MEASURES OF GOODNESS OF FIT FOR EXTRAPOLATIONS: INITIAL RESULTS USING THE INDIVIDUAL TAX MODEL DATABASE

by

Robert E. Gillette
Office of Tax Analysis
U.S. Treasury Department

OTA Paper 62 February 1989

OTA Papers and Briefs are circulated so that the preliminary findings of tax research conducted by staff members and others associated with the Office of Tax Analysis may reach a wider audience. The views expressed are those of the authors, and do not reflect Treasury policy. Comments are invited, but OTA Paper and Briefs should not be quoted without permission from the authors. Additional copies of this publication may be purchased from the National Technical Information Service 5285 Port Royal Road, Springfield, VA 221161. Phone (703) 487-4660.

Office of Tax Analysis
U.S. Treasury Department, Room 4040
Washington, DC 2022

ABSTRACT

Recent experiences during Tax Reform suggest that estimates of the impact of various tax proposals are quite sensitive to the extrapolation of the tax data to future years. In this paper various aspects of the extrapolation process are examined. We discuss several alternative extrapolation methods and describe the extrapolation procedure currently used by the Office of Tax Analysis. A set of statistics for evaluating the accuracy of an extrapolation are developed and then used to compare several extrapolations of the 1981 Individual Tax Model Database to 1983 levels.

MEASURES OF GOODNESS OF FIT FOR EXTRAPOLATIONS: INITIAL RESULTS USING THE INDIVIDUAL TAX MODEL DATABASE

Contents

Page

Tables:

Figures:

MEASURES OF GOODNESS OF FIT FOR EXTRAPOLATIONS: INITIAL RESULTS USING THE INDIVIDUAL TAX MODEL DATABASE

I. INTRODUCTION

During the course of tax reform, there was a great deal of interest in how revenue estimates for future tax years were obtained. The estimated tax impact of many of the tax provisions considered during tax reform were found to be sensitive to the extrapolation of current and prior year tax return data to reflect the years being considered in the reform proposal. Thus, the accuracy of the extrapolation process used in conjunction with the Treasury's Office of Tax Analysis (OTA) Individual Tax Model Database has become increasingly important (for a description of the Tax Model, see Wyscarver and Cilke [1987]).

The purpose of this paper is to present some initial results of a study of the extrapolation of databases used in microeconomic simulations. The eventual goal is to develop an improved extrapolation procedure that can generate a five year panel file for both analytic and revenue use. The initial results, however, are more limited. The results presented in this paper are based on a set of test extrapolations of the Treasury's 1981 Individual Tax Model Database to 1983 levels. The year 1983 was chosen as the target year for the initial extrapolation test since it was the most recent year for which extensive IRS Statistics of Income (SOI) information was available when the project began.

The body of the paper is divided into four sections. In the first section of the paper, we review the various techniques that have been used to extrapolate tax databases. In the second section, we present some statistical measures that may be used to examine the success or 'fit' of an extrapolation. In the third section, we discuss the extrapolations made to adjust or 'age' the Individual Tax Model Database to 1983 levels. Finally in the fourth section, we discuss some of the conclusions that can be drawn from these initial extrapolation tests and suggest some areas for future research.

II. METHODS OF EXTRAPOLATION

In general. there are two approaches by which a tax database can be extrapolated. The database can be reweighted so that the individual variables on each tax return are adjusted together in order to achieve some set of aggregate and/or distributional targets. Reweighting changes the number of tax returns represented by each return in the database. Alternatively. the individual variables on the database can be adjusted separately to match a set of targets. Each variable can be multiplied by some constant adjustment factor for all the returns in the database.

Most of OTA's extrapolation procedures. including the one used for most recent extrapolations (i.e.. the extrapolation creating the tax reform database in 1985) have emphasized reweighting each record as the best procedure for adjusting the database. For example. in the extrapolation from 1981 to 1983. a single growth factor (the Consumer Price Index) was applied to all income items except the itemized deductions (which were adjusted separately). All of the other adjustments to the database were made by re-weighting the individual records. For a more complete description of the current extrapolation procedures used by OTA. see Wyscarver and Cilke [1987].

Extrapolation procedures that emphasize the reweighting approach derive from a concern for preserving the actual information reported on a taxpayer's return. Maintaining the integrity of a return is important since there exist significant correlations between the various items reported on a return that should be preserved. Therefore. each return should be treated as a unit. rather than as a group of independent items.

In order to maintain the relationships of the variables within a given return. any change in the levels reported on a return should only be adjusted through the application of a uniform growth factor to all of the items on the return. Of course. the use of a uniform growth rate does not allow the composition (i.e., the relative mix) of the various items on the database as a whole to change. Adjustment of the composition of the database must be done through modification of the weights on the various returns found on the database.

Lindsey (1985). however. argues that extrapolation procedures which are based on the reweighting approach have two fundamental flaws. First. the use of a uniform growth factor for all income items does not take into account possible changes in the functional distribution of income (i.e.. changes in the relative mix of the various income components due to changes in the economy. such as an increase in the rate of return). Second. making changes in the functional distribution of income through altering the database's sample weights implicitly assumes that all changes in the relative values of the components of income are caused by a change in the number of returns containing that component of income, rather than a change in the level of that income component reported on the returns (i.e., the mean value of the income component remains constant).

In order to correct for these problems. Lindsey developed an extrapolation process for the National Bureau of Economic Research (NBER) TAXSIM model's database that emphasizes individual adjustments to the levels of the various items on the database.[1] In particular Lindsey suggests that a better estimate of the change in the income distribution resulting from a change in a component of income (such as capital gains) would be obtained through an increase (or decrease as the case may be) of the level of that income component for the existing recipients rather than through an increase (decrease) in the number of the recipients.

