Comparison of Imputation Accuracy Based on Imputation Class Definitions for IPEDS Finance Survey

Marcus Berzofsky

Joint Statistical Meetings

Minneapolis, MN August 7-11, 2005

3040 Cornwallis Road ■ P.O. Box 12194 ■ Research Triangle Park, North Carolina, USA 27709
Phone 919-316-3752 Fax 919-541-5966 e-mail berzofsky@rti.org

Contents

- General background
- Overview of problem
- Methodology
  - Test if response is missing completely at random
  - Comparison of imputation class designs
  - Analysis of precision and accuracy of imputation classes
- Results & Conclusions

General Background

- IPEDS, Integrated Postsecondary Education Data System, is a set of annual surveys that institutions that participated in title IV student aid programs are required to complete
- The Finance survey is a required component that obtains information on an institution's assets as well as their revenue and expenditures
- Different survey form for Public, Private not-for-Profit and For-Profit institutions

General Background (cont.)

- Nonresponding institutions' data are imputed
- Survey contains over 200 variables that are simultaneously imputed (i.e. one donor institution) to maintain consistency across variables
- Missing data imputed by 3 different methods
- Will focus on nearest neighbor imputation method
  - Procedure used when nonrespondent institution does not have prior year data

Problem

- Two ways to form imputation classes:
  - A non-statistical imputation class using subject matter expertise
  - A statistical imputation class using information from principle component analysis (PCA) and Chi-square automatic detector indicator (CHAID)
- **Need to determine which way results in more accurate imputation**

Methodology

- Test if nonresponse is missing completely at random
  - Define universe of eligible institutions as those that responded to 2005 survey (N=5,084)
  - Define nonrespondent universe as those that responded to 2005 survey, but were nonrespondents in either 2002, 2003 or 2004 surveys (N=388)

Methodology (cont.)

- Select 500 sets of samples (n=49) from both sets of institutions
  - Random nonresponse samples selected via stratified SRS proportionally to the distribution of institutions by institutional level and control from among all eligible institutions
  - Nonrandom nonresponse samples selected by SRS among nonrespondent institutions
  - Sample size based on number of total nonrespondents to 2005 survey

Methodology (cont.)

- For each sample, removed respondent data and imputed using both the PCA/CHAID and the non-statistical imputation classes
- The nearest neighbor procedure was used to compute imputations
  - Uses distance formula to select donor institution from imputation class
- Regress the imputed value onto the respondent value for each variable using the no intercept model
    - IMPUTED = B*RESPONSE + e

Methodology (cont.)

- Compare distribution of beta estimates for each variable across all 500 samples and across all finance variables
  - Calculate MSE of beta estimates for each variable
    - $MSE(B_i) = (B_i - 1)^2 + V(B_i)$
  - Compare distribution of MSEs across all variables

Distribution of Eligible Population and Samples by Institutional Level and Control

Institutional Level and Control	% Dist. of Eligible Population (N=5,084)	% Dist. of Nonrespondent Universe (N=388)	Avg. % Random Sample Dist. (n=49)	Avg. % Nonrandom Sample Dist. (n=49)
Administrative unit only	0.00	0.00	0.00	0.00
Public, 4 year and above	11.15	1.29	11.14	1.31
Private not-for-profit, 4 year and above	28.40	15.98	28.41	16.04
Private for-profit, 4 year and above	4.45	2.58	4.44	2.57
Public, 2 year	18.17	10.05	18.17	10.10
Private not-for-profit, 2 year	2.68	4.38	2.69	4.08
Private for-profit, 2-year	11.53	16.24	11.50	16.05
Public, less than 2-year	3.11	7.73	3.20	7.76
Private not-for-profit, less than 2-year	1.24	1.55	1.19	1.54
Private for-profit, less than 2-year	19.28	40.21	19.27	40.55

