Forecast Evaluation for Multivariate Time-Series Models: The U.S. Cattle Market

Timothy Park

A set of rigorous diagnostic techniques is used to evaluate the forecasting performance of five multivariate time-series models for the U.S. cattle sector. The root-mean-squared-error criterion along with an evaluation of the rankings of forecast errors reveals that the Bayesian vector autoregression (BVAR) and the unrestricted VAR (UVAR) models generate forecasts which are superior to both a restricted VAR (RVAR) and a vector autoregressive moving-average (VARMA) model. Two methods for calculating a test evaluating the ability to forecast directional changes are implemented. The BVAR models and the UVAR model unambiguously outperform the VARMA model in forecasting directional change.

Key words: Bayesian vector autoregression, forecast evaluation, Henriksson-Merton test, multivariate time-series models.

A prime objective of this paper is to present a set of rigorous diagnostic tools which can be used to evaluate forecasting performance and, in turn, guide the selection of an appropriate forecasting model. Two criteria originally suggested by Granger and Newbold are presented to motivate the techniques used to evaluate forecasting performance. The methods put forth complement work by Kaylen (1988) which examined a different set of techniques for comparing forecasts. The complete set of techniques implemented in this paper has not been applied previously.

The first criterion proposed by Granger and Newbold examines whether a set of forecasts is significantly better than its competitors. The root mean-squared error is the traditional statistic for comparing forecasts from alternative models. However, simple comparisons of root mean-squared errors provide limited information for guiding the selection of a model. To examine whether the observed differences in root mean-squared errors are statistically significant, a formal test is applied. A second test evaluates forecasting performance based on the relative size of forecast errors. The time-series models are compared based on rankings of the forecast errors using a test procedure developed by Stekler.

The second criterion proposed by Granger and Newbold attempts to assess the worth of the forecasts to a decision maker in an absolute sense, an evaluation that is often very difficult. Merton developed the theoretical foundations for one type of test which meets this criterion. The test evaluates the ability of time-series models to forecast directional changes in the variables of interest. The test is implemented in a nonparametric procedure following Henriksson and Merton and also using a regression framework developed by Cumby and Modest.

The forecasting performances of five multivariate time-series models which have been proposed as methods for generating optimal forecasts are investigated. The models examined here were chosen for two reasons. First, in previous empirical work these models have been shown to outperform other models in forecasting accuracy. Second, the model selection and specification procedures for each

Timothy Park is an assistant professor in the Department of Agricultural Economics at the University of Nebraska, Lincoln.

The author is indebted to Art Havenner of the University of California, Davis, for advice and comments and to a referee for detailed suggestions.

model are well defined, avoiding to a large degree the need for judgmental decisions in choosing an appropriate model.

First, a set of five time-series models is specified and identified. Then methods for evaluating the forecast performance of the models are developed and applied. This paper represents the first attempt to compare these models on a common data set gathered from the U.S. livestock sector.

Data

The livestock sector has been buffeted by a series of shocks in the past 15 years. Movements in cattle-on-feed, beef produced, and prices for steers have been exacerbated by changes in agricultural policy such as the Dairy Termination Program and grain embargoes along with structural shifts in the demand for meat. The impact of these events on the livestock sector provides a stringent test for the forecasting performance of time-series models.

The time-series models are estimated using four series from the U.S. livestock industry. The variables are: beef production, total cattle-on-feed, steer prices for 900- to 1,100-pound steers, and feeder steer prices for 600- to 700-pound feeders. The data series were chosen from a U.S. Department of Agriculture (USDA) model for the U.S. livestock industry developed by Stillman. Data used were available on a continuous monthly basis from January 1970 through October 1987. The beef production variable corresponds with the primary production sector of the USDA structural model. The total cattle-on-feed variable is the capital stock variable representing the total number of animals available for marketing. The price variables are key components in the USDA model. Steers are the most highly valued slaughter animals in the livestock sector. Feeder steer prices comprise the largest cost in cattle feedings for the cow-calf operator.

The specification of time-series models begins with series which are stationary. Yet formal tests for stationary series often have been neglected in formulating time-series models. A test procedure developed by Dickey and Pantula was used to check for stationarity. The test can be motivated by noting that hypotheses about the presence of unit roots can be tested based on parameters from the sequence of regressions specified in table 1.

This application of the procedure begins with the third difference of each series, following Dickey and Pantula's presentation which is appropriate for economic time series. Each third difference is regressed on a constant and the lagged value of the second difference for that series. Compared to the τ_μ table of Fuller, the t-statistics for the lagged second difference of each series in Step 1 are lower than the critical values. The null hypothesis that the third difference for each series is nonstationary is rejected.

In a sequential test procedure, the degree of differencing is decreased by one to examine whether the second difference is sufficient to induce stationarity. In Step 2, the third difference of each series is regressed on a constant, the lagged second difference, and the lagged first difference. Based on the t-statistics for both the lagged first difference and the lagged second difference, the null hypothesis that the second difference of each series is nonstationary is also rejected.

