An Improved Technique for Obtaining Current Sub-state Income Estimates

*E. Anthon Eff**

Abstract: This paper compares some specifications for estimating current sub-state (county or MSA) income. A large number of trials are generated, using the data for Tennessee, in a procedure akin to a series of Monte-Carlo experiments. The out-of-sample forecast errors are examined to see which specifications perform best, and then to obtain some insight into the approximate error one would obtain using the best specification.

In many local economies, regional economists are routinely asked—by the press, by business groups, or by policy makers—to discuss current economic conditions. Most of these discussions are constrained by the availability of data and tend to center on employment issues, since employment statistics for small areas (counties and MSAs) are relatively current. However, the number of employed or unemployed persons is of no more intrinsic interest than other indicators of the local economy, such as total personal income, per capita income, total earnings, or average earnings per worker. Indeed, any economist who tries to keep in view changes in the quality of life needs to have at least some idea of the ways in which income is changing in the local economy.

There are a variety of data sources that cast light on some aspect of income at the county level. The decennial census provides detailed income data by race and age, including measures of household income. The drawback, of course, is that these data are only updated every 10 years. The Bureau of Economic Analysis (BEA) produces annual estimates of income and earnings in its Regional Economic Information System (REIS), though with a lag of about two years, so that the most recent data represent conditions two years or more in the past. These data are considered by most economists to be the definitive annual county-level income measures, and include detailed sub-categories of transfer payments and earnings by industry, though no detail by demographic characteristics (Bailey 1997). The Internal Revenue Service (IRS) also issues annual estimates of income for counties, and these can be used to proxy household income or per capita income, but do not include any separate estimate of employment earnings. These data have an average lag of about a year.[1]

The Current Employment Statistics (CES) survey, produced by the state affiliates of the Bureau of Labor Statistics (BLS), produces estimates of average hourly and weekly earnings for selected sectors in selected MSAs (U.S. Depart-

*Assistant Professor, Department of Economics and Finance, Middle Tennessee State University. The author would like to thank Joachim Zietz, Reuben Kyle, David Walls, Thomas Knapp, and three anonymous reviewers for their comments and suggestions.

[1]Figures for individual counties are accessible via Internet from Syracuse University (http://trac.syr.edu/tracirs/index.html).

ment of Labor, BLS 1997a). Unlike the previous data sources, these figures appear monthly, with a lag of only one or two months. Their main drawback is that they present earnings data for only a few sectors, and then only for production workers, and therefore do not give a good picture of overall income growth in the local economy. For this reason, many economists prefer working with the ES202 data, which are also produced by BLS affiliates, and which provide earnings and employment data for each firm paying premiums for unemployment insurance (U.S. Department of Labor, BLS 1997b). These data are not published, but analysts affiliated with state governments are often able to gain access to four digit SIC series for individual counties-series that can easily be aggregated to show total earnings and average earnings per worker at the county level, though one must always exercise caution not to release sensitive information. Data are generally available with about a six-month lag (though this varies from state to state).

Both the CES and the ES202 are limited to earnings from employment, and to place of work earnings. This makes them narrower in scope than the Census, BEA, and IRS series, which all provide place of residence estimates for personal income (this includes not only employment earnings, but transfer payments and dividends, interest, and rent). In addition, these series are not without their own sources of error. The CES figures are based on a state-wide sample and allocated to specific MSAs; revisions are often rather large. Though the ES202 series is more accurate than the CES, it still has a number of problems. Most important among these are the workers who are excluded: about half of those in agriculture, some state and local government employees, and most workers at non-profit organizations, including the many employees of religious institutions (U.S. Department of Labor, BLS 1997b). In addition, multi-establishment firms often choose not to report separate figures for each establishment (U.S. Department of Labor, BLS 1997b). Thus, one reporting unit can be assigned the employment and income of establishments scattered throughout the state. This can dramatically skew employment and earnings figures for individual counties.

