Research Article

Received: 19 June 2021 / Accepted: 25 August 2021 / Published: 5 September 2021

Forecasting Albanian Time Series with Linear and Nonlinear Univariate Models

Blerina Vika

Statistics and Applied Informatics Department,
Faculty of Economy, University of Tirana,
Arben Broci St., Tirana, Albania

Ilir Vika

Research Department,
Bank of Albania,
Tirana, Albania

DOI: https://doi.org/10.36941/ajis-2021-0140

Abstract

Albanian economic time series show irregular patterns since the 1990s that may affect economic analyses with linear methods. The purpose of this study is to assess the ability of nonlinear methods in producing forecasts that could improve upon univariate linear models. The latter are represented by the classic autoregressive (AR) technique, which is regularly used as a benchmark in forecasting. The nonlinear family is represented by two methods, i) the logistic smooth transition autoregressive (LSTAR) model as a special form of the time-varying parameter method, and ii) the nonparametric artificial neural networks (ANN) that mimic the brain's problem solving process. Our analysis focuses on four basic economic indicators – the CPI prices, GDP, the T-bill interest rate and the lek exchange rate – that are commonly used in various macroeconomic models. Comparing the forecast ability of the models in 1, 4 and 8 quarters ahead, we find that nonlinear methods rank on the top for more than 75 percent of the out-of-sample forecasts, led by the feed-forward artificial neural networks. Although the loss differential between linear and nonlinear model forecasts is often found not statistically significant by the Diebold-Mariano test, our results suggest that it can be worth trying various alternatives beyond the linear estimation framework.

Keywords: *Albania, linearity, time-varying, nonlinear models, artificial neural networks*

1. Introduction

The Albanian economy has experienced various structural changes in the past three decades. During its transition to a market-oriented economy, the country had to undertake many reforms in order to liberalize its economy and develop the financial system (such as privatizations, removal of price administration, free trade agreements, financial integration, capital account liberalization, etc.). Besides, a number of stabilization policies have been taken in particular to cope with the domestic 1997 collapse of pyramid schemes and later on with the long-lasting effects of the 2009 global

financial crisis. These socio-economic changes have had important influence on the macroeconomic performance, leaving visible footprints on financial and real economic time series.

The sudden jumps or irregular patterns that are witnessed in many statistical data can increase instability of model parameters and negatively affect policy analyses with linear methods. For that reason, time-varying estimations as well as the newly developing nonlinear methods with neural networks have been tried in the literature as additional tools to help improve economic and forecast analyses. By construction, nonlinear methods commonly use more parameters than their traditional linear counterpart and are, therefore, expected to provide better in-sample model explanation. For that reason, many seminal studies including Stock and Watson (1996) have relied on the out-of-sample forecasts in order to shed light on the usefulness of nonlinear methods.

This article aims to contribute to the literature on nonlinearity by assessing the forecast ability of nonlinear methods and compare it to the performance of popular linear models. The latter are represented by the traditional linear autoregressive (AR) models, whereas nonlinear methods make use of time-varying econometric techniques with smooth transition autoregressive models, as well as feed-forward neural networks that are a computer artificial intelligence method. Moreover, the findings should provide some insights to macroeconomic modelers whether to enhance their suite of models with nonlinear techniques or not. Therefore, our analysis focuses on four illustrative statistical series – namely consumer prices, GDP, interest rate and the exchange rate – which are commonly used in various macroeconomic models and their developments and prospects are regularly discussed by policymakers. To keep it simple, the analysis concentrates only on univariate models, and leaves the scrutiny with multivariate specifications for further research.

To preview our findings, nonlinear methods rank on top three models for more than 75 percent of the out-of-sample forecasts, particularly led by the artificial neural network models. The result is consistent in all of the intended forecast horizons: the 1, 4 and 8 quarters ahead forecasts. On the other hand, the linear AR models appear successful and outperforming nonlinear methods when forecasting CPI prices at almost all horizons, and perhaps in predicting gross domestic production at four quarters ahead. Interestingly, the loss differential between linear and nonlinear model forecasts is often found not statistically significant by the Diebold-Mariano test. Nevertheless, the analysis points out to the potentiality of nonlinearities in the Albanian economic indicators, suggesting that it can be worth trying various alternatives beyond the linear estimation models.

The rest of the paper is organized as follows. Sections 2 describes the various methods employed to estimate and forecast in the face of probable structural changes. Section 3 presents some stylized facts and preliminary tests of the data, followed by a discussion of model selection and forecasting procedure. Section 4 presents the forecast evaluation results and compares the relative performance for each method. Section 5 carries out a number of tests that check for parameter constancy in linear models. Section 6 offers some concluding remarks.

2. A Glimpse at the Time Series Forecasting Frameworks

The literature introduces us with various methods that are commonly applied in forecasting economic time series. Unlike structural models that rely deeply on formal economic theory, time series models are bound to include only a handful of explanatory variables and have often shown similar or even better near-term forecast performance than the more sophisticated structural models. A similar accomplishment is also evidenced in the case of the simpler and less time-consuming univariate models that only explore the past behavior of its own data. Because of their timeliness as well as the ability to often generate satisfactory economic forecasts, univariate models are widely used as attractive benchmarks in evaluating the relative performance of various multivariate forecast models.

In time series forecasting models, the variable y to be forecast at time t for h-periods ahead is generally a function of a vector of predictor variables Z_t and a vector of unknown parameters θ, plus the forecast error ε_{t+h}, as formulated below:

$$y_{t+h} = f(Z_t\,;\,\theta_{ht}) + \varepsilon_{t+h} \quad (1)$$

Parameters θ_{ht} could be linearly estimated or allowed to evolve over time, thereby distinguishing the functional form of the forecasting methods. Then, each method can use a variety of alternative models depending on the choice of the predictor variables and the stationarity transformation applied to the dependent variable. As our study focuses on univariate models, predictor variables Z_t consist solely of current and past values of y_t [$Z_t = (y_t,..., y_{t-p}, \Delta y_t,..., \Delta y_{t-p}, 1, t)$].

Linear univariate models have been widely used in forecasting economic and financial time series, following the seminal work by Box and Jenkins (1976). A key assumption in this framework is that variable y_t has a normal distribution and it is stationary, meaning that y_t has a stable mean and variance and its correlation with its past values is also stable over time (Ghysels & Marcellino, 2018). Some of the linear univariate models introduced in the literature include autoregressive (AR) models, moving average (MA) models, autoregressive-integrated moving average (ARIMA) models, regressions with forecast errors, fractionally differenced models for long-range dependence, and unobserved components models that are useful for extracting time series cyclical as well as seasonal components.

