IMPUTING MISSING DATA AND THE BOOTSTRAP:
SOME PRELIMINARY RESULTS

Irma A. Gomez, Diana M. Burton, and H. Alan Love¹

Abstract: Gathering information on forest inventories is expensive, but lack of data inhibits forestry sector modeling. Most published work has focused on drawing broader inventory estimates from small survey plot data. Other studies use predictions from linear equations with parameters derived from ordinary least squares estimation. This research develops a method to impute missing inventory and growth observations when survey plot data are not available, using cross-section time-series inventory data. A one-way error component model is estimated and missing inventory values are imputed using an optimally weighted combination of forward and backward projections. Confidence intervals are formed using the bootstrap. Preliminary results for a region in Louisiana are presented.

Introduction

Gathering inventory information in many resource industries is expensive. On-the-ground surveys of forest biomass are performed in the U.S. only every six to ten years. Lack of data for years between ground surveys seriously hinders resource modeling and policy analysis of resource markets. As growing populations and economic development put increasing pressure on resource stocks worldwide, accurate information about these resources becomes crucial. There is a need for methods to impute missing inventory and other needed data from existing information, and to obtain an idea about the reliability of such estimates from confidence intervals.

Literature

Most attempts to impute missing inventory data for years between field surveys emphasize making broader inventory estimates from observable small sample plot data. For example, Lynch (1995) estimates components of forest growth for individual plots. Often, however, annual survey data are not available and these methods cannot be used.

Other studies use predictions from an ordinary least squares (OLS) equation to impute missing tree growth data between survey periods. Starting from a beginning inventory level, missing stock data are imputed using predicted growth in conjunction with the inventory identity. While predictions from these models are unbiased, they often suffer from being incompatible with observed ending inventory level (Flewelling 1981).

The Bootstrap Method

This paper develops a method to impute missing inventory and growth data when survey plot data are not available. The method utilizes available cross-sectional time-series data (panel data) to estimate growth and inventories. A one-way error component estimator is used to predict growth in periods

¹Respectively, Assistant Professor, Department of Economics, Instituto Tecnologico y Estudios Superiores de Monterrey, Monterrey, Mexico; Assistant Professor, Department of Forest Science; and Associate Professor, Department of Agricultural Economics, Texas A&M University, College Station, TX 77843. Paper presented at the 1996 Southern Forest Economics Workshop, Gatlinburg, Tennessee, March 27-29.
between field surveys. Compatibility among beginning-period inventories, ending-period inventories, and predicted growth is maintained using optimally weighted forward and backward forecasts based on observed beginning-period and ending-period inventories. A bootstrap resampling method is used to compute small sample confidence intervals for imputed inventory values. An application to regional data for Louisiana is presented.

**Model Specification**

The model uses the cross-section time-series inventory identity:

\[ S_{kct} = S_{kct-1} + G_{kct-1} - H_{kct-1} \]  

(1)

for resource type \( k = 1 \) to \( K \); individual region \( c = 1 \) to \( C \); and time period \( t = 1 \) to \( T \). \( S \) is stock level, \( G \) is net growth and \( H \) is harvest. Harvest, \( H_{kct} \), is observed in each period. Thus, predicting growth, \( G_{kct} \), is equivalent to predicting change in stocks, \( S_{kct} - S_{kct-1} \). This leads to computation of the growth equation:

\[ G_{kct} = \alpha_{0kc} + \alpha_{1kc} z_{kct} + \alpha_{2kc} S_{kct} + e_{kct} \]  

(2)

where \( e_{kct} \) is a stochastic error term, \( z_{kct} \) is a vector of observable exogenous variables related to resource growth of type \( k \) in geographic area \( c \) at time \( t \), like weather, resource prices, and management costs, and \( \alpha_{id} \) are unknown parameters.

The model uses observable data. Harvests are observed annually. Stocks are observed at multi-year increments. Using observed stock data, net growth can be observed as a total for multi-year periods. The average growth per year can be calculated from observed total growth. The growth equation can be represented as averaged growth over time intervals between observed resource stocks:

\[ \bar{G}_{kct} = \alpha_{0kc} + \alpha_{1kc} \bar{z}_{kct} + \alpha_{2kc} \bar{S}_{kct} + \bar{e}_{kct} \]  

(3)

for intervals \( \tau = 1 \) to \( T \). \( \bar{G}_{kct} \), \( \bar{z}_{kct} \), \( \bar{S}_{kct} \), and \( \bar{e}_{kct} \) are average annual values for the interval \( \tau \). Since average stocks are unavailable, beginning stocks \( S_{kct}^0 \) can be substituted.

