Forecasting the US CPI: Does Nonlinearity Matter?

Marcos Álvarez-Díaz* and Rangan Gupta**

Abstract

The objective of this paper is to predict, both in-sample and out-of-sample, the consumer price index (CPI) of the United States (US) economy based on monthly data covering the period of 1980:1-2013:12, using a variety of linear (random walk (RW), autoregressive (AR) and seasonally-adjusted autoregressive moving average (SARIMA)) and nonlinear (artificial neural network (ANN) and genetic programming (GP)) univariate models. Our results show that, while the SARIMA model is superior relative to other linear and nonlinear models, as it tends to produce smaller forecast errors; statistically, these forecasting gains are not significant relative to higher-order AR and nonlinear models, though simple benchmarks like the RW and AR(1) models are statistically outperformed. Overall, we show that in terms of forecasting the US CPI, accounting for nonlinearity does not necessarily provide us with any statistical gains.

JEL Codes: C22, C45, C53, E31

Keywords: Linear, Nonlinear, Forecasting, Consumer Price Index.

1. Introduction

Inflation forecasts are of paramount importance for central banks around the world for proper designing of monetary policy. Hence, the Federal Reserve Bank of the United States (US) produces regular inflation forecasts at various forecast horizons. From a theoretical perspective, Faust and Wright (2013) assert that the non-neutrality of monetary policy in the New-Keynesian framework has led to the importance of inflation forecasting from a policy perspective. Many methodologies have been applied in forecasting US inflation. For an excellent summary of models see Stock and Watson (2009), which compares the forecasting performance of various Phillips curve specifications, and also Koop and Korobilis (2012) and Faust and Wright (2013) for the

* Corresponding author. Department of Economics, University of Vigo, Galicia, Spain. Email: marcos.alvarez@uvigo.es.

** Department of Economics, University of Pretoria, Pretoria, 0002, South Africa. Email: rangan.gupta@up.ac.za.

performance comparison of newer methods. These methodologies range from simple linear models to large data-driven econometric models and from structural models to nonlinear specifications.

Stock and Watson (2010) indicates that there have been tremendous changes in inflation dynamics in the US, both due to transformations in the US economy in general, and, as well as, the stance of monetary policy . Hence, it is not surprising that research over the past decade has documented considerable instability in inflation forecasting models. Given this instability, inflation forecasters seem to have a dearth of reliable multivariate models for forecasting inflation. In fact, as Stock and Watson (2010) points out, it is the simple univariate forecasting models that tends to perform better relative to their complex multivariate counterparts. Against this backdrop, the objective of this paper is to predict (in-sample and out-of-sample) consumer price index (CPI) of the US economy based on monthly data covering the period of 1980:1-2013:12, using a variety of linear and nonlinear univariate models. Specifically, we compare the forecasting abilities of two nonlinear models, namely the artificial neural network (ANN) and genetic Programming (GP) models with the best seasonally autoregressive integrated moving average (SARIMA) model. Moreover, our comparison is also completed by the forecasts of two naive forecasting models: the random walk model and the autoregressive model. Though large varieties of nonlinear models exist, we decided to choose the ANN and GP models over the other nonlinear models because they have the main advantage of being extremely flexible, since the forecaster does not impose any a priori and discretional assumptions on the functional form of the model. Indeed, it is the method itself which determines the functional form of the model. Therefore, the forecaster just "*lets the data speak for itself*". This functional flexibility allows us to exploit the possible non-linearity existing in the data and, therefore, it would be possible

to increase our predictive capacity. ANNs have been successfully applied to predict the evolution of the US CPI (McAdams and McNelis, 2005); however, this is the first time that genetic programming is considered into the analysis.

At this stage, it is important to emphasize the reasons behind the choice of our sample period. While the end point of the sample is purely driven by data availability at the time of writing this paper, the starting point of the sample is in line with a major break point corresponding to significant changes in US monetary policy (Ireland, 2004). In addition, 1980 also roughly coincides with the end of the Volcker stabilization and disinflation era. Therefore, the period of 1980-2013 is characterized by a more stable monetary and financial structure, to the extent that the US economy was believed to be in an indeterminate equilibrium before, but not after, October 1979 (Benati and Surico, 2008). To the best of our knowledge, this is the first attempt to compare in-sample and out-of-sample inflation forecasting performances of ANN and GP models relative to standard linear benchmark models over a period which not only includes the stable US economic environment, but also the unusual period of the "Great Recession".