The test procedure continues by considering whether a first difference is sufficient to induce stationarity. The model for testing the null hypothesis uses the same set of variables as in Step 2, supplemented with the lagged value of the dependent variable. The appropriate model is presented in table 1 as Step 3. For the logarithm of beef production and the logarithm of cattle-on-feed, each of the t-statistics for the complete set of lagged variables indicates that no differencing is required. For the steer prices and feeder steer prices series, the t-statistics on the lagged value of the dependent variable in each equation indicate that the first difference is necessary to induce stationarity. Graphical analysis of the data confirms that the transformed series are stationary.

Estimation of Multivariate Vector Autoregressive Moving-Average Models

The first forecasting model estimated is a vector generalization of univariate autoregressive moving-average (VARMA) models proposed by Tiao and Box. Recent empirical work by Fackler and Krieger has shown that these models can substantially outperform unre-

Table 1. Dickey-Pantula Stationarity Tests for Times Series

	Dependent Variable[a]			
Explanatory Variables[b]	Beef Production	Cattle-on-Feed	Steer Prices	Feeder Prices
STEP 1:				
H_0: Three unit roots[c]				
Constant	0.000652 (0.092)	0.000135 (0.040)	0.000873 (0.00398)	−0.00631 (−0.031)
$DIF(Y^2_{t-1})$	−1.746 (−35.087)*	−1.301 (−18.086)*	−1.323 (−18.465)*	−1.453 (−21.554)*
STEP 2:				
H_0: Two unit roots				
Constant	0.00129 (0.296)	0.000720 (0.242)	0.1509 (0.814)	0.119 (0.658)
$DIF(Y_{t-1})$	−2.221 (−17.128)*	−0.637 (−7.520)*	−0.787 (−8.582)*	−0.699 (−7.438)*
$DIF(Y^2_{t-1})$	−0.636 (−8.883)*	−0.984 (−13.029)*	−0.927 (−8.582)*	−1.104 (−14.652)*
STEP 3:				
H_0: One unit root				
Constant	1.438 (2.885)	0.907 (3.841)	1.271 (1.799)	0.951 (1.565)
Y_{t-1}	−0.190 (−2.883)*	−0.099 (−3.838)*	−0.022 (−1.642)	−0.016 (−1.434)
$DIF(Y_{t-1})$	−2.031 (−14.187)*	−0.538 (−6.287)*	−0.769 (−8.369)*	−0.686 (−7.284)*
$DIF(Y^2_{t-1})$	−0.695 (−9.512)*	−1.064 (−14.067)*	−0.939 (−12.380)*	−1.112 (−14.764)*

[a] The dependent variable is the third difference for each series.
[b] Notation for the explanatory variables in the sequence of models: Y_{t-1} is the lagged value of the dependent variable; $DIF(Y_{t-1})$ is the lagged first difference of the dependent variable; $DIF(Y^2_{t-1})$ is the lagged second difference of the dependent variable.
[c] Asymptotic t-values are in parentheses. Asterisk indicates significance at .05 level. Critical value for $\tau_{.05} = -2.89$.

stricted VAR models on the basis of forecasting accuracy. The VARMA model is specified as

(1) $\Phi_p(L)Z_t = C + \Theta_q(L)a_t$, where
$$\Phi_p = I - \Phi_1 L - \ldots - \Phi_p L^p \text{ and}$$
$$\Theta_q = I - \Theta_1 L - \ldots - \Theta_q L^q,$$

where I represents the unit matrix, L is the lag operator, and p and q denote the orders of the autoregressive and the moving-average processes, respectively. For this model, C is a $k \times 1$ vector of constants, Φ_p and Θ_q are $k \times k$ matrices, Z_t is a k-element vector, and a_t is a k-element vector of identical and independently distributed random shocks that have mean zero and covariance Σ.

The tentative specification of the VARMA model begins with the sample cross-correlation matrices, which are used to identify the order of the moving-average process. If Z_t follows an MA(q) model, then the cross-covariances and cross-correlations are zero for $k > q$. Second, to identify autoregressive models, Tiao and Box proposed a generalized partial autocorrelation function, $P(\nu)$. If Z_t follows a vector AR(p) model, then $P(\nu) = 0$ for $\nu \geq p$. The partial autoregression matrices and associated statistics used for identifying pure autoregressive processes indicate that a low-order AR specification would not adequately model the data. The sample autocorrelations and cross-correlations for the series are persistently high, also ruling out a low-order MA model.

The extended sample cross-correlation (ESCC) is a key diagnostic tool developed by Tiao and Tsay to identify mixed models which contain both moving-average and autoregressive components. The procedures developed by Tiao and Tsay using the ESCC methodology lead to the identification of a VARMA(1,1) model from which forecasts are generated. The

Table 2. Likelihood Ratio Test Statistics for Lag Length in VAR Models

k_2	k_1	$M(k_2, k_1)^{a,b}$	Degrees of Freedom
1	0	471.11	16
2	1	129.99	16
3	2	64.26	16
4	3	33.64	16
5	**4**	24.30	16
7	6	36.64	16
8	7	43.95	16
9	8	54.10	16
10	**9**	13.75	16
11	10	28.42	16
12	11	38.21	16
9	4	183.55	80

Note: H_0: All coefficients in lag $k_1 + 1, \ldots, k_2$ equal zero. Boldface indicates candidate lag lengths for testing.
[a] $M(k_2, k_1)$ is approximately distributed as a χ^2 random variable with $m^2(k_2 - k_1)$ degrees of freedom.
[b] Critical value for $\chi^2_{16,.05} = 26.296$. Critical value for $\chi^2_{80,.05} = 101.879$.

final model is estimated by maximum likelihood.