However, if variable y_t is not Gaussian and its future is not similar to the past behavior, forecasts can be improved by using nonlinear time series models. In that case, the model is said to have time-varying parameters as θ coefficients can change to new values at a known date T. Some of the parametric nonlinear models include the bilinear models, the state-space model, the threshold autoregressive (TAR) models, the Markov-switching model, and the functional-coefficient autoregressive model. An alternative approach uses nonparametric and semiparametric methods to explore the nonlinearity in time series by treating the functional form f as unknown. Such methods include nearest neighbor, kernel regression and artificial neural networks models (Tsay, 2002, p. 127).

3. Data Analysis, Model Estimation, and Forecasting Procedure

This section initially presents the characteristics of our selected economic indicators in terms of their variability, normal distribution, and stationarity. The analysis should serve as a motivation for identifying the 'appropriate' framework and model specification, which are defined in a comprehensible way in the following subsection.

3.1 Data characteristics

Our analysis focuses on four main economic time series, namely the CPI index, real GDP, the lek/euro exchange rate, and the 12-month Treasury bill interest rate.[1] The analysis is based on quarterly frequency data during the period from 1996Q1 to 2018Q4. Figures 1 through 4 graph the variables in levels and in quarter-on-quarter percent changes. Developments in each of the time series display high volatility around the domestic 1997 financial turmoil, during which there was a swift exchange rate depreciation, an upsurge in inflation and interest rates, as well as a pronounced decline in gross domestic production up to the first quarter of 1998. The next few years evidence a return of the variables to rather "normal" trends, which characterize most of the remaining period.

The descriptive statistics in Table 1 suggest that time series developments have not been constant throughout the sample. The average quarterly changes of most variables except GDP points to positive and more rapidly developments in the 1990s, followed by a slower pace or even a change in direction in the consequent subsamples. Also, GDP experienced a higher growth rate from 2000 to 2009, but its pace

[1] *The CPI and GDP data are taken from the Institute of Statistics, whereas the exchange rate and T-bill rate from the Bank of Albania website. Because national accounts are only published as annual data before 2008, real output is quarterly interpolated such as to match the annual figures. Lastly, all variables have been seasonally adjusted by using X-12 additive procedure.*

sort of halved in the 2010s. These features cast doubt on the relevance of key assumptions made for linear univariate models (such as variables are Gaussian and have stable mean and variance), therefore justifying the estimation of alternative methods that explore nonlinearity in time series.

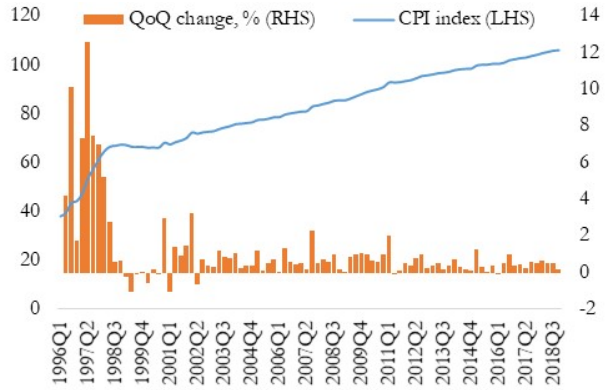

Figure 1. Consumer Prices

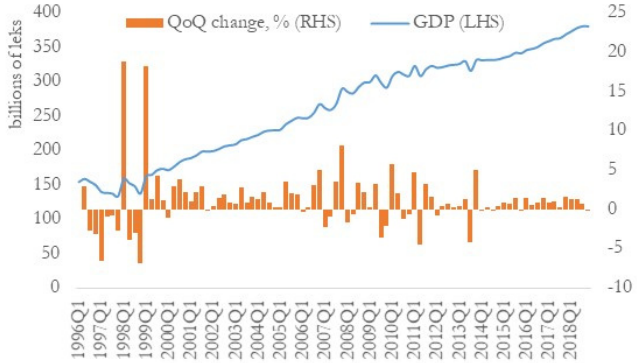

Figure 2. Gross Domestic Product

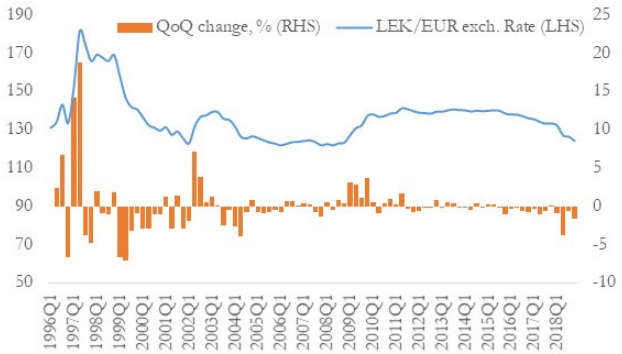

Figure 3. The LEK/EUR Exchange Rate

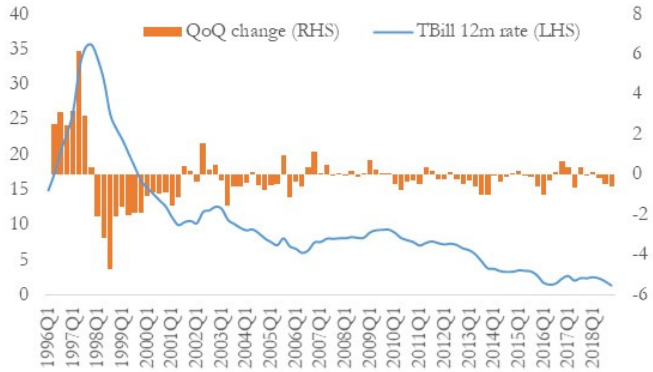

Figure 4. Treasury Bill, 12-month rate

Table 1. Descriptive Statistics

	CPI q/q % chg	GDP q/q % chg	Exch. Rate q/q % chg	T-Bill q/q chg
		Mean		
Whole sample	1.2	1.1	0.0	-0.2
1996q1:1999q4	3.7	1.1	0.7	0.0
2000q1:2009q4	0.7	1.3	0.0	-0.2
2010q1:2018q4	0.5	0.8	-0.3	-0.2
		Standard deviation		
Whole sample	2.1	3.7	3.3	1.3
1996q1:1999q4	4.0	7.8	7.3	2.9
2000q1:2009q4	0.8	2.1	2.2	0.7
2010q1:2018q4	0.4	2.0	0.9	0.4
		Normality test[#)] (Jarque-Bera probability)		
Whole sample	0.00	0.00	0.00	0.00