On the state level, there is one excellent source of current income data: BEA's quarterly state income series. These are typically available with a lag of one or (at most) two quarters, and provide detailed information on income by source. Beginning with the example of Michael A. Conte (1989), these data have been used to estimate county-level and MSA-level income. The advantages of the series generated in this way are:

- Current estimates, with a lag of less than two quarters (making them superior to the REIS data).
- Estimates of total personal income, rather than simply earnings from labor or certain types of labor (making them superior to the CES and ES202 data).
- Place of residence estimates, rather than place of work estimates (making them superior to the CES and ES202 data).
- No potential confidentiality problems (making them superior to the ES202 data).

This method is especially valuable for producing estimates of either total personal income or net earnings (post-FICA earnings of resident workers). The former series can be divided by population (though, here again, county-level data on population are only available with a lag of between six months and a year and a half) to give per capita income. The latter series can be divided by the Current Population Survey (CPS) estimate of resident workers to give average earnings per resident worker.[2]

Conte (1989, p. 89) suggested that his estimates could be used to obtain quarterly estimates of local-area personal income. It seems, however, that the best use of these estimates is not to produce a measure of quarterly fluctuations, but rather to provide a current estimate of annual income or earnings. Other series provide excellent measures of quarterly fluctuations (most notably the various employment series); the unique value of this series is its ability to present current estimates reflecting the standard of living, such as average earnings or per capita income.

Estimating Sub-State Income from State Quarterly Income

Conte (1989) proposed an estimating equation as follows:

$$(1) \qquad \frac{Y^{q}_{c,t}}{Y_{state,t}} = \alpha_0 + \alpha_1 \frac{L_{c,t}}{L_{state,t}} + \sum \gamma_i X_{i,t} + \epsilon_t$$

where Y= Income (or earnings); L= Employment; X_i =other independent variables; t= time period; c denotes county or county combination; and q indicates quarterly interpolation.

In this model (equation 1), a sub-state region's share of state income (or earnings) is a function of the region's share of state employment, as well as some other unspecified variables. In practice, Conte (1989, p. 91) considers only employment share, ignoring the other independent variables. This gives the model (equation 2) below:

$$(2) \qquad \frac{Y^{q}_{c,t}}{Y_{state,t}} = \alpha_0 + \alpha_1 \frac{L_{c,t}}{L_{state,t}} + \mu_t.$$

The data for this estimating equation are assembled as follows:

1. Obtain state income data ($Y_{state,t}$), from the quarterly BEA state income series.
2. Obtain sub-state region income data ($Y_{c,t}$), from the annual REIS series. Interpolate to quarterly frequency.
3. Obtain sub-state region employment data ($L_{c,t}$), from the monthly CPS series. Seasonally adjust and convert to quarterly frequency.
4. Obtain state employment data ($L_{state,t}$), from the monthly CPS series. Seasonally adjust and convert to quarterly frequency.

[2]Since the net earnings figure includes military earnings, and the employment figure excludes military personnel, some correction is always needed when producing a local area series on net earnings per resident worker.

The model will have missing values for the dependent variable in the most current two years (since the REIS data are not available), but no missing values for the independent variable. Thus, one can estimate the model and—using the resultant parameter estimates—forecast the dependent variable into the two-year period for which the data are missing. Then, by multiplying state income by the forecast of regional income share, one can obtain an estimate of regional income (equation 3).

$$(3) \qquad Y^{q}_{ct} = Y_{state,t} * \text{forecast}\left(\frac{Y^{q}_{ct}}{Y_{state,t}}\right).$$

This paper seeks to extend Conte's (1989) method, by considering the other independent variables X_i in equation 1. A variety of other specifications are possible, including lags of the dependent variable and lags of the employment ratio. One available data source is the state-level composition of income and earnings, broken out into such categories as transfer payments, earnings by one-digit SIC sector, and property income. Since each sub-state region differs in its sources of income—particularly in the importance of the SIC sectors—it stands to reason that any shift in the relative fortunes of a sub-state region would be reflected in a shift in the relative shares of these different income categories. This suggests an expansion of the Conte method, in which the independent variable (sub-state changes in employment shares) is supplemented by variables measuring state-level changes in income composition. A "full" model, containing all easily obtained independent variables, would appear as in equation 4:

$$(4) \qquad \frac{Y^{q}_{c,t}}{Y_{state,t}} = \gamma_0 + \sum_{j=1}^{4} \gamma_j \frac{Y^{q}_{c,t-j}}{Y_{state,t-j}} + \sum_{j=0}^{4} \alpha_j \frac{L_{c,t-j}}{L_{state,t-j}} + \sum_{i=1}^{k} \beta_i \frac{Y_{state,i,t}}{L_{state,t}} + \mu_t$$

where Y= Income; L= employment; t= time period; j= lags; i=1,..., k components of state income, as follows:

