RESEARCH REPORT SERIES
(Statistics #2008-09)

Research To Model Field Of Degree Information For College Graduates, Using The 2003 NSCG File With Linked Census 2000 Long-Form Data

Elizabeth Huang
Donald Malec
Lynn Weidman

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: September 30, 2008

Research To Model Field Of Degree Information For College Graduates, Using The 2003 NSCG File With Linked Census 2000 Long-Form Data[1]

Elizabeth Huang, Donald Malec, and Lynn Weidman

1. Background

The 2009 American Community Survey (ACS) is being considered as a sampling frame for the 2010 National Survey of College Graduates (NSCG). Although the current ACS questionnaire contains the same information as the census long-form, the 2009 ACS will, in addition, include a detailed question on field of degree. Field of degree (FOD) is an important subdomain for the survey, and its inclusion could greatly reduce screening rates for the 2010 NSCG.

In order to evaluate possible sample designs, the Demographic Statistical Methods Division (DSMD) has requested a "mock-up," person-level file consisting of the 2005/2006 ACS samples (i.e., before FOD is included in the survey) along with a reasonable value of FOD included on each record. This will be accomplished by modeling the relationship between FOD and census long-form variables based on the 2003 NSCG and applying this relationship to the ACS.

In order to model the relationship between FOD and the ACS responses, DSMD has created a file of the 2003 NSCG that has been matched to the 2000 Census long-form responses. Census data was used because the long-form was the actual sampling frame for the 2003 NSCG (ACS was not fully implemented for housing units until 2005). The field of degree information was collected as part of the 2003 NSCG and has been limited to degrees earned prior to April 1, 2000 for use in this research. Although this file could be used to develop sampling specifications by selectively estimating FOD crossed with selected characteristics from the long-form, specifying which domains to cross with field of degree is problematic because of the multi-objective nature of the design. For example, estimates are not only needed for three types of degree field (Science and Engineering, Science and Engineering Related, and Non-Science and Engineering) but are also crossed with the corresponding three-way category of employment sector (i.e., education, government or business). In addition, estimates are published for a number of other cross classifications based on highest degree, type of employment (public sector, private sector, etc), as well as basic estimates of the number of college graduates by demographic groups such as age, race/ethnicity and sex. Because of the number of relationships between FOD and the census long-form responses, the easiest approach is to create a person-level file in which any variety of cross-tabulations can be viewed.

Specifically, DSMD is considering imputing FOD onto the 2005/2006 ACS using a model developed from the 2003 NSCG/2000 Census long-form. Since the population of characteristics will surely have changed from 2000 to 2005/2006, using a model to impute FOD by population characteristics will, at least, take into account changes in population characteristics through time. DSMD has requested the assistance of SRD in researching the use of modeling techniques for imputing FOD.

---

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

Possible Methods

Of interest is the development of a model to impute FOD status on the entire 2005/2006 ACS files (and later ACS files), for the purpose of evaluating the impact of using FOD as a design variable for the NSCG. Upon review, a procedure using classification trees is appropriate for the problem. This procedure identifies partitions of the data which form relatively homogeneous groups, with respect to categorical dependent variables. Partitions of the data are made using accompanying independent variables (either discrete or continuous) and the optimal partition is identified using pre-specified measures of homogeneity. An important feature of this type of modeling is that the number and degree of interactions of the model (i.e., two-way, three-way interactions) are not predetermined, but instead are estimated using a computer-intensive, data-driven approach. The identification of higher-order interaction in modeling FOD seems important if the imputation of FOD to an entire record of personal ACS information is to be representative of the NSCG sampling frame. Other procedures such as Alternating Conditional Expectations (ACE) and projection pursuit regression are geared more towards continuous outcome variables. The procedure, Multivariate Adaptive Regression Splines (MARS), is an extension of classification trees in that lower level interactions can be removed and splines may be incorporated. However, software to do MARS is not currently available at the Bureau, and classification trees can still do the same predictions but at a cost of needing more parameters. Random Forests are another extension of classification trees. Instead of relying on the usual procedure of building one classification tree, with independent variables selected in one forward stepwise procedure, Random Forests allow many alternative classification trees constructed by randomly forcing independent variables out of consideration. Both procedures, Classification Trees and Random Forests, should be able to identify important interactions. Heuristically, the difference between Classification Trees and Random Forests seems to be similar to the difference between stepwise procedures and all possible subset selection procedures in linear model building. Since Classification Tree methodology is relatively mature (Random Forests are new), we will use Classification Trees. One other procedure, neural nets, may also be useful, but, at this point, learning about and developing neural nets for this project would be too time-consuming. The usefulness of this procedure for future related projects could be pursued.

In summary, the Classification Tree approach will be used to model the relationship between field of degree and the Census long-form variables. It is appropriate for modeling categorical data, it automatically determines the number and degree of interactions that may be present in the model, and it uses software that is available (through the statistical package called "R"). The remainder of the section describes how the classification tree modeling is implemented for this study, an overview of the classification tree procedure and, lastly, limitations of this modeling approach. The following sections provide details on the final choice of the model, conclusions and comparisons.

Implementation

The Bureau will include FOD as an open-ended question on the ACS beginning in 2009. For use in the NSCG, the Bureau will code the FOD responses into 142 FOD categories. However, the main objective of the NSCG is to categorize individuals by broader categories of FOD such as "Science and Engineering" (S&E), "Science and Engineering Related" (S&E-R), and "Neither

Science and Engineering nor Science and Engineering related" (non-S&E). Equally important is the non-response rate (NR) because a differential non-response rate can greatly affect survey costs if sample selection is also differential. The resulting modeling will aim at predicting and imputing values for all of the 142 categories based on classification tree methods. A model for predicting and imputing non-response will be formed independently, since no records containing the FOD for non-respondents are available to model. Throughout this report, an important distinction is made between the terms "prediction" and "imputation;" we use the term "prediction" to denote the general problem of filling in an unknown FOD value. We report on two specific, different ways of filling in these values. One, termed "modal prediction," uses the Classification Tree methodology of the most likely value (the mode) from the estimated distribution of FOD within a node of a final classification tree. The other, termed "imputation," draws new values from the estimated distributions of FOD. Since the goal of this project is to fill in the unknown values for FOD that match up with the estimated distribution of FOD within the nodes (not their most likely value), we use the "imputations" as our final predictions. The term "predictions" without the qualifier "modal" refers to the general problem of filling in the unknown FOD values.

Implementation of rpart to predict FOD

Classification tree modeling can be applied to any data where the dependent variable is categorical, such as FOD or non-response. Independent variables can consist of both continuous and categorical variables. Predictions will be made by using the rpart procedure in "R." Classification trees can be described as a type of discriminant analysis. As in discriminant analysis, independent variables are used to partition the population into groups with each group being as homogeneous as possible. In terms of FOD, groups of people sharing the same values (or range of values) of independent variables are formed so that each group contains predominately people in only one field of degree category. Classification Tree methods search for these groups by recursively finding variables and their values which divide the previous groups into more homogeneous groups.

Many classification tree methods, including rpart, pick a best fitting tree using cross-validation methods. As described in Breiman, Friedman, Olshen and Stone (1984), cross-validation avoids overfitting a model to data by fitting a model to only part of the data and then checking the model's predictive power on the remaining part of the data. The cross-validation samples are used to estimate a penalty on the size of the model, called the "complexity parameter." This penalty is then applied to the original tree to pick the final tree from the entire data set.

Cross-validation and model fitting are automatically accomplished within rpart, as is the selection of a final classification tree. What needs to be specified are the variables for possible inclusion in the model and what kind of prediction errors to guard against.

Candidate Variables

As a computer intensive procedure, rpart will take any set of variables and fit a data-determined non-linear model with data-determined non-linear interactions. It would be nice if one could include all the questionnaire outcomes from the long-form and have rpart "pick the best" model;

to some degree, this is possible. However, for variables with too many categories and also when too many potential variables are available, the convergence of the program becomes questionable. When a variable has many categories, rpart will look at all possible partitions of these categories into two groups. Also, as more variables are included, the number of possible interactions to evaluate for inclusion into the model increases exponentially. As part of our preliminary testing, the full-category occupation coded in the long-form along with about ten other variables was used to develop a regression tree. The program ran overnight on SRD's SGI computer and was terminated in the morning. This experience illustrates that, a smaller set of variables needs to be used in the model selection procedure if we want to complete the project in a timely manner.

Loss functions

In the model development, we will use the Gini index (the rpart default) to decide which groups should be split to make more homogeneous subgroups. The Gini index for a group is defined as $\sum_{i \neq j} p_i p_j$, or, equivalently, as $1-\sum_i (p_i)^2$ (Breiman, et al., p. 104), where, for the FOD model, i ranges over all the FOD groups, and p_i is the sample proportion of the group. (For the non-response model, i ranges over the two response categories.) A low Gini index indicates a good way to split the sample into homogeneous groups (it is zero if everyone is in the same FOD category). It is highest when all p_i's are equal.

Although the Gini index is used to decide which group of relatively homogeneous records are broken into subgroups, the misclassification rate is used to decide how large of a model to use. That is, the misclassification rate is used to decide how many (and what type of) subgroups to keep in the final model. For the model selection employed here, the misclassification rate is defined to be the proportion of records whose predicted FOD status is different from their observed FOD status. The misclassification rate is minimized using cross-validation methods (described above).

Both the Gini index and the misclassification rate can be modified by weighting heterogeneity and misclassification more heavily towards specific types of errors (e.g., misclassifying non-science and engineering FOD into another category may be more costly than misclassifying a science and engineering degree). Instead of choosing specific weights for specific categories of FOD, we use equal weights but investigate the predictive power of modeling categories of FOD collapsed into two categories at a time, and then forming models on the collapsed categories in a stepwise fashion.

Appendix 1 contains a simple example to illustrate how the rpart procedure is used to select a final model.

Dependent variables: Field of Degree

Besides predicting the 142 category FOD variable called FOD2 (see Appendix 2 for the definition of each of the FOD categories), DSMD has requested that we look at the misclassification rates for several different versions of collapsed FOD categories.

The broadest categories considered consist of how much the degree relates to Science and Engineering. Table 1.1 shows how the 142 detailed FOD categories are classified into these three categories, called CFOD2.

Table 1.1. Categorizing Field of Degree into 3 categories

CFOD2	Description	Detailed FOD categories
1	Science and Engineering	671, 673, 674, 676, 677, 841, 842, 843, 844, 845, 605, 606, 607, 608, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 680, 681, 871, 872, 873, 874, 875, 876, 877, 878, 879, 601, 620, 704, 771, 861, 891, 892, 893, 894, 895, 896, 897, 902, 921, 922, 923, 924, 925, 927, 928, 929, 930, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 733, 734, 735, 736, 737, 738, 739, 740, 741
2	Science and Engineering Related	781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791, 610, 652, 672, 675, 702, 706, 709, 712, 751, 752, 753, 754
0	Neither Science and Engineering nor Science and Engineering Related	602, 651, 653, 654, 655, 656, 657, 658, 659, 661, 662, 663, 682, 690, 701, 703, 705, 707, 708, 710, 711, 713, 760, 772, 800, 810, 820, 830, 850, 862, 901, 903, 910, 926, 941, 942, 943, 944, 995

The collapsing that was actually used further partitions the three broad science and engineering categories into 8 categories that were used in stratifying the NSCG cases for sample selection. These categories, indexed by FOD3, are defined below, in terms of the detailed 142 FOD categories.