[#)] *Normality tests are rejected even if tried on variables in levels.*

The graphical inspection in Figure 1-4 suggested that the mean of most variables in levels increases dramatically over time, thus giving the impression that they may be nonstationary and a unit root should be imposed. Nevertheless, stationarity examinations based on the Augmented Dickey-Fuller test and the Elliott-Rothenberg-Stock DF-GLS test often bring to dissimilar inferences (please see Table 2). The two tests suggest opposing orders of integration in the case of consumer prices and the exchange rate – inferring they are stationary, $I(0)$, or difference-stationary, $I(1)$. The Treasury bill rate is shown to be stationary according to the DF-GLS test but is determined as trend-stationary according to the ADF test. On the other side, gross domestic production is generally found difference-stationary from both tests.

Table 2. Stationarity Tests

	CPI index		Real GDP		Lek/Eur exch. rate		T-Bill, 12M rate	
Sample period	ADF	DF-GLS	ADF	DF-GLS	ADF	DF-GLS	ADF	DF-GLS
1996Q1:2012Q4	L/C	D/C	L/T	D/C	D/C	L/C	D/C	L/C
1996Q1:2013Q4	L/C	D/C	D/C	D/C	D/C	L/C	L/T	L/C
1996Q1:2014Q4	L/C	D/C	D/C	D/C	D/C	L/C	L/T	L/C
1996Q1:2015Q4	L/C	D/C	D/C	D/C	D/C	L/C	L/T	L/C

	CPI index		Real GDP		Lek/Eur exch. rate		T-Bill, 12M rate	
Sample period	ADF	DF-GLS	ADF	DF-GLS	ADF	DF-GLS	ADF	DF-GLS
1996Q1:2016Q4	L/C	D/C	D/C	D/C	D/C	L/C	L/T	L/C
1996Q1:2017Q4	L/C	D/C	D/C	D/C	D/C	L/C	L/T	L/C
1996Q1:2018Q4	L/C	D/C	D/C	L/T	D/C	L/C	L/T	L/C

Note: i) ADF = Augmented Dickey-Fuller test; DF-GLS = Dickey-Fuller GLS (ERS) test; ii) L/C = stationary in levels, I(0), including a constant; L/T = I(0), trend stationary; D/C = stationary in difference I(1), including a constant; iii) Lag length is selected automatically on the basis of Schwarz information criterion, BIC.

3.2 Selected time series forecasting models

In what follows, we make a brief description of the characteristics and estimation issues of our selected forecasting models, along the lines of Stock and Watson (1998) and Marcellino (2004). Each of the three methods as discussed in Section 2 is represented in our analysis here by a single commonly-used model. Due to its simplicity and timeliness, the classic autoregressive (AR) model is chosen as an indispensable forecasting tool within the linear framework. Pertaining to the nonlinear family, the time-varying parameter method will be represented by the logistic smooth transition autoregressive (LSTAR) model, which is special form of threshold autoregressive techniques.[2] Lastly, nonlinearity in the data is alternatively modeled by the nonparametric feed-forward artificial neural network (ANN), which is a computer artificial intelligence method that has gained considerable attention in recent decades as a promising forecast tool. The analysis uses a variety of model specifications, which are based on the choice of lags and some additional considerations.

3.2.1 Autoregression (AR)

An autoregressive model of order p takes the following linear form:

$$y_{t+1} = \alpha + \beta_1 y_t + \ldots + \beta_p y_{t-p+1} + \varepsilon_{t+1} = \alpha + \beta(L) y_t + \varepsilon_{t+1} \quad (2)$$

where $\beta(L)$ is a lag polynomial, and α , $\beta_1 \ldots \beta_p$ are parameters that can be linearly estimated by the ordinary least squares method. The specification of the model differs in terms of a) the lag length, b) stationarity transformation, and c) the inclusion of an additional dummy variable that takes into account the time trend.

In many studies, the lag length is commonly determined on the basis of Bayesian and Akaike information criteria. Due to a rather short estimation period (starting from 17 years of quarterly data), we follow a forecast-based model selection approach using 1 up to 4 alternative time lags ($1 \geq p \leq 4$ quarters). The results are, however, further compared with the model specifications as suggested by the BIC and AIC criteria to shed some light whether the latter can serve as a convenient shortcut to model lag length selection for forecast purposes.

Figures 1 to 4 displayed considerable persistency in the economic data, particularly in prices and gross domestic production. As data persistency could lead to spurious regressions, all equations are estimated with variables in their level, as well as the first difference form. In their empirical illustration, Clements and Hendry (1996) have pointed out that differencing the variables could do away with structural changes like level shift and also improve economic predictions, particularly through reductions in forecast bias. Yet another common approach to determine the specification of the forecasting models is based upon unit root pretests. Although Christiano and Eichenbaum (1989) and Rudebusch (1993) question their ability in selecting the "true" model, other authors (e.g., Stock & Watson, 1998; Allen & Fildes, 2005; Marcellino, 2008; Diebold & Kilian, 2000) report that pretesting can improve relative forecast accuracy and provide conditions under which it can happen. In view of

[2] *There are no straightforward rules for selecting a particular nonlinear family (Teräsvirta, 2005), therefore we select the LSTAR model in line with other studies.*

this, we conduct our forecast analysis based on the models in levels and differences, and later extend the discussion with suggestions from models as if having been pretested for the presence of unit root.