Estimation of the Vector Autoregression Models

Two alternative versions of the VAR model are estimated. In the unrestricted VAR (UVAR) model, each variable in the system depends on lagged values of itself and lagged values of all the other variables. A common lag length is specified for the variables in the system.

The likelihood ratio test statistic developed by Tiao and Box is used to determine the order of the lag length for the UVAR model. The test starts with a maximum lag length of 12 months to capture any yearly pattern in the series for the cattle industry. For testing lag length k_1 versus k_2, the likelihood ratio test statistic is

$$(2) \quad M(k) = [N - .5 - m \cdot k][\ln S(k_1) - \ln S(k_2)],$$

where N is the effective number of observations, m is the number of endogenous series in the model, k is the order of the longer lag length in the model, and $S(k)$ is the determinant of the matrix of the residual sum of squares and cross products from the model with lags 1 through k. Under the null hypothesis that all of the coefficients in $k_1 + 1, \ldots, k_2$ are zero, $M(k)$ is asymptotically χ^2 with $m^2(k_2 - k_1)$ degrees of freedom.

The likelihood ratio test statistics for testing lags 1 through 12 are presented in table 2. The results of table 2 indicate that the lag lengths of four and nine are candidate lag lengths for the UVAR. When testing lag 9 versus lag 4, the null hypothesis that all of the coefficients in lags 5 through 9 were zero was not rejected. The lag length for the UVAR is set at nine.

Fackler and Krieger identified potential problems with using the UVAR model for forecasting in multivariate time-series models. Lengthening the lag length by one variable increases the number of estimated parameters by the square of the number of variables in the model. Choice of lag length may be constrained by the available degrees of freedom in the data set. A lag length which is underspecified leads to biased coefficient estimates. The specification that each variable in the system has an identical lag length is also restrictive and may be inappropriate.

An alternative method for specifying a type of VAR model is based on a method proposed by Webb to overcome these difficulties. The method, which was designed to restrict significantly the number of estimated coefficients relative to the UVAR model, is here termed RVAR. For each series, a search procedure designed to minimize a goodness-of-fit criterion is used to choose the lag length. Kaylen (1988) has shown that a similar model-fitting procedure using an alternative criterion for choosing lag length yields forecasts for the U.S. hog market which outperformed a variety of other models.

The procedure developed by Webb is designed to efficiently identify the specification that minimizes an appropriate goodness-of-fit criterion while limiting the role of judgment in the process. The optimal criterion for identifying the order of any type of VAR model should be based on the success of the criterion in identifying the most accurate forecasting model.

Lütkepohl showed that the Schwarz criterion was the most effective criterion for identifying the correct order of lag length. In a test of various criteria for forecasting performance, Engle and Brown demonstrated that the Schwarz criterion resulted in the smallest mean-squared forecasting error. Using the Schwarz criterion, the chosen model specification minimizes:

(3) $$Schwarz = T^{(K/T)} * ESS,$$

where T is the number of observations used in the model, K is the number of lags in the model, and ESS is the sum of squared residuals.

Selection of the lag length based on the RVAR specification proposed by Webb consists of the following steps. The starting specification for each equation contains lagged values of the dependent variable based on the Tiao-Box criterion. The lag length for each variable is increased by one period. A decrease in the Schwarz criterion determines whether the lag length for one variable is increased by one period. To avoid converging to a local minimum, additional lag lengths also are examined to see whether the Schwarz criterion declines. When a lower Schwarz criterion cannot be attained, no additional terms are added. For the final specification, the values in each lag are examined to see if removing a lag lowers the criterion. The final specifications of the UVAR and RVAR models are presented in table 3.

Table 3. Lags Structure for Vector Autoregression Forecasting Models

	Dependent Variable			
Lagged Variables	Beef Production	Cattle-on-Feed	Steer Prices	Feeder Prices
Unrestricted VAR				
Beef Production	9	9	9	9
Cattle-on-Feed	9	9	9	9
Steer Prices	9	9	9	9
Feeder Prices	9	9	9	9
Restricted VAR				
Beef Production	9		1	
Cattle-on-Feed		9		1
Steer Prices		1	9	
Feeder Prices	1			9

Estimation of the Bayesian Vector Autoregression

A modification of the VAR model imposes Bayesian priors for the parameter values and their underlying distributions. Litterman (1986a) showed that the Bayesian vector autoregression (BVAR) can produce forecasts which perform as well as structural econometric models. The Bayesian approach has been especially effective in dealing with specification uncertainty inherent in time-series modeling. Nickelsburg and Ohanian demonstrated that imposing Bayesian random-walk priors reduced forecast error variance even in the presence of misspecification. A final strength of the BVAR model has been the emergence of a consistent method for specifying the Bayesian priors, including formal statistical criteria for examining the performance of alternative specifications.