Lindsey's extrapolation method is quite similar to the extrapolation procedure used by OTA in the mid-1970's. This earlier version of the OTA extrapolation process allowed for individual adjustment factors to be applied to the various components of income prior to any reweighting of the file. Thus. the emphasis in the earlier OTA extrapolations seems to have been on individual adjustment of the income variables. rather than on reweighting records. In recent extrapolations. OTA choose to emphasize reweighting. and the multiple adjustment factors were eliminated.[2]

The differences between the reweighting approach and the individual adjustment approach derives from different views of portfolio adjustment.[3] The assumptions behind the individual adjustment approach suggest that. in the short run at least, an individual's portfolio is inelastic with respect to changes in the relative rates of return for the various income components. In other words. by not allowing new recipients of a particular income component to appear on the extrapolated database, the individual adjustment approach assumes that an individual cannot or will not adjust his or her portfolio in response to changes in the rates of return on alternative assets.

This assumption is in sharp contrast to the extremely responsive portfolio behavior implied by the reweighting approach. If all aggregate changes on the tax database result from adjustments in the number of taxpayers with different portfolios found on the database. then no change in the relative rates of return for the various income components is assumed. This implies that any economic circumstance that might change the rate of return of a given component of income will be compensated for through an adjustment in the taxpayers' portfolios. Thus, by not allowing the relationships between the various components of income on a record to change, the reweighting approach assumes that all taxpayer's portfolios are completely flexible.

Currently, however, there is no means for choosing between the assumption of flexible portfolios and the assumption that, in the short run at least, portfolios are inflexible. In fact, it seems more reasonable to assume that, given the differences in the liquidity of assets composing the portfolios, the portfolios are in the 'short run' more flexible with respect to some components of income (e.g., capital gains and dividends) and less flexible with respect to other income components (e.g., wages and pensions). In addition it seems reasonable to assume that the flexibility of a taxpayer's portfolio will vary with the taxpayer's place in his (or her) life cycle. Therefore, an accurate extrapolation procedure would need to combine the use of individual adjustment factors with the reweighting approach.

OTA currently implements an extrapolation routine that combines the reweighting approach and the individual adjustment approach. Specifically, OTA's most recent extrapolation procedure operates in two stages. In the first stage, the levels of various components of income (such as wages and salaries, capital gains, interest, dividends, pensions, and business income) are independently adjusted to hit various macro-targets. In the second stage, the weights on the database are adjusted to achieve other aggregate targets (such as the two earner deduction, the Foreign Tax Credit [FTC], the Investment Tax Credit [ITC], itemized deductions, and the distribution of Adjusted Gross Income [AGI]. By using various combinations of first and second stage targets, it is possible create a large, if not infinite, number of extrapolations for a given set of targets. These extrapolations range from extrapolation based solely upon reweighting to extrapolation using only adjustment factors.

The determination of the best mix of individual adjustment and reweighting, however, requires an examination of the accuracy of the different extrapolations with respect to the target values. This examination requires some means for comparing the fit of the different extrapolations being considered. The next section describes some possible measures of the 'goodness of fit' of an extrapolation.

III. MEASURES OF GOODNESS OF FIT

A key step in the development of some means for evaluating alternative extrapolations is constructing some measure of how well the adjusted database fits the actual targets. In this initial study the targets are provided by the 1983 SOI. Since we were concerned with both the aggregate values of selected variables and the distribution of the variables by AGI class, it was necessary to have 'goodness of fit' measures that considered both the aggregate value as well as the distribution of the variables. Further, each extrapolation consists of many variables, so it was desirable to have a measure that could be used to compare both individual variables and the extrapolation as a whole (e.g., something like T-statistics and F-statistics in linear regression equations).

Clearly, no one measure could satisfy all of these requirements. Therefore, three separate measures for the fit of an extrapolation were developed. Two of these measures examine the fit of individual variables: the percent error and the information gain (for the aggregate value and the distribution, respectively). The third measure examines the fit of the extrapolation as a whole: the multiplicative decomposition of the variance.

A Measure of the Fit of a Value

There are any number of possible mechanisms for measuring the fit of a value. One obvious measure of the error in the predicted value of a variable is the percent error:

$$(1) \qquad PE(B) = (\beta\text{-}b)/b$$

where β is the estimated value of some variable B and b is the true value. The percent error has several desirable properties. The most important of these properties is that the percent error is a relative measure of the deviation from the true value and so can be used to measure the relative accuracy of the various extrapolated variables.

A Measure of the Fit of a Distribution

The most commonly used non-parametric statistic for comparing two distributions is the Kolmogorov-Smirnov two sample test. Unfortunately, the Kolmogorov-Smirnov test compares the empirical cumulative distribution functions of the two distributions being considered. Since we are interested in a variable's distribution by AGI class, the Kolmogorov-Smirnov test is inappropriate. Further, the standard alternative, the Chi-squared test, is not invariant to changes in the scale of measurement. Therefore an alternative measure, the information gain, was chosen (see Theil [1967]).

The information gain measure is derived from information theory which examines the relationship between probabilities and events. For example, suppose that some event E_i is expected to occur with some probability p_i. If at some latter point in time a message is received that states that event E_i has occurred, then some amount of information has been received. Intuitively, the amount of information that is received from such a message is inversely related to the probability of E_i's occurrence (i.e., the more probable an event is, the less information that is obtained from that event's occurrence). To formalize this relationship, let:

$$h(p_i) = -\ln(p_i) \qquad 0 \leq p_i \leq 1 \tag{2}$$

where $h(p_i)$ is a measure of the amount of information generated by knowledge of the event and $\ln(p_i)$ is the natural log of probability p_i.