1) Transfer Payments
2) Durable Goods Manufacturing
3) Dividends, Interest, and Rent
4) Nondurable Goods Manufacturing
5) Construction
6) Services
7) Retail Trade
8) Wholesale Trade
9) T.C.P.U.
10) F.I.R.E.
11) Government: Military
12) Government: Federal Civilian
13) Government: State and Local
14) Farm
15) Residence Adjustment
16) Mining
17) Agricultural Services

The data on components of state income are readily available, published each quarter, and included in the same files that contain the state total personal income and earnings. This makes the specification convenient.

Evaluation of Model

This section compares some of the specifications mentioned above for estimating sub-state income. A large number of trials are generated, using the data for Tennessee, in a procedure akin to a series of Monte-Carlo experiments. The out-of-sample forecast errors are examined to see which specifications perform best, and then to obtain some insight into the approximate error one would obtain using the best specification.

Forecast Statistic

This paper assumes that the purpose of the estimation is to obtain a current annual estimate of local-area personal income, using the state-level quarterly series. Thus, the method seeks an estimate of the county-level REIS annual income figure, not some unknown quarterly series. Forecast error for a given year can therefore be written as the mean income estimate during a year minus the actual income that year:

$$\text{(5a)} \qquad E_c = \frac{1}{4}\sum_{t=1}^{4} \hat{Y}^q_{c,t} - Y_c$$

where E_c is the forecast error for county c, Y_c is the actual income for county c (from the REIS data), and $\hat{Y}^q_{c,t}$ is the estimated quarterly income for county c in quarter t. Equation 5a can be decomposed into two parts: that portion due to estimation error, and that portion due to conversion of the annual series to a quarterly series:

$$\begin{aligned}\text{(5b)} \qquad E_c &= \frac{1}{4}\sum_{t=1}^{4} \hat{Y}^q_{c,t} - Y_c \\ &= \frac{1}{4}\sum_{t=1}^{4} (\hat{Y}^q_{c,t} - Y^q_{c,t} + Y^q_{c,t} - Y_c) \\ &= \frac{1}{4}\sum_{t=1}^{4} (\hat{Y}^q_{c,t} - Y^q_{c,t}) + \left(\frac{1}{4}\sum_{t=1}^{4} Y^q_{c,t} - Y_c\right) \\ &= E_f + E_a.\end{aligned}$$

E_f is the estimation error for county c: it measures the fit of the model to the values used as dependent variables. These values, however, are interpolated from an annual series, and the interpolation process itself introduces error. E_a measures that interpolation error: it gives the difference between the mean interpolated value and the actual annual value for county c.

Dividing both sides by Y_c gives the percent forecast error:

$$\text{(5c)} \qquad \frac{E_c}{Y_c} = \frac{E_f}{Y_c} + \frac{E_a}{Y_c}.$$

Taking the absolute value of each term then gives an indication of the relative magnitudes of the forecast error and the interpolation error.

Simulation Procedure

In practice, these estimates will usually be performed for large counties or combinations of counties; estimates will employ REIS county-level income data ending about two years before the present. In the evaluation of each model, efforts were made to duplicate these conditions. Simulations were conducted using Tennessee data. Tennessee has 95 counties, so that the number of sub-state regions consisting of two counties[3] is 95!/(2!93!)=4,465. For each of these regions, two-year-ahead ex-post forecasts were conducted for each of the years 1988 through 1993. This gives a total of 26,790 forecasts conducted for each model. The result resembles a Monte-Carlo experiment in that a very large number of trials are run; it differs in that the data are not wholly contrived.[4]

Data were drawn from the three sources mentioned above: the BEA's REIS annual estimates of county income; the BEA's quarterly estimates of state income and its components; and the BLS's CPS monthly estimates of employed persons per county, covering all civilian workers, including farm. The data span the period 1980:1 through 1995:4, for a total of 64 quarters. Employment was seasonally adjusted—using the X11 procedure—and then averaged for each quarter. All income measures were converted to real figures using the Gross Domestic Consumption implicit price deflator. Total personal income (TPI) was chosen as the series to be estimated, though net earnings would have served just as well.