The model is written as an nth order autoregressive form for the n-vector of dependent variables denoted by Y

(4) $$Y_t = D_t + \sum_{j=1}^{m} B_j Y_{t-j} + e_t,$$

where D_t is the constant term for each component of Y. Let b_{ij} be the i,jth element of the autoregressive matrix B. Choice of lag length using the appropriate test statistic has been discussed previously.

The Litterman prior is based on the assumption that the behavior of most economic variables can be approximated as a random walk around an unknown, deterministic component. For each equation in the system, this specification is written as

(5) $$Y_t = D_t + B_j Y_{t-1} + e_t.$$

Additional assumptions on the B matrix are that: (*a*) the elements of the B matrix are jointly normally distributed; (*b*) the means of the b_{ij} elements are zero, except for b_{ii} which has a mean of one; and (*c*) the b_{ij} are uncorrelated across all i and j.

Following the specification adopted by Litterman (1986a, b), the initial own-lag coefficients, b_{ii}, are equal to one for both the series specified in logarithms and for the differenced series. Kaylen (1988) suggested that for series which are differenced, it may be appropriate to center initial own-lag coefficients on zero. Litterman (1986a) also noted that modifications to these values might be considered but argued that the forecasting performance should be relatively insensitive to the specification of the prior on the first lag of the dependent variable. Alternative specifications of the b_{ii} can be considered in future work.

The random-walk prior is supplemented with additional assumptions on the form of the distribution of the prior means. Variable lags further in the past have less explanatory power than the more recent lags; standard deviations on the lag coefficients decrease as the lag lengthens.

The standard deviations of the estimated parameters are specified based on the following information:

λ represents the constant overall tightness of the prior. Low values of λ imply a tight prior in which the distributions of the estimated coefficients are tightly spiked around the prior means.

γ_1 determines the rate at which standard deviations decrease on coefficients in the lag distributions. The decay pattern of the lagged standard deviations can take the form of an harmonic lag, a geometric lag, or no decay pattern. A tighter decay pattern implies that standard deviations on higher lags receive less weight and are more tightly restricted about prior means.

γ_2 represents the relative tightness on standard deviations of own lags of dependent variables compared to lags of other variables in the system.

Given the parameters (λ, γ_1, γ_2), the Litterman prior for the standard deviation of coefficient i,j at lag l is

$$(6) \qquad \delta_{ij}^l = -\begin{cases} \dfrac{\lambda}{l^{\gamma_1}} & \text{if } i = j \\[2ex] \dfrac{\lambda\gamma_2\hat{\sigma}_i}{l^{\gamma_1}\hat{\sigma}_j} & \text{if } i \neq j. \end{cases}$$

In the prior, $\hat{\sigma}_i$ is the standard error of the residuals from the univariate autoregressions for variable i of the lag length chosen for the VAR. The priors are scaled by a ratio of standard errors from the univariate autoregressions. The scaling is necessary so that the units in which the variables of the original series are measured do not bias the specification of the BVAR.

In the specification of the symmetric prior, the tightness parameters for the coefficients of variable i in equation j are the same for all i and j. The nonsymmetric prior relaxes this restriction, allowing varying degrees of interactions among the variables in the model based on additional information about the relationships of the variables. The specification of the nonsymmetric prior is based on a matrix of $\gamma_2(i, j)$ coefficients, which reflects the tightness on the coefficient of variable j in the ith equation of the system. Values of $\gamma_2(i, j)$ near one are chosen when the variables tightly interact; values near zero reflect less interaction among the variables.

The BVAR models are fit using a procedure developed by Bessler and Kling. The starting point for the BVAR specification is based on the lag length chosen by the UVAR model. A grid search for the settings of (λ, γ_1, γ_2) is conducted using values from (.00, .01, .25, .50, .75, 1) based on data over the period January 1970 through December 1981. Out-of-sample forecasts are generated for each parameter setting using the Kalman filter, as described in Doan and Litterman, for the period January 1982 through December 1984. The parameter setting for (λ, γ_1, γ_2) which minimizes the log determinant of the error covariance matrix over the out-of-sample period is (.50, .00, .50). The nonsymmetric prior uses the same value for $\lambda = .50$ and $\gamma_1 = .00$ with values of $\gamma_2(i, j)$ specified in table 4.

Selection of the BVAR model is based on the forecasting performance of the models estimated using a subset of the complete data. The final BVAR models for both the symmetric prior and the nonsymmetric prior are specified over the period January 1970 through December 1984 and forecast out of sample from January 1985 through October 1987.

Out-of-Sample Forecast Evaluation

The first criterion proposed by Granger and Newbold for evaluating forecasts examines whether a set of forecasts is significantly better than its competitors. The root mean-squared error (RMSE) reflects the decision maker's concern with the variability of forecast errors. The traditional use of the RMSE criterion is complemented with an additional method for evaluating the comparative forecast performance of the models. A second test proposed by Stekler evaluates forecasting ability based on the rankings of forecast errors from the models.