Next, suppose there are N possible events, that collectively exhaust the outcome space (i.e., one of the N events must occur). Then, prior to actually receiving a message the expected information inherent in the message is:

$$H(p) = \sum_{i=1}^{N} -p_i \ln(p_i) \tag{3}$$

where

$$\sum_{i=1}^{N} p_i = 1 \tag{4}$$

and H(p) is the expected information or entropy measure. Note that if the empirical probabilities, p_i, are calculated as the share of the total value of a given variable (v_j) that falls into a given AGI class (this limits the test to variables whose values have the same sign for each AGI class), then the distribution of a variable by AGI class can be viewed as an algorithm for allocating expenditures into different classes and therefore can be summarized using the entropy measure. Two distributions can be compared by examining the expected change in information that results from moving from one distribution to the other. Let,

$$(5) \qquad h(x_i) - h(y_i) = -\ln(x_i/y_i) \qquad x_i . y_i > 0$$

where x_i is the prior probability (1983 SOI's i-th AGI class share of v_j) and y_i is the posterior probability (the adjusted Tax Model Database's i-th AGI class share of v_j). Then the expected information gain can be measured by

$$(6) \qquad I(y:x) = \sum_{i=1}^{N} -y_i \ln(x_i/y_i)$$

where

$$(7) \qquad \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i = 1 \text{ and } x_i . y_i \geq 0.$$

The information gain measure is a useful tool for examining the differences between the shapes of distributions.[4] Since the calculations are all based on shares (technically, allocation probabilities) the measure is completely independent of the unit of measurement or the aggregate value of the variable. The only requirement is that the number of cells into which the distributions are divided be identical. Thus, the information gain is a very powerful tool for examining broad questions about distributions for it can be used to compare the distribution of any two variables (e.g., one could examine how closely the distribution of dividends by AGI resembles the distribution of interest income).[5]

A Measure of the Overall Fit of an Extrapolation

Clearly in addition to having measures of how well individual variables are predicted by an extrapolation. it is desirable to have a measure of the overall fit of an extrapolation. Most common measures of the fit of a predictor are based on the variance. Unfortunately none of standard measures are applicable. Therefore. we have modified a measure suggested by Theil (1967).

In developing a general measure of the fit of an extrapolation we shall make use of several basic definitions. Any given extrapolation t (or, more generally, any predictor t where $t=1,\ldots,\tau$) generates an IxJ matrix of variables, V_t, where the ij-th element of V_t represents the t-th extrapolation's prediction of the i-th variable's value in the j-th AGI class. Now, let X_t be the natural log of V_t, where the natural log is taken element by element. Then, assuming that

(8) $$E(X_t - X_\tau) = 0$$

where X_τ is a matrix of the natural logs of the true values, the error variance of X_t can be written as:

(9) $$E((X_t - X_\tau)(X_t - X_\tau)^T) = \Theta_t:$$

and the variance for each element of X_t can be written as,

(10) $$E((x_{ijt} - x_{ij\tau})^2) = \sigma_{ijt}.$$

Note that if $E(x_{ijt} - x_{ij\tau}) = \mu_{ijt} \neq 0$, then estimates of the error variance based on the mean square error will overstate the true error variance and the efficiency of the extrapolation will be understated. In the absence of any information to the contrary, however, we assume that the predictors are unbiased. Unfortunately, even with the assumption of unbiasedness in the predictions of the extrapolation, this specification still has too many estimable parameters. Additional simplifying assumptions are necessary.

Let us assume that the error variance can be decomposed in the following multiplicative manner:

(11) $$E((x_{ijt} - x_{ij\tau})^2) = \alpha_i^2 \beta_j^2 \gamma_t^2$$

where α_i measures the inaccuracy corresponding to the i-th variable, β_j the inaccuracy corresponding to the j-th AGI class, and γ_t the inaccuracy of the t-th extrapolation. The use of a multiplicative decomposition instead of the additive decomposition (i.e., $\delta_i^2 + \varepsilon_j^2 + \zeta_t^2$) generally used in variance analysis has a major benefit. With the multiplicative decomposition, a change in one of the decomposition factors has the same percentage effect on all the variances. Thus if γ_t^2 changes from 1 to 1/2 this reduces all σ_{ijt} by half, while if ζ_t^2 changes from 1 to 1/2 the percentage effect on σ_{ijt} depends on the values of δ_i^2 and ε_j^2.

The estimation of α_i, β_j, and γ_t is relatively straight forward. By definition:

$$\text{(12)} \qquad E\,\frac{(x_{ijt}-x_{ij\tau})^2}{\alpha_i^2\,\beta_j^2\,\gamma_t^2} = 1.$$

Using this identity we can derive the following estimators:

$$\text{(13a)} \qquad (1/n_i)\sum_{j=1}^{J}\sum_{t=1}^{\tau-1}\frac{(x_{ijt}-x_{ij\tau})^2}{b_j^2\,c_t^2} = a_i^2 \qquad n_i = J+\tau-1$$

$$\text{(13b)} \qquad (1/n_j)\sum_{i=1}^{I}\sum_{t=1}^{\tau-1}\frac{(x_{ijt}-x_{ij\tau})^2}{a_i^2\,c_t^2} = b_j^2 \qquad n_j = I+\tau-1$$

$$\text{(13c)} \qquad (1/n_t)\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(x_{ijt}-x_{ij\tau})^2}{a_i^2\,b_j^2} = c_t^2 \qquad n_t = I+J$$

where a_i^2. b_j^2, and c_t^2 are estimators for α_i^2. β_j^2. and γ_t^2 respectively. Since equation 11 has two multiplicative degrees of freedom (as one can multiply every α_i by λ_1. every β_j by λ_2. and every γ_t by $1/\lambda_1\lambda_2$ without changing the left hand side of the equation) the estimates of b_j and c_t must be normalized.

