January 2007 Labor Estimates Methodology

USDA NASS Research and Development Division
Stephen Busselberg, Mathematical Statistician
3251 Old Lee Highway, Room 305
Fairfax, Virginia 22030
Phone: 703-877-8000, Ext. 141

Background

Historically, NASS has used probability sampling to produce labor estimates. NASS labor estimates for January 2007 consist of multivariate time series forecasts within domains formed by the following crossed factors: worker type, labor unit, and geographic region. Worker type is estimated at the level of field, field and livestock, and all hired workers. Labor unit includes the wage rate, number of workers, and number of hours worked. Geographic regions are composed of groupings of states that lie within close proximity to each other and have similar agricultural production. Three of the regions contain only one state and those regions are Hawaii, California, and Florida. The United States is also considered a region.

Regions	Labor Unit	Worker Type
Northeast I	Wage Rates	All Hired
Northeast II	Number of Workers	Field & Livestock
Appalachian I	Hours Worked	Field
Appalachian II		
Southeast		
Lake		
Corn Belt I		
Corn Belt II		
Delta		
Northern Plains		
Southern Plains		
Mountain I		
Mountain II		
Mountain III		
Pacific		
California		
Florida		
Hawaii		
US		

Three worker types, three labor units, and nineteen regions equate to 171 required estimates.

Historical Data

The inputs for the labor time series models consist of published quarterly data for each region, labor unit, and worker type. These data are available in an uninterrupted time sequence from January 1997 to December 2006 for hours worked and number of workers; and from January 1989 to December 2006 for wage rates. The published estimates from these years and quarters served as the inputs to the time series modeling, and the ultimate outputs of the models served as survey indications, which were used in setting the official estimates for January 2007. The time series model used in generating the inputs to the estimation process produced very consistent results that reflect the periodic trends of the published estimates.

Methodology

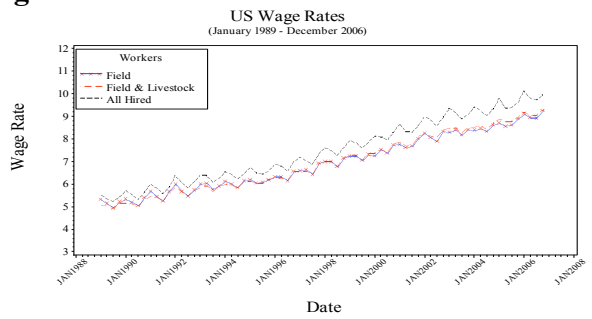
The chosen class of models for the January 2007 labor estimates is Vector Autoregressive Moving Average (VARMA). The endogenous dependent variables are the levels of worker type. Wage rates, number of workers, and hours worked are assumed uncorrelated. Schematic plots of the autocovariance structure and minimum corrected Akaike's Information Criterion (AICc) for test values of the autoregressive and moving average orders show that the autocovariance structure is similar across many of the regions.

If the United States region is considered an averaging out or smoothing over the other regions, its correlation structure should be representative of all the other regions combined and give a good indication of the model structure to be used for all regions for a given labor unit. Assuming the autocovariance structures between the worker types within

each labor unit are the same across all regions, at most three independent time series model structures were considered appropriate – one for each level of labor unit. The same model structure could then be used for each region within each labor unit. Only the model parameters differed for computation of the step-ahead estimates.