For each of the 26,790 runs, for each model, in each of the two out-of-sample forecast years, the absolute values of E_c, E_a, and E_f are divided by Y_c, in order to give the absolute percent forecast errors. Thus, for each model, there are a total of 53,580 calculations of absolute percent forecast error. These measures are used to assess the relative performances of the model specifications.

Model Specifications

A variety of issues were examined in the simulations. Each run differed in the following respects: specification of the independent variables; specification of the dependent variable; the quarterly interpolation technique; and whether or not irrelevant independent variables were excluded.

Specification of the independent variables

Four basic models were formulated:

(1) the simple regional employment share a la Conte (1989);

$$\text{(Model 1)} \qquad \frac{Y^{q}_{c,t}}{Y_{state,t}} = \alpha_0 + \alpha_1 \frac{L_{c,t}}{L_{state,t}} + \mu_t$$

[3]Areas are aggregated without respect to contiguity since there are a number of reasons why sub-state estimates might be based on noncontiguous counties. For example, income might be estimated for all counties in Tennessee heavily dependent on tobacco, or heavily dependent on auto production.

[4]This simulation method can be used to evaluate forecasts for many regional economic series (e.g., Eff 1998).

(2) a model with state-level sectoral income shares;

$$\text{(Model 2)} \qquad \frac{Y^{q}_{c,t}}{Y_{state,t}} = \gamma_0 + \sum_{i=1}^{k} \beta_i \frac{Y_{state,i,t}}{L_{state,t}} + \mu_t$$

(3) current and lagged (four quarters) regional employment shares with state-level sectoral income shares;

$$\text{(Model 3)} \qquad \frac{Y^{q}_{c,t}}{Y_{state,t}} = \gamma_0 + \sum_{j=0}^{4} \alpha_j \frac{L_{c,t-j}}{L_{state,t-j}} + \sum_{i=1}^{k} \beta_i \frac{Y_{state,i,t}}{L_{state,t}} + \mu_t$$

(4) the "full" model with lagged (four quarters) regional income shares, current and lagged (four quarters) regional employment shares, and state-level sectoral income shares.

$$\text{(Model 4)} \qquad \frac{Y^{q}_{c,t}}{Y_{state,t}} = \gamma_0 + \sum_{j=1}^{4} \gamma_j \frac{Y^{q}_{c,t-j}}{Y_{state,t-j}} + \sum_{j=0}^{4} \alpha_j \frac{L_{c,t-j}}{L_{state,t-j}} + \sum_{i=1}^{k} \beta_i \frac{Y_{state,i,t}}{L_{state,t}} + \mu_t$$

Specification of the dependent variable

Two specifications were tested. The logistic formulation has the advantage that the share of the region in the state is always constrained to lie between zero and one.

$$\text{Linear: } \frac{Y^{q}_{c,t}}{Y_{state,t}}$$

$$\text{Logistic: } \ln\left(\frac{z}{(1-z)}\right) \text{where } z = \frac{Y^{q}_{c,t}}{Y_{state,t}}$$

Quarterly interpolation technique

The amount of error due to conversion of county income to quarterly frequency can be expected to differ from one interpolation technique to another. One choice would be to dispense with any interpolation and simply use the annual REIS figure for each of the four quarters. E_a would then equal zero, but it seems likely that E_f would be larger than it would have been had some quarterly adjustment been performed. Here, three alternatives are tested:

1) no interpolation;
2) the method used by Conte (1989, p. 92n) and attributed to Diz (1970); and
3) a cubic spline curve transformation performed by SAS (SAS Institute Inc. 1988, pp. 272-273).

The last alternative was included because there is a well-developed literature on spline curves in mathematics (e.g., de Boor 1978; Nonweiler 1984) and in finance (e.g., Waggoner 1997) and because the procedure is easily implemented by anyone with access to SAS/ETS.