These equations are nonlinear in the estimators and must be solved recursively. Although recursive estimation of equations 13a - 13c does not guarantee convergence. in all the cases considered the estimates of the decomposition factors converged within twenty iterations.

In order to engage in statistical testing using the decomposition factors it is necessary to make two restrictive distributional assumptions: that the prediction errors are normally distributed. and that the predictors are uncorrelated. Given the nature of the extrapolation procedure. it is improbable that these two assumptions hold. Despite this. the decomposition factors can be used to rank the different variables. AGI classes. or extrapolations according to their relative efficien9cies (e.g.. to compare the relative efficiency of different extrapolations simply by comparing their ranking by c_t).[6]

IV. EXTRAPOLATING THE TAX MODEL DATABASE

The initial test of OTA's extrapolation process was a series of extrapolations of the Individual Tax Model Database based on the 1981 SOI to the year 1983. Since the 1983 SOI was available, all of the targets were derived from the actual values found on the SOI.[7] Thus, the fit of the different extrapolations reflected the mix of targeted variables chosen rather than the accuracy of the targets themselves.

The key problem in extrapolating the Individual Tax Model Database to 1983 SOI levels was the selection of the mix of targets to be used in the extrapolation process. It was decided that the initial group of targets should be chosen from the set of variables that are important in the new tax law but are in the Tax Model Database. These variables were: AGI, capital gains, partnership income, rental income, pensions, dividends, interest, wages and salaries, the earned income credit, the investment credit, and the foreign tax credit.

As was mentioned previously, the extrapolation process allows variables to be adjusted in two different procedures. A variable can be "blown up" (multiplied) by some constant adjustment factor for all the returns in the Tax Model Database -- this is the Stage One adjustment in the extrapolation process. Alternatively, a variable can be adjusted by changing the weights (the number of tax returns represented by each return in the database) for some subset of the returns in the database -- not surprisingly, this is called the Stage Two adjustment in the extrapolation process.

There are, of course, costs and benefits to using either the adjustment factor approach or the reweighting approach to extrapolate a database. To a large extent, the decision to use one of the extrapolation method depends on how the value of a variable is known or believed to have changed in the actual population. The modification of a variable only through the use of an adjustment factor assumes that the incidence of the variable in the population has not changed and that modification can be accounted for by an increase or decrease in the mean value of the variable per return. Changing the value of a variable using only the reweighting approach, on the other hand, assumes that the average value of the variable per return has not changed, and that the change can be explained by an increase or decrease in the incidence of the variable in the population. Using both adjustment factors and reweighting, allows the change to be caused by changes in both the incidence and the level of the variable.

The first extrapolation attempt adjusted the Tax Model Database using only the Stage Two reweighting method. This led to major changes in the weights on the adjusted database. This implied a massive change in the types of returns filed by taxpayers. Since OTA believed that such a massive change was unlikely in a two year period, this initial extrapolation was rejected.

In order to minimize the weight changes resulting from the extrapolation process all of the monetary targets were blown up by the inflation rate in a Stage One adjustment prior to the Stage Two reweighting. This significantly reduced both the size of the weight changes and the number of iterations the extrapolation routines required to meet the target levels. Hereafter, this extrapolation will be referred to as the initial extrapolation.

Although this initial extrapolation was successful in achieving the required aggregate levels for the target variables, the distribution of the targets by AGI class did not resemble the target variables' 1983 distributions. Therefore, it was decided to explicitly directly target both capital gains and partnership income by AGI class in the Stage Two reweighting (AGI had already been targeted by AGI class). This, however, resulted in more targets for the Stage Two reweighting than the program was designed to handle.

In order to achieve both the aggregate and distributional targets, the next extrapolation was divided into two subproblems: adjusting aggregate values and adjusting distributions. The first subproblem adjusted selected target values to the correct aggregate level. In addition, in the first subproblem, two distributional items were targeted: AGI by AGI class and the population by age strata.

The second subproblem was intended to correctly distribute selected target variables by AGI class. Therefore, no further Stage One adjustments were made for this problem. The subproblem's second stage was used to target the distribution of various targets. In addition, in order to correct for potential errors caused by reweighting, AGI was also targeted by class as were filing status and some income aggregates.

Many attempts were made to adjust the Tax Model Database to the 1983 SOI using this type of extrapolation. Although many different extrapolation runs were made the overall results of the adjustment process can be summarized by extrapolations 1-4 presented in Tables 1 and 2. Specifically, Table 1 presents the percent error in the prediction of the aggregate value of

Table 1: PERCENT ERROR IN THE PREDICTION OF THE TOTAL VALUE

Data	Extrapolation 1/				
Items	Initial	1	2	3	4
Adjusted Gross Income	1.14	0.0	- 0.71	- 0.01	- 0.01
Capital Gain	-21.63	0.0	0.0	0.0	0.0
Pensions	- 6.47	0.0	0.0	0.0	0.0
Dividends	4.38	- 0.64	- 2.78	- 5.56	- 4.17
Interest	8.28	0.0	0.0	0.0	0.0
Wages	0.98	- 0.61	- 0.78	- 1.54	- 1.19
Partnership Gain	- 1.68	0.0	0.0	0.0	0.0
Partnership Loss	- 9.46	- 0.1	0.0	- 0.01	0.0
Rental Gain	- 2.99	0.78	0.06	- 3.22	18.64
Rental Loss	-24.02	- 2.96	- 2.12	- 4.23	- 0.37
Other Schedule E Gain	-17.41	52.09	0.02	35.52	28.52
Other Schedule E Loss	- 1.45	- 6.44	0.0	- 7.87	10.12
Earned Income Credit	- 8.8	-23.8	0.11	0.0	0.0
Investment Tax Credit	1.77	2.26	5.70	4.33	4.42
Foreign Tax Credit	4.70	- 0.16	0.0	0.0	0.0
Two Earner Deduction	101.5	102.1	- 7.09	-51.39	1.24
Medical Deductions	0.50	38.09	30.78	27.13	-15.43
Single Return	2.05	- 0.05	0.05	- 0.05	- 0.04
Joint Return	- 0.57	0.08	0.08	0.08	0.08
Married, Filing Separately	18.18	0.0	0.0	0.0	0.0
Head of Household	- 5.02	- 0.47	- 0.47	- 0.47	- 0.47

[1] Each of the extrapolations are described in the text.