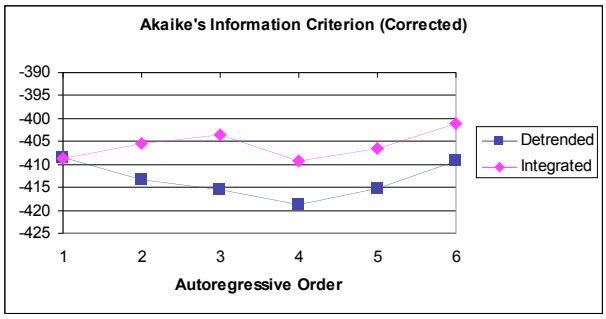
Wage Rates

Figure 1



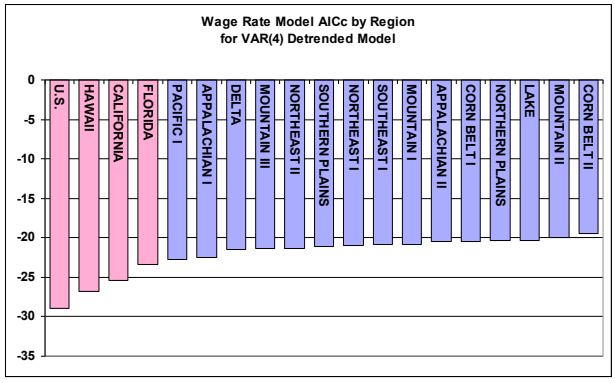
The graph in Figure **1** demonstrates that the levels of worker type would be good candidates for the endogenous variables in a VARMA time series model as they are highly correlated over time. Minimum AICc selection criteria for the preliminary orders of the autoregressive (AR) and moving average (MA) polynomials produced an order of zero for the MA polynomial for all regions. Further investigation results conclude any MA polynomial of positive order does not produce maximum likelihood coefficients due to lack of convergence. This condition holds for estimation of all labor unit levels and therefore all final models contain only an AR polynomial or an autoregressive order greater than zero.

Figure 2



The sums of the AICc statistics over the regions for preliminary AR orders are shown in the Figure **2** for two different transformation options for stationarity. Field workers, field and livestock workers, and all hired workers appear to be nonstationary with a linearly increasing trend. Options for transforming the data to a wide sense stationary series include integration through a differencing matrix or a vector of detrending parameters. The annual cycle of the wage rates suggests a natural seasonal differencing of four data points (quarters) and equates to a first difference lag 4. Figure **2** suggests that a detrended model is superior to an integrated model for all AR orders, and that a detrended model of AR order 4 provides the best fit.

Figure 3



The chart in Figure **3** breaks down the summed AICc for a VAR(4) model by region. The region with the lowest AICc statistic is the U.S. It is interesting to note that the three regions that are comprised of only one state make up the next three lowest AICc statistics. The final

wage rate model takes the following form for all regions:

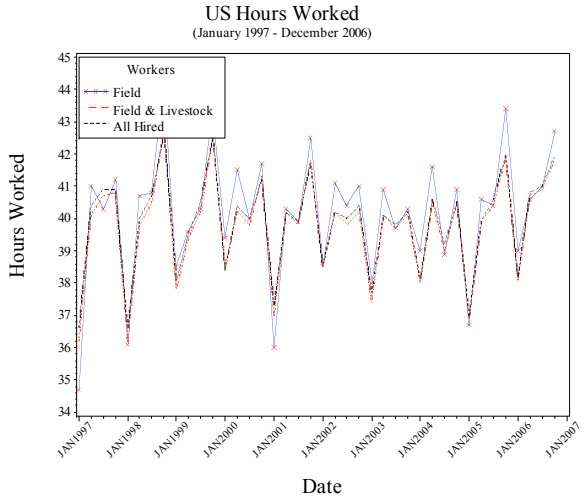
$$\mathbf{x}_t = \exp\left(\mu + \beta t + \sum_{k=1}^{P} \Phi_k \left(\ln \mathbf{x}_{t-k} - \mu\right)\right)$$

where Φ is a 3x3 matrix of autoregressive coefficients, β is a 3x1 vector of linear trend coefficients, μ is a 3x1 vector of centering coefficients, and P is equal to the model AR order of 4. An AR order of 4 uses the last historical year of wage rate data to estimate a step-ahead forecast. This model reduces the autocorrelation in the residuals much more so than models of other orders, including subsets such as seasonal models ($\Phi_1 = \Phi_2 = \Phi_3 = \mathbf{0}$).