Test of Missing Completely at Random

- Using a chi-square test of homogeneity, the distributions of the entire population and the nonrespondent population were not equal
  - Test Statistic = 215.36
  - P-value less than 0.0001

- Concluded that response is not missing completely at random

Non-statistical Imputation Classes

- Based on subject matter expertise of the data set
- Identified following variables for classes
  - Form used (3 different forms; Public, Private not-for-profit, Private for-profit)
  - Institutional level and Control
  - Medical school
  - Level of offering (first-professional, graduate, undergraduate)
  - Census division (for public institutions only)
- Minimum allowable size of class was 9 institutions
  - If class was less than 9 institutions, it was collapsed with closely related class
- 31 imputation classes created

PCA/CHAID Imputation Classes

- Selected summary variables from each form (about 20 variables across all three forms)
- PCA provides means to summarize all 200 variables into one single value
  - This allows pooling of institutions with similar responses across all variables
- PCA was applied to generate weight for each variable
- Index for each institution created by summing the product of weight*variable across all summary variables

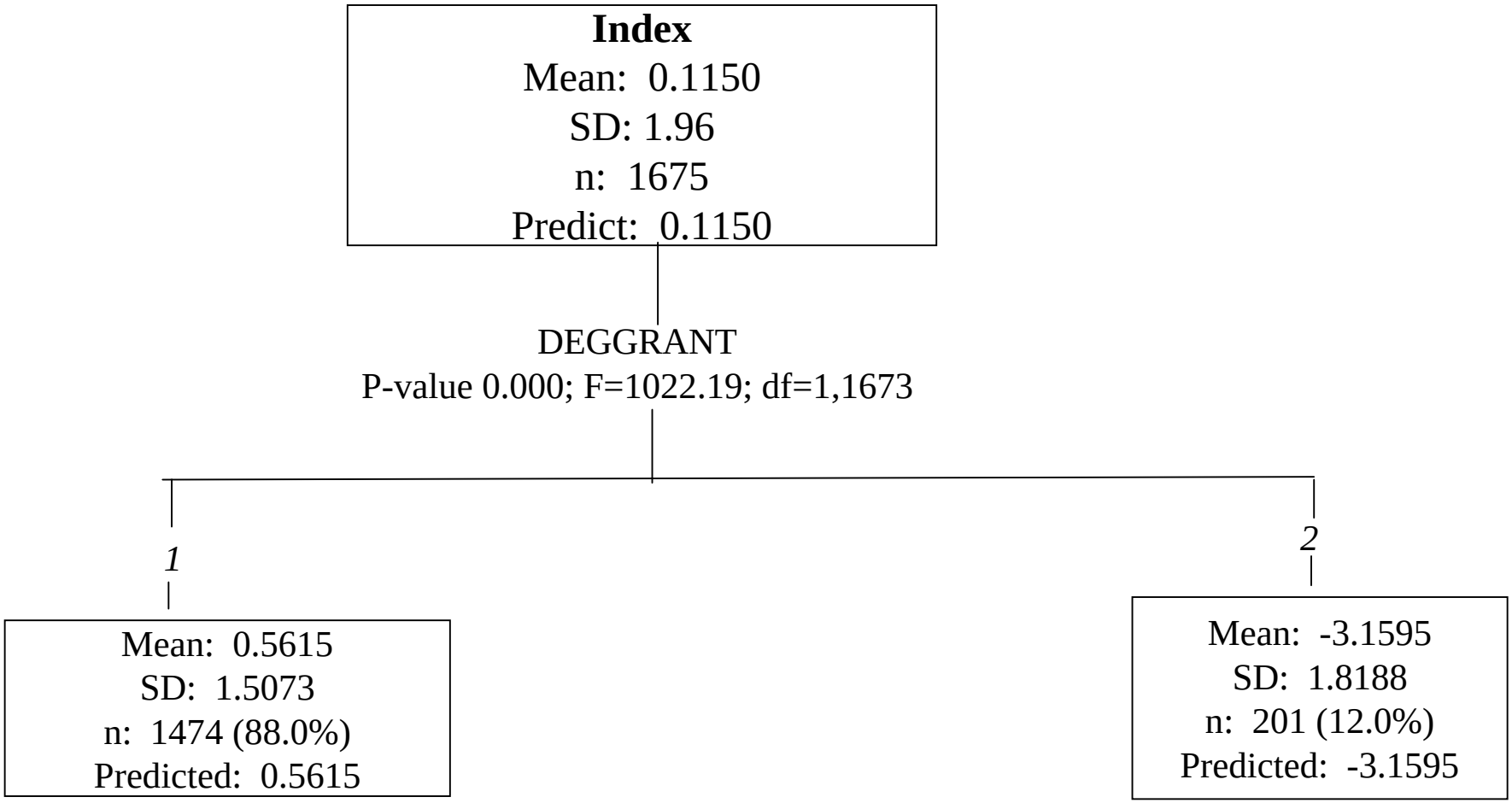
PCA/CHAID Imputation Classes (cont.)