The rest of the article is presented as follows: After this introduction, Section 2 presents the two non-linear forecast methods used in our study. In Section 3, we describe the data and how we assess the forecasting ability. The results obtained are presented in Section 4 and discussed in Section 5. Finally, in Section 6 we draw our conclusions.

2. Nonlinear forecasting methods

Nonlinear autoregressive neural networks

Artificial neural networks (ANN) are information processing procedures that have received a great deal of attention in the last years. These procedures, which are based on how biological neurons work, have been successful applied to model and predict different financial and economic time series, included the modeling and forecasting of the levels of inflation for different economies (Aiken, 1999; Binner et al., 2005; McAdam and McNelis, 2005; Choudary and Haider, 2012, among others). The reason that explains this popularity is that these procedures are very powerful to detect nonlinear structures that might be hidden in the data, even under conditions of incomplete data or where the presence of noise is important. In fact, many theoretical studies have demonstrated that this method can approximate any non-linear function with a specific degree of accuracy (Cybenko, 1989; Hornik *et al.*, 1989; White, 1990).

In the specialized literature we can find different kinds of networks (Rumelhart and McClelland, 1986; Zhang, 2004), although the feed-forward multi-layer network is by far the most popular network employed in forecasting economic and financial data (Wong *et al.*, 1995; Wong *et al.*, 2000; Yao, Li and Tan, 1997). This is the type of network employed in our forecasting study. Actually, we use a non-linear autoregressive neural (NAR) network, which can be considered as a general nonlinear autoregressive model. To be more precise, given any time series$\{y_t\}$, the NAR neural network performs a non-linear functional mapping from past observations $\left(y_{t-1}, y_{t-2},..., y_{t-p}\right)$ to the future value $\left(y_t\right)$ as follows

$$y_t = f\left(\beta_0 + \sum_{h=1}^{H} \beta_h \cdot g_h\left(\alpha_{h0} + \sum_{p=1}^{P} \vartheta_{hp} \cdot y_{t-p}\right)\right) + \varepsilon_t \qquad (1)$$

where $f(\cdot)$ is the output function, $g_h(\cdot)$ is the activation function, P is the number of lags of the variable, and H is the number of activation functions in the network. The parameters β_0, β_h, α_{h0} and ϑ_{hp} are known as the weights of the network. These weights are initially determined randomly within a given range of values. An iterative training process based on the back-propagation technique (Rumelhart et al., 1986) is used to find the weights that minimize a certain predictive error measure that, in our specific case, was the sum of the squared errors (SSE). The iterative process finishes after a fixed number of iterations is achieved, or when the error measure falls below some specific value. Finally, ε_t is a disturbance term which is assumed to be an independent and identically distributed random variable.

The forecasting performance of our NAR not only depends on the complexity of the data, but also on selecting correctly an adequate specification of the network represented in equation (1). This selection process, which is called network training, is not an easy task because of the large number of factors related to the selection process of an optimal specification (Zhang et al., 2001). The design of a network requires choosing an adequate functional form for $f(\cdot)$ and $g_h(\cdot)$, and to determine the number of lags (P) and activation functions (H). An optimal choice of these parameters ensures the success of the network to make good forecasts.

The first key decision, therefore, is to choose an optimal mathematical structure for the output function and the activation functions. Regarding this, it is recommended to choose a linear functional form for $f(\cdot)$, and a sigmoid function for $g_h(\cdot)$ (Faraway and Chatfield, 1998; Zhang and Kline, 2007). Probably, the most important and difficult decision is to specify correctly the number of lags of the variable (P), and the number activations functions (H) of the network. A correct specification of these factors is

crucial for the forecasting success of the network. Indeed, too low values of *P* and/or *H* may lead to the network not being capable to adequately capture the dynamic of the time series. On the contrary, too high values may lead to produce a spuriously good fit which does not lead to better forecasts when new data are considered. The latter problem is called overfitting, and is quite common in forecasting time series using neural networks. To avoid this problem, it is necessary to find a parsimonious specification that fits the data well. Unfortunately, there is no general rule universally accepted that allows us to automatically select the optimal number of lags and activation functions. Nevertheless, a common practice usually recommended in the literature is to select these values through a trial and error process (Zhang et al., 2001; Zhang and Kline, 2007). Specifically, in our forecasting study, we assume the fixed perspective explained in Kaastra and Boyd (1996) to select these values. This approach lies in constructing several networks by combining different values of *H* and *P*. In our case, the value of *P* varies from 1 to 20, and the value of *H* ranges from 1 to 5. In total, we have taken into account more than one hundred networks. The predictive performance of each one of these networks is assessed in a specific set of data, called validation set. The network which performs best in the validation set was finally selected to make out-of-sample predictions.