Table 1.2. Categorizing Field of Degree into 8 nested categories

FOD3	Description	Detailed FOD categories
1	Computer and Mathematical Sciences	671, 673, 674, 676, 677, 841, 842, 843, 844, 845
2	Life and Related Sciences	605, 606, 607, 608, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 680, 681
3	Physical and Related Sciences	871, 872, 873, 874, 875, 876, 877, 878, 879
4	Social and Related Sciences	601, 620, 704, 771, 861, 891, 892, 893, 894, 895, 896, 897, 902, 921, 922, 923, 924, 925, 927, 928, 929, 930
5	Engineering	721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 733, 734, 735, 736, 737, 738, 739, 740, 741'
6	Science and Engineering Related Health	781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791
7	Science and Engineering Related non-Health	610, 652, 672, 675, 702, 706, 709, 712, 751, 752, 753, 754
8	Non-Science and Engineering	602, 651, 653, 654, 655, 656, 657, 658, 659, 661, 662, 663, 682, 690, 701, 703, 705, 707, 708, 710, 711, 713, 760, 772, 800, 810, 820, 830, 850, 862, 901, 903, 910, 926, 941, 942, 943, 944, 995

Note that FOD3 = 1-5 correspond to S&E degrees, FOD3 = 6-7 correspond to S&E-related degrees, and FOD3 = 8 corresponds to non-S&E degrees.

The 142 detailed FOD categories can also be broadly classified by more detailed subject matter and, again, by the three Science & Engineering categories: Science & Engineering, Science & Engineering Related, and Neither Science & Engineering nor Science & Engineering Related. Table 1.3 presents this cross classification using letters to represent subject matter versus italic,

bold, or plain text to represent science and engineering, science and engineering related, or non-science and engineering, respectively.

Table 1.3. Categorizing Field of Degree by Subject and Science& Engineering

FOD1	Description	Detailed FOD categories
A	Biological, Agricultural, Physical, Earth, or Other Natural Sciences	*605* , *606* , *607* , *608* , *631* , *632* , *633* , *634* , *635* , *636* , *637* , *638* , *639* , *640* , *641* , *642* , *680* , *681* , *871* , *872* , *873* , *874* , *875* , *876* , *877* , *878* , *879*
B	Health, Nursing, or Medical Fields	**781** , **782** , **783** , **784** , **785** , **786** , **787** , **788** , **789** , **790** , **791**
C	Engineering, Computer Sciences, or Mathematical Sciences	*671* , **672** , *673* , *674* , **675** , *676* , *677* , *721* , *722* , *723* , *724* , *725* , *726* , *727* , *728* , *729* , *730* , *731* , *732* , *733* , *734* , *735* , *736* , *737* , *738* , *739* , *740* , *741* , **751** , **752** , **753** , **754** , *841* , *842* , *843* , *844* , *845*
D	History, Arts, or Humanities	662 , 760 , 772 , 820 , 862 , 926 , 941 , 942 , 943 , 944
E	Psychology, Economics, or Other Social Sciences	*601* , *620* , *704* , *771* , *861* , *891* , *892* , *893* , *894* , *895* , *896* , *897* , *902* , *921* , *922* , *923* , *924* , *925* , *927* , *928* , *929* , *930*
F	Business or Management	602 , 651 , 653 , 654 , 655 , 656 , 657 , 658 , 659
G	Education or Education Administration	701 , **702** , 703 , 705 , **706** , 707 , 708 , **709** , 710 , 711 , **712** , 713
H	Some other major field	**610** , **652** , 661 , 663 , 682 , 690 , 800 , 810 , 830 , 850 , 901 , 903 , 910 , 995

Note that all degrees within categories A and E along with most in category C correspond to S&E degrees (*italicized* numbers in Table 1.3), while all degrees in category B along with a few in categories C, G, and H are S&E-related (**bolded** numbers in Table 1.3). Numbers that are not italicized or bolded correspond to non-S&E degrees.

Table 1.4 cross-classifies FOD1 with CFOD2 into 11 non-empty groups as follows.

Table 1.4. Field of Degree by Subject and Broad Science& Engineering Categories

	Description	S and E Classification		
		S&E (1)	S&ER (2)	NS&E (0)
A	Biological, Agricultural, Physical, Earth, or Other Natural Sciences	A1		
B	Health, Nursing, or Medical Fields		B2	
C	Engineering, Computer Sciences, or Mathematical Sciences	C1	C2	
D	History, Arts, or Humanities			D0
E	Psychology, Economics, or Other Social Sciences	E1		
F	Business or Management			F0
G	Education or Education Administration		G2	G0
H	Some other major field		H2	H0

Dependent variable: Non-response

Apart from FOD, the other dependent variable to be predicted is non-response. A separate model for non-response was developed to determine who is likely to not respond to the NSCG. Since there is no FOD information for non-response, the two models can only be constructed

independently of each other and their accuracy is based on the assumption that response status and FOD are independent of each other, conditional on the long-form covariates.

Independent Variables Used

The variables used to form the 2003 NSCG design strata (except AGE1927) were used as potential predictors in the model, as were the design strata for the 2000 long-form and the 2005 ACS (to avoid model misspecification), plus a variety of not-too-detailed long-form questionnaire responses. Single year of age and income were included as continuous variables. All other variables were included as factors.

2003 NSCG sampling stratification variables named on file AIAN, ASIAN, BLACK, DEMGRP03, GENDER03, HIDEG03, NEWCSAB, NEWDIS, NEWHISP, NEWRACE, NHPI, OCCUP03, OTHER, WHITE

Other variables as named on data file
qinctot - total income
pov - poverty status
qrel - relationship
qspanx - Spanish origin
qms - marital status
ESR – employment status recode
qpowccs - place of work, metro status
qcow - class of worker
disable – disability status
ESP - employment status of parents
xage – age in single years (long-form: agelong)
xqwk - wklyrhr – usual hours worked last year (long-form: wklyrhr)
qcit – citizenship (long-form: qcitizen)
qeng – English ability (long-form: qengabil)
Place of Birth (if outside of U.S.)
Year entry into U.S (coded zero if non-immigrant)
Age entry into U.S. (coded zero if non-immigrant)
Census Region
Census Division
PAOCF (presence and age of own children-female)
XPERN (person total earnings)
BSAM – long-form sampling strata
TRACTMOS - long-form tract size (used to form ACS strata)
GUMOS - long-form governmental unit size (used to form ACS strata)
ACS strata
Collapsed occupation code
Collapsed industry code

Appendix 3 lists the detailed coding for ACS strata, collapsed occupation code and collapsed industry code.

Limitations of the Study

Due to differences between the ACS and 2000 Census long-form design and to the targeted sampling used in the NSCG selection, we do not have a data set that matches up exactly with what will be available from the 2009 ACS. The major limitations to modeling are:

1) The relationship of FOD within housing unit is lost. NSCG samples persons from a list. The relationship of FOD among persons within housing unit cannot be modeled, since the NSCG will rarely (if ever) have collected FOD for more than one person per housing unit. When the models developed here are used to specify FOD for all college graduates in a housing unit, it must be remembered that the housing unit designation is meaningless.
2) Predictions of FOD are extrapolations (not interpolations). A model of FOD (as of 2000) between the 2003 NSCG and the 2000 long-form will be applied to the 2005-2006 ACS. Using these predictions assumes that persons that share the same independent variables in that model will behave the same over time, regardless of whether they were a long form or an ACS sample unit. In summary, the possible affects missed are:
    i) time changes in the relationship between FOD and the long form outcomes.
    ii) differences in response due to survey procedures (e.g., difference in non-response)
3) Predictions from a Census Long-form Model cannot account for ACS survey differences. There are operational differences between the ACS and the long-form, including mode of interview and non-response follow-up. These differences may differentially affect the non-response rate and the type of follow-up responses to the NSCG that cannot be predicted from the 2000 long-form/2003 NSCG data set.

Due to the strata constructed for the NSCG and resulting disproportionate sample selection within each stratum, the unweighted sample no longer reflects the long-form sample. To avoid modeling the NSCG instead of the long-form sample, we will include the NSCG design stratification variables as potential covariates in the model. Doing so will render the first-order selection probability conditionally independent from the design, enabling models to be applied to any set of sample data that is based on these same variables. We will also include the ACS stratification variables as potential model variables. The ACS strata are nested within the long-form strata so that using the ACS strata in the model will correct for possible design effect differences between the ACS and the long-form sample stratification.

Limitations of the Method

Most classification tree procedures, including the implementation in rpart, focus on providing classification rules for individual observations. For example, classification trees have been constructed on data sets to determine what characteristics may predict whether someone defaults on their loan. The resulting classification tree is then applied to new credit applicants and a prediction of their "credit risk" is made. As mentioned in the "implementation" section, above, the de facto prediction from rpart is "modal prediction"; i.e., using rpart, the predicted value for a new individual who has been classified into a terminal node of the tree, is based on the response that is the majority response of the records in the node (from the sample that was used to make the classification tree). However, for this project, we do not use modal prediction for FOD value; instead we impute the FOD using the empirical distribution of the FOD under the final node.

Also, the results of this research are based upon a model and results may vary depending upon the random assignment of FOD from the derived empirical distribution in each node of the classification tree.

2.0 Results: Modeling FOD and Non-response

2.1 Results: Modeling FOD

The ultimate goal of this project is to impute values of the 142 FOD categories, as well as of non-response, onto every record in a sampling frame. Classification trees are used as a partial way to achieve this goal. In particular, classification tree methodology is used as a way to identify homogenous groups of records. The more homogenous these groups are, the more precise the imputations can be. As an extreme example, if the sample can be accurately partitioned into 142 groups, each group consisting only of records with the same FOD, then the resulting imputations of FOD will be free from variability.

Before describing the relatively simple modeling of non-response, the approach taken to model FOD is summarized. The procedure rpart will find classification trees to predict the categorical variable: FOD. The criterion for model selection is based on a cross-validation estimate of total misclassification error of the predicted FOD categories (note: the Gini index, mentioned above, is used to select the variables). Although rpart can weight certain types of misclassification errors more than others, these weights need to be specified as input. Although DSMD staff could state that the incorrect classification of type of degree into non-science and engineering was the most costly, quantifying the weights was difficult. As an alternative to fitting one model to predict all FOD categories, we investigated stepwise models that concentrated on reducing the misclassification rates of broad FOD categories, followed up by selecting additional classification trees to predict more detailed categories of FOD. Since the misclassification among predictions of the individual 142 category FOD was not of particular interest to DSMD staff, we concentrated on selecting classification trees that predict broader classes of FOD, such as those mentioned in tables 1.1 through 1.4. However, we did fit a classification tree to directly predict the 142 category FOD as one of the alternative modeling procedures. Given a classification tree (or classification trees in the stepwise modeling), the final 142 FOD category is predicted by imputing FOD using its empirical distribution within each final node of the classification tree.

2.1.1 Stepwise Dependent Variable Model Selection: FOD

In general, the rpart program will automatically select a best model, once the dependent variable is identified and the set of possible covariates are specified. The program finds the best model (also called a "tree") using cross-validated estimates of misclassification error. As mentioned previously, modal predictions used to calculate the misclassification error are the modal values of the data set within each of the partition subsets. (Note: the rpart program can only select the final model using this definition of prediction.)
The entire 142 categories of FOD need to be imputed. However, misclassification errors are more important between the three major types of degrees (S&E, S&E-R and non-S&E) than within each of these groups. To explicitly guard against misclassification errors between the

three major types of degrees (and other broad degree-type categories), we evaluated different models that successively predicted more detailed categories of the FOD variables. Our aim was to obtain models that predicted the eight FOD3 categories well, with an emphasis on reducing the misclassification error for the non-S&E group.

A number of sequential modeling procedures were tried since it was believed that it was easier to empirically compare the estimated misclassification rates from the models rather than to provide a theoretical reason why one sequential strategy should be preferred to another. In this initial part of the development, the estimated misclassification rates were not based on the imputed FOD values, but instead were based on the modal predicted values from the model. For example, if the model is developed to predict all eight FOD3 categories, then the best predicted value for all records in a particular terminal node would consist of modal value from the empirical distribution of FOD3 in that node. This way of prediction is the default approach in rpart and is used by rpart as part of the final tree selection. Comparisons of different models from rpart were made using the misclassification rate (the percentage of the predicted FOD value that is different from the observed survey FOD value in the response sampling units) with respect to the three category variable CFOD2.