Table 2: INFORMATION GAIN IN THE PREDICTED DISTRIBUTION (IN THOUSANDS)

Data Items	Extrapolation 1/				
	Initial	1	2	3	4
Adjusted Gross Income[1]	N/C	N/C	N/C	N/C	N/C
Capital Gain	1.61	2.1	2.38	2.58	1.47
Pensions	6.97	10.7	8.25	8.48	26.1
Dividends	9.4	14.3	10.4	14.5	43.8
Interest	9.99	9.44	6.78	8.37	17.9
Wages	3.27	0.59	1.3	0.77	1.43
Partnership Gain	9.84	8.98	8.34	10.4	63.7
Partnership Loss	5.9	17.6	90.8	16.0	14.4
Rental Gain	14.2	9.45	11.0	9.84	26.9
Rental Loss	14.3	11.3	36.4	9.07	42.7
Other Schedule E Gain	6.94	83.2	44.6	108.0	147.0
Other Schedule E Loss	70.4	98.6	29.8	106.0	70.0
Earned Income Credit	N/C	N/C	N/C	N/C	N/C
Investment Tax Credit	N/C	N/C	N/C	N/C	N/C
Foreign Tax Credit[2]	N/C	N/C	N/C	N/C	N/C
Two Earner	1050.0	1080.0	19.5	1120.0	1310.0
Medical Deduction[2]	N/C	N/C	N/C	N/C	N/C
Single Return	1.57	0.72	0.37	0.61	1.25
Joint Return	5.20	1.69	2.01	1.87	2.83
Married, Filing Separately	81.0	91.1	87.0	87.3	104.0
Head of Household	4.90	5.04	4.98	5.30	0.87

[1] Each of the extrapolations are described in the text.

N/C = Not calculated due to zero or negative values in cell.

selected variables for the initial and the four experimental extrapolations. Table 2 presents the information gain in the predicted distributions for the four experimental extrapolations. The major adjustments and additions made to the targets for each extrapolation, as well as the results of the different target mixes, are summarized below.

Extrapolation 1's relative efficiency proved to be superior to the initial extrapolation. The error variance attributable to the extrapolation (c_t^2) is 24.4% smaller than the variance attributable to the initial extrapolation (see Figure 1). In examining the individual predictions of the extrapolation 1, however, it was decided that the extrapolation's major flaws lie in its prediction of the two earner deduction and its prediction of the Other Schedule E income and loss. Therefore, the second extrapolation included five additional targets in the second pass' second stage to improve the extrapolation's prediction of the distribution of Other Schedule E income. In addition, the two earner deduction's income share imputation was replaced with an imputation based on 1983 data.

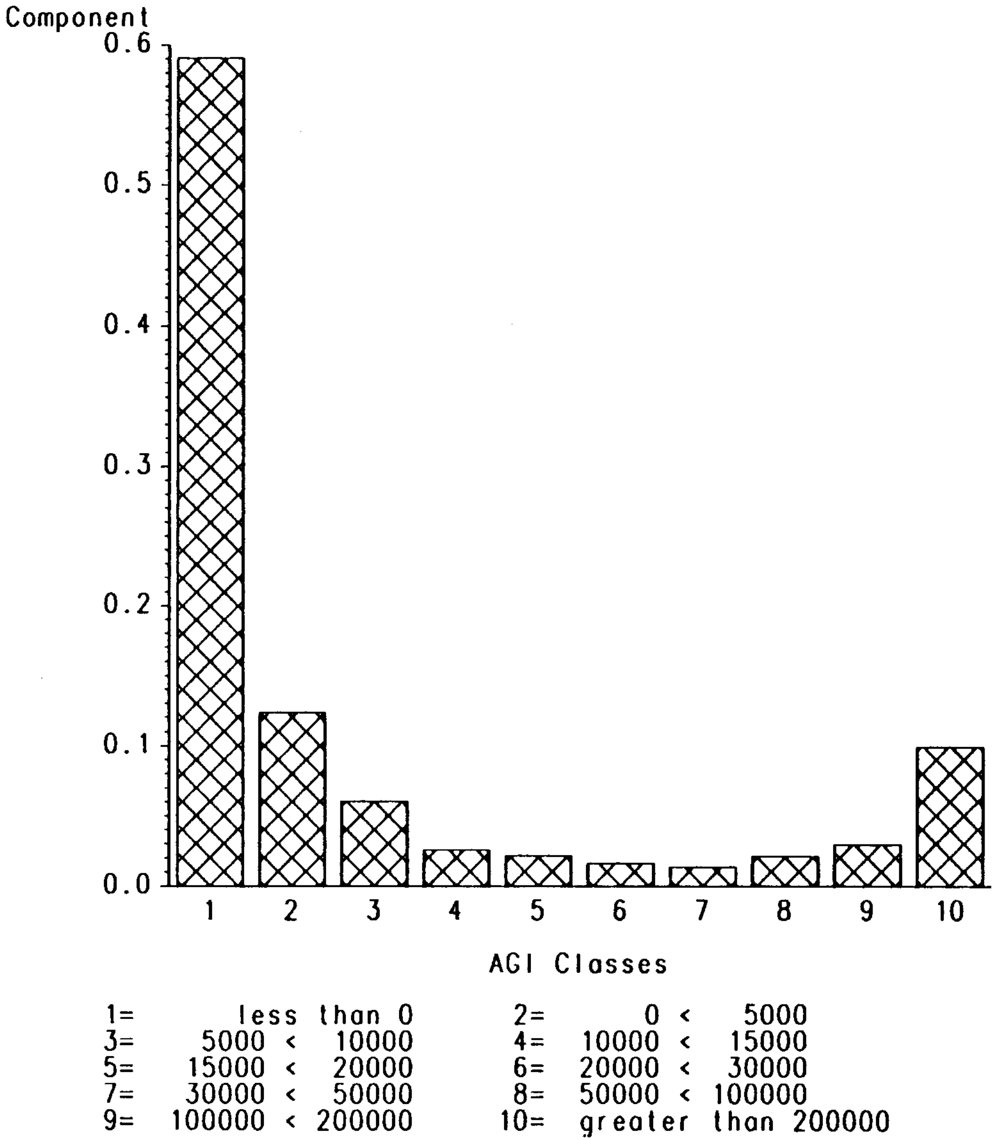
The relative efficiency of extrapolation 2 was also far better than that of the initial extrapolation. As Figure 5 shows, the error variance attributable to extrapolation 2 was 2.1% smaller than the error variance of extrapolation 1. In addition, extrapolation 2 improved the prediction of Other Schedule E income (see Tables 1 and 2). Unfortunately, extrapolation 2 had a poorer prediction of the levels of wages and salaries and dividends than did extrapolation 1. Further, extrapolation 2's prediction of the medical deduction was significantly worse than that of the initial extrapolation.