Omission of irrelevant variables

Several of the models were examined to see if dropping irrelevant variables would lower forecast error. First, an unrestricted model was estimated in which the p-values of all coefficient t-statistics were recorded. Next, a restricted model was built in which the independent variables were introduced in batches, as defined by threshold p-value (beginning with p-value 0.10 and below), a new regression run, and an F-statistic (H_0: omitted variables do not belong in the model) calculated. If H_0 was rejected, threshold p-value was incremented by .10, a new restricted regression was run, and so on, until one could no longer reject H_0. The procedure thus follows the basic top-down approach to model building and can be performed with loops programmed in TSP or some other econometric software.

Other Issues Examined in the Simulations

There are various factors, other than model specification, that can affect the accuracy of sub-state income estimation. First, as established by Conte (1989), error increases as the size of the sub-state area decreases. The two-county areas used in the simulations vary in size from 0.1 percent to 33.7 percent of Tennessee's total income. This range, together with the large number of simulations, allows some determination of how error might vary with the size of the sub-state region.

Second, forecasts are made over a two-year period: the results are used to produce two separate annual estimates of income (one for the current year, one for the previous year). It seems reasonable that the forecast accuracy should be higher in the first year than in the second.

Third, forecasts made from different estimation periods could vary in accuracy. For example, a three-year-ahead forecast using data ending in 1988 might differ in error from one made with 1993 as the end year. These "annual effects" could be due to structural change or to different phases of the business cycle. For example, if the current period is recessionary, and the most recent REIS data are from an expansionary period, then one might expect a relatively high forecast error. The years used in the simulation (1988-1993) include one recession (the trough for Tennessee occurred in the first quarter of 1991), and allow some insight into how these annual effects might affect forecast accuracy.

Results

Using the Tennessee data, thirty models were tested, each with a different specification of dependent or independent variables, different quarterly interpolation technique, or different treatment of irrelevant variables. For each model, 53,580 estimates of annual income were produced. Absolute percent error was calculated, as well as its components (estimation error and quarterly interpolation error). Errors were ranked, from lowest absolute percent error to highest, and the 95th percentile error (APE95) was found (i.e., that level of error for which the absolute percent error was lower 95 percent of the time). Table 1 presents these errors. The table ranks models from lowest APE95 to highest.

TABLE 1

Results of Simulations: 95th Percentile Error, All Models

No.	Independent Variable Specification	Dependent Variable Specification	Quarterly Interpolation Method	Treatment of Irrelevant Variables	Absolute Percent Error Total	Absolute Percent Error Estimation	Absolute Percent Error Quarterly Interpolation
1	Model 4	Logistic	SAS Spline	Restricted	2.64%	2.57%	0.41%
2	Model 4	Logistic	Diz-Conte	Restricted	2.64%	2.64%	0.03%
3	Model 4	Linear Ratio	Diz-Conte	Restricted	2.75%	2.75%	0.03%
4	Model 4	Linear Ratio	SAS Spline	Restricted	2.78%	2.71%	0.41%
5	Model 4	Logistic	SAS Spline	Unrestricted	3.49%	3.39%	0.41%
6	Model 4	Logistic	Diz-Conte	Unrestricted	3.51%	3.49%	0.03%
7	Model 4	Linear Ratio	SAS Spline	Unrestricted	3.65%	3.56%	0.41%
8	Model 4	Linear Ratio	Diz-Conte	Unrestricted	3.70%	3.68%	0.03%
9	Model 2	Logistic	SAS Spline	Restricted	5.15%	5.08%	0.41%
10	Model 2	Logistic	Diz-Conte	Restricted	5.24%	5.23%	0.03%
11	Model 2	Linear Ratio	SAS Spline	Restricted	5.38%	5.31%	0.41%
12	Model 3	Logistic	SAS Spline	Restricted	5.41%	5.32%	0.41%
13	Model 2	Linear Ratio	Diz-Conte	Restricted	5.47%	5.46%	0.03%
14	Model 3	Logistic	Diz-Conte	Restricted	5.60%	5.60%	0.03%
15	Model 3	Linear Ratio	SAS Spline	Restricted	5.62%	5.53%	0.41%
16	Model 3	Linear Ratio	Diz-Conte	Restricted	5.79%	5.79%	0.03%
17	Model 4	Logistic	None	Restricted	5.88%	5.88%	0.00%
18	Model 4	Linear Ratio	None	Restricted	6.07%	6.07%	0.00%
19	Model 2	Logistic	None	Restricted	6.72%	6.72%	0.00%
20	Model 4	Logistic	None	Unrestricted	6.74%	6.74%	0.00%
21	Model 4	Linear Ratio	None	Unrestricted	6.99%	6.99%	0.00%
22	Model 2	Linear Ratio	None	Restricted	7.11%	7.11%	0.00%
23	Model 3	Logistic	None	Restricted	7.75%	7.75%	0.00%
24	Model 3	Linear Ratio	None	Restricted	8.00%	8.00%	0.00%
25	Model 1	Logistic	SAS Spline	NA	11.80%	11.76%	0.41%
26	Model 1	Logistic	None	NA	11.82%	11.82%	0.00%
27	Model 1	Logistic	Diz-Conte	NA	11.82%	11.84%	0.03%
28	Model 1	Linear Ratio	SAS Spline	NA	11.93%	11.90%	0.41%
29	Model 1	Linear Ratio	Diz-Conte	NA	11.95%	11.97%	0.03%
30	Model 1	Linear Ratio	None	NA	11.99%	11.99%	0.00%

