Forecasting Chilean Industrial Production and Sales with Automated Procedures[1]

Rómulo A. Chumacero[2]

February 2004

[1]I thank Ernesto Pastén, Klaus Schmidt-Hebbel, and Rodrigo Valdés for helpful comments and suggestions. Financial support from FONDECYT is gratefully acknowledged. The usual disclaimer applies.

[2]Department of Economics of the University of Chile and Research Department of the Central Bank of Chile. *Address*: Diagonal Paraguay 257. Santiago - CHILE. *Phone*: (56-2) 678-3436. *E-mail address*: rchumace@econ.uchile.cl

Abstract

This paper presents a rigurous framework for evaluating alternative forecasting methods for Chilean industrial production and sales. While nonlinear features appear to be important for forecasting the very short term, simple univariate linear models perform about as well for almost every forecasting horizon.

Key Words: Forecasting, Time Series, Threshold, Artificial Neural Networks, Reality Check, Bootstrap.

JEL Classification: C22, C45, C53.

1 Introduction

Forecast accuracy is important because forecasts are often used to guide decisions. As a wide range of forecasting methods is available, a rigorous methodological approach for assessing their relative strengths is needed.

This paper considers several time series models and their respective automated selection procedures for forecasting Chilean Industrial Production and Sales. Some of the models try to capture nonlinear features that may be present in the data and not captured with simple linear models. Yet, the lack of parsimony and extensive specification searches may seriously damage the usefulness of complex models; thus requiring a sound approach for comparing forecasting accuracy.

The remainder of the paper is organized as follows. Section 2 briefly describes the data. Section 3 presents the various types of models used. Section 4 uses several methodologies to compare the forecasting accuracy of the models. Finally, Section 5 concludes.

2 The Data

I use monthly observations of the 12 month variations of the industrial production and sales reported by the National Bureau of Statistics for the period 1991:12-2003:11.[1]

	Production	Sales
Mean (%)	3.27	3.46
Standard Deviation (%)	4.85	5.09
First order autocorrelation	0.56	0.50
Jarque-Bera (p-value)	0.25	0.86
ADF (p-value)	0.04	0.02

Table 1: Summary Statistics for Chilean Production and Sales Growth Rates (1991:12-2003:11)

Table 1 and Figure 1 present the evolution of both series and some summary statistics. Production and sales growth present positive, yet volatile,

[1]If Y_t is the raw data, the variable of interest is $y_t = \ln(Y_t/Y_{t-12})$. No additional treatment of the data is conducted.

growth rates for most of the sample. While, as expected, unit roots are strongly rejected, both series present relatively high persistence for variables that are already expressed in terms of growth rates. A sign that forecasting these series may be challenging is that the variation coefficient (standard deviation over mean) profusely exceeds unity. Finally, both series are highly correlated (simple correlation 0.76) and there is marginal evidence of unidirectional Granger causality from Production to Sales (not reported).

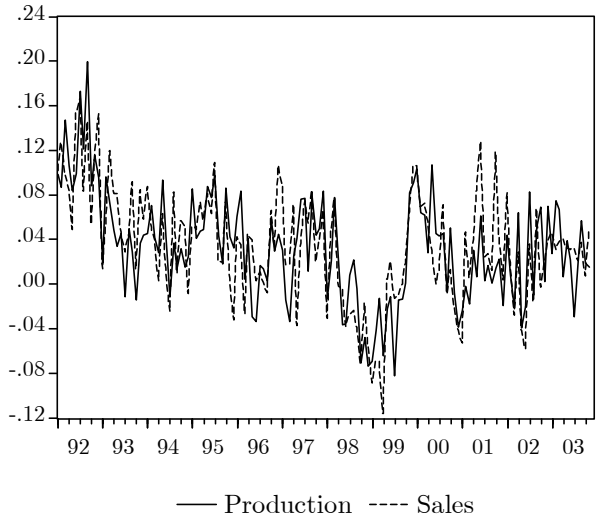

Figure 1: Growth Rates of Chilean Industrial Production and Sales (1991:12-2003:11)

The next section presents the forecasting models that are used. If it not were for the inclusion of the 12 month variation of the number of working days in the month (denoted by d), most of them are otherwise univariate time series representations of the data.