The following models based on different stepwise dependent variables were evaluated, using the estimated misclassification error rate. (Recall that a misclassification occurs when the predicted FOD value is not same as the observed FOD value).

1) One-step modeling of the 11- category FOD1 by CFOD2 cross-classified variable as outlined in Table 1.4. (termed Model B1)

2) One-step modeling of the 3 category field of degree variable CFOD2 in Table 1.1. (termed Model B2)

3) One-step modeling of the 8 category field of degree variable FOD3 in Table 1.2. (termed Model B3)

4) One-step modeling of the detailed 142-category field of degree variable FOD2. (termed Model B4)

5) Two-step modeling of the 3 category field of degree variable CFOD2 in Table 1.1. First model the binary variable of S&E (CFOD2=1) or not S&E . Then, given not S&E, model the binary variable S&E-R (CFOD2=2) versus non-S&E (CFOD2=0). (termed Model C).

6) Two-step modeling of the 3 category field of degree variable CFOD2 in Table 1.1. First model the binary variable of S&E-R(CFOD2=2) or not S&E-R. Then, given not S&E-R,, model the binary variable S&E (CFOD2=1) versus non-S&E (CFOD2=0). (termed Model D)

7) Two-step modeling of the 3 category field of degree variable CFOD2 in Table 1.1. First model the binary variable of non-S&E(CFOD2=0) or not non-S&E. Then, given not non-S&E, model the binary variable S&E (CFOD2=1) versus S&E-R (CFOD2=2). (termed Model E)

8) Two-step modeling of the 8 category field of degree variable FOD3 in Table 1.2. First model the binary variable of non-S&E, (i.e. CFOD2=0 or equivalently FOD3=8) or not non-S&E. Then, given not non-S&E, model the remaining seven categories of FOD3. (termed Model F)

The predicted value of "FOD" under each model is obtained from the modal assignment under the final nodes of the respective model from rpart. We then cross tabulate the observed and the predicted "FOD" categories from the respondents in the sample. The "FOD" categories in the respective model are collapsed into 3-category CFOD2. The total misclassification error from each model is the summation of the off diagonal percent table of the observed versus the predicted 3-category CFOD2 table.

The total misclassification rate based on CFOD2 (in percent) for the models considered is listed in Table 2.1.1.

Table 2.1.1. Misclassification rate (in percent) of the models

Model	B1	B2	B3	B4	C	D	E	F
Rate	34.36	32.99	34.18	34.46	33.16	32.87	33.26	33.31

In summary, the estimated misclassification rates arising from the different stepwise dependent variable models were not too different from one another. However, the model to be used was picked because it gave a slight estimated edge in classifying non-S&E degrees correctly.

After consulting with DSMD staff, model F was chosen to create partitions for FOD3. This model had one of the lowest misclassification errors of non-S&E (not shown here, but can be seen from the detailed observed versus predicted CFOD2 tables for each model). The fact that the model directly discriminates between the eight FOD3 categories was also important for NSCG sample design planning purposes. A final procedure for model fitting and imputing the 142 FOD values was implemented. This procedure, a two-step model fitting with final imputation (not modal assignment) from the empirical FOD distribution, is described below.

2.1.2 Strategy for Modeling and Imputing FOD

Stepwise Modeling of FOD

Step 1: Recode all respondents into a binary dependent variable consisting of non-S&E degrees versus all others (i.e.,FOD3=8 or not). Use rpart to determine the best fitting classification tree. Call this model, with the resulting rules for splitting the sample frame into those cases with FOD3=8 or not, "Tree(FOD3=8, FOD3 ≠ 8)."

Step 2: For each terminal node in the model "Tree(FOD3=8, FOD3 ≠ 8)" calculate the empirical distribution function of the detailed FOD variable. Call the empirical distribution at node k "P(FOD| node=k, Tree(FOD3=8, FOD3 ≠ 8))".

Step 3: Recode all respondents who have either a Science & Engineering degree or a Science & Engineering Related degree into the remaining seven category variable (i.e., FOD3=1,…,7). Use rpart, again, to determine the best fitting classification tree. Call this model with the resulting rules for splitting the sample frame into those cases with FOD3=1,…,7, given that FOD3 ≠ 8, "Tree(FOD3=1,…,7 | FOD3 ≠ 8)".

Step 4: For each terminal node in the model "Tree(FOD3=1,…,7 | FOD3 ≠ 8)" calculate the empirical distribution function of the detailed FOD variable. Call the empirical distribution at node k "P(FOD| node=k, "Tree(FOD3=1,…,7 | FOD3 ≠ 8)")".

Stepwise Imputation of FOD

Given a new frame with the same set of independent variables used in the model, impute values of FOD as follows.

Step 5: Determine the node for each observation using "Tree(FOD3=8, FOD3 ≠ 8)." For an observation assigned to node k, by the model, randomly assign a value of FOD using the frequency distribution: "P(FOD| node=k, "Tree(FOD3=8, FOD3 ≠ 8)")." If the realized value of FOD falls into the collapsed category, FOD3=8, keep the assigned FOD value. If, the realized value of FOD falls into a collapsed category of FOD3 ≠ 8, throw it away and proceed to step 6. Since the model of step 1 is used to identify and remove the non-S&E degree only, the second tree model needs to be used to refine the distribution of FOD3=1,…,7, before a final imputation is made.

Step 6: Given that the realized value of FOD, from step 5, is in a collapsed category such that FOD3 ≠ 8, determine the final node of the observation using the tree: "Tree(FOD3=1,…,7 | FOD3 ≠ 8)." For an observation assigned to node k, by this model, randomly assign a value of FOD using the frequency distribution: "P(FOD| node=k, "Tree(FOD3=1,…,7 | FOD3 ≠ 8)")."

In their entirety, step 5 and step 6 determine the imputed value FOD.

2.1.3 Summary of Classification Tree Developed For FOD

Appendix 4 lists the entire classification tree for modeling FOD. One tree is developed to classify records into non-science and engineering degrees (coded as "0") and all others (coded as "N"). The other tree is developed to classify the others into the remaining 7 categories of science and engineering degrees in FOD3.

Variables used in the tree that classifies non-science and engineering from all others are the following:
ROCCUP03, DEMGRP03, IND21, PAOCF, XAGE, HIDEG03, QINCTOT, and POBREV

Variables used in further classifying the science and engineering (and related) degrees into the remaining 7 categories are:
ROCCUP03, POBREV, PAOCF, HIDEG03, IND21, REGION, XAGEIN, GUMOS, XAGE, XQWK, QMS and XPERN

2.1.4 Estimated Misclassification Error of Imputed FOD

Misclassification rates of the imputed FOD of the sampling frame were estimated by, first, using the model to impute FOD values for the original 2003 NSCG sample, then comparing the imputed value with the actual value and, finally, considering it a match if both imputed and actual values are in the same collapsed FOD group of interest and weight it up.

Table 2.1.4.1.a. Percentage table of observed CFOD2 (row) values x imputed CFOD2 (column) values

	non-S&E	S&E	S&E -R	Total
non-S&E	34.08	17.47	3.77	55.33
S&E	16.34	15.71	2.90	34.95
S&E-R	3.74	3.05	2.93	9.72
Total	54.17	36.23	9.60	100.00

The total misclassification rate is 47.28% (100%-52.72%) in the response data. Notice that the total percentage of imputed CFOD2 category is very close to the total observed CFOD2 category. In the response data, the percentage of the observed non-S&E category is 55.33 %. The percentage of the imputed non-S&E category is 54.17%. This is because we use the empirical distribution of FOD to impute the FOD value in the final node. The percentage of the imputed S&E category is 36.23% as compared to 34.95% of the observed S&E category. The percentage of the imputed S&E-R category is 9.60% as compared to the percentage of the observed 9.72%.

The following table presents the misclassification rate relative to the three-category CFOD2. Columns sum to 100% since the error rate of the predicted values is the quantity of interest.

Table 2.1.4.1.b. Percentage table of observed CFOD2 (row) values x imputed CFOD2 (column) values, conditional on the imputed CFOD2

	non-S&E	S&E	S&E-R
non-S&E	62.93	48.22	39.25
S&E	30.16	43.37	30.25
S&E-R	6.91	8.41	30.50
Total	100.00	100.00	100.00

In the imputed non-S&E category, 62.93% are correctly imputed as compared to the observed FOD value. In the S&E degree category, 43.37% are correctly imputed. And in the S&E-R category, 30.25% are correctly imputed.

Table 2.1.4.2.a. Percentage of observed FOD3 (row) values by the imputed FOD3 (column) values

	S &E					S&E-R		Non-S&E	
	1	2	3	4	5	6	7	8	Total
1	0.64	0.21	0.14	0.50	0.56	0.11	0.18	1.61	3.95
2	0.17	0.98	0.38	0.79	0.32	0.67	0.17	2.70	6.19
3	0.15	0.35	0.28	0.32	0.35	0.17	0.11	1.14	2.88
4	0.53	0.70	0.36	2.73	0.69	0.68	0.35	7.99	14.03
5	0.51	0.38	0.38	0.60	2.67	0.12	0.33	2.89	7.91
6	0.13	0.70	0.17	0.73	0.16	2.52	0.08	2.20	6.70
7	0.15	0.17	0.10	0.36	0.37	0.08	0.25	1.55	3.02
8	2.31	2.75	1.20	8.52	2.70	2.26	1.51	34.08	55.33
Total	4.58	6.24	3.02	14.56	7.83	6.62	2.99	54.17	100.00

The total misclassification rate of imputed FOD3 is 55.85% (100%-44.15%) by adding the off diagonals. The correct classification rate for non-S&E category is 34.08%, the largest in all 8 categories of FOD3.

The imputed non-S&E in the response is 54.17% as compared with the observed non-S&E of 55.33%. The next largest percentage of the imputed FOD3 category is 4 (Social and Related Science) with 14.56% as compared with the observed of 14.03%. The percentage of the imputed category of FOD3 is very close to the observed category of FOD3 for all 8 categories.

The following table presents the misclassification rate relative to the eight-category FOD3. Columns sum to 100% since the error rate of the predicted values is the quantity of interest.

Table 2.1.4.2.b. The percentage of observed FOD3 (row) by imputed FOD3 (column) conditional on the imputed FOD3 category

	S&E					S&E-R		non-S&E
	1	2	3	4	5	6	7	8
1	14.08	3.31	4.70	3.40	7.21	1.63	6.04	2.97
2	3.66	15.65	12.71	5.43	4.14	10.19	5.82	4.99
3	3.24	5.67	9.39	2.18	4.41	2.60	3.82	2.11
4	11.50	11.17	11.92	18.78	6.86	10.21	11.79	14.75
5	11.17	6.17	10.66	4.14	34.14	1.88	11.06	5.34
6	2.83	11.22	5.70	5.03	2.06	38.12	2.71	4.06
7	3.17	2.78	3.39	2.48	4.70	1.21	8.24	2.85
8	50.36	44.03	39.53	58.56	34.47	34.17	50.52	62.93
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Conditional on the imputed FOD3 category, the correct classification rate of the imputed non-S&E (FOD3=8) is 62.93 %.

2.1.5 Estimated Misclassification Error Based on rpart Predictions

The following tables show the estimated misclassification rates if we had used the rpart procedure to predict FOD3 (i.e. using modal values from each subset partition).