The results of Table 1 establish the following points:

• **Independent Variable Specification:** The highest errors are associated with model 1 (the simple model proposed by Conte [1989]), and the lowest are associated with model 4 (the "full" model). Results for the other two models seem mixed.

• **Dependent Variable Specification:** Logistic dependent variables appear to perform somewhat better than the linear dependent variables.

• **Quarterly Interpolation Method:** Not interpolating produces higher total errors than either of the quarterly interpolation techniques. The Diz-Conte method results in lower quarterly interpolation error, but the SAS Spline has lower estimation error, so that they are very similar in size of total error.

• **Treatment of Irrelevant Variables:** Higher errors are found in the unrestricted models (where irrelevant variables are not removed).

The greatest gains in forecast accuracy are associated with changes in the independent variable specification. Compare lines 1, 9, 12, and 25 in Table 1. These four lines are alike in that they all have a logistic specification of the dependent variable, all use the SAS spline quarterly interpolation, and all drop irrelevant

independent variables; they differ only in the independent variable specification. Model 1 has an APE95 of 11.8 percent, model 2 falls to 5.15 percent, model 3 is 5.41 percent, and model 4 lies at 2.64 percent—a reduction of over 9 percentage points as one moves from the simplest model to the model with the most independent variables. Note that model 3 does not perform as well as model 2; the gains in accuracy in model 4 apparently owe more to the presence of lagged dependent variables than they do to the employment ratio terms.

The following tables report results for model 4, the best overall specification of independent variables. These tables supply some insight into how errors are affected by the following factors: quarterly interpolation technique, specification of dependent variable, and treatment of irrelevant variables.

TABLE 2

Quarterly Interpretation Method: 95th Percentile Error, Comparing All Versions of Model 4

Quarterly Interpolation Method	Absolute Percent Error Total	Absolute Percent Error Estimation	Absolute Percent Error Quarterly Interpolation
None	6.46%	6.46%	0.00%
Diz-Conte	3.20%	3.19%	0.03%
SAS-Spline	3.19%	3.10%	0.41%

The worst choice is clearly no interpolation. Of the two interpolation techniques tested, the Diz-Conte method creates only about one-fifteenth of the quarterly interpolation error of the SAS spline procedure. However, the SAS spline creates smaller estimation error, and it appears that often the SAS spline error has a sign opposite of the estimation error, so that the total error is slightly smaller (a difference of 0.01 percentage points) for the SAS spline method. Overall, there is so little difference among the methods that it seems one could justifiably choose one or the other based on convenience.

TABLE 3

Dependent Variable Specification: 95th Percentile Error, Comparing All Versions of Model 4

Dependent Variable Specification	Absolute Percent Error Total	Absolute Percent Error Estimation	Absolute Percent Error Quarterly Interpolation
Logistic	4.57%	4.56%	0.30%
Linear Ratio	4.77%	4.76%	0.30%

The logistic specification outperforms the linear, reducing total APE95 about 0.2 percentage points. The use of the logistic specification might be especially important in cases where, for whatever reason, one is unable to drop irrelevant variables.

