Enduse 5	10518718 Weather Zone 4	76177084 BLCL2			

Table 5.2. (Continued)

Category	$\sqrt{MSE}$	N	R^2	CV	Avg Y	B_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8
(31) Mixed Retail/Wholesale	404476	40	.93	34	1172493	-1163712	365730 #Emp Cat	307718 Sq ft Cat	-1782810 Enduse 4	2703130 Enduse 5	453411 CLIMZON 4	-782417 Weather Zone 4	845791 Region3	3029425 BLCL1
(32) Auto Sales/Service	944184	106	.65	103	919036	-2299503	226 HDD50	911761 #Emp Cat	218045 Sq ft Cat	682824 Region 3	-654212 BLCL1			
(33) General Office	3191498	79	.90	104	3066516	1441515	87 Sq ft Heat	-79 Sq ft Cool	74 Sq ft Vacant	-32604 Hrs Op				
(34) Professional Office < 75 Emps	1095201	141	.54	145	752732	106733	90 Sq ft Heat	12765018 TD7	2484927 TC4	1782296 Enduse 5				
(35) Professional Office ≥ 75 Emps	14024497	134	.56	167	8383945	-4939796	1264 HDD50	-134632 CDD30	315554 #Floors	14 Sq ft Heat	97 Sq ft Cool			
(36) Financial Office	2960995	89	.95	72	4124058	-602659	28 Sq ft	687 Sq ft Res	-4937 #Employees	34 Sq ft Heat	15 Sq ft Cool	92 Sq ft Vacant		
(37) Mixed Use Office	13506205	115	.71	241	5635362	12443311	-1332704 #Floors	-39181 #Employees	255 Sq ft Heat	-11981048 Enduse 1				
(38) Residential Bldgs	2499336	55	.84	88	2932425	-9043637	532192 #Floors	-30 Sq ft Res	2283541 #Emp Cat	45 Sq ft Heat	1295625 % Glass	2734214 Enduse 1		
(39) Residential Mixed Usel	2604595	136	.96	155	1717906	-2546186	1093092 #Floors	-256 Sq ft	290 Sq ft Heat	157314 #Employees	-21 Sq ft Res	-13027 Hrs Op	-5735334 Enduse 4	1524318 CLIMZON3
(40) Commercial Lodging	3912836	72	.99	41	9604108	-2144754	394408 CDD75	1090082 Sq ft Cat	31 Sq ft Heat	18570695 BLCL2				
(41) Other Lodging	2543001	54	.98	44	5741511	-4614264	1033689 Bldg Age Cat	-512 Sq ft Res	7047 #Employees	51 Sq ft Heat	197 Sq ft Cool	-7550297 Enduse 2	2217665 Region 2	14823323 BLCL1
(42) Refrigerated Warehouses	2530603	105	.75	85	3097992	-493543	-1615 Cond	12409 #Employees	38 Sq ft Heat	129 Sq ft Cool	2873232 Enduse 5	1762617 Region 2		
(43) Nonrefrigerated Warehouses	17577456	120	.74	235	7479049	6373757	241055 #Employees	-2683464 Sq ft Cat						

Table 5.3. Additional notes about the regression equation variables from Tables 5.1 and 5.2

Regression category	Notes
All categories of natural gas or electricity	ENDUSE1 = 1, if the building uses natural gas in the natural gas model (or electricity in the electricity model) for heating, else = 0
	ENDUSE2 = As for ENDUSE1 but for space cooling
	ENDUSE3 = As for ENDUSE1 but for water heating
	ENDUSE4 = As for ENDUSE1 but for electricity generation
	ENDUSE5 = As for ENDUSE1 but for manufacturing
	ENDUSE6 = As for ENDUSE1 but for cooking
	Region i = 1, 2, 3, or 4 depending on the census region (1 = northeast, 2 = north central, 3 = south, 4 = west)
	CLIMZONi = 1, 2, 3, 4, or 5 depending on the weather zone. This is based on heating degree-days (HDDs) and cooling degree-days (CDDs) for 1979.
	Emp cat (no.) = Number of employees category rather than actual number of employees [Employees (no.)]
	Weather Zone i = As for weather zone but based on a 40-year average of the HDDs and CDDs
	Sq ft res = Number of residential square feet in a building
	Oper.h/w = Weekly number of hours of operation

Table 5.3 (Continued)

Natural gas category	Building type
Personal services	BLCL1 = 1 if building class = 951
Other	BLCL2 = 1 if building class = 1600
Mixed retail/wholesale	BLCL1 = 1 if building class = 1054
Auto sales/service	BLCL1 = 1 if building class = 936
Professional office, <75 employees	TD7 = 1 if the building has heating panels in the walls or floor, and
	TC4 = 1 if the building has a cooling system other than window units, packaged units, or central air
Commercial lodging	BLCL2 = 1 if building class = 1411
Other lodging	BLCL1 = 1 if building class = 1400
Electricity category	
Assembly	BLCL2 = 1 if building class = 251
Education	BLCL1 = 1 if building class = 340
Food sales	BLCL1 = 1 if building class = 441
Retail sales/service	BLCL1 = 1 if building class = 910
Residential	BLCL1 = 1 if building class = 1310
Lodging	BLCL1 = 1 if building class = 1415
Warehouse storage	BLCL1 = 1 if building class = 1530
Other	BLCL1 = 1 if building class = 800
	BLCL2 = 1 if building class = 1260

and Weather Zone 4 variables. Four electricity and five natural gas regression categories were involved in the substitution.

The following electricity regression categories exhibited reproduced models that were different from the EIA models: retail, raw goods industrial, assembly, and food sales. The ORNL reproductions that do not reasonably match the original EIA regression models are summarized in Table 5.4. The estimates that do not match well are connected by arrows. "Not matching well" applies to coefficient estimates that have opposite signs or whose ORNL estimate is outside three times the standard error of the EIA estimate (even though the error term is not normally distributed, it was used as a tool) or applies to estimates of the square root of the model mean square error ($\sqrt{\text{model MSE}}$) that differ by at least 30,000, an arbitrary number.

For the assembly buildings (electricity) category, the previous EIA model was calculated with an additional two observations (building identification numbers are unknown) with large consumption values. These two observations are the probable cause of the difference. Three of the four categories where the Weather Zone i variable is substituted for CLIMZONi do not match well (i.e., all listed in Table 5.4, except for the lodging category). The substitution is assumed to be the source of the mismatching.

The following natural gas regression categories exhibited reproduced models that were different from the EIA models: education, health care, ≥350,000 ft^2, assembly plants, other industrial, raw goods industrial, mixed retail/wholesale, mixed use office, residential mixed use, shopping, retail sales <3 floors, and retail sales ≥3 floors. The ORNL reproductions that do not reasonably match the original EIA regression models are summarized in Table 5.5. Estimates that do not match well are connected by arrows. "Match well" is as defined above for Table 5.4.

As discussed in Sect. 4, some additional observations were in the ORNL data set. Some of these observations had to be deleted from calculations to preserve the EIA model coefficient estimates. Notably, buildings 1982 and 5539 had to be omitted from the raw goods industrial (natural gas) category, and buildings 5628, 5574, and 6586 were omitted from the professional office building with <75 employees (natural gas) category.

CLIMZONi was replaced by Weather Zone i in five natural gas categories. Of these five, only one (education) exhibited a significantly different coefficient. It was not possible to estimate a CLIMZON4 coefficient for the mixed retail/wholesale category because Weather Zone 4 was also part of the EIA model.

As discussed in Sect. 3, "Methodology," regression diagnostics included the following: checking the normality of the residuals (i.e., stem-and-leaf plot, skewness, and kurtosis), analyzing the residual plots per Draper and Smith (1981), and checking for influence, multicollinearity, and autocorrelation.

Table 5.4. Comparison of original EIA electricity regression models and Ornl reproductions which do not match

Raw Goods Industrial Estimates	EIA		ORNL
√Model MSE	10,175,950	<—>	10,430,451
intercept	650,186	<—>	-1,717,500
Sq ft	20		21
#Employees	25,926		25,313
Sq ft Cool	-67		-66
Enduse 3	5,813,467		5,759,654
CLIMZON4	-5,284,553	<—>	2,714,744

Retail Estimates	EIA		ORNL
√Model MSE	2,615,693	<—>	2,656,937
intercept	-48,514	<—>	176,036
#Floors	-142,705	<—>	-109,999
Sq ft	2		2
#Employees	3909		3858
Sq ft Heat	4		4
Enduse 6	1,330,309		1,385,460
CLIMZON3	1,073,217	<—>	-96,058
BLCL1	1,605,570		1,696,822

Assembly Building Estimates	EIA		ORNL
√Model MSE	1,762,321	<—>	1,706,367
intercept	232,331		272,186
#Floors	-233,002		-255,122
Sq ft	18		19
#Employees	13,154		15,008
Sq ft Cool	-16		-18
Enduse 5	-3,139,396		-2,994,450
BLCL2	1,584,706		450,367

Food Sales Estimates	EIA		ORNL
√Model MSE	280,030		283,633
intercept	-294,758		-330,070
HDD80	-12		-21
Sq ft Res	47		47
#Emp. Category	164,966		162,764
Sq ft Category	72,852		71,396
Sq ft Heat	22		22
CLIMZON4	-116,976	<———>	42,080
BLCL1	231,580		233,354

Table 5.5. Comparison of original EIA natural gas regression models and ORNL reporductions which do not match

Education Estimates	EIA		ORNL
$\sqrt{\text{Model MSE}}$	6,717,425	<—>	6,773,306
intercept	-1,951,660		-2,060,902
HDD60	400		592
CDD80	-138,765		-146,683
#Employees	17,670		19,504
Sq ft Heat	55		54
Sq ft Cool	30		30
CLIMZON4	2,169,726	<—>	-57,654

Health Care > 350,000 Sq Ft Estimates	EIA		ORNL
$\sqrt{\text{Model MSE}}$	56,363,236	<—>	56,596,762
intercept	4,269,622		4,225,377
Sq ft Heat	111		111
Sq ft Cool	153		152

Assembly Plant Estimates	EIA		ORNL
$\sqrt{\text{Model MSE}}$	27,224,909	<—>	27,283,539
intercept	-13,148,574		-12,829,157
Sq ft Heat	94		94
Hrs. Op.	258,355		253,185

Other Industrial Estimates	EIA		ORNL
$\sqrt{\text{Model MSE}}$	48,685,765	<—>	45,406,755
intercept	-28,886,858		-22,497,850
Sq ft	177		154
#Employees	-83,910		-61,084
Sq ft Heat	286		415
Hrs. Op.	370,859		177,171
Enduse 2	39,789,617		33,998,451

Table 5.5 (Continued)

Raw Goods Industrial Estimates	EIA		ORNL
√Model MSE	69,571,526	<—>	68,309,947
intercept	-9,455,862		-9,979,970
#Employees	336,438		339,573
Sq ft Vacant	5,451		5,744
Enduse 6	-40,988,112		-54,253,244

Mixed Retail/ Wholesale Estimates	EIA		ORNL
√Model MSE	404,476	<—>	445,946
intercept	-1,163,712		-1,126,505
#Emp. Cat	365,730		423,830
Sq ft Cat	307,718		343,471
Enduse 4	-1,782,810		-2,551,086
Enduse 5	2,703,130		2,125,161
CLIMZON4	458,411	<—>	--
Weather Zone 4	-782,417		-989,865
Region 3	846,791		637,990
BLCL1	3,029,826		3,075,624

Mixed Use Office Estimates	EIA		ORNL
√Model MSE	13,606,025	<—>	13,667,603
intercept	12,443,311		12,438,571
#Floors	-1,382,704		-1,381,497
#Employees	-39,181		-39,196
Sq ft Heat	255		255
Enduse 1	-11,981,048		-11,968,403

Residential Mixed Use Estimates	EIA		ORNL
√Model MSE	2,664,595	<—>	2,186,835
intercept	-2,546,185	<—>	-1,582,715
#Floors	1,003,092	<—>	502,804
Sq ft	-256	<—>	-382
Sq ft Heat	290	<—>	228
#Employees	157,314	<—>	199,616
Sq ft Res	-21	<—>	309
Hrs Op	-18,027	<—>	-5,519
Enduse 4	-573,584	<—>	-613,650
Weather Zone 3	1,524,318	<—>	242,870

Table 5.5 (Continued)

Shopping Estimates	EIA		ORNL
$\sqrt{\text{Model MSE}}$	9,150,223	<—>	9,095,416
intercept	-15,283,287		-15,224,248
Sq ft Cat.	2,800,204		2,785,764
Sq ft Vacant	175		176
Enduse 5	35,699,807		35,711,711
Weather Zone 2	5,314,035		5,358,051

Retail Sales > 3 Floors Estimates	EIA		ORNL
$\sqrt{\text{Model MSE}}$	964,229	<—>	1,013,160
intercept	291,263		221,009
#Employees	3,283		2,875
Sq ft Heat	25		25
Weather Zone 5	-1,259,121		-1,158,884

Retail Sales < 3 Floors Estimates	EIA		ORNL
$\sqrt{\text{Model MSE}}$	523,622		535,915
intercept	-86,423	<—>	172,781
CDD60	312		295
#Emp. Cat.	214,577		218,343
Sq ft Heat	23		24
Sq ft Cool	-40		-38
Enduse 4	-4,436,995		-4,392,505
Enduse 6	-298,427		-350,732
CLIMZON5	-303,964		-150,966
Weather Zone 2	260,159		338,896

For a normal distribution, the stem-and-leaf plot should resemble a symmetric, bell-shaped curve; the skewness and kurtosis statistics should equal 0. The Durbin-Watson D-statistic for autocorrelation should be about 2 (indicating that $\rho = 0$). These results are summarized in Table 5.6. If the skewness, kurtosis, and Durbin-Watson D-statistics are significant at the $\alpha = 0.05$ level then "yes" is entered in the respective column of Table 5.6. For kurtosis, it is noted whether the distribution is "pointed" with most values clustered at the center or "flat" with more values at the shoulders of the distribution than would be observed in a normal distribution. The conclusion for or against a normally distributed error term is based on the stem-and-leaf plot, skewness, and kurtosis statistics. The Durbin-Watson D-statistic for autocorrelation will provide one of two conclusions: autocorrelation or inconclusive results. For significant autocorrelation, the estimated correlation coefficient $\hat{\rho}$ is provided. The NBECS data are not time-series data and should not inhibit significant autocorrelation; however, the D-statistic may be significant when true autocorrelation exists or when important variables are missing from the models. The latter case is assumed in this study. In general, the error terms of the models were not normally distributed and were pointed, skewed distributions. Detectable autocorrelation existed in only five cases, and the estimated values of ρ are listed.

Multicollinearity (or collinearity) is difficult to diagnose but can be detected via such analyses as correlation coefficients and the method of Belsley, Kuh, and Welsch (1980). The collinearity results are not summarized in a table because so few regression categories had condition indices of 10 or more. Additionally, because the regression equations are intended for prediction purposes and not to calculate coefficients for the purpose of elasticity estimation, multicollinearity is a reduced problem as long as satisfactory predictions are possible. The stepwise regression used in this analysis also helped to reduce the incidence of selecting variables that would show multicollinearity.

With respect to the error term, a plot of the residuals versus the corresponding predicted values should reveal a horizontal band if the regression model has the correct form [see Draper and Smith (1981)]. Additional plots of the residuals vs each independent variable also helped to pinpoint any problem areas. For this study, the plots revealed nonhorizontal, erratic bands; skewed patterns with many negative residuals; and a few somewhat-fanned patterns. Fanned patterns might normally suggest the need for WLS analysis, but these patterns appeared in only three or four category plots. Because of this, the WLS analysis as a methodology to be used for all building regression categories was dropped in favor of nonlinear and transformation analyses. The plots did not indicate the need for any linear polynomial models.

Table 5.6. Diagnostic statistics for normality and autocorrelation for the ordinary least squares models in Tables 5.1 and 5.2

Regression Category	Stem-and-leaf plot	Skewness	Kurtosis	Conclusion: normal error	Autocorrelation present
Natural gas					
Assembly	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Education	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Food sales	Symmetric, pointed	No	Yes, pointed	No	Not conclusive
Health care, <350,000 sq ft	Skewed	Yes	No	No	Not conclusive
Health care, ≥350,000 sq ft	Skewed, pointed	Yes	Yes, pointed	No	Yes, $\hat{\rho}$ = 0.248
Assembly plants	Symmetric, pointed	No	Yes, pointed	No	Not conclusive
Raw goods, industrial	Symmetric, pointed	Yes	Yes, pointed	No	Not conclusive
Other industrial	Skewed	No	Yes, undetermined	No	Not conclusive
Shopping centers	Symmetric, pointed	No	Yes, pointed	No	Not conclusive
Retail sales, <3 floors	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Retail sales ≥3 floors	Symmetric, pointed	No	Yes, pointed	No	Not conclusive
Personal services	Skewed, pointed	No	Yes, pointed	No	Not conclusive
Other buildings	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Mixed retail/ wholesale	Symmetric	No	No	Yes	Not conclusive
Auto sales/ service	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
General office	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Professional office, <75 employees	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Professional office ≥75 employees	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Financial office	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Mixed-use office	Symmetric, pointed	No	Yes, pointed	No	Not conclusive
Residential	Symmetric, bell	No	No	Yes	Not conclusive
Residential mixed use	Symmetric, pointed	Yes	Yes, pointed	No	Not conclusive

Table 5.6. (Continued)

Regression Category	Stem-and-leaf plot	Skewness	Kurtosis	Conclusion: normal error	Autocorrelation present
Commercial lodging	Truncated, pointed	Yes	Yes, pointed	No	Not conclusive
Other lodging	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Refrigerated warehouses	Symmetric, pointed	Yes	Yes, pointed	No	Not conclusive
Nonrefrigerated warehouses	Symmetric, pointed	No	Yes, pointed	No	Not conclusive
<u>Electricity</u>					
Assembly	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Education	Skewed, pointed	Yes	Yes, pointed	No	Yes, $\hat{\rho} = 0.22$
Food sales	Skewed	Yes	Yes, undetermined	No	Not conclusive
Assembly plants	Symmetric, pointed	Yes	Yes, pointed	No	Not conclusive
Raw goods industrial	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Other industrial	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Retail sales/ services	Symmetric, pointed	Yes	Yes, pointed	No	Yes, $\hat{\rho} = 0.22$
Auto sales/ services	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
General office	Symmetric, pointed	Yes	Yes, pointed	No	Not conclusive
Professional office	Skewed, pointed	Yes	Yes, pointed	No	Yes, $\hat{\rho} = -0.18$
Financial office	Skewed, pointed	Yes	Yes, pointed	No	Yes, $\hat{\rho} = 0.36$
Mixed use office	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Residential	Slightly Skewed	Yes	Yes, flat	No	Not conclusive
Mixed use Residential	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Lodging	Skewed[a]	Yes	Yes, pointed	No	Not conclusive
Warehouse storage	Skewed, pointed	Yes	Yes, pointed	No	Not conclusive
Other	Skewed,[a] pointed	Yes	Yes, pointed	No	Not conclusive

[a]One or two large residuals seem to cause the skewed distribution.

In summary, the regression models previously developed by the EIA provide important information relative to delineating building categories and indicating important climatic and building dummy variables. The models, however, produced negative imputed values and nonnormal error terms. The diagnostics showed that while the error terms were nonnormal, problems such as autocorrelation and multicollinearity are of no serious consequence. Finally, the residual analysis indicated that neither polynomial nor WLS techniques are necessary. Nonlinear regression techniques then became a starting point for the 1983 NBECS regression model methodology.

5.2 DATA STRUCTURE STUDY

The data structure study is a continuation of the data base development study discussed in Sect. 4 and the EIA model development reviewed in Sect. 5.1. Familiarity with and understanding of data bases are imperative if modeling efforts are to succeed. This section discusses additional analyses and suggests exercises for "uncovering" variables that should be included in the regression model development.

The structure of the NBECS data set was studied to learn more about data dependencies and to include more factors that are potentially important, especially from an engineering standpoint. The importance of the variables included in the EIA models of Sect. 5.1 was reviewed (Sect. 5.2.1); the distributions of the independent variables were checked (Sect. 5.2.1); and additional conservation variables that should be included in the model building were tabulated (Sect. 5.2.2).

5.2.1 Importance of the EIA Models

The EIA models are the result of Statistical Analysis System (SAS) stepwise regression analyses on this set of 24 variables:

1. heating degree days,
2. cooling degree days,
3. estimated year of construction,
4. number of floors,
5. square footage,
6. estimated square footage (interval recode),
7. square footage heated by this fuel,
8. square footage cooled by this fuel,
9. square footage, residential,
10. square footage vacant during previous year,
11. number of employees,
12. estimated number of employees (interval recode),
13. weekly number of hours of operation,
14. fuel used for space heating (Yes, No),
15. fuel used for air conditioning (Yes, No),
16. fuel used for water hearing (Yes, No),
17. fuel used for electricity generation (Yes, No),
18. fuel used for manufacturing (Yes, No),
19. fuel used for cooking (Yes, No),

20. census region (coded as a set of dummy variables),
21. weather zone (coded as a set of dummy variables),
22. climate zone (coded as a set of dummy variables),
23. percent glass on outside walls, and
24. detailed four-digit building code (dummy variable).

The models' error terms are not normally distributed, so the statistics and significance levels used by the stepwise techniques are incorrect (an unknown amount). These models do, however, provide information on potentially important variables that should be included in future modeling efforts.

Table 5.7 ranks the independent variables by the number of times they appeared in building regression models. Rankings are provided for the electricity, natural gas, and combined data sets. Next to each variable is the number of models in which it appeared (i.e., its frequency). Both the electricity and natural gas data sets seem to have a break point between variables with a frequency of eight or more and the remaining variables. The four most frequently used modeling variables for both natural gas and electricity were: total square feet, square feet cooled, square feet heated, and number of employees. A regression equation that tries to form the basis of a model for all building categories needs to contain these four variables as a minimum. For natural gas models, the variable ENDUSE5 (fuel used for manufacturing) is also included in the top ranks. Next in frequency were the following variables: number of floors, hours of operation, CDDs, HDDs, and square feet vacant. These EIA models, summarized in Tables 5.1 through 5.3, show for which building categories additional variables like Region i, CLIMZONi, and the building dummies (e.g., BLCL1) are important.

The regression model described in Sect. 5.3 is designed to be a basis for developing final models for each electricity and natural gas building category. The variables included in this model are: ENDUSEi, heating degree days (65°F), cooling degree days (65°F), square feet heated, square feet cooled, number of floors, percent of glass, weekly number of hours of operation, number of employees, and square feet. The additional independent variables listed in Tables 5.7 and 5.3 should be included in the final model formulation and testing. Section 5.2.3 notes additional conservation variables that should be considered.

The continuous independent variables were analyzed for potential nonnormal distributions for each regression category. Each, in fact, was nonnormal, and most showed very skewed distributions.

Table 5.7. Frequency of appearance of independent variables in the EIA building category models (variable name/frequency)[a]

Electricity models		Natural gas models		Combined models	
Employees (no.)	14	Sq feet heat	19	Employees (no.)	28
Sq ft	13	Employees (no.)	14	Sq ft heat	28
Sq ft cool	10	Sq ft cool	10	Sq ft	22
Sq ft heat	9	Sq ft	9	Sq ft cool	20
-------------------		ENDUSE5	8	-----------	
HDD	5	----------------		Floors (no.)	10
Floors (no.)	5	CDD	5	Sq ft vacant	10
Sq ft vacant	5	Floors (no.)	5	Oper. (h/w)	9
Oper. (h/w)	4	Oper.(h/w)	5	ENDUSE5	9
CLIMZONi	4	Sq ft vacant	5	CLIMZONi	8
Weather zone i	3	Weather zone i	5	Region i	8
Region i	3	Region i	5	Weather zone i	8
Glass (%)	3	CLIMZONi	4	CDD	7
Bldg Age	2	Sq ft res	4	HDD	7
CDD	2	ENDUSE4	3	Sq ft res	6
Sq ft res	2	ENDUSE6	3	Glass (%)	4
ENDUSE1	1	HDD	2	ENDUSE4	4
ENDUSE2	1	ENDUSE1	2	ENDUSE6	4
ENDUSE3	1	ENDUSE2	2	Bldg Age	3
ENDUSE4	1	TD7	1	ENDUSE1	3
ENDUSE5	1	TC4	1	ENDUSE2	3
ENDUSE6	1	Glass (%)	1	TC4	1
		Bldg Age	1	TD7	1
				ENDUSE3	1

[a]Definitions:

Employees (no.) - Number of employees.
Sq ft - square feet (ft^2).
Sq ft cool - Number of square feet cooled.
Sq ft heat - Number of square feet heated.
HDD - Heating degree days.
Floors (no.) - Number of floors in building.
Sq ft vacant - Unoccupied square footage.
Oper.(h/w) - Weekly number of hours of operation.
CLIMZONi - Climate zone depending on weather zone, based on heating degree days for 1979.
Weather zone i - Weather zone based on a 40-year average of heating degree days and cooling degree days.
Glass (%) - Percent glass on outside walls.
Bldg age - Number of years of building existence.
CDD - Cooling degree days.
res - Residential.
ENDUSE1, TD7, and TCA: See Table 5.3.

5.2.2 Conservation Variables

The frequency counts of additional conservation variables (dummy variables) that can be added to the regression model in Sect. 5.3 are presented in Tables 5.8 and 5.9 for natural gas and electricity, respectively. Large buildings are included in these calculations. These show the types of conservation activities that are employed in nonresidential buildings and provide limited information on heating and cooling equipment. Table 5.10 defines the variable names used in Tables 5.8 and 5.9.

An example of how to read the tables is given for Table 5.8 and building category 20 (Education). The Education (20) column contains frequency counts for each variable. The frequency is increased each time the respondent's answer to the question is "Yes" for all variables but the exceptions (which are defined subsequently). If two variables are listed e.g., FOCH1 x EDUSE1, the respondent must have answered "Yes" to both questions before the frequency count is increased. The "5" corresponding to FOCH1 x ENDUSE1 means that five buildings use natural gas for heating and converted from fuel oil to some other fuel for heating. It is assumed, perhaps incorrectly, that the change might be to natural gas. The exceptions are described next. ENDUSE2 X CST1 gives the name of the air conditioning types with nonzero frequency counts. INSULATE states the number of respondents (out of the total number in the building category) that did not add any insulation: "0" - 338/417. ENDUSE1 x %GLASS and ENDUSE2 x %GLASS rank the most frequent to least frequent responses for the percent glass question. For education buildings, for example, the rank for ENDUSE1 x %GLASS is: "0," "3," "4," "2," indicating that most buildings did not use natural gas for heating ("0"). Of those that did use natural gas for heating, most had 25% to 50% glass on exterior walls ("3") followed by 0% to 25% ("4"), and the smallest number of buildings had 25% to 50% exterior glass ("2"). For SQ FT RES the quantity "0" - 444/448" means that 444 of 448 respondents reported that no percentage of the building was used for residential purposes. The SQ FT VACANT item is interpreted similarly.

Not all heat generation and distribution and air conditioning dummy variables (see Table 5.10) need to be added to a regression model just because they have sizeable nonzero frequencies. For heat distribution systems, either forced air system (CAU1) (i.e., air handling units with self-contained fans that distribute heat to only part of the building or single central air handling units separate from the energy conversion system) should be about equally efficient (although the central system may be slightly more efficient) and can be collapsed into one dummy variable. However, the forced air systems will generally be more efficient than the radiant or naturally circulated air systems: electric baseboards (EB1), baseboard heating using hot water (HWB1), baseboard heating using steam (SB1), radiators or convectors (RAD1), or heating panels in the walls or floor (WOFP1). If WOFP1 is electric, then EB1 and WOFP1 should have about the same efficiency and thus be represented by one dummy variable. HWB1 and SB1 should have approximately the same efficiency and be less efficient,

Table 5.8 Frequency counts of additional conservation variables for natural gas data[a]

Variables	Building Regression Category (Category Number) Assembly (10)	Education (20)	Food Sales (30)	Health ≥ 350,000 sq ft (40)	Health < 350,000 sq ft (50)
FOCH1 x ENDUSE1	2	5	0	1	1
FOCAC1 x ENDUSE2	0	0	0	1	0
FOCW1 x ENDUSE3	0	1	0	1	0
FOCG1 x ENDUSE4	0	0	0	0	0
FOCM1 x ENDUSE5	0	0	0	0	0
FOCC1 x ENDUSE6	0	0	0	0	0
ENDUSE1 x VHCR1	10	29	14	5	2
ENDUSE1 x RESNHR1	1	1	2	1	1
ENDUSE1 x NRNHR1	168	241	96	21	29
ENDUSE2 x VHCR1	2	6	0	2	0
ENDUSE2 x RESNCR1	0	0	0	0	0
ENDUSE2 x NRNCR1	14	24	12	9	7
ENDUSE1 x CPU1	87	81	49	19	34
ENDUSE2 x CST1	Ctrl	Ctrl;pkg.	Ctrl;pkg.	Ctrl;pkg.	Ctrl;pkg.
ENDUSE1 x CONU1	162	222	67	30	44
ENDUSE1 x COFFU1	28	47	9	32	27
ENDUSE1 x HCCON1	73	140	31	44	39
ENDUSE2 x HCCON1	10	18	5	19	10
ENDUSE3 x HWB1	17	41	7	8	8
INSULATE	Few	"0"=338/417	"0"=121/203	"0"=58/85	"0"=71/92
LTCON1	83	189	60	63	46
ENDU[illegible]1 x OHD1	46	95	24	29	34
ENDUSE1 x OTU1	9	3	5	0	0
ENDUSE1 x %GLASS		"0","3","4","2"	"4","3","0"	"3","0","4","2"	"3","4","0"
ENDUSE2 x %GLASS		"0","4","3"	"0","4"	"0","3","2"	"0","3"
ENDUSE1 x RAD1	47	103	17	26	18
SQ FT RES	"0"=283/291	"0"=444/448	"0"=186/203	"0"=82/86	"0"=94/98
SQ FT VACANT	"0"=275/291	"0"=394/444	"0"=182/202	"0"=66/86	"0"=87/99
ENDUSE1 x SB1	5	28	2	8	4
ENDUSE1 x SONU1	64	63	96	2	14

Table 5.8 (Continued)

	Assembly Plants (60)	Raw Goods Industrial (70)	Other Industrial (80)	Shopping Centers (90)	Retail ≥ 3 Floors (100)	Retail < 3 Floors (110)
FOCH1 x ENDUSE1	0	3	4	0	1	0
FOCAC1 x ENDUSE2	0	0	0	0	0	0
FOCW1 x ENDUSE3	0	1	1	0	0	0
FOCG1 x ENDUSE4	0	0	0	0	0	0
FOCM1 x ENDUSE5	0	3	1	0	0	0
FOCC1 x ENDUSE6	0	1	0	0	0	0
ENDUSE1 x VHCR1	2	0	3	31	7	9
ENDUSE1 x RESNHR1	0	0	0	0	0	0
ENDUSE1 x NRNHR1	93	71	48	56	53	118
ENDUSE2 x VHCR1	0	0	0	2	2	1
ENDUSE2 x RESNCR1	0	0	0	0	0	0
ENDUSE2 x NRNCR1	11	10	6	6	7	9
ENDUSE1 x CAU1	26	27	22	29	17	44
ENDUSE2 x CST1	Ctrl;pkg.	Ctrl;pkg.	Ctrl;pkg.	Ctrl;pkg.	Ctrl;pkg.	Ctrl;pkg.
ENDUSE1 x CONU1	71	59	38	33	50	67
ENDUSE1 x COFFU1	23	9	12	5	5	4
ENDUSE1 x HCCON1	60	40	32	38	11	37
ENDUSE2 x HCCON1	6	4	6	4	2	1
ENDUSE3 x HWB1	4	7	4	1	3	5
INSULATE	"0"=105/182	"0"=99/142	"0"=72/112	"0"=80/122	"0"=54/83	"0"=129/176
LTCON1	57	50	45	74	34	72
ENDUSE1 x OHD1	52	37	29	15	18	21
ENDUSE1 x OTU1	7	4	6	2	4	2
ENDUSE1 x %GLASS	"4","3","0"	"4","0","3"	"4","0","3"	"4","0","3"	"4","0","3"	"4","3",0-2
ENDUSE2 x %GLASS	"0","4"	"0","4"	"0","4"	"0","4"	"0","4"	"0","4"
ENDUSE1 x RAD1	31	17	11	4	22	6
SQ FT RES	"0"=179/190	"0"=1/145	"0"=0/145	"0"=1/131	"0"=80/95	"0"=0/193
SQ FT VACANT	"0"=179/190	"0"=139/145	"0"=111/116	"0"=77/131	"0"=76/94	"0"=175/193
ENDUSE1 x SB1	8	7	2	1	1	1
ENDUSE1 x SONU1	91	72	53	80	31	120

Table 5.8 (Continued)

	Personal Service (120)	Mxd. Ret./Whole. (130)	Auto Sales (140)	General Office (150)	Professional ≥ 75 Employees (160)
FOCH1 x ENDUSE1	0	1	0	0	2
FOCAC1 x ENDUSE2	0	0	0	0	0
FOCW1 x ENDUSE3	0	0	0	0	0
FOCG1 x ENDUSE4	0	0	0	0	0
FOCM1 x ENDUSE5	0	0	0	0	0
FOCC1 x ENDUSE6	0	0	0	0	1
ENDUSE1 x VHCR1	5	4	2	7	25
ENDUSE1 x RESNHR1	0	1	0	0	0
ENDUSE1 x NRNHR1	70	31	90	51	77
ENDUSE2 x VHCR1	1	0	1	1	1
ENDUSE2 x RESNCR1	0	0	0	0	0
ENDUSE2 x NRNCR1	5	3	4	7	9
ENDUSE1 x CAU1	17	13	27	23	36
ENDUSE2 x CST1	Ctrl	End.2[a] only	End.2 only	Ctrl;pkg.	End.2;Ctrl only
ENDUSE1 x CONU1	43	25	57	43	76
ENDUSE1 x COFFU1	6	0	1	11	19
ENDUSE1 x HCCON1	21	6	29	31	64
ENDUSE2 x HCCON1	2	1	2	6	6
ENDUSE3 x HWB1	1	1	1	7	13
INSULATE	"0"=88/111	"0"=30/49	"0"=100/135	"0"=68/101	"0"=134/185
LTCON1	38	14	45	45	107
ENDUSE1 x OHD1	21	8	27	16	33
ENDUSE1 x OTU1	3	1	7	3	6
ENDUSE1 x %GLASS	"4","0","3"	"4","3"	"4","3","2"	"4","0","3"	Even
ENDUSE2 x %GLASS	"0"	"0","4"	"0"	"0"	"0"
ENDUSE1 x RAD1	13	9	7	15	24
SQ FT RES	"0"=5/120	"0"=40/47	"0"=1/141	"0"=4/107	"0"=5/192
SQ FT VACANT	"0"=111/120	"0"=41/49	"0"=135/141	"0"=80/107	"0"=133/192
ENDUSE1 x SB1	0	0	0	3	5
ENDUSE1 x SONU1	51	25	78	26	28

[a] ENDUSE 2 - all buildings using natural gas for air conditioning use the same type of equipment.