RMSEs for the three forecast horizons (one month, three months, and six months ahead) are presented in table 5. Monthly forecasts are generated for the period January 1985 through September 1987 for the untransformed series. Based on the RMSEs, no model clearly outperforms the other models across all horizons for all series. The VARMA model has the poorest performance, achieving the lowest ranking for each of the four series in both the

Table 4. Tightness Parameter Setting for BVAR—Nonsymmetric Prior

	Dependent Variable			
Lagged Variables	Beef Production	Cattle-on-Feed	Steer Prices	Feeder Prices
Beef Production	1.00	0.80	0.80	0.10
Cattle-on-Feed	0.80	1.00	0.80	0.10
Steer Prices	0.80	0.80	1.00	0.80
Feeder Prices	0.10	0.10	0.80	1.00

three-month and six-month forecasts. For the one-month forecasts, the VARMA model provides the worst forecasts for the beef production and cattle-on-feed series and the most accurate forecasts for the steer price series. Based on the RMSE criterion, the BVAR models perform well across all forecast horizons.

The key issue is whether the differences between the RMSEs are statistically significant across the models. Although the RMSE is often used to distinguish between forecasting performance of time-series models, formal statistical tests for significant differences in this statistic are often neglected.

Granger and Newbold presented a method for testing the equality of the mean-squared errors across forecasting models. The test is based on the correlation between $(\epsilon^1 + \epsilon^2)$ and $(\epsilon^1 - \epsilon^2)$ where ϵ^1 is the forecast error from model 1 and ϵ^2 is the forecast error from model 2. If the forecasting models are unbiased (i.e., $E(\epsilon^1) = E(\epsilon^2) = 0$) and the errors are not autocorrelated, then $E[(\epsilon^1 + \epsilon^2)(\epsilon^1 - \epsilon^2)] = \mathrm{VAR}(\epsilon^1) - \mathrm{VAR}(\epsilon^2)$. Thus, the mean-squared errors are equal if and only if the correlation between $(\epsilon^1 + \epsilon^2)$ and $(\epsilon^1 - \epsilon^2)$ is zero. The statistic for testing whether this correlation is zero is

$$(7) \qquad Z = [\ln(1 + r) - \ln(1 - r)]\,\frac{(T - 3)^{1/2}}{2},$$

where r is the sample correlation and T is the number of out-of-sample predictions. Under the null hypothesis of no correlation, Z is approximately distributed $N(0, 1)$.

For each variable, pairwise comparisons are made between the model with the minimum RMSE and the alternative models. Comparisons in which the minimum RMSEs are significantly lower than an alternative model are denoted by an underline in table 5. The results confirm that the VARMA model achieves significantly higher RMSEs across all forecast horizons for all variables, except for the one-step forecasts of steer prices. The results also suggest that choosing a model based solely on the magnitude of the RMSE is not sufficient.

An alternative measure of forecasting accuracy defined by Stekler is based on a ranking

Table 5. Out-of-Sample Root Mean-Squared Error Forecasts, 1985–87

Model[a,b]	BVAR-S	BVAR-NS	VARMA	UVAR	RVAR
One Period Ahead					
Beef Production	112.27	111.88	125.84	115.51	110.23*
Cattle-on-Feed	250.92*	286.49	347.26	251.44	310.21
Steer Prices	2.53	2.58	2.34*	2.68	2.39
Feeder Prices	2.03	1.86*	2.09	2.16	1.99
Three Periods Ahead					
Beef Production	102.78	105.05	127.65	105.53	102.31*
Cattle-on-Feed	400.47	513.86	737.77	369.12*	534.86
Steer Prices	2.47*	2.51	4.92	2.51	2.48
Feeder Prices	2.03	1.95*	4.15	2.10	2.02
Six Periods Ahead					
Beef Production	106.64	109.25	141.25	102.90*	118.46
Cattle-on-Feed	563.86	682.64	891.83	532.63*	581.39
Steer Prices	2.49	2.46*	6.07	2.48	2.71
Feeder Prices	1.95	1.87*	5.86	2.00	1.96

[a] Asterisk indicates the minimum root mean-squared error (RMSE) for each series.
[b] The underscore indicates that the minimum RMSE, denoted by an asterisk, was significantly smaller than the RMSE from the alternative model.
Note: BVAR-S—Bayesian VAR model with a symmetric prior; BVAR-NS—Bayesian VAR model with a nonsymmetric prior; VARMA—vector ARMA model (Tiao and Box); UVAR—unrestricted VAR; and RVAR—restricted VAR using Schwarz criterion.

Table 6. Test for Equal Forecast Ability—Out-of-Sample Forecasts

Variables[a]	One Period Ahead	Three Periods Ahead	Six Periods Ahead
Beef Production	4.682	4.042	6.595
Cattle-on-Feed	4.200	27.720*	15.547*
Steer Prices	6.406	12.537*	30.070*
Feeder Prices	4.386	11.983*	21.782*

[a] Asterisk indicates significance at .05 level. Critical value for $\chi^2_{4,.05} = 9.488$.

of forecast errors from the models. This measure reflects the decision maker's concern with the comparative size of forecast errors generated by alternative time-series models. Each of the forecasts from the models is ranked one through five according to their accuracy in predicting the four series. A score equal to the ranking is assigned to each variable. Aggregate scores are obtained for each of the series by summing the rankings across the given forecast horizon.