3.2.2 Logistic smooth transition autoregression (LSTAR)

The nonlinear method with time-varying parameters extends equation (2) as follows:

$$y_{t+1} = \alpha + d_t\beta(L)y_t + (1 - d_t)\varphi(L)y_t + \varepsilon_{t+1} \quad (3)$$

where β and φ denote parameters estimated on past data in different 'regimes', $\beta(L)$ and $\varphi(L)$, while d_t is a delay parameter that is expressed in a nonlinear function to determine whether the switch between the regimes occurs in a sharp manner (the threshold autoregressive model) or in a smoother way (smooth transition autoregression). The smooth transition function can take various forms, such as logistic, exponential, or cumulative distributive function. We assume the logistic form as in Stock and Watson (1998), where $d_t = 1/(1 + \exp[\gamma_0 + \gamma_1\zeta_t])$ and the threshold variable $\zeta_t = (1, y_t, y_{t-1}, \ldots y_{t-p+1})$ if y_t enters the model in levels or $\zeta_t = (1, \Delta y_t, \Delta y_{t-1}, \ldots \Delta y_{t-p+1})$ if y_t is differenced. Also, parameter γ_1 denotes the smoothness, or the shape of the parameter over time: a very large value of γ_1 makes the model look similar to a self-exciting threshold (SETAR) model with swift changes in parameters, and if $\gamma_1 = 0$ the model becomes linear.

As with the AR models, we estimate various LSTAR models with up to four time lags, specified in levels or differences, but no additional deterministic component other than a constant. To determine a fixed threshold variable for each model, we attempt alternative threshold variables of up to eight quarters as well as three different modes of choosing the starting parameter values and select among those that provide superior short-term forecasts.

3.2.3 Artificial neural networks (ANN)

The nonparametric nonlinear method is represented in our study by the feed-forward neural network models with up to two hidden layers. As a computer artificial intelligence method, they search for patterns in the data, learn them, and classify new patterns, which can be used to make predictions. The specification of ANNs consists of three layers, namely the input(s), hidden layer(s) and output. The inputs and output layers can be equivalently viewed as the regressors and the dependent variables, respectively. Contrary to them, the hidden layer has no parallel in econometrics, even though it behaves like output by allowing the information processed from the input node(s) to the node(s) in the hidden layer(s) and the output by certain "activation" functions. The *j*th node of the hidden layer *h* in a feed-forward ANN has a function as

$$h_j = f_j\left(\alpha_{0j} + \sum_{i \to j} w_{ij}x_i\right) \quad (4)$$

where α_{oj} is called the bias, w_{ij} denotes the weight of input node x_i feeding to j, and the activation function $f_j(.)$ is typically the logistic function, $f_j(z) = \exp(z)/(1+\exp(z))$. The network learns by applying various weights in each iteration, until forecast errors are minimized. The activation function of the output layer o in an ANN with j nodes in the hidden layer is defined as

$$o = f_o\left(\alpha_{0o} + \sum_{j \to o} w_{jo}h_j\right) \quad (5)$$

Combining the activation functions (4) and (5), a feed-forward ANN model with a single hidden layer h_1 (and a linear component) can then be written as

$$y_t = f_o\left[\alpha_{0o} + \sum_{i \to o} w_{io}x_i + \sum_{j \to o}^{h_1} w_{jo}f_j\left(\alpha_{0j} + \sum_{i \to j} w_{ij}x_i\right)\right] \quad (6)$$

where y_t is the variable of interest, and in a univariate model the inputs $x_i = (1, y_t, y_{t-1}, y_{t-p+1})$. The summation of input nodes $\sum_{i \to o} w_{io}x_i$ makes up the linear component that allows for a direct relationship between the input layer and the output layer, thus having the usual interpretation as in linear modeling. On the other hand, the summation of nodes in the hidden layer h depicts the nonlinear ANN component, which can be viewed as a set of time-varying intercepts that evolve according to their logistic functions. The higher the number of nodes in h, the better the fit of the model to any type of temporal evolution. By increasing the number of nodes in the hidden layer, one

can actually parametrize a general continuous nonlinear function (Tsay, 2002). Augmenting equation (6) with another hidden layer h_2 with k nodes could enhance the flexibility of the neural network (Ghysels & Marcellino, 2018):

$$y_t = f_o\left[\alpha_{0o} + \sum_{i\to o} w_{io}x_i + \sum_{k\to o}^{h_2} w_{ko} f_k\left(\sum_{j\to k}^{h_1} w_{jk} f_j\left(\alpha_{0j} + \sum_{i\to j} w_{ij}x_i\right)\right)\right] \quad (7)$$

Again, there is no theoretical basis for selecting the network, therefore the number of nodes in the input layer and the hidden layer(s) are often chosen through experimentation or by trial-and-error method. In line with Stock and Watson (1998), we use a neural network with single and double layers. The models are differentiated then in terms of the number of input nodes, the hidden nodes, and the specification in levels and differences. Both variants of the neural network include up to four lags in the input layer, $p = \{1,2,3,4\}$. Next, the single layer network has four nodes in hidden layer, $h_1 = \{1,2,3,4\}$; however, as the ANN models are data-intensive models and require a sufficient number of observations the number of nodes in the network with two hidden layers changes to $h_1 = \{2,3,4\}$ and $h_2 = \{1,2\}$.

Forecasting with neural networks involves two steps. In the first step, a specific ANN structure is trained by using a fixed number of nodes and choosing their biases and weights (i.e. estimating the intercepts and parameters, respectively) through a minimization of some fitting criterion. We use the Levenberg-Marquardt optimization, which makes a recursive estimation on different starting values of the biases and weights until a global minimum value is achieved. We test up to 200 configurations of biases and weights and select the one that provides the lowest RMSE in the last 8 quarters. Once the ANN structures are trained and tested, they are then used to compute out-of-sample forecasts in order to select the best performing network. This method is similar to the cross-validation test in statistical analysis intended to identify problems like overfitting or selection bias.

3.3 Forecasting procedure

Table 3 summarizes the forecasting methods and model specifications. As described in the previous subsection, we compare the forecast ability among three methods, specifically the linear autoregressive (AR) technique, the time-varying LSTAR and the nonlinear artificial neural network (ANN). All models are estimated with variables in levels or transformed in first differences, irrespective of their stationarity condition. Further, the column on lag length specification indicates the number of lags, p, that are included in each alternative specification (please note that in the case of ANN models, p denotes the number of input nodes). Although the Bayesian and the Akaike information criteria would generally recommend using more than four lags[3] particularly in the time-varying models, the lag length in our analysis is restrained to range from 1 to 4 quarterly lags. The last column of the table displays the division or augmentation of models in terms of the deterministic and nonlinear components. The AR models are augmented with a time trend, while the LSTAR equations are regressed on 1 to 6 nonlinear thresholds each at a time. Finally, the ANN models are initially trained on a single hidden layer with 1 to 4 nodes; in the next exercise we include a double hidden layer, in which the first layer h_1 starts with 2 to 4 nodes while the second layer h_2 has up to 2 nodes. As a consequence, our forecast comparison exercise comprises a total of 144 models; of which, 16 belong to the linear method, 48 are time-varying, while the remaining 80 models pertain to neural networks (the ANNs with single and double hidden layer(s) consist of 32 and 48 models, respectively).

