EVALUATION OF IMPUTATION METHODS FOR STATE EDUCATION FINANCE DATA

David Monaco and Stanley Weng, Synectics for Management Decisions, Inc., Frank Johnson, National Center for Education Statistics

David Monaco, Synectics for Management Decisions, Inc., 3030 Clarendon Blvd #305, Arlington VA 22201

KEY WORDS: Missing values, Imputation distributions, Percent error, Log transformation

INTRODUCTION

The purpose of this study is to identify, develop, and analyze appropriate methods for imputing missing data in the National Public Education Financial Survey (NPEFS), collected by the National Center for Education Statistics (NCES).

NPEFS is part of the Common Core of Data (CCD), a series of surveys collected annually from State education agencies. NPEFS provides detailed State-level information about revenues and expenditures for public elementary and secondary education. These data are used to allocate $7 billion in federal funds for education to the states, therefore all states submit data for this survey. The need for imputation is not to correct for non-reporting states, but to correct for missing items in the states' submissions. The goal is a complete dataset that is comparable across states.

Each state has a unique accounting structure for tracking revenues and expenditures for public education. Even in states following the most recent 1990 accounting handbook there are revenues or expenditures which are reported as aggregate amounts with other items. NCES works with states to improve reporting and have developed state specific software to crosswalk finance data from states' accounting systems to NCES's. However even with these efforts, imputation operations were required for 37 states for the FY 1992 collection.

In most cases these imputations were used to disaggregate a single value reported for two or more items. For example a state may not distinguish between student fees for transportation, textbooks, and summer school but only track student fees in general which they might report as student fees for transportation because state officials know that transportation fees are larger than book or summer school fees. NCES would then perform an imputation to disaggregate the reported single value and distribute it to the three separate student fee items. NCES performed 148 separate imputation operations for the FY 1992 collection, of which 129 involved disaggregating a reported value to two or more items. The remaining operations involved imputing values for items that states do not track.

This study looks at the two similar methods for imputing data that were developed by NCES (NCES I and NCES II) along with a variation of this method (NCES III). In addition, time series, regression, and nearest neighbor methods are discussed.

These methods were analyzed in order to determine the affects of each and to select one method as being "better" in disaggregating the data. The analysis focuses on the distribution of Revenues from Nonproperty Taxes (R1D) to Revenues from Tuition (R1F) and Summer School (R1N). This particular operation was chosen because of the variability of these revenues across states. Unlike expenditures for education where the proportions spent for salaries, instruction, etc. are fairly consistent across states, revenues for education come through a variety of revenue collecting activities.

IMPUTATION METHODS

NCES I Imputation Method

The NCES I method for distributing aggregate amounts is to calculate a ratio of each appropriate item in the distribution to the sum of the items in the distribution. For example, one state reports tuition fees (R1F) and summer school fees (R1N) as a Non-property tax (R1D), then the ratio of R1D to the sum of R1D + R1F + R1N is calculated for each state reporting both items. The ratios of R1F to the sum, and R1N to the sum are also calculated and then the average of each set of ratios across states is determined. This ratio is then used to disaggregate the reported amount.

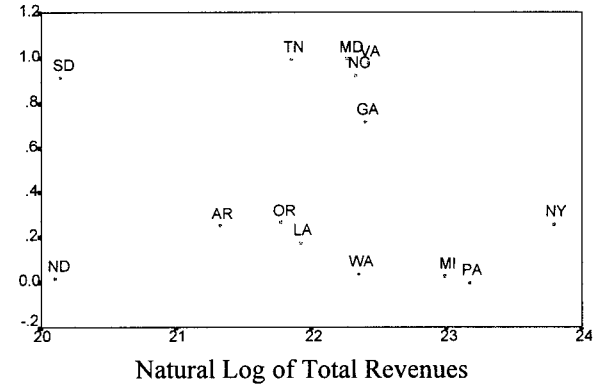
Table 1 demonstrates this method. State A is the state reporting the three revenues as R1D. States B through E etc. are the states whose reported amounts are used for the imputation. The R1D ratio for state B is 18.0/(18.0+3.2+3.9). The average ratio is for all of the states used in the imputation, of which only four are shown in the table. The average ratio times the amount reported for R1D yields the imputed amounts for each of the three variables (at the bottom of the table.)