3 Models

This Section briefly describes the four types of models under scrutiny: a linear autoregressive model, an artificial neural network model, a self-exciting

threshold autoregressive model, and a combination of the three.

3.1 Linear Autoregressive Model

The linear autoregressive (AR) model for series y reads

$$y_t = \alpha + \sum_{j=1}^{p} \beta_j y_{t-j} + \delta d_t + u_t, \tag{1}$$

where y denotes the variable of interest (production or sales), d is as defined above, and u is a white noise.

To determine the order of the process, p is selected by minimizing the Hannan-Quinn information criterion (HQ)

$$HQ_i \simeq \ln\left(\widehat{\sigma}_i^2\right) + 2\frac{k_i}{T}\ln\left(\ln\left(T\right)\right), \tag{2}$$

where k is the number of parameters being estimated, T is the sample size, and $\widehat{\sigma}_i^2$ is a consistent estimate of the variance of u for model i. This choice has some desirable properties over other candidates, as it lays between the Akaike criterion (that tends to overfit) and the Schwarz criterion (that chooses parsimonious models).[2] In the empirical application the minimum and maximum orders of p are set equal to 1 and 36. In the case of Industrial production, the model chosen was an AR(12). For industrial sales, the model chosen was an AR(3).

3.2 Artificial Neural Network Model

The artificial neural network (NN) model with K hidden units (layers) is defined as:

$$y_t = \alpha_0 + \sum_{j=1}^{p} \beta_{j,0} y_{t-j} + \delta_0 d_t + \sum_{k=1}^{K} \left[\phi_k \psi\left(\alpha_k + \sum_{j=1}^{p} \beta_{j,k} y_{t-j} + \delta_k d_t\right)\right] + u_t, \tag{3}$$

where $\psi\left(v\right) = \left(1 + e^{-v}\right)^{-1}$ is a logistic activation function and ϕ_k is the "weight" of the hidden unit k.

[2] Furthermore, this information criterion is consistent in the sense that it chooses the "correct" model with probability 1 as $T \to \infty$. As is well known, the Akaike criterion does not fulfill this requirement (Inoue and Kilian, 2003).

As pointed out by Tkcaz (2001), considering NN models has several advantages: First, neural networks are data-driven and flexible tools that are particularly useful when there are no prior beliefs about functional forms. Second, when properly specified, NN are universal functional approximators. Finally, neural networks are nonlinear and nest linear models.

For a given value of K, (3) can be estimated using nonlinear least squares (Kuan and White, 1994). Large values of K may be difficult to estimate as the number of parameters to be estimated increases linearly. The choice of K is also conducted by minimizing (2). To make the search manageable, the value of p is set equal to the linear model, thus forcing the NN model to nest the AR model. In the empirical application the minimum and maximum orders of K are set equal to 1 and 3. In the case of Industrial production, the model chosen had $K = 3$. For sales, the model chosen had $K = 1$.

3.3 Threshold Model

The two-regime self-exciting threshold autoregressive (TAR) model reads

$$y_t = \begin{cases} \alpha_1 + \sum_{j=1}^{p} \beta_{j,1} y_{t-j} + \delta_1 d_t + u_t & \text{if } y_{t-r} \leq \theta \\ \alpha_2 + \sum_{j=1}^{p} \beta_{j,2} y_{t-j} + \delta_2 d_t + u_t & \text{if } y_{t-r} > \theta \end{cases}, \tag{4}$$

where θ is the threshold value and y_{t-r} is the threshold variable.

This model is popular because it provides an easy-to-estimate alternative to the regime-switching model (in fact it is a special case of the latter), is consistent with nonlinear features of the data, and provides asymmetric impulse-response functions.[3]

Given a choice of r, θ in (4) can be estimated by direct search. The choice of r is conducted by minimizing (2). As above, the value of p is set to coincide with the value for the linear model and the minimum and maximum values of r are set equal to 1 and 12. In both, production and sales, the value obtained for r was 1.

3.4 Combined Forecast

As Fang (2003) puts it, forecasting models differ in structure and data used. If their forecasts are not perfectly correlated with each other, they may provide

[3]See Hansen (1997) or Siliverstovs and van Dijk (2003), and references therein for further details.

different insights of the dynamics of a series. Combining competing forecasts often leads to increased forecasting accuracy.