- Answer Tree 2.0 software was used in generating imputation classes
- CHAID uses Chi-square or F statistics to identify optimal splits
- Index was used as outcome variable (i.e., institutions with similar index values would be grouped together)
- Predictor variables were determined by prior knowledge about the institutional characteristics. The significance of the variables on the finance data set were determined through regression analysis using the index variable as the dependent outcome

PCA/CHAID Imputation Classes (cont.)

- PCA/CHAID analysis was computed for the past three years of finance data. Variables consistently significant across all three years of data were used to create the final imputation classes
- Imputation classes were defined by common branches generated from three separate Decision Trees for each survey form
- Variables that were used in defining imputation classes include: degree granting status, institutional level and control, Offer graduate/first-professional, medical school, FTE groupings, student services and sports variables
- State (FIPS) was never part of the main branches and, therefore, not included in defining imputation groups
- Total of 45 imputation classes were created from Decision Tree

CHAID Tree Example



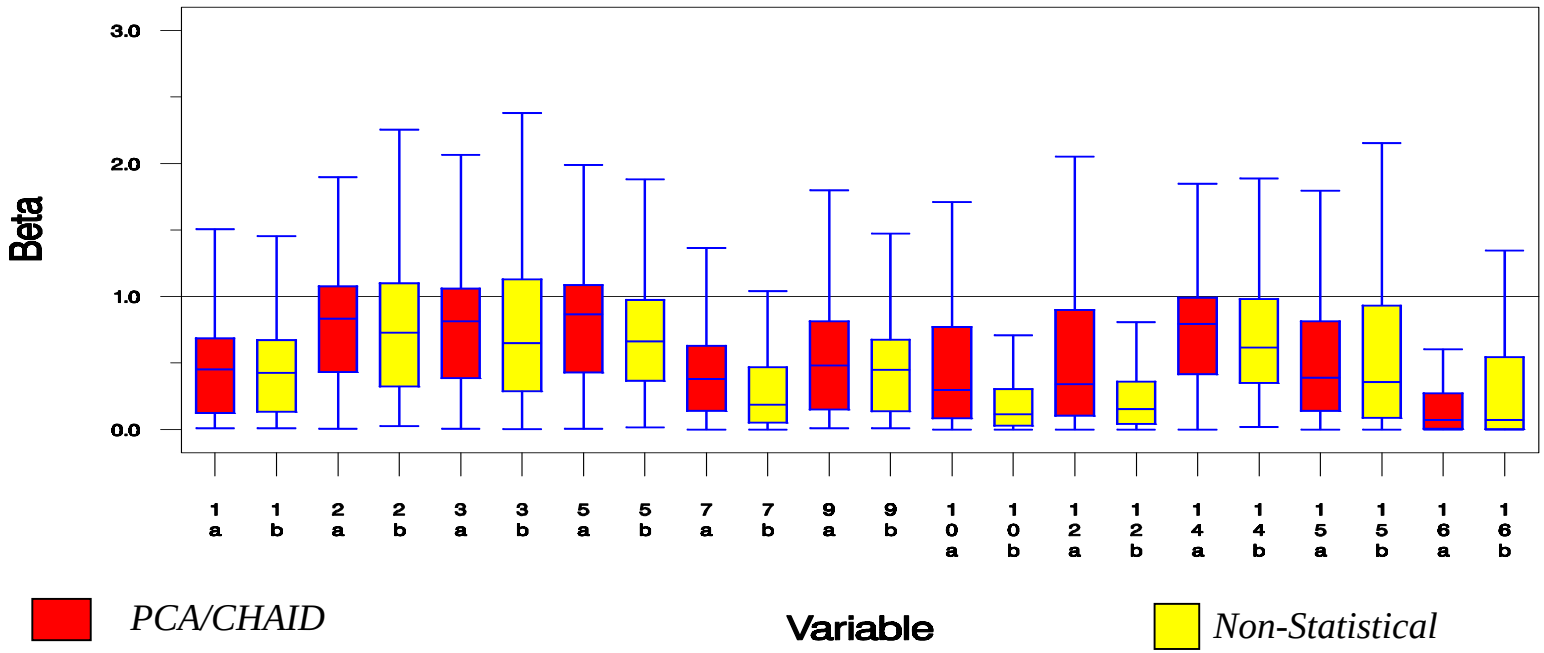
Analysis

Following tables based on results from nonrespondent set of samples

PCA/CHAID Method vs. Non–statistical Method

Public Institutions Net Assets Variables

Nonrandom Sample

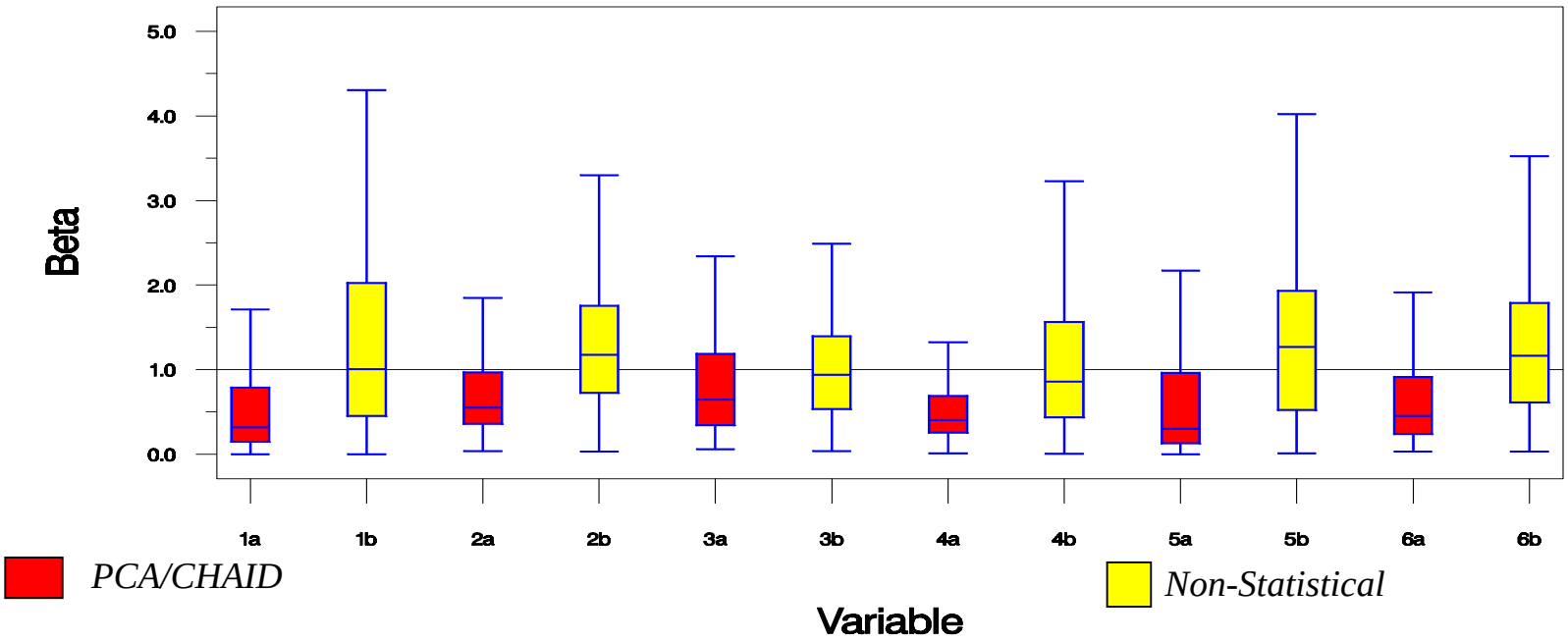


1=Current Assests	*2=Capital Assests*	*3=Accumulated Depreciation*	*5=Total Noncurrent Assets*
7=Long Term Debt, current portion	*9=Total Current Liabilities*	*10=Long Term Debt*	*12=Total Noncurrent Liabilities*
14=Invested in Capital Assets	*15=Restricted-expendable*	*16=Restricted-nonexpendable*	