Table 2.1.5.1.a. The Percentage of observe (row) CFOD2 values x modal predictions (column) of CFOD2 values

	non-S&E	S&E	S&E-R	Total
non-S&E	48.29	5.85	1.19	55.33
S&E	19.19	14.04	1.73	34.95
S&E-R	3.86	1.50	4.36	9.72
Total	71.33	21.38	7.28	100.00

Conditional on the response data, the total correct classification rate for 3 categories of CFOD2 from model F (adding the diagonals from the above table) is 66.69% (or misclassification rate is 33.31%). The correct classification rates from model F for categories of non-S&E, S&E and S&E-R are 48.29% 14.04% and 4.36% , respectively. The predicted percentages for categories of non-S&E , S&E and S&E-R are 71.33%,21.38% and 7.28%, respectively.

It presents the misclassification rate relative to the three-category CFOD2. The columns sum to 100% since the error rate of the predicted values is the quantity of interest.

Table 2.1.5.1.b. Percentage of observed CFOD2 (row) values x modal predictions (column) of CFOD2 values

	non-S&E	S&E	S&E-R
Non-S&E	67.70	27.34	16.33
S&E	26.90	65.65	23.75
S&E-R	5.41	7.00	59.93
Total	100.00	100.00	100.00

The correct classification rates conditional on the predicted CFOD2 categories are 67.70%, 65.65% and 59.93% for non-S&E, S&E , and S&E-R, respectively.

Table 2.1.5.2.a. The percentage of observed (row) FOD3 values x modal predictions (column)of FOD3 values

	S &E					S&ER		Non-S&E	
	1	2	3	4	5	6	7	8	Total
1	1.26	0.05	0.02	0.19	0.61	0.03	0.00	1.80	3.95
2	0.12	1.34	0.15	0.40	0.36	0.79	0.03	3.00	6.19
3	0.12	0.29	0.38	0.21	0.52	0.15	0.02	1.18	2.88
4	0.39	0.22	0.02	1.57	0.61	0.54	0.04	10.65	14.03
5	0.53	0.08	0.26	0.22	4.31	0.06	0.07	2.56	7.91
6	0.05	0.45	0.02	0.16	0.06	4.05	0.00	1.90	6.70
7	0.19	0.02	0.00	0.10	0.44	0.04	0.27	1.96	3.02
8	1.36	0.32	0.04	2.02	2.11	1.08	0.11	48.29	55.33
Total	4.02	2.76	0.71	4.88	9.02	6.74	0.54	71.33	100.00

Conditional on the response data, the correct classification rate of FOD3 (adding all the diagonals) is 61.47%, or the misclassification rate is 38.53%. In the observed NSCG 03 data, 55.33% is in category 8 (Non-S&E), in comparing with the predicted FOD3 of 71.33% is in category 8 (Non-S&E). The percentage of the observed category 4 of FOD3 is 14.03%, while the percentage of the predicted category 4 (Social & Related Science) of FOD3 is 4.88%.

The following table presents the misclassification rate relative to the eight-category FOD3. Columns sum to 100% since the error rate of the predicted values is the quantity of interest.

Table 2.1.5.2.b. The percentage of observed (row) FOD3 values x modal predictions (column)of FOD3 values

	S&E					S&E-R		non-S&E
	1	2	3	4	5	6	7	8
1	31.32	1.69	2.53	3.86	6.73	0.49	0.50	2.52
2	3.04	48.52	20.78	8.23	3.97	11.79	5.22	4.21
3	2.96	10.59	53.78	4.30	5.80	2.18	3.99	1.66
4	9.63	7.89	3.49	32.25	6.72	7.94	6.59	14.93
5	13.15	2.84	10.47	4.57	47.85	0.84	13.85	3.58
6	1.19	16.28	2.92	3.36	0.71	60.03	0.54	2.67
7	4.82	0.68	0.67	2.03	4.84	0.66	49.80	2.74
8	33.90	11.52	5.35	41.40	23.38	16.07	19.52	67.70
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

As can be seen, using the modal values for predicting FOD3 results in lower misclassification errors than does imputing from the empirical distribution of FOD (comparing table 2.1.5.2.b with table 2.1.4.2.b). However, as can be seen in the table of unconstrained predictions (Table

2.1.4.2.a), the empirical distribution of the imputed values matches the empirical distribution of the observed data more closely that the empirical distribution of the modal predicted values. This feature of the imputed values was important to the DSMD staff and so the imputed values (i.e., values from the empirical distribution) were selected as the method to predict the final FOD categories.

Being a two-category variable, non-response can be modeled in a more straightforward manner. The classification tree model fitting procedure closely follows the simple example in Appendix 1. The steps for imputing non-response in a new sampling follow.

2.2.1 Strategy for Modeling and Imputing Non-response

Stepwise Modeling of Non-response

Step 1: Use rpart to determine the best fitting classification tree to non-response. Call this model, with the resulting rules for splitting the sample, "Tree(non-response)."

Step 2: For each terminal node, k, in the model "Tree(non-response)," calculate the proportion that are responses, p_k .

Stepwise Imputation of Non-response

Given a new frame with the same set of independent variables used in the model, impute values of non-response as follows.

Step 3: Determine the node for each of the new observations using "Tree(non-response)." For an observation assigned to node k by the model, use a Bernoulli random variable with proportion success p_k to assign a value of "Y" for response, if a success, or "N" for non-response, if a failure.

2.2.2 Summary of Classification Tree Developed For Non-response

Appendix 5 lists the entire classification tree for non-response. The following variables were used to form the tree.
DEMGRP03, XQYR2US, ROCCUP03, POV, QREL, QCOW, FOCC, XAGE, QMS, GUMOS, IND21, XAGEIN and ESR .

Of the 30 independent variables that could be included in the model, 21 were used. The nine independent variables which were not included are:
AIAN, ASIAN, BLACK, NEWCSAB, NEWHISP, NHPI, OTHER, qpowccs ,ESP

2.2.3 Estimated Misclassification Error of Imputed Non-response

As in section 2.1.4, weighted misclassification rates for non-response can be used to estimate the misclassification rate in the entire sampling frame. This is accomplished by imputing non-response and then comparing the imputed values to the actual values of the response variable. The following table presents the misclassification rates (in percent) for non-response.

Table 2.2.3. Response (row) x Imputed Response (column)

	N	Y
N	44.97	37.72
Y	55.03	62.28
Total	100.00	100.00

2.2.4 Estimated Misclassification Error of Modal Predictions

As in section 2.1.5, weighted misclassification rates based on the modal predicted values from the rpart program can be calculated. The following table presents the misclassification rates (in percent) for non-response based on mode assignment.

Table 2.2.4.1. Response (row) x Modal Predicted Response (column)

	N	Y
N	61.35	33.80
Y	38.65	66.20
Total	100.00	100.00

As with FOD, the modal predictions have a smaller estimated misclassification error. Again, as evidenced in the following two tables, the empirical distribution of the imputed values is much closer to the observed than the model predictions.

Table 2.2.4.2. Full classification: Response (row) x Imputed Response (column)

Observed	N	Y	Total
N	17.68	22.89	40.57
Y	21.64	37.79	59.43
Total	39.32	60.68	100

Table 2.2.4.3. Full classification: Response (row) x Modal Predicted Response (column)

Observed	N	Y	Total
N	15.13	25.44	40.57
Y	9.54	49.89	59.43
Total	24.67	75.33	100.00

2.3 Software Used

As an example, the R software used to impute non-response is included in Appendix 6. The R batch file to model non-response is included. In that batch file, there is a call to rpart, a call to the user-defined function “cpvalue”- a cost-complexity measure (see Breiman, et al. (1984), page 66) , used to pick the minimum cp statistic used in the R function “prune” to pick a final tree.

Lastly, the user-defined R function, "rpart2sas_all," which outputs the SAS statements needed to assign the classification tree nodes to a new data set is also included.

3.0 Cross-validated comparisons with predictions from NSCG strata

A quick way to provide imputed FOD values is to use the distribution of the sampled FOD values and non-responses in the 2003 NSCG strata to impute for the sampling frames. DSMD staffs have used this approach while the classification tree models were being developed and it is instructive to compare the two, since the strata approach is relatively easy to implement and the classification tree approach is lengthy and involved.

As the components that make up the sampling strata were included as independent variables in the classification tree modeling, the classification tree should do as well as, if not better than, the strata approach in imputing values. However, there are several reasons why this may not be true. First, rpart is a stepwise procedure, so that optimal models may still be missed. Second, we evaluate the misclassification of imputed values, not the modal predicted values that rpart was designed for. Third, empirically, classification tree methods appear to be fairly robust to over-specified models, as reported in Breiman, et al. (1984). Thus gains in prediction can be realized by including enough of variables in the model while including too many does not seem to increase the error for an appreciable range of model sizes.

In this comparison, weighted estimates of misclassification were cross-validated to guard against possibly over-fitting a larger model. Specifically, the sample was randomly split in half, the distribution of FOD (or of non-response) was derived from one of the half-samples, the detailed FOD (or non-response) was imputed to the records in the other half-sample, and weighted estimates of misclassification were made.

The table of classification rates in percent due to imputation in terms of CFOD2 from both the classification tree model and from the NSCG03 strata model are as follows.

Table 3.0.1.a. The percentage of observed CFOD2 (row) x imputed CFOD2 (column) from the classification tree model

	non-S&E	S&E	S&E-R	Total
non-S&E	34.29	17.41	3.85	55.56
S&E	16.00	15.82	2.86	34.69
S&E-R	3.75	2.99	3.02	9.76
Total	54.04	36.22	9.73	100.00

The misclassification rate in percent is 46.87% by adding off diagonals.

Table 3.0.1.b. The percentage of observed CFOD2 (row) x imputed CFOD2 (column) from the classification tree model, conditional on the imputed category of CFOD2.

	non-S&E	S&E	S&E-R
non-S&E	63.46	48.07	39.53
S&E	29.61	43.67	29.42
S&E-R	6.93	8.25	31.05
Total	100.00	100.00	100.00

Table 3.0.2 a. The percentage of observed CFOD2 (row) x imputed CFOD2 (column) from the NSCG03 sampling strata model

	non-S&E	S&E	S&E-R	Total
non-S&E	35.29	16.43	3.84	55.56
S&E	16.22	15.36	3.11	34.69
S&E-R	3.73	3.10	2.93	9.76
Total	55.23	34.89	9.88	100.00

The misclassification rate of CFOD2 using NSCG03 sampling strata model is 46.42%.

Table 3.0.2 b. The percentage of observed CFOD2 (row) x imputed CFOD2 (column) from the NSCG03 sampling strata model, conditional on the imputed category of CFOD2

	non-S&E	S&E	S&E-R
non-S&E	63.89	47.08	38.86
S&E	29.36	44.03	31.49
S&E-R	6.75	8.89	29.65
Total	100.00	100.00	100.00

Since the tree model was developed based on the FOD3 categories, cross validated estimates of the classification tree model and the sampling strata model are compared next.

Table 3.0.3.a. The percentage of observed FOD3 (row) x imputed FOD3 (column) from the classification tree model

	1	2	3	4	5	6	7	8	Total
1	0.63	0.22	0.16	0.48	0.60	0.11	0.18	1.55	3.93
2	0.20	1.00	0.36	0.85	0.34	0.70	0.15	2.67	6.27
3	0.15	0.36	0.27	0.30	0.36	0.15	0.09	1.19	2.86
4	0.47	0.74	0.40	2.67	0.70	0.73	0.30	7.73	13.75
5	0.51	0.34	0.42	0.69	2.61	0.14	0.32	2.86	7.88
6	0.11	0.68	0.18	0.73	0.13	2.61	0.05	2.24	6.73
7	0.16	0.14	0.12	0.33	0.41	0.09	0.27	1.50	3.02
8	2.23	2.91	1.26	8.47	2.53	2.33	1.51	34.29	55.56
Total	4.45	6.41	3.17	14.51	7.68	6.86	2.87	54.04	100.00

The misclassification rate of FOD3 from model F is 55.65% (adding the off diagonals).