[3] *To save space, the results about information criteria on models with up to eight lags are not shown here, but can be at the readers' disposal upon request. Table 7 displays, nevertheless, a summary of the lag length selection by BIC and AIC on models regressed on up to four lags.*

Table 3. Summary of Methods and Forecasting Models

No. of models	Stationarity assumptions	No. of models	Stationarity assumptions	Method/ technique	Lag length specification	Deterministic and Nonlinear components
1	Level	73	Differenced	Linear/ AR	1	Constant
2		74			2	
3		75			3	
4		76			4	
5		77			1	Constant and Trend
6		78			2	
7		79			3	
8		80			4	
9-14		81-86		Time-varying/ LSTAR	1	Constant and 1 to 6 thresholds
15-20		87-92			2	
21-26		93-98			3	
27-32		99-104			4	
33-36		105-108		Nonlinear/ ANN	1	Single hidden layer, $h_1=\{1,2,3,4\}$
37-40		109-112			2	
41-44		113-116			3	
45-48		117-120			4	
49-54		121-126			1	Double hidden layers, $h_1=\{2,3,4\}$; $h_2=\{1,2\}$
55-60		127-132			2	
61-66		133-138			3	
67-72		139-144			4	

The forecasting experience has often demonstrated that a model with satisfactory in-sample forecasts does not guarantee similarly successful outcomes in the out-of-sample period. Therefore, the whole sample period under investigation, which ranges from 1996Q1 to 2018Q4 has been divided into the so-called training period up to 2012Q4 (68 quarters), and the forecast evaluation period that runs from 2013Q1 through 2018Q4 (24 quarters). To evaluate the forecast performance we rely on the smallest estimated forecast errors, as measured by the root mean squared error (RMSE). More precisely, the forecasting procedure starts with the model estimations over the period 1996Q1:2012Q4 where for each of its specifications according to the form of variables and the number of lags we note down its forecast performance for the 1, 4 and 8 quarters ahead. The original estimation period is then recursively extended by a quarter (1996Q1:2013Q1), wherefrom the RMSE of forecasts for each of the desired horizons is computed and retained correspondingly. The evaluation process for every model is repeated 23, 20, and 16 times for the 1, 4 and 8 quarters forecast horizon, respectively, until the last quarter of 2018 is reached.

4. Empirical Results and Model Comparison

We now evaluate the forecast performance of the 144 models for each of the 4 economic variables as described above. Table 4 displays the ranking of the competing models based on the average RMSE over the whole set of variables.[4] Clearly, the majority of best performing models belong to the nonlinear method. The artificial neural network models with a double hidden layer structure seem to provide the best results in forecasting 1 quarter ahead, while the time-varying LSTAR estimations

[4] *The RMSE of model m is computed as the average* $RMSE^h_{n,m}$ *over the whole set of variables n for each forecast horizon h:* $RMSE^h_m = \frac{1}{N}\sum_{n=1}^{N} \frac{RMSE^h_{n,m}}{RMSE^h_{n,1}}$ *All models are compared to the benchmark model AR(1) with a constant, specified in levels.*

specified in differences perform better in 4 and 8 quarters ahead. On the other hand, the linear model (AR(3) in differences) is only ranked third when forecast horizon is equal to 8, but it does not reach the top five places for the short-term forecasts.

Table 4. Best Performing Models for the Whole Set of Variables

Rank\Horizon	h=1	h=4	h=8
	All methods		
1	ANN,D,4-3-2-1	LSTAR,D,1	LSTAR,D,2
2	ANN,D,1-4-2-1	LSTAR,D,2	LSTAR,D,1
3	ANN,D,3-3-1-1	LSTAR,D,4	AR,D,3
4	ANN,D,1-2-1-1	ANN,D,1-2-1	ANN,D,1-2-1
5	ANN,D,3-4-2-1	LSTAR,D,3	LSTAR,D,3
	Linear method (AR)		
1	AR,D,T,1	AR,D,3	AR,D,3
2	AR,D,1	AR,D,T,3	AR,D,2
	Time-varying (LSTAR)		
1	LSTAR,D,1	LSTAR,D,1	LSTAR,D,2
2	LSTAR,D,3	LSTAR,D,2	LSTAR,D,1
	Non-linear (ANN)		
1	ANN,D,4-3-2-1	ANN,D,1-2-1	ANN,D,1-2-1
2	ANN,D,1-4-2-1	ANN,D,1-3-1	ANN,D,3-1-1

Table 5 shows the ranking of the methods based on the fraction of variables for which it is found the lowest RMSE. It turns out that the nonlinear method offers the best predictions in 83 percent of the cases, being dominated by feed-forward ANN models (67 percent). On the other side, the linear modeling technique is only useful for around one-sixth of the cases. The potential of nonlinear models is evident in all forecast horizons, particularly for 1 quarter ahead predictions (100 percent of the cases, totally controlled by ANNs). Computing the second and third best options brings little changes to the above picture, with the linear AR models performance swinging from 8 to 25 percent, respectively.

Table 5. Fraction of Variables for Which a Forecast Method Has the Lowest RMSE

Rank	Horizon	AR	LSTAR	ANN
1		0.17	0.17	0.67
	h=1	0.00	0.00	1.00
	h=4	0.25	0.25	0.50
	h=8	0.25	0.25	0.50
2		0.08	0.08	0.83
	h=1	0.00	0.00	1.00
	h=4	0.00	0.25	0.75
	h=8	0.25	0.00	0.75
3		0.25	0.00	0.75
	h=1	0.00	0.00	1.00
	h=4	0.50	0.00	0.50
	h=8	0.25	0.00	0.75

Because nonlinear models are still expensive to maintain, particularly the ANNs, it is important to know whether their forecast gains are statistically different from linear methods or not. For that reason, we conduct the Diebold-Mariano (DM) test to formally determine if the forecasts errors from two methods are statistically the same or show different predictive accuracy. In the DM test, the null hypothesis is that the loss differential between the two forecasts has zero expectation for all t, i.e. both models have equal accuracy.