Hours Worked

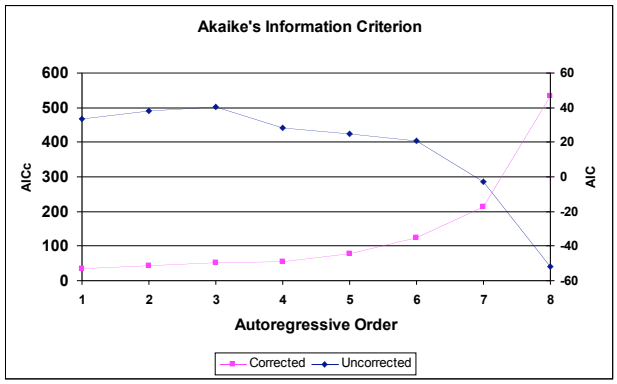
A plot of US hours worked for all three worker types shows a process that appears stationary

Figure 4



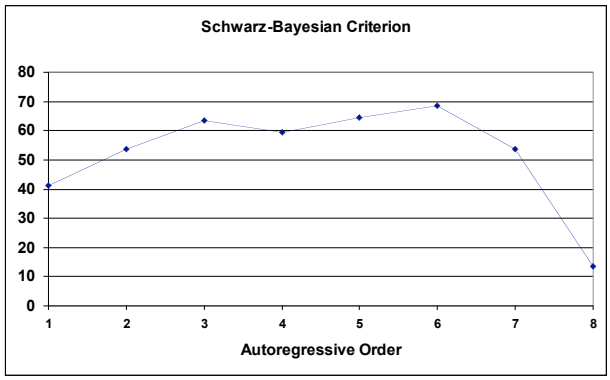
containing a seasonal element similar to the wage rates with a period of four quarters. Schematic plots of the autocorrelation function also show this four quarter seasonal periodicity. Although the series in Figure **4** appears stationary, a seasonal differencing of lag 4 (implying a yearly differencing) is recognizably appropriate from the autocorrelation function plots. Preliminary estimates of an autoregressive order are at first unclear, as there is no concrete minimum in the sum of information criteria over the regions.

Figure 5



The summed AICc in Figure **5** suggests an AR order of zero, and the uncorrected AIC suggests the higher the order, the better the fit. An examination of the Schwarz-Bayesian Criterion shown in Figure **6** is equally unclear.

Figure 6



However, adding a seasonal AR order to the seasonally differenced hours worked series produces a minimum at 4.

Figure 7

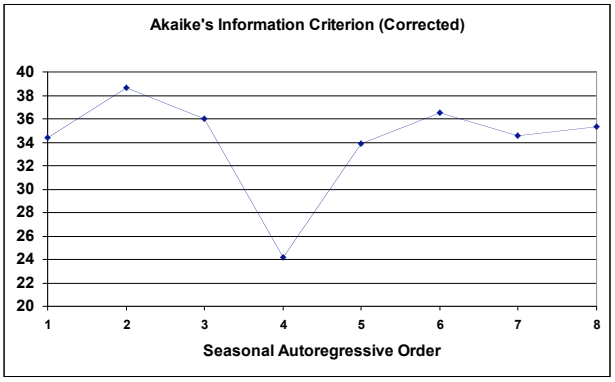
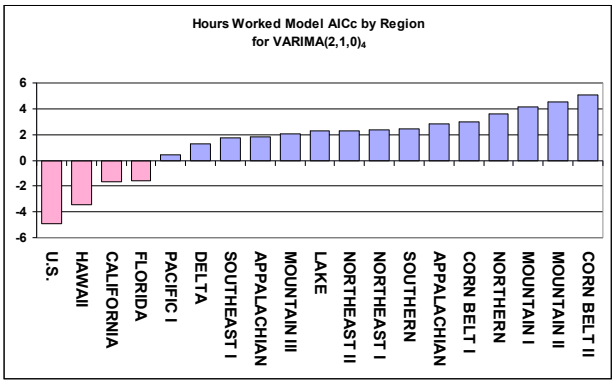


Figure **7** implies a model where the coefficient matrices of lags 1-3 are zero. Further examination of the autocorrelations in the residuals leads to a seasonal AR order of 2 with a seasonal component of 4, and the resulting model is VARIMA$(2,1,0)_4$.