Table 5.8 (Continued)

	Professional < 75 Employees (170)	Financial Office (180)	Mxd. Use Office (190)	Residential Subset (200)	Mxd. Use Residential (210)
FOCH1 x ENDUSE1	2	1	2	0	2
FOCAC1 x ENDUSE2	0	0	0	0	0
FOCW1 x ENDUSE3	0	0	1	0	0
FOCG1 x ENDUSE4	0	0	0	0	0
FOCM1 x ENDUSE5	0	0	0	0	0
FOCC1 x ENDUSE6	0	0	1	0	0
ENDUSE1 x VHCR1	14	12	17	7	22
ENDUSE1 x RESNHR1	1	1	2	5	11
ENDUSE1 x NRNHR1	126	70	79	27	69
ENDUSE2 x VHCR1	0	1	2	0	2
ENDUSE2 x RESNCR1	0	0	0	0	1
ENDUSE2 x NRNCR1	16	11	9	1	8
ENDUSE1 x CAU1	63	32	32	10	23
ENDUSE2 x CST1	Ctrl;pkg.	Ctrl;pkg.	Ctrl;pkg.	Ctrl (only 1)	Few of each
ENDUSE) x CONU1	107	63	66	35	102
ENDUSE1 x COFFU1	12	2	15	6	7
ENDUSE1 x HCCON1	45	38	52	18	27
ENDUSE2 x HCCON1	4	8	9	0	2
ENDUSE3 x HWB1	13	9	9	3	12
INSULATE	"0"=140/192	"0"=79/99	"0"=87/143	"0"=48/74	"0"=87/154
LTCON1	68	57	58	31	43
ENDUSE1 x OHD1	19	15	32	12	37
ENDUSE1 x OTU1	4	2	3	4	2
ENDUSE1 x %GLASS	"4","3","0"	"4","3","0","2"	"4","3","0","2"	"4","0","3"	"4","0","3"
ENDUSE2 x %GLASS	"0","4"	"0"	"0"	"0"	"0"
ENDUSE1 x RAD1	27	10	29	27	44
SQ FT RES	"0"=195/201	"0"=2/107	"0"=127/155		
SQ FT VACANT	"0"=170/202	"0"=83/107			
ENDUSE1 x SB1	3	2			
ENDUSE1 x SONU1	62	26			

Table 5.8 (Continued)

	Commercial Lodging (220)	Long Term Lodging (230)	Refrigerated Warehouse (240)	Nonrefrig. Warehouse (250)	Other (260)
FOCH1 x ENDUSE1	0	1	2	0	0
FOCAC1 x ENDUSE2	0	0	0	0	0
FOCW1 x ENDUSE3	0	0	0	0	0
FOCG1 x ENDUSE4	0	0	0	0	0
FOCM1 x ENDUSE5	0	0	1	0	0
FOCC1 x ENDUSE6	0	0	0	0	0
ENDUSE1 x VHCR1	3	4	8	12	18
ENDUSE1 x RESNHR1	0	0	0	0	3
ENDUSE1 x NRNHR1	26	14	70	86	95
ENDUSE2 x VHCR1	1	1	2	2	2
ENDUSE2 x RESNCR1	0	0	0	0	0
ENDUSE2 x NRNCR1	6	1	4	13	6
ENDUSE1 x CAU1	6	7	22	26	29
ENDUSE2 x CST1	Few of each	Few of each	Ctrl;pkg.	P&C[a]	P&C
ENDUSE1 x CONU1	30	28	57	63	101
ENDUSE1 x COFFU1	5	10	23	5	16
ENDUSE1 x HCCON1	19	1	2	34	53
ENDUSE2 x HCCON1	1	1	2	5	4
ENDUSE3 x HWB1	3	11	1	4	17
INSULATE	"0"=65/90	"0"=52/61	"0"=94/143	"0"=102/154	"0"=165/223
LTCON1	52	34	48	51	85
ENDUSE1 x OHD1	14	7	32	38	35
ENDUSE1 x OTU1	1	1	3	5	8
ENDUSE1 x %GLASS	"0","4","3"	"0","4","3"	"4","3","0"	"4","3","0"	"4","0","3","2"
ENDUSE2 x %GLASS	"0"	"0"	"0"	"0","4"	"0","4"
ENDUSE1 x RAD1	8	10	18	13	39
SQ FT RES			"0"=141/143	"0"=153/154	
SQ FT VACANT			"0"=126/143	"0"=138/154	
ENDUSE1 x SB1			2	3	
ENDUSE1 x SONU1			77	95	

[a] Pkg-Ctrl combination.

Table 5.9. Frequency counts of additional conservation variables for electricity data[a]

	Building Regression Category (Category Number)					
Variables	Assembly (270)	Education (280)	Food Sales (290)	Health Care (300)	Assembly Plants (310)	Raw Goods Industrial (320)
FOCH1 x ENDUSE1	0	0	0	0	1	0
FOCAC1 x ENDUSE2	0	0	0	0	0	2
FOCW1 x ENDUSE3	0	0	0	0	1	0
FOCG1 x ENDUSE4	0	0	0	0	0	0
FOCM1 x ENDUSE5	0	0	0	0	0	3
FOCC1 x ENDUSE6	0	0	0	0	0	1
ENDUSE1 x VHCR1	5	10	1	1	1	1
ENDUSE1 x RESNHR1	0	1	1	0	0	0
ENDUSE1 x NRNHR1	65	64	46	15	27	18
ENDUSE2 x VHCR1	10	37	8	10	2	0
ENDUSE2 x RESNCR1	0	1	0	0	0	0
ENDUSE2 x NRNCR1	174	202	109	53	93	77
ENDUSE1 x CAU1	33	31	21	13	12	4
ENDUSE2 x CST1	WPC[a]	WPC	WPC	WPC	WPC	WPC;W&P[b]
ENDUSE1 x CONU1	42	41	[illegible]	14	19	15
ENDUSE1 x CUFFU1	9	16	6	12	4	0
ENDUSE1 x HCCON1	27	49	25	22	22	13
ENDUSE2 x HCCON1	90	197	58	97	81	57
ENDUSE3 x HWB1	7	12	3	2	2	3
INSULATE	"0"=301/414	"0"=490/591	"0"=214/309	"0"=149/204	"0"=144/237	"0"=133/187
LTCON1	133	247	102	122	79	69
ENDUSE1 x OHU1	24	39	22	16	20	16
ENDUSE1 x OTU1	5	4	5	2	4	0
ENDUSE1 x %GLASS	"0","4"	"0","4"	"0","3","4"	"0"	"0","4"	"0","4"
ENDUSE2 x %GLASS	"0","4"	"0","3","4"	"0","2","3","4"	"4","3","2"	"4","3","0"	"4","0","3"
ENDUSE1 x RAD1	9	15	5	10	5	2
SQ FT RES	"0"=435/445	"0"=622	"0"=318/340	"0"=202/211	"0"=1/248	"0"=1/192
SQ FT VACANT	"0"=410/445	"0"=541/623	"0"=302/338	"0"=1[illegible]4/212	"0"=235/248	"0"=186/192
ENDUSE1 x SB1	1	4	0	3	2	2
ENDUSE1 x SONY1	56	65	64	12	42	29

[a] Window; Pkg and Ctrl

[b] Window-Pkg combination

Table 5.9 (Continued)

	Other Industrial (330)	Retail Sales/Service (340)	Auto Sales (350)	General Office (360)	Professional Office (370)	Financial Office (380)
FOCH1 x ENDUSE1	0	0	0	0	3	0
FOCAL'1 x ENDUSE2	0	0	0	0	2	0
FOCW1 x ENDUSE3	0	0	0	0	1	0
FOCG1 x ENDUSE4	0	0	0	0	1	0
FOCM1 x ENDUSE5	1	0	0	0	0	0
FOCC1 x ENDUSE6	0	0	0	0	1	0
ENDUSE1 x VHCR1	0	23	0	5	30	8
ENDUSE1 x RESNHR1	0	1	0	0	0	0
ENDUSE1 x NRNHR1	14	105	26	28	104	48
ENDUSE2 x VHCR1	4	73	1	21	67	15
ENDUSE2 x RESNCR1	0	0	0	1	0	2
ENDUSE2 x NRNCR1	57	312	32	84	309	107
ENDUSE1 x CAU1	9	57	9	12	67	45
ENDUSE2 x CS/1	WPC[a]	WPC	WPC	WPC	Every type	WPC
ENDUSE1 x CONU1	6	44	9	18	48	36
ENDUSE1 x COFFU1	5	8	0	3	39	7
ENDUSE1 x HCCON1	9	74	8	21	104	41
ENDUSE2 x HCCON1	40	181	23	68	251	80
ENDUSE3 x HWB1	2	10	0	6	12	4
INSULATE	"0"=103/158	"0"=557/774	"0"=175/237	"0"=116/161	"0"=443/606	"0"=141/191
LTCON1	60	332	67	74	288	99
ENDUSE1 x OHD1	6	43	15	10	50	19
ENDUSE1 x OTU1	2	11	4	3	17	5
ENDUSE1 x %GLASS	"0","4"	"4"	"0","4","3"	"0","4","3"	"0","4","3","2"	"0","4","3"
ENDUSE2 x %GLASS	"4","0","3"	"0"	"0","4","3"	Even	"4","3","0","2"	Even
ENDUSE1 x RAD1	1	3	0	4	7	5
SQ FT RES	"0"=1/167	"0"=810/846	"0"=3/249	"0"=5/172	"0"=625/639	"0"=2/196
SQ FT VACANT	"0"=160/167	"0"=698/840	"0"=238/249	"0"=128/172	"0"=494/636	"0"=163/196
ENDUSE1 x SB1	0	0	0	1	2	0
ENDUSE1 x SONY1	16	189	31	26	115	44

[a] Window; Pkg and Ctrl

Table 5.9 (Continued)

	Mxd. Use Office (390)	Residential Subset (400)	Mxd. Use Residential (410)	Lodging (420)	Warehouse/ Storage (430)	Other Buildings (440)
FOCH1 x ENDUSE1	0	0	0	0	0	1
FOCAC1 x ENDUSE2	1	0	0	0	0	0
FOCW1 x ENDUSE3	0	0	0	0	0	1
FOCG1 x ENDUSE4	0	0	0	0	0	0
FOCM1 x ENDUSE5	0	0	0	0	0	0
FOCC1 x ENDUSE6	0	0	0	0	0	0
ENDUSE1 x VHCR1	5	5	4	6	6	10
ENDUSE1 x RESNHR1	0	0	0	0	0	0
ENDUSE1 x NRNHR1	28	11	18	42	60	41
ENDUSE2 x VHCR1	20	13	24	9	18	28
ENDUSE2 x RESNCR1	1	0	7	1	1	0
ENDUSE2 x NRNCR1	93	17	47	51	139	114
ENDUSE1 x CAU1	18	6	9	12	27	28
ENDUSE2 x CST1	WPC[a];W&P[b]	WPC;W&P	WPC;W&C[c]	WPC;W&P	WPC;W&P;P&C[d]	WPC;W&P;P&C
ENDUSE1 x CONU1	21	9	9	21	35	31
ENDUSE1 x COFFU1	8	2	1	1	8	10
ENDUSE1 x HCCON1	21	4	8	23	31	27
ENDUSE2 x HCCON1	76	23	33	52	87	94
ENDUSE3 x HWB1	3	1	3	6	6	11
INSULATE	"0"=121/196	"0"=78/113	"0"=107/194	"0"=160/210	"0"=367/491	"0"=318/404
LTCON1	84	35	56	107	167	134
ENDUSE1 x OHD1	16	6	9	24	33	39
ENDUSE1 x OTU1	6	4	2	2	4	7
ENDUSE1 x %GLASS	"0","4","3"	"0"	"0","4"	"0","4","3"	"0","4"	"0","4","3"
ENDUSE2 x %GLASS	"4","3","0","2"	"3","0","4"	"4","0","3"	"0","4","3"	"4","0","3"	"0","4","3"
ENDUSE1 x RAD1	6	5	0	2	6	13
SQ FT RES	"0"=174/210	"0"=2/131	"0"=7/213	"0"=186/227	"0"=527/533	"0"=410/434
SQ FT VACANT	"0"=165/208	"0"=104/131	"0"=163/213	"0"=197/233	"0"=472/531	"0"=290/430
ENDUSE1 x SB1	0	0	1	0	3	0
ENDUSE1 x SONY1	30	17	22	83	92	58

[a] Window; Pkg and Ctrl

[b] Window-Pkg combination

[c] Window-Ctrl combination

[d] Pkg-Ctrl combination

Table 5.10. Additional variable definitions for Tables 5.8 and 5.9

Variable name	NBECS[a] question no.	Description
FOCH1	70a	Converted from fuel oil to some other source for heating
FOCAC1	70a	Converted from fuel oil to some other source for space cooling
FOCG1	70a	Converted from fuel oil to some other source for water heating
FOCM1	70a	Converted from fuel oil to some other source for electricity generation
FOCC1	70a	Converted from fuel oil to some other source for manufacturing
VHCR1	63	Converted from fuel oil to some other source for cooking
RESNHR1	59	Night heat is reduced for residential areas
NRNHR1	52	Night heat is reduced for nonresidential areas
RESNCR1	57	Night cooling is reduced for residential areas
NRNCR1	60	Night cooling is reduced for nonresidential areas
CAU1	46c.I	Uses forced hot air as heat distribution system
CST1	54	Air conditioning system: (1) Window = Window only, (2) Pkg = One or more packaged units (i.e., built and assembled at a factory and installed as a unit at the building), (3) Ctrl = single central system
CONU1	46b	Uses a central system located in the building to generate heat but needs an additional system for heat distribution
COFFU1	46b	Uses a central system located outside the building to generate heat but needs an additional system for distribution
HCCON1	65	The building's heating or cooling systems have features designed to help conserve energy
EB1	46c.II	Uses electric baseboards to circulate heat
HWB1	46c.II	Uses baseboard heating with hot water to circulate heat
INSULATE	21	Year insulation was last added
LTCON1	67	The building's lighting system has features designed to help conserve energy
OHD1	46c.II	Uses some heat distribution method other than: CAU1, EB1, HWB1, SB1, RAD1, or WOFP1
OTU1	46b	Uses some heating method other than: SONU1, CONU1 or COFFU1
RAD1	46c.II	Uses radiators or convectors to circulate heat
SB1	46c.II	Uses baseboard to circulate steam heat
SONU1	46b	Uses a self-contained unit to generate and deliver heat. Unit may be internal or external to the building
%GLASS	22	Percent exterior glass: (1) 75% or more, (2) at least 50% but less than 75%, (3) at least 25% but less than 50%, and (4) less than 25%.

[a]NBECS = Nonresidential buildings energy consumption survey

and RAD1 should be least efficient. The rank efficiencies (highest to lowest) are:

1. forced air (CAU1),
2. electric baseboards (EB1) and wall or floor panels (WOFP1),
3. hot water baseboards (HWB1) and steam baseboards (SB1),
4. radiators (RAD1), and
5. other (OTU1).

For heat generation systems, natural gas is the predominant (about 3 to 1) fuel. The self-contained unit (SONU1) is the most efficient because there are no distribution losses. The central system located within the building (CONU1) is next on the efficiency list because there are no distance losses. The central system outside the building (COFFU1) suffers from distribution, distance, and environmental losses. Other (OTU1) is expected to be the least efficient. All of the heat generation variables, then, should have individual dummy variables for modeling purposes.

The cooling systems (CST1) involve three main types of equipment: window units, packaged units, and central systems. The types are listed in order of decreasing efficiency and should have separate dummy variables. However, for combinations of these systems, it is suggested that they be coded the same as the least efficient equipment in the combination. For example, window-package equipment can be included in the dummy variable for package equipment.

Frequency counts such as these provide good information about the data bases and about new variables that can be included in the fine-tuning of models for each building category. Tables 5.8 and 5.9 show that the fuel oil change variables have little to offer to model building because all the frequency counts are five or less, and many are zero. The dummy variables in Tables 5.8 and 5.9 can be added to the models if the frequency is large enough, say seven or more. If the frequency count is smaller, it may not be good to model the variable but to consider it as an explanatory factor for potential data anomalies. If the frequency count is too large (close to the sample size), it may also produce computational problems and should be excluded from the model. As an example of variable selection from Tables 5.8 and 5.9, the frequency counts underlined for refrigerated warehouses (building category 240) that use natural gas are underlined in Table 5.8. ENDUSE1 x %GLASS and ENDUSE2 x %GLASS are underlined because %GLASS is used in the model described in Sects. 5.3 and 5.4 but should not be included as an additional dummy variable. The INSULAT variable should be recoded to indicate the age of the insulation (i.e., 1979-INSULATE) or as a dummy indicating the presence of new insulation. The SQ FT VACANT variable (and SQ FT RES, when appropriate) should be multiplied by the appropriate fuel-use variable so that it will equal zero when the fuel is not used but equal the square feet when the fuel is used. Further study can suggest other variations.

5.3 ENGINEERING MODELS

The Efficiency and Renewables Research Section of the Energy Division at ORNL performs engineering studies concerning the energy consumption of residential and commercial buildings and appliances. These ORNL personnel guided the NBECS study team in using an engineering approach that attempts to provide regression models for each building category based on a generalized building model. This generalized model accounts for the basic energy flows into and out of buildings, and it was developed from basic engineering calculations and estimates. The regression models use the building model as a framework that is filled in for each building category. Here, we will only sketch the development of the engineering models without completely deriving or providing engineering justification for each. A good reference for these and other more elaborate building studies is **ASHRAE Handbook; 1981 Fundamentals**. The initial equations are presented in the Methodology Section, and Eqs. 3.2 through 3.4 are repeated here.

The heating component equation is:

$$Q_{heat} = \frac{24\ h/d \times HDD65}{E_H} \times [\underbrace{(U^* \times Ae)}_{Q_C} \times \underbrace{(1 - B_W \times \frac{G}{100})}_{Q_E}] \quad (3.2)$$

$$+ [\underbrace{AV_W \times \rho \times Cp \times A_W}_{Q_D}] - [\underbrace{(\lambda \times S) \times \mu \times A_W] - (P \times \mu)}_{Q_I}] ,$$

where

U^*	-	effective envelope heat transfer coefficient (Btu/h.ft^2°F),
Ae	-	surface area of envelope (ft^2),
AV_W	-	volumetric air flow rate of outside winter air per square foot of heated floor area (ft^3/hr.ft^2),
$\rho \times C_p$	-	conversion from ft^3/h.ft^2 to Btu/h.°F.ft^2,
A_W	-	heated floor area (ft^2),
B_W	-	solar heating reduction fraction (estimated at 0.25),
G	-	percent glass (12.5%, 37.5%, 62.5%, or 87.5%),
$\lambda \times S$	-	internal load heating reduction (Btu/hr.ft^2.°F),
μ	-	occupied fraction - total weekly hours operation + 168 hours per week,
P	-	number of employees x (Btu/h.person.°F),
HDD65	-	number of heating degree days with base 65°F, and
E_H	-	efficiency of heating equipment.

The Q_p component, the heat necessary to protect the building's contents when empty, is part of the space heating component as defined by Eq. 3.1 but is not included in Eq. 3.2 because of a lack of appropriate data. This quantity becomes part of the regression model intercept and error term. This model assumes that the annual heating energy use can be normalized for all buildings in a building category using HDD65 in a linear equation.

The cooling component equation is:

$$Q_{cool} = \frac{24\ h/d \times CDD65}{E_C} \left\{ \left[\underbrace{(U^* \times Ae) \times (1}_{Q_C} + \underbrace{B_S \times \frac{G}{100})}_{Q_E} \right] + \underbrace{(AV_S \times \rho \times Cp \times A_S)}_{Q_D} + \underbrace{(\lambda \times S \times A_S \times \mu) + (P \times \mu)}_{Q_I} \right\} , \qquad (3.3)$$

where

CDD65 = cooling degreee days with base 65°F,

E_C = cooling equipment efficiency,

B_S = solar fraction cooling increase (estimated at 1),

AV_S = volumetric air flow rate of outside summer air per square foot of cooled floor area ($ft^3/h.ft^2$), and

A_S = cooled floor area (ft^2).

The other variables are as defined for space heating in Eq. 3.2. However, since cooling energy use is not so well related to temperature as heating, the internal load quantities, Q_I, have the temperature dimension absent in the denominator. Notice that there are no reductions to the cooling load as there are for heating (i.e., from internal gains Q_I and external gains Q_E). Instead, Q_I and Q_E represent additions to the cooling load. A variation in Eq. 3.3, which was also considered, involves substituting equivalent full-load hours (EFLH) for CDD65, where EFLH = (0.5 CDD65 + 300), in accordance with the discussion in 1981 **ASHRAE Handbook, Fundamentals**. The total annual fuel consumption (Q_{TOTAL}) for electricity or natural gas can be represented by:

$$Q_{TOTAL} = Q_{heat} + Q_{cool} + Q_{INTERNAL} + Q_{EXTERNAL} , \qquad (3.4)$$

where

$Q_{INTERNAL}$ = Fuel needed for internal uses (such as cooling, manufacturing and lighting), that affect the internal heat component Q_I; and

$Q_{EXTERNAL}$ = Fuel required for external uses (such as water heating, electricity generation, outdoor lighting, and other uses) that do not affect the internal heat component Q_I.

It is not possible to calculate a regression model in the above form because the NBECS does not supply data for all the necessary variables. It is possible to estimate some quantities from data available in NBECS, and the remaining quantities will be accounted for via other regression model coefficients and the error term.

For the heating model, these variables do have NBECS counterparts: Ae, A_W, G, P, and HDD65. Equation 3.2 may then be rewritten as:

$$Q_{heat_i} = \text{ENDUSE 2} \times \text{HDD65} \times \left([U^* \times Ae \times (1 - B_W \times \frac{G}{100})] \right.$$

$$\left. + (\gamma_H \times A_W) - (C_1 \times \text{INTERNAL}) - (\Delta_{H_4} \times \text{WKWK}) \right) \quad , \qquad (5.1)$$

where

ENDUSE1 = 1 if fuel i is used for heating, else = 0,

INTERNAL = $(\Delta_{H4} \times \text{SFHTPWK}) + (\Delta_{H2} \times \text{COOKWK}) + (\Delta_{H3} \times \text{MANUWKSF})$, (5.2)

SFHTPWK = $A_W \times \mu$,

COOKWK = COOK x number of employees,

COOK = 1 if any fuel is used for cooking, else = 0,

MANUWKSF = MANUF x number of employees x A_W,

MANUF = 1 if any fuel is used for manufacturing, else = 0,

WKWK = $P \times \mu$,

and γ_H, C_1, Δ_{H1}, Δ_{H2}, Δ_{H3}, and Δ_{H4} are regression coefficients.

The building surface area (Ae) is, of course, unknown, but can be crudely approximated using square-footage and number-of-floor data. Then, the effect of U* x Ae can be measured by estimating the following regression coefficients:

$$\begin{aligned} U^* \times Ae = \; & (\alpha_{11} \times \text{BANDSA}_1) + (\alpha_{12} \times \text{ROOFSA}_1) + \\ & (\alpha_{21} \times \text{BANDSA}_2) + (\alpha_{22} \times \text{ROOFSA}_2) + \\ & (\alpha_{31} \times \text{BANDSA}_3) + (\alpha_{32} \times \text{ROOFSA}_3) + \\ & (\alpha_{42} \times \text{BANDSA}_4) + (\alpha_{41} \times \text{ROOFSA}_4), \end{aligned} \qquad (5.3)$$

where

BANDSA_1 = surface area of outside walls for free-standing buildings built before 1961 (otherwise = 0),

BANDSA_2 = surface area of outside walls for free-standing buildings built during or after 1961 (otherwise = 0),

$BANDSA_3$ = surface area of outside walls for attached buildings built before 1961 (otherwise = 0),

$BANDSA_4$ = surface area of outside walls for attached buildings built during or after 1961 (otherwise = 0),

$$BANDSA_i = (12\text{-ft-high wall} \times 4 \text{ walls})^{*} \times \left[\frac{\text{total sq feet}}{\text{number of floors}}\right]^{1/2} \times \text{number of floors},$$

$ROOFSA_i$ = estimated roof area for case i, as with $BANDSA_i$, and is calculated as:

total square feet/number of floors, and

α_{ij} = regression coefficients.

In this manner, several coefficients will represent the regression coefficient that would have been estimated had all variables been available and are multiplied or divided by the missing variables.

For the cooling model, these variables do have NBECS counterparts: Ae, A_W, G, P, and CDD65. Equation 3.3 may, then, be rewritten as

$$Q_{cooli} = ENDUSE2 \times CDD65 \times [U^{\star} \times Ae \times (1 + B_S \times \frac{G}{100})]$$
$$+ (\gamma_c \times A_S) + (C_2 \times INTERNAL) + (\Delta_c \times WKWK) \quad ,$$

where,

ENDUSE2 = 1 if fuel i is used for space cooling, else = 0,

$U^{\star} \times Ae$ = as for heating, and

γ_C, C_2, and Δ_C are regression coefficients.

As for the space-heating component, the coefficients of the cooling component will represent the regression coefficient that would have been estimated had all variables been available and are multiplied or divided by the missing variables.

The $Q_{INTERNAL}$ is the fuel needed for internal uses such as cooking, manufacturing, and lighting (electricity consumption model only). In the engineering regression model approach for natural gas, this is represented as:

$$Q_{INTERNAL_{NG}} = (\Delta_{H2} \times COOKWK \times ENDUSE6) + (\Delta_{H3} \times MANUWKSF \times ENDUSE5), \quad (5.5)$$

*This quantity is included in the $\sim_{ijh}$ regression coefficients and is not included in the variable $BANDSA_i$ calculation.

where

ENDUSE6 = 1 if fuel i is used for cooking, else = 0,

ENDUSE5 = 1 if fuel i is used for manufacturing, else = 0; and

Δ_{H2} and Δ_{H3} are regression coefficients.

Similarly, for electricity, the $Q_{INTERNAL}$ component is represented as:

$$Q_{INTERNAL_{NE}} = (\Delta_{H2} \times COOKWK \times ENDUSE6) + (\Delta_{H3} \times MANUWKSF \times ENDUSE5), \quad (5.5)$$

where

ENDUSE6 = 1 if fuel i is used for cooking, else = 0,

ENDUSE5 = 1 if fuel i is used for manufacturing, else = 0, and

Δ_{H2} and Δ_{H3} are regression coefficients.

Similarly, for electricity, the $Q_{INTERNAL}$ component is represented as:

$$Q_{INTERNAL_{E}} = (\Delta_{H1} \times SFHTPWK) + (\Delta_{H2} \times COOKWK \times ENDUSE6) + (\Delta_{H3} \times MANUWKSF \times ENDUSE5), \quad (5.6)$$

where

Δ_{H1} is a regression coefficient.

While electricity for lighting is an internal use, appropriate data are not available for modeling. Unfortunately, this quantity will be part of some or all of the model coefficients and the model error term for the electricity model only. No energy studies exist which show approximate electricity use for lighting per square foot for the building categories defined by the EIA.

$Q_{EXTERNAL}$ is the fuel required for external uses that do not contribute to the internal heating/cooling component, such as hot water heaters, electricity generation, and external lighting. No data are collected in the NBECS for external lighting, but this effect will be part of some or all of the model coefficients and the model error term for the electricity model only. The external component for natural gas is, therefore, represented as:

$$Q_{EXTERNAL_{NG}} = (\Delta_{E3} \times ENDUSE3) + (\Delta_{E4} \times ENDUSE4), \quad (5.7)$$

where

ENDUSE3 = 1 if fuel i is used to heat water, else = 0,
ENDUSE4 = 1 if fuel i is used to generate electricity, else = 0, and
Δ_{E3} and Δ_{E4} are regression coefficients.

The external component for electricity is represented as:

$$^{Q}EXTERNAL_{E} = (\Delta_{E1} \times PRMTR) + (\Delta_{E3} \times ENDUSE3) \quad , \tag{5.8}$$

where

PRMTR (perimeter) = $\sqrt{\text{total square feet} + \text{number of floors}}$, which is taken to be roughly proportional to the external use of electricity for lighting, and Δ_{E1} is a regression coefficient.

Finally, the total fuel consumption for fuel i, then, is represented by:

$$^{Q}TOTALij = \mu + {}^{Q}heatij + {}^{Q}coolij + {}^{Q}EXTERNAL_{i} + {}^{Q}INTERNAL_{i}, \tag{5.9}$$

where

i = ith fuel,
j = jth building using fuel i, and
μ = a constant included to improve the overall fit of the model.

Additional limitations to the models are expected because of limitations in the data. With respect to the industrial sector, 70% to 80% of the total energy use is for manufacturing. The actual building energy use is typically a small portion of the total use for an individual building. No variables that correlate well with the magnitude of the manufacturing energy use were collected by NBECS, so it is expected that accurate estimation of the fuel consumption for this building category may not be possible. Similarly for establishments in the food sales sector, it will probably not be possible to estimate the fuel consumption accurately for those buildings with a sizeable cooking effort, because no variables that correlate well with cooking were collected by NBECS for such buildings.