Table 3.0.3.b. The percentage of observed FOD3 (row) x imputed FOD3 (column) from the classification tree model, conditional on the imputed total

	1	2	3	4	5	6	7	8
1	14.21	3.51	4.99	3.30	7.79	1.56	6.17	2.87
2	4.41	15.56	11.48	5.84	4.46	10.18	5.28	4.94
3	3.35	5.67	8.50	2.04	4.69	2.14	3.07	2.20
4	10.47	11.62	12.59	18.36	9.16	10.70	10.54	14.31
5	11.43	5.32	13.16	4.73	33.96	1.97	11.20	5.28
6	2.45	10.67	5.55	5.05	1.66	38.07	1.65	4.15
7	3.57	2.16	3.86	2.29	5.32	1.38	9.33	2.78
8	50.11	45.48	39.86	58.38	32.97	34.00	52.75	63.46
total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 3.0.4.a.. The percentage of observed FOD3 (row) x imputed FOD3 (column) from the NSCG03 sampling strata model

	1	2	3	4	5	6	7	8	total
1	0.58	0.21	0.14	0.47	0.57	0.09	0.18	1.69	3.93
2	0.15	0.95	0.37	0.81	0.30	0.82	0.19	2.67	6.27
3	0.11	0.40	0.28	0.30	0.36	0.16	0.12	1.14	2.86
4	0.38	0.77	0.34	2.52	0.76	0.71	0.35	7.92	13.75
5	0.44	0.39	0.35	0.62	2.80	0.14	0.35	2.79	7.88
6	0.11	0.75	0.17	0.76	0.14	2.59	0.08	2.12	6.73
7	0.15	0.18	0.07	0.41	0.36	0.07	0.18	1.60	3.02
8	1.75	2.61	1.23	8.11	2.73	2.19	1.65	35.29	55.56
total	3.68	6.25	2.94	14.00	8.02	6.78	3.10	55.23	100.00

The misclassification rate of FOD3 using NSCG03 sampling strata model is 54.81%.

Table 3.0.4.b.. The percentage of observed FOD3 (row) x imputed FOD3 (column)from the NSCG03 sampling strata model, conditional on the imputed FOD3 category

	1	2	3	4	5	6	7	8
1	15.75	3.31	4.67	3.36	7.14	1.32	5.80	3.07
2	4.17	15.20	12.52	5.78	3.78	12.14	6.13	4.84
3	3.04	6.40	9.44	2.18	4.43	2.29	3.84	2.06
4	10.42	12.25	11.46	17.98	9.51	10.51	11.22	14.34
5	11.98	6.21	11.94	4.40	34.89	2.11	11.34	5.05
6	3.05	12.01	5.79	5.44	1.70	38.29	2.72	3.84
7	4.17	2.86	2.25	2.95	4.47	1.00	5.86	2.91
8	47.41	41.76	41.92	57.91	34.08	32.35	53.09	63.89
total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Finally the classification tree model developed for non-response is compared to the strata model. The results follow.

Table 3.0.5.a.. The percentage of observed response (row) x imputed response (column) from the classification tree model

	No	Yes	Total
No	17.43	23.26	40.68
Yes	21.36	37.95	59.32
Total	38.79	61.21	100.00

The misclassification rate of response category from cross validation of non-response model of weighted data is 44.62%.

Table 3.0.5.b.. The percentage of observed response (row) x imputed response (column) from the classification tree model, conditional on the imputed response category

	No	Yes
No	44.93	38.00
Yes	55.07	62.00
Total	100.00	100.00

The predicted response (in percent) for each category via cross validation of NSCG03 strata model is given in table below.

Table 3.0.6.a. The percentage of the observed response(row) x imputed response (column) from the NSCG03 sampling strata model

Observed	No	Yes	Total
No	17.61	22.90	40.50
Yes	22.54	36.96	59.50
Total	40.14	59.86	100.00

The misclassification rate of non-response model from NSCG03 sampling strata model of weighted data is 45.44%.

Table 3.0.6.b. The percentage of the observed response (row) x imputed response (column) from the NSCG03 sampling strata model, conditional on the imputed response category

	No	Yes
No	43.86	38.25
Yes	56.14	61.75
Total	100.00	100.00

In summary, the classification tree models and the models using the strata model are comparable in terms of the estimated classification rates.

Although the methods appear to have equal predictive power, the classification tree models are much smaller in terms of terminal nodes (cells). The strata model is based on making estimates from 565 strata, while the classification tree model for FOD3 uses a combined set of 51 terminal nodes. The classification tree model for non-response uses only 27 terminal nodes. This savings in model complexity could be of benefit to estimation issues if a smaller number of homogeneous cells with a relatively large sample size can be identified by the classification tree model.

4.0 Summary

Classification tree methods were applied to model and impute FOD for future sampling frames. A number of conclusions were reached. First, the sampling strata developed for the 2003 NSCG were comparable to classification trees in forming homogeneous subpopulations of FOD. However, the classification tree approach resulted in a much smaller collection of subgroups. Second, fitting a series of classification trees appears to be a way to select trees that guard against specific misclassification errors. The modal predictions, which are the default predictions from rpart, make predictions which do not necessarily match the population distribution, while drawing imputations from the final model partition can closely match the population distribution with some loss in classification.

Acknowledgement

We would like to thank John Finamore and David Warren Hall for monitoring the project and working closely in providing feedback, sharing ideas, and satisfying all our data requests.

5.0 References

Breiman, Leo, Friedman, Jerome H., Olshen, Richard A., and Stone, Charles J. (1984). *Classification and Regression Trees*, Wadsworth International Group, Belmont, California.

Ciampi, A., Chang, C. H., Hogg, S. and McKinney, S. (1987). Recursive Partition: A versatile method for exploratory data analysis, *Festschrift in Honor of Professor V.M. Joshis's 70th Birthday*, Vol. V: Biostatistics, D. Reidel Publishing Company, Dordrecht.

Clark, Linda A. and Daryl Pregibon (1991). "Tree-Based Models S," Chapter 9 in *Statistical Models in S*, John M. Chambers and Trevor Hastie (eds.), Chapman & Hall Computer Science Series.

Ripley, B.D. Classification and Regression Trees in R , *Documentation for package "tree" version 1.0-26*. http://finzi.psych.upenn.edu/R/library/tree/html/00Index.html

Thernean, Terry M., and Atkinson, Elizabeth J. (1997). An Introduction to Recursive Partitioning Using the RPART Routines. Mayo Foundation.

Venables, W.N. and Ripley, B.D. (2002). Tree-based Methods, Chapter 9 in *Modern Applied Statistics with S-Plus*, Fourth Edition. Springer.

Appendix 1. An example of how classification trees are formed

The following represents a simple example on how classification trees are formed and how cross validation methods are used to pick a final tree. This example is provided to give a flavor for the method and to allow this report to be a relatively complete and accessible document. There are many other sources that cover classification trees more completely, with many examples. For example, Breiman, et al. (1984), Chambers, et al (1991), Venables and Ripley (2002), Thernean and Atkinson (1997) and Ciampi, et al. (1987).

Suppose the dependent variable is "job satisfaction" and two independent variables each with two categories are available. Let the first independent variable be occupation code consisting of Science related (SR) or not (NR). Let the second independent variable be industry code consisting, again, of science related (SR) or not (NS). Suppose, further, that a sample of 400 is available and, to keep things simple, that the four cross-classified cells each have a sample of 100.

Sample size table:

		Occupation	
		SR	NS
Industry	SR	100	100
	NS	100	100

The following table contains the proportion in each cell who are satisfied with their job.

Proportion satisfied by Industry and Occupation type:

		Occupation	
		SR	NS
Industry	SR	.80	.40
	NS	.60	.40

Stepwise model building: Starting with a model with no independent variables. The sample is split in two by all possible independent variables (in this case there are two independent variables and only one split each).

Step 0: No split yet
Proportion satisfied with job:

.55

Gini index=.55x.45=0.2475

Step 1:
1) Split on Occupation:

Occupation	
SR	NS
.70	.40

Gini index weighted by sample size:
(1/2)x.70x.30+(1/2)x.40x.60=0.225

2) Split on Industry:

Industry	SR	.60
	NS	.50

Gini index weighted by sample size:
(1/2)x.60x.40+(1/2)x.50x.50=0.245

Step 1: best split is on Occupation since this offers the greatest reduction in the average Gini index

Model after step 1:

Occupation	
SR	NS
.70	.40

Step 2
1) split Occupation = SR cell:

		Occupation	
		SR	NS
Industry	SR	.80	.40
	NS	.60	

Weighted Gini index:
(1/4).80x.20+(1/4)x.60x.40+(1/2)x.40x.60=.22

2) split Occupation =NS cell:

		Occupation	
		SR	NS
Industry	SR	.70	.40
	NS		.40

Weighted Gini index:
(1/2)x.70x.30+(1/4)x.60x.40+(1/4)x.40x.60=.225

Since the split in on Occupation=SR provides the greatest reduction in the Gini index, the final two-level model split first on Occupation, then splits within the Occupation=SR cell.

Since there is only one more split available, the final three-level model has to split the Occupation=NS cell.

Selecting a final model: The stepwise procedure provides a final model for each of four levels: no splits, 1 split, two splits and 3 splits.
If there was no concern about over-fitting the model to the table, the above classification into four final nodes could be used as a convenient summary of the data. However, since classification tree methods use the data intensively, it is important to guard against over-fitting the model. The final choice of model is accomplished using cross-validation.

Picking the final model using cross-validation

Cross validation is used to pick one of the final models from those picked above. The basic idea of cross-validation is to randomly split the sample, model one part and check its predictivity on the other. Sample size is recovered by repeating this on each of many samples. Goodness of fit statistics that use the same data to model and check fit show a better fit as the model gets bigger, if the model is flexible enough to account for the noise in the data. Cross-validated goodness of fit statistics do not continue to get smaller since some of the noise is not available to model.

Although the rpart default will divide the data into 10 nearly equal size samples, only two samples will be used here.

Cross validation:

Suppose the sample is randomly split in two and that the resulting sample sizes are as follows.

Sub-sample 1 crossed by occupation and industry:

Sample size for first subsample:

		Occupation	
		SR	NS
Industry	SR	60	50
	NS	50	50

Hence, the remaining sample sizes in the sub-sample are:

Other subsample size:

		Occupation	
		SR	NS
Industry	SR	40	50
	NS	50	50

Suppose, that the proportions of the first subsample who are satisfied with their job are as follows.

Subsample one:

		Occupation	
		SR	NS
Industry	SR	.9	.3
	NS	.6	.5

And the proportions of the remaining subsample who are satisfied with their job are as follows.

Subsample two:

		Occupation	
		SR	NS
Industry	SR	.2	.5
	NS	.6	.3

Classification tree models are used to fit each subsample independently of the other. The following summarizes the steps of model fitting.

Step 0:
Subsample one:
Overall proportion with job satisfaction: 0.5905
Gini index: 0.2418
The standard prediction rule used for classification trees is to assign every case the modal value on the node, in this case as being satisfied.
Applying this prediction to the other sample, the total number of miss-classified cases in the other sample is 112.

Subsample two:
Overall proportion with job satisfaction: 0.4105
Gini index: 0.2420
In this case everyone will be classified as not satisfied (the majority in subsample 2).
Applying this prediction to sample one, the total number of misclassified cases in the other sample is 124.

With no splitting, the total misclassification error is 236.

Step 1:

Subsample 1: Gini index splitting on Occupation:
(11/21)x.76x.24+(10/21)x.4x.6= 0.2098
Gini index splitting on Industry
(11/21)x.63x.37+(10/21)x.55x.45= 0.2400
So that splitting on Occupation is optimal

In this case, records with occupation of SR are predicted as satisfied, the other as not satisfied, giving a misclassification error when applied to sample two as 92.