Genetic programming

Genetic algorithms are a broad class of related computational procedures that have in common to be inspired by biological concepts based on the theory of evolution of species (Holland, 1975; Goldberg, 1989). These population-based, probabilistic procedures look for an optimal solution to a problem from a population of possible

solutions. This search is based on the Darwinian biological principles of natural selection and survival of the fittest individual. In our specific forecasting study we use one of these computational procedures, called genetic programming (Koza, 1992). This type of genetic algorithm simulates artificially in a computer an evolutionary process so as to find an optimal model that adequately represents the underlying dynamic of an observed time series. This procedure has theoretically demonstrated a good performance in modeling the data-generating process of one-dimensional time series (Szpiro, 1997a; Yadavalli et al., 1999; Álvarez et al., 2001), and spatio-temporal time series (Álvarez et al. 2000). Moreover, genetic programming has empirically demonstrated to forecast accurately time series from many different disciplines, including economics (Koza, 1995; Szpiro, 1997b; Beenstock and Szpiro, 2002; Álvarez-Díaz and Caballero, 2008), Finance (Neely et al., 1997; Allen and Karjalainen, 1999; Fyfe et al., 1999; Kaboudan, 2000; Álvarez-Díaz and Álvarez, 2003; Álvarez-Díaz and Álvarez, 2005; Álvarez-Díaz, 2010). The procedure has some important advantages over other nonlinear forecasting methods such as artificial neural networks. Firstly, genetic programming is more robust and easy-to-use than artificial neural networks. This means that genetic programming simplifies the heavy burden involved in selecting correctly an adequate specification of the network. Secondly, unlike neural network approach, genetic programming provides explicitly a mathematical equation which is assumed to optimally describe the evolution of a time series.

In our study, we use a specific genetic programming programmed in FORTRAN following the computational routines explained in Álvarez et al. (2001). The evolutionary process starts creating a random initial population of N mathematical equations in the following way:

$$Equation_j: \ ((A\ Op\ B)Op\ (C\ Op\ D)) \quad \forall\ 1 \leq\ j \leq\ N \qquad (2)$$

where *A*, *B*, *C*, and *D* are the arguments (real numbers or lagged values of the time series),and *Op* represents one of the four arithmetic operators (sum, rest, multiplication and division). The subscript *j* refers to each one of the *N* equations belonging to the initial population. In a second step, the genetic program evaluates the fitness of each equation *j* of the initial population according to a specific criterion. In our study, we assume as fitness criterion the sum of the square errors

$$SSE_j = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \quad \forall \mathrm{j} = 1,..., \mathrm{N} \qquad (3)$$

where SSE is the sum of the square errors (SSE) presented by the equation j-th ($\forall\ 1 \leq\ j \leq\ N$), y_t represents the time series, y_t is the predicted value and T is the total number of observations reserved to evolve the genetic program. Those equations whose SSE is very high are dismissed while those with a low value are more likely to survive.

The third step is to randomly select some of the survival equations. These equations are used to create a new population of *N* equations by using the genetic operators: cloning, crossover, and mutation. With cloning, the "fittest" equations are included in the new group of equations without any modification. With crossover, pairs of equations exchange part of their arguments and their mathematical operators to create new equations. The mutation means that any operator or argument can be randomly replaced in a small number of equations.

This evolutionary process of evaluation, selection and creation of equations is iteratively repeated. After a number of iterations given by the user, the procedure ceases and the genetic program provides an optimal model which is the strongest mathematical equation of the last population.

Finally, the technical configuration of the genetic programming employed in this study was similar to that one described in Álvarez et al. (2001). A maximum number of 10,000 generations were considered. Each generation had a maximum population of 260 equations, and the maximum number of arguments and mathematical operator in each equation were 26. The crossover and mutation rates have been 0.2 and 0.1, respectively. The adequacy of this setup is guaranteed by previous work (Álvarez-Díaz, 2010; Álvarez-Díaz and Caballero, 2008; Álvarez-Díaz and Álvarez, 2003), and it was also confirmed afterwards by a sensitivity study.