5.4 REGRESSION MODELS AND ALGORITHM

It would seem natural that the variability, σ^2, of building energy consumption should increase with the mean, Θ. The two most familiar candidates for modeling the relationship between σ^2 and Θ are (1) $\sigma^2 \alpha \Theta$ and (2) $\Theta^2 \alpha \Theta^2$. These relationships suggest weighted analyses. Alternatively, transformations can also be used to stabilize consumption variance. The transformations square root and natural log correspond to (1) and (2) respectively (via a first-order Taylor expansion about Θ).

To fit the model (Eq. 5.9) to the energy survey data, a functional form for the error structure must be specified. Thus, we might consider the models

$$Y = \Theta + \epsilon \quad , \tag{5.10}$$

where Y denotes energy consumption, $\Theta = \mu_{\Theta} + {}^{Q}heat + {}^{Q}cool + {}^{Q}INTERNAL + {}^{Q}EXTERNAL$ as in Eq. 5.9, and j is random with mean zero and variance proportional to Θ or Θ^2. Alternatively, we could write

$$f(Y) = f(\theta) + \epsilon \quad , \tag{5.11}$$

where f denotes the square-root or log function and ϵ has mean zero and (constant) variance, σ^2. An approximation to Eq. 5.11, which will also be useful, is obtained by first-order Taylor expanding f about μ_θ on the right side of Eq. 5.11 to yield, in the case of the log transform,

$$\log(y) = \log(\mu_\theta) + \frac{1}{\mu_\theta}(y - \mu_\theta) + \epsilon \tag{5.12}$$

and similarly for the square-root transform.

Were it not for the coefficients c_1 and c_2 in the internal load adjustments in expressions 5.1 for Q_{heat} and 5.4 for Q_{cool}, the model 5.9 and thus models 5.10 and 5.12 would be linear in the parameters. Linearity represents a distinct advantage in terms of ease of computing and flexibility of the available software, SAS. Therefore, at least as a starting point, we also consider the models 5.10-5.12 with the terms involving c_1 and c_2 modified to make θ linear. For example, the term c_1 * Δ_H1 x SFHTPWK become Δ'_H1 x SFHTPWK.

The choice of one among several candidate models must be based on validity, which will be assessed here primarily through plots of residual vs predicted values and the independent variables. Only when two or more models both appear valid should the choice be based on statistics such as R_2 or f, or considerations such as ease of computing. Obviously, an f-statistic is meaningless if the error term is non-normal. It is easy to construct examples of regression models where Y is say log-normal, but application of the log transformation actually reduces the value of R^2. One explanation for this phenomenon is the following. Transformations such as log (or square-root) often tend to increase the importance of the intercept in the model, a feature obscured by R^2 because it is corrected for the mean. This can be seen heuristically for the log transformation in the present example as follows. Without reference to any independent variables, suppose that $\log Y \sim N(v, \tau^2)$. (This would be the case, for example, if the independent variables were themselves normally distributed.) Then

$$E(Y) = e^{v+\tau^2/2} \quad ,$$

and

$$Var(Y) = (e^{\tau^2} - 1)(e^{2v+\tau^2}) \, ,$$

as is well known. It follows that

$$E(Y)^2/Var(Y) = 1/\ (e^{\tau^2}-1) \ll v^2/\tau^2 \text{ whenever } v \gg 1 \ .$$

Letting $\bar{y}$ play the role of v, and

$$\sum_{i=1}^{n}(y_i-\bar{y})^2$$

play the role of r^2, since $\bar{y} >> 1$ in all of the regression categories, we see that

$$C = n\bar{y}^2 \bigg/ \sum_{i=1}^{n}(y_i-\bar{y})^2$$

increases considerably with the log transformation. Now, another reasonable measure of the goodness of a model is the *uncorrected R-squared*,

$$R_U^2 = \sum_{i=1}^{n}\hat{y}_i^2 \bigg/ \sum_{i=1}^{n}y_i^2 \quad .$$

Since for purposes of prediction the intercept is as important as any other parameter, one could argue that R_U^2 is a more appropriate measure of goodness than R^2 is. It is easy to show that $R_U^2 = (R^2+C)/(1+C)$. Thus, R_U^2 tends to increase with the log transformation, but unless it increases considerably, R^2 will naturally decrease. This was in fact the case in most of the categories.

On the other hand, any meaningful regression model should at least display overall (corrected) significance. Overall significance is important here--more so than the significance of individual terms--because all terms in the model (other than the intercept) are considered a priori to be important. In contrast with stepwise regression, parameters cannot be validly discarded from our model simply because they are not significant in a particular analysis. Another product of valid inference, whose importance is discussed in Sect. 6, is the standard errors of predicted values.

First, consider model 5.10. This model suggests a multistage weighted analysis with weights proportional to $\hat{y}^{-1}$ or $\hat{y}^{-2}$ ($\hat{y}$ = predicted value). At stage 1, let all weights equal 1, or else for some small $\delta > 0$, let the weights be either $\delta + y^{-1}$ or $\delta + y^{-2}$. This procedure could be carried out in two stages (once to start, then once more) or several iterations up to convergence. For large sample sizes, any of these procedures would be best linear unbiased since the stage 1 estimator is consistent. The fully iterated estimator has some intuitive appeal. However, unlike many iteratively reweighted least-square estimators, the fully iterated estimator does not maximize the likelihood, as can be seen from the following simple example:

Suppose $z \sim N(\theta,\theta^2)$, $\theta \geq 0$, and θ is to be estimated upon observing z. The weighted least-squares estimator is then max(z,0), but the maximum likelihood estimator, $\hat{\theta}_M$, is

$$\hat{\theta}_M = \begin{cases} \frac{1}{2} \times (\sqrt{5} - 1)z, & \text{if } z \geq 0, \\ \frac{1}{2} \times (-\sqrt{5} + 1)\, z, & \text{if } z < 0. \end{cases}$$

Seeing no particular reason to compute the fully iterated estimator, and since requiring convergence complicates matters, we arbitrarily elected to base the weighted estimates on five iterations.

Actually effecting the iterative scheme requires one further consideration. It is implicit in model 5.10 that $\theta > 0$ for the log transform and $\theta \geq 0$ for the square-root transform. Nevertheless, there is nothing to preclude $\hat{y} < 0$. Thus, we are forced to redefine the weights as $\max(\hat{y}^{-1},b^{-1})$ or $\max(\hat{y}^{-2},b^{-2})$ where b is the lower bound (different for gas and electricity models) discussed in Appendix F. Even with this lower bound, if certain predicted values are low because of noise in the data, it seems possible that they could receive discordantly high weights and that this problem could be compounded with each iteration.

The weighted residual plots resulting from these analyses did not look as good as we had hoped, nor as good as the plots in subsequent analyses. Usually, there were one or two extreme outliers, often associated with negative predicted values. These extremes, coupled with the following fact, leads us to recommend against the weighted approach: the validity of the inference associated with the weighted analyses, which relies completely on asymptotic arguments, has not to our knowledge been adequately assessed. It seems intuitive that inferences might be sensitive to vagaries associated with the weighting scheme.

Next, consider the model 5.11. It has a potential problem in that the argument of f on the right-hand side should be constrained to be positive, since we are considering the transformations $f(x) = \log x$ and $f(x) = \sqrt{x}$. Probably, the ideal way to rectify this problem would be to fit the model

$$f(Y) = f(\theta) + \epsilon$$

$$\text{subject to } \theta \geq b \quad ,$$

where θ is as defined after Eq. 5.10 and b is the lower bound also discussed previously. Unfortunately, the software in SAS (PROC NLIN) will not permit this fit. Instead, we used the model

$$f(Y) = f[\psi(\theta)] + \epsilon \quad , \tag{5.13}$$

where ψ forces a lower bound on θ while preserving the differentiability of the right side of Eq. 5.13 with respect to the model parameters (necessary for the computational algorithms used in PROC NLIN). Such a function, ψ, is the cubic spline

$$\psi(x) = \begin{cases} b, & x \le 0, \\ b + \dfrac{4x^3}{27b^2} & 0 \le x \le \dfrac{3b}{2}, \\ x, & x \ge \dfrac{3b}{2}. \end{cases}$$

where, again, b is the lower bound.

Fitting a model like 5.13 involves minimizing the expression

$$\sum_{i=1}^{n} \{f(y_i) - f[\psi(\theta_i)]\}^2 , \qquad (5.14)$$

where y_i is the consumption value and $\theta_i = \mu_\theta + {}^{\alpha}heat_i + {}^{\alpha}cool_i + {}^{\alpha}EXTERNAL_i + {}^{\alpha}INTERNAL_i$ for the ith set of predictor variables. If a zero of the gradient of Eq. 5.14 (with respect to the parameters) is found and if Eq. 5.14 is convex--or more weakly, pseudoconvex--then a minimum of Eq. 5.14 has also been found, and one can expect computational algorithms to behave reasonably well. A function h is pseudoconvex if $\nabla h(\theta)'(\theta^* - \theta) \ge 0$ implies $h(\theta^*) \ge h(\theta)$. (If h is increasing at θ in the direction of θ^*, then h continues to increase in that direction.) Unfortunately, Eq. 5.14 need not even be pseudoconvex, as the following example shows: Consider the function

$$h(\theta) = \sum_{i=1}^{2} [\log (y_i - \log \theta_i]^2 = \sum_{i=1}^{2} [\log(y_i/\theta_i)]^2 \text{ at } \theta_2 = 1/2,\ y_1 = y_2 =$$

1/16, and $\theta_1^* = \theta_2^* = 1$. Then

$$\nabla h(\theta) = -2 \times \begin{bmatrix} 1/\theta \log(y_1/\theta_1) \\ 1/\theta_2 \log(y_2/\theta_2) \end{bmatrix} , \text{ and}$$

$$\nabla h(\theta)'(\theta^* - \theta) = 2 \sum_{i=1}^{2} \log(y_i/\theta_i) \left(1 - \frac{\theta_i^*}{\theta_i}\right)$$

$$= 0.69 > 0 .$$

But $h(\theta) = 16.34 > h(\theta^*) = 15.37$.

The fact that Eq. 5.14 is not pseudoconvex does not preclude a meaningful nonlinear analysis. If good starting values can be found, a nonlinear algorithm will converge to a global minimum, despite irregular features of the objective function. On the other hand, the fact that Eq. 5.14 is not pseudoconvex must be construed as a severe warning that convergence to the solution could be extremely difficult. With this in mind, we spent much time in fitting model 5.13, using PROC NLIN. Starting values had to be coded manually for each nonlinear run. For starting values, we tried fitted values from the weighted analyses already discussed as well as ordinary least-squares estimates and values obtained after fitting model 5.12 to be discussed. The Gauss-Newton and Marquardt methods were employed, and the DUD method was also used several times as a check on the accuracy of the derivative calculations (see SAS manual).

Overall, results of the nonlinear analyses do not look promising. For certain regression categories, the procedure converged to apparently good solutions. For many others, NLIN's convergence criteria either were met or were not met, but the fixed point was clearly unsatisfactory (sometimes yielding $R^2 < 0$). Because we are seeking a mechanical and reasonably easy approach to imputation and because the nonlinear approach, which requires a lot of time coding starting values, is not particularly easy even without the difficulties associated with a nonpseudoconvex objective, we strongly recommend that this approach not be considered further.

Finally, consider Eq. 5.12, the linear approximation to Eq. 5.11. As a linear model Eq. 5.12 is easy to fit to the data (using PROC REG or GLM), and quantities such as predicted values and their standard errors can be readily output for use with other procedures. Consider, for $f(x) = \log X$, the linear approximation $ax + b$ to $f(x)$, over the range $[x_0, x_2]$. Since $\log X$ is strictly concave, the maximum error (E) occurs at a maximum of three points and is minimized when occurring at exactly three. In that case, two of the points are x_0 and x_2. Denoting the other point by x_1, we have

$$E = -[\log x_0 - (ax_0 + b)] = \log x_1 - (ax_1 + b) = -[\log x_2 - (ax_2 + b)].$$

Solving for E, we have

$$E = 1/2 \left[\log\left(\frac{x_2 - x_0}{\log x_2 - \log x_0} \right) - 1 + \frac{x_0 \log x_2 - x_2 \log x_0}{x_2 - x_0} \right].$$

For example, for natural gas regression category 10 (assembly buildings), the 0.01-percentile is 10268 ft^3 and the 99-percentile is 18,181,424 ft^3. The corresponding natural log values are 9.21 and 16.72, and E = 2.24. The root mean squared error for category 10 turns out to be 1.27. Other categories are similar and, in general, we believe that linearizing Eq. 5.11 provides an adequate approximation.

Plots of residuals vs predicted values and the independent variables look fair for the square root transform and good for the log transform. The log plots look better since a few of the square-root plots of residuals vs predicted values fan out, suggesting that the square-root transform is not quite strong enough to bring the consumption error structure into line. Generally, significance levels for uncorrected models (i.e., no intercept) also favor the log transform. The overall corrected f-tests are significant for the log model. The log model has one other advantage over the square-root model: Both analyses give negative predicted values in terms of the transformed variables. However, in backtransforming, the antilog transformation handles a negative argument with no problem, but the square transform cannot be applied without special consideration for negative predicted values. In general, this consideration would involve simply enforcing the lower bound on predicted values. However, a disadvantage to this enforcement is that it is not clear how to compute the standard error of bounded predicted values.

In summary, we prefer and recommend the log transform analysis of model 5.12 because it provides a highly significant fit to the data in every category and has the best residual plots. The linearization approximation is adequate, as discussed previously, and the log transform linearized analysis is extremely easy to perform. The weighted analyses do not provide good residual plots, and the inference associated with them is questionable. The nonlinear analyses, which seem to have a lot of potential, are unsatisfactory in practice--most likely because of the nonpseudoconvexity of the objective Eq. 5.14 and because the data is sufficiently noisy that adequate starting values cannot be found (even the linear model estimates). Even without these problems, the nonlinear analyses are much more cumbersome to perform than their linearized analogues. The SAS programs for performing the linearized log-transform analyses for the various regression categories are given in Appendices J (natural gas) and K (electricity).

6. EMPIRICAL STUDY OF THE REGRESSION MODEL

The theoretical assessment of an imputation procedure is quite difficult. The complexity of the NBECS data made it even more difficult to assess the imputation effect of the ORNL engineering model approach. The scope of most theoretical work on the evaluation of imputation procedures is limited to fairly simple data and procedures. However, it is often important to quantify the imputation error under normal production conditions. In general, imputation procedures must be evaluated by empirical studies where a clean data set is created to act as a population, and a percentage of this population is reserved as the subset of nonrespondents with missing data.

The object of this exercise is to simulate imputation to check model bias. A stratified random sample is drawn from the Group 3 population to reserve as a test imputation group. Then, the regression model is recalculated with the remaining Group 3 observations. The recalculated model is used to impute consumption values for the test imputation group, and the actual consumption values are compared with the imputed consumption values.

To assess the ORNL imputation procedure, a limited effort was made in conducting an empirical study on the 1979 NBECS data set. First, education buildings using natural gas were arbitrarily selected to create a test population. The test population is the Group 3 records that passed all the data edits described in Sect. 4. This clean data set, with 201 education buildings, is complete, with respect to natural gas-consumption values as well as all the independent variables for the engineering model. Thus, this data set is also the input data set that was used to develop the regression model for the education buildings.

Next, Group 1 records were cleaned by deleting records with previously imputed independent variables. The clean Group 1 records contain 21 educational buildings, each with a complete set of independent variable values that can be used in computing the predicted consumption values, but none of these buildings had a utility bill coverage of over 30 days.

In this empirical study, the clean Group 1 records cannot be used, because these records do not have the actual (reported) natural gas consumption values for comparison. Therefore, the missing data set must be selected from the test population of 201 education building records.

To create a set of records of missing consumption data, observe that (1) the actual percentage of consumption nonresponses among the clean education building records is

$$\frac{\text{number of clean Group 1 records}}{\text{total number of clean Group 1 and Group 3 records}}$$
$$= 21/(21 + 201) = 9.46\%, \text{ and}$$

(2) the distribution of the Group 1 records over several independent variable categories given in Table 6.1, is quite different from the corresponding distribution of the Group 3 data set as shown in the same table.

Table 6.1. Distributions of Group 1 and Group 3 education buildings with variable categories: natural gas-use buildings with all edits

SQFT1	NWKER1	HDD651*ENDUSE1	Group 1 records after all edits	Group 3 records after all edits	Total
10,000 < SQFT1 ≤ 82,857	0 ≤ NWKER1 ≤ 73	= 0	5	9	14
		> 0	1	54	55
	73 < NWKER1 ≤ 140	= 0	1	2	3
82,857 < SQFT1 ≤ 229,857	0 ≤ NWKER1 ≤ 73	= 0	1	3	4
	73 < NWKER1 ≤ 140	= 0	2	14	16
		> 0	3	17	20
	140 < NWKER1 ≤ 999	= 0	1	13	14
229,857 < SQFT1 < 500,000	73 < NWKER1 ≤ 140	= 0	1	2	3
	140 < NWKER1 ≤ 999	= 0	5	11	16
		> 0	1	25	26
		Other	0	51	51
		Total	21	201	222

Based on the above observations, an effort was made to create the consumption nonresponses from the test population of 201 records. To follow the patterns of the actual consumption nonresponses (i.e., the Group 1 records), 9.46% of the test population should be selected, and the distribution of the created missing data set should be close to the Group 1 distribution of Table 6.1. As a result, the missing data set was created by taking a stratified random sample from the 201 records with a sample distribution close to the actual group of consumption nonresponses. The 19 (9.46%) selected building records were then treated as the consumption responses for the empirical study.

The empirical imputation error has two components: (1) the bias of the imputation estimator from the actual total consumption, and (2) the variance of the imputation estimator. The first component can be measured in one trial; however, to estimate the expected value of the imputation estimator or to estimate the variance of the imputation estimator requires many replications of the experiment. A complete assessment requires a Monte Carlo simulation study where each replication of the experiment can be generated by computer algorithms. The limited time and resources available for this project did not allow conducting the Monte Carlo simulation study.

Estimates of the total natural gas consumption for the single test population were computed using the two imputation approaches:

1. Impute the 19 missing consumption values using the input data set of 182 buildings (Table 6.2) by following the engineering regression model approach of Sect. 5.11.

Table 6.2. Distribution of the test population of education buildings with selected variable categories: natural gas use buildings

			Test Population		
SWFT1	NWKER1	HDD651*ENDUSE1	Consumption nonresponses[a] from the test population (missing data set)	Input data set of the test population	Total
10,000 < SQFT1 ≤ 82,857	0 ≤ NWKER1 ≤ 73	= 0	4	5	9
		> 0	1	53	54
	73 < NWKER1 ≤ 140	= 0	1	1	2
82,857 < SQFT1 ≤ 229,857	0 ≤ NWKER1 ≤ 73	= 0	1	2	3
	73 < NWKER1 ≤ 140	= 0	2	12	14
		> 0	3	14	17
	140 < NWKER1 ≤ 999	= 0	1	12	13
229,857 < SQFT1 < 500,000	73 < NWKER1 ≤ 140	= 0	1	1	2
	140 < NWKER1 ≤ 999	= 0	4	7	11
		> 0	1	24	25
		OTHER	0	51	51
		TOTAL	19	182	201

[a]Missing data set is a stratified random sample from the test population of 201 buildings.

2. Impute the 19 missing consumption values by adjusting the sampling weights of those reported. This adjustment is another simple method of treating unit nonresponses in unequal probability sampling. In this case, the 182 buildings were adjusted for their sampling weights, on the assumption that the 19 missing units occurred at random. The total consumption of the test population is estimated by weighting consumption values with the adjusted weights.

Table 6.3 gives imputation estimates of the total natural gas-consumption value for the test population. Imputation estimates using the ORNL modeling approach and the reweighting adjustment approach were calculated. The ORNL estimate is 1.2% away from the actual consumption value, and the weighting adjustment method is 7.6% away from the target value. Thus, in this experiment, the weighting method yields a target value biased 6.4% more than the modeling approach. The poorer performance of the weighting approach is probably caused by the nonrandom pattern of the consumption nonresponses, as can be seen in Table 6.1. The modeling approach adjusts the self-selection bias using an assumed population model of energy-consumption relationships. The modeling approach will perform well if the model is sensitive to the independent variables and if the group of nonrespondents as well as the group of respondents follow approximately the same response surface.

Table 6.3. Imputation estimates of the total natural gas consumption value for the test population

	Test population (201 buildings)		
	Consumption nonresponses (19 buildings)	Input data set (182 buildings)	Total consumption value of the test population
Actual total weighted consumption value (target value)	1,888,162,918.27	119,184,210,892	121,072,373,810 (100%)
Consumption estimate using the ORNL model approach	3,328,053,280.89	119,184,210,892	122,512,264,173 (101.2%)
Consumption estimate using the weighting adjustment approach	11,100,340,666	119,184,210,892	130,284,551,558 (107.6%)

7. IMPUTATION PROCEDURES

7.1 AN IMPUTATION MODEL

The following imputation model is concerned with statistical methods of dealing with missing consumption values after data collection and regression modeling are completed. At the data-collection stage, survey data should be collected as completely and as accurately as possible, using callbacks and follow-ups as needed. The imputation model described below is only an alternative approach when information cannot be obtained at the data-collection stage.

To produce total energy-consumption estimates based on the building records in the NBECS sample, it is necessary to assign–one way or another–a consumption value to each of the units in the sample. The recommended approach (Sect. 5) replaces missing consumption data with data that can be said to have response errors determined by the imputation. However, as with most imputation procedures in complex surveys, the model will apply to only a major portion of the missing data set. The difference in the distributions of independent variables of the respondent Group 3 buildings vs the respondent Group 1 buildings, caused by incompleteness, often leads to cases where the data from the respondents cannot reasonably be used to impute some of the missing consumption data. In cases where the regression models do not apply, some ad hoc imputations are needed to clean up the data so that the job is done expediently.

The overall imputation model divides the missing records into segments so that appropriate imputations can be made according to the condition of the records in each segment. This segmentation makes the imputation task less formidable and, in those cases where the regression models do not directly apply, allows for remedial measures to be implemented before the final assignments of imputed consumption values. For example, additional follow-up calls might be made to obtain information on missing key independent variables so that a predicted consumption value can be calculated using the regression model. Actual square footage or actual number of floors, instead of truncated values for large buildings, can be examined by subcontractors for analytical purposes or for estimating some lower bound of the building energy consumption.

The flow diagram in Fig. 7.1 [at the end of this subsection (7.1)] displays a breakdown of Group 1 records that are considered to have missing consumption values. The Group 1 records were first divided into segments according to the conditions of the key independent variables. Each record will terminate with one of the ten imputation conditions where some action is recommended for each, according to the amount of information available to the analyst. The conditions (denoted Conditions I,...,X) are described as:

A. Conditions where the ORNL regression model can directly apply:

Condition VII: The case with these features: (A.1) records contain no known invalid auxiliary information, (A.2) locations of the independent variables are near the donor data used to fit the model, and (A.3) the variance estimate of the imputed value is relatively small. The majority of missing records will fall into this group.

Conditions V and VI: The cases where the model will produce predicted consumption values with unknown reliability. Currently, EIA does not have the statistical software to check whether a record will violate prerequisite A.2. Also, the establishment of an exact criterion to determine at what point a large regression variation indicates an unreliable prediction requires further research efforts.
If records can be classified as Condition V or VI, then the regression models may not apply. Other than the weighting adjustments, some hot-decking methods may be tried for these records. However, records that fail to satisfy conditions A.2, and A.3 are likely to be records that input a combination of independent variables dissimilar to the input records. Therefore, the hot-decking method may not be better than the weighting adjustment method.

B. Conditions where the ORNL regression model can partially apply:

Conditions II and IV: The cases where nothing else can be done with the truncated square footage or the truncated number of floor values. The most convenient way to handle these missing data is by the weighting adjustment method. In the same situation, it is also possible to estimate a lower bound of the fuel consumption. An example has these features: (B.1) the building is an education building with over 1 million square feet and over 50 floors and (B.2) the donor records of education buildings has an upper limit of 600,000 square feet and 30 floors.

One can replace the building square footage with 600,000 and replace the number of floors with 30 and obtain a predicted value from the regression model. This procedure tends to underestimate the actual consumption value but it protects against misuse of the model. If the actual square footage and the actual number of floors can be obtained for analysis purposes, then one might be able to develop a revised model that includes all the large buildings. This revised model might be directly applicable to the large buildings with missing consumption values. Unfortunately, even with full knowledge about the two variables, the modeling approach may not perform well for large buildings because of the limited number of donor records, as can be seen in Tables 4.16 and 4.17. Neither can any kinds of hot-decking or matching techniques help much in such situations, because they require a large reservoir of donor records.

It is important to note that large buildings may have large contributions to the overall consumption totals because of their size, especially when they sampling weights assigned to these buildings are large. Additional follow-up efforts will be worthwhile and desirable if it is possible to collect valid consumption values for these buildings.

Conditions I, III, VIII, and IX: The cases where imputed key independent variable(s), if important, can seriously bias the results of consumption estimates using the regression models. Weighting adjustment is a simple way to treat these records with unusable auxiliary information.

Condition X: The unique case where the imputed key independent variable is not important in the regression model. One may assign some average value to that variable and continue to check through decision box 6 in Fig. 7.1.

The 1979 data can be used for outlier checks, as described in Sect. 4.9, and they also provide a good source of prior information for imputing missing consumption values. For example, if a building had a known 1979 consumption value and its 1983 consumption value could not be collected, then it is possible to impute the 1983 value on the basis of the 1979 value. However, one must check to see if the building changed its consumption-related characteristics before or during 1983, because these changes might have caused a large difference between the 1979 and 1983 consumption values. In particular, variables such as square footage, number of floors, number of employees, and fuel end use should be checked.

Explanatory Notes for Fig. 7.1:

1. BUILDING RECORDS WITH COMPLETELY MISSING CONSUMPTION VALUES:

   Natural gas-or electricity- use building records in the WORKING.NATGAS or WORKING.ELECT data set with completely missing consumption values. The records are Group 1 records in the data set with A = 1 (either natural gas or electricity is one of the first three primary fuels used in the building). In the 1979 NBECS file, 576 natural gas-use building records and 779 electricity-use building records are in this category (see Tables 4.2 and 4.3).

2. LARGE BUILDINGS WITH SQFT1 >=1,000,000: Large buildings with more than or equal to 1 million square feet.

3. HIGH-RISE BUILDINGS WITH NFLOOR1 >= 50: A decision box that determines whether e number of floors are truncated to 50 floors.

4. KEY INDEPENDENT VARIABLES IMPUTED: Includes those independent variables that appear in the regression equation. This decision box determines whether any of the key independent variables are values that are derived from imputed values.

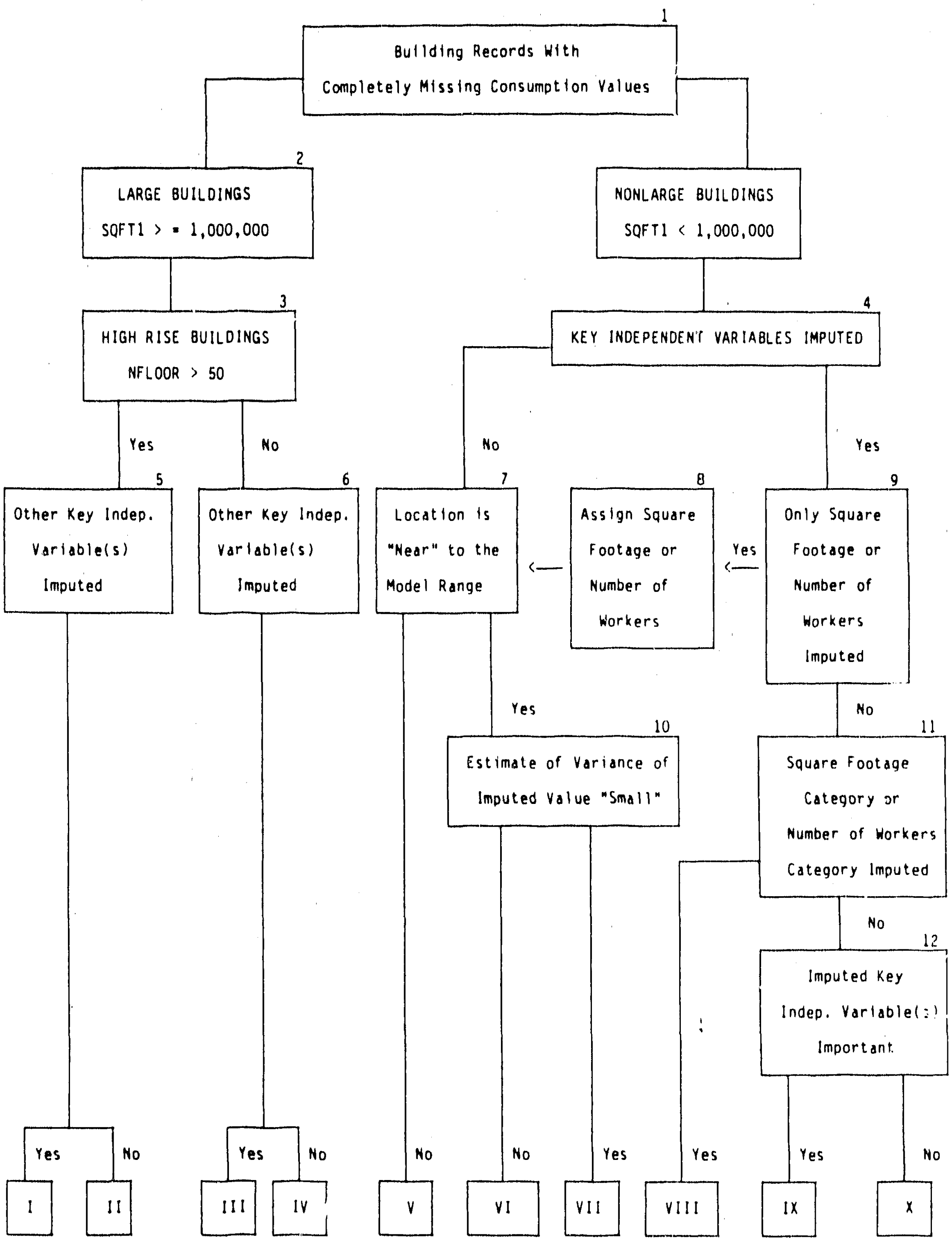

Figure 7-1. FLOW DIAGRAM OF THE IMPUTATION MODEL

5. OTHER KEY INDEPENDENT VARIABLE(S) IMPUTED: Decision box checks for imputed key independent variables other than square footage or number of floors.

6. OTHER KEY INDEPENDENT VARIABLES IMPUTED: Same as Note 5 above.

7. LOCATION IS NEAR THE MODEL RANGE: A determination is to be made of the location of the corresponding independent variables relative to those used to determine the regression model. A criterion for measuring the nearness is presented in Sect. 7.2.

8. ASSIGN SQUARE FOOTAGE OR NUMBER OF WORKERS: Because only the square footage or number of workers variable is an imputed value, the category variable SQFTC1 or NWKERC1 may still be valid. One can either use the given imputed value or replace it with the mean value of the data from input records. The bias will somehow be bounded because the imputed or substituted value will fall within range limits of the corresponding category variable. It is best, of course, to contact the respondent and obtain the information.

9. ONLY SQUARE FOOTAGE OR NUMBER OF WORKERS IMPUTED: Only the square footage variable, SQFT1(SQFTX), or the number of workers variable, NWKER1 (NWKERX), are imputed values.

10. ESTIMATE OF VARIANCE OF IMPUTED VALUE SMALL: Because independent variables for values to be imputed are available, variance estimates for predicted values are straightforward to compute (Draper and Smith 1981, p. 210). The criterion to determine the degree of variation is yet to be determined.

11. SQUARE FOOTAGE CATEGORY OR NUMBER OF WORKERS CATEGORY IMPUTED: This box determines whether a category variable has an imputed value.