PCA/CHAID Method vs. Non–statistical Method

Private Not–for–Profit Institutions Asset Variables

Nonrandom Sample

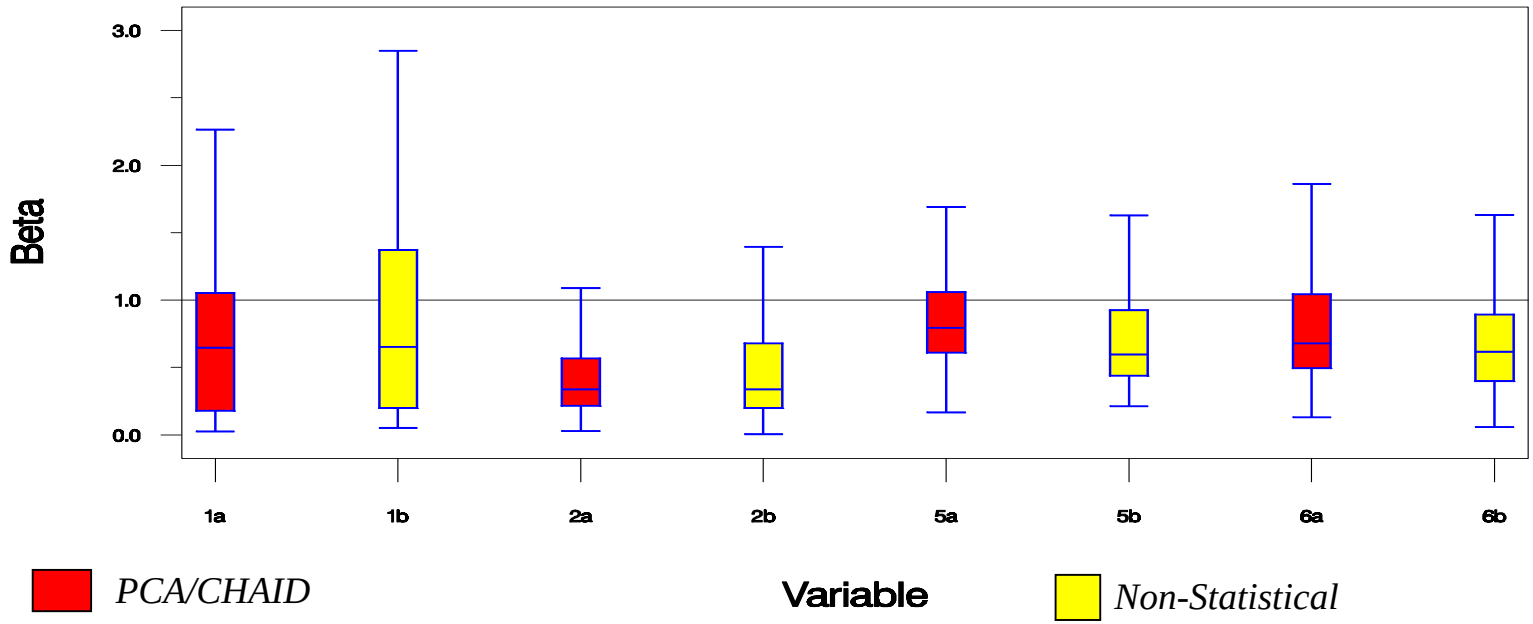


1=Long-term investment	*2=Total assets*	*3=Total liabilities*