3. Data description, sample division and assessment of forecasting performance

Our database consists of monthly data of the Consumer Price Index (CPI) of the United States, and they were taken from the Global Financial Database. The sample covers a period that goes from January 1980 to December 2013. Therefore, we have a total of 408 observations to carry out our predictive exercise. As usual in time series forecasting, and specifically in financial forecasting, the first difference of the price logarithm is assumed as our variable of interest. Therefore, given the original time series $\{Y_t\}_{t=1}^{T}$, the variable under study is

$$y_t = \log(Y_t) - \log(Y_{t-1}) \qquad (4)$$

where Y_t is the CPI value at time t, $\log(Y_t)$ is its logarithmic transformation and y_t is the variable after taking first differences. This data processing has become quite standard as it simplifies and improves the modeling procedure. The logarithm transformation is helpful to reduce the variability and asymmetry existing in the data distribution, and differencing is necessary to get a stationary time series, which is a statistical requirement to use the forecasting methods applied in this study. Moreover, the variable y_t is extremely interesting

since it can be interpreted as a proxy of the growth rate of the CPI, or in other words, the month-on-month inflation rate. Nevertheless, it is also recognised that this transformation can increase the existing noise in the series and, therefore, reduce our forecasting accuracy (Soofi and Cao, 1999).

[Figure 1]

[Table 1]

Figure 1(a) depicts the time evolution of the CPI over the sample period. The most remarkable characteristic is that the series shows a strong and smooth upward trend. Figure 1(b) shows the evolution of the transformed series, i.e., the inflation rate, where we can observe how the trend has been completely removed. Table 1 presents the summary statistics of the original data, and of the transformed data as indicated in expression (3). The sample partial autocorrelation coefficients (PACs) show a strong linear dependence at lags 1 and 12 for both series, and this characteristic is also corroborated by the Ljung-Box test. Additionally, the Jarque-Bera test leads to reject the null hypothesis that the data are normally distributed can be rejected at the 1% significance level. Additionally, the values of the Augmented Dickey-Fuller (ADF) and Phillips-Perron (P-P) unit root tests reveal that the CPI series is non-stationary at frequency zero, whereas the inflation rate is stationary[1].

In order to know more about the statistical properties of the CPI, we apply the Brock-Decker-Scheinkman (BDS) test for detecting nonlinear dependence in the data over our sample period. This test checks the null hypothesis that our observations are independent and identically distributed (*i.i.d.*, in short) and, as indicated by Brock et al. (1996), has high power

[1]This result is also corroborated by other unit root tests such as the Ng-Perron unit root test (Ng-Perron, 2001). Additionally, the HEGY test (Hylleberg, et al. 1990) does not detect the presence of unit roots at seasonal frequencies for any of the considered series.

against a wide range of linear and nonlinear alternatives. As we are interested in detecting non-linear patterns in the conditional mean, we apply an autoregressive model to remove the linearity on the data, and the BDS test is computed on the residuals of an autoregressive model[2]. However, this linear filter of the data is not enough since there can be a deterministic dependence in the conditional second moments, which is not useful to improve our predictive capacity.

[Table 2]

Table 2 shows the results of the ARCH test for the residuals of the autoregressive used to bleach the original data. According to these results, there seems to be a deterministic structure in the conditional variance. This characteristic can make the BDS test rejects the *i.i.d.* hypothesis, but this rejection would not imply the presence of non-linear structures in the mean. It is for this reason that we have applied the BDS test to the standardized residuals

$$Z_t = \frac{u_t}{\hat{h}_t} \qquad (5)$$

where Z_t is the standardized residuals which represent the data that are purged of linear dependences in mean and of nonlinear dependences in variance, u_t is the residuals of the autoregressive applied to bleach the original data and $\hat{h}_t^2$ is the conditional variance which is estimated using the GARCH(1,1) model

[2] The order of the autoregressive model was chosen so that minimized a generalization of the Akaike Information Criterion (AIC) where an additional term was included so as to penalize the use of extra coefficients that do not reduce significantly the error (Álvarez-Díaz et al., 2010; Álvarez-Díaz et al., 2014). The optimal order of the autoregressive was p=14, and the residuals did not exhibit significant autocorrelation.