12. IMPUTED KEY INDEPENDENT VARIABLE(S) IMPORTANT: Examine the regression model to see if the variable with imputed values makes an important contribution to the model; that is, small changes in the independent variable may cause large changes in the dependent variable, which is not the same as the concepts of statistical significance. A variable may be significant, but both the true coefficient and the magnitude of that variable may be so small that the product makes a relatively small contribution to the estimated magnitude of the dependent variable.

7.2 EXTRAPOLATION VS INTERPOLATION

The quality of consumption imputation estimates must be measured by their nearness, in some sense, to the particular missing values that they are to represent. Of course, that quality is difficult to assess for the very reason that those values are missing. Nevertheless, meaningful information can be obtained from the data because the independent variables for consumption values to be imputed are available. Variance estimates for predicted consumption values, which are straightforward to compute, then measure the adequacy of the predictions, assuming that the regression model holds.

More realistic than the assumption that the model holds is the assumption that it provides an adequate approximation over or, at least, near the range of data used to fit it. Then, before considering an imputed value or its variance estimate, a determination must be made of the location of the corresponding independent variables relative to those used to determine the model. Even defining this range is somewhat difficult.

As noted under Methodology (Sect. 3), it was planned to use discriminant analysis to verify or negate that the population requiring imputation could be represented by the modeled population. The discriminant analysis could not be used unless the X-variables formed a multivariate normal distribution. Because the individual X-variables were not distributed normally, this condition was not met, and discriminant analysis was not utilized. However, another method was developed.

Let X denote the usual n x p matrix of regression independent variables (but with no column of 1's) or, perhaps, the matrix used to generate the usual matrix. Also, let x* denote the corresponding vector of independent variables for a value to be imputed. In one-dimension, the range is clear; it is $[x_{min}, x_{max}]$. We might trust the regression model if $\max[0, x_{min} - x^*, x^* - x_{max}]$ is not too large relative to, say,

$$\left[\frac{1}{n-1} \Sigma(X_i - X)^2 \right]^{1/2} ,$$

perhaps less than 10%.

In analogy with the one-dimensional case, we define the range of the independent variables used to determine the model to be their <u>convex hull</u>, $H = (x = X'\lambda \mid \Sigma\lambda = 1, \lambda >= 0)$. We will be interested in determining $d(x^*,H) + \min_{x \in H} d(x,x^*)$ for some appropriate distance function (d). There are many reasonable choices for d. We consider just one here, though we feel that it is more natural than many others. This is

$$d(x,x^*) = [(x - x^*)' Q^{-1} (x - x^*)]^{1/2} ,$$

where $Q = \frac{1}{n-1} X'[I - \frac{1}{n} \underset{\sim}{1}\underset{\sim}{1}']X$ (I is the n x n identity matrix, and $\underset{\sim}{1}$ is an n x 1 vector of 1's). This distance is standardized and thus unit-free. In one-dimension, it is just

$$|x - x^*| + \left[\frac{1}{n-1} \Sigma(x_i - \bar{x})^2\right]^{1/2}.$$

The distance $d(x^*,H)$ can then be determined by solving the quadratic programming (QP) problem:

$$\text{minimize } (X'\lambda - x^*)'Q^{-1}(X'\lambda - x^*)$$

$$\text{subject to } \Sigma\lambda = 1,\ \lambda \geq 0. \tag{7.1}$$

In analogy with the one-dimensional case, we could choose to trust the regression model only if $d(x^*,H) < 0.1$.

The solution to Eq. 7.1 can be determined quickly and easily with QP software. Unfortunately, QP routines are not available in SAS nor, as far as we know, on the EIA's computer. To experiment with this approach, we wrote a short program (Appendix L) in PROC MATRIX, using Newton's method. The program solves the problem (Eq. 7.1), though with no guarantee of convergence. The program did compute the solution in each of the few cases that we tried, but slowly. Thus, the approach seems possible, but we recommend the use of usual QP software rather than Newton's method, perhaps linking the former with SAS.

8. SUMMARY AND CONCLUSIONS

The ORNL study team has developed regression models that can successfully impute missing electricity- and natural gas- consumption values with two important features: (1) only positive imputed values are produced, and (2) the model error terms are, approximately, normally distributed. Feature 1 is important because no ad hoc procedures are necessary to deal with missing values. Feature 2 is important because the EIA can correctly test the significance of each independent variable, quote an error term that is constant over the range of the independent variables, and assign random normal deviates to imputed values.

The study team developed a model that: (1) is a generalized building model based on basic energy flows into and out of buildings; (2) uses the building model as a framework that is filled in for each building category; and (3) is a linearization of a nonlinear, logarithmic model. (The linearization is much easier to calculate and check diagnostically than the nonlinear model.)

The study also included related model-development activities. Data base calculations and edits are described in Sect. 4, and detailed conclusions appear in Sects. 4.6 through 4.9. The NBECS variables that are most important to an imputation study are listed in Sects. 5.3 and 5.4, and it is suggested that greater effort be spent to obtain these values. Sections 5.1 and 5.2 suggest additional NBECS variables that should be included in the models developed in Sect. 5.4.

A flowchart in Sect. 7.1 shows how to use the ORNL models in the total imputation procedure. Recommended ad hoc procedures to be used when regression adjustment is appropriate include sample weight adjustment (first choice) and lower bound estimation. A method to determine inappropriateness is given in Sect. 7.2.

The reader is encouraged to review the final segment of each section for specific recommendations.

APPENDIX A:

BUILDING REGRESSION CATEGORIES AND BUILDING CLASSES FOR THE 1979 NBECS DATA BASE

Appendix A. Building Regression Categories and Building Classes for the 1979 NBECS Data Base

Building Class Name (Major Building Activity)	Building Class Number (BCLASS1)
<u>Assembly Category</u>	
Assembly Buildings	0200
Social/Public/Civil	0210
Religious Assembly	0220
Recreational Facility	0230
Gymnasium/Indoor-Athletic	0231
Pool Room	0232
Amusement Arcade	0233
Skating Rink	0234
Bowling Alley	0235
Indoor Pool	0236
Other Recreational	0237
Entertainment Building	0240
Archive/Library/Museum, etc.	0241
Observatory/Planetarium	0242
Concert Hall	0243
Coliseum/Arena (enclosed)	0244
Theater/Movie/Cinema	0245
Radio-TV Studio/Station	0246
Nightclub	0247
Other Entertainment	0248
Other Enclosed Assembly Building	0250
Passenger Terminal	0251
Armory	0252
Other Assembly (enclosed)	0253
Non-enclosed or Partial Structure	0260
Stadium	0261
Grandstand	0262
Other Assembly (non-enclosed)	0263
<u>Education Category</u>	
Educational Buildings	0300
Preschool	0310
Elementary School	0320
Junior High School	0330
Senior High School	0340
College or University	0350
Vocational School	0360
<u>Food Sales Category</u>	
Food-related Sales and Service	0400
Cafeteria	0410
Full-service Restaurant	0420
Carry-out Service	0430

Continued

Appendix A. Continued

Building Class Name (Major Building Activity)	Building Class Number (BCLASS1)
Food Sales Category (cont)	
Retail Food Sales	0440
Supermarket	0441
Specialty-food Store	0442
Meat/Seafood Market	0443
Retail Bakery	0444
Farmers Market	0445
Other Retail-food Store	0446
Food Related (except residential)	1030
Food Sales/Other Retail Sales	1031
Food Sales/Other Service	1032
Food Sales/Non-food Service	1033
Food Sales/Other Activity	1034
Health Category	
Health - In-patient Care	0500
Medical-care Hospital	0510
Menial-Health Facility	0520
Rehabilitation Center	0530
Veterinary Hospital/Kennel	0540
Health - Out-patient Care	0600
Medical Clinic	0610
Mental-health Clinic	0620
Dental Clinic	0630
Veterinary Clinic	0640
Industrial Category: Assembly Plants	
Light-assembly - Factory	0730
Heavy-assembly - Factory	0740
Industrial Category: Raw Goods Industrial	
Paper/Chemical, etc - Factory	0750
Metalworks, Glassworks, etc.	0760
Printing/Publishing	0770
Utility or Sanitary Services	0780
Construction/Natural Resource	0790
Industrial Category: Other Industrial	
Industrial Buildings	0700
Food-processing plant	0710
Leather/Textile Mill	0720

continued

Appendix A. Continued

Building Class Name (Major Building Activity)	Building Class Number (BCLASS1)
Retail Sales and Service Category: Shopping	
Shopping Mall	0910
Strip-shopping Center	0920
Retail Sales and Service Category: Retail Sales	
Mercantile/Service	0900
Retail Sales	0930
Hardware, etc - Retail Sales	0931
Department Store - Retail	0933
Furniture, etc - Retail	0934
Drugstore	0935
Multi-retail Establishment	0937
Other Retail Stores	0938
Retail Sales and Service Category: Personal Service	
Services (except food)	0950
Laundry/Car Wash	0951
Post Office	0953
Personal Service	0954
Multi-service Establishment	0955
Other Non-food Service	0956
Retail Sales and Service Category: Mixed Retail/Wholesale	
Non-food Wholesale Goods	0940
Real-estate/Other Commercial	1025
Two or More Services	1050
Service/Retail	1051
Retail/Wholesale	1052
Service/Wholesale	1053
Retail/Wholesale/Service	1054
Automobile Sales Category	
Gas Station	0932
Automobile Dealer	0936
Motor-Vehicle Repair	0952
Office Building Category: General Office	
Office Building	1100

Continued

Appendix A. Continued

Building Class Name (Major Building Activity)	Building Class Number (BCLASS1)
Office Building Category: Professional Office	
Professional Office Building	1110
Office Building Category: Financial Office	
Financial Office Building	1120
Office Building Category: Mixed Use Office	
Data Processing	1130
Computer Center	1131
Other Data-processing	1132
Residental Category: Residential Only	
Residential Housekeeping	1300
Multi-family	1310
High-rise Apartments	1311
Low-rise Apartments	1312
Single Family:	1320
Single Family: Detached	1321
Single-Family: Duplex	1322
Single-Family: Triplex	1323
Single-Family: Guadraplex	1324
Townhouse/Rowhouse	1325
Mobile-home	1330
Residential Category: Residential Mixed Use	
Residential/Other	1010
Residential/Food	1011
Residential/Sales (non-food)	1012
Residential/Office Sales	1013
Residential/Service Activity	1014
Residential/Other Use	1015
Lodging Category: Commercial Lodging	
Short-term Residence	1410
Shelter-home	1411
Motel	1412
Tourist-home	1413
Motel	1414
Convention-Hotel	1415
Inn	1416
Other Short-term Residence	1417

Continued

Appendix A. Continued

Building Class Name (Major Building Activity)	Building Class Number (BCLASS1)
Lodging Category (cont): Other Long Term Lodging	
Residential Non-housekeeping	1400
Long-term Residence	1420
Boarding-house	1421
Orphanage	1422
Home-for-aged/Nursing-home	1423
Convent/Monastery	1424
Dormitory/Sorority/Fraternity	1425
Other Long-term Residence	1426
Warehouse/Storage Category: Refrigerated Warehouse	
Storage/Sales/Manufacturing	1040
Storage/Food Processing	1041
Storage/Non-Food Retail Sales	1042
Storage/Non-Food Wholesale	1043
Storage/Non-Food Manufacturing	1044
Storage	1500
Agricultural Storage	1510
Refrigerated Storage	1530
Other Storage	1540
Warehouse/Storage Category: Non-refrigerated Warehouse	
Non-refrigerated Warehouse	1520
Other Buildings Category	
Agriculture Buildings	0100
Agriculture on Farm	0110
Livestock (Non-Farm)	0120
Agricultural Service	0130
Laboratory	0800
Mechanical/Electrical Laboratory	0810
Medical/Dental Laboratory	0820
Agricultural Laboratory	0830
Other Laboratory	0840
Mixed-Use	1000
Other Mixed-Use Building	1060
Public-order and Safety	1200
Fire Station	1210
Police Station	1220
Jail	1230
Reformatory	1240
Penitentiary	1250
Courthouse	1260
Sheriffs Office	1270
Other Public Order/Safety	1280
Other	1600

Continued

Appendix A. Continued

Building Class Name (Major Building Activity)	Building Class Number (BCLASS1)
<u>Other Buildings Category (cont)</u>	
Crematorium	1610
Parking Garage	1620
Hangar	1630
Telephone Exchange	1640
Rest Rooms	1650
Other	1660
Don't Know	9998
Not Ascertained	9999
<u>Other Buildings Category: Vacant</u>	
Vacant	1700

APPENDIX B:

SOURCE STATEMENTS FOR THE CREATION OF CONVERT.NATGAS

```
//HT1UHJT JOB (6616,X10,2,,,,),
// 'HOW TSAO ** ORNL ',TIME=(,20)
/*JOBPARM LINES=10
/*ROUTE  PRINT RMT030
// EXEC SAS,REGION=1024K,OPTIONS='MACRO DQUOTE MPRINT',TIME=(,20)
//CLASS DD DSN=CN6616.RL2.GENE.NBECS79.TAPE2.OAKR.SASTEST3,DISP=SHR
//SASLIB DD DSN=CN6616.HT1.NBECS79.SASLIB3,DISP=(OLD,KEEP)
//CONVERT DD DSN=CN6616.HT1.CONVERT.NATGAS.DATA73,DISP=(NEW,CATLG),
// UNIT=DASD,SPACE=(TRK,(800,400),RLSE)
//SYSIN DD *
DATA NBECS79 ;
  SET CLASS.NBECS79;
  RENAME ES11 = ES1                  SPLID11 = SPLID1
         ES21 = ES2                  SPLID21 = SPLID2
         ES31 = ES3                  SPLID31 = SPLID3
         ES41 = ES4                  SPLID41 = SPLID4
         ES51 = ES5                  SPLID51 = SPLID5
         ES61 = ES6                  SPLID61 = SPLID6
         ES71 = ES7                  SPLID71 = SPLID7
         ES81 = ES8                  SPLID81 = SPLID8
         ES91 = ES9                  SPLID91 = SPLID9
         CNSMP11 = CNSMP1            NSUPL11 = NSUPL1
         CNSMP21 = CNSMP2            NSUPL21 = NSUPL2
         CNSMP31 = CNSMP3            NSUPL31 = NSUPL3
         CNSMP41 = CNSMP4            NSUPL41 = NSUPL4
         CNSMP51 = CNSMP5            NSUPL51 = NSUPL5
         CNSMP61 = CNSMP6            NSUPL61 = NSUPL6
         CNSMP71 = CNSMP7            NSUPL71 = NSUPL7
         CNSMP81 = CNSMP8            NSUPL81 = NSUPL8
         CNSMP91 = CNSMP9            NSUPL91 = NSUPL9
         BTUS11 = BTUS1              MLTBL11 = MLTBL1
         BTUS21 = BTUS2              MLTBL21 = MLTBL2
         BTUS31 = BTUS3              MLTBL31 = MLTBL3
         BTUS41 = BTUS4              MLTBL41 = MLTBL4
         BTUS51 = BTUS5              MLTBL51 = MLTBL5
         BTUS61 = BTUS6              MLTBL61 = MLTBL6
         BTUS71 = BTUS7              MLTBL71 = MLTBL7
         BTUS81 = BTUS8              MLTBL81 = MLTBL8
         BTUS91 = BTUS9              MLTBL91 = MLTBL9
         CNSD11 = CNSD1              NBLS11 = NBLS1
         CNSD21 = CNSD2              NBLS21 = NBLS2
         CNSD31 = CNSD3              NBLS31 = NBLS3
         CNSD41 = CNSD4              NBLS41 = NBLS4
         CNSD51 = CNSD5              NBLS51 = NBLS5
         CNSD61 = CNSD6              NBLS61 = NBLS6
         CNSD71 = CNSD7              NBLS71 = NBLS7
         CNSD81 = CNSD8              NBLS81 = NBLS8
         CNSD91 = CNSD9              NBLS91 = NBLS9
         HEAT11 = HEATU1             BLCOV11 = BLCOV1
         HEAT21 = HEATU2             BLCOV21 = BLCOV2
         HEAT31 = HEATU3             BLCOV31 = BLCOV3
         HEAT41 = HEATU4             BLCOV41 = BLCOV4
         HEAT51 = HEATU5             BLCOV51 = BLCOV5
         HEAT61 = HEATU6             BLCOV61 = BLCOV6
         HEAT71 = HEATU7             BLCOV71 = BLCOV7
         HEAT81 = HEATU8             BLCOV81 = BLCOV8
```

```
HEAT91 = HEATU9
COOL11 = COOLU1
COOL21 = COOLU2
COOL31 = COOLU3
COOL41 = COOLU4
COOL51 = COOLU5
COOL61 = COOLU6
COOL71 = COOLU7
COOL81 = COOLU8
COOL91 = COOLU9
WATER11 = WATERU1
WATER21 = WATERU2
WATER31 = WATERU3
WATER41 = WATERU4
WATER51 = WATERU5
WATER61 = WATERU6
WATER71 = WATERU7
WATER81 = WATERU8
WATER91 = WATERU9
GENER11 = GENERU1
GENER21 = GENERU2
GENER31 = GENERU3
GENER41 = GENERU4
GENER51 = GENERU5
GENER61 = GENERU6
GENER71 = GENERU7
GENER81 = GENERU8
GENER91 = GENERU9
MANUF11 = MANUFU1
MANUF21 = MANUFU2
MANUF31 = MANUFU3
MANUF41 = MANUFU4
MANUF51 = MANUFU5
MANUF61 = MANUFU6
MANUF71 = MANUFU7
MANUF81 = MANUFU8
MANUF91 = MANUFU9
COOK11 =COOKU1
COOK21 =COOKU2
COOK31 =COOKU3
COOK41 =COOKU4
COOK51 =COOKU5
COOK61 =COOKU6
COOK71 =COOKU7
COOK81 =COOKU8
COOK91 =COOKU9
BOILR11 = BOILRU1
BOILR21 = BOILRU2
BOILR31 = BOILRU3
BOILR41 = BOILRU4
BOILR51 = BOILRU5
BLCOV91 = BLCOV9
NMETR11 = NMETR1
NMETR21 = NMETR2
NMETR31 = NMETR3
NMETR41 = NMETR4
NMETR51 = NMETR5
NMETR61 = NMETR6
NMETR71 = NMETR7
NMETR81 = NMETR8
NMETR91 = NMETR9
NCUST11 = NCUST1
NCUST21 = NCUST2
NCUST31 = NCUST3
NCUST41 = NCUST4
NCUST51 = NCUST5
NCUST61 = NCUST6
NCUST71 = NCUST7
NCUST81 = NCUST8
NCUST91 = NCUST9
UNIT11 = UNIT1
UNIT21 = UNIT2
UNIT31 = UNIT3
UNIT41 = UNIT4
UNIT51 = UNIT5
UNIT61 = UNIT6
UNIT71 = UNIT7
UNIT81 = UNIT8
UNIT91 = UNIT9
COST11 = COST1
COST21 = COST2
COST31 = COST3
COST41 = COST4
COST51 = COST5
COST61 = COST6
!OCOST71 = COST7
COST81 = COST8
COST91 = COST9
CSTD11 = CSTD1
CSTD21 = CSTD2
CSTD31 = CSTD3
CSTD41 = CSTD4
CSTD51 = CSTD5
CSTD61 = CSTD6
CSTD71 = CSTD7
CSTD81 = CSTD8
CSTD91 = CSTD9
WRQ11 = WRQ1
WRQ21 = WRQ2
WRQ31 = WRQ3
WRQ41 = WRQ4
WRQ51 = WRQ5
```

```
BOILR61 = BOILRU6        WRQ61 = WRQ6
BOILR71 = BOILRU7        WRQ71 = WRQ7
BOILR81 = BOILRU8        WRQ81 = WRQ8
BOILR91 = BOILRU9        WRQ91 = WRQ9
WOBT11 = WOBT1
WOBT21 = WOBT2
WOBT31 = WOBT3
WOBT41 = WOBT4
WOBT51 = WOBT5
WOBT61 = WOBT6
WOBT71 = WOBT7
WOBT81 = WOBT8
WOBT91 = WOBT9;
DATA CONVERT.NATGAS;
SET NBECS79;
ARRAY ES (I) ES1-ES9;
ARRAY BTUS (I) BTUS1-BTUS9;
ARRAY CNSD (I) CNSD1-CNSD9;
ARRAY CNSMP (I) CNSMP1-CNSMP9;
ARRAY WATERU (I) WATERU1-WATERU9;
ARRAY HEATU (I) HEATU1-HEATU9;
ARRAY COOLU (I) COOLU1-COOLU9;
ARRAY GENERU (I) GENERU1-GENERU9;
ARRAY MANUFU (I) MANUFU1-MANUFU9;
ARRAY COOKU (I) COOKU1-COOKU9;
ARRAY BOILRU (I) BOILRU1-BOILRU9;
ARRAY WOBT (I) WOBT1-WOBT9;
ARRAY SPLID (I) SPLID1-SPLID9;
ARRAY NSUPL (I) NSUPL1-NSUPL9;
ARRAY MLTBL (I) MLTBL1-MLTBL9;
ARRAY NBLS (I) NBLS1-NBLS9;
ARRAY BLCOV (I) BLCOV1-BLCOV9;
ARRAY NMETR (I) NMETR1-NMETR9;
ARRAY NCUST (I) NCUST1-NCUST9;
ARRAY UNIT (I) UNIT1-UNIT9;
ARRAY COST (I) COST1-COST9;
ARRAY CSTD (I) CSTD1-CSTD9;
ARRAY WRQ (I) WRQ1-WRQ9;
RETAIN B1-B9 0;
ARRAY B (I) B1-B9;
TOTB = 0;
DO OVER B;
  B = 0;
END;
DO OVER ES;
  DO I = 1 TO 9;
    IF ES = '22' THEN DO;
      B = 1;
      TOTB = TOTB + B;
    END;
  END;
END;
```

```
          IF TOTB >=1;
          * ONLY BUILDINGS REPORTED USE OF NATURAL GAS ARE KEPT;
          IF TOTB = 1 THEN DO;
            DO OVER ES;
               DO I = 1 TO 9;
                 IF ES = '22' THEN DO;
                   BTU = BTUS;
                   DAYS = CNSD;
                   CNSUNIT = CNSMP;
                   HEATX = HEATU;
                   COOLX = COOLU;
                   WATERX = WATERU;
                   GENERX = GENERU;
                   MANUFX = MANUFU;
                   COOKX = COOKU;
                   BOILRX = BOILRU;
                   WOBTX = WOBT;
                   SPLIDX = SPLID;
                   NSUPLX =NSUPL;
                   MLTBLX = MLTBL;
                   NBLSX = NBLS;
                   BLCOVX = BLCOV;
                   NMETRX = NMETR;
                   NCUSTX = NCUST;
                   UNITX = UNIT;
                   COSTX = COST;
                   CSTDX = CSTD;
                   WRQX =WRQ;
                   END;
                 END;
               END;
             END;
* VARIABLES CREATED IN THE DO LOOP ABOVE ARE NEW VARIABLES;
DROP ES1-ES9 BTUS1-BTUS9 CNSD1-CNSD9 CNSMP1-CNSMP9 WATERU1-WATERU9;
DROP HEATU1-HEATU9 COOLU1-COOLU9 BOILRU1-BOILRU9 WOBT1-WOBT9;
DROP GENERU1-GENERU9 MANUFU1-MANUFU9 COOKU1-COOKU9 SPLID1-SPLID9;
DROP NSUPL1-NSUPL9 MLTBL1-MLTBL9 NBLS1-NBLS9 BLCOV1-BLCOV9;
DROP NMETR1-NMETR9 NCUST1-NCUST9 UNIT1-UNIT9 COST1-COST9 ;
DROP CSTD1-CSTD9 WRQ1-WRQ9;
FORMAT BTU COMMA20. CNSUNIT COMMA17. WOBTX WRQX MISS1CH.
  NSUPLX MISS2CH. NBLSX NCUSTX NMETRX MISS4CH.
  MLTBLX $MLTBL. BOILRX $BOILR. BLCOVX $BLCOV.
  UNITX $UNIT. COSTX COMMA12.
  HEATX COOLX WATERX GENERX MANUFX COOKX $USE.;
PROC CONTENTS DATA = CONVERT.NATGAS;
//
```

APPENDIX C:

SOURCE STATEMENTS FOR THE CREATION OF CONVERT.ELECT

```
//HT1UHJT JOB (6616,X10,2,,,,),
// 'HOW TSAO ** ORNL ',TIME=(,20)
/*JOBPARM LINES=10
/*ROUTE  PRINT RMT030
// EXEC SAS,REGION=1024K,OPTIONS='MACRO DQUOTE MPRINT',TIME=(,20)
//CLASS DD DSN=CN6616.RL2.GENE.NBECS79.TAPE2.OAKR.SASTEST3,DISP=SHR
//SASLIB DD DSN=CN6616.HT1.NBECS79.SASLIB3,DISP=(OLD,KEEP)
//CONVERT DD DSN=CN6616.HT1.CONVERT.ELECT.DATA73,DISP=(NEW,CATLG),
// UNIT=DASD,SPACE=(TRK,(800,400),RLSE)
//SYSIN DD *
DATA NBECS79 ;
  SET CLASS.NBECS79;
  RENAME ES11 = ES1                SPLID11 = SPLID1
         ES21 = ES2                SPLID21 = SPLID2
         ES31 = ES3                SPLID31 = SPLID3
         ES41 = ES4                SPLID41 = SPLID4
         ES51 = ES5                SPLID51 = SPLID5
         ES61 = ES6                SPLID61 = SPLID6
         ES71 = ES7                SPLID71 = SPLID7
         ES81 = ES8                SPLID81 = SPLID8
         ES91 = ES9                SPLID91 = SPLID9
         CNSMP11 = CNSMP1          NSUPL11 = NSUPL1
         CNSMP21 = CNSMP2          NSUPL21 = NSUPL2
         CNSMP31 = CNSMP3          NSUPL31 = NSUPL3
         CNSMP41 = CNSMP4          NSUPL41 = NSUPL4
         CNSMP51 = CNSMP5          NSUPL51 = NSUPL5
         CNSMP61 = CNSMP6          NSUPL61 = NSUPL6
         CNSMP71 = CNSMP7          NSUPL71 = NSUPL7
         CNSMP81 = CNSMP8          NSUPL81 = NSUPL8
         CNSMP91 = CNSMP9          NSUPL91 = NSUPL9
         BTUS11 = BTUS1            MLTBL11 = MLTBL1
         BTUS21 = BTUS2            MLTBL21 = MLTBL2
         BTUS31 = BTUS3            MLTBL31 = MLTBL3
         BTUS41 = BTUS4            MLTBL41 = MLTBL4
         BTUS51 = BTUS5            MLTBL51 = MLTBL5
         BTUS61 = BTUS6            MLTBL61 = MLTBL6
         BTUS71 = BTUS7            MLTBL71 = MLTBL7
         BTUS81 = BTUS8            MLTBL81 = MLTBL8
         BTUS91 = BTUS9            MLTBL91 = MLTBL9
         CNSD11 = CNSD1            NBLS11 = NBLS1
         CNSD21 = CNSD2            NBLS21 = NBLS2
         CNSD31 = CNSD3            NBLS31 = NBLS3
         CNSD41 = CNSD4            NBLS41 = NBLS4
         CNSD51 = CNSD5            NBLS51 = NBLS5
         CNSD61 = CNSD6            NBLS61 = NBLS6
         CNSD71 = CNSD7            NBLS71 = NBLS7
         CNSD81 = CNSD8            NBLS81 = NBLS8
         CNSD91 = CNSD9            NBLS91 = NBLS9
         HEAT11 = HEATU1           BLCOV11 = BLCOV1
         HEAT21 = HEATU2           BLCOV21 = BLCOV2
         HEAT31 = HEATU3           BLCOV31 = BLCOV3
         HEAT41 = HEATU4           BLCOV41 = BLCOV4
         HEAT51 = HEATU5           BLCOV51 = BLCOV5
         HEAT61 = HEATU6           BLCOV61 = BLCOV6
```

```
HEAT71 = HEATU7
HEAT81 = HEATU8
HEAT91 = HEATU9
COOL11 = COOLU1
COOL21 = COOLU2
COOL31 = COOLU3
COOL41 = COOLU4
COOL51 = COOLU5
COOL61 = COOLU6
COOL71 = COOLU7
COOL81 = COOLU8
COOL91 = COOLU9
WATER11 = WATERU1
WATER21 = WATERU2
WATER31 = WATERU3
WATER41 = WATERU4
WATER51 = WATERU5
WATER61 = WATERU6
WATER71 = WATERU7
WATER81 = WATERU8
WATER91 = WATERU9
GENER11 = GENERU1
GENER21 = GENERU2
GENER31 = GENERU3
GENER41 = GENERU4
GENER51 = GENERU5
GENER61 = GENERU6
GENER71 = GENERU7
GENER81 = GENERU8
GENER91 = GENERU9
MANUF11 = MANUFU1
MANUF21 = MANUFU2
MANUF31 = MANUFU3
MANUF41 = MANUFU4
MANUF51 = MANUFU5
MANUF61 = MANUFU6
MANUF71 = MANUFU7
MANUF81 = MANUFU8
MANUF91 = MANUFU9
COOK11 =COOKU1
COOK21 =COOKU2
COOK31 =COOKU3
COOK41 =COOKU4
COOK51 =COOKU5
COOK61 =COOKU6
COOK71 =COOKU7
COOK81 =COOKU8
COOK91 =COOKU9
BOILR11 = BOILRU1
BOILR21 = BOILRU2
BOILR31 = BOILRU3
BOILR41 = BOILRU4
BOILR51 = BOILRU5
BOILR61 = BOILRU6
BLCOV71 = BLCOV7
BLCOV81 = BLCOV8
BLCOV91 = BLCOV9
NMETR11 = NMETR1
NMETR21 = NMETR2
NMETR31 = NMETR3
NMETR41 = NMETR4
NMETR51 = NMETR5
NMETR61 = NMETR6
NMETR71 = NMETR7
NMETR81 = NMETR8
NMETR91 = NMETR9
NCUST11 = NCUST1
NCUST21 = NCUST2
NCUST31 = NCUST3
NCUST41 = NCUST4
NCUST51 = NCUST5
NCUST61 = NCUST6
NCUST71 = NCUST7
NCUST81 = NCUST8
NCUST91 = NCUST9
UNIT11 = UNIT1
UNIT21 = UNIT2
UNIT31 = UNIT3
UNIT41 = UNIT4
UNIT51 = UNIT5
UNIT61 = UNIT6
UNIT71 = UNIT7
UNIT81 = UNIT8
UNIT91 = UNIT9
COST11 = COST1
COST21 = COST2
COST31 = COST3
COST41 = COST4
COST51 = COST5
COST61 = COST6
COST71 = COST7
COST81 = COST8
COST91 = COST9
CSTD11 = CSTD1
CSTD21 = CSTD2
CSTD31 = CSTD3
CSTD41 = CSTD4
CSTD51 = CSTD5
CSTD61 = CSTD6
CSTD71 = CSTD7
CSTD81 = CSTD8
CSTD91 = CSTD9
WRQ11 = WRQ1
WRQ21 = WRQ2
WRQ31 = WRQ3
WRQ41 = WRQ4
WRQ51 = WRQ5
WRQ61 = WRQ6
```

```
          BOILR71 = BOILRU7        WRQ71 = WRQ7
          BOILR81 = BOILRU8        WRQ81 = WRQ8
          BOILR91 = BOILRU9        WRQ91 = WRQ9
          WOBT11 = WOBT1
          WOBT21 = WOBT2
          WOBT31 = WOBT3
          WOBT41 = WOBT4
          WOBT51 = WOBT5
          WOBT61 = WOBT6
          WOBT71 = WOBT7
          WOBT81 = WOBT8
          WOBT91 = WOBT9;
DATA CONVERT.ELECT;
  SET NBECS79;
  ARRAY ES (I) ES1-ES9;
  ARRAY BTUS (I) BTUS1-BTUS9;
  ARRAY CNSD (I) CNSD1-CNSD9;
  ARRAY CNSMP (I) CNSMP1-CNSMP9;
  ARRAY WATERU (I) WATERU1-WATERU9;
  ARRAY HEATU (I) HEATU1-HEATU9;
  ARRAY COOLU (I) COOLU1-COOLU9;
  ARRAY GENERU (I) GENERU1-GENERU9;
  ARRAY MANUFU (I) MANUFU1-MANUFU9;
  ARRAY COOKU (I) COOKU1-COOKU9;
  ARRAY BOILRU (I) BOILRU1-BOILRU9;
  ARRAY WOBT (I) WOBT1-WOBT9;
  ARRAY SPLID (I) SPLID1-SPLID9;
  ARRAY NSUPL (I) NSUPL1-NSUPL9;
  ARRAY MLTBL (I) MLTBL1-MLTBL9;
  ARRAY NBLS (I) NBLS1-NBLS9;
  ARRAY BLCOV (I) BLCOV1-BLCOV9;
  ARRAY NMETR (I) NMETR1-NMETR9;
  ARRAY NCUST (I) NCUST1-NCUST9;
  ARRAY UNIT (I) UNIT1-UNIT9;
  ARRAY COST (I) COST1-COST9;
  ARRAY CSTD (I) CSTD1-CSTD9;
  ARRAY WRQ (I) WRQ1-WRQ9;
  RETAIN B1-B9 0;
  ARRAY B (I) B1-B9;
  TOTB = 0;
  DO OVER B;
    B = 0;
  END;
  DO OVER ES;
    DO I = 1 TO 9;
      IF ES = '21' THEN DO;
        B = 1;
        TOTB = TOTB + B;
      END;
    END;
  END;
IF TOTB >=1;
* ONLY BUILDINGS REPORTED USE OF NATURAL GAS ARE KEPT;
IF TOTB = 1 THEN DO;
```

```
DO OVER ES;
  DO I = 1 TO 9;
    IF ES = '21' THEN DO;
      BTU = BTUS;
      DAYS = CNSD;
      CNSUNIT = CNSMP;
      HEATX = HEATU;
      COOLX = COOLU;
      WATERX = WATERU;
      GENERX = GENERU;
      MANUFX = MANUFU;
      COOKX = COOKU;
      BOILRX = BOILRU;
      WOBTX = WOBT;
      SPLIDX = SPLID;
      NSUPLX =NSUPL;
      MLTBLX = MLTBL;
      NBLSX = NBLS;
      BLCOVX = BLCOV;
      NMETRX = NMETR;
      NCUSTX = NCUST;
      UNITX = UNIT;
      COSTX = COST;
      CSTDX = CSTD;
      WRQX =WRQ;
    END;
  END;
 END;
END;
* VARIAbLES CREATED IN THE DO LOOP ABOVE ARE NEW VARIABLES;
DROP ES1-ES9 BTUS1-BTUS9 CNSD1-CNSD9 CNSMP1-CNSMP9 WATERU1-WATERU9;
DROP HEATU1-HEATU9 COOLU1-COOLU9 BOILRU1-BOILRU9 WOBT1-WOBT9;
DROP GENERU1-GENERU9 MANUFU1-MANUFU9 COOKU1-COOKU9 SPLID1-SPLID9;
DROP NSUPL1-NSUPL9 MLTBL1-MLTBL9 NBLS1-NBLS9 BLCOV1-BLCOV9;
DROP NMETR1-NMETR9 NCUST1-NCUST9 UNIT1-UNIT9 COST1-COST9 ;
DROP CSTD1-CSTD9 WRQ1-WRQ9;
FORMAT BTU COMMA20. CNSUNIT COMMA17. WOBTX WRQX MISS1CH.
  NSUPLX MISS2CH. NBLSX NCUSTX NMETRX MISS4CH.
  MLTBLX $MLTBL. BOILRX $BOILR. BLCOVX $BLCOV.
  UNITX $UNIT. COSTX COMMA12.
  HEATX COOLX WATERX GENERX MANUFX COOKX $USE.;
PROC CONTENTS DATA = CONVERT.ELECT;
//
```