The combined forecast (C) model is:

$$\widehat{y}_t^C = \sum_{l=1}^{L} \widehat{\omega}_l \widehat{y}_t^l, \tag{5}$$

where ω_l is the weight associated with forecast l and $\widehat{y}_t^l$ is the forecast of model l.

Given L forecasts, the weights are obtained as follows:

$$\widehat{\omega} = \arg\min_{\omega} \sum_{t=T_0}^{T} \left[y_t - y_t^C \left(\omega_1, \ldots, \omega_L \right) \right]^2, \quad \omega_l \geq 0, \quad \sum_{l=1}^{L} \omega_l = 1,$$

For the application, L is set equal to 3 and combines the forecasts of the three models described above.

4 Forecast Evaluation

Several evaluation criteria are available to judge the performance of a forecasting model. Here we concentrate on point forecast evaluation for h-steps ahead forecasts of the models discussed.[4]

For evaluation of point forecasts, the Root Mean Squared Forecast Error (RMSE) and the Mean Absolute Forecast Error (MAFE) are considered. Let $\{\widehat{u}_{i,t}\}_{t=T_0}^{T}$ denote the sequence of h-steps forecast errors and define $T^* = T - T_0 + 1$. The RMSE and MAFE of model i are defined as

$$\begin{aligned} \text{RMSE}_i &= \sqrt{\frac{1}{T^*} \sum_{t=T_0}^{T} \widehat{u}_{i,t}^2} \\ \text{MAFE}_i &= \frac{1}{T^*} \sum_{t=T_0}^{T} \left| \widehat{u}_{i,t} \right|. \end{aligned}$$

As several competing forecast models are considered, one set of them will appear more successful than another in a given dimension (say, one model

[4] The predictive performance of models can also be judged in dimensions such as interval and density forecasts (Siliverstovs and van Dijk, 2002).

has the smallest MAFE for 2-steps ahead forecasts). It is inevitable then to ask how likely it is that this result is due to chance. Diebold and Mariano (1995), approach forecast comparison in this framework.

Consider the pair of h-steps ahead forecasts of models i and j $(\widehat{u}_{i,t}, \widehat{u}_{j,t})$ for $t = T_0, \ldots, T$; whose quality is to be judged by the loss function $g(\widehat{u}_{i,t})$.[5] Defining $d_t = g(\widehat{u}_{i,t}) - g(\widehat{u}_{j,t})$, under the null hypothesis of equal forecast accuracy between models i and j, we have $\mathcal{E}d_t = 0$. Given the covariance-stationary realization $\{d_t\}_{t=T_0}^{T}$, it is natural to base a test on the observed sample mean:

$$\overline{d} = \frac{1}{T^*} \sum_{t=T_0}^{T} d_t.$$

Even with optimal h-steps ahead forecasts, the sequence of forecast errors follows a MA$(h-1)$ process. If the autocorrelations of order h and higher are zero, the variance of $\overline{d}$ can be consistently estimated as follows:

$$\overline{V} = \frac{1}{T^*} \left(\widehat{\gamma}_0 + 2 \sum_{j=1}^{h-1} \widehat{\gamma}_j \right),$$

where $\widehat{\gamma}_j$ is an estimate if the j-th autocovariance of d_t.

The Diebold-Mariano (DM) statistic is given by

$$DM = \frac{\overline{d}}{\sqrt{\overline{V}}} \xrightarrow{d} \mathcal{N}(0,1)$$

under the null of equal forecast accuracy. Harvey et al (1997) suggest to modify the DM test and use instead:

$$HLN = DM \cdot \left[\frac{T^* + 1 - 2h + h(h-1)/T^*}{T^*} \right]^{1/2}$$

to correct size problems of DM. They also suggest to use a Student's t with $T^* - 1$ degrees of freedom instead of a standard normal to account for possible fat-tailed errors.

To test if model i is not dominated by model j in terms of forecasting accuracy for the loss function $g(\cdot)$, a one-sided test of DM or HLN can be conducted, where under the null $\mathcal{E}d_t \leq 0$. Thus, if the null is rejected, we conclude that model j dominates model i.