$$\hat{h}_t^2 = \alpha_0 + \alpha_1 \cdot u_{t-1}^2 + \beta_1 \cdot h_{t-1}^2 \qquad (6)$$

Table 2 also reveals that the standardized residuals do not have any structure in variance. It seems, therefore, that the GARCH(1,1) model has captured adequately the nonlinear dependence in variance[3]. The BDS test is applied on these standardized residuals to discover nonlinear patterns in the conditional mean. Table 3 displays the values of the BDS test for different values of the embedding dimension (m) and different values of the distance (ϵ), the two technical parameters that must be arbitrarily chosen by the researcher (Kocenda, 2001) [4]. It is not evident from the results of the BDS test that the data reject the null hypothesis of *i.i.d.* There seems to have only a small evidence of nonlinearity in the data for the specific case of $\varepsilon = 0.5 \cdot \sigma$ and m = 8, 10, 11, 12, 13, 14 and 15. That is, it is possible the presence of a nonlinear seasonal pattern in the data, but it is not conclusive. The use of nonlinear forecasting methods such as the NAR neural networks and genetic programming would be adequate to extract this potential nonlinearity and, therefore, we would be capable of improving our forecasts of the CPI.

It is true that neural networks and genetic programming are extremely powerful to fit data and exploit the nonlinearities existing in the data, but this desirable characteristic can be also an important problem. These methods can tend to memorize the specific characteristics of individual observations rather than the general pattern in the data, which can lead to a low predictive ability when new data are considered (Zhang and Yu, 1998). In order to ensure the

[3] The choice of a GARCH(1,1) model is common in the literature to model the volatility of a financial time series. It is considered the simplest and most robust way to approximate the market volatility (Engle, 2001).

[4] Specifically, and following the recommendations given in Hsieh and LeBaron (1988), the distance parameter ϵ is chosen in a range between 0.5 and 1.5 of the data's standard deviation. The choice of the embedding dimension *(m)* depends on the lags that we wish to examine for serial dependence. Given that we work with monthly data, we have considered values of m from 2 to 15 in order to include delays around lag 12 (Mcnelis, 2005).

possibility of a good generalization of our non-linear forecasting and avoid overfitting problems, we follow the recommendation given in the literature to split the whole sample period into three different sub-periods (Bishop, 1995; Kaastra and Boyd, 1996; Kajitani et al., 2005; Binner et al., 2010). Figure 1 gives us information about the sample division that we have considered in our study. The first one sub-sample is the training/evolutionary period, and covers the period from January 1980 to December 2003, having a total of 286 observations. This subsample is used to design different NAR networks, and to get different equations through the evolutionary process generated by the genetic program. The second sub-period, the selection period, goes from January 2004 to September 2010, and it is used to select the NAR network and the evolutionary equation that perform best on these 81 observations. This sub-period plays a fundamental role because it allows us to discard overfitted models. The addition of training/evolution and selection sub-periods is usually known as the in-sample data set. The last subsample is the out-of-sample period and spans from October 2010 to December 2013. These 41 observations are completely new since they have not been used in the modeling process. In order to carry out a useful and fair predictive exercise, the accuracy showed in this sub-sample is the only valid to evaluate and compare the forecasting performance of the methods considered in our study[5].

Another important question that we have to define before starting our predictive study is how to assess and compare the predictive power of the different forecasting methods applied in our study. There are many possible measures available, but most of the handbooks recommend the use of the mean absolute percentage error (MAPE) to

[5] Even though there is no consensus on how to divide the whole a sample, the general practice is to assign more data to the evolutionary/training and validation samples (Zhang , 2004). Most studies split the sample according to a convenient percentage. In our case, we follow a common practice of allocating 70% of the data to the evolutionary/training period (286 observations), 20% to the validation period (81 observations), and the last 10% to the out-of-sample period (41 observations) (Yao and Tan, 2000).