APPENDIX D:

SOURCE STATEMENTS FOR THE CREATION OF WORKING.NATGAS

```
//HT1UHJT JOB (6616,X10,2,,,,),
// 'HOW TSAO ** ORNL ',TIME=(1,30),CLASS=0
/*JOBPARM LINES=10
/*ROUTE  PRINT RMT030
// EXEC SAS,REGION=1024K,OPTIONS='MACRO DQUOTE MPRINT',TIME=(1,30)
//CONVERT DD DSN=CN6616.HT1.CONVERT.NATGAS.DATA73,DISP=SHR
//SASLIB DD DSN=CN6616.HT1.NBECS79.SASLIB3,DISP=(OLD,KEEP)
//DISKOUT DD DSN=CN6616.HT1.IMPUTED.UPDATE,DISP=SHR
//WORKING DD DSN=CN6616.HT1.WORKING.NATGAS.DATA109,DISP=(NEW,CATLG),
// UNIT=DASD,SPACE=(TRK,(800,400),RLSE)
//WORKEL DD DSN=CN6616.HT1.WORKING.ELECT.DATA924,DISP=SHR
//SAS.WORK DD UNIT=SYSDA,SPACE=(6160,(800,400),,,ROUND)
//SYSIN DD *
OPTIONS GEN=2;
DATA ELECT;
  SET WORKEL.ELECT;
  ELCONS = CNSUNIT;
  ENDUSEZ1 = ENDUSE1;
  ENDUSEZ2 = ENDUSE2;
  ENDUSEZ3 = ENDUSE3;
  ENDUSEZ3 = ENDUSE3;
  ENDUSEZ4 = ENDUSE4;
  ENDUSEZ5 = ENDUSE5;
  ENDUSEZ6 = ENDUSE6;
  KEEP BLDGID1 ELCONS ENDUSEZ1-ENDUSEZ6;
DATA IMPUTE;
  SET DISKOUT.IMPUTE;
DATA NATGAS1;
  SET CONVERT.NATGAS;
PROC SORT NODUP DATA = IMPUTE;
  BY BLDGID1;
PROC SORT DATA = ELECT;
  BY BLDGID1;
PROC SORT DATA = NATGAS1;
  BY BLDGID1;
DATA NATGAS;
  MERGE NATGAS1(IN= F1) IMPUTE(IN=F2);
  BY BLDGID1;
  IF F1;
PROC SORT DATA = NATGAS;
  BY BLDGID1;
DATA WORKING.NATGAS;
  MERGE NATGAS(IN=F1) ELECT(IN=F2);
  BY BLDGID1;
  IF F1;
  IF TOTB = 1;
  A = B1 + B2 + B3;
*-------------------------------------------------------------------;
* A = 1 MEANS THAT THE FUEL IS ONE OF THE THREE PRIMARY ENERGY      ;
*       SOURCES                                                     ;
* BUILDINGS THAT REPORTED NATURAL GAS USE ON A SINGLE ENERGY        ;
* SOURCE FIELD WILL HAVE TOTB = 1, AND WILL BE INCLUDED IN THE      ;
* DATA BASE CREATED HERE.                                           ;
*-------------------------------------------------------------------;
* WE NOW CREATE VARIABLE "DAYCLASS"                                 ;
  IF 0<= DAYS < 31 THEN DAYCLASS = 1;
```

```
    ELSE IF 31 <= DAYS < 331 THEN DAYCLASS = 2;
    ELSE IF 331 <= DAYS <= 365 THEN DAYCLASS = 3;
    ELSE DAYCLASS = .;
*-----------------------------------------------------------------;
* CREATE VARIABLE "BTUCLASS"                                      ;
  IF BTU = 0 THEN BTUCLASS = 0;
    ELSE IF BTU > 0 THEN BTUCLASS = 1;
    ELSE BTUCLASS = .;
*-----------------------------------------------------------------;
* CONVERT SURVEY VARIABLE VALUES TO WORKING VARIABLE VALUES:      ;
  IF HEATP1 > 100 OR HEATP1 < 0 THEN HEATP1 = .;
  IF COOLP1 > 100 OR COOLP1 < 0 THEN COOLP1 = .;
  IF RESP1 > 100 OR RESP1 < 0 THEN RESP1 = .;
    IF RESU1 = '1' OR RESUSE1 = '1' THEN PCTRES = RESP1;
    ELSE IF RESU1 = '2' AND RESUSE1 = '2' THEN PCTRES = 0;
    ELSE PCTRES = .;
  IF VACP1 > 100 OR VACP1 < 0 THEN VACP1 = .;
    IF PORVAC1 = '1' THEN PCTVAC = VACP1;
    ELSE IF PORVAC1 = '2' THEN PCTVAC = 0;
    ELSE PCTVAC = .;
  IF GLASPC1 > 4 THEN GLASPC1 = .;
  IF AVGNHR1 > 168 OR AVGNHR1 < 0 THEN AVGNHR1 = .;
*-----------------------------------------------------------------;
* CREATE WORKING VARIABLES ENDUSE1 - ENDUSE6.        ;
  IF HEATX ='1' THEN ENDUSE1 = 1;
    ELSE IF HEATX ='2' THEN ENDUSE1 = 0;
    ELSE ENDUSE1 = .;
  IF COOLX ='1' THEN ENDUSE2 = 1;
    ELSE IF COOLX ='2' THEN ENDUSE2 = 0;
    ELSE ENDUSE2 = .;
  IF WATERX ='1' THEN ENDUSE3 = 1;
    ELSE IF WATERX ='2' THEN ENDUSE3 = 0;
    ELSE ENDUSE3 = .;
  IF GENERX ='1' THEN ENDUSE4 = 1;
    ELSE IF GENERX ='2' THEN ENDUSE4 = 0;
    ELSE ENDUSE4 = .;
  IF MANUFX ='1' THEN ENDUSE5 = 1;
    ELSE IF MANUFX ='2' THEN ENDUSE5 = 0;
    ELSE ENDUSE5 = .;
  IF COOKX ='1' THEN ENDUSE6 = 1;
    ELSE IF COOKX ='2' THEN ENDUSE6 = 0;
    ELSE ENDUSE6 = .;
*-----------------------------------------------------------------;
* CREATE VARIABLES SFHEAT, SFCOOL, AND SFVAC                      ;
  SFHEAT = ENDUSE1 * SQFT1 * (HEATP1/100);
  SFCOOL = ENDUSE2 * SQFT1 * (COOLP1/100);
  SFVAC = SQFT1 * (PCTVAC/100);
  SFRESI = SQFT1 * (PCTRES/100);
*-----------------------------------------------------------------;
* WE NOW CREATE WORKING VARIABLE "PCTGLA";
  IF GLASPC1 = '1' THEN PCTGLA = 1;
  IF GLASPC1 = '2' AND GLASP1 = '1' THEN PCTGLA = 2;
  IF GLASPC1 = '3' AND GLASP1 = '2' THEN PCTGLA = 3;
  IF GLASPC1 = '4' THEN PCTGLA = 4;
```

```
*---------------------------------------------------------------------;
* WE NOW CREATE WORKING VARIABLES "WTRZON1 - WTRZON5";
  IF CLIMAT1 = '1' THEN WTRZON1 = 1;
  ELSE IF CLIMAT1 NE . THEN WTRZON1 = 0;
  IF CLIMAT1 = '2' THEN WTRZON2 = 1;
  ELSE IF CLIMAT1 NE . THEN WTRZON2 = 0;
  IF CLIMAT1 = '3' THEN WTRZON3 = 1;
  ELSE IF CLIMAT1 NE . THEN WTRZON3 = 0;
  IF CLIMAT1 = '4' THEN WTRZON4 = 1;
  ELSE IF CLIMAT1 NE . THEN WTRZON4 = 0;
  IF CLIMAT1 = '6' THEN WTRZON5 = 1;
  ELSE IF CLIMAT1 NE . THEN WTRZON5 = 0;
*---------------------------------------------------------------------;
* WE NOW CREATE WORKING VARIABLES REG1 - REG4;
  IF REGION1 = '1' THEN REG1 = 1;
  ELSE REG1 = 0;
  IF REGION1 = '2' THEN REG2 = 1;
  ELSE REG2 = 0;
  IF REGION1 = '3' THEN REG3 = 1;
  ELSE REG3 = 0;
  IF REGION1 = '4' THEN REG4 = 1;
  ELSE REG4 = 0;
*---------------------------------------------------------------------;
* WE NOW CREATE WORKING VARIABLES YRCAT,SFCAT AND EMPCAT.;
      YRCAT = 0; SFCAT = 0; EMPCAT = 0;
      YRCAT = YRCONC1;
      IF YRCAT = 98 OR YRCAT = 99 THEN YRCAT = .;
      SFCAT = SQFTC1;
      IF SFCAT = 98 OR SFCAT = 99 THEN SFCAT = .;
      EMPCAT = NWKERC1;
      IF EMPCAT = 98 OR EMPCAT = 99 THEN EMPCAT = .;
*---------------------------------------------------------------------;
* WE NOW CREATE WORKING VARIABLES TD7 AND TC4;
      TD7 = 0; TC4 = 0;
      IF WOFP1 = '1' THEN TD7 = 1;
      IF CST1 = '4' THEN TC4 = 1;
*---------------------------------------------------------------------;
%MACRO REGCAT;
      CLASS = BCLASS1;
      REGCAT = 001;
      IF BLDGID1 = 2756 THEN SQFT1 = 759;
      IF BLDGID1 = 4158 THEN SQFT1 = 2074;
      IF BLDGID1 = 4158 THEN SFCAT = 2;
      IF BLDGID1 = 163 OR
         BLDGID1 = 2870 OR
         BLDGID1 = 4358 OR
         BLDGID1 = 5239 OR
         BLDGID1 = 5243 OR
         BLDGID1 = 7114 OR
         BLDGID1 = 3766 OR
         BLDGID1 = 4973 OR
         BLDGID1 = 5015 OR
         BLDGID1 = 5019 OR
         BLDGID1 = 5021 OR
```

```
         BLDGID1 = 2646 OR
         BLDGID1 = 1764 OR
         BLDGID1 = 3843 OR
         BLDGID1 = 5058 THEN REGEDIT = 001;
%*-----------------------------------------------------------------;
%* THIS MACRO PROGRAM DEFINES THE 26 REGRESSION CATEGORIES FOR ;
%* NATURAL GAS USE BUILDINGS. THE VARIABLE REGCAT IS A GROUP ID;
%* WHICH ASSIGNS A BUILDING IN THE SAMPLE TO ONE OF THE 26     ;
%* BUILDING GROUPS. THE VARIABLE REGEDIT IDENTIFIES OUTLIERS OF;
%* THE CORRESPONDING REGRESSION CATEGORY.                      ;
%* EXAMPLE: SUPPOSE A BUILDING IS SELECTED FROM REGCAT = 060.  ;
%*          IF ITS VALUE OF REGEDIT = 060 THEN THIS BUILDING   ;
%*             IS AN OUTLIER WHICH IS TO BE DELETED FROM       ;
%*             REGRESSION ANALYSIS.                            ;
%*          IF ITS VALUE OF REGEDIT = 001 THEN THIS BUILDING   ;
%*             IS TO BE DELETED FROM THE MASTER DATA SET.      ;
%*          IF ITS VALUE OF REGEDIT = . THEN THIS BUILDING IS  ;
%*             NOT AN IDENTIFIED OUTLIER.                      ;
%*-----------------------------------------------------------------;
 %*ASSEMBLY BUILDINGS CATEGORY;
    IF BCWM1 = '02' THEN REGCAT = 010;
 %*-----------------------------;
 %*EDUCATIN BUILDINGS CATEGORY;
    IF BCWM1 = '03' THEN DO; REGCAT = 020;
    IF CNSUNIT < 10000000 AND SQFT1 > 600000 THEN REGEDIT = 020;
    END;
 %*-----------------------------;
 %*FOOD SALES CATGORY;
    IF BCWM1 = '04' OR
       BCLASS1 = '1030' OR
       BCLASS1 = '1031' OR
       BCLASS1 = '1032' OR
       BCLASS1 = '1033' OR
       BCLASS1 = '1034' THEN DO; REGCAT = 030;
    IF CNSUNIT > 20000000 THEN REGEDIT = 030;
    END;
 %*-----------------------------;
 %*HEALTH CARE CATEGORY WITH SQ.FT. >= 350,000 ;
    IF BCWM1 = '05' AND SQFT1 >= 350000 THEN DO; REGCAT = 040;
    IF CNSUNIT > 500000000 THEN REGEDIT = 040;
    END;
 %*-----------------------------;
 %*HEALTH CARE CATEGORY WITH SQ.FT. < 350,000 ;
    IF BCWM1 = '05' AND SQFT1 < 350000 THEN REGCAT = 050;
 %*-----------------------------;
 %*ASSEMBLY PLANTS CATEGORY;
    IF BCLASS1 = '0730' OR
       BCLASS1 = '0740' THEN DO; REGCAT =060;
    IF (CNSUNIT < 100000000 AND COSTX > 600000) OR
       COSTX > 1000000 OR
       NWKER1 > 5000 THEN REGEDIT = 060;
    END;
```

```
%*-----------------------------;
%*RAW GOODS INDUSTRIAL CATEGORY;
    IF BCLASS1 = '0750' OR
       BCLASS1 = '0760' OR
       BCLASS1 = '0770' OR
       BCLASS1 = '0780' OR
       BCLASS1 = '0790' THEN DO; REGCAT = 070;
    IF (CNSUNIT < 50000000 AND COSTX > 500000) OR
       (CNSUNIT < 50000000 AND SQFT1 > 1000000) OR
       NWKER1 > 5000 OR
       (CNSUNIT > 200000000 AND SQFT1 < 100000) THEN REGEDIT = 070;
    END;
%*-----------------------------;
%*OTHER INDUSTRIAL CATEGORY;
    IF BCLASS1 = '0700' OR
       BCLASS1 = '0710' OR
       BCLASS1 = '0720' THEN DO; REGCAT = 080;
    IF (CNSUNIT > 700000000 AND SQFT1 < 500000) OR
       NWKER1 > 2400 THEN REGEDIT = 080;
    END;
%*-----------------------------;
%*SHOPPING CENTER CATEGORY;
    IF BCLASS1 = '0910' OR
       BCLASS1 = '0920' THEN DO; REGCAT = 090;
    IF (CNSUNIT > 50000000 AND SQFT1 < 900000) OR
       CNSUNIT > 200000000 THEN REGEDIT = 090;
    END;
%*-----------------------------;
%*RETAIL SALES CATEGORY WITH NO. OF FLOORS >= 3;
    IF NFLOOR1 >= 3 AND
      (BCLASS1 = '0900' OR
       BCLASS1 = '0930' OR
       BCLASS1 = '0931' OR
       BCLASS1 = '0933' OR
       BCLASS1 = '0934' OR
       BCLASS1 = '0935' OR
       BCLASS1 = '0936' OR
       BCLASS1 = '0937' OR
       BCLASS1 = '0938') THEN DO; REGCAT = 100;
    IF SQFT1 > 2700000 OR
       (CNSUNIT > 30000000 AND SQFT1 < 600000) THEN REGEDIT = 100;
    END;
%*-----------------------------;
%*RETAIL SALES CATEGORY WITH NO. OF FLOORS < 3;
    IF NFLOOR1 < 3 AND
      (BCLASS1 = '0900' OR
       BCLASS1 = '0930' OR
       BCLASS1 = '0931' OR
       BCLASS1 = '0933' OR
       BCLASS1 = '0934' OR
       BCLASS1 = '0935' OR
       BCLASS1 = '0937' OR
       BCLASS1 = '0938') THEN DO; REGCAT = 110;
```

```
      IF CNSUNIT > 10000000 OR
         SQFT1 > 600000 OR
         NWKER1 > 1000 OR
         (CNSUNIT > 100000000 AND SQFT1 < 1000000) THEN REGEDIT = 110;
      END;
%*----------------------------;
%*PERSONAL SERVICES BUILDINGS;
      IF BCLASS1 = '0950' OR
         BCLASS1 = '0951' OR
         BCLASS1 = '0953' OR
         BCLASS1 = '0954' OR
         BCLASS1 = '0955' OR
         BCLASS1 = '0956' THEN DO; REGCAT = 120;
      BLCL1 = 0;
      IF BCLASS1 = '0951' THEN BLCL1 = 1;
      END;
%*----------------------------;
%*MIXED RETAIL/WHOLESALE CATEGORY;
      IF BCLASS1 = '0940' OR
         BCLASS1 = '1050' OR
         BCLASS1 = '1051' OR
         BCLASS1 = '1052' OR
         BCLASS1 = '1053' OR
         BCLASS1 = '1054' THEN DO; REGCAT = 130;
      BLCL1 = 0;
      IF BCLASS1 = '1054' THEN BLCL1 = 1;
      IF CNSUNIT > 36000000 THEN REGEDIT = 130;
      END;
%*----------------------------;
%*AUTOMOBILE SALES/SERVICES;
      IF BCWM1 = '18' THEN DO; REGCAT = 140;
      BLCL1 = 0;
      IF BCLASS1 = '0936' THEN BLCL1 = 1;
      IF CNSUNIT > 10000000 THEN REGEDIT = 140;
      END;
%*----------------------------;
%*GENERAL OFFICE CATEGORY;
      IF BCLASS1 = '1100' THEN DO; REGCAT = 150;
      IF NFLOOR1 > 49 OR
         NWKER1 > 10000 OR
         SQFT1 > 1500000 OR
         CNSUNIT > 80000000 OR
         (CNSUNIT > 75000000 AND COSTX < 50000) OR
         (CNSUNIT > 49000000 AND SQFT1 < 150000) OR
         CNSUNIT = 89388 THEN REGEDIT = 150;
      END;
%*----------------------------;
%*PROFESSIONL OFFICE CATEGORY WITH 75 EMPLOYEES OR MORE;
      IF BCLASS1 = '1110' AND NWKER1 >= 75 THEN DO; REGCAT = 160;
      IF (CNSUNIT > 100000000 AND SQFT1 < 200000) OR
         CNSUNIT > 250000000 THEN REGEDIT = 160;
      END;
%*-----------------------------;
```

```
%*PROFESSIONL OFFICE CATEGORY WITH LESS THAN 75 EMPLOYEES;
   IF BCLASS1 = '1110' AND NWKER1 < 75 THEN REGCAT = 170;
%*-----------------------------;
%*FINANCIAL OFFICE CATEGORY;
   IF BCLASS1 = '1120' THEN DO; REGCAT = 180;
   IF CNSUNIT > 200000000 OR
      SQFT1 > 2500000 OR
      NWKER1 > 10000 OR
      (CNSUNIT < 10000000 AND SQFT1 > 1600000) OR
      (CNSUNIT < 10000000 AND NWKER1 > 6500) OR
      (CNSUNIT > 75000000 AND COSTX < 50000) OR
      (CNSUNIT > 60000000 AND SQFT1 < 400000) OR
      COSTX > 225000 THEN REGEDIT = 180;
   END;
%*-----------------------------;
%*MIXED USE OFFICE;
   IF BCLASS1 = '1020' OR
      BCLASS1 = '1021' OR
      BCLASS1 = '1022' OR
      BCLASS1 = '1023' OR
      BCLASS1 = '1024' OR
      BCLASS1 = '1130' OR
      BCLASS1 = '1131' OR
      BCLASS1 = '1132' THEN DO; REGCAT = 190;
   IF SQFT1 > 1600000 THEN REGEDIT = 190;
   END;
%*-----------------------------;
%*RESIDENTIAL BUILDINGS CATEGORY;
   IF BCLASS1 = '1300' OR
      BCLASS1 = '1310' OR
      BCLASS1 = '1311' OR
      BCLASS1 = '1312' OR
      BCLASS1 = '1320' OR
      BCLASS1 = '1321' OR
      BCLASS1 = '1322' OR
      BCLASS1 = '1323' OR
      BCLASS1 = '1324' OR
      BCLASS1 = '1325' OR
      BCLASS1 = '1330' THEN REGCAT = 200;
%*-----------------------------;
%*RESIDENTIAL MIXED USE CATEGORY;
   IF BCLASS1 = '1010' OR
      BCLASS1 = '1011' OR
      BCLASS1 = '1012' OR
      BCLASS1 = '1013' OR
      BCLASS1 = '1014' OR
      BCLASS1 = '1015' THEN REGCAT = 210;
%*-----------------------------;
%*COMMERCIAL LODGING (SHORT TERM) CATEGORY;
   IF BCLASS1 = '1410' OR
      BCLASS1 = '1411' OR
      BCLASS1 = '1412' OR
      BCLASS1 = '1413' OR
```

```
      BCLASS1 = '1414' OR
      BCLASS1 = '1415' OR
      BCLASS1 = '1416' OR
      BCLASS1 = '1417' THEN DO; REGCAT = 220;
   BLCL2 = 0;
   IF BCLASS1 = '1411' THEN BLCL2 = 1;
   END;
%*-----------------------------;
%*OTHER LODGING (LONG TERM) CATEGORY;
   IF BCLASS1 = '1400' OR
      BCLASS1 = '1420' OR
      BCLASS1 = '1421' OR
      BCLASS1 = '1422' OR
      BCLASS1 = '1423' OR
      BCLASS1 = '1424' OR
      BCLASS1 = '1425' OR
      BCLASS1 = '1426' THEN DO; REGCAT = 230;
   BLCL1 = 0;
   IF BCLASS1 = '1400' THEN BLCL1 = 1;
   END;
%*-----------------------------;
%*REFRIGERATED WAREHOUSES AND OTHER STORAGE;
   IF (('1040' <= BCLASS1 <= '1044') OR
      ('1500' <= BCLASS1 <= '1590')) AND
     (BCLASS1 NE '1520') THEN DO; REGCAT = 240;
   IF CNSUNIT > 40000000 OR
      SQFT1 > 600000 OR
      NWKER1 > 1600 OR
      (CNSUNIT > 25000000 AND SQFT1 < 180000) THEN REGEDIT = 240;
   END;
%*-----------------------------;
%*NONREFRAGERATED WAREHOUSES CATEGORY;
   IF BCLASS1 = '1520' THEN DO; REGCAT = 250;
   IF CNSUNIT > 500000000 OR NWKER1 > 2000 THEN REGEDIT = 250;
   END;
%*-----------------------------;
%*OTHER BUILDINGS CATEGORY;
   IF BCWM1 = '01' OR
      ('0800' <= BCLASS1 < '0900') OR
      BCLASS1 = '1000' OR
      BCLASS1 = '1050' OR
      ('1200' <= BCLASS1 < '1300') OR
      BCLASS1 >= '1600' THEN DO; REGCAT = 250;
      BLCL2 = 0;
      IF BCLASS1 = '1250' THEN BLCL2 = 1;
      IF (BCLASS1 < '1500' AND CNSUNIT > 500000000) OR
         (BCLASS1 = '1640' AND CNSUNIT > 500000000) OR
         (BCLASS1 > '1599' AND CNSUNIT > 17500000 AND
          COSTX < 30000) OR
         (BCLASS1 = '1660' AND CNSUNIT > 500000000) THEN
          REGEDIT = 260;
   END;
%*-----------------------------;
```

```
%MEND REGCAT;
*-----------------------------------------------------------------;
* THE MACRO CALL " %REGCAT" DEFINES THE 26 REGRESSION CATEGORIES. ;
* IT ALSO DEFINES THE OUTLIER EDITS PROVIDED BY EIA.              ;
%REGCAT
*-----------------------------------------------------------------;
*-------------------------------------------------------------;
*    CREATE WORKING INDEPENDENT VARIABLES FOR NONLINEAR        ;
*    REGRESSION MODELING USE.                                  ;
*-------------------------------------------------------------;
BANDSA = SQRT (SQFT1*NFLOOR1);
IF (SQFT1 GE 0 AND NFLOOR1 GT 0) THEN PRMTR = SQRT (SQFT1/NFLOOR1);
ROOFSA = (SQFT1/NFLOOR1);
GLASS = (1 - .25 * (0.25 * (5-PCTGLA) - 0.125));
* CREATE INDICATOR VARIABLES BASED ON FREE STANDING OR YEAR

IND1 = 0; IND2 = 0; IND3 = 0; IND4 = 0;
IF FRESTA1 = '1' AND YRCONC1 < '05' THEN IND1 = 1;
  ELSE IF FRESTA1 = '1' AND '05' <= YRCONC1 <= '07' THEN IND2 = 1;
  ELSE IF FRESTA1 = '2' AND YRCONC1 < '05' THEN IND3 = 1;
  ELSE IF FRESTA1 = '2' AND '05' <= YRCONC1 <= '07' THEN IND4 = 1;
  ELSE DO; IND1 = .; IND2 = .; IND3 = .; IND4 = .; END;
BANDSA1 = IND1 * BANDSA;
BANDSA2 = IND2 * BANDSA;
BANDSA3 = IND3 * BANDSA;
BANDSA4 = IND4 * BANDSA;
ROOFSA1 = IND1 * ROOFSA;
ROOFSA2 = IND2 * ROOFSA;
ROOFSA3 = IND3 * ROOFSA;
ROOFSA4 = IND4 * ROOFSA;
HRSPERWK = AVGNHR1/168;
SFHTPWK = SFHEAT * HRSPERWK;
COOKWK = ENDUSE6 * NWKER1;
MANUWKSF = ENDUSE5 * NWKER1 * SFHEAT;
EFLH = 0.5 * CDD651 + 300;
WKWK = NWKER1 * HRSPERWK;
IF CNSUNIT GT 0 THEN LNCNSP = LOG(CNSUNIT);
*-----------------------------------------------------------------;
LABEL A = FIRST THREE SOURCES
      B1 = FIRST SOURCE OF ENERGY
      B2 = SECOND SOURCE OF ENERGY
      B3 = THIRD SOURCE OF ENERGY
      B4 = FOURTH SOURCE OF ENERGY
      B5 = FIFTH SOURCE OF ENERGY
      B6 = SIXTH SOURCE OF ENERGY
      B7 = SEVENTH SOURCE OF ENERGY
      B8 = EIGHTH SOURCE OF ENERGY
      B9 = NINETH SOURCE OF ENERGY
      BTU = NATURAL GAS CONSUMPTION IN BTU
      BTUCLASS = ZERO OR NONZERO BTU
      CNSUNIT = NATURAL GAS CONSUMPTION/PHYSICAL UNITS
      COOKX = NATURAL GAS USED FOR COOKING
      COOLX = NATURAL GAS USED FOR COOLING
      COSTX = COST OF NATURAL GAS CONSUMED
```

```
      CSTDX = BILL COVERAGE OF COSTS
      DAYCLASS = THE THREE CONSUMP. DAYS RANGES
      DAYS = BILL COVERAGE OF NATURAL GAS CONS.
      EMPCAT = NUMBER OF WORKERS CATEGORY
      ENDUSE1 = NATURAL GAS USED FOR SPACE HEATING
      ENDUSE2 = NATURAL GAS USED FOR COOLING
      ENDUSE3 = NATURAL GAS USED FOR WATER HEATING
      ENDUSE4 = NATURAL GAS USED FOR ELECTRICITY GEN.
      ENDUSE5 = NATURAL GAS USED FOR MANUFACTURING
      ENDUSE6 = NATURAL GAS USED FOR COOKING
      PCTGLA = PERCENT GLASS
      PCTRES = PERCRNT RESIDENTIAL
      PCTVAC = PERCENT VACANT
      REG1 = CENSUS REGION1
      REG2 = CENSUS REGION2
      REG3 = CENSUS REGION3
      REG4 = CENSUS REGION4
      SFCAT = SQUARE FOOTAGE CATEGORY
      SFCOOL = SQUARE FOOTAGE COOLED
      SFHEAT = SQUARE FOOTAGE HEATED
      SFRESI = SQUARE FOOTAGE RESIDENTIAL
      SFVAC = SQUARE FOOTAGE VACANT
      TC4 = OTHER COOLIN  SYSTEM TYPE
      TD7 = WALL OR FLOOR PANELS
      YRCAT = YEAR BUILT CATEGORY
      WTRZON1 = CLIMATZON NO.1
      WTRZON2 = CLIMATZON NO.2
      WTRZON3 = CLIMATZON NO.3
      WTRZON4 = CLIMATZON NO.4
      WTRZON5 = CLIMATZON NO.5
      IMPSFX = IMPUTED SQUARE FOOTAGE(EIA)
      IMPNWX = IMPUTED NUMBER OF WOKERS(EIA)
      ELCONS = ELECTRICITY CONSUMPTION/PHYSICAL UNITS
      ENDUSEZ1 = ELECTRICITY USED FOR SPACE HEATING
      ENDUSEZ2 = ELECTRICITY USED FOR COOLING
      ENDUSEZ3 = ELECTRICITY USED FOR WATER HEATING
      ENDUSEZ4 = ELECTRICITY USED FOR ELECTRICITY GEN.
      ENDUSEZ5 = ELECTRICITY USED FOR MANUFACTURING
      ENDUSEZ6 = ELECTRICITY USED FOR COOKING;
*-------------------------------------------------------------;
DROP HS24011--HS24241 JANA1--SUNHRC1 INTMO1--INTWT1 NTANKS1--BKLR1;
PROC CONTENTS DATA=WORKING.NATGAS;
```