[5] For example, in case of Mean Squared Error comparison, $g(\cdot)$ is a quadratic loss function $g(\widehat{u}_{i,t}) = \widehat{u}_{i,t}^2$ and in the case of MAFE, it is the absolute value loss function $g(\widehat{u}_{i,t}) = |\widehat{u}_{i,t}|$.

4.1 In-Sample Evaluation

One way of evaluating competing models is by judging their in-sample forecasting accuracy. For in-sample forecasting (IS), each model is estimated with the full sample, the "best" of each category is chosen by minimizing (2), and the coefficients thus obtained are used to conduct h-steps ahead forecasts. For example, a 3-step ahead forecast for period y_t uses the coefficients estimated with the full sample $\widehat{\beta}_T$, but only uses the observations of y up to period $t-3$.

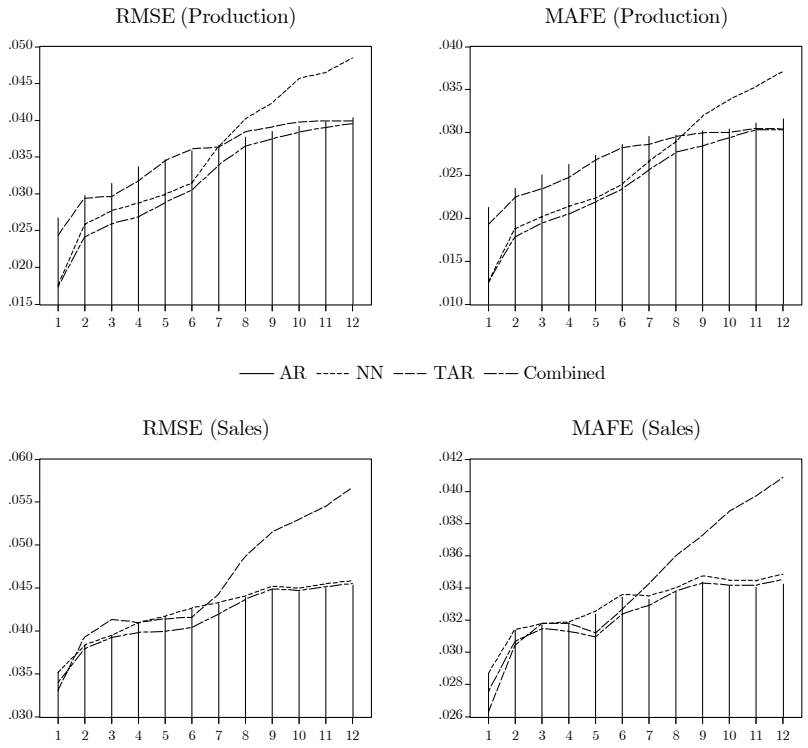

Figure 2: In-sample h-step Ahead Forecast Evaluation

Figure 2 and Table 2 present the RMSE, MAFE, and HLN tests of forecasting accuracy for h-steps ahead forecasts. All display the typical ascending pattern (increasing RMSE and MAFE as the forecasting horizon increases).

		Production				Sales			
Model i−**Model** j		RMSE		MAFE		RMSE		MAFE	
AR-	**NN**	2	**6**	2	**5**	1	**0**	1	**0**
	TAR	2	**0**	1	**0**	1	**0**	1	**0**
	Combined	4	**0**	5	**0**	2	**0**	2	**0**
NN-	**TAR**	7	**1**	4	**2**	0	**0**	0	**0**
	Combined	10	**0**	8	**1**	0	**0**	0	**0**
TAR-**Combined**		4	**0**	5	**0**	3	**0**	1	**0**

Table 2: Evaluation of In-sample Point Forecasts using the HLN Test. Entries in plane font indicate the number of forecast horizons (of the 12 possible) in which the null hypothesis that model i is not dominated by model j is rejected. The reverse hypothesis is reported in bold font.

In both cases, nonlinear models appear to provide better forecasts than the linear model for short term forecasting (1 and 2-steps ahead forecasts) but deteriorate faster as the forecasting horizon increases and end up, either being dominated by or providing comparable forecasts to the linear model. Thus, univariate models signal that nonlinear features are important solely for short term forecasting.