assess the forecasting performance and make comparisons among different methods (Hyndman and Koehler, 2006). This measure calculates the forecasting error as a percentage of the actual value, and is computed from the following expression:

$$MAPE = \sum_{t=1}^{T} \left| \frac{Y_t - \hat{Y}_t}{\hat{Y}_t} \right| \cdot \frac{100}{T} \qquad (5)$$

where Y_t is the actual value of the US CPI, $\hat{Y}_t$ is the predicted value, and T is the sample size of the period in which the forecasting performance is evaluated. It is important to see that the predictions provided by the different forecasting methods ($\hat{y}_t$) are rescaled back following the reverse of the data transformation, and the forecasting accuracy is determined on the basis of the original scale of the CPI data (Y_t). The MAPE has several desirable advantages that make it very attractive to evaluate forecasting performance: (i) it is less sensitive to outliers distortion than other measures based on the squared errors (Yafee and McGee, 2000), (ii) is simple to calculate and easy to interpret (Amstrong, 1982), (iii) is independent of the scale of the variable being predicted, and this characteristic explains why this measure is frequently used to compare the forecasting capacity of different competing models and data sets (Hyndman and Koehler, 2006).

4. Forecasting study

Modeling setup and description of the benchmark model

As was already mentioned in the methodological section, the modeling procedure followed in our forecasting study is based on constructing multiple networks and evolutionary equations according to different parameters such as the number of lags (p), and also the number of activation functions (H) for the specific case of the NAR

network. The network and evolutionary equation finally chosen to correctly represent the dynamic of the CPI series will be those that best fit the data belonging to the selection period. However, as Zhang et al. (2001) point out, the most important parameter that must be chosen is the optimal number of model lags. In fact, this choice is the most critical decision when designing a forecasting model because too many lags could lead to overfitting problems, and too few could not provide the most accurate forecasts. It is for this reason that it is recommendable to examine the forecasting responsiveness of the non-linear methods to different lags. Figure 2 shows the MAPE of the best NAR network and genetic equation in the selection period for each one of the delays from 1 to 20. A simple look at this figure reveals a positive characteristic of the proposed nonlinear methods: both methods have certain predictive stability in the selection period regardless the number of lags that are considered. Following the recommendation given by Kaastra and Boyd (1996), we have chosen the number of lags that provided the lowest forecasting error for the NAR network, which was achieved at lag p=12, and the value for the activation function was equal to *H*=3. For the case of the genetic program, the value that minimized the MAPE was attained at lag *p*=13. The mathematical equation that best fit the CPI data in the selection period after the evolutionary process was

$$\widehat{\boldsymbol{y}}_{\boldsymbol{t}} = \mathbf{3} \cdot \boldsymbol{y}_{\boldsymbol{t-1}} + \left(\frac{\boldsymbol{y}_{\boldsymbol{t-11}}}{\boldsymbol{y}_{\boldsymbol{t-13}} + \mathbf{7.20}} - \boldsymbol{y}_{\boldsymbol{t-1}} \right) \cdot (\mathbf{10} \cdot \boldsymbol{y}_{\boldsymbol{t-12}} + \mathbf{2.46}) \qquad (\mathbf{6})$$

It is important to note that the fact of getting explicitly a mathematical equation is an advantage over other nonlinear methods such as artificial neural networks. The structure of the equation allows us to know more about the dynamics of the time series. Specifically, the equation defined in (6) reflects an apparently strong non-linear

behavior of the inflation rate, and shows the important influence of the previous values of the variable at lags 1, 12 and 13. The influence of lags 12 and 13 is indicative of a possible non-linear seasonal pattern, which is apparently well caught by the genetic program.

[Figure 2]

The choice of the optimal models has been done on the basis of maximizing the forecasting performance in the selection period. However, this is not the only criterion that must be met to ensure a good out-of-sample forecasting performance of the models calibrated in our study. It is also necessary to guarantee that the errors of the models finally chosen do not show significant signs of autocorrelation; otherwise, the forecasts would be sub-optimal (Granger and Newbold, 1974). If the errors were correlated, then the network and/or the genetic program would not be adequately designed, and the modeling process should be done again. Figure 3 depicts the in-sample autocorrelation function of the residuals of the models, as well as their respective empirical intervals constructed by means of the surrogate method with a level of confidence of 99 percent (Theiler et al., 1992). From these figures, it can be inferred that the errors do not have any strong systematic pattern since none of the sample autocorrelation coefficients is significantly different from zero at 5 percent level[6]. Therefore, the lack of autocorrelation allows us to assert that the NAR network and the evolutionary equation are optimal to make out-of-sample forecasts, at least from a linear point of view.

[6] The lack of autocorrelation is also corroborated by the Ljung-Box test. The results of this test have not been reported to save space, but the details are available upon request from the authors.