APPENDIX E:

SOURCE STATEMENTS FOR THE CREATION OF WORKING.ELECT

```
//HT1UHJT JOB (6616,X10,2,,,,),
// 'HOW TSAO ** ORNL ',TIME=(1,30),CLASS=0
/*JOBPARM LINES=10
/*ROUTE  PRINT RMT030
// EXEC SAS,REGION=1024K,OPTIONS='MACRO DQUOTE MPRINT',TIME=(1,30)
//CONVERT DD DSN=CN6616.HT1.CONVERT.ELECT.DATA73,DISP=SHR
//SASLIB DD DSN=CN6616.HT1.NBECS79.SASLIB3,DISP=(OLD,KEEP)
//DISKOUT DD DSN=CN6616.HT1.IMPUTED.UPDATE,DISP=SHR
//WORKING DD DSN=CN6616.HT1.WORKING.ELECT.DATA109,DISP=(NEW,CATLG),
// UNIT=DASD,SPACE=(TRK,(800,400),RLSE)
//SAS.WORK DD UNIT=SYSDA,SPACE=(6160,(800,400),,,ROUND)
//SYSIN DD *
OPTIONS GEN=2;
DATA IMPUTE;
  SET DISKOUT.IMPUTE;
PROC SORT NODUP DATA = IMPUTE;
  BY BLDGID1;
DATA ELECT;
  SET CONVERT.ELECT;
PROC SORT DATA = ELECT;
  BY BLDGID1;
DATA WORKING.ELECT;
  MERGE ELECT(IN=F1) IMPUTE(IN=F2);
  BY BLDGID1;
  IF F1;
  IF TOTB = 1;
  A = B1 + B2 + B3;
*------------------------------------------------------------------------;
* A = 1 MEANS THAT THE FUEL IS ONE OF THE THREE PRIMARY ENERGY    ;
*       SOURCES                                                          ;
* BUILDINGS THAT REPORTED ELECTRICITY USE ON A SINGLE ENERGY    ;
* SOURCE FIELD WILL HAVE TOTB = 1, AND WILL BE INCLUDED IN THE    ;
* DATA BASE CREATED HERE.                                                ;
*------------------------------------------------------------------------;
* WE NOW CREATE VARIABLE "DAYCLASS"                                      ;
  IF 0<= DAYS < 31 THEN DAYCLASS = 1;
    ELSE IF 31 <= DAYS < 331 THEN DAYCLASS = 2;
    ELSE IF 331 <= DAYS <= 365 THEN DAYCLASS = 3;
    ELSE DAYCLASS = .;
*---------------------------------------------------------------------;
* CREATE VARIABLE "BTUCLASS"                                             ;
  IF BTU = 0 THEN BTUCLASS = 0;
    ELSE IF BTU > 0 THEN BTUCLASS = 1;
    ELSE BTUCLASS = .;
*---------------------------------------------------------------------;
* CONVERT SURVEY VARIABLE VALUES TO WORKING VARIABLE VALUES:        ;
  IF HEATP1 > 100 OR HEATP1 < 0 THEN HEATP1 = .;
  IF COOLP1 > 100 OR COOLP1 < 0 THEN COOLP1 = .;
  IF RESP1 > 100 OR RESP1 < 0 THEN RESP1 = .;
    IF RESU1 = '1' OR RESUSE1 = '1' THEN PCTRES = RESP1;
    ELSE IF RESU1 = '2' AND RESUSE1 = '2' THEN PCTRES = 0;
    ELSE PCTRES = .;
  IF VACP1 > 100 OR VACP1 < 0 THEN VACP1 = .;
    IF PORVAC1 = '1' THEN PCTVAC = VACP1;
    ELSE IF PORVAC1 = '2' THEN PCTVAC = 0;
    ELSE PCTVAC = .;
  IF GLASPC1 > 4 THEN GLASPC1 = .;
```

```
  IF AVGNHR1 > 168 OR AVGNHR1 < 0 THEN AVGNHR1 = .;
*-----------------------------------------------------------------;
* CREATE WORKING VARIABLES ENDUSE1 - ENDUSE6.        ;
  IF HEATX ='1' THEN ENDUSE1 = 1;
    ELSE IF HEATX ='2' THEN ENDUSE1 = 0;
    ELSE ENDUSE1 = .;
  IF COOLX ='1' THEN ENDUSE2 = 1;
    ELSE IF COOLX ='2' THEN ENDUSE2 = 0;
    ELSE ENDUSE2 = .;
  IF WATERX ='1' THEN ENDUSE3 = 1;
    ELSE IF WATERX ='2' THEN ENDUSE3 = 0;
    ELSE ENDUSE3 = .;
  IF GENERX ='1' THEN ENDUSE4 = 1;
    ELSE IF GENERX ='2' THEN ENDUSE4 = 0;
    ELSE ENDUSE4 = .;
  IF MANUFX ='1' THEN ENDUSE5 = 1;
    ELSE IF MANUFX ='2' THEN ENDUSE5 = 0;
    ELSE ENDUSE5 = .;
  IF COOKX ='1' THEN ENDUSE6 = 1;
    ELSE IF COOKX ='2' THEN ENDUSE6 = 0;
    ELSE ENDUSE6 = .;
*-----------------------------------------------------------------;
* CREATE VARIABLES SFHEAT, SFCOOL, AND SFVAC                       ;
  SFHEAT = ENDUSE1 * SQFT1 * (HEATP1/100);
  SFCOOL = ENDUSE2 * SQFT1 * (COOLP1/100);
  SFVAC = SQFT1 * (PCTVAC/100);
  SFRESI = SQFT1 * (PCTRES/100);
*-----------------------------------------------------------------;
* WE NOW CREATE WORKING VARIABLE "PCTGLA";
  IF GLASPC1 = '1' THEN PCTGLA = 1;
  IF GLASPC1 = '2' AND GLASP1 = '1' THEN PCTGLA = 2;
  IF GLASPC1 = '3' AND GLASP1 = '2' THEN PCTGLA = 3;
  IF GLASPC1 = '4' THEN PCTGLA = 4;
*-----------------------------------------------------------------;
* WE NOW CREATE WORKING VARIABLES "WTRZON1 - WTRZON5";
  IF CLIMAT1 = '1' THEN WTRZON1 = 1;
  ELSE IF CLIMAT1 NE . THEN WTRZON1 = 0;
  IF CLIMAT1 = '2' THEN WTRZON2 = 1;
  ELSE IF CLIMAT1 NE . THEN WTRZON2 = 0;
  IF CLIMAT1 = '3' THEN WTRZON3 = 1;
  ELSE IF CLIMAT1 NE . THEN WTRZON3 = 0;
  IF CLIMAT1 = '4' THEN WTRZON4 = 1;
  ELSE IF CLIMAT1 NE . THEN WTRZON4 = 0;
  IF CLIMAT1 = '6' THEN WTRZON5 = 1;
  ELSE IF CLIMAT1 NE . THEN WTRZON5 = 0;
*-----------------------------------------------------------------;
* WE NOW CREATE WORKING VARIABLES REG1 - REG4;
  IF REGION1 = '1' THEN REG1 = 1;
  ELSE REG1 = 0;
  IF REGION1 = '2' THEN REG2 = 1;
  ELSE REG2 = 0;
  IF REGION1 = '3' THEN REG3 = 1;
  ELSE REG3 = 0;
```

```
  IF REGION1 = '4' THEN REG4 = 1;
  ELSE REG4 = 0;
*-------------------------------------------------------------;
* WE NOW CREATE WORKING VARIABLES YRCAT,SFCAT AND EMPCAT.;
      YRCAT = 0; SFCAT = 0; EMPCAT = 0;
      YRCAT = YRCONC1;
      IF YRCAT = 98 OR YRCAT = 99 THEN YRCAT = .;
      SFCAT = SQFTC1;
      IF SFCAT = 98 OR SFCAT = 99 THEN SFCAT = .;
      EMPCAT = NWKERC1;
      IF EMPCAT = 98 OR EMPCAT = 99 THEN EMPCAT = .;
*-------------------------------------------------------------;
* WE NOW CREATE WORKING VARIABLES TD7 AND TC4;
      TD7 = 0; TC4 = 0;
      IF WOFP1 = '1' THEN TD7 = 1;
      IF CST1 = '4' THEN TC4 = 1;
*----------------------------------------------------------------;
%MACRO ELECAT;
      CLASS = BCLASS1;
      REGCAT = 002;
      IF BLDGID1 = 2756 THEN SQFT1 = 759;
      IF BLDGID1 = 4158 THEN SQFT1 = 2074;
      IF BLDGID1 = 4158 THEN SFCAT = 2;
      IF BLDGID1 = 1036 OR
         BLDGID1 = 2909 OR
         BLDGID1 = 3017 OR
         BLDGID1 = 3291 OR
         BLDGID1 = 163 OR
         BLDGID1 = 1766 OR
         BLDGID1 = 1772 OR
         BLDGID1 = 1773 OR
         BLDGID1 = 1788 OR
         BLDGID1 = 2334 OR
         BLDGID1 = 2338 OR
         BLDGID1 = 4913 OR
         BLDGID1 = 4064 OR
         BLDGID1 = 433 OR
         BLDGID1 = 3693 THEN REGEDIT = 002;
 %*-----------------------------------------------------------;
 %* THIS MACRO PROGRAM DEFINES THE 18 REGRESSION CATEGORIES FOR ;
 %* ELECTRICITY USE BUILDINGS. THE VARIABLE REGCAT IS A GROUP ID;
 %* WHICH ASSIGNS A BUILDING IN THE SAMPLE TO ONE OF THE 18     ;
 %* BUILDING GROUPS. THE VARIABLE REGEDIT IDENTIFIES OUTLIERS OF;
 %* THE CORRESPONDING REGRESSION CATEGORY.                      ;
 %* EXAMPLE: SUPPOSE A BUILDING IS SELECTED FROM REGCAT = 060.  ;
 %*          IF ITS VALUE OF REGEDIT = 300 THEN THIS BUILDING   ;
 %*             IS AN OUTLIER WHICH IS TO BE DELETED FROM       ;
 %*             REGRESSION ANALYSIS.                            ;
 %*          IF ITS VALUE OF REGEDIT = 002 THEN THIS BUILDING   ;
 %*             IS TO BE DELETED FROM THE MASTER DATA SET.      ;
 %*          IF ITS VALUE OF REGEDIT = . THEN THIS BUILDING IS  ;
 %*             NOT AN IDENTIFIED OUTLIER.                      ;
 %*-----------------------------------------------------------;
```

```
  %* ASSEMBLY BUILDINGS CATEGORY;
     IF BCWM1 = '02' THEN DO; REGCAT = 270;
     BLCL2 = 0;
     IF BCLASS1 = 'C251' THEN BLCL2 = 1;
     IF (CNSUNIT > 50000000 AND NWKER1 < 400) OR
        (CNSUNIT < 10000000 AND NWKER1 > 3800) THEN REGEDIT = 270;
     END;
*------------------------------;
  %* EDUCATIN BUILDINGS CATEGORY;
     IF BCWM1 = '03' THEN DO; REGCAT = 280;
     BLCL1 = 0;
     IF BCLASS1 = '0340' THEN BLCL1 = 1;
     IF CNSUNIT > 50000000 THEN REGEDIT = 280;
     END;
*------------------------------;
  %* FOOD SALES CATGORY;
     IF BCWM1 = '04' OR
        BCLASS1 = '1030' OR
        BCLASS1 = '1031' OR
        BCLASS1 = '1032' OR
        BCLASS1 = '1033' OR
        BCLASS1 = '1034' THEN DO; REGCAT = 290;
     BLCL1 = 0;
     IF BCLASS1 = '0441' THEN BLCL1 = 1;
     END;
*------------------------------;
  %* HEALTH CARE CATEGORY;
     IF BCWM1 = '05' THEN REGCAT = 300;
*------------------------------;
  %* ASSEMBLY PLANTS CATEGORY;
     IF BCLASS1 = '0730' OR
        BCLASS1 = '0740' THEN DO; REGCAT = 310;
     IF CNSUNIT > 100000000 THEN REGEDIT = 310;
     END;
*------------------------------;
  %* RAW GOODS INDUSTRIAL CATEGORY;
     IF BCLASS1 = '0750' OR
        BCLASS1 = '0760' OR
        BCLASS1 = '0770' OR
        BCLASS1 = '0780' OR
        BCLASS1 = '0790' THEN DO; REGCAT = 320;
     IF (CNSUNIT > 100000000 AND COSTX > 4500000) OR
        CNSUNIT > 160000000 THEN REGEDIT = 320;
     END;
*------------------------------;
  %* OTHER INDUSTRIAL CATEGORY;
     IF BCLASS1 = '0700' OR
        BCLASS1 = '0710' OR
        BCLASS1 = '0720' THEN DO; REGCAT = 330;
     IF CNSUNIT > 160000000 OR NFLOOR1 > 15 THEN REGEDIT = 330;
     END;
*------------------------------;
```

```
  %* RETAIL SALES/SERVICES CATEGORY;
     IF BCLASS1 = '0910' OR
        BCLASS1 = '0920' OR
        BCLASS1 = '0900' OR
        BCLASS1 = '0930' OR
        BCLASS1 = '0931' OR
        BCLASS1 = '0933' OR
        BCLASS1 = '0934' OR
        BCLASS1 = '0935' OR
        BCLASS1 = '0937' OR
        BCLASS1 = '0938' OR
        BCLASS1 = '0950' OR
        BCLASS1 = '0951' OR
        BCLASS1 = '0953' OR
        BCLASS1 = '0954' OR
        BCLASS1 = '0955' OR
        BCLASS1 = '0956' OR
        BCLASS1 = '0940' OR
        BCLASS1 = '1025' OR
        BCLASS1 = '1050' OR
        BCLASS1 = '1051' OR
        BCLASS1 = '1052' OR
        BCLASS1 = '1053' OR
        BCLASS1 = '1054' THEN DO; REGCAT = 340;
     BLCL1 = 0;
     IF BCLASS1 = '0910' THEN BLCL1 = 1;
     IF (BCLASS1 = '0920' AND CNSUNIT > 50000000) OR
        (BCLASS1 = '0910' AND CNSUNIT > 50000000) THEN REGEDIT =

     END;
*------------------------------;
  %* AUTOMOBILE SALES/SERVICES;
     IF BCWM1 = '18' THEN DO; REGCAT = 350;
     IF CNSUNIT > 5000000 OR
        COSTX > 100000 THEN REGEDIT = 350;
     END;
*------------------------------;
  %* GENERAL OFFICE CATEGORY;
     IF BCLASS1 = '1100' THEN DO; REGCAT = 360;
     IF CNSUNIT < 5000000 AND SQFT1 > 1000000 THEN REGEDIT = 360;
     END;
*------------------------------;
  %* PROFESSIONL OFFICE CATEGORY;
     IF BCLASS1 = '1110' THEN DO; REGCAT = 370;
     IF CNSUNIT > 100000000 THEN REGEDIT = 370;
     END;
*-------------------------------;
  %* FINANCIAL OFFICE CATEGORY;
     IF BCLASS1 = '1120' THEN DO; REGCAT = 380;
     IF NWKER1 > 28000 OR CNSUNIT > 60000000 THEN REGEDIT = 380;
     END;
*-------------------------------;
```

```
%* MIXED USE OFFICE;
    IF BCLASS1 = '1016' OR
       BCLASS1 = '1020' OR
       BCLASS1 = '1021' OR
       BCLASS1 = '1022' OR
       BCLASS1 = '1023' OR
       BCLASS1 = '1024' OR
       BCLASS1 = '1026' OR
       BCLASS1 = '1130' OR
       BCLASS1 = '1131' OR
       BCLASS1 = '1132' THEN DO; REGCAT = 390;
    IF NWKER1 > 24000 THEN REGEDIT = 390;
    END;
*------------------------------;
%* RESIDENTIAL BUILDINGS CATEGORY;
    IF BCLASS1 = '1300' OR
       BCLASS1 = '1310' OR
       BCLASS1 = '1311' OR
       BCLASS1 = '1312' OR
       BCLASS1 = '1320' OR
       BCLASS1 = '1321' OR
       BCLASS1 = '1322' OR
       BCLASS1 = '1323' OR
       BCLASS1 = '1324' OR
       BCLASS1 = '1325' OR
       BCLASS1 = '1330' THEN DO; REGCAT = 400;
    BLCL1 = 0;
    IF BCLASS1 = '1310' THEN BLCL1 = 1;
    END;
*------------------------------;
%* RESIDENTIAL MIXED USE CATEGORY;
    IF BCLASS1 = '1010' OR
       BCLASS1 = '1011' OR
       BCLASS1 = '1012' OR
       BCLASS1 = '1013' OR
       BCLASS1 = '1014' OR
       BCLASS1 = '1015' THEN REGCAT = 410;
*------------------------------;
%* COMMERCIAL LODGING CATEGORY;
    IF BCLASS1 = '1410' OR
       BCLASS1 = '1411' OR
       BCLASS1 = '1412' OR
       BCLASS1 = '1413' OR
       BCLASS1 = '1414' OR
       BCLASS1 = '1415' OR
       BCLASS1 = '1416' OR
       BCLASS1 = '1417' OR
       BCLASS1 = '1400' OR
       BCLASS1 = '1420' OR
       BCLASS1 = '1421' OR
       BCLASS1 = '1422' OR
       BCLASS1 = '1423' OR
       BCLASS1 = '1424' OR
```

```
         BCLASS1 = '1425' OR
         BCLASS1 = '1426' THEN DO; REGCAT = 420;
      BLCL1 = 0;
      IF BCLASS1 = '1415' THEN BLCL1 = 1;
      END;
*-------------------------------;
  %* WAREHOUSES AND OTHER STORAGE CATEGORY;
      IF '1040' <= BCLASS1 <= '1044' OR
         '1500' <= BCLASS1 <= '1599' THEN DO; REGCAT = 430;
      BLCL1 = 0;
      IF BCLASS1 = '1530' THEN BLCL1 = 1;
      IF (CNSUNIT > 25000000 AND COSTX < 400000) OR
         (CNSUNIT < 5000000 AND COSTX > 1000000) OR
         (BCLASS1 < '1500' AND CNSUNIT < 5000000 AND NWKER1 > 1000)
         OR (BCLASS1 < '1500' AND SQFT1 > 1000000) OR
         (BCLASS1 < '1500' AND NWKER1 > 1200) OR
         (BCLASS1 < '1500' AND CNSUNIT > 20000000) OR
         (BCLASS1 = '1520' AND CNSUNIT < 5000000 AND NWKER1 > 1000)
         OR (BCLASS1 > '1499' AND BCLASS1 NE '1520' AND
         COSTX > 1500000 AND SQFT1 < 200000) OR (BCLASS1 > '1499' AND
         BCLASS1 NE '1520' AND COSTX > 1000000 AND NWKER1 < 500) AND
         REGEDIT = 430;
      END;
*-------------------------------;
  %* OTHER BUILDINGS CATEGORY;
      IF BCWM1 = '01' OR
         ('0800' <= BCLASS1 < '0900') OR
         BCLASS1 = '1000' OR
         BCLASS1 = '1060' OR
         ('1200' <= BCLASS1 < '1300') OR
         BCLASS1 >= '1600' THEN DO; REGCAT = 440;
      BLCL1 = 0;
      IF BCLASS1 = '0800' THEN BLCL1 = 1;
     BLCL2 = 0;
     IF BCLASS1 = '1260' THEN BLCL2 = 1;
      END;
*-------------------------------;
%MEND ELECAT;

*-----------------------------------------------------------------;
* THE MACRO CALL "%ELECAT" WILL DEFINE THE 18 REGRESSION CATEGORIES.;
* IT WILL ALSO IDENTIFY THE OUTLIERS IDENTIFIED BY EIA.             ;
* VARIABLE "CLASS" IS CREATED TO GIVE NUMERICAL CODES FOR "BCLASS1" ;
%ELECAT

*-----------------------------------------------------------------;
*-------------------------------------------------------;
*    CREATE WORKING INDEPENDENT VARIABLES FOR NONLINEAR    ;
*    REGRESSION MODELING USE.                              ;
*-------------------------------------------------------;
BANDSA = SQRT (SQFT1*NFLOOR1);
IF (SQFT1 GE 0 AND NFLOOR1 GT 0) THEN PRMTR = SQRT (SQFT1/NFLOOR1);
ROOFSA = (SQFT1/NFLOOR1);
GLASS = (1 - .25 * (0.25 * (5-PCTGLA) - 0.125));
```

```
* CREATE INDICATOR VARIABLES BASED ON FREE STANDING OR YEAR
CATEGORY;
IND1 = 0; IND2 = 0; IND3 = 0; IND4 = 0;
IF FRESTA1 = '1' AND YRCONC1 < '05' THEN IND1 = 1;
  ELSE IF FRESTA1 = '1' AND '05' <= YRCONC1 <= '07' THEN IND2 = 1;
  ELSE IF FRESTA1 = '2' AND YRCONC1 < '05' THEN IND3 = 1;
  ELSE IF FRESTA1 = '2' AND '05' <= YRCONC1 <= '07' THEN IND4 = 1;
  ELSE DO; IND1 = .; IND2 = .; IND3 = .; IND4 = .; END;
BANDSA1 = IND1 * BANDSA;
BANDSA2 = IND2 * BANDSA;
BANDSA3 = IND3 * BANDSA;
BANDSA4 = IND4 * BANDSA;
ROOFSA1 = IND1 * ROOFSA;
ROOFSA2 = IND2 * ROOFSA;
ROOFSA3 = IND3 * ROOFSA;
ROOFSA4 = IND4 * ROOFSA;
HRSPERWK = AVGNHR1/168;
SFHTPWK = SFHEAT * HRSPERWK;
COOKWK = ENDUSE6 * NWKER1;
MANUWKSF = ENDUSE5 * NWKER1 * SFHEAT;
EFLH = 0.5 * CDD651 + 300;
WKWK = NWKER1 * HRSPERWK;
IF CNSUNIT GT 0 THEN LNCNSP = LOG(CNSUNIT);
*--------------------------------------------------------------;
LABEL A = FIRST THREE SOURCES
      B1 = FIRST SOURCE OF ENERGY
      B2 = SECOND SOURCE OF ENERGY
      B3 = THIRD SOURCE OF ENERGY
      B4 = FOURTH SOURCE OF ENERGY
      B5 = FIFTH SOURCE OF ENERGY
      B6 = SIXTH SOURCE OF ENERGY
      B7 = SEVENTH SOURCE OF ENERGY
      B8 = EIGHTH SOURCE OF ENERGY
      B9 = NINETH SOURCE OF ENERGY
      BTU = FUEL CONSUMPTION IN BTU
      BTUCLASS = ZERO OR NONZERO BTU
      CNSUNIT = FUEL CONSUMPTION/PHYSICAL UNITS
      COOKX = FUEL USED FOR COOKING
      COOLX = FUEL USED FOR COOLING
      COSTX = COST OF FUEL CONSUMED
      CSTDX = BILL COVERAGE OF COSTS
      DAYCLASS = THE THREE CONSUMP. DAYS RANGES
      DAYS = BILL COVERAGE OF FUEL CONSUMED
      EMPCAT = NUMBER OF WORKERS CATEGORY
      ENDUSE1 = FUEL USED FOR SPACE HEATING
      ENDUSE2 = FUEL USED FOR COOLING
      ENDUSE3 = FUEL USED FOR WATER HEATING
      ENDUSE4 = FUEL USED FOR ELECTRICITY GEN.
      ENDUSE5 = FUEL USED FOR MANUFACTURING
      ENDUSE6 = FUEL USED FOR COOKING
      PCTGLA = PERCENT GLASS
      PCTRES = PERCRNT RESIDENTIAL
      PCTVAC = PERCENT VACANT
```

```
      REG1 = CENSUS REGION1
      REG2 = CENSUS REGION2
      REG3 = CENSUS REGION3
      REG4 = CENSUS REGION4
      SFCAT = SQUARE FOOTAGE CATEGORY
      SFCOOL = SQUARE FOOTAGE COOLED
      SFHEAT = SQUARE FOOTAGE HEATED
      SFRESI = SQUARE FOOTAGE RESIDENTIAL
      SFVAC = SQUARE FOOTAGE VACANT
      TC4 = OTHER COOLING SYSTEM TYPE
      TD7 = WALL OR FLOOR PANELS
      YRCAT = YEAR BUILT CATEGORY
      WTRZON1 = CLIMATZON NO.1
      WTRZON2 = CLIMATZON NO.2
      WTRZON3 = CLIMATZON NO.3
      WTRZON4 = CLIMATZON NO.4
      WTRZON5 = CLIMATZON NO.5
      IMPSFX = IMPUTED SQUARE FOOTAGE(EIA)

*------------------------------------------------------------------------;
DROP HS24011--HS24241 JANA1--SUNHRC1 INTMO1--INTWT1 NTANKS1--BKLR1;
PROC CONTENTS DATA=WORKING.ELECT;
//
```

APPENDIX F

COST/CONSUMPTION RATIO ANALYSIS

COST/CONSUMPTION RATIO ANALYSIS

An important variable needed in the regression model analysis is the annual fuel consumption (the response variable). Erroneous consumption values can bias or reduce the sensitivity of the resulting models.

It is difficult to determine the appropriate upper or lower limits for the raw fuel consumption values. However, if the cost data can be used in conjunction with the consumption data, then an approximate estimate of the fuel price can be calculated by the cost-consumption ratio C, where

C = Annual Total Cost of Fuel ÷ Annual Consumption of Fuel,

provided that the individual building has an annual consumption of fuel greater than 0 and that the building utility bill record covers 331 or more days for consumption and cost. With the values of C calculated for each eligible record, EIA identified outliers of the 1979 NBECS data based on the two range edits:

1. For natural gas use buildings, the acceptance range for C is
   0.001 (dollars/C.F.) <= C <= 0.01 (dollars/C.F.), and
2. For electricity use buildings, the acceptance range for C is
   0.001 (dollars/Kw) <= C <= 0.02 (dollars/Kw),

where C.F. = cubic foot.

Reasons for choosing these limits were not documented. As years pass by, these bounds need to be updated to remain effective. The following procedures are recommended to establish new bounds when checking the validity of consumption values. Outliers, once identified, should be checked against data processing error and respondent error as well as possible error on the corresponding cost variable before the outlier is to be deleted from the input data set. For example, sometimes a very low fuel consumption followed by a fixed service charge from the utility company can induce an unreasonably high value of C.

Procedures to establish new bounds are as follows:

1. Identify natural gas records with possible invalid consumption records. Two types of range checks are suggested.

   The first type specifies a cut-point, M, as a lower bound for the natural gas consumption values where

   M = (800 Btu/hr)x(1 C.F./1020 Btu)x(8 hrs/day)x(120 days).

   = 753 C.F.

   M estimates the total consumption of operating a pilot light in a gas heating furnance for eight hours a day and over a four-month period.

   The second type of range check is based on the two limits of C where a lower limit is

   CL = 1979 National average wellhead natural gas price

   = \$ 1.18 per 1000 C.F.

   = 0.00118 dollars per C.F.

   and an upper limit can be calculated from

   CU = 3.5 x (1979 National average residential natural gas price)

   = 3.5 x (\$ 2.98 per 1000 C.F.)

   = 3.5 x (0.00298 dollars per C.F.)

   = 0.01043 dollars per C.F.

Records with either CNSUNIT > M or C > CU or C < CL are candidates for invalid consumption value, and they should be checked for possible errors.

2. To identify electricity records with possible invalid consumption records. Two types of range checks are suggested.

The first type specifies a cut-point, M, as a lower bound for the electricity consumption values where

M = (100 Watts/hr)x(8 hr/day)x(365 days)

= 116.8 Kilowatts.

M estimates the total consumption of "operating a pair of four feet long flourescent tube" for eight hours a day for one year.

The second type of range check is based on the two limits of C where a lower limit is

CL = 1/3 x (1979 National average industrial electricity price)

= 1/3 x ($ 3.05 per Kilowatt)

= 0.00305 dollars per Kilowatt.

An upper limit can be calculated from

CU = 3 x (1979 National average commercial electricity price)

= 3 x ($ 4.68 per Kilowatt)

= 3 x (0.00468 dollars per Kilowatt)

= 0.014 dollars per Kilowatt.

Records with CNSUNIT > M or C > CU or C < CL are candidates for invalid electricity consumption records, and they should be checked for possible errors.

For the 1983 NBECS survey, these suggested edits can be adjusted by changing the parameters corresponding to the 1983 values. The two data sources for the energy price and energy conversion information are the Monthly Energy Review and the State Energy Price and Expenditure Report. Both reports are published by EIA.