4=Total unrestricted net assets	*5=Total restricted net assets*	*6=Total net assets*

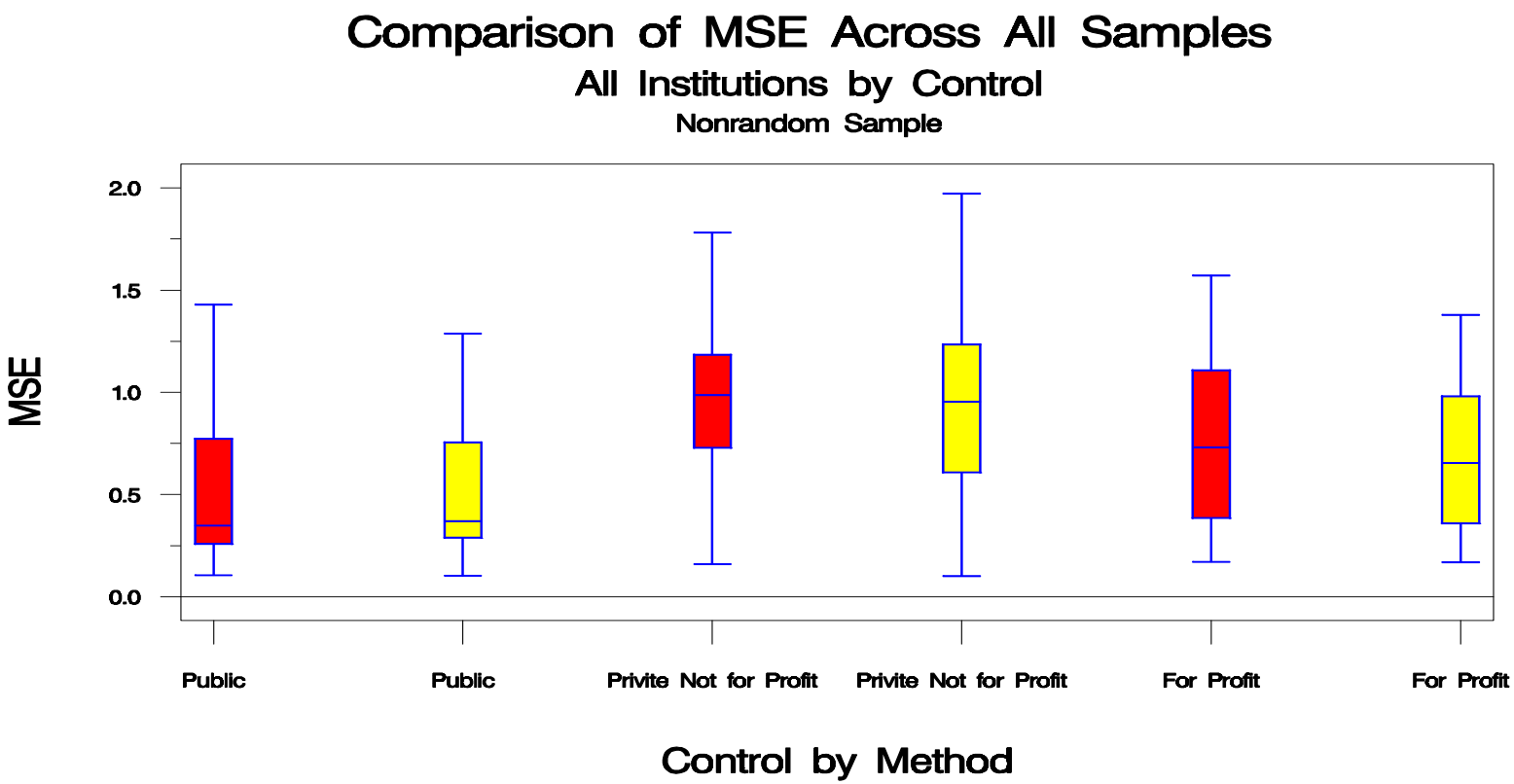
PCA/CHAID Method vs. Non–statistical Method