A rigorous non-linear forecasting study also requires a comparison with some linear statistical model. It is for this reason that we have estimated an autoregressive integrated moving average (SARIMA) model. As indicated by Zhang et al. (2008), this kind of linear model is one of the most popular in seasonal time series forecasting, and is very often used as a benchmark for predictive comparison. The explanation for this popularity is that SARIMA models are relatively simple, ease of use and they have showed a good accuracy in predicting economic and financial time series when monthly or quarterly data are used. The model building process followed in our study to get an optimal specification of the SARIMA was based on the construction of a collection of models with different non-seasonal and seasonal orders. The final specification of the model that was finally chosen to make out-of-sample forecasts was the one that minimized the Bayesian information criterion (BIC) in the in-sample period (Shen et al., 2008)[7]. According to this criterion, the optimum model was the $SARIMA(3,1,1)x(2,1,2)_{12}$ whose mathematical estimated expression is

$$(1 - 1.08 \cdot B + 0.43 \cdot B^2 - 0.06 \cdot B^3) \cdot (1 - 1.67 \cdot B^{12} + 0.67 \cdot B^{24}) \cdot \nabla \log(Y_t) =$$
$$= (1 + 0.51 \cdot B) \cdot (1 + 1.73 \cdot B^{12} - 0.78 \cdot B^{24}) \cdot \varepsilon_t \qquad (6)$$

where B is the backshift (lag) operator, ∇ is the first-difference operator, $\log(Y_t)$ is the logarithm of the CPI, and ε_t is the error term. Figure 3 also displays the autocorrelation function for the residuals of the estimated SARIMA model. This figure puts on view that these residuals are not significantly related in time and, therefore, we can say that

[7] The BIC is strongly recommended to select the best model from candidate models having different numbers of parameters. Nevertheless, in our case, the specification of the SARIMA model would not change if we had chosen other criteria such as the Akaike information criterion (AIC). Moreover, the selected SARIMA specifications were also those that presented the lowest MAPE in the in-sample period.

the forecasts provided by the estimated SARIMA model are a good proxy of the best linear forecasts (Theiler et al., 1993).

Out-of-sample forecasting assessment

Up to this point we have verified that the technical setting of the different forecasting methods proposed in our study is adequate to predict the US CPI; therefore, it seems that all of them can be used to make accurate out-of-sample predictions. Figure 4 shows graphically the out-of-sample forecasts of each one of the methods. This is an obvious and simple plot that allows us to examine very roughly the forecasting ability of the proposed models. At first glance, it draws attention the fact that the forecasts of the different methods are rather similar. They are all seemingly accurate since the forecasts follow the actual values of the time series quite closely. Although this figure can be very illustrative, it is not very precise in assessing the predictive differences between competing models (Kajitani et al., 2005)

[Figure 4]

Table 4 gives us much more precise information about the forecasting performance of the competing methods. The out-of-sample value of MAPE was 0.2069 percent for the evolutionary equation, and 0.2098 percent for the network. These low values are a clear sign that the difference between the predicted and actual values is very small. According to the scale proposed by Lewis (1982), the CPI forecasts of these non-linear methods can be deemed as highly accurate. However, as we can also see in this table, the SARIMA model yields an out-of-sample MAPE of 0.1925 percent, which is even lowest than those obtained by the nonlinear methods. Therefore, it seems that the linear SARIMA is able to beat the forecasts of much more complex non-linear

methods such as an NAR network and the equation obtained by means of a genetic program.

[Table 5]

The forecasting comparison done until now is not entirely complete since we must also explore if the predictive gain showed by the SARIMA model is statistically significant. That is, we must check if the SARIMA model outperforms the forecasts of the non-linear methods from a statistical point of view. To do so, we apply the test statistic proposed by Diebold and Mariano (1995) (DM, hereinafter) which has been widely used in time series forecasting to check the null hypothesis that two competing methods have the same predictive capacity. In spite of its widespread use, this test has the problem that its asymptotic distribution could be inaccurate for small samples (Diebold and Mariano, 1995), as happens in our case. To overcome this problem and be more accurate in the comparison between competing models, we applied the procedure followed in Ferson *et al.* (2013). Specifically, these authors recommend estimating empirically the p-values associated to the DM test by using the bootstrap technique[8]. Table 5 shows the values of the DM test for the different comparisons among methods, and their respective confidence intervals and *p*-values using the bootstrap method. The most remarkable result from this table is that we cannot reject the null hypothesis of