APPENDIX G

COMPUTER PROGRAM TO REPRODUCE THE EIA ELECTRICITY CONSUMPTION MODELS

```
//HT1UHJT JOB (6616,X10,2,,,,),
// 'HOW TSAO ** ORNL ',TIME=(,20)
/*JOBPARM LINES=10
/*ROUTE  PRINT RMT030
// EXEC SAS,REGION=1024K,OPTIONS='MACRO DQUOTE MPRINT',TIME=(,20)
//CLASS DD DSN=CN6616.RL2.GENE.NBECS79.TAPE2.OAKR.SASTEST3,DISP=SHR
//WORKING DD DSN=CN6616.HT1.WORKING.ELECT.DATA109,DISP=SHR
//DISKOUT DD DSN=CN6616.HT1.IMPUTED.UPDATE,DISP=SHR
//SASLIB DD DSN=CN6616.HT1.NBECS79.SASLIB3,DISP=SHR
//SAS.WORK DD UNIT=SYSDA,SPACE=(6160,(800,400),,,ROUND)
//SYSIN DD *
%MACRO MASTER1;
  %* CREATE MASTER DATA SET BASED ON CRITERIA GIVEN BY EIA;
  %* FOR THE 1979 IMPUTATION;
    DATA MASTER;
      SET WORKING.ELECT;
      YRCAT = 0; SFCAT = 0; EMPCAT = 0;
      YRCAT = YRCONC1;
      IF YRCAT = 98 OR YRCAT = 99 THEN YRCAT = .;
      SFCAT = SQFTC1;
      IF SFCAT = 98 OR SFCAT = 99 THEN SFCAT = .;
      EMPCAT = NWKERC1;
      IF EMPCAT = 98 OR EMPCAT = 99 THEN EMPCAT = .;
      TD7 = 0; TC4 = 0;
      IF WOFP1 = 1 THEN TD7 = 1;
      IF CST1 = 4 THEN TC4 = 1;
      PCTGLA = 5 - PCTGLA;
      IF DAYS >= 330;
      IF CSTDX >= 330;
      IF BLDGID1 = 2756 THEN SQFT1 = 759;
      IF BLDGID1 = 4158 THEN SQFT1 = 2074;
      IF BLDGID1 = 4158 THEN SFCAT = 2;
      IF BLDGID1 = 1036 OR
         BLDGID1 = 2909 OR
         BLDGID1 = 3017 OR
         BLDGID1 = 3291 OR
         BLDGID1 = 163 OR
         BLDGID1 = 1766 OR
         BLDGID1 = 1772 OR
         BLDGID1 = 1773 OR
         BLDGID1 = 1788 OR
         BLDGID1 = 2334 OR
         BLDGID1 = 2338 OR
         BLDGID1 = 4913 OR
         BLDGID1 = 4064 OR
         BLDGID1 = 433 OR
         BLDGID1 = 3693 THEN DO;
         PUT _ALL_;
         DELETE;
```

```
      END;
%MEND MASTER1;
*----------------------------------------------------------------;
%MACRO MASTER2;
  %* CREATE MASTER DATA SET TO TEST ORNL REGRESSION RUNS;
  DATA MASTER;
    MERGE WORKING.ELECT(IN=F1) DISKOUT.IMPUTE(IN=F2);
    IF F1;
    IF A = 1;
  %* A = 1 MEANS THAT THE FUEL IS ONE OF THE THREE PRIMARY;
  %* ENERGY SOURCES;
    IF SQFTC1 NE '10';
    IF DAYCLASS = 3;
    IF IMPSFC1 = '1' THEN DELETE;
    IF IMPNWC1 = '1' THEN DELETE;
%MEND MASTER2;
*---------------------------------------------------------------;
*-----------------------------------------------------------------;
* CREATE REGRESSION INPUT DATA CATEGORIES FOR ELECTRICITY         ;
* USE BUILDINGS.                                                  ;
*-----------------------------------------------------------------;
%MACRO ASSEMBLY;
  %* ASSEMBLY BUILDINGS CATEGORY;
    DATA ASSEMBLY;
      SET MASTER;
      IF BCWM1 = '02';
      BLCL2 = 0;
      IF BCLASS1 = '0251' THEN BLCL2 = 1;
      IF (CNSUNIT > 50000000 AND NWKER1 < 400) OR
         (CNSUNIT < 10000000 AND NWKER1 > 3800) THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*------------------------------;
%MACRO EDUCATIN;
  %* EDUCATIN BUILDINGS CATEGORY;
    DATA EDUCATIN;
      SET MASTER;
      IF BCWM1 = '03';
      BLCL1 = 0;
      IF BCLASS1 = '0340' THEN BLCL1 = 1;
      IF CNSUNIT > 50000000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
%MEND;
*------------------------------;
%MACRO FOODSALE;
  %* FOOD SALES CATGORY;
    DATA FOODSALE;
      SET MASTER;
      IF BCWM1 = '04' OR
         BCLASS1 = '1030' OR
         BCLASS1 = '1031' OR
```

```
         BCLASS1 = '1032' OR
         BCLASS1 = '1033' OR
         BCLASS1 = '1034';
      BLCL1 = 0;
      IF BCLASS1 = '0441' THEN BLCL1 = 1;
  %MEND;
*------------------------------;
%MACRO HEALTH;
  %* HEALTH CARE CATEGORY;
    DATA HEALTH;
      SET MASTER;
      IF BCWM1 = '05';
  %MEND;
*------------------------------;
%MACRO ASSPLANT;
  %* ASSEMBLY PLANTS CATEGORY;
    DATA ASSPLANT;
      SET MASTER;
      IF BCLASS1 = '0730' OR
         BCLASS1 = '0740';
      IF CNSUNIT > 100000000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*------------------------------;
%MACRO RGINDUST;
  %* RAW GOODS INDUSTRIAL CATEGORY;
    DATA RGINDUST;
      SET MASTER;
      IF BCLASS1 = '0750' OR
         BCLASS1 = '0760' OR
         BCLASS1 = '0770' OR
         BCLASS1 = '0780' OR
         BCLASS1 = '0790';
      IF (CNSUNIT > 100000000 AND COSTX > 4500000) OR
         CNSUNIT > 160000000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*------------------------------;
%MACRO OTINDUST;
  %* OTHER INDUSTRIAL CATEGORY;
    DATA OTINDUST;
      SET MASTER;
      IF BCLASS1 = '0700' OR
         BCLASS1 = '0710' OR
         BCLASS1 = '0720';
      IF CNSUNIT > 160000000 OR NFLOOR1 > 15 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
*_%MEND;-----------------------;
```

```
%MACRO RETAIL;
  %* RETAIL SALES/SERVICES CATEGORY;
    DATA RETAIL;
      SET MASTER;
      IF BCLASS1 = '0910' OR
         BCLASS1 = '0920' OR
         BCLASS1 = '0900' OR
         BCLASS1 = '0930' OR
         BCLASS1 = '0931' OR
         BCLASS1 = '0933' OR
         BCLASS1 = '0934' OR
         BCLASS1 = '0935' OR
         BCLASS1 = '0937' OR
         BCLASS1 = '0938' OR
         BCLASS1 = '0950' OR
         BCLASS1 = '0951' OR
         BCLASS1 = '0953' OR
         BCLASS1 = '0954' OR
         BCLASS1 = '0955' OR
         BCLASS1 = '0956' OR
         BCLASS1 = '0940' OR
         BCLASS1 = '1025' OR
         BCLASS1 = '1050' OR
         BCLASS1 = '1051' OR
         BCLASS1 = '1052' OR
         BCLASS1 = '1053' OR
         BCLASS1 = '1054';
      BLCL1 = 0;
      IF BCLASS1 = '0910' THEN BLCL1 = 1;
      IF (BCLASS1 = '0920' AND CNSUNIT > 50000000) OR
         (BCLASS1 = '0910' AND CNSUNIT > 50000000) THEN DO;
         PUT _ALL_;
         DELETE;
        END;
%MEND;
*-----------------------------;
%MACRO AUTOSALE;
  %* AUTOMOBILE SALES/SERVICES;
    DATA AUTOSALE;
      SET MASTER;
      IF BCWM1 = '18';
      IF CNSUNIT > 5000000 OR
         COSTX > 100000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*-----------------------------;
%MACRO OFFICEGE;
  %* GENERAL OFFICE CATEGORY;
    DATA OFFICEGE;
      SET MASTER;
      IF BCLASS1 = '1100';
      IF CNSUNIT < 5000000 AND SQFT1 > 1000000 THEN DO;
         PUT _ALL_;
```

```
        DELETE;
       END;
  %MEND;
*------------------------------;
%MACRO OFFPROF;
  %* PROFESSIONL OFFICE CATEGORY;
    DATA OFFPROF;
      SET MASTER;
      IF BCLASS1 = '1110';
      IF CNSUNIT > 100000000 THEN DO;
         PUT _ALL_;
         DELETE;
       END;
  %MEND;
*--------------------------------;
%MACRO OFFICEFN;
  %* FINANCIAL OFFICE CATEGORY;
    DATA OFFICEFN;
      SET MASTER;
      IF BCLASS1 = '1120';
      IF NWKER1 > 28000 OR CNSUNIT > 60000000 THEN DO;
         PUT _ALL_;
         DELETE;
       END;
  %MEND;
*--------------------------------;
%MACRO OFFICEMX;
  %* MIXED USE OFFICE;
    DATA OFFICEMX;
      SET MASTER;
      IF BCLASS1 = '1016' OR
         BCLASS1 = '1020' OR
         BCLASS1 = '1021' OR
         BCLASS1 = '1022' OR
         BCLASS1 = '1023' OR
         BCLASS1 = '1024' OR
         BCLASS1 = '1026' OR
         BCLASS1 = '1130' OR
         BCLASS1 = '1131' OR
         BCLASS1 = '1132';
      IF NWKER1 > 24000 THEN DO;
         PUT _ALL_;
         DELETE;
       END;
  %MEND;
*--------------------------------;
%MACRO RESIDENT;
  %* RESIDENTIAL BUILDINGS CATEGORY;
    DATA RESIDENT;
      SET MASTER;
      IF BCLASS1 = '1300' OR
         BCLASS1 = '1310' OR
         BCLASS1 = '1311' OR
         BCLASS1 = '1312' OR
         BCLASS1 = '1320' OR
```

```
        BCLASS1 = '1321' OR
        BCLASS1 = '1322' OR
        BCLASS1 = '1323' OR
        BCLASS1 = '1324' OR
        BCLASS1 = '1325' OR
        BCLASS1 = '1330';
     BLCL1 = 0;
     IF BCLASS1 = '1310' THEN BLCL1 = 1;
  %MEND;
*--------------------------------;
%MACRO RESIDMX;
  %* RESIDENTIAL MIXED USE CATEGORY;
    DATA RESIDMX;
      SET MASTER;
      IF BCLASS1 = '1010' OR
         BCLASS1 = '1011' OR
         BCLASS1 = '1012' OR
         BCLASS1 = '1013' OR
         BCLASS1 = '1014' OR
         BCLASS1 = '1015';
  %MEND;
*--------------------------------;
%MACRO COMLODGS;
  %* COMMERCIAL LODGING CATEGORY;
    DATA COMLODGS;
      SET MASTER;
      IF BCLASS1 = '1410' OR
         BCLASS1 = '1411' OR
         BCLASS1 = '1412' OR
         BCLASS1 = '1413' OR
         BCLASS1 = '1414' OR
         BCLASS1 = '1415' OR
         BCLASS1 = '1416' OR
         BCLASS1 = '1417' OR
         BCLASS1 = '1400' OR
         BCLASS1 = '1420' OR
         BCLASS1 = '1421' OR
         BCLASS1 = '1422' OR
         BCLASS1 = '1423' OR
         BCLASS1 = '1424' OR
         BCLASS1 = '1425' OR
         BCLASS1 = '1426';
      BLCL1 = 0;
      IF BCLASS1 = '1415' THEN BLCL1 = 1;
  %MEND;
*--------------------------------;
%MACRO WARESTO;
  %* WAREHOUSES AND OTHER STORAGE CATEGORY;
    DATA WARESTO;
      SET MASTER;
      IF '1040' <= BCLASS1 <= '1044' OR
         '1500' <= BCLASS1 <= '1599';
      BLCL1 = 0;
      IF BCLASS1 = '1530' THEN BLCL1 = 1;
* NOTE: BUILDING CLASS '1530' IS NOT INCLUDED IN THIS CATEGORY;
```

```
      IF (CNSUNIT > 25000000 AND COSTX < 400000) OR
         (CNSUNIT < 5000000 AND COSTX > 1000000) OR
         (BCLASS1 < '1500' AND CNSUNIT < 5000000 AND NWKER1 > 1000)
         OR (BCLASS1 < '1500' AND SQFT1 > 1000000) OR
         (BCLASS1 < '1500' AND NWKER1 > 1200) OR
         (BCLASS1 < '1500' AND CNSUNIT > 20000000) OR
         (BCLASS1 = '1520' AND CNSUNIT < 5000000 AND NWKER1 > 1000)
         OR (BCLASS1 > '1499' AND BCLASS1 NE '1520' AND
         COSTX > 1500000 AND SQFT1 < 200000) OR (BCLASS1 > '1499' AND
         BCLASS1 NE '1520' AND COSTX > 1000000 AND NWKER1 < 500) THEN
         DO;
         PUT _ALL_;
         DELETE;
        END;
%MEND;
*------------------------------;
%MACRO OTHER;
  %* OTHER BUILDINGS CATEGORY;
    DATA OTHER;
      SET MASTER;
      IF BCWM1 = '01' OR
         ('0800' <= BCLASS1 < '0900') OR
         BCLASS1 = '1000' OR
         BCLASS1 = '1060' OR
         ('1200' <= BCLASS1 < '1300') OR
         BCLASS1 >= '1600';
      BLCL1 = 0;
      IF BCLASS1 = '0800' THEN BLCL1 = 1;
     BLCL2 = 0;
     IF BCLASS1 = '1260' THEN BLCL2 = 1;
  %MEND;
*------------------------------;
%MACRO REGRES(DEP=,IND=,INA=);
  PROC REG DATA = &INA;
    MODEL &DEP = &IND/VIF COVB CORRB COLLIN R DW;
    OUTPUT OUT = DATRES P = PRED R = RES;
    FORMAT P PRED R RES CLM CLI BEST12.;
  PROC UNIVARIATE DATA = DATRES NORMAL PLOT;
    VAR RES &IND;
  PROC PLOT DATA = DATRES;
    PLOT RES*(PRED &IND);
  PROC SORT DATA = &INA;
    BY &DEP;
  PROC PRINT DATA = &INA;
    VAR BLDGID1 &DEP &IND;
%MEND REGRES;
*-------------------------------------------------------------;
* EACH OF THE FOLLOWING SIX MACRO SUBROUTINES WILL GENERATE ;
* REGRESSION DIAGNOSTICS FOR SEVERAL OF THE REGRESSION      ;
* CATEGORIES DEFINED ABOVE.                                 ;
*-------------------------------------------------------------;
%MACRO ONE;
  %ASSEMBLY
  %REGRES(DEP=CNSUNIT,IND=NFLOOR1 SQFT1 NWKER1 SFCOOL ENDUSE5 BLCL2,
          INA=ASSEMBLY)
```

```
  %EDUCATIN
  %REGRES(DEP=CNSUNIT,IND=HEATDD1 YRCAT SQFT1 SFRESI NWKER1 SFHEAT
          SFCOOL AVGNHR1 WTRZON1 REG1 BLCL1,INA=EDUCATIN)
  %FOODSALE
  %REGRES(DEP=CNSUNIT,IND=HEATDD10 SFRESI EMPCAT SFCAT SFHEAT WTRZON4
          BLCL1,INA=FOODSALE)
* NOTE: ORIGINALLY WTRZON1 WAS USED;
%MEND ONE;
*-------------------------------------------------------------;
%MACRO TWO;
  %ASSPLANT
  %REGRES(DEP=CNSUNIT,IND=SQFT1 NWKER1 SFCOOL PCTGLA,INA=ASSPLANT)
  %RGINDUST
  %REGRES(DEP=CNSUNIT,IND=SQFT1 NWKER1 SFCOOL ENDUSE3 WTRZON4,
          INA=RGINDUST)
  %OTINDUST
  %REGRES(DEP=CNSUNIT,IND=SQFT1 SFCOOL AVGNHR1 ENDUSE2,
          INA=OTINDUST)
%MEND TWO;
*-------------------------------------------------------------;
%MACRO THREE;
  %RETAIL
  %REGRES(DEP=CNSUNIT,IND=NFLOOR1 SQFT1 NWKER1 SFHEAT ENDUSE6 WTRZON3
          BLCL1,INA=RETAIL)
  %AUTOSALE
  %REGRES(DEP=CNSUNIT,IND=EMPCAT SFHEAT SFCOOL ENDUSE4,INA=AUTOSALE)
  %OFFICEGE
  %REGRES(DEP=CNSUNIT,IND=SQFT1 COOLDD9 NWKER1 SFVAC AVGNHR1 WTRZON5 ,
          INA=OFFICEGE)
%MEND THREE;
*-------------------------------------------------------------;
%MACRO FOUR;
  %OFFPROF
  %REGRES(DEP=CNSUNIT,IND=SQFT1 NWKER1 SFHEAT SFCOOL SFVAC REG4,
          INA=OFFPROF)
  %OFFICEFN
  %REGRES(DEP=CNSUNIT,IND=NFLOOR1 SFRESI NWKER1 SFHEAT SFCOOL AVGNHR1
          WTRZON5,INA=OFFICEFN)
* NOTE: ORIGINALLY WTRZON6 WAS USED;
  %OFFICEMX
  %REGRES(DEP=CNSUNIT,IND=NFLOOR1 SQFT1 NWKER1,
          INA=OFFICEMX)
%MEND FOUR;
*-------------------------------------------------------------;
%MACRO FIVE;
  %RESIDENT
  %REGRES(DEP=CNSUNIT,IND=COOLDD3 YRCAT EMPCAT SFHEAT SFCOOL SFVAC
         PCTGLA BLCL1,INA=RESIDENT)
  %RESIDMX
  %REGRES(DEP=CNSUNIT,IND=NWKER1 SFHEAT SFCOOL SFVAC,
          INA=RESIDMX)
%MEND FIVE;
*-------------------------------------------------------------;
%MACRO SIX;
  %COMLODGS
```

```
  %REGRES(DEP=CNSUNIT,IND=HEATDD1 NFLOOR1 SQFT1 NWKER1 SFVAC PCTGLA
          WTRZON5 BLCL1,INA=COMLODGS)
  %WARESTO
  %REGRES(DEP=CNSUNIT,IND=HEATDD1 SQFT1 SFHEAT AVGNHR1 ENDUSE1 REG3
          BLCL1,INA=WARESTO)
  %OTHER
  %REGRES(DEP=CNSUNIT,IND=HEATDD10 NFLOOR1 SQFT1 NWKER1 SFHEAT SFCOOL
          BLCL1 BLCL2,INA=OTHER)
%MEND SIX;
*-------------------------------------------------------------;
*------------------------------------------------------------;
*            TEST RUNS                                        ;
*------------------------------------------------------------;
%MASTER1
HEATDD1 = ENDUSE1 * HDD601;
HEATDD2 = ENDUSE1 * HDD621;
HEATDD3 = ENDUSE1 * HDD641;
HEATDD4 = ENDUSE1 * HDD651;
HEATDD5 = ENDUSE1 * HDD661;
HEATDD6 = ENDUSE1 * HDD681;
HEATDD7 = ENDUSE1 * HDD701;
HEATDD8 = ENDUSE1 * HDD731;
HEATDD9 = ENDUSE1 * HDD751;
HEATDD10 = ENDUSE1 * HDD801;
COOLDD1 = ENDUSE2 * CDD601;
COOLDD2 = ENDUSE2 * CDD621;
COOLDD3 = ENDUSE2 * CDD641;
COOLDD4 = ENDUSE2 * CDD651;
COOLDD5 = ENDUSE2 * CDD661;
COOLDD6 = ENDUSE2 * CDD681;
COOLDD7 = ENDUSE2 * CDD701;
COOLDD8 = ENDUSE2 * CDD731;
COOLDD9 = ENDUSE2 * CDD751;
COOLDD10 = ENDUSE2 * CDD801;
  %FOODSALE
  %REGRES(DEP=CNSUNIT,IND=HEATDD10 SFRESI EMPCAT SFCAT SFHEAT WTRZON4
          BLCL1,INA=FOODSALE)
```

APPENDIX H

COMPUTER PROGRAM TO REPRODUCE THE EIA NATURAL GAS CONSUMPTION MODELS

```
//HT1UHJT JOB (6616,X10,2,,,,),
// 'HOW TSAO ** ORNL ',TIME=(1,)
/*JOBPARM LINES=10
/*ROUTE  PRINT RMT030
// EXEC SAS,REGION=1024K,OPTIONS='MACRO DQUOTE MPRINT',TIME=(1,)
//CLASS DD DSN=CN6616.RL2.GENE.NBECS79.TAPE2.OAKR.SASTEST3,DISP=SHR
//WORKING DD DSN=CN6616.HT1.WORKING.NATGAS.DATA109,DISP=SHR
//DISKOUT DD DSN=CN6616.HT1.IMPUTED.UPDATE,DISP=SHR
//SASLIB DD DSN=CN6616.HT1.NBECS79.SASLIB3,DISP=SHR
//SAS.WORK DD UNIT=SYSDA,SPACE=(6160,(800,400),,,ROUND)
//SYSIN DD *
%MACRO MASTER1;
  %* CREATE MASTER DATA SET BASED ON CRITERIA GIVEN BY EIA;
  %* FOR THE 1979 IMPUTATION;
    DATA MASTER;
      SET WORKING.NATGAS;
      YRCAT = 0; SFCAT = 0; EMPCAT = 0;
      YRCAT = YRCONC1;
      IF YRCAT = 98 OR YRCAT = 99 THEN YRCAT = .;
      SFCAT = SQFTC1;
      IF SFCAT = 98 OR SFCAT = 99 THEN SFCAT = .;
      EMPCAT = NWKERC1;
      IF EMPCAT = 98 OR EMPCAT = 99 THEN EMPCAT = .;
      TD7 = 0; TC4 = 0;
      IF WOFP1 = 1 THEN TD7 = 1;
      IF CST1 = 4 THEN TC4 = 1;
      IF DAYS >= 330;
      IF CSTDX >= 330;
      IF BLDGID1 = 2756 THEN SQFT1 = 759;
      IF BLDGID1 = 4158 THEN SQFT1 = 2074;
      IF BLDGID1 = 4158 THEN SFCAT = 2;
      IF BLDGID1 = 163 OR
         BLDGID1 = 2870 OR
         BLDGID1 = 4358 OR
         BLDGID1 = 5239 OR
         BLDGID1 = 5243 OR
         BLDGID1 = 7114 OR
         BLDGID1 = 3766 OR
         BLDGID1 = 4973 OR
         BLDGID1 = 5015 OR
         BLDGID1 = 5019 OR
         BLDGID1 = 5021 OR
         BLDGID1 = 2646 OR
         BLDGID1 = 1764 OR
         BLDGID1 = 3843 OR
         BLDGID1 = 5058 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
```

```
%MEND MASTER1;
*---------------------------------------------------------------;
%MACRO MASTER2;
  %* CREATE MASTER DATA SET TO TEST ORNL REGRESSION RUNS;
  DATA MASTER;
    MERGE WORKING.NATGAS(IN=F1) DISKOUT.IMPUTE(IN=F2);
    IF F1;
    IF A = 1;
  %* A = 1 MEANS THAT THE FUEL IS ONE OF THE THREE PRIMARY;
  %* ENERGY SOURCES;
    IF SQFTC1 NE '10';
    IF DAYCLASS = 3;
    IF IMPSFC1 = '1' THEN DELETE;
    IF IMPNWC1 = '1' THEN DELETE;
%MEND MASTER2;
*---------------------------------------------------------------;
*-----------------------------------------------------------------;
* CREATE REGRESSION INPUT DATA CATEGORIES FOR NATURAL GAS         ;
* USE BUILDINGS.                                                  ;
*-----------------------------------------------------------------;
%MACRO ASSEMBLY;
  %* ASSEMBLY BUILDINGS CATEGORY;
    DATA ASSEMBLY;
      SET MASTER;
      IF BCWM1 = '02';
*    NOTE: WE NEED TO DELETE THE 110 TH BUILDING ON THE DF DATA BASE;
  %MEND;
*------------------------------;
%MACRO EDUCATIN;
  %* EDUCATIN BUILDINGS CATEGORY;
    DATA EDUCATIN;
      SET MASTER;
      IF BCWM1 = '03';
      IF CNSUNIT < 10000000 AND SQFT1 > 600000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
%MEND;
*------------------------------;
%MACRO FOODSALE;
  %* FOOD SALES CATGORY;
    DATA FOODSALE;
      SET MASTER;
      IF BCWM1 = '04' OR
         BCLASS1 = '1030' OR
         BCLASS1 = '1031' OR
         BCLASS1 = '1032' OR
```

```
         BCLASS1 = '1033' OR
         BCLASS1 = '1034';
      IF CNSUNIT > 20000000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*-----------------------------;
%MACRO HEALTHHI;
  %* HEALTH CARE CATEGORY WITH SQ.FT. >= 350,000 ;
    DATA HEALTHHI;
      SET MASTER;
      IF BCWM1 = '05';
      IF SQFT1 >= 350000;
      IF CNSUNIT > 500000000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*-----------------------------;
%MACRO HEALTHLO;
  %* HEALTH CARE CATEGORY WITH SQ.FT. < 350,000 ;
    DATA HEALTHLO;
      SET MASTER;
      IF BCWM1 = '05';
      IF SQFT1 < 350000;
  %MEND;
*-----------------------------;
%MACRO ASSPLANT;
  %* ASSEMBLY PLANTS CATEGORY;
    DATA ASSPLANT;
      SET MASTER;
      IF BCLASS1 = '0730' OR
         BCLASS1 = '0740';
      IF (CNSUNIT < 100000000 AND COSTX > 600000) OR
         COSTX > 1000000 OR
         NWKER1 > 5000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*-----------------------------;
%MACRO RGINDUST;
  %* RAW GOODS INDUSTRIAL CATEGORY;
    DATA RGINDUST;
      SET MASTER;
      IF BCLASS1 = '0750' OR
         BCLASS1 = '0760' OR
         BCLASS1 = '0770' OR
```

```
          BCLASS1 = '0780' OR
          BCLASS1 = '0790';
       IF (CNSUNIT < 50000000 AND COSTX > 500000) OR
          (CNSUNIT < 50000000 AND SQFT1 > 1000000) OR
          NWKER1 > 5000 OR
          (CNSUNIT > 200000000 AND SQFT1 < 100000) THEN DO;
          PUT _ALL_;
          DELETE;
         END;
  %MEND;
*-------------------------------;
%MACRO OTINDUST;
  %* OTHER INDUSTRIAL CATEGORY;
    DATA OTINDUST;
      SET MASTER;
      IF BCLASS1 = '0700' OR
         BCLASS1 = '0710' OR
         BCLASS1 = '0720';
      IF (CNSUNIT > 700000000 AND SQFT1 < 500000) OR
         NWKER1 > 2400 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*-------------------------------;
%MACRO SHOPPING;
  %* SHOPPING CENTER CATEGORY;
    DATA SHOPPING;
      SET MASTER;
      IF BCLASS1 = '0910' OR
         BCLASS1 = '0920';
      IF (CNSUNIT > 50000000 AND SQFT1 < 900000) OR
         CNSUNIT > 200000000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*-------------------------------;
%MACRO RETAILHI;
  %* RETAIL SALES CATEGORY WITH NO. OF FLOORS >= 3;
    DATA RETAILHI;
      SET MASTER;
      IF NFLOOR1 >= 3;
      IF BCLASS1 = '0900' OR
         BCLASS1 = '0930' OR
         BCLASS1 = '0931' OR
         BCLASS1 = '0933' OR
         BCLASS1 = '0934' OR
         BCLASS1 = '0935' OR
```

```
          BCLASS1 = '0936' OR
          BCLASS1 = '0937' OR
          BCLASS1 = '0938';
       IF SQFT1 > 2700000 OR
          (CNSUNIT > 30000000 AND SQFT1 < 600000) THEN DO;
          PUT _ALL_;
          DELETE;
         END;
  %MEND;
*-----------------------------;
%MACRO RETAILLO;
  %* RETAIL SALES CATEGORY WITH NO. OF FLOORS < 3;
    DATA RETAILLO;
      SET MASTER;
      IF NFLOOR1 < 3;
      IF BCLASS1 = '0900' OR
         BCLASS1 = '0930' OR
         BCLASS1 = '0931' OR
         BCLASS1 = '0933' OR
         BCLASS1 = '0934' OR
         BCLASS1 = '0935' OR
         BCLASS1 = '0937' OR
         BCLASS1 = '0938';
      IF CNSUNIT > 10000000 OR
         SQFT1 > 600000 OR
         NWKER1 > 1000 OR
         (CNSUNIT > 100000000 AND SQFT1 < 1000000) THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*-----------------------------;
%MACRO PERSONAL;
  %* PERSONAL SERVICES BUILDINGS;
    DATA PERSONAL;
      SET MASTER;
      IF BCLASS1 = '0950' OR
         BCLASS1 = '0951' OR
         BCLASS1 = '0953' OR
         BCLASS1 = '0954' OR
         BCLASS1 = '0955' OR
         BCLASS1 = '0956';
      BLCL1 = 0;
      IF BCLASS1 = '0951' THEN BLCL1 = 1;
  %MEND;
*-----------------------------;
%MACRO MIXEDRW;
  %* MIXED RETAIL/WHOLESALE CATEGORY;
    DATA MIXEDRW;
      SET MASTER;
```

```
      IF BCLASS1 = '0940' OR
         BCLASS1 = '1050' OR
         BCLASS1 = '1051' OR
         BCLASS1 = '1052' OR
         BCLASS1 = '1053' OR
         BCLASS1 = '1054';
      BLCL1 = 0;
      IF BCLASS1 = '1054' THEN BLCL1 = 1;
      IF CNSUNIT > 36000000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*------------------------------;
%MACRO AUTOSALE;
  %* AUTOMOBILE SALES/SERVICES;
    DATA AUTOSALE;
      SET MASTER;
      IF BCWM1 = '18';
      BLCL1 = 0;
      IF BCLASS1 = '0936' THEN BLCL1 = 1;
      IF CNSUNIT > 10000000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*------------------------------;
%MACRO OFFICEGE;
  %* GENERAL OFFICE CATEGORY;
    DATA OFFICEGE;
      SET MASTER;
      IF BCLASS1 = '1100';
      IF NFLOOR1 > 49 OR
         NWKER1 > 10000 OR
         SQFT1 > 1500000 OR
         CNSUNIT > 80000000 OR
         (CNSUNIT > 75000000 AND COSTX < 50000) OR
         (CNSUNIT > 49000000 AND SQFT1 < 150000) OR
         CNSUNIT = 89388 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*------------------------------;
%MACRO OFFPFHI;
  %* PROFESSIONL OFFICE CATEGORY WITH 75 EMPLOYEES OR MORE;
    DATA OFFPFHI;
      SET MASTER;
      IF BCLASS1 = '1110';
```

```
      IF NWKER1 >= 75;
      IF (CNSUNIT > 100000000 AND SQFT1 < 200000) OR
         CNSUNIT > 250000000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*-------------------------------;
%MACRO OFFPFLO;
  %* PROFESSIONL OFFICE CATEGORY WITH LESS THAN 75 EMPLOYEES;
    DATA OFFPFLO;
      SET MASTER;
      IF BCLASS1 = '1110';
      IF NWKER1 < 75;
  %MEND;
*-------------------------------;
%MACRO OFFICEFN;
  %* FINANCIAL OFFICE CATEGORY;
    DATA OFFICEFN;
      SET MASTER;
       IF BCLASS1 = '1120';
      IF CNSUNIT > 200000000 OR
         SQFT1 > 2500000 OR
         NWKER1 > 10000 OR
         (CNSUNIT < 10000000 AND SQFT1 > 1600000) OR
         (CNSUNIT < 10000000 AND NWKER1 > 6500) OR
         (CNSUNIT > 75000000 AND COSTX < 50000) OR
         (CNSUNIT > 60000000 AND SQFT1 < 400000) OR
         COSTX > 225000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
  %MEND;
*-------------------------------;
%MACRO OFFICEMX;
  %* MIXED USE OFFICE;
    DATA OFFICEMX;
      SET MASTER;
      IF BCLASS1 = '1020' OR
         BCLASS1 = '1021' OR
         BCLASS1 = '1022' OR
         BCLASS1 = '1023' OR
         BCLASS1 = '1024' OR
         BCLASS1 = '1130' OR
         BCLASS1 = '1131' OR
         BCLASS1 = '1132';
      IF SQFT1 > 1600000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
```

```
  %MEND;
*-------------------------------;
%MACRO RESIDENT;
  %* RESIDENTIAL BUILDINGS CATEGORY;
    DATA RESIDENT;
      SET MASTER;
      IF BCLASS1 = '1300' OR
         BCLASS1 = '1310' OR
         BCLASS1 = '1311' OR
         BCLASS1 = '1312' OR
         BCLASS1 = '1320' OR
         BCLASS1 = '1321' OR
         BCLASS1 = '1322' OR
         BCLASS1 = '1323' OR
         BCLASS1 = '1324' OR
         BCLASS1 = '1325' OR
         BCLASS1 = '1330';
  %MEND;
*-------------------------------;
%MACRO RESIDMX;
  %* RESIDENTIAL MIXED USE CATEGORY;
    DATA RESIDMX;
      SET MASTER;
      IF BCLASS1 = '1010' OR
         BCLASS1 = '1011' OR
         BCLASS1 = '1012' OR
         BCLASS1 = '1013' OR
         BCLASS1 = '1014' OR
         BCLASS1 = '1015';
  %MEND;
*-------------------------------;
%MACRO COMLODGS;
  %* COMMERCIAL LODGING (SHORT TERM) CATEGORY;
    DATA COMLODGS;
      SET MASTER;
      IF BCLASS1 = '1410' OR
         BCLASS1 = '1411' OR
         BCLASS1 = '1412' OR
         BCLASS1 = '1413' OR
         BCLASS1 = '1414' OR
         BCLASS1 = '1415' OR
         BCLASS1 = '1416' OR
         BCLASS1 = '1417';
      BLCL2 = 0;
      IF BCLASS1 = '1411' THEN BLCL2 = 1;
  %MEND;
*-------------------------------;
%MACRO OTHLODGL;
  %* OTHER LODGING (LONG TERM) CATEGORY;
    DATA OTHLODGL;
```

```
      SET MASTER;
      IF BCLASS1 = '1400' OR
         BCLASS1 = '1420' OR
         BCLASS1 = '1421' OR
         BCLASS1 = '1422' OR
         BCLASS1 = '1423' OR
         BCLASS1 = '1424' OR
         BCLASS1 = '1425' OR
         BCLASS1 = '1426';
      BLCL1 = 0;
      IF BCLASS1 = '1400' THEN BLCL1 = 1;
  %MEND;
*-------------------------------;
%MACRO WAREREF;
  %* REFRIGERATED WAREHOUSES AND OTHER STORAGE;
    DATA WAREREF;
      SET MASTER;
      IF ('1040' <= BCLASS1 <= '1044') OR
         ('1500' <= BCLASS1 <= '1590');
      IF BCLASS1 NE '1520';
      IF CNSUNIT > 40000000 OR
         SQFT1 > 600000 OR
         NWKER1 > 1600 OR
         (CNSUNIT > 25000000 AND SQFT1 < 180000) THEN DO;
         PUT _ALL_;
         DELETE;
        END;
%MEND;
*-------------------------------;
%MACRO WARENREF;
  %* NONREFRAGERATED WAREHOUSES CATEGORY;
    DATA WARENREF;
      SET MASTER;
      IF BCLASS1 = '1520';
      IF CNSUNIT > 500000000 OR NWKER1 > 2000 THEN DO;
         PUT _ALL_;
         DELETE;
        END;
%MEND;
*-------------------------------;
%MACRO OTHER;
  %* OTHER BUILDINGS CATEGORY;
    DATA OTHER;
      SET MASTER;
      IF BCWM1 = '01' OR
         ('0800' <= BCLASS1 < '0900') OR
         BCLASS1 = '1000' OR
         BCLASS1 = '1060' OR
         ('1200' <= BCLASS1 < '1300') OR
         BCLASS1 >= '1600';
         BLCL2 = 0;
```

```
        IF BCLASS1 = '1250' THEN BLCL2 = 1;
        IF (BCLASS1 < '1500' AND CNSUNIT > 500000000) OR
           (BCLASS1 = '1640' AND CNSUNIT > 500000000) OR
           (BCLASS1 > '1599' AND CNSUNIT > 17500000 AND
            COSTX < 30000) OR
           (BCLASS1 = '1660' AND CNSUNIT > 500000000) THEN DO;
           PUT _ALL_;
           DELETE;
          END;
  %MEND;
*------------------------------;
%MACRO AGRI;
  %* AGRICULTURE BUILDINGS CATEGORY;
    DATA AGRI;
      SET MASTER;
      IF BCWM1 = '01';
%MEND;
*------------------------------;
%MACRO REGRES(DEP=,IND=,INA=);
  PROC REG DATA = &INA;
    MODEL &DEP = &IND/VIF COVB CORRB COLLIN R DW;
    OUTPUT OUT = DATRES P = PRED R = RES;
    FORMAT P PRED R RES CLM CLI BEST12.;
  PROC UNIVARIATE DATA = DATRES NORMAL PLOT;
    VAR RES &IND;
  PROC PLOT DATA = DATRES;
    PLOT RES*(PRED &IND);
  PROC SORT DATA = &INA;
    BY &DEP;
  PROC PRINT DATA = &INA;
    VAR BLDGID1 &DEP &IND;
%MEND REGRES;
*------------------------------------------------------------;
* EACH OF THE FOLLOWING EIGHT MACRO SUBROUTINES WILL GENERATE ;
* REGRESSION DIAGNOSTICS FOR FOUR OF THE 24 REGRESSION        ;
* CATEGORIES DEFINED ABOVE.                                   ;
*------------------------------------------------------------;
%MACRO ONE;
  %ASSEMBLY
  %REGRES(DEP=CNSUNIT,IND=COOLDD10 NFLOOR1 SFCOOL SFVAC ENDUSE5,
          INA=ASSEMBLY)
  %EDUCATIN
  %REGRES(DEP=CNSUNIT,IND=HEATDD1 COOLDD10 NWKER1 SFHEAT SFCOOL WTRZON4,
          INA=EDUCATIN)
  %FOODSALE
  %REGRES(DEP=CNSUNIT,IND=SFHEAT  AVGNHR1,INA=FOODSALE)
  %HEALTHHI
  %REGRES(DEP=CNSUNIT,IND=SFHEAT SFCOOL,INA=HEALTHHI)
```

```
%MEND ONE;
*-----------------------------------------------------------------;
%MACRO TWO;
  %HEALTHLO
  %REGRES(DEP=CNSUNIT,IND=COOLDD10 SFHEAT ENDUSE5 ENDUSE6 WTRZON1,
          INA=HEALTHLO)
* NOTE: ORIGINALLY WTRZON2 WAS USED;
  %ASSPLANT
  %REGRES(DEP=CNSUNIT,IND=SFHEAT AVGNHR1,INA=ASSPLANT)
  %RGINDUST
  %REGRES(DEP=CNSUNIT,IND=NWKER1 SFVAC ENDUSE6,INA=RGINDUST)
  %OTINDUST
  %REGRES(DEP=CNSUNIT,IND=SQFT1 NWKER1 SFHEAT AVGNHR1 ENDUSE2,
          INA=OTINDUST)
%MEND TWO;
*-------------------------------------------------------------------;
%MACRO THREE;
  %SHOPPING
  %REGRES(DEP=CNSUNIT,IND=SFCAT SFVAC ENDUSE5 WTRZON2,INA=SHOPPING)
* NOTE: ORIGINALLY WTRZON2 WAS USED;
  %RETAILHI
  %REGRES(DEP=CNSUNIT,IND=NWKER1 SFHEAT WTRZON5,INA=RETAILHI)
  %RETAILLO
  %REGRES(DEP=CNSUNIT,IND=COOLDD1 EMPCAT SFHEAT SFCOOL ENDUSE4 ENDUSE6
          WTRZON5 WTRZON2,INA=RETAILLO)
* NOTE: ORIGINALLY WTRZON2 WAS USED;
  %PERSONAL
  %REGRES(DEP=CNSUNIT,IND=SQFT1 EMPCAT SFHEAT ENDUSE5 REG3 BLCL1,
          INA=PERSONAL)
%MEND THREE;
*----------------------------------------------------------------------;
%MACRO FOUR;
  %MIXEDRW
  %REGRES(DEP=CNSUNIT,IND=EMPCAT SFCAT ENDUSE4 ENDUSE5 WTRZON4 REG3
          BLCL1,INA=MIXEDRW)
* NOTE: ORIGINALLY WTRZON4 WAS USED;
  %AUTOSALE
  %REGRES(DEP=CNSUNIT,IND=HEATDD1 EMPCAT SFCAT REG3 BLCL1,INA=AUTOSALE)
  %OFFICEGE
  %REGRES(DEP=CNSUNIT,IND=SFHEAT SFCOOL SFVAC AVGNHR1,
          INA=OFFICEGE)
%OFFPFHI
%REGRES(DEP=CNSUNIT,IND=HEATDD1 COOLDD10 NFLOOR1 SFHEAT SFCOOL,
        INA=OFFPFHI)
%OFFPFLO
%REGRES(DEP=CNSUNIT,IND=SFHEAT TD7 TC4 ENDUSE5,INA=OFFPFLO)
%MEND FOUR;
*-------------------------------------------------------------------;
%MACRO FIVE;
  %OFFICEFN
  %REGRES(DEP=CNSUNIT,IND=SQFT1 SFRESI NWKER1 SFHEAT SFCOOL SFVAC,
          INA=OFFICEFN)
  %OFFICEMX
  %REGRES(DEP=CNSUNIT,IND=NFLOOR1 NWKER1 SFHEAT ENDUSE1,
          INA=OFFICEMX)
```

```
  %RESIDENT
  %REGRES(DEP=CNSUNIT,IND=NFLOOR1 SFRESI EMPCAT SFHEAT PCTGLA ENDUSE1
         ,INA=RESIDENT)
  %RESIDMX
  %REGRES(DEP=CNSUNIT,IND=NFLOOR1 SQFT1 SFHEAT NWKER1 SFRESI AVGNHR1
          ENDUSE4 WTRZON3,INA=RESIDMX)
%MEND FIVE;
*------------------------------------------------------------------;
%MACRO SIX;
  %COMLODGS
  %REGRES(DEP=CNSUNIT,IND=COOLDD9 SFCAT SFHEAT BLCL2,INA=COMLODGS)
  %OTHLODGL
  %REGRES(DEP=CNSUNIT,IND=YRCAT SFRESI NWKER1 SFHEAT SFCOOL ENDUSE2
          REG2 BLCL1,INA=OTHLODGL)
%WAREREF
%REGRES(DEP=CNSUNIT,IND=COOLDD6 NWKER1 SFHEAT SFCOOL ENDUSE5 REG2,
        INA=WAREREF)
%WARENREF
%REGRES(DEP=CNSUNIT,IND=NWKER1 SFCAT,INA=WARENREF)
  %OTHER
  %REGRES(DEP=CNSUNIT,IND=SFHEAT SFCOOL ENDUSE5 WTRZON4 BLCL2,
          INA=OTHER)
* NOTE: ORIGINALLY WTRZON4 WAS USED;
%MEND SIX;
*----------------------------------------------------------------;
*             TEST RUNS                                          ;
*----------------------------------------------------------------;
%MASTER1
HEATDD1 = ENDUSE1 * HDD601;
HEATDD2 = ENDUSE1 * HDD621;
HEATDD3 = ENDUSE1 * HDD641;
HEATDD4 = ENDUSE1 * HDD651;
HEATDD5 = ENDUSE1 * HDD661;
HEATDD6 = ENDUSE1 * HDD681;
HEATDD7 = ENDUSE1 * HDD701;
HEATDD8 = ENDUSE1 * HDD731;
HEATDD9 = ENDUSE1 * HDD751;
HEATDD10 = ENDUSE1 * HDD801;
COOLDD1 = ENDUSE2 * CDD601;
COOLDD2 = ENDUSE2 * CDD621;
COOLDD3 = ENDUSE2 * CDD641;
COOLDD4 = ENDUSE2 * CDD651;
COOLDD5 = ENDUSE2 * CDD661;
COOLDD6 = ENDUSE2 * CDD681;
COOLDD7 = ENDUSE2 * CDD701;
COOLDD8 = ENDUSE2 * CDD731;
COOLDD9 = ENDUSE2 * CDD751;
COOLDD10 = ENDUSE2 * CDD801;
%OTHER
  %REGRES(DEP=CNSUNIT,IND=SFHEAT SFCOOL ENDUSE5 WTRZON4 BLCL2,
          INA=OTHER)
```

```
%OFFPFLO
  IF BLDGID1 = '5628' OR BLDGID1 = '5574' OR BLDGID1 = '6586'
     THEN DELETE;
%REGRES(DEP=CNSUNIT,IND=SFHEAT TD7 TC4 ENDUSE5,INA=OFFPFLO)
  %RGINDUST
  IF BLDGID1 = '1982' OR BLDGID1 = '5539' THEN DELETE;
  %REGRES(DEP=CNSUNIT,IND=NWKER1 SFVAC ENDUSE6,INA=RGINDUST)
  %FOODSALE
  %REGRES(DEP=CNSUNIT,IND=SFHEAT  AVGNHR1,INA=FOODSALE)
  %HEALTHLO
  %REGRES(DEP=CNSUNIT,IND=COOLDD10 SFHEAT ENDUSE5 ENDUSE6 WTRZON1,
          INA=HEALTHLO)
```