Table 6 summarizes the loss differential between the best models in the linear and nonlinear

groups and the statistical significance based on the DM test. Initially, we present the relative RMSE in order to show the prediction accuracy of one model in relation to the others. Because it is measured as the RMSE of AR models over nonlinear techniques, a value below 1 indicates that AR forecasts have outperformed the LSTAR predictions; otherwise, a ratio that is above 1 indicates that AR predictions have on average performed worse.

The results on the whole set of variables (part 6.A) give us the impression of a general underperformance of linear models, particularly in comparison to time-varying models. The RMSE ratio between the best AR and LSTAR forecast models varies from 1.08 to 1.13, indicating that the best linear autoregressive model forecasts have underperformed around 8 to 13 percent over the 2013q1:2018 period. Nevertheless, the p-values of the Diebold-Mariano statistic suggest that the loss differential between the two models could only be statistically different when forecasting 8 quarters ahead, but they might not be very different at h=1 and h=4. Perhaps the worst linear models performance is against neural network models at one quarter ahead forecasts – as the RMSE ratio (1.37) and the DM test p-value (0.03) suggest – though they may be a useful tool in predicting four to eight quarters ahead.

Table 6. Statistical Significance of the Loss Differential between Forecasts

A. Whole set of variables

Forecast horizon	h=1	h=4	h=8
Relative RMSE: AR over LSTAR	1.08	1.12	1.13
Diebold-Mariano Test:			
Order	4	4	4
DM Stat	0.64	0.07	2.35
p-value	0.52	0.95	0.02
Relative RMSE: AR over ANN	1.37	1.01	0.95
Diebold-Mariano Test:			
Order	5	4	4
DM Stat	2.18	-0.02	-0.06
p-value	0.03	0.98	0.95

B. Individual variables

Variables	CPI			GDP			Exch. Rate			T-Bill		
F'cast horizon	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8	h=1	h=4	h=8
RMSE ratio: AR/LSTAR	0.87	0.66	0.43	1.07	1.05	1.06	1.03	1.21	1.16	1.03	1.00	0.94
DM test:												
Order	4	4	4	4	4	5	4	4	4	4	4	4
DM Stat	13.93	4.38	5.79	21.09	0.19	2.57	0.67	1.17	0.94	2.01	0.13	0.21
p-value	0.00	0.00	0.00	0.00	0.85	0.01	0.51	0.24	0.35	0.04	0.89	0.83
RMSE ratio: AR/ANN	1.11	0.75	0.60	1.70	0.92	0.98	1.79	1.28	1.38	1.36	1.14	1.21
DM test:												
Order	4	4	4	4	4	4	4	4	4	4	4	4
DM Stat	1.33	4.74	8.05	1.62	3.56	0.76	1.22	1.56	0.77	0.06	1.47	1.37
p-value	0.18	0.00	0.00	0.11	0.00	0.45	0.22	0.12	0.44	0.95	0.14	0.17

Note: 1) the AR/LSTAR ratio uses models that are ranked best in each method; 2) The DM test is based on the squared-errors method; when this is not possible to compute, the absolute-error method is used instead.

The lower part of the table (6.B) presents the forecast gains in individual variables between models that are ranked best in each method. It turns out that the AR model has outperformed both the LSTAR and ANN models in forecasting CPI prices. With the exception of relative RMSE calculated for AR/ANN at h=1, the other forecast error ratios range considerably below 1 at between 0.87 and 0.43. The statistical significance derived from the DM test strongly approves the superiority of the linear AR model in predicting consumer prices (p-values are closer to zero for all forecast horizons). It is

found the linear method could also be effective in predicting gross domestic production at four quarters ahead. Nevertheless, its forecast ability appears to have generally underperformed with regard to the other variables. The relative RMSE values are higher than one in most cases, especially the exchange rate ratios. However, the poor performance of linear models is found statistically significant only when forecasting GDP at h=1 and h=8, as well as the interest rate at 1 quarter ahead. Perhaps surprisingly, their seemingly bad performance in the exchange rate case (with the RMSE often higher than 20 percent) is not shown statistically relevant by the DM test.

Last but not least, we now turn to the discussion about the relevance of using unit root tests or certain information criteria in model specification for forecasting purposes. Results from Table 4 gave us the impression that specifying the models in differences and using a couple of own lags might help in attaining lower RMSEs within each forecast method. Yet, the ADF and DF-GLS tests on stationarity (in Table 2) were only clear about using differences in the case of real GDP. They provided opposing recommendations for prices and the exchange rate, while both concluded that the interest rate is stationary in levels. Moreover, the Bayesian and Akaike information criteria on lag length selection were mostly in favor of a plenteous number of lags, rather than a couple of them as might be inferred from Table 4.

Table 7 combines model selections as recommended by stationarity tests and information criteria for modeling each variable. In general, they do not match with our findings on the top two best models within each group, i.e., the linear, time-varying and artificial neural network[5] method. For example, the best AR models ranked first and second for predicting consumer prices correspond to model specification in levels, which are regressed on a couple of own lags and a time trend component. However, the unit root tests suggested that CPI is stationary in level (ADF) or in first difference (DF-GLS) without a deterministic trend, and both information criteria (BIC/AIC) determined to choose an optimal number of at least three lags. This discrepancy is evidenced throughout the linear AR models, with very few exceptions such as the model selection for GDP, whose order of integration and the number of lags match all with the best AR model performers. Furthermore, the picture with the time-varying LSTAR models is similarly unclear. Although the DF-GLS (ADF) stationarity test suggest the right variable transformation in three (two) cases, they fail to be accurate in all of them. Also, employing the BIC (AIC) information criterion for the optimal lag length selection turns out to be appropriate for only two (one) out of four variables. The inability to find an appropriate measure for model selection in our univariate analysis is consistent with Vika's (2018) concluding remarks in a vector autoregression analysis. The problem might be related to the insufficiently large number of observations at disposal for economic time series in Albania. Yet, numerous simulations in the book by McQuarrie and Tsai (1998) suggest that there might not be any generally "best" criterion.