TABLE 4

Treatment of Irrelevant Variables: 95th Percentile Error, Comparing All Versions of Model 4

Treatment of Irrelevant Variables	Absolute Percent Error Total	Absolute Percent Error Estimation	Absolute Percent Error Quarterly Interpolation
Restricted	4.26%	4.25%	0.30%
Unrestricted	5.04%	5.03%	0.30%

Dropping irrelevant variables lowers total APE95 by 0.79 percentage points. This effect is four times as great as the effect of using the logistic specification.

Tables 1–4 establish that the best overall specification is model 4, purged of irrelevant variables, with a logistic dependent variable, and interpolated with the SAS spline technique. The following two tables use the simulation results for this best model to examine the effect of some non-specification factors on APE95.[5] One item of particular interest would be the annual effects: i.e., the amount by which error varies according to the last year of the estimation period.

TABLE 5

Position in Business Cycle: 95th Percentile Error, Best Model

Last Year of REIS Data Before Forecast	Absolute Percent Error Total	Absolute Percent Error Estimation	Absolute Percent Error Quarterly Interpolation	Annual Real Income Growth Rate During Forecast Period (two years ahead)	Absolute Value of Difference in Annual Growth Rates in Forecast Period
1988	2.65%	2.63%	0.30%	1.8%	1.0%
1989	2.81%	2.78%	0.41%	1.2%	0.2%
1990	3.28%	3.06%	0.51%	3.3%	4.5%
1991	2.74%	2.78%	0.48%	4.4%	2.3%
1992	2.09%	2.05%	0.34%	3.6%	0.8%
1993	1.69%	1.66%	0.24%	4.1%	0.2%

Table 5 shows a difference of 1.58 percentage points in total APE95 between the years of 1990 (highest error) and 1993 (lowest error). In 1990, the forecasts estimated income for 1991 and 1992, the former a year of declining growth in Tennessee, the latter the first year of recovery. From these results it appears that the phase of business cycle during the forecast period might be an important determinant of forecast error. The last two columns in Table 5 shed some light on this supposition: they show the mean growth rate of Tennessee's TPI during the forecast period and the absolute value of the difference between the Tennessee TPI growth rates in the two forecast years. It appears that APE95 is highest not when forecasting into a period of low growth, but when forecasting into a two-year period where the two years deviate markedly from each other in phase of business cycle.

It is difficult to provide any guidance on the amount of error generated by phase of business cycle other than to caution that error increases as one forecasts into a period characterized by growth rate fluctuations. The simulation results do, however, allow some sense of how forecast error changes for size of sub-region and forecast year. Table 6 gives APE95 for each decile size of sub-region, for each of the two years of the forecast.

[5]An alternative approach would be that employed by West (1998), West and Fullerton (1996), and Lenze (1998) in assessing ex ante forecasts: regress the absolute percent error on a set of variables depicting the non-specification factors. This could allow an estimate of probable error conditional on the values of these variables.

TABLE 6

95th Percentile Error, Size of Sub-Region and Years Ahead, Best Model

Upper Limit on Size of Sub-State Region (% State Income)	Absolute Percent Error Total: First Year	Absolute Percent Error Total: Second Year	Absolute Percent Error Total: Both Years
0.38%	2.1%	3.7%	3.1%
0.54%	2.1%	3.4%	2.8%
0.67%	2.0%	3.1%	2.6%
0.81%	1.9%	3.1%	2.6%
0.97%	1.8%	3.0%	2.5%
1.17%	1.8%	3.1%	2.7%
1.57%	1.9%	3.1%	2.6%
2.16%	1.8%	3.0%	2.6%
3.41%	1.6%	2.9%	2.4%
33.19%	1.7%	2.6%	2.3%
All Sub-Regions	1.9%	3.1%	2.6%

The first of the two forecast years has total APE95 about 1.2 percentage points lower than the second year. Between the largest decile and smallest decile region there is a difference of 0.8 percentage points in total APE95. Table 6 can be used as a guide for practitioners employing this specification (model 4, purged of irrelevant variables, with a logistic dependent variable, and interpolated with the SAS spline technique) for other states. Find the size of sub-region relative to the state in the left-hand column. Read across the row to find the APE95 in the first year and in the second year: 95 percent of the time, one's estimate should have an error lower than that indicated.