For extrapolation 3, therefore, it was decided to add adjustment factors for the medical expense deduction to Stage One of the first subproblem and to add an aggregate target for this variable to Stage Two to improve the extrapolation's prediction of these variables. After an initial attempt, subproblem 1's blowups of Other Schedule E Gains and Losses were removed to improve the extrapolation's predictive power. These modifications to the subproblem 1 of the extrapolation did not improve the relative efficiency of the extrapolation. In fact, the error variance attributable to the third extrapolation was 14.4% larger than that of the second extrapolation and the predictions for many of the individual variables were much worse (see Tables 1 and 2, and Figure 1).

For extrapolation 4, an initial attempt was made to fix the third extrapolation by adding targets for the medical expenses deduction and the two earner deduction to subproblem 2's second stage. Under this target specification, however, the extrapolation procedures did not converge. After some

Figure 1

Portion of Total Variance Attributed to AGI Classes



experimentation. it was discovered that the key set of targets was the Other Schedule E distribution. Removal of the Other Schedule E targets resulted in convergence of the extrapolation procedure.

The results of extrapolation 4 are much worse than those of any of the previous extrapolations, as the error variance of the fourth extrapolation was 88.7% larger than the error variance attributable to the initial extrapolation. Although the predictions of the individual variables that were poorly predicted by the prior extrapolations were improved in the fourth extrapolation, other predictions of individual variables are worse than those in both the prior extrapolations and the initial extrapolation.

The results of extrapolations 1 through 4 indicate that there are significant trade-offs to be made between individual variables in the extrapolation (see Figures 2 and 3) and it is impossible to choose between them without some external criterion. In fact, without some sort of lexicographic preference ordering about the fit of different variables the only reasonable selection criteria for choosing the best extrapolation procedure is the relative efficiency of the various extrapolations. Based on this criteria, the second extrapolation provides the best adjustment of the Tax Model Database to 1983 SOI levels.

V. CONCLUSIONS

The use of extrapolation methods to adjust the OTA's Individual Tax Model Database has in some regards been only partially successful. Although it has proven possible to improve the efficiency of the Tax model Database as a predictor of the 1983 SOI, it has not been possible to totally eliminate the error variance of the database. In the course of this process, however, a great deal has been learned about the nature of the extrapolation procedures. First and foremost, it has proven impossible to improve the predictive power of the extrapolation by simply adding more targets. As the number of targets grows, the extrapolation procedures produce an increasingly distorted database as seen in both the changes in the pattern of the sample weights and in the distribution of the examined variables. The addition of more targets to adjust for the increased distortion eventually leads to nonconvergence of the extrapolation routines. This is especially true when the targets are sources of income (e.g., note the effect of targeting Other Schedule E income when most of the other income sources were already targeted). Perhaps the extrapolation procedure needs some 'degrees of freedom' in order to successfully adjust a database. In particular