Subsample 2:
Gini index splitting on Occupation:
(9/19)x.42x.58+(10/19)x.4x.6= 0.2417

Gini index splitting on Industry
(9/19)x.63x.37+(10/19)x.55x.45=0.2407

So that splitting on Industry is optimal.

The misclassification error applying the optimal prediction to sample one is 124.

With one split, the total misclassification error is: 216, lower than no split.

Step 2

For subsample one:
The next split is of Occupation=SR.
(6/21)x.9x.1+(3/21)x.6x.4+(10/21)x.4x.6=0.1743
For the split of Occupation=NS:
(11/21)x.76x.24+(50/210)x.3x.7+(50/210)x.5x.5= 0.2051

So that next split is to Occupaton=SR. The misclassification applied to sample 2 is again 92.

For subsample two:
The next split is of Industry=SR
(4/19)x.2x.8+(5/19)x.5x.5+(10/19)x.45x.55= 0.2297

For the split of Industry=NS :
(9/19)x.37x.63+(5/19)x.6x.4+(5/19)x.3x.7= 0.2288
So the next split is on Industry=NS , giving total miss-classification error based on sample one is 114.

So, the total misclassification error at step two is 206

Step 3:

With only one, possible, split left, the total misclassification error is 188. In this example, using the complete model with four terminal nodes is justified.

Appendix 2. Detailed Field of Degree (FOD) definition

FOD	Description
601	Agricultural economics
602	OTHER agricultural business and production
605	Animal sciences
606	Food sciences and technology
607	Plant sciences
608	OTHER agricultural sciences
610	Architecture/Environmental Design
620	Area and Ethnic Studies
631	Biochemistry and biophysics
632	Biology, general
633	Botany
634	Cell and molecular biology
635	Ecology
636	Genetics, animal and plant
637	Microbiological sciences and immunology
638	Nutritional sciences
639	Pharmacology, human and animal
640	Physiology and pathology, human and animal
641	Zoology, general
642	OTHER biological sciences
651	Accounting
652	Actuarial science
653	Business administration and management
654	Business, general
655	Business and managerial economics
656	Business marketing/marketing management
657	Financial management
658	Marketing research
659	OTHER business management/administrative services
661	Communications, general
662	Journalism
663	OTHER communications
671	Computer and information sciences
672	Computer programming
673	Computer science
674	Computer systems analysis
675	Data processing
676	Information services and systems
677	OTHER computer and information sciences
680	Environmental science or studies

681	Forestry sciences
682	OTHER natural resources and conservation
690	Criminal Justice/Protective Services
701	Education administration
702	Computer teacher education
703	Counselor education and guidance services
704	Educational psychology
705	Elementary teacher education
706	Mathematics teacher education
707	Physical education and coaching
708	Pre-school/kindergarten/early childhood teacher education
709	Science teacher education
710	Secondary teacher education
711	Special education
712	Social science teacher education
713	OTHER education
721	Aerospace, aeronautical and astronautical engineering
722	Agricultural engineering
723	Architectural engineering
724	Bioengineering and biomedical engineering
725	Chemical engineering
726	Civil engineering
727	Computer and systems engineering
728	Electrical, electronics and communications engineering
729	Engineering sciences, mechanics and physics
730	Environmental engineering
731	Engineering, general
732	Geophysical and geological engineering
733	Industrial and manufacturing engineering
734	Materials engineering, including ceramics and textiles
735	Mechanical engineering
736	Metallurgical engineering
737	Mining and minerals engineering
738	Naval architecture and marine engineering
739	Nuclear engineering
740	Petroleum engineering
741	OTHER engineering
751	Electrical and electronic technologies
752	Industrial production technologies

753	Mechanical engineering-related technologies
754	OTHER engineering-related technologies
760	English Language, literature and letters
771	Linguistics
772	OTHER foreign languages and literature
781	Audiology and speech pathology
782	Health services administration
783	Health/medical assistants
784	Health/medical technologies
785	Medical preparatory programs (e.g. pre-dentistry,-medical,-veterinary)
786	Medicine (dentistry,optometry,osteopathic,podiatry,veterinary)
787	Nursing (4 years or longer program)
788	Pharmacy
789	Physical therapy and other rehabilitation/therapeutic services
790	Public health (including environmental health and epidemiology)
791	OTHER health/medical sciences
800	Home Economics
810	Law/Prelaw/Legal Studies
820	Liberal Arts/General Studies
830	Library Science
841	Applied mathematics
842	Mathematics, general
843	Operations research
844	Statistics
845	OTHER mathematics
850	Parks, Recreation, Leisure, and Fitness Studies
861	Philosophy of science
862	OTHER philosophy, religion, theology
871	Astronomy and astrophysics
872	Atmospheric sciences and meteorology
873	Chemistry, except biochemistry
874	Earth sciences
875	Geology
876	Geological sciences, other
877	Oceanography
878	Physics
879	OTHER physical sciences
891	Clinical psychology
892	Counseling psychology
893	Experimental psychology

894	General psychology
895	Industrial/Organizational psychology
896	Social psychology
897	OTHER psychology
901	Public administration
902	Public policy studies
903	OTHER public affairs
910	Social Work
921	Anthropology and archaeology
922	Criminology
923	Economics
924	Geography
925	History of science
926	History, other
927	International relations
928	Political science and government
929	Sociology
930	OTHER social sciences
941	Dramatic arts
942	Fine arts, all fields
943	Music, all fields
944	OTHER visual and performing arts
991	Science, unclassified
995	OTHER FIELDS (Not Listed)

Appendix 3. Derived independent variables

ACS strata:

```
/* ACS05 subsampling stratum(SBSTR);*/
if GUMOS < 200 then SBSTR='5';
else if 200 <= GUMOS < 800 then SBSTR='2';
else if 800 <= GUMOS < 1200 then SBSTR='3';
else if TRACTMOS >= 2000 then SBSTR='4';
else SBSTR='1';
```

collapsed occupation code:

```
If OCCUP03 = '16' and QOCC in
('300','301','304','305','306','312','314','330','332','341','351','352','353')
 then REVOCCUP03 = '16A';
Else if OCCUP03 = '17' and QOCC in ('011','101','120') then REVOCCUP03 = '17A';
Else if OCCUP03 = '17' and QOCC in ('196','200','201','202') then REVOCCUP03 = '17B';
Else if OCCUP03 = '17' and QOCC in ('030','035','036') then REVOCCUP03 = '17C';
Else if OCCUP03 = '20' and QOCC in ('086','440','592','701','790') then REVOCCUP03 = '20A';
Else if OCCUP03 = '20' and QOCC in ('062','204','205','206','464','525')
then REVOCCUP03 = '20B';
Else if OCCUP03 = '20' and QOCC in ('001','002','003','004','005','006','010','012','013','014',
'015','016','020','021','022','023','031','032','033','034','040','041','042','043','981')
then REVOCCUP03 = '20C';
Else if OCCUP03 = '20' and QOCC in ('210','211','214','215') then REVOCCUP03 = '20L';
Else if OCCUP03 = '20' and QOCC in ('284','286','290','292','296','400','401','403','405',
'471','485','503','563','580','601','602','604','605','610','612','613','620','621','672','674',
'676','680','682','683','684','691','694','700','702','703','704','705','710','712','714','733',
'741','742','743','760','770','771','772','783','804','806','813','874','876','884','904','931',
'951','952','982') then REVOCCUP03 = '20S';
Else if OCCUP03 = '20' and QOCC in ('230','231','233','234','254','255','460')
then REVOCCUP03 = '20T';
Else REVOCCUP03 = OCCUP03;
 ROCCUP03 =REVOCCUP03;
```

collapsed industry code:

```
if xind=0                          then ind21='notinuni';
if xind ge 17 and xind le 29       then ind21='I17t29  ';
if xind ge 37 and xind le 49       then ind21='I37t49  ';
if xind ge 57 and xind le 69       then ind21='I57t69  ';
if xind = 77                       then ind21='I77     ';
if xind ge 107 and xind le 399 then ind21='I107t399';
if xind ge 407 and xind le 459 then ind21='I407t459';
if xind ge 467 and xind le 579 then ind21='I467t579';
```

```
if xind ge 607 and xind le 639 then ind21='I607t639';
if xind ge 647 and xind le 679 then ind21='I647t679';
if xind ge 687 and xind le 699 then ind21='I687t699';
if xind ge 707 and xind le 719 then ind21='I707t719';
if xind ge 727 and xind le 749 then ind21='I727t749';
if xind ge 757 and xind le 779 then ind21='I757t779';
if xind ge 786 and xind le 789 then ind21='I786t789';
if xind ge 797 and xind le 847 then ind21='I797t847';
if xind ge 856 and xind le 859 then ind21='I856t859';
if xind ge 866 and xind le 869 then ind21='I866t869';
if xind ge 877 and xind le 929 then ind21='I877t929';
if xind ge 937 and xind le 987 then ind21='I937t987';
if xind  = 992                 then ind21='unemploy';
```

Appendix 4. Classification trees for predicted FOD3

Tree to classify non-science and engineering (coded as "0") versus all others(coded as "N"):

```
 1) root 97730 41054 N (0.4200757 0.5799243)
   2) ROCCUP03=04,11,17B,18,19,20,20A,20B,20C,20L,20S,20T 55856 22188 0 (0.6027643 0.3972357)
     4) ROCCUP03=19,20,20B,20T 26629  8300 0 (0.6883097 0.3116903) *
     5) ROCCUP03=04,11,17B,18,20A,20C,20L,20S 29227 13888 0 (0.5248229 0.4751771)
      10) DEMGRP03=1,2,4,5,6 23847 10551 0 (0.5575544 0.4424456)
        20) IND21=I407t459,I467t579,I607t639,I647t679,I687t699,I707t719,I866t869,I877t929 5061  1724 0 (0.6593559
                  0.3406441) *
        21) IND21=I107t399,I17t29,I37t49,I57t69,I727t749,I757t779,I77,I786t789,I797t847,I856t859,I937t987 18786  8827 0
                  (0.5301288 0.4698712)
          42) PAOCF=1,2,3,4 8613  3514 0 (0.5920121 0.4079879)
            84) XAGE>=34.5 6236  2348 0 (0.6234766 0.3765234) *
            85) XAGE< 34.5 2377  1166 0 (0.5094657 0.4905343)
             170) ROCCUP03=20A,20C 437   147 0 (0.6636156 0.3363844) *
             171) ROCCUP03=04,11,17B,18,20L,20S 1940   921 N (0.4747423 0.5252577)
               342) IND21=I107t399,I57t69,I727t749,I77 445   191 0 (0.5707865 0.4292135) *
               343) IND21=I17t29,I37t49,I757t779,I786t789,I797t847,I856t859,I937t987 1495   667 N (0.4461538 0.5538462) *
          43) PAOCF=0 10173  4860 N (0.4777352 0.5222648)
            86) HIDEG03=1,2 8495  4246 N (0.4998234 0.5001766)
             172) XAGE>=29.5 7565  3674 0 (0.5143424 0.4856576)
               344) QINCTOT< 63461.5 3915  1752 0 (0.5524904 0.4475096) *
               345) QINCTOT>=63461.5 3650  1728 N (0.4734247 0.5265753)
                 690) IND21=I17t29,I727t749,I757t779,I786t789,I856t859 1662   793 0 (0.5228640 0.4771360) *
                 691) IND21=I107t399,I37t49,I57t69,I77,I797t847,I937t987 1988   859 N (0.4320926 0.5679074) *
             173) XAGE< 29.5 930   355 N (0.3817204 0.6182796) *
            87) HIDEG03=3 1678   614 N (0.3659118 0.6340882) *
      11) DEMGRP03=3,7 5380  2043 N (0.3797398 0.6202602)
        22) HIDEG03=1,2 4102  1746 N (0.4256460 0.5743540)
          44) POBREV=00,01,63,64,66,67,68,69 2468  1191 N (0.4825770 0.5174230)
            88) IND21=I407t459,I57t69,I647t679,I687t699,I757t779,I77,I786t789,I856t859,I866t869,I877t929 1335   610 0
                      (0.5430712 0.4569288)
             176) XAGE>=27.5 1169   503 0 (0.5697177 0.4302823) *
             177) XAGE< 27.5 166    59 N (0.3554217 0.6445783) *
            89) IND21=I107t399,I17t29,I37t49,I467t579,I607t639,I707t719,I727t749,I797t847,I937t987 1133   466 N (0.4112974
                      0.5887026) *
          45) POBREV=58,65 1634   555 N (0.3396573 0.6603427) *
        23) HIDEG03=3 1278   297 N (0.2323944 0.7676056) *
   3) ROCCUP03=01,02,03,05,06,07,08,09,10,12,13,14,15,16,16A,17,17A,17C 41874  7386 N (0.1763863 0.8236137) *
```