Table 1. NCES I Method (amounts in $ millions)

State	R1D	R1F	R1N	R1D Ratio	R1F Ratio	R1N Ratio
State A	302.8	---	---			
State B	18.0	3.2	3.9	0.72	0.13	0.16
State C	1,069.1	3.2	2.5	0.99	0.01	0.00
State D	55.1	156.3	2.5	0.26	0.73	0.01
State E	500.5	1.1	2.3	0.99	0.00	0.06
ect.						
Average Ratio	---	---	---	0.42	0.51	0.06
State A	R1D	R1F	R1N			
Imputed	127.90	155.50	19.40			

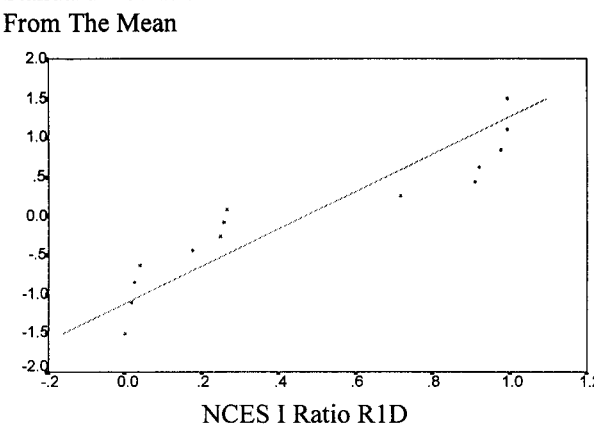
The ratio containing R1D is plotted against the natural logarithm of Total Revenues for every state. This gives us an indication of the characteristics of this model and the weight that the variable R1D has on the imputation involving the three variables (R1D, R1F and R1N). The plot for Method I is presented in Figure 1. The ratio plotted on the Y axis is that of R1D/(R1D+R1F+R1N), and the natural logarithm of Total Revenues is plotted on the x axis. Two groups of states are apparent, one group of six states where the ratio of R1D to the sum of the three variables is between .70 and 1.00 and another group where the ratio is .30 or less.

Figure 1. NCES I Method: Plot of R1D Ratio to Log of Total Revenues



The large gap that exists between the two groups of states would indicate that an average of these ratios would not be representative of the data for either of the two groupings. This conclusion is supported further by a normal probability plot (Figure 2), where the ratios are arranged in increasing order of magnitude and then plotted against normal distributed values.

Figure 2. NCES I Method: Normal probability plot of R1D ratio



If the data are from a normal distribution, this plot will resemble a straight line. The state with the lowest R1D value is approximately 1.5 standard deviations below the mean. The state reporting a slightly higher value for R1D is found to be slightly higher than 1 standard deviation below the mean, and so on. The resulting plot is curved. The departure of the data points from the straight line exhibits the departure of the data from normality.

NCES II Imputation Method

NCES II method was developed as an improvement over NCES I, but is very similar. We will use the same example where State A reports the value for R1F and R1N aggregated in the value reported for R1D. This time the ratios calculated are of each value divided by total revenues (TR). (If the items were expenditures the ratios would be calculated with total expenditures as the denominator.) Only states reporting values greater than 0 for each of the 3 revenues are used in the operation. States in which any of the 3 revenues are changed by other imputations are excluded from the operation. The average of these ratios is calculated, and then the relative distribution of these averages is determined. This distribution is then used to disaggregate the reported revenue amount.

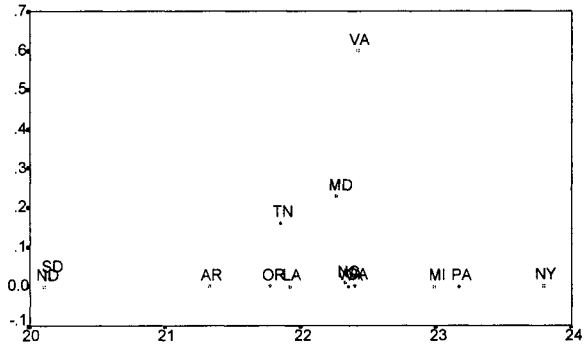
Table 2. NCES II Method (amounts in $ million's)

State	R1D	R1F	R1N	TR	R1D Ratio	R1F Ratio	R1N Ratio
State A	302.8	---	---				
State B	18.0	3.2	3.9	5,332	0.00	0.00	0.00
State C	1,069.1	3.2	2.5	4,692	0.23	0.00	0.00
State D	55.1	156.3	2.5	21,574	0.00	0.01	0.00
State E	500.5	1.1	2.3	3,094	0.16	0.00	0.00
ect.							
Average Ratio	---	---	---	---	0.04	0.01	0.00
Percent distribution of avg. ratios					0.85	0.15	0.01
State A	R1D	R1F	R1N				
Imputed	255.8	44.5	2.5				

The plot of R1D/TR (NCES II ratios) by the natural logarithm of TR is presented in Figure 3. This plot shows most points are scattered about a horizontal level with a few outliers. The average ratio would shift from that stable level and therefore not represent the majority of the ratios.