In the case of Industrial Production, the NN model dominates all the other forecasting models for up to 2-steps ahead forecast and its performance deteriorates markedly up to the point of being outperformed by the linear model for horizons exceeding seven months. On the other hand, in the case of Industrial Sales, the NN model (which is not as nonlinear as in the model for production) basically coincides with the linear model. The nonlinear short-term features of the series are better approximated by the TAR model, while again, this model succumbs for further forecasting horizons. However, in both cases, combining forecasts is prudent as this mixture exploits the nonlinear features that are absent in the linear model for very short-term forecasts.

Although useful as evaluation tools (Inoue and Kilian, 2002), in-sample forecasts are not why models are used in practice. Next, we evaluate the out-sample forecasting performance of the models presented.

4.2 Out-of-Sample Evaluation

The total sample for both series comprises 144 observations. For out-of-sample forecasting (OS), we obtain estimates for each model beginning with the first 100 observations, produce a forecast for the relevant horizon with them, add one more observation, produce the next forecast, and so on until we use the full sample. The model used in each category, is the "best" obtained with the full sample and not chosen again with each new observation.

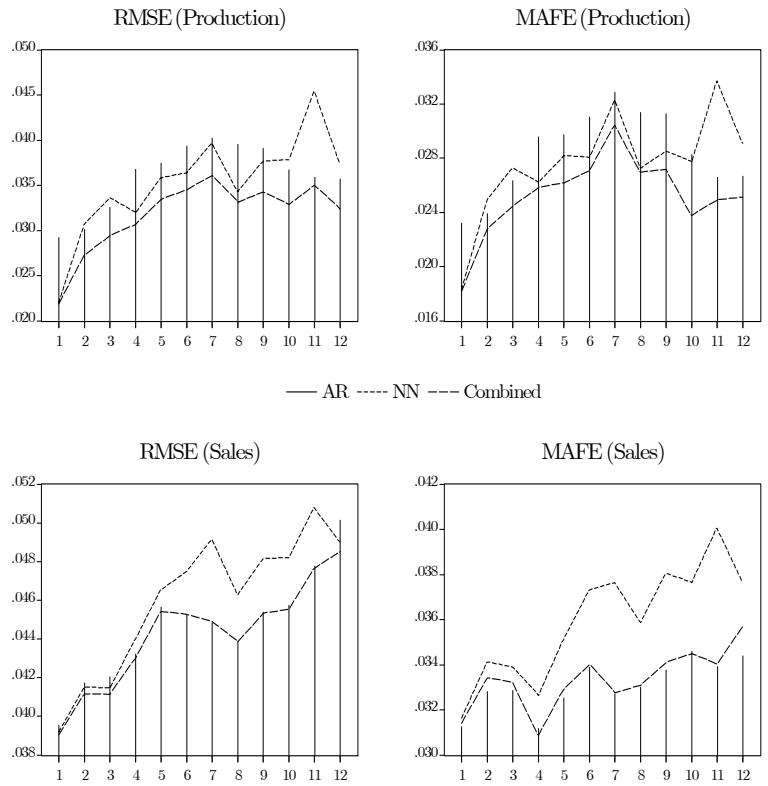

Figure 3: Out-of-sample h-step Ahead Forecast Evaluation

Figure 3 and Table 3 compare the out-of-sample performance of the models.[6] The results here are very robust: 1-step ahead forecasts for Industrial

[6] The performance of the TAR models deteriorates very rapidly for forecast horizons of

		Production				Sales			
Model i−**Model** j		RMSE		MAFE		RMSE		MAFE	
AR-	**NN**	1	**0**	1	**0**	0	**0**	0	**0**
	TAR	0	**5**	0	**3**	0	**5**	0	**5**
	Combined	3	**0**	2	**0**	0	**0**	0	**0**
NN-	**TAR**	0	**3**	0	**3**	0	**2**	0	**1**
	Combined	1	**0**	0	**0**	0	**0**	0	**0**
TAR-**Combined**		9	**0**	9	**0**	6	**0**	6	**0**

Table 3: Evaluation of Out-of-sample Point Forecasts using the HLN Test. Entries in plane font indicate the number of forecast horizons (of the 12 possible) in which the null hypothesis that model i is not dominated by model j is rejected. The reverse hypothesis is reported in bold font.