In an error-components model, the random error term \( \bar{e}_{kct} \) has two components. \( \mu_{kc} \) represents an individual regional effects error term and \( v_{kct} \) represents a remaining random error term. Average growth becomes:

\[ \bar{G}_{kct} = \alpha_{0kc} + \alpha_{1kc} \bar{z}_{kct} + \alpha_{2kc} S_{kct}^0 + \mu_{kc} + v_{kct} \]  

(4)

An average growth equation is formulated for each resource type \( k \) and region \( c \), and is estimated over time interval 1 through \( T \). The resulting set of equations is estimated using Zellner's Seemingly Unrelated Regression (SUR) technique as adapted for an error components (EC) model, as in Baltagi (1995) and in Kinal and Lahiri (1990).
Forecasts

Coefficient estimates from the ECSUR model are used to impute missing observations between inventories. Annual growth forecasts are computed from the starting inventory level and annual observations of z (observable variables related to growth). Three types of forecasts can be used: forward, backward and combined. The variance of the forecast is expected to increase for periods farther from the origin point of the forecast. The combined forecast, a weighted combination of the forward and backward forecasts, is expected to perform best.

The forward forecast is easily derived:

$$f^f_{t+m} = (1 + \alpha_2)^m S_t + \sum_{j=1}^{m} (1 + \alpha_2)^{m-j} (\alpha_0 + \alpha_1 z_{t-j-1} - h_{t-j-1})$$

(5)

where \( t \) is the forecast starting period and \( m \) is the number of forecast periods. The backward forecast is similarly:

$$f^b_{t-m} = \frac{1}{(1 + \alpha_2)^m} \left[ S_t - \sum_{j=1}^{m} (1 + \alpha_2)^{m-j} (\alpha_0 + \alpha_1 z_{t-m+j-1} - h_{t-m+j-1}) \right]$$

(6)

The combined forecast is a weighted combination of the forward forecast \( f^f_t \) and the backward forecast \( f^b_t \) (Diebold and Pauly 1987). Combined forecasts are restricted to \( C_t = \phi f^f_t + (1 - \phi) f^b_t \). The weight is calculated so that the 1-step ahead combined prediction error satisfies the equality \( e_t^{c} = \phi e_t^f + (1 - \phi) e_t^b \) because \( e_t^c = y_t - C_t \). Thus:

$$\text{var}(e_t^{c}) = \phi^2 \sigma_1^2 + (1 - \phi)^2 \sigma_2^2 + 2 \phi (1 - \phi) \sigma_{12}$$

(7)

and the weight \( \phi \) that minimizes this expression is given by:

$$\phi^* = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 + 2 \sigma_{12}}$$

(8)

Forecast variance and covariance can be computed using the delta method. Denoting the forecasting equations for period \( t \) as \( f^q_t(X\alpha) \) where \( q \) is forward, backward and combined forecasts, the forecast variance-covariance matrix for each time period is:

$$\sum_t = \left[ \begin{array}{c} \frac{\partial f^q_1(X\alpha)}{\partial \alpha} \\
\quad \vdots \\
\frac{\partial f^q_{\delta}(X\alpha)}{\partial \alpha} \\
\end{array} \right] \sum_{\delta} \left[ \begin{array}{c}
\frac{\partial f^q_1(X\alpha)}{\partial \alpha} \quad \ldots \quad \frac{\partial f^q_{\delta}(X\alpha)}{\partial \alpha} \\
\end{array} \right]$$

(9)
where $\Sigma$ is the variance-covariance matrix associated with $\hat{\alpha}_{ECSUR}$. $\Sigma_i$ is computed for each time period.

**Bootstrapping**

The bootstrap method can be used to establish confidence intervals for the forecasts. Confidence intervals are necessary to evaluate small sample properties of the imputed data. Bootstrapping requires that the model is assumed correctly specified and the population of estimated residuals can be resampled with replacement to generate new pseudo samples of observations. After each resampling, the estimation is performed again. The sampling distribution of the bootstrap estimators, conditional on the data, then mimics the small sample distribution of the estimators (for a thorough description of the procedure see Efron (1994) and Freedman and Peters (1984a, 1984b). Bootstrapping has been widely used in various studies, but has only recently been used in panel data models in Simar (1992).

To obtain forecast confidence intervals requires bootstrapping the model parameters $\hat{\alpha}_{ECSUR}$ since the stochastic properties of the forecast only depend on the stochastic properties of $\hat{\alpha}_{ECSUR}$. To bootstrap the ECSUR model, we need to maintain the model error structure so that small sample standard errors generated from resampling are consistent with the model’s own assumptions. To implement the bootstrap, first compute ECSUR residuals $\varepsilon = \mathbf{y} - \mathbf{x} \hat{\alpha}_{ECSUR}$. Then resample $\varepsilon$ with replacement while preserving the variance-covariance structure. The error structure is preserved by randomly selecting individuals with replacement until the total of C individuals is obtained, and then randomly selecting T intervals with replacement until T are obtained. We then form 1,000 pseudo data sets $\tilde{\mathbf{y}}^* = \mathbf{x}\hat{\alpha}_{ECSUR} + \varepsilon^*$ and reestimate the model using each pseudo data set. Forecast missing growth and stock values using each forecast method. Small sample standard errors are constructed from the sample of parameter estimates and forecasts.

Confidence intervals can be constructed from the bootstrap standard error estimates for each forecast period or can be obtained directly from the sampling distribution of the forecast (Efron 1994).