Private For–Profit Institutions Asset Variables

Nonrandom Sample



1=Total assets	*2=Total liabilities*	*5=Total revenues*	*6=Total expenses*

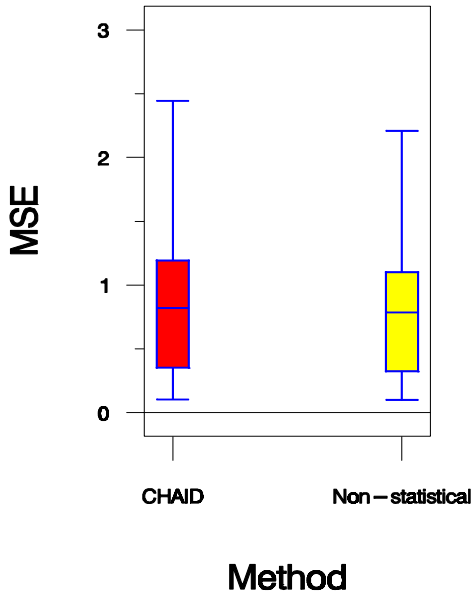


PCA/CHAID *Non-Statistical*

Comparison of MSE Across All Samples

All Institutions

Nonrandom Sample



PCA/CHAID

Non-Statistical

Distribution of MSE Across All Variables by Control and Imputation Class Method

	Public		Private not-for-Profit		Private for-Profit		Overall	
	PCA/ CHAID	Non-Stats	PCA/ CHAID	Non-Stats	PCA/ CHAID	Non-Stats	PCA/ CHAID	Non-Stats
Mean	46,201,054	10,108.86	37.12	1,571.54	5.67	7.49	22537115	5,570.8
Min	0.10408	0.10312	0.15912	0.09906	0.16982	0.16721	0.10408	0.46153
25th %	0.30496	0.33615	0.96128	0.92724	0.54107	0.55595	0.56407	0.46153
Median	**0.99953**	**0.92544**	**1.44954**	**1.32948**	**0.93824**	**0.75244**	**1.19236**	**1.03093**
75th %	23.2158	14.3219	10.5242	5.50132	1.57107	1.04083	10.9853	7.27908
Max	2.5 Billion	296,119.62	1,884.01	110,965	95.1646	152.71	2.5 Billion	296,119

Conclusions

- The PCA/CHAID method was hampered by the following limitations and constraints:
  - Require one donor institution per imputee for all variables
  - About 20 key summery finance variables were used in PCA to represent over 200 finance variables. These variables only explained 53% to 70% of the variation
  - Only kept common major branches in the decision tree due to year to year consistency issues

Conclusions (cont.)

- When specific knowledge about the data is known prior to imputation, it is possible to create imputation classes that generate as accurate imputations as the PCA/CHAID produce

- We recommend the non-statistical method because it is easier to implement and provides homogeneous imputation classes across years

Questions?