Figure 8



The AICc of the individual regions in Figure **8** show a strikingly similar pattern to the wage rate model (Figure **3**) in the rankings of the AICc's. The three states and the U.S. that make up their own region have the lowest scores, hence best fit.

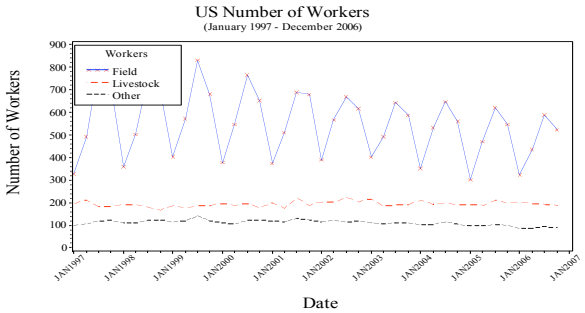
Number of Workers

The levels of labor type are a summation of parts with embedded variables for livestock workers and other workers. Field workers are present in all three levels. Livestock workers could be determined by subtracting field from field & livestock. "Other workers" could be calculated by subtracting field & livestock from all hired workers.

Figure 9

A graph of the data for levels of labor type is displayed in Figure **9**. There is a strong correlation between all three parts which could be attributed to either the common elements in each part, or actual correlations between the field workers, livestock workers, and other workers, or both. The transformed variables which show the independent levels for field workers, livestock workers, and other workers are displayed in Figure **10**. This dissection explains the field workers as the single major contributor to the periodicity within the data and to the correlations between levels shown in Figure **9**.

Figure 10



Due to the size of the field worker numbers relative to livestock and other workers, it is not apparent in the graph above that all three variables are still highly correlated over time. Mathematically, the estimates and resulting information criterion computed from either data source are the same, whether the original nested data in Figure **9** is modeled, or the independent transformed dataset shown in Figure **10** is used. However, residual diagnostics prove more conclusive with the original nested dataset of field, field & livestock, and all hired workers as the final model almost completely eliminates significant correlations. The final model structure for the number of workers is similar to the structure of the wage rates model, except that the model includes an integration vector instead of linear trend parameters. The final model in Box Jenkins notation is a VARIMA(4,0,0)x$(0,1,0)_4$.

The AICc statistics for each region present a different picture for number of workers than for hours worked or wage rates. The sheer worker size differences between the regions create AICc statistics on a non-comparable scale. The U.S. and California regions are relatively larger compared with the rest of the regions, and Hawaii is smaller. This size relationship makes comparison of the model structure fit not useful for diagnostics. However, almost complete elimination of autocorrelation in the residuals by region is supportive of the model fit.

The region Corn Belt II produces an indication that is too low in comparison to historical data with the chosen model. For this reason, this region is fitted with a VARIMA(1,0,0)x$(0,1,0)_4$ model which produces a more realistic indication.

APPENDIX: MODEL FITS

Wage Rates

VARIMA(4,0,0)

$$\mathbf{x}_t = \exp\left(\mu + \beta t + \sum_{k=1}^{p} \Phi_k \left(\ln \mathbf{x}_{t-k} - \mu\right)\right)$$

Hours Worked

VARIMA$(2,1,0)_4$

$$\mathbf{x}_t = \mu + \mathbf{x}_{t-s} + \sum_{k=1}^{P} \Phi_k \left(\mathbf{x}_{t-sk} - \mathbf{x}_{t-s(k+1)} - \mu\right)$$

Number of Workers

VARIMA$(4,0,0)x(0,1,0)_4$

$$\mathbf{x}_t = \mu + \mathbf{x}_{t-s} + \sum_{k=1}^{p} \Phi_k \left(\mathbf{x}_{t-k} - \mathbf{x}_{t-k-s} - \mu\right)$$

Explanation of Box - Jenkins (1970) notation

VARMA$(p,d,q)x(P,D,Q)_s$

Parameter	Description
p	Autoregressive Order
d	Differencing Order
q	Moving Average Order
P	Seasonal Autoregressive Order
D	Seasonal Differencing Order
Q	Seasonal Moving Average Order
s	Seasonal Parameter