Table 7. Comparing Model Selection by Information Criteria and Stationarity Tests

		CPI		GDP		Exch. Rate		T-Bill	
Models	Assumption	Stationary	BIC/AIC	Stationary	BIC/AIC	Stationary	BIC/AIC	Stationary	BIC/AIC
AR	Level	ADF	3		3/4	DF-GLS	4	DF-GLS	2
AR	Level, with trend		4		1/4		4	ADF	2
AR	Differenced	DF-GLS	4	ADF/ DF-GLS	2/3	ADF	3/4		1/4
AR	Differenced, with trend		4		2/3		3/4		1/4
LSTAR	Level	ADF	3		3/4	DF-GLS	4	DF-GLS	2/3
LSTAR	Differenced	DF-GLS	1/4	ADF/ DF-GLS	3	ADF	4		4

[5] *The discussion on ANN models is limited to the unit root test. It could be interesting to assess which neural network structure(s) would the BIC or AIC recommend, but our information criteria analysis has focused here only on the popular econometric/parametric AR and LSTAR models.*

5. Instability Tests and Some Additional Robustness Check

The forecast evaluation revealed that nonlinear models could be better suited to forecast the main economic indicators in Albania. As the underperformance of linear specifications might result from potential instabilities in the data during the twenty-three years of investigation, testing for parameter stability can provide important information about model adequacy and out-of-sample forecasting accuracy (Stock, 1994). In this section, we attempt to shed light on the parameter constancy in our best forecasting linear models. In line with Stock and Watson (1996), we apply a number of stability tests that check whether the variation in coefficients remains in control and there are unknown structural breaks in the sample period.

The Breusch-Pagan-Godfrey (BPG) test examines on the assumption of homoscedasticity in linear regressions, against the alternative of heteroscedastic, or "differently scattered" errors. Other popular tests make use of the recursive residuals in order to inspect for possible structural changes over time. We focus on the techniques that are considered appropriate for time series data and do not require a prior knowledge about when a structural break might have occurred. Two widely used applications, in this respect, are the cumulative sum (CUSUM) and cumulative sum of squares (CUSUMSQ) statistics of Brown, Durbin and Evans (1975), which basically test if the cumulative sum (of squares) of recursive residuals moves outside of its confidence interval. Additional to these two recursive least squares methods we employ certain stability diagnostics that test for one or more regime changes in our estimation sample. Although there may be obvious periods at which economic series give indications of structural changes, we assume no knowledge about break dates on individual series and thus consider the Quandt-Andrews unknown breakpoint tests, namely the Maximum statistic (MLR), the Exponential statistic (ELR) and the Average statistic (ALR) (see Quandt, 1960; Andrews, 1993; and Andrews & Ploberger, 1994), as well as a multiple breakpoint test based on the global information criteria method. As structural breaks can be hard to identify near the beginning and the end of a sample, Andrews recommends a 15 percent "trimming" of the data if there is no reason to assume break date(s). We, nevertheless, leave out the first and last 10 percent of the observations, so as to allow the testing procedure to include at least some of the adverse effects in the variables around the 1997 economic turmoil.

Table 8 presents the results on parameter stability in the best performing linear AR models. We discuss them variable by variable and make inference with respect to their relative performance as reported in Table 6, part B. The stability tests for the CPI and T-bill rate linear models suggest that the estimated parameters are in general not constant and could suffer from heteroscedasticity and structural breaks. The most likely breakpoint location for both variables is around the middle of 1998, as suggested by the Quandt-Andrews maximum statistic and the Schwarz criterion in the multiple breakpoint test, too. Despite the strong instability in coefficients and the suggested breakpoints, the best AR linear model has still outperformed nonlinear models in the case of consumer prices, while showing similar ability to forecast the T-bill 12-month interest rate.

Next, the stability tests reveal somewhat incongruous results for GDP and the LEK/EUR exchange rate. The assumption of homoscedasticity in the regression is rejected by the BPG test, and the CUSUMSQ test suggests a deviation of parameters outside the 95% confidence interval. On the contrary, the remaining six stability diagnostics find no structural breaks in parameters. Interestingly, though the stability tests might show favor of using linear models for these two economic series, their forecast ability is found to be poorer than that of the best time-varying-parameter models.

Table 8. Stability Tests for Best Performing AR models[a)]

	CPI	GDP	Exch. Rate	T-Bill, 12m
Heteroscedasticity test:				
BPG, F prob.	0.00	0.09	0.00	0.00
Cumulative sum (of squares) tests. Period of moving outside 5% confidence interval:				
CUSUM test	Yes: 1999q1 2000q2	NO	NO	Yes: 1998Q2 2018Q4
CUSUMSQ test	Yes: 1998q3 2014q3	Yes: 1998q4 2002q1	Yes: 1997q2 2014q3	Yes: 1998q1 2014q3
Quandt-Andrews unknown breakpoint test. Null hypothesis: No breakpoints within 10% trimmed data:				
Max LR F, prob.[b)]	0.00	0.35	0.10	0.00
Break location	1998Q3	1999Q1	1999Q2	1998Q3
Exp LR F, prob.[b)]	0.00	0.69	0.33	0.00
Ave LR F, prob.[b)]	0.03	0.66	0.40	0.20
Multiple breakpoint test: Selected breaks based on global information criteria method (trimming 10%):				
Schwarz crit.	1	0	0	2
LWZ crit.	1	0	0	1
Est. break dates	1998Q2	-	-	1998Q2; 2012Q4

[a)] Best linear models are represented by the AR(1) type specified in levels for the T-bill rate; in differences for GDP; in levels and including a trend for the CPI; and in difference and including a trend for the lek exchange rate.
[b)] Probabilities calculated using Hansen method in the EViews software.

Because stability tests were not distinctly defined and some of them evidence parameter instabilities or structural breakpoints following the collapse of pyramid schemes in 1997, we attempt an additional robust analysis to safeguard the forecast ability of linear models. Initially, the AR models are re-estimated by using a shorter sample period that leaves out the 1990s. We tried different starting periods from the beginning of every year from 2000 to 2003 and checked in vain whether the new forecasts could improve their performance in Table 5. The exercise produced better forecasts with respect to the whole estimation period (starting from 1996), especially for consumer prices and the short-term interest rate; yet, in neither of the re-estimated periods were they able to increase the fraction of variables for which a lower RMSE is computed.