Figure 3. NCES II Method: Plot of R1D ratio vs. Log of Total Revenues

NCES II Ratio R1D

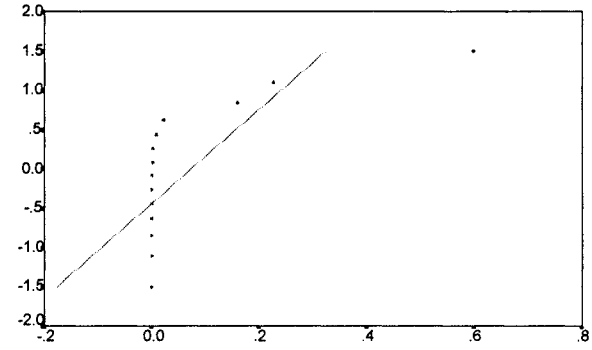


Natural Log of Total Revenues

In addition, the normal probability plot for the NCES II ratios for R1D shows a curve differing from the expected straight line, indicating the data are not normally distributed. (Figure 4)

Figure 4. NCES II Method, Normal probability plot of R1D ratio

Standard Deviations
From The Mean



NCES II R1D Ratio

NCES III Method

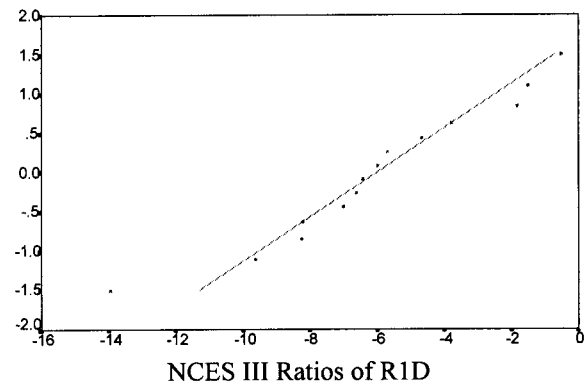
A variation of Method II was designed for this analysis. This method calculates the natural logarithms of the ratios (of item to total revenues (or expenditures)). The average of these logs is computed, and then the natural exponent of the average is determined. The distribution of these exponents is calculated, and the resulting values are used to distribute the aggregated amount. The log transformation of the ratios should stabilize the variance of the ratios. An example of the NCES III method is shown in Table 3. Note that the averages are calculated from more data than are shown.

Table 3. NCES III Method (amounts in $ million's)

State	R1D	R1F	R1N	TR	R1D Ratio	R1F Ratio	R1N Ratio
State A	302.8	---	---				
State B	18.0	3.2	3.9	5,332	0.00	0.00	0.00
State C	1,069.1	3.2	2.5	4,692	0.23	0.00	0.00
State D	55.1	156.3	2.5	21,574	0.00	0.01	0.00
State E	500.5	1.1	2.3	3,094	0.16	0.00	0.00
ect.							
*	*	*	*	*	*	*	*
					Log of R1D Ratio	Log of R1F Ratio	Log of R1N Ratio
State B					-5.6890	-7.4261	-7.2080
State C					-1.4791	-7.2886	-7.5391
State D					-5.9701	-4.9272	-9.0689
State E					-1.8214	-7.9184	-7.2218
ect.							
Average					-6.4080	-6.0309	-7.9524
Natural exponent (of average Log)					0.0016	0.0024	0.0004
Distribution					0.4743	0.5458	0.0799
State A	R1D	R1F	R1N				
Imputed	255.8	44.5	2.5				

A normal probability plot of the logs of the ratios is presented in Figure 5. This figure demonstrates that random discrepancy and normality is significantly improved with the log-transformation model.

Figure 5. NCES III Method, Normal Probability plot of R1D ratio



ADDITIONAL METHODS EXPLORED

The following sections cover Time Series, Regression, and the Nearest Neighbor methods which, after initial exploration, were found not to be suitable candidates for CCD Finance Data Imputations.

Time Series Method

The problem encountered using time series is that there is not enough data to get good diagnostic plots which are critical in determining which model should be fit. At present there are only 4 years of CCD Finance Data were available to fit a model and at least 6 to 8 more years are needed in order to determine what model should be fit.