Figure 2

Portion of Total Variance Attributed to Different Extrapolations

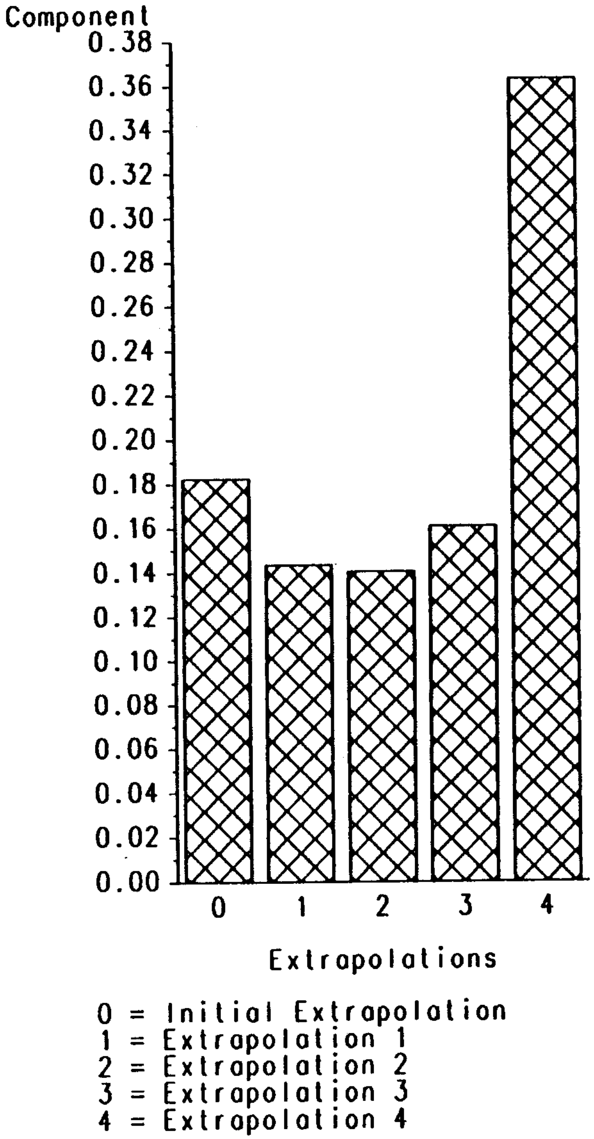
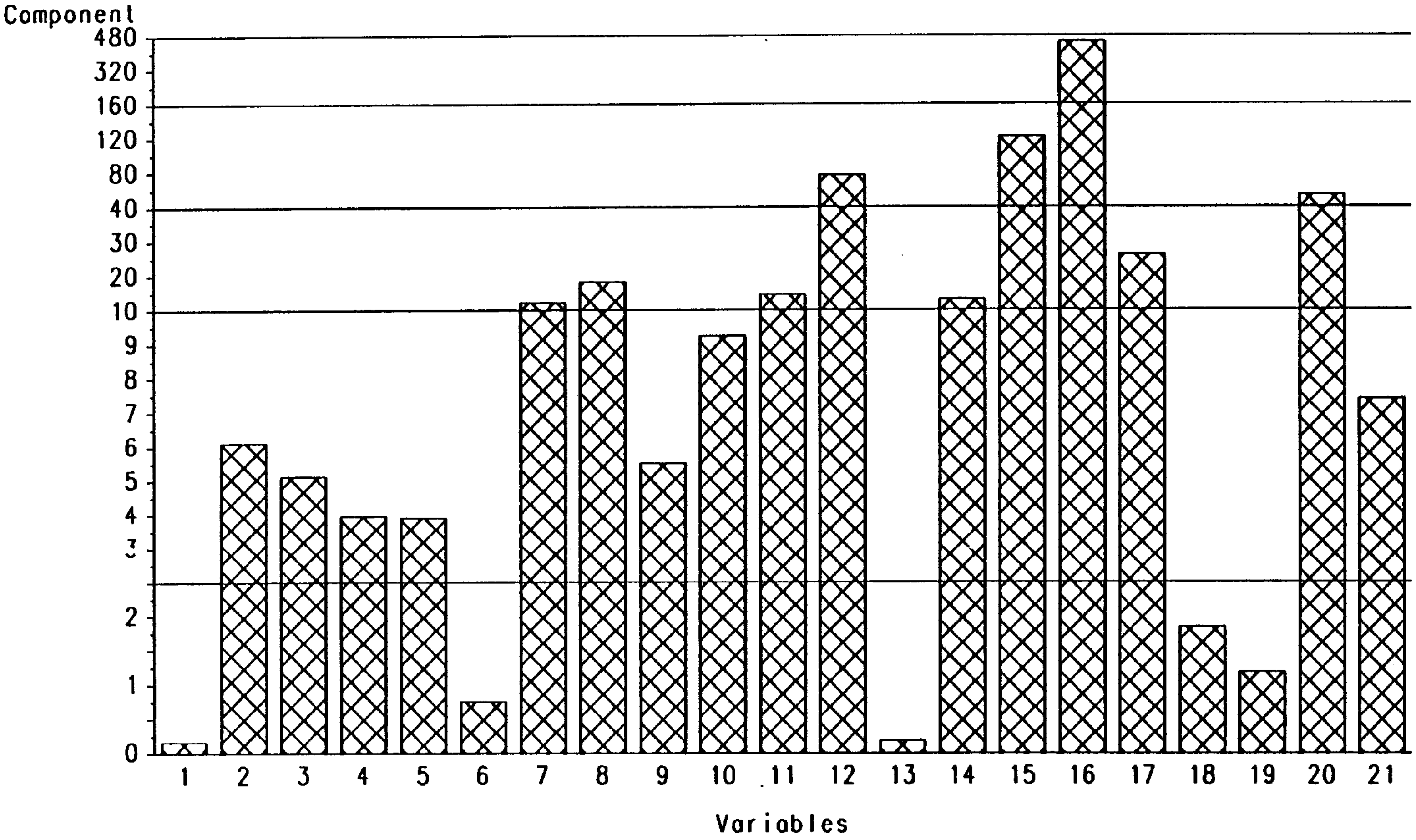


Figure 3

Portion of Total Variance Attributed to Data Items



1= Adjusted Gross Income 2= Capital Gain 3= Pensions
4= Dividends 5= Interest 6= Wages
7= Partnership Gain 8= Partnership Loss 9= Rent Gain
10= Rent Loss 11= Other Schedule E Gain 12= Other Schedule E Loss
13= Earned Income Credit 14= Investment Tax Credit 15= Foreign Tax Credit
16= Two Earner 17= Medicaid Deduction 18= Single Return
19= Joint Return 20= Married Filing Separately 21= Head of Household

it appears that at least three sources of income must be untargeted (i.e., three 'degrees of freedom') if the extrapolation procedure is to converge.[8]

Second, the changes caused by the second pass indicate that Stage Two of the extrapolation process does not preserve the distributions of either targeted or untargeted variables. Due to the nature of the criterion function (i.e., to minimize

$$(14) \qquad L = (w_1/w_0)^4 + (w_1/w_0)^{-4} - 2$$

where w_1 is the adjusted weight and w_0 is the initial weight) and the fact that variables are not identically distributed across AGI classes, the reweighting scheme cannot conserve the distribution of a variable by AGI class unless each class is targeted independently.