Production can still be better captured with nonlinear (NN) models that once again deteriorate for extended forecast horizons. Nonlinear models do not provide better information than linear models for forecasting Sales. However, the dominance of some in-sample forecasts is not present out-of-sample. In fact, all models are preferred to the TAR model.

4.3 Forecast Encompassing

The DM and HLN tests are useful to asses if a model dominates another in the dimension chosen. Forecast encompassing tests seek to evaluate whether competing forecasts may be fruitfully combined to produce a forecast superior to the individual forecasts. One of such tests prescribes to regress the actual level of y_t on the predicted values of y by the competing models (Clements and Hendry, 1998). For example, an encompassing test between models i and j can be conducted with the regression model:

$$y_t = \rho_1 \widehat{y}_t^i + \rho_2 \widehat{y}_t^j + v_t$$

4 periods or more, thus they are excluded from Figure 3 to make a better visual comparison of the other models.

and test for $\rho_1 = 1$, (or $\rho_2 = 0$) conditional on $\rho_1 + \rho_2 = 1$.[7] The former specification is indeed equivalent to the regression:

$$\widehat{u}_{j,t} = \rho\left(\widehat{u}_{i,t} - \widehat{u}_{j,t}\right) + v_t \tag{6}$$

and test the null hypothesis ($\rho = 0$). If the null is rejected, model j could be improved by incorporating some of the features of model i.[8]

		Production				Sales			
Model i−**Model** j		IS		OS		IS		OS	
AR-	**NN**	9	**11**	8	**12**	1	**1**	0	**6**
	TAR	3	**0**	0	**12**	4	**8**	1	**11**
	Combined	11	**0**	8	**0**	3	**1**	0	**0**
NN-	**TAR**	11	**9**	10	**12**	7	**11**	1	**11**
	Combined	11	**0**	12	**0**	7	**0**	6	**0**
TAR-**Combined**		9	**0**	12	**0**	10	**0**	11	**1**

Table 4: Forecast Encompassing Tests. Entries in plane font indicate the number of forecast horizons (of the 12 possible) in which the null hypothesis that model i can not be improved by incorporating features of model j is rejected. The reverse hypothesis is reported in bold font. IS=In-sample Forecast. OS=Out-of-sample forecast.

Table 4 presents the results of estimations of (6) for different forecast horizons.[9] As would be expected, the case for combining linear and nonlinear models appears to be strong.

4.4 White's Reality Check

As noted by White (2000), whenever a "good" forecasting model is obtained by an extensive specification search, there is always the danger that the observed good performance results from luck and not from actual forecasting

[7] Including a constant and not imposing the constraint are also other possibilities (Fang, 2003).

[8] As mentioned, as the forecast errors tend to be correlated for 2-steps ahead forecasts or more, a HAC covariance matrix should be used to test the null.

[9] For example, the TAR model would help improve the forecasts of the AR model only for 1 of the 12 horizon periods in the case of industrial sales. However, the AR model would help improve the forecasts of the TAR model in 11 of the 12 horizon periods.

ability. Even when no exploitable relation exists, looking long enough at a given data set will often reveal inexistent patterns that are in fact useless. The practice of conducting extensive searches and their consequences for inference is called "data snooping" or "data mining" and may induce naive practitioners to mistake the spurious for the substantive.

White (2000) provides a formal framework with which to test the null hypothesis that the best model encountered during a specification search has no predictive superiority over a benchmark model.[10] The test is appropriately named a "Reality Check" and can be conducted in several ways. Here, we briefly describe the so-called "Bootstrap Reality Check", which is the one we perform. The steps involved in it are:[11]

- Obtain the DM or HLN test for each model j against a benchmark model and denote it by A_j.
- Generate $i = 1, \ldots, M$ artificial samples of y and the other variables involved in estimating the models. In this case, we require a resampling procedure applicable to dependent processes. Here we use Politis and Romano's (1994) stationary bootstrap.[12]
- Conduct the specification search, minimizing (2), to select the "best" model of each category and each bootstrapped sample.
- Compare the forecast obtained by each model j with a (fixed) benchmark model using either the DM or HLN tests and denote it by D_i^j.[13]
- After obtaining the M values of D_i^j for each j, denote the sorted values of D_i^j as S_i^j. Find F such that $S_F^j \leq A_j < S_{F+1}^j$. Then, the Bootstrap Reality Check p-value is $1 - F/M$.