The tree that classifies the remaining seven categories of FOD3 (coded as 1 to 7) is

```
  1) root 56676 40673 5 (0 0.11 0.13 0.088 0.22 0.28 0.1 0.059)
    2) ROCCUP03=01,02,03,04,05,06,07,08,09,10,11,16,16A,17,17A,17B,17C,18,19,20,20A,20B,20C,20L,20S,20T 46062 33767 4
(0 0.13 0.16 0.098 0.27 0.15 0.12 0.061)
      4) ROCCUP03=04,05,09,10,11,17,17A,17B,17C,18,19,20,20A,20B,20C,20L,20S,20T 35227 23741 4 (0 0.17 0.11 0.07 0.33
0.19 0.061 0.078)
        8) ROCCUP03=05,17A 9601  5642 1 (0 0.41 0.046 0.063 0.13 0.28 0.016 0.058)
         16) POBREV=00,01,57,58,59,61,63,64,66,67,69 7189  4021 1 (0 0.44 0.05 0.057 0.15 0.22 0.017 0.061)
           32) PAOCF=1,2,3,4 2099  1091 1 (0 0.48 0.067 0.037 0.22 0.098 0.042 0.052) *
           33) PAOCF=0 5090  2930 1 (0 0.42 0.043 0.065 0.12 0.28 0.0067 0.064)
             66) HIDEG03=1 3572  1911 1 (0 0.47 0.038 0.045 0.11 0.26 0.0078 0.076) *
             67) HIDEG03=2,3 1518  1019 1 (0 0.33 0.057 0.11 0.14 0.32 0.004 0.037)
               134) IND21=I37t49,I467t579,I687t699,I77,I786t789,I866t869,I877t929,I937t987 363   217 1 (0 0.4 0.055 0.094 0.2 0.2
0.0028 0.047) *
               135) IND21=I107t399,I17t29,I407t459,I57t69,I607t639,I647t679,I727t749,I757t779,I797t847,I856t859 1155   742 5 (0 0.31
0.057 0.12 0.12 0.36 0.0043 0.034) *
         17) POBREV=60,65,68 2412  1343 5 (0 0.33 0.033 0.082 0.052 0.44 0.012 0.05)
           34) PAOCF=1,2,3,4 689   424 1 (0 0.38 0.064 0.086 0.097 0.29 0.03 0.048) *
           35) PAOCF=0 1723   854 5 (0 0.31 0.02 0.081 0.034 0.5 0.0046 0.051) *
        9) ROCCUP03=04,09,10,11,17,17B,17C,18,19,20,20A,20B,20C,20L,20S,20T 25626 15344 4 (0 0.082 0.13 0.072 0.4 0.15
0.078 0.085)
         18) ROCCUP03=09,10,17B,20B,20L 3970   847 4 (0 0.02 0.066 0.026 0.79 0.032 0.049 0.02) *
         19) ROCCUP03=04,11,17,17C,18,19,20,20A,20C,20S,20T 21656 14497 4 (0 0.093 0.14 0.081 0.33 0.18 0.084 0.097)
           38) PAOCF=1,2,3,4 8670  5130 4 (0 0.094 0.15 0.053 0.41 0.049 0.16 0.083)
```

76) ROCCUP03=04,11,17C,18,20,20A,20C,20S,20T 7712 4340 4 (0 0.091 0.14 0.051 0.44 0.048 0.18 0.056)
152) ROCCUP03=04,11,18,20,20A,20C,20S,20T 7247 3977 4 (0 0.096 0.14 0.052 0.45 0.047 0.15 0.059) *
153) ROCCUP03=17C 465 206 6 (0 0.019 0.086 0.028 0.22 0.073 0.56 0.017)
306) IND21=I107t399,I37t49,I407t459,I467t579,I57t69,I647t679,I727t749 47 20 5 (0 0.15 0.043 0.064 0.085 0.57 0.043 0.043) *
307) IND21=I786t789,I797t847,I937t987 418 161 6 (0 0.0048 0.091 0.024 0.23 0.017 0.61 0.014) *
77) ROCCUP03=17,19 958 668 7 (0 0.12 0.22 0.073 0.18 0.055 0.052 0.3)
154) IND21=I107t399,I17t29,I37t49,I467t579,I57t69,I607t639,I647t679,I687t699,I757t779,I786t789,I797t847,I877t929,I937t987 820 613 7 (0 0.13 0.25 0.076 0.19 0.044 0.056 0.25)
308) IND21=I107t399,I17t29,I467t579,I57t69,I607t639,I687t699,I757t779,I877t929 92 51 2 (0 0.043 0.45 0.15 0.043 0.23 0.054 0.033) *
309) IND21=I37t49,I647t679,I786t789,I797t847,I937t987 728 524 7 (0 0.15 0.22 0.066 0.21 0.021 0.056 0.28)
618) REGION=1,4 314 226 4 (0 0.15 0.22 0.096 0.28 0.019 0.048 0.18) *
619) REGION=,2,3 414 268 7 (0 0.14 0.22 0.043 0.15 0.022 0.063 0.35) *
155) IND21=I727t749,I77,I856t859 138 55 7 (0 0.014 0.087 0.058 0.087 0.12 0.029 0.6) *
39) PAOCF=0 12986 9367 4 (0 0.092 0.13 0.099 0.28 0.26 0.031 0.11)
78) IND21=I17t29,I407t459,I467t579,I607t639,I647t679,I687t699,I707t719,I757t779,I786t789,I797t847,I856t859,I866t869,I877t929,I937t987,unemploy 8623 5769 4 (0 0.1 0.15 0.1 0.33 0.18 0.038 0.093)
156) XAGEIN< 13.5 6874 4358 4 (0 0.099 0.16 0.093 0.37 0.14 0.038 0.099)
312) ROCCUP03=17,17C,18,19,20T 3322 2407 4 (0 0.11 0.18 0.13 0.28 0.14 0.046 0.12)
624) ROCCUP03=17,17C,18 2432 1740 4 (0 0.1 0.19 0.14 0.28 0.16 0.054 0.064)
1248) IND21=I17t29,I467t579,I687t699,I786t789,I797t847,I856t859,I866t869,I877t929 2144 1513 4 (0 0.1 0.2 0.15 0.29 0.13 0.057 0.062) *
1249) IND21=I407t459,I607t639,I647t679,I707t719,I757t779,I937t987 288 167 5 (0 0.083 0.097 0.08 0.21 0.42 0.028 0.08) *
625) ROCCUP03=19,20T 890 641 7 (0 0.13 0.16 0.096 0.25 0.056 0.024 0.28)
1250) GUMOS>=16976 394 269 4 (0 0.14 0.14 0.089 0.32 0.071 0.028 0.21) *
1251) GUMOS< 16976 496 328 7 (0 0.12 0.18 0.1 0.2 0.044 0.02 0.34) *
313) ROCCUP03=04,11,20,20A,20C,20S 3552 1951 4 (0 0.088 0.15 0.059 0.45 0.15 0.03 0.077)
626) XAGE< 57.5 2996 1561 4 (0 0.094 0.14 0.049 0.48 0.13 0.03 0.076)
1252) IND21=I17t29,unemploy 99 50 2 (0 0.081 0.49 0.051 0.27 0.04 0.02 0.04) *
1253) IND21=I407t459,I467t579,I607t639,I647t679,I687t699,I707t719,I757t779,I786t789,I797t847,I856t859,I866t869,I877t929,I937t987 2897 1489 4 (0 0.095 0.13 0.049 0.49 0.13 0.03 0.077) *
627) XAGE>=57.5 556 390 4 (0 0.058 0.16 0.11 0.3 0.26 0.034 0.086)
1254) IND21=I17t29,I786t789,I797t847,I856t859,I937t987 230 147 4 (0 0.061 0.18 0.13 0.36 0.12 0.013 0.13) *
1255) IND21=I407t459,I467t579,I607t639,I647t679,I687t699,I707t719,I757t779,I866t869,I877t929,unemploy 326 211 5 (0 0.055 0.14 0.092 0.25 0.35 0.049 0.052) *
157) XAGEIN>=13.5 1749 1156 5 (0 0.12 0.11 0.13 0.19 0.34 0.038 0.072) *
79) IND21=I107t399,I37t49,I57t69,I727t749,I77,notinuni 4363 2535 5 (0 0.07 0.093 0.096 0.18 0.42 0.017 0.13)
158) ROCCUP03=04,11,17,18,20,20A,20C,20S,20T 3869 2435 5 (0 0.075 0.1 0.097 0.2 0.37 0.018 0.14)
316) ROCCUP03=04,11,18,20,20A,20C,20S,20T 3384 2068 5 (0 0.083 0.11 0.099 0.21 0.39 0.02 0.09) *
317) ROCCUP03=17 485 241 7 (0 0.016 0.078 0.08 0.07 0.24 0.0082 0.5)
634) IND21=I107t399,I37t49,I57t69,I77 162 109 5 (0 0.049 0.15 0.17 0.093 0.33 0.025 0.19) *
635) IND21=I727t749 323 109 7 (0 0 0.043 0.034 0.059 0.2 0 0.66) *
159) ROCCUP03=17C 494 100 5 (0 0.036 0.014 0.089 0.016 0.8 0.002 0.045) *
5) ROCCUP03=01,02,03,06,07,08,16,16A 10835 7106 2 (0 0.015 0.34 0.19 0.075 0.042 0.33 0.007)
10) ROCCUP03=01,02,03,06,07,08 4355 2539 2 (0 0.02 0.42 0.37 0.053 0.078 0.051 0.0078)
20) ROCCUP03=06,07,08 1503 427 2 (0 0.0086 0.72 0.098 0.066 0.027 0.079 0.006) *
21) ROCCUP03=01,02,03 2852 1373 3 (0 0.026 0.26 0.52 0.047 0.1 0.036 0.0088)
42) ROCCUP03=01,02 1114 351 3 (0 0.02 0.17 0.68 0.016 0.065 0.034 0.0081) *
43) ROCCUP03=03 1738 1022 3 (0 0.03 0.32 0.41 0.066 0.13 0.037 0.0092)
86) PAOCF=1,2,3,4 538 293 2 (0 0.022 0.46 0.26 0.1 0.078 0.065 0.013) *
87) PAOCF=0 1200 625 3 (0 0.033 0.25 0.48 0.049 0.15 0.025 0.0075) *
11) ROCCUP03=16,16A 6480 3145 6 (0 0.011 0.3 0.066 0.089 0.018 0.51 0.0065)
22) ROCCUP03=16A 4201 2540 2 (0 0.014 0.4 0.094 0.081 0.022 0.39 0.0071)
44) IND21=I107t399,I57t69,I607t639,I687t699,I727t749,I757t779,I786t789,I797t847,I877t929,I937t987 3709 2122 2 (0 0.016 0.43 0.1 0.085 0.023 0.34 0.0075)
88) XAGEIN< 19.5 3046 1644 2 (0 0.017 0.46 0.11 0.095 0.024 0.29 0.0072)
176) XQWK>=47 1269 625 2 (0 0.02 0.51 0.13 0.089 0.04 0.2 0.0047) *
177) XQWK< 47 1777 1019 2 (0 0.015 0.43 0.085 0.098 0.013 0.35 0.009)
354) QMS=4,5 324 141 2 (0 0.012 0.56 0.062 0.11 0.0093 0.23 0.0093) *
355) QMS=1,2,3 1453 878 2 (0 0.015 0.4 0.09 0.095 0.014 0.38 0.0089)
710) XPERN>=90001 409 216 2 (0 0.017 0.47 0.12 0.13 0.0073 0.24 0.0049) *
711) XPERN< 90001 1044 590 6 (0 0.014 0.37 0.078 0.08 0.016 0.43 0.011) *
89) XAGEIN>=19.5 663 292 6 (0 0.011 0.28 0.078 0.044 0.02 0.56 0.009) *
45) IND21=I407t459,I467t579,I647t679 492 128 6 (0 0.002 0.15 0.045 0.051 0.0081 0.74 0.0041) *
23) ROCCUP03=16 2279 566 6 (0 0.0053 0.11 0.014 0.1 0.01 0.75 0.0053)