If the models have equal forecasting ability, the scores would have the same expected value for each model. A χ^2 goodness-of-fit statistic is used to test for differences in forecasting ability by examining whether the aggregate score differs significantly from the expected score assuming the models had equal forecast ability. This criterion explicitly compares the complete set of forecasts over each period from each model.

Table 6 presents the χ^2 values for the scores from the forecasting models. The χ^2 values are calculated separately across each forecasting horizon for the four series. For the one-month forecasts, the calculated values for each of the four series do not exceed the 5% critical value. The null hypothesis that the models have equal scores and thus equal forecasting ability cannot be rejected.

The calculated statistics for the three-month and six-month forecasts indicate that for longer range forecasts the models differ in forecasting ability. Significant χ^2 values are obtained for cattle-on-feed, steer prices, and feeder prices at the three-month and six-month forecasting horizons. The null hypothesis that the models have equal forecasting ability for these three series is rejected. For the beef production series, no model significantly outperforms the other models. The χ^2 statistics do not exceed the critical values for either the three-month or the six-month forecast horizons.

The rankings of the forecast errors reinforce the results obtained based on the RMSE criterion. The VARMA model yields the largest forecast errors and ranks last by the Stekler criterion. For the three-month and six-month horizons, the VARMA models rank last for each of the four variables. The symmetric and nonsymmetric BVAR models compare favorably with the other models for the one-month forecasts. The UVAR model is clearly the best model for longer term forecasts; this model has the lowest ranking for the variables beef production, cattle-on-feed, and steer prices.

The second criterion proposed by Granger and Newbold for evaluating forecasts attempts to assess the value of the forecasts to a decision maker in some absolute sense. The Henriksson-Merton test, which evaluates the ability of the time-series model to predict directional changes in the forecast variable, meets this criterion. Merton provided the theoretical justification for such a test. He suggested that if a forecast has any value, it must cause a rational observer to modify prior beliefs about the distribution of subsequent movements in the variable being forecast.

Cumby and Modest reformulated the test for forecasting ability and proposed a method to implement the test in a simpler and more intuitive regression framework. To implement the test, let $\gamma_{it} = 1$ if the forecast change for a particular series is nonnegative, and $\gamma_{it} = 0$ otherwise. Under the null hypothesis that the forecast has no value, Henriksson and Merton showed that the following condition must hold:

$$(8) \quad \text{Prob}[\gamma_{it} = 0 \mid Z_{it} < 0] + \text{Prob}[\gamma_{it} = 1 \mid Z_{it} \geq 0] = 1,$$

where Z_{it} represents the actual change in the variable.

A method to carry out the Henriksson-Merton test is based on the regression

$$(9) \qquad Z_{it} = \alpha + \beta X_{it} + \epsilon_t,$$

where $X_{it} = 1$ if the forecast change for a series is positive and $X_{it} = 0$ if the forecast change is negative, α and β represent coefficients to be estimated, and ϵ is the random error term. When the forecasting model contains no information about movements in the variable, then $\beta = 0$. If the model is able to forecast

Table 7. Regression Tests for Out-of-Sample Forecast Evaluation

Model:[a,b]	BVAR-S	BVAR-NS	VARMA	UVAR	RVAR
One Period Ahead					
Beef Production	91.07	99.94	−37.12	81.42	87.60
	(1.81)	(1.90)	(−0.66)	(1.52)	(1.58)
Cattle-on-Feed	187.14	113.91	163.82	114.39	132.78
	(1.72)	(1.04)	(1.54)	(1.05)	(1.17)
Steer Prices	2.70	2.20	0.40	2.31	1.49
	(3.99)	(3.05)	(0.52)	(3.19)	(1.84)
Feeder Prices	1.17	1.01	0. 86	1.33	0.86
	(1.77)	(1.51)	(1.27)	(2.03)	(1.30)
χ^2 Test Slopes = 0	20.555*	14.599*	4.671	15.377*	7.744
Three Periods Ahead					
Beef Production	126.03	162.55	54.96	143.20	145.78
	(2.32)	(3.03)	(0.90)	(2.66)	(2.78)
Cattle-on-Feed	181.32	224.49	25.71	329.01	76.56
	(1.71)	(2.10)	(0.22)	(3.61)	(0.68)
Steer Prices	1.81	1.47	−0.16	1.66	0.77
	(2.58)	(2.11)	(−0.20)	(2.45)	(0.98)
Feeder Prices	1.45	0.60	−1.49	0.95	0.72
	(2.10)	(0.81)	(−2.12)	(1.30)	(1.00)
χ^2 Test Slopes = 0	18.760*	19.159*	5.373	24.487*	9.056
Six Periods Ahead					
Beef Production	205.46	217.33	6.38	153.96	153.96
	(3.65)	(4.13)	(0.10)	(4.32)	(2.61)
Cattle-on-Feed	60.42	−82.69	−57.47	107.97	257.80
	(0.43)	(−0.61)	(−0.50)	(0.79)	(2.22)
Steer Prices	1.18	1.31	1.50	1.29	0.46
	(1.27)	(1.47)	(1.84)	(1.41)	(0.53)
Feeder Prices	−0.75	0.18	1.31	0.40	0.50
	(0.98)	(0.23)	(1.71)	(0.53)	(0.62)
χ^2 Test Slopes = 0	10.796*	13.560*	6.316	13.808*	12.069*

[a] Asymptotic t-values in parentheses.
[b] Asterisk indicates significance at .05 level. Critical value for $\chi^2_{4,.05} = 9.488$.
Note: See the note to table 5 for model definitions.

directional changes in the variable, then $\beta > 0$.