Regression Method

In employing the regression method, individual regression relationships need to be identified for each variable to be imputed and the auxiliary variable have to be identified. These variables in turn may have to be imputed. In addition, the imputed values for the missing components of an aggregate, provided by separate regressions, would not sum up to the reported value of the aggregate. Though, seemingly, a proportional adjustment can be taken to the imputed values to make their sum matching the aggregate value, the validity of such adjustment is in question.

Nearest Neighbor Method

The Nearest Neighbor method uses the financial data to group States in order to apply separate imputation distributions. Each group of States would have its own imputation distribution. As recognized on the Original ratio plot for R1D for the fiscal year 1992 data, two clusters of points appear. (Figure 1) This pattern displays a classification of the States.

The Nearest Neighbor method incorporates this information of classification into the imputation operation. The reporting States are grouped into two classes, and imputation distribution would be found for each class of States. When imputing for a missing value, the imputing State's class needs to be identified before applying the corresponding imputation distribution.

This model has been rejected in the past because of the difficulty in determining the class of states. In addition, for some survey items, there are only a small number of states which have the specific revenues or expenditures for which we are imputing. Dividing this small number of observations (states) into groups results in too small a grouping upon which an imputation can be based.

ANALYSIS AND RESULTS

The selection of the best imputation method depends on the uses to which the data are to be put. For each of the imputation methods described in the previous sections, groups of variables of importance to NPEFS were evaluated across fiscal years 1989 through 1992. The objective used in the evaluation was to minimize the average percent error across the largest set of variables of interest. Percent error is defined as the absolute difference of the reported value from the imputed value, divided by the reported value.

Three groups of variables are used in the evaluation of three NCES methods using data for fiscal years 1989-1992. Group 1 is a small group of revenue variables which consisted of R1D, R1F, and R1N (which were used in demonstrating the methods). Group 2 is a larger group of revenue variables. Group 3 consists of expenditures for Food Services.

For Group 1 variables, the NCESII method performed best across all years, yielding the smallest average percentage error as highlighted in Table 4.

Table 4. Results of Analysis using group 1 variables

Method	Year	Percent Error R1D	Percent Error R1F	Percent Error R1N	Average Percent Error
NCES I	1989	3.69	40.65	3.19	15.84
NCES II	**1989**	**10.35**	**14.06**	**0.74**	**8.38**
NCES III	1989	3.71	40.59	3.10	15.80
NCES I	1990	3.04	32.07	15.05	16.72
NCES II	**1990**	**7.84**	**5.93**	**0.80**	**4.68**
NCES III	1990	3.94	29.12	6.63	13.23
NCES I	1991	3.20	34.18	12.12	16.50
NCES II	**1991**	**5.73**	**6.82**	**0.86**	**4.47**
NCES III	1991	4.29	22.45	6.50	11.08
NCES I	1992	4.90	26.97	10.95	14.27
NCES II	**1992**	**8.99**	**5.44**	**0.88**	**5.10**
NCES III	1992	6.14	20.87	6.56	11.19

Similar analysis was performed for Group 2 and Group 3 variables. The resulting average percent errors from this analysis and from the Group 1 analysis is presented in Table 5.

Table 5. Summary of analysis using groups 1, 2 and 3

		Average Percent Error		
Method	Year	Group 1	Group 2	Group 3
NCES I	1989	15.84	274.49	---
NCES II	1989	**8.38**	170.89	---
NCES III	1989	15.80	**36.56**	---
NCES I	1990	16.72	10.17	---
NCES II	1990	**4.68**	7.16	---
NCES III	1990	13.23	**2.34**	---
NCES I	1991	16.50	15.11	---
NCES II	1991	4.47	10.78	---
NCES III	1991	11.80	**5.65**	---
NCES I	1992	14.27	2.24	2.25
NCES II	1992	**5.10**	2.48	2.27
NCES III	1992	11.19	**1.35**	**1.28**

Conclusion

The NCES III method of imputation appears to be the best method for imputing data for the NPEFS survey. It does better a majority of the time and always does better than the other methods for larger groups of variables. The overall average percent error is the smallest using the NCES III method for the majority of the variable groups considered. The logarithmic transformation works to minimize the amount of variability encountered in the data.

References

National Center for Education Statistics. (1995). *The National Education Finance Survey Booklet.* Washington DC.

National Center for Education Statistics. (1995). *Statistics in Brief: Revenues and Expenditures for Public Elementary and Secondary Education: School Year 1992-93* . Washington DC.

Weisburg, S. (1985). *Applied Linear Regression. 2nd ed.* New York: John Wiley.