46) IND21=I17t29,I57t69,I727t749 97 45 2 (0 0.021 0.54 0.041 0.062 0.021 0.32 0) *
47)
IND21=I107t399,I407t459,I467t579,I607t639,I647t679,I687t699,I757t779,I786t789,I797t847,I856t859,I866t869,I877t929,I937t987
2182 500 6 (0 0.0046 0.092 0.013 0.1 0.0096 0.77 0.0055) *
3) ROCCUP03=12,13,14,15 10614 1691 5 (0 0.026 0.017 0.047 0.016 0.84 0.0038 0.049) *

Appendix 5. Classification tree for non-response

Tree that classifies response (coded as "Y") versus non-response (coded as "N")

```
  1) root 170624 72894 Y (0.4272201 0.5727799)
    2) DEMGRP03=1,2,4,7 69288 30938 N (0.5534869 0.4465131)
      4) XQYR2US>=1989.5 18338  6161 N (0.6640310 0.3359690) *
      5) XQYR2US< 1989.5 50950 24777 N (0.5136997 0.4863003)
       10) ROCCUP03=16,16A,20,20A,20B,20S 18934  7901 N (0.5827084 0.4172916)
         20) POV< 231.5 3446  1061 N (0.6921068 0.3078932) *
         21) POV>=231.5 15488  6840 N (0.5583678 0.4416322)
           42) XQYR2US>=1973.5 5301  2051 N (0.6130919 0.3869081) *
           43) XQYR2US< 1973.5 10187  4789 N (0.5298910 0.4701090)
             86) QREL=03,04,05,06,07,09,10,11,12,13,15,16,17,18,19,21,23 1340   472 N (0.6477612 0.3522388) *
             87) QREL=01,02,08 8847  4317 N (0.5120380 0.4879620)
              174) QCOW=0,1,6,7,9 6161  2849 N (0.5375751 0.4624249) *
              175) QCOW=2,3,4,5,8 2686  1218 Y (0.4534624 0.5465376) *
       11) ROCCUP03=01,02,03,04,05,06,07,08,09,10,11,12,13,14,15,17,17A,17B,17C,18,19,20C,20L,20T 32016 15140 Y
(0.4728886
                            0.5271114)
         22) FOCC=4,6,7 2212   755 N (0.6586799 0.3413201) *
         23) FOCC=0,1 29804 13683 Y (0.4590994 0.5409006)
           46) XAGE< 37.5 11247  5374 N (0.5221837 0.4778163)
             92) QMS=2,3,4,5 5473  2330 N (0.5742737 0.4257263) *
             93) QMS=1 5774  2730 Y (0.4728091 0.5271909)
              186) GUMOS>=342696 678   282 N (0.5840708 0.4159292) *
              187) GUMOS< 342696 5096  2334 Y (0.4580063 0.5419937)
                374) POV< 187.5 272   109 N (0.5992647 0.4007353) *
                375) POV>=187.5 4824  2171 Y (0.4500415 0.5499585) *
           47) XAGE>=37.5 18557  7810 Y (0.4208654 0.5791346)
             94) IND21=I17t29,I37t49,I407t459,I467t579,I607t639,I687t699,I707t719,I727t749,I757t779,I797t847,I856t859
                          ,I866t869, , I877t929 6777  3194 Y (0.4713000 0.5287000)
              188) POV< 500.5 2312  1088 N (0.5294118 0.4705882) *
              189) POV>=500.5 4465  1970 Y (0.4412094 0.5587906)
                378) XAGEIN>=25.5 1263   595 N (0.5288994 0.4711006) *
                379) XAGEIN< 25.5 3202  1302 Y (0.4066209 0.5933791) *
             95) IND21=I107t399,I57t69,I647t679,I77,I786t789,I937t987 11780  4616 Y (0.3918506 0.6081494) *
    3) DEMGRP03=3,5,6 101336 34544 Y (0.3408858 0.6591142)
      6) QMS=3,4,5 31321 13360 Y (0.4265509 0.5734491)
       12) FOCC=2,4,6,7 1954   700 N (0.6417605 0.3582395) *
       13) FOCC=0,1,3 29367 12106 Y (0.4122314 0.5877686)
         26) XAGE< 34.5 14151  6625 Y (0.4681648 0.5318352)
           52) ESR=2,3,4,6 1344   522 N (0.6116071 0.3883929) *
           53) ESR=1 12807  5803 Y (0.4531116 0.5468884)
            106) ROCCUP03=04,05,16A,20,20A,20B,20C,20S 6607  3248 Y (0.4915998 0.5084002)
              212) QCOW=1,3,6,7,8 5729  2824 N (0.5070693 0.4929307)
                424) DEMGRP03=3,5 1024   416 N (0.5937500 0.4062500) *
                425) DEMGRP03=6 4705  2297 Y (0.4882040 0.5117960)
                  850) IND21=I407t459,I467t579,I57t69,I607t639,I757t779,I77,I786t789,I866t869,I877t929 1511   683 N (0.5479815
                              0.4520185) *
                  851) IND21=I107t399,I17t29,I37t49,I647t679,I687t699,I707t719,I727t749,I797t847,I856t859,I937t987 3194  1469 Y
                              (0.4599249 0.5400751)
                   1702) QREL=03,04,05,06,11,16,17,18,19,21,23 1180   559 N (0.5262712 0.4737288) *
                   1703) QREL=01,08,12,13 2014   848 Y (0.4210526 0.5789474) *
              213) QCOW=2,4,5 878   343 Y (0.3906606 0.6093394) *
             107) ROCCUP03=01,02,03,06,07,08,09,10,11,12,13,14,15,16,17,17A,17B,17C,18,19,20L,20T 6200  2555 Y (0.4120968
                              0.5879032) *
         27) XAGE>=34.5 15216  5481 Y (0.3602129 0.6397871) *
      7) QMS=1,2 70015 21184 Y (0.3025637 0.6974363)
       14) FOCC=2,4,6,7 3954  1894 Y (0.4790086 0.5209914)
         28) ESR=1,2,4 2585  1214 N (0.5303675 0.4696325) *
         29) ESR=3,6 1369   523 Y (0.3820307 0.6179693) *
       15) FOCC=0,1,3 66061 19290 Y (0.2920028 0.7079972) *
```

Appendix 6. R Programs

R Batch file to build the non-response model, select final model and output SAS code to assign nodes to new cases:

```
load(".RData")
set.seed(982347120)
library(rpart)
resp.rpart<-
rpart(RESP~QINCTOT+POV+QREL+QSPANX+QMS+QSPEAK+ESR+QPOWCCS+QCOW+FOCC+DISABLE+PAOCF+DEMGRP
03+HIDEG03+GENDER03+AIAN+ASIAN+BLACK+NHPI+OTHER+WHITE+NEWRACE+NEWHISP+NEWCSAB+NEWDIS+XAGE+
XQWK+QCIT+QENG+ROCCUP03+XAGEIN+XPERN+XQYR2US+POBREV+REGION+DIV+IND21+SBSTR+GUMOS+TRACTMO
S,data=lf_fod,method='class',control=rpart.control(xval=10,minbucket=10,minsplit=100,cp=0,maxsurrogate=0,maxdepth=10))
resp.prune<-prune(resp.rpart,cp=cpvalue(resp.rpart))
library(mvpart)
sink("resp.rpart2sastree_all")
rpart2sastree_all(resp.prune)
sink()
quit(save="yes")
```

Note: This batch calls the following r-functions developed for this project:

5.1 R function that picks final tree: cpvalue:

```
  cpvalue
function(x) {
table<-x$cptable
indx<-table[,4]==min(table[,4])
# table[,4]-mean(table[indx,4]+table[indx,5])
sq<-seq(1:length(table[,4]))
tempdif<-table[,4]-mean(table[indx,4]+table[indx,5])
firstneg<-min(sq[tempdif<0])
x<-mean(table[indx,4]+table[indx,5])
y1<-table[firstneg-1,1]
x1<-table[firstneg-1,4]
y2<-table[firstneg,1]
x2<-table[firstneg,4]
y<-y1+((x-x1)*(y2-y1)/(x2-x1))
y
}
```

5.2 R function that creates SAS statements for assigning new case to nodes of the classification tree

- this is a modification of "rparttoSAS" which is freely available through the internet:
  http://www.biostat.wustl.edu/archives/html/s-news/2001-09/msg00033.html
- thanks to Rolando Rodreguez (SRD) for help on coding categorical variable node assignments

```
rpart2sastree_all
function(x, minlength=0, spaces=3, cp,
            digits=.Options$digits, ...) {
    if(!inherits(x, "rpart")) stop("Not legitimate rpart object")
    if (!is.null(x$frame$splits)) x <- rpconvert(x)  #help for old objects

    if (!missing(cp)) x <- prune.rpart(x, cp=cp)
    frame <- x$frame
    node <- as.numeric(row.names(frame))
    depth <- tree.depth(node)
    indent <- paste(rep(" ," spaces * 32), collapse = "")
    #32 is the maximal depth
    if(length(node) > 1) {
       indent <- substring(indent, 1, spaces * 1:max(depth))
       indent <- c("," indent[depth])
       }
    else stop("Tree has only 1 node")

    # this is the ending part of each line
    term <- ifelse(frame$var == "<leaf>," paste("then do; rpnode=," node, "; output; end;"),
          paste("then do;  rpnode=," node, "; output;" ))
    # the first part of the line
    z <- labels(x, digits=digits, minlength=minlength, ...)
    indx<- -grep("[<>],"z)
    if(length(indx)==0) indx<-1:length(z)
```

```
z[indx]<-paste(gsub(',',"',',"sub("'=,"" in (',"z[indx])),"'),"sep="")
z <- paste(indent, "if (," z, '') ," sep='')

# add in the "end" statements that go with each "do"
delta  <- -diff(c(depth,1))   #leftward movement of the indent
temp <- paste(rep("end;," max(delta)), collapse=' ')
endlist <- substring(temp, 1, 5*delta)
endstring <- ifelse(delta>0, paste("\n," indent, endlist, sep=''), "")
z <- paste(z, term, endstring, sep='')

cat(z[-1], sep = "\n")  # the -1 prevents listing the root node
return(invisible(x))
}
```