The regression tests of forecasting ability are based on the joint modeling of all four variables in the time-series models, allowing for cross-variable interactions in the livestock sector. Tests are performed for each of the five time-series models using out-of-sample forecasts for the one-month, three-month, and six-month horizons.

The definitions for forecast and actual changes for each series are based on Kaylen (1986). The forecast direction compares the forecast for k periods in advance made at time t with the actual value at time t. The actual direction compares the difference between the actual value for k periods in advance and the actual at time t. Table 7 presents the estimated slope coefficients from the regression tests of forecasting ability along with asymptotic t-statistics. The tests for forecasting ability are based on the χ^2 statistics which jointly test the hypothesis that the slope coefficients are zero for each of the time-series models.

The regression tests reveal strong evidence of forecasting ability for the symmetric and nonsymmetric BVAR models and the UVAR model. For each forecast horizon, the null hypothesis of no value in the forecasts is rejected for these three time-series models. By contrast, the VARMA model shows no evidence of forecasting ability across any of the forecast horizons. The RVAR model, while not able to forecast directional changes over either the one-

Table 8. Henriksson-Merton Nonparametric Out-of-Sample Forecast Evaluation[a]

Model:[b]	c	n_1	n	N_1	N_2
One Period Ahead					
BVAR-S	.99964	43	64	66	62
BVAR-NS	.99791	40	61	66	62
VARMA	.89438	36	62	66	62
UVAR	.99875	42	64	66	62
RVAR	.98978	39	62	66	62
Three Periods Ahead					
BVAR-S	.99831	41	61	64	56
BVAR-NS	.99476	40	61	64	56
VARMA	.70838	34	60	64	56
UVAR	.99913	41	60	64	56
RVAR	.98086	36	56	64	56
Six Periods Ahead					
BVAR-S	.95308	37	57	61	47
BVAR-NS	.97834	39	59	61	47
VARMA	.97716	35	52	61	47
UVAR	.97096	37	56	61	47
RVAR	.99815	39	55	61	47

[a] c is the confidence level defined in the text; n_1 is the number of successful predictions given a positive revision; n is the number of successful predictions given a positive revision plus the number of unsuccessful predictions given a nonpositive revision; N_1 is the number of observations with positive revisions; and N_2 is the number of observations with nonpositive revisions.
[b] See the note to table 5 for model definitions.

month or the three-month horizon, is able to forecast directional changes over the six-month horizon.

The Henriksson-Merton test provides additional information for evaluating the forecasting ability of the successful BVAR models and the UVAR model. Calculation of this test provides information about the number of times each model correctly and incorrectly predicts both upward and downward directional changes in the variables of interest. Such information would clearly be of use to a decision maker in evaluating a time-series model.

For a series of N observed out-of-sample forecasts for the variables, define N_1 = number of observations with positive revisions; N_2 = number of observations with nonpositive revisions; $N = N_1 + N_2$; n_1 = number of successful predictions given a positive revision; n_2 = the number of unsuccessful predictions given a nonpositive revision; and $n = n_1 + n_2$.

The Henriksson-Merton test for forecasting ability examines whether the observed number of successful predictions is unlikely under the null hypothesis of no forecasting ability. Let x represent the number of correct predictions. The null hypothesis of no forecasting ability is rejected when the probability of observing n_1 or more correct signals is unacceptably small. For a given confidence level (c), the null hypothesis of no value in the forecasts is rejected when $n_1 \geq x^*$, where x^* is the solution to

$$(10)\quad \sum_{x=x^*}^{\bar{n}_1} \binom{N_1}{x}\binom{N_2}{n-x} \Big/ \binom{N}{n}$$

$$= 1 - \text{confidence level with } \bar{n}_1 = \min(N_1, n).$$

These results, summarized in table 8, indicate significant forecasting ability for the symmetric and nonsymmetric BVAR, UVAR, and RVAR models across all the forecast horizons at the 5% significance level. The VARMA models reveal no evidence of ability to predict directional changes for either one-month or three-month forecast horizons at the 5% significance level.

Given that the variable actually experienced an increase, the symmetric and nonsymmetric BVAR models and UVAR model predict such directional changes over 60% of the time across all forecast horizons. Downward directional changes are correctly predicted with a success rate over 60% at both the one-month and three-month forecast horizons, dropping off to about 40% at the six-month horizon. The VARMA model achieves a lower success rate in predicting both upward and downward directional changes across all the forecast horizons.

Summary

The forecasting performance of five multivariate time-series models for the livestock industry is evaluated. Formal comparisons of time-series models are based on two criteria. The first criterion examines whether a set of forecasts is significantly better than its competitors. The RMSEs for each model are evaluated and complemented with a statistical test which examines whether the observed differences in RMSE are statistically significant. A second measure of forecast performance based on this criterion examines the comparative rankings of forecast errors. Both tests provide insight into the comparative forecasting ability of alternative time-series models.