Third, although there are no large (i.e., larger than 0.1) covariances (as calculated using the Pearson covariance estimator, the nonparametric covariances are somewhat higher) between the various sources of income for the entire population, there do seem to be such correlations within various subgroups of the population. This may explain some of the changes in the non-targeted variables that result from the extrapolation process. In addition, the absence of correlations for the whole database and their presence for subgroups suggests that it might be desirable to re-weight subgroups rather than the entire database.

Together these three findings suggest that it would be fruitful to continue research into improvements on the extrapolation process in two general directions. First, since it is impossible to target all variables in the database, a procedure must be developed that determines what are the key variables on the database to be used as targets. One possible source of these variables is to examine the highly correlated variables that are found for the subpopulations. Another possibility is to use artificial variables derived from principle component analysis as targets.

A second area where the extrapolation process may be improved is in its distributional effects. To the extent that some distortion of distributions is inevitable, it can be minimized through the proper selection of a criterion function. Perhaps a criterion function designed to minimize the change in the shape of the distribution of a group of variables will have less effect on the distribution than does a criterion function intended to minimize relative weight changes. Alternatively, perhaps the database ought to be extrapolated by subgroups rather than as a whole.
An additional area of long-term research is the mechanism for selecting targets. The research in this paper was based on using known SOI values as targets. For most extrapolations, the targets

will be forecasts rather than actual values. and the accuracy of the extrapolation will be dependent on the accuracy of the forecast. At this time. however. there has been little investigation of the properties needed in the forecast of extrapolation targets.

To summarize. the recent use of the extrapolation process to adjust the Individual Tax Model Database to 1983 SOI levels has resulted in a substantial increase in our knowledge of the behavior of the extrapolation procedure and has suggested several ways in which the extrapolation process might be improved.

ENDNOTES

[1] It is important to note that there are. in fact. some items on the tax database that cannot be adjusted through inflation factors. In general. such items are either discrete variables or items that are dependent on the value of other items on the record. Some examples of these are the number of deductions. the value of various income dependent credits. and the itemized deductions. Lindsey (1985) does not suggest that these items should be extrapolated using inflation factors: his arguments in favor of adjustment factors are limited to the components of income.

[2] I wish to thank John Wilkins for informing me of the existence of this earlier OTA extrapolation procedure.

[3] I am using the term 'portfolio adjustment' in a much broader sense than is common in the finance literature. Specifically. I am including all sources of income in an individual's portfolio rather than just the individual's capital assets.

[4] Unfortunately. traditional hypothesis testing cannot be used with the expected information statistic. There does exist a weaker concept -- the minimum discrimination information -- which is analogous to the Cramer-Rao inequality (Theil [1971]. and Kullback [1959]). Essentially. this concept determines the minimum value of I(y:x) which allows one to reject the null hypothesis that the two distributions are identical. For our purposes. the discriminating level for I(y:x) is zero (i.e. if I(y:x) > 0.0. then the two distributions are not the same). Note. however. that the discriminating concept does not provide a mechanism for calculating confidence intervals or for comparing the fit of two distributions relative to a third: in these cases. analysis is limited to an ordinal ranking of the values of the expected information statistics.

[5] I am grateful to Roy Wyscarver for pointing out some of the potential uses of the expected information concept (e.g. in the imputation process when imputed variables are being distributed using another variable's distribution). Another potential use of the expected information concept is to examine changes in the distribution of variables over time.

[6] Two modifications were made to the formula so that the squared logarithmic prediction errors could be calculated for all the cell values. First. since there were no cases where the predicted cell value had a different sign than the true cell value all logs were taken using absolute values. Second, for those cells where either the true or the predicted value equaled zero, one was added to both cell values before the logs were calculated.

[7] An interesting set of comparison values for the extrapolations were the estimated errors resulting from the subsampling process used to generate the reduced SOI file used in the Tax Model. Not surprisingly. the errors found for the reduced SOI were far smaller than those found for any of the extrapolations.

[8] Not surprisingly the requirement that there be approximately three degrees of freedom implies that AGI and the various components of income are related in an extremely nonlinear fashion. Unfortunately the exact nature of this nonlinear relationship is not clear. This relationship may be clarified through future research.

REFERENCES

Cilke. James. and Roy Wyscarver (1987) "The Treasury Individual Income Tax Simulation Model" in Compendium of Tax Research. U.S. Department of the Treasury. Washington. DC: Government Printing Office.

Kullback. Solomon (1959) Information Theory and Statistics. New York: John Wiley and Sons. Inc.

Lindsey. Lawrence (1985) 'Creating a Baseline Income Distribution for Tax Data'. Unpublished Manuscript.

Theil. Henri (1967) Economics and Information Theory. Amsterdam. The Netherlands: North-Holland Publishing Co.

Theil. Henri (1971) Principles of Econometrics. New York: John Wiley and Sons. Inc.