**Empirical Application**

This method is applied to inventory and growth (average net annual growth) data for hardwood and softwood in Louisiana for the years 1964, 1974, 1984, and 1991. For estimation the parish inventory data are grouped into five regions consistent with the price regions designated by the Louisiana Department of Agriculture and Forestry. All other data are available annually. For estimation in the average growth equation (4), data are averaged over the periods 1964-1973, 1974-1983, and 1984-1991 ($\tau = 1$ to 3).

Growth of trees ($k = $ hardwood and softwood) in each region ($c = 1$ to 5) of Louisiana is specified as a function of stumpage price and management costs formed into a Divisia price index, initial inventory, and growing season temperature and precipitation. The initial inventory is equal to the inventory of trees reported in the last survey. Each variable contains an observation for each parish for each $\tau$. The a priori expected sign on temperature is negative while the expected signs on precipitation, stock and price index are all positive.

Roy-Zellner tests indicated that the data could be pooled into two groups for estimation. Group A contains the northwest, southwest and southeast regions of Louisiana and is generally in loblolly-shortleaf pine forests. Group B contains the north and south delta regions and is largely oak-gum-cypress forests. Pooling is justifiable when both time series regressions across regions, and cross-region regressions over time are stable. Tests did not reject the stability of the cross-regions.
Table 1. ECSUR and bootstrapped estimates for northwest, southwest and southeast Louisiana.

<table>
<thead>
<tr>
<th>Equation Variable</th>
<th>ECSUR Estimates</th>
<th>Bootstrapped Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Standard Error</td>
</tr>
<tr>
<td>a. Hardwood</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>3.9606</td>
<td>70.9188</td>
</tr>
<tr>
<td>Average precipitation</td>
<td>-1.0260</td>
<td>0.2799</td>
</tr>
<tr>
<td>Average temperature</td>
<td>-0.0624</td>
<td>0.8356</td>
</tr>
<tr>
<td>Observed beginning stocks</td>
<td>0.0226</td>
<td>0.0038</td>
</tr>
<tr>
<td>Price index</td>
<td>686.7330</td>
<td>507.7910</td>
</tr>
<tr>
<td>b. Softwood</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>522.9450</td>
<td>126.9570</td>
</tr>
<tr>
<td>Average precipitation</td>
<td>0.5043</td>
<td>0.7930</td>
</tr>
<tr>
<td>Average temperature</td>
<td>-6.5006</td>
<td>1.6170</td>
</tr>
<tr>
<td>Observed beginning stocks</td>
<td>0.0399</td>
<td>0.0052</td>
</tr>
<tr>
<td>Price index</td>
<td>31.3974</td>
<td>245.4480</td>
</tr>
</tbody>
</table>

regressions over time for both groups A and B. The stability of time-regression across regions was not tested due to the lack of degrees of freedom.

A Hausman (1978) specification test was used to test whether the unobservable individual-region effects are uncorrelated with the exogenous variables. Following Baltagi (1995), the test was performed using GLS and OLS parameter estimations. The test fails to reject the null hypothesis that $E(u_{it}/X_{it}) = 0$. Thus, the individual-region effects are uncorrelated with the exogenous variables for both groups A and B.

Table 1 shows the coefficients for the ECSUR model estimation with standard errors and the mean bootstrapped estimates with bootstrap standard errors for Group A. In all cases, the bootstrap procedure significantly reduced the standard errors. For hardwoods, the coefficient on precipitation is negative and significant. The coefficient on temperature is negative, but insignificant. Coefficients on beginning stocks are significant. The coefficient on the price index is not significant in the ECSUR estimate, but is so for the bootstrapped estimate. The signs on the coefficients are as expected for softwoods. Temperature and beginning stocks are significant in both equations. Precipitation is significant only in the bootstrapped model.

Figures 1a, 1b, and 1c present the forward forecast, the backward forecast and the combined forecast results, respectively, for the southwest region. Each forecast is framed by a 95 percent confidence interval developed using the bootstrap standard errors. The confidence intervals widen substantially as the forecast moves away from the point of origin for both the forward and backward forecasts. The combined forecast has very narrow confidence intervals, barely visible on the graph.

**Forecast Validation**

In an attempt to assess the accuracy of the forecast, the bootstrap was reestimated using years 1974 and 1991 as pivotal years for the forward, backward, and combined forecasts. Forecasts for the year
Figure 1a. Forward forecast with 95 percent confidence interval.

Figure 1b. Backward forecast with 95 percent confidence interval.

Figure 1c. Combined forecast with 95 percent confidence interval.

Figure 1d. Validation forecast.
1984 were obtained. The result for the combined forecast, Figure 1d, shows that the annual forecast line indeed moves toward the 1984 actual value. The estimated value is 95 percent of the actual value.

Conclusion

This paper presents preliminary results for a method to impute missing inventory data points using panel data techniques. The bootstrap is used to generate confidence intervals for the estimates. Statistical results for a region of Louisiana are presented. The combined forecast is a superior method of imputing missing inventory. The standard errors of the estimate generated from the bootstrap are useful in assessing forecast reliability and in determining the impact of estimate reliability on other models in which the inventory estimate is used.

Literature Cited