Conclusions

The simulation results establish that model 4 is the best specification of the independent variables; model 4 reduces APE95 by 2.5 percentage points from the next best alternative, model 2. Both of these vastly outperform model 1 (the simple model proposed by Conte [1989]), which has an APE95 over nine percentage points higher than model 4.

Using model 4, error can be further reduced by dropping irrelevant variables (APE95 declines by 0.8 percentage points) and by using a logistic specification of the dependent variable (APE95 drops by 0.2 percentage points). However, there is virtually no difference in APE95 between the two quarterly interpolation techniques tested here. The best overall specification is model 4, purged of irrelevant variables, with a logistic dependent variable, and interpolated with the SAS spline technique.

The results also show that a number of non-specification factors affect forecast error. These include annual effects, size of sub-state region, and whether one is forecasting the first or second of the two years in the horizon. Using only the simulation results for the best model, APE95 was calculated to study these factors. For the annual effects—which are here interpreted as having to do with either structural change or the phase of the business cycle—the difference in APE95 between the lowest year and the highest is 1.6 percentage points. Next in

importance is the horizon year: the first year has an APE95 1.2 percentage points lower than the second year. Finally, size of sub-state region plays a role in determining error, with smaller error for larger regions. After ranking sub-state regions by size, the difference in total APE95 between the highest and lowest deciles is 0.8 percentage points.

REFERENCES

Bailey, Wallace K. "Comprehensive Revision of Local Area Personal Income, 1969-95." http://www.bea.doc.gov/bea/ar/0997rem/maintext.htm. Bureau of Economic Analysis, 1997.

Boor, Carl de. *A Practical Guide to Splines.* Applied Mathematical Sciences, vol. 27. New York: Springer-Verlag, 1978.

Conte, Michael A. "A Simple Method for Estimating Quarterly Local-Area Personal Income." *Journal of Regional Science* 29(1989), 89-105.

Diz, Adolfo C. "Money and Prices in Argentina, 1935-1962." In David I. Meiselman (ed.) *Varieties of Monetary Experience.* Chicago: University of Chicago Press, 1970.

Eff, E. Anthon. "Evaluating State-Level Forecasts: A Gross State Product Example." Paper presented at the 37th Southern Regional Science Association Meeting, Savannah, Georgia, April 1998.

Lenze, David G. "An Evaluation of *Ex Ante* Sub-State Long-Term Economic Forecast Accuracy and Efficiency." Paper presented at the 37th Southern Regional Science Association Meeting, Savannah, Georgia, April 1998.

Nonweiler, T.R.F. *Computational Mathematics: An Introduction to Numerical Approximation.* New York: Halsted Press, 1984.

SAS Institute Inc. *SAS/ETS User's Guide, Version 6.* First ed. Cary, North Carolina: SAS Institute, Inc., 1988.

U.S. Department of Commerce. Bureau of Economic Analysis, Regional Economic Analysis Division, 1997. http://www.bea.doc.gov/.

U.S. Department of Labor. "BLS Handbook of Methods. Chapter 5: Employment and Wages Covered by Unemployment Insurance." Bureau of Labor Statistics. 1997a. http://www.bls.gov/cewchap5.htm.

_____. "BLS Handbook of Methods. Chapter 2: Employment, Hours, and Earnings from the Establishment Survey." Bureau of Labor Statistics. 1997b. http://www.bls.gov/790meth.htm.

Waggoner, Daniel F. "Spline Methods for Extracting Interest Rate Curves from Coupon Bond Prices." *Federal Reserve Bank of Atlanta, Working Paper 97-10.* November 1997.

West, Carol Taylor. "Structural Regional Factors that Determine Absolute and Relative Accuracy of U.S. Regional Labor Market Forecasts." Paper presented at the 37th Southern Regional Science Association Meeting, Savannah, Georgia, April 1998.

West, Carol Taylor and T.M. Fullerton, Jr. "Assessing the Historical Accuracy of Regional Economic Forecasts." *Journal of Forecasting* 15(1996), 19-36.