APPENDIX J

SAS PROGRAM FOR FITTING MODEL (5.11) TO NATURAL GAS CATEGORIES

```
//SCNUSCN JOB (6616,X10,2,,,,),
// 'STAN CANTOR ** ORNL ',TIME=(2,40)
/*JOBPARM LINES=30
/*ROUTE PRINT RMT030
// EXEC SAS,REGION=1024K,OPTIONS='MACRO DQUOTE MPRINT',TIME=(2,40)
//SAS.WORK DD UNIT=SYSDA,SPACE=(6160,(800,400),,,ROUND)
//SASLIB DD DSN=CN6616.HT1.NBECS79.SASLIB3,DISP=SHR
//WORKING DD DSN=CN6616.HT1.WORKING.NATGAS.DATA109,DISP=SHR
//SYSIN DD *
%MACRO MASTER;
DATA MASTER;
  SET WORKING.NATGAS;
  IF A = 1;
IF DAYCLASS=1 OR DAYCLASS=3;
  IF DAYCLASS = 1 THEN LNCNSP=.;
  IF SQFT1 <1000000;
  IF NFLOOR1 < 50;
  IF IMPSFC1 NE '1';
  IF IMPNWC1 NE '1';
  IF IMPSF1 NE '1';
  IF IMPNW1 NE '1';
  IF IMPSFX NE '1';
  IF IMPNWX NE '1';
  IF IMPNOF1 NE '1';
  IF IMPPG1 NE '1';
  IF IMPPGC1 NE '1';
  IF IMPYRC1 NE '1';
  IF REGCAT NE REGEDIT;
  IF REGEDIT NE 001;
IF BLCOVX NE '2';
IF COOK1='1' THEN COOK1=1;
ELSE IF COOK1='9' THEN COOK1=.; ELSE COOK1=0;
IF MANUF1='1' THEN MANUF1=1;
ELSE IF MANUF1='9' THEN MANUF1=.; ELSE MANUF1=0;
COOKWK=COOK1*NWKER1;
MANUWKSF=MANUF1*NWKER1*SFHEAT;
GLASS1=.875+(5-PCTGLA)*.25;
IF REGCAT=40 THEN REGCAT=50;
IF REGCAT=130 THEN REGCAT=140;
IF REGCAT=90 THEN REGCAT=100;
IF REGCAT=200 THEN REGCAT=210;
IF REGCAT=220 THEN REGCAT=230;
%MEND MASTER;
%MACRO REG(CODE=);
 %MASTER
  IF REGCAT = &CODE;
%MEND REG;
%MACRO LINEAR;
A11=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*BANDSA1;
A12=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*ROOFSA1;
A21=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*BANDSA2;
A22=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*ROOFSA2;
```

```
A31=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*BANDSA3;
A32=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*ROOFSA3;
A41=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*BANDSA4;
A42=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*ROOFSA4;
INTH2=COOKWK*HDD651*ENDUSE1;
INTH3=MANUWKSF*HDD651*ENDUSE1;
INTC2=COOKWK*CDD651*ENDUSE2;
INTC3=MANUWKSF*CDD651*ENDUSE2;
GAMH=ENDUSE1*HDD651*SFHEAT;
DELH2=COOKWK*ENDUSE6;
DELH3=MANUWKSF*ENDUSE5;
DELH4=-ENDUSE1*HDD651*WKWK;
GAMC=ENDUSE2*CDD651*SFCOOL;
DELC=ENDUSE2*CDD651*WKWK;
IF LNCNSP=. THEN PREDCAT=1; ELSE PREDCAT=0;
KEEP A11--PREDCAT BLDGID1 REGCAT LNCNSP ENDUSE1-ENDUSE6 SQFT1 NFLOOR1
NWKER1 HDD651 CDD651;
PROC SORT; BY REGCAT PREDCAT;
PROC REG;
ID BLDGID1;
MODEL LNCNSP=
A11 A12 A21 A22 A31 A32 A41
GAMH
DELH2 DELH3 INTH2 INTH3
DELH4
GAMC DELC INTC2
ENDUSE3
ENDUSE4;
OUTPUT OUT=MASTER
PREDICTED=PRED
STDP=STDPRED
RESIDUAL=RESID;
BY REGCAT;
PROC PRINT DATA=MASTER;
VAR BLDGID1 LNCNSP RESID PRED STDPRED;
BY REGCAT;
PROC PLOT DATA=MASTER;
PLOT RESID*(PRED SQFT1 NWKER1 NFLOOR1 HDD651 CDD651)=BLDGID1;
BY REGCAT;
%MEND LINEAR;
%MASTER
%LINEAR
//
```

APPENDIX K

SAS PROGRAM FOR FITTING MODEL (5.11) TO ELECTRICITY CATEGORIES

```
//SCNUSCN JOB (6616,X10,2,,,,),
// 'STAN CANTOR ** ORNL ',TIME=(2,40)
/*JOBPARM LINES=30
/*ROUTE PRINT RMT030
// EXEC SAS,REGION=1024K,OPTIONS='MACRO DQUOTE MPRINT',TIME=(2,40)
//SAS.WORK DD UNIT=SYSDA,SPACE=(6160,(800,400),,,ROUND)
//SASLIB DD DSN=CN6616.HT1.NBECS79.SASLIB3,DISP=SHR
//WORKING DD DSN=CN6616.HT1.WORKING.ELECT.DATA109,DISP=SHR
//SYSIN DD *
%MACRO MASTER;
DATA MASTER;
  SET WORKING.ELECT;
  IF A = 1;
IF DAYCLASS=1 OR DAYCLASS=3;
  IF DAYCLASS = 1 THEN LNCNSP=.;
  IF CNSUNIT NE 0;
  IF SQFT1 <1000000;
  IF NFLOOR1 < 50;
  IF IMPSFC1 NE '1';
  IF IMPNWC1 NE '1';
  IF IMPSF1 NE '1';
  IF IMPNW1 NE '1';
  IF IMPSFX NE '1';
  IF IMPNWX NE '1';
  IF IMPNOF1 NE '1';
  IF IMPPG1 NE '1';
  IF IMPPGC1 NE '1';
  IF IMPYRC1 NE '1';
  IF REGCAT NE REGEDIT;
  IF REGEDIT NE 002;
IF BLCOVX NE '2';
IF COOK1='1' THEN COOK1=1;
ELSE IF COOK1='9' THEN COOK1=.; ELSE COOK1=0;
IF MANUF1='1' THEN MANUF1=1;
ELSE IF MANUF1='9' THEN MANUF1=.; ELSE MANUF1=0;
COOKWK=COOK1*NWKER1;
MANUWKSF=MANUF1*NWKER1*SFHEAT;
IF REGCAT=400 THEN REGCAT=410;
GLASS1=.875+(5-PCTGLA)*.25;
%MEND MASTER;
%MACRO REG(CODE=);
 %MASTER
  IF REGCAT = &CODE;
%MEND REG;
%MACRO LINEAR;
A11=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*BANDSA1;
A12=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*ROOFSA1;
A21=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*BANDSA2;
A22=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*ROOFSA2;
A31=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*BANDSA3;
A32=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*ROOFSA3;
A41=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*BANDSA4;
```

```
A42=(ENDUSE1*HDD651*GLASS+ENDUSE2*CDD651*GLASS1)*ROOFSA4;
INTH2=COOKWK*HDD651*ENDUSE1;
INTH3=MANUWKSF*HDD651*ENDUSE1;
INTC2=COOKWK*CDD651*ENDUSE2;
INTC3=MANUWKSF*CDD651*ENDUSE2;
GAMH=ENDUSE1*HDD651*SFHEAT;
DELH1=SFHTPWK;
DELH2=COOKWK*ENDUSE6;
DELH3=MANUWKSF*ENDUSE5;
DELH4=-ENDUSE1*HDD651*WKWK;
GAMC=ENDUSE2*CDD651*SFCOOL;
DELC=ENDUSE2*CDD651*WKWK;
IF LNCNSP=. THEN PREDCAT=1; ELSE PREDCAT=0;
KEEP A11--PREDCAT BLDGID1 REGCAT LNCNSP ENDUSE1-ENDUSE6 SQFT1 NFLOOR1
NWKER1 HDD651 CDD651 PRMTR;
PROC SORT; BY REGCAT PREDCAT;
PROC REG;
ID BLDGID1;
MODEL LNCNSP=
A11 A12 A21 A22 A31 A32 A41
GAMH DELH1 INTH2 INTH3
DELH2 DELH3
DELH4
GAMC DELC INTC2
PRMTR
ENDUSE3
ENDUSE4;
OUTPUT OUT=MASTER
PREDICTED=PRED
STDP=STDPRED
RESIDUAL=RESID;
BY REGCAT;
PROC PRINT DATA=MASTER;
VAR BLDGID1 LNCNSP RESID PRED STDPRED;
BY REGCAT;
PROC PLOT DATA=MASTER;
PLOT RESID*(PRED SQFT1 NWKER1 NFLOOR1 HDD651 CDD651)=BLDGID1;
BY REGCAT;
%MEND LINEAR;
%MASTER
%LINEAR
//
```

APPENDIX L

SAS PROGRAM TO COMPUTE DISTANCES OF VECTORS FROM THE CONVEX HULL OF A SET OF VECTORS

```
//SCNUSCN JOB (6616,X10,2,,,,),
// 'STAN CANTOR ** ORNL ',TIME=(1,30)
/*JOBPARM LINES=10
/*ROUTE  PRINT RMT030
// EXEC SAS,REGION=1024K,OPTIONS='MACRO DQUOTE MPRINT',TIME=(1,30)
//SASLIB DD DSN=CN6616.HT1.NBECS79.SASLIB3,DISP=(OLD,KEEP)
//WORKING DD DSN=CN6616.HT1.WORKING.NATGAS.DATA109,DISP=SHR
//SAS.WORK DD UNIT=SYSDA,SPACE=(6160,(800,400),,,ROUND)
//SYSIN DD *
 DATA X;
   SET WORKING.NATGAS;
   IF A = 1;
   IF DAYCLASS = 3;
   IF SQFT1 <1000000;
   IF NFLOOR1 < 50;
   IF IMPSFC1 NE '1';
   IF IMPNWC1 NE '1';
   IF IMPSF1 NE '1';
   IF IMPNW1 NE '1';
   IF IMPSFX NE '1';
   IF IMPNWX NE '1';
   IF IMPNOF1 NE '1';
   IF IMPPG1 NE '1';
   IF IMPPGC1 NE '1';
   IF IMPYRC1 NE '1';
   IF BLCOVX NE '2';
   IF REGCAT = 250;
 KEEP HDD651 CDD651 SQFT1 NFLOOR1 NWKER1;
PROC MATRIX FUZZ;
FETCH X DATA=X;
XSTAR=9 600000 200 8000 2000;
NSTAR=NROW(XSTAR);
N=NROW(X);
P=NCOL(X);
CENTER=J(1,N,1#/N)*X;
X=X-J(N,1,1)*CENTER;
XSTAR=XSTAR-CENTER;
SVD U DV V X;
Q=(N-1)#(U*U');
DO RS=1 TO NSTAR;
LSTAR=U*(DIAG(1#/DV))*V'*(XSTAR(RS,)');
L=(LSTAR<>0)><1;
LARGE_NO=P#1E05;
L=L#/SUM(L);
S=SQRT(2#L);
M=Q*(L-LSTAR);
DO ITER=1 TO 100;
DL=Q*(L-LSTAR)-M+J(N,1,LARGE_NO#(SUM(L)-1));
DS=M#S;
DM=(S#S#/2)-L;
QSTAR=Q+J(N,N,LARGE_NO);
YS=SOLVE(DIAG(M)+DIAG(S)*QSTAR*DIAG(S),DS+S#(DL+QSTAR*DM));
YL=S#YS-DM;
YM=QSTAR*(S#YS-DM)-DL;
L=L-YL;
```

```
S=S-YS;
M=M-YM;
SUML=SUM(L);
DISTANCE=SSQ(U'*(L-LSTAR));
OBJECTIV=DISTANCE+LARGE_NO#(SUML-1)#(SUML-1);
DSSQ=SSQ(DL)+SSQ(DS)+SSQ(DM);
PRINT ITER DSSQ OBJECTIV;
IF DSSQ LT 1E-8 THEN GO TO DISTANCE;
END;
DISTANCE: DISTANCE=SQRT(DISTANCE);
PRINT ITER DSSQ DISTANCE SUML;
PRINT L S M XSTAR;
END;
//
```

APPENDIX M

DISTRIBUTION OF IMPUTATION FLAG COUNTS

Table M-1. Distribution of imputation flag counts for natural gas use buildings by the 26 regression categories

REGCAT = 10

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	22
No	No	No	105
No	No	Yes	99

IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	2
No	No	No	206
No	No	Yes	18

REGCAT = 20

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	9
No	No	No	265
No	No	Yes	79

IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	7
No	No	No	294
No	No	Yes	52

REGCAT = 30

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	10
No	No	No	84
No	No	Yes	58

IMPNW1	IMPNWC1	IMPNWX	FREQ
No	Yes	Yes	1
No	No	No	145
No	No	Yes	6

REGCAT = 40

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	1
No	No	No	54
No	No	Yes	6

IMPNW1	IMPNWC1	IMPNWX	FREQ
No	No	No	52
No	No	Yes	9

(Continued)

Table M-1. Continued

REGCAT	IMPSF1	IMPSFC1	IMPSFX	FREQ
REGCAT = 50	Yes	Yes	No	5
	No	No	No	53
	No	No	Yes	11
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	1
	No	No	No	57
	No	No	Yes	11
REGCAT = 60	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	1
	No	No	No	122
	No	No	Yes	13
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	1
	No	No	No	131
	No	No	Yes	4
REGCAT = 70	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	4
	No	No	No	91
	No	No	Yes	18
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	No	No	No	106
	No	No	Yes	7
REGCAT = 80	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	2
	No	No	No	61
	No	No	Yes	15
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	1
	No	No	No	74
	No	No	Yes	3

(Continued)

Table M-1. Continued

REGCAT = 90	IMPSF1	IMPSFC1	IMPSFX	FREQ
	No	No	No	79
	No	No	Yes	11
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	1
	No	Yes	Yes	2
	No	No	No	55
	No	No	Yes	32
REGCAT = 100	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	2
	No	No	No	45
	No	No	Yes	25
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	No	No	No	67
	No	No	Yes	5
REGCAT = 110	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	2
	No	No	No	109
	No	No	Yes	38
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	1
	No	No	No	143
	No	No	Yes	5
REGCAT = 120	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	2
	No	No	No	70
	No	No	Yes	25
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	No	No	No	90
	No	No	Yes	7

(Continued)

Table M-1. Continued

REGCAT	IMPSF1	IMPSFC1	IMPSFX	FREQ
REGCAT = 130	Yes	Yes	No	1
	No	No	No	24
	No	No	Yes	15
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	No	No	No	36
	No	No	Yes	4
REGCAT = 140	**IMPSF1**	**IMPSFC1**	**IMPSFX**	**FREQ**
	Yes	Yes	No	1
	No	No	No	66
	No	No	Yes	38
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	No	No	No	105
REGCAT = 150	**IMPSF1**	**IMPSFC1**	**IMPSFX**	**FREQ**
	Yes	Yes	No	2
	No	No	No	64
	No	No	Yes	13
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	No	No	No	68
	No	No	Yes	11
REGCAT = 160	**IMPSF1**	**IMPSFC1**	**IMPSFX**	**FREQ**
	Yes	Yes	No	1
	No	No	No	117
	No	No	Yes	16
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	No	No	No	98
	No	No	Yes	36

(Continued)

Table M-1. Continued

REGCAT	IMPSF1	IMPSFC1	IMPSFX	FREQ
REGCAT = 170	Yes	Yes	No	4
	No	No	No	125
	No	No	Yes	35
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	No	No	No	148
	No	No	Yes	16
REGCAT = 180	**IMPSF1**	**IMPSFC1**	**IMPSFX**	**FREQ**
	Yes	Yes	No	2
	No	No	No	65
	No	No	Yes	23
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	No	No	No	78
	No	No	Yes	12
REGCAT = 190	**IMPSF1**	**IMPSFC1**	**IMPSFX**	**FREQ**
	Yes	Yes	No	5
	No	No	No	85
	No	No	Yes	25
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	Yes	Yes	No	2
	No	No	No	99
	No	No	Yes	14
REGCAT = 200	**IMPSF1**	**IMPSFC1**	**IMPSFX**	**FREQ**
	Yes	Yes	No	3
	No	No	No	14
	No	No	Yes	38
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	Yes	Yes	No	1
	No	No	No	45
	No	No	Yes	9

(Continued)

Table M-1. Continued

REGCAT = 210	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	13
	No	No	No	58
	No	No	Yes	65
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	Yes	Yes	No	2
	No	No	No	123
	No	No	Yes	11

REGCAT = 220	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	9
	No	No	No	37
	No	No	Yes	26
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	Yes	Yes	No	1
	No	No	No	68
	No	No	Yes	3

REGCAT = 230	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	2
	No	No	No	39
	No	No	Yes	13
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	No	No	No	49
	No	No	Yes	5

REGCAT = 240	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	1
	No	No	No	90
	No	No	Yes	14
	IMPNW1	**IMPNWC1**	**IMPNWX**	**FREQ**
	No	No	No	101
	No	No	Yes	4

(Continued)

Table M-1. Continued

REGCAT = 250	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	1
	No	No	No	104
	No	No	Yes	15
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	1
	No	No	No	115
	No	No	Yes	4
REGCAT = 260	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	5
	No	No	No	113
	No	No	Yes	31
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	No	No	No	131
	No	No	Yes	18

Table M-2. Distribution of imputation flag counts for electricity use buildings by the 18 regression categories

REGCAT = 270

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	27
No	No	No	168
No	No	Yes	141

IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	7
No	No	No	300
No	No	Yes	29

REGCAT = 280

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	13
No	No	No	340
No	No	Yes	95

IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	9
No	No	No	375
No	No	Yes	64

REGCAT = 290

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	17
No	No	No	132
No	No	Yes	83

IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	1
No	No	No	222
No	No	Yes	9

REGCAT = 300

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	4
No	No	No	112
No	No	Yes	17

IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	1
No	No	No	111
No	No	Yes	21

(Continued)

Table M-9. Continued

REGCAT = 310	IMPSF1	IMPSFC1	IMPSFX	FREQ
	No	No	No	150
	No	No	Yes	20
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	1
	No	No	No	162
	No	No	Yes	7
REGCAT = 320	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	5
	No	No	No	115
	No	No	Yes	22
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	2
	No	No	No	134
	No	No	Yes	6
REGCAT = 330	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	2
	No	No	No	97
	No	No	Yes	18
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	1
	No	No	No	111
	No	No	Yes	5
REGCAT = 340	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	8
	No	No	No	421
	No	No	Yes	153
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	2
	No	No	No	519
	No	No	Yes	61

(Continued)

Table M-2. Continued

REGCAT = 350

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	4
No	No	No	99
No	No	Yes	68

IMPNW1	IMPNWC1	IMPNWX	FREQ
No	No	No	168
No	No	Yes	3

REGCAT = 360

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	2
No	No	No	99
No	No	Yes	19

IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	1
No	No	No	97
No	No	Yes	22

REGCAT = 370

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	8
No	No	No	373
No	No	Yes	69

IMPNW1	IMPNWC1	IMPNWX	FREQ
Yes	Yes	No	1
No	No	No	366
No	No	Yes	83

REGCAT = 380

IMPSF1	IMPSFC1	IMPSFX	FREQ
Yes	Yes	No	2
No	No	No	114
No	No	Yes	37

IMPNW1	IMPNWC1	IMPNWX	FREQ
No	No	No	139
No	No	Yes	14

(Continued)

Table M-2. Continued

REGCAT	IMPSF1	IMPSFC1	IMPSFX	FREQ
REGCAT = 390	Yes	Yes	No	4
	No	No	No	110
	No	No	Yes	30
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	2
	No	No	No	125
	No	No	Yes	17
REGCAT = 400	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	9
	No	No	No	23
	No	No	Yes	50
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	3
	No	No	No	70
	No	No	Yes	9
REGCAT = 410	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	16
	No	No	No	77
	No	No	Yes	71
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	2
	No	No	No	154
	No	No	Yes	8
REGCAT = 420	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	12
	No	No	No	97
	No	No	Yes	45
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	1
	No	No	No	141
	No	No	Yes	12

(Continued)

Table M-2. Continued

REGCAT = 430	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	5
	No	No	No	293
	No	No	Yes	59
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	3
	No	No	No	341
	No	No	Yes	13

REGCAT = 440	IMPSF1	IMPSFC1	IMPSFX	FREQ
	Yes	Yes	No	11
	No	No	No	190
	No	No	Yes	66
	IMPNW1	IMPNWC1	IMPNWX	FREQ
	Yes	Yes	No	2
	Yes	No	No	2
	No	No	No	233
	No	No	Yes	30

END

DATE FILMED

12 / 12 / 90