[10] There is a subtle but important difference with the DM and HLN tests presented above. Those tests compare the forecasts of two (fixed) models, while White compares the forecasts of a model obtained through extensive specification searches and a benchmark model.

[11] This procedure can be computationally very demanding. For example, obtaining the Bootstrap Reality Check p-value (defined below) for the NN model takes up to 30 hours with a Pentium 4 computer. All the results of this paper were performed using GAUSS. The code is available upon request.

[12] Fixed block bootstrapping is also commonly used, however it does not guarantee stationarity of the sample so generated. The stationary bootstrap resamples blocks of random length (length drawn from a geometric distribution) and the resulting series is

$h-$steps ahead	Production		Sales	
	AR	NN	AR	NN
1	.01	.00	.42	.48
2	.36	.01	.51	.48
3	.80	.36	.74	.75
4	.81	.50	.77	.77
5	.82	.60	.84	.87
6	.79	.68	.85	.86
7	.76	.88	.85	.84
8	.74	.77	.79	.79
9	.71	.80	.81	.78
10	.55	.85	.81	.77
11	.59	.83	.87	.83
12	.71	.77	.90	.88

Table 5: Bootstrap Reality Check p-values. An AR(1) Model is used as the benchmark model.

Table 5 presents the Bootstrap Reality Check p-values for the AR(p) and NN Models against an AR(1) benchmark model. The results show that except for one (two) case [1-step (1 and 2 steps) ahead forecast(s) for Industrial Production], the best AR and NN models do not beat (in mean squared error) the AR(1) benchmark model. The results thus confirm that nonlinear features, are important only for very short-term forecasting.

5 Concluding Remarks

This paper develops a methodology for comparing linear and nonlinear univariate forecasting models for Chilean Industrial Production and Sales.

The results suggest that nonlinear features may be relevant for forecasting only in the very short run, and that combining forecasts improves the forecasting performance of each model.

stationary.

[13] White (2000) also uses a variant of the DM test due to West (1996).

References

Clements, M. and D. Hendry (1998). *Forecasting Economic Time Series*, Cambridge University Press.

Diebold, F. and R. Mariano (1995). "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, 253-65.

Fang, Y. (2003). "Forecasting Combination and Encompassing Tests," *International Journal of Forecasting* 19, 87-94.

Hansen, B. (1997). "Inference in TAR Models," *Studies in Nonlinear Dynamics and Econometrics* 2, 1-14.

Harvey, D., S. Leybourne, and P. Newbold (1997). "Testing the Equality of Prediction Mean Square Errors," *International Journal of Forecasting* 13, 281-91.

Inoue, A. and L. Kilian (2002). "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?," *Manuscript*, North Carolina State University.

Inoue, A. and L. Kilian (2003). "On the Selection of Forecasting Models," *Manuscript*, North Carolina State University.

Kuan, C. and H. White (1994). "Artificial Neural Networks: An Econometric Perspective," *Econometric Reviews* 13, 1-91.

Patton, A. and A. Timmermann (2002). "Properties of Optimal Forecasts," *Manuscript*, University of California, San Diego.

Politis, D. and J. Romano (1994). "The Stationary Bootstrap," *Journal of the American Statistical Association* 89, 1303-13.

Siliverstovs, B. and D. van Dijk (2002). "Forecasting Industrial Production with Linear, Non-Linear, and Structural Breaks Models," *Manuscript*, Erasmus University.

Tkacz, G. (2001). "Neural Network Forecasting of Canadian GDP Growth," *International Journal of Forecasting* 17, 57-69.

West, K. (1996). "Asymptotic Inference About Predictive Ability," *Econometrica* 64, 1067-84.

White, H. (2000). "A Reality Check for Data Snooping," *Econometrica* 68, 1097-126.