The symmetric and nonsymmetric BVAR models perform well across all forecast horizons based on the RMSE criterion. The VAR-

MA model performs poorest, achieving the lowest ranking for each of the four series in both the three-month and six-month forecasts. The VARMA model provides the worst forecasts for the beef production and cattle-on-feed series for the one-month forecasts.

The rankings of the forecast errors confirm the results of the RMSE criterion. The BVAR models compare favorably with the other models for the one-month forecasts. The UVAR model is clearly the best model for the six-month forecasts, with the lowest ranking for beef production, cattle-on-feed, and steer prices. The poor performance of the VARMA model is again apparent as the model results in the largest forecast errors. For the three-month and six-month horizons, the VARMA models rank last for each of the four variables.

The second criterion attempts to assess the worth of the forecasts to a decision maker in some absolute sense. A formal test for forecasting ability evaluates the ability of the time-series models to forecast directional changes in the variables of interest. The test reveals that the BVAR models and the UVAR model show unambiguous evidence of forecasting ability, again outperforming the VARMA model.

The results reported here may not be generalized outside the livestock industry. Based on the test criteria examined in this paper, both the BVAR and the UVAR models generate forecasts which compare favorably with forecasts from other models. Forecasts from both the symmetric and nonsymmetric BVAR models along with the UVAR model dominate both alternative VAR specifications and VARMA time-series models. The model specification procedures used to identify appropriate BVAR models have been extensively developed and implemented. These procedures yield forecasts which can clearly outperform a range of alternative time-series models.

[*Received August 1988; final revision received October 1989.*]

References

Bessler, D. A., and J. L. Kling. "Forecasting Vector Autoregressions with Bayesian Priors." *Amer. J. Agr. Econ.* 68(1986):144–51.

Cumby, R. E., and D. M. Modest. "Testing for Market Timing Ability: A Framework for Forecast Evaluation." *J. Financ. Econ.* 19(1987):169–90.

Dickey, D. A., and S. G. Pantula. "Determining the Order of Differencing in Autoregressive Processes." *J. Bus. and Econ. Statist.* 5(1987):455–63.

Doan, T., and R. B. Litterman. *Users Manual–RATS,* revision 4.30. Minneapolis MN: VAR Econometrics, 1984.

Engle, R. F., and S. E. Brown. "Model Selection for Forecasting." *Appl. Math. and Comput.* 20(1986):313–29.

Fackler, J. S., and S. C. Krieger. "An Application of Vector Time Series Techniques to Macroeconomic Forecasting." *J. Bus. and Econ. Statist.* 4(1986):71–80.

Fuller, W. A. *Introduction to Statistical Time Series.* New York: John Wiley and Sons, Inc., 1976.

Granger, C. W. J., and P. Newbold. *Forecasting Economic Time Series,* 2nd ed. Orlando FL: Academic Press, Inc., 1986.

Henriksson, R. D., and R. C. Merton. "Market Timing and Mutual Fund Performance: An Empirical Investigation." *J. of Bus.* 57(1983):73–96.

Kaylen, M. S. "A Note on the Forecasting of Turning Points." *N. Cent. J. Agr. Econ.* 8(1986):155–58.

———. "Vector Autoregression Forecasting Models: Recent Developments Applied to the U.S. Hog Market." *Amer. J. Agr. Econ.* 70(1988):701–12.

Litterman, R. B. "Forecasting with Bayesian Vector Autoregressions: Five Years of Experience." *J. Bus. and Econ. Statist.* 4(1986a):25–38.

———. "Specifying Vector Autoregressions for Macroeconomic Forecasting." *Bayesian Inference and Decision Techniques,* eds., P. K. Goel and A. Zellner. Amsterdam: Elsevier Science Publishers, 1986b.

Lütkepohl, H. "Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process." *J. Time Ser. Anal.* 6(1985):35–52.

Merton, R. C. "On Market Timing and Investment Performance. I. An Equilibrium Theory of Value for Market Forecasts." *J. of Bus.* 54(1981):363–406.

Nickelsburg, G., and L. Ohanian. "VAR Priors and Economic Theory." Proceedings of the Business and Economic Statistics Section, American Statistical Association, (1987):423–26.

Stekler, H. O. "Who Forecasts Better?" *J. Bus. and Econ. Statist.* 5(1987):155–58.

Stillman, R. P. "A Quarterly Model of the Livestock Industry." Tech. Bull. No. 1711, U.S. Department of Agriculture, Economic Research Service, 1985.

Tiao, G. C., and G.E.P. Box. "Modeling Multiple Time Series With Applications." *J. of the Amer. Statist. Assoc.* 75(1981):802–16.

Tiao, G. C., and R. S. Tsay. "Multiple Time Series Modeling and Extended Sample Cross-Correlations." *J. Bus. and Econ. Statist.* 1(1983):43–56.

Webb, R. "Toward More Accurate Macroeconomic Forecasts from Vector Autoregressions." *Econ. Rev. Fed. Res. Bank of Richmond* 71(1985):3–11.