Another way to improve the forecast performance of AR models in the presence of structural breaks is to enter the variables in second differences, as it is proposed by Clements and Hendry (1999). This method is expected to take into account much of the economic and institutional changes that have taken place during the transition period or in the aftermath of the global financial crisis. The results only proved important for the exchange rate forecast performance at all horizons and for GDP at h=4. This helped the AR models to double the fraction of variables with the lowest RMSE in the first rank to 33 percent (though figures in the second rank did not change whereas in the third rank deteriorated), while reducing both the LSTAR and ANN model performance by 10 percentage points to 8 and 58 percent, respectively.[6]

To summarize, stability tests do not support linear modeling in two out of four variables in our analysis. Yet, heteroscedasticity and structural changes do not seem to affect their forecast performance against the nonlinear methods, as in the case of consumer prices. Alternative autoregressive specifications, such as treating variables as integrated of order $I(2)$, may yield fruitful results even in spite of good-performing nonlinear models.

[6] The results about the robustness analysis are not shown in the material, but can be provided by the authors upon request.

6. Concluding Remarks

Nonlinear forecasts have performed better than linear forecasts for most variables at almost all forecast horizons. Analyzing the whole set of variables, the feed-forward ANN models performed especially well in predicting 1 quarter ahead, while the time-varying parameter LSTAR models showed better results at longer horizons of 4 and 8 quarters.

Although it is tempting to use classic autoregressive models due to their relatively low costs, the linear framework has shown limited forecast ability in the context of Albanian economic indicators. We find that AR models outperformed both the LSTAR and ANN models only in forecasting CPI prices, and could also be useful in predicting GDP at four quarters ahead. Perhaps the successful performance of AR models with respect to the CPI index may be related to the relatively low volatility of prices around their trend, especially during the forecast evaluation period from 2013 to 2018.

The stability tests for the linear models suggest that parameters generally suffer from heteroscedasticity and likely structural changes, especially in the case of consumer prices and the T-bill interest rate. Potential instabilities in the data during the twenty-three years of investigation may often show favor of using nonlinear models, yet they do not seem to affect the relative AR model forecast performance. Neither the omission of the rather volatile period in the 1990s, nor the alternative estimation with variables in second differencing are able to change the worse forecast performance of linear autoregressive models.

On the whole, our findings should instigate macroeconomic modelers to enhance their suite of models with alternatives beyond the linear estimation framework, with due considerations in the direction of artificial neural networks. In this context, an interesting future research would certainly be to optimize on time efficiency. As trying just a few starting values of biases and weights was not very promising, an increase to 200 trials produced substantial improvements in finding the best forecast configuration. Yet, training and selecting the best network out of 80 structures specified in level and in difference necessitated about two days for forecasting each variable using Matlab program in a common desktop PC. That is an enormous time to deal with if forecasters need to prepare projections on a number of economic variables within, let say a couple of weeks. Therefore, improving the neural network tools to optimize efficiency in forecasting is important.

7. Acknowledgement

Disclaim: Research Articles are considered as preliminary work that aim at stimulating debate and critical comments. Therefore, they express the views of the authors and do not necessarily represent those of the institutions where they work. The authors are especially grateful to Luca Rossi for the invaluable discussion on the paper during the "14th SEE Research Workshop" organized by the Bank of Albania in December 10-11, 2020.

References

Allen, G. P. & Fildes, R. (2005). Levels, differences and ECMs – Principles for improved econometric forecasting. *Oxford Bulletin of Economics & Statistics,* Vol. 67, No. S1 (December 2005) 881-904.

Andrews, D. W.K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, Vol. 61, No. 4 (July 1993), 821-856.

Andrews, D. W.K. & Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, Vol. 62, No. 6 (Nov. 1994), 1383-1414.

Brown, R. L., Durbin, J. & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 37, No. 2 (1975), 149-192.

Christiano, L. J. & Eichenbaum, M. (1989). Unit roots in real GNP: Do we know, and do we care? *NBER Working Paper No.* 3130.

Clements, M. P. & Hendry, D. F. (1996). Intercept corrections and structural change. *Journal of Applied Econometrics*, Vol. 11, No. 5, Special Issue: Econometric Forecasting (Sep. – Oct., 1996), pp. 475-494.

Clements, M. P. & Hendry, D. F. (1999). Forecasting non-stationary economic times series. *MIT Press*, Cambridge, Massachusetts.

Diebold, F. X. & Kilian, L. (2000). Unit-root tests are useful for selecting forecasting models. *Journal of Business & Economic Statistics*, Vol. 18, No. 3 (July 2000), pp. 265-273.

Ghysels, E. & Marcellino, M. (2018). *Applied economic forecasting using time series methods*. Book published by Oxford University Press, 2018.

Marcellino, M. (2004). Forecasting EMU macroeconomic variables. *International Journal of Forecasting*, 20, 359-72.

Marcellino, M. (2008). A linear benchmark for forecasting GDP growth and inflation? in *Journal of Forecasting*, Volume 27, Issue 4, pp. 305-340, July 2008.

McQuarrie, A. D. R. & Tsai, C. (1998). *Regression and time series model selection*. Book published by World Scientific Publishing Co. Pte. Ltd.

Quandt, R.E. (1960). Tests of hypotheses that a linear system obeys two separate regimes. *Journal of the American Statistical Association*, 55, 324-330.

Rudebusch, G. D. (1993). The uncertain unit root in real GNP. *American Economic Review*, Vol. 83, No. 1, (Mar. 1993), pp. 264-272.

Stock, J. H. (1994). Unit roots, structural breaks and trends. in *Engle R, McFadden D Handbook of Econometrics*, Vol. IV, Chapter 46, 1994, pp. 2740-2843.

Stock, J. H. & Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *American Statistical Association, Journal of Business & Economic Statistics*, January 1996, Vol. 14, No. 1, pp. 11-30.

Stock, J. H. & Watson, M. W. (1998). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. *NBER Working Paper No*. 6607.

Teräsvirta, T. (2005). Forecasting economic variables with nonlinear models. *SEE/EFI Working Papers Series in Economics and Finance No*. 598.

Tsay, R. S. (2002). *Analysis of financial time series*. Book published by John Wiley & Sons, Inc.

Vika, I. (2018). Practical issues in forecasting with vector autoregressions. *Bank of Albania, Economic Review*, 2018 H2.

Zivot, E. (2003). *Lectures on structural change*. University of Washington, Department of Economics, April 5, 2003. https://faculty.washington.edu/ezivot/book/structuralchangeslides1.pdf