[8] The procedure followed in our study for the estimation of the *p*-values begins with the construction of 10,000 artificial samples by re-sampling with replacement of the original data. The DM test is calculated for each one of these artificial samples. These 10,000 values of the test are used to estimate an empirical probability density function using the Kernel method (Bowman, 1997). The estimated *p*-value of the DM test is the probability that this test leaves in the tails of the empirical distribution. We have also applied the test proposed by Harvey et al. (1997) that implies a small-sample modification of the Diebold-Mariano test. The values of the modified BDS test do not modify substantially the results reported in our study using the bootstrapped p-values.

equal predictive accuracy among methods. That is, the predictive gain showed by the SARIMA is not statistically significant, concluding that all methods considered in our study have the same accuracy to predict the CPI time series. The value showed by the bootstrapped *p*-value of the DM test for the comparison between the SARIMA model and the network (0.55) and the genetic equation (0.56) reflects that both the linear method and the non-linear methods have practically the same out-of-sample forecasting capacity. It is also important to note that the high value of the *p*-value of the DM test for the predictive comparison between the evolutionary equation and the NAR network (0.9793) is indicative that both nonlinear methods offer the same out-of-sample accuracy. That is, genetic programming provides at least as good forecasts as neural networks, but it must be emphasized that genetic programming is less time consuming and easier to design than neural networks. Therefore, at least for our case, our results give a clear argument in favor of considering the use of genetic programming in forecasting the CPI time series.

5. Discussion

Our forecasting results provide evidences that support the general belief that non-linear methods, and specifically neural networks, do not perform significantly better than linear models when the goal is to predict macroeconomic data (Swanson and White, 1997). However, it is also true that most of previous research on forecasting inflation in the United States have reported that nonlinear methods, and particularly artificial neural networks, usually outperform linear models. Regarding this point, Nakamura (2005) and Binner et al. (2006) found that neural networks did comparatively better than autoregressive models at predicting US inflation by using quarterly data. McAdams and

McNelis (2005) applied neural networks to forecast monthly inflation rates in the US, Japan and the euro area. These authors concluded that neural networks perform well relative to the linear benchmark, but they also recognize that the network approximation is not the only alternative or the best among a variety of alternatives. Binner et al. (2010) also employed neural networks to predict monthly US inflation, they confirmed the predictive superiority of the network against the naive random walk model. Choudhary and Haider (2012) explored the power of neural networks to predict monthly inflation rates for some countries belonging to the Organization for Economic Cooperation and Development (OECD), where the US case is included. These authors came to the conclusion that, in general, neural networks showed a higher predictive capacity compared to a simple autoregressive model of order one. However, this simple autoregressive model was able to perform better for some of the countries considered in the study.

Our findings differ from those obtained in the studies mentioned above. There are some possible reasons that explain this discrepancy:

(a) The main advantage of genetic programming and neural networks is that they are excellent forecasting methods when data are nonlinear and show complex, or even chaotic, dynamics. That is, these methods are good at extracting nonlinear deterministic patterns existing in the series in order to improve forecasts. However, in our case, there is no a strong evidence that the monthly series of the US CPI is nonlinear. At least, this conclusion can be inferred from the results of the BDS test. If the series does not show a clear nonlinear pattern, then there is no room for a predictive improvement by using nonlinear methods. Moreover, the predictive dominance showed by the SARIMA model

in our study would be consistent with the idea that linear models provide better forecasts of linear data than neural networks (Kuan and White, 1994). Additionally, it is important to underline here that the optimal equation found by the genetic programing, as represented in (6), shows a nonlinear structure. However, it is possible that the output of this equation is close to being linear.

(b) It is very important to underline that many of the above studies could have failed to select the most accurate linear benchmark model to make forecasting comparisons. It does not seem appropriate to compare a nonlinear model with a simple autoregressive model or with a naïve random walk. Moreover, many times the order of the autoregressive is not the most adequate (for example, it does not seem appropriate to choose an autoregressive of order one (AR(1)) when the data are on a monthly basis). These models are so simple that they should not be used as benchmark. Table 6 shows the forecasting results in terms of MAPE of these simple benchmark models: the random walk model, the autoregressive of order one (AR(1)), and the best autoregressive model according to the BIC criterion which was an autoregressive of order 14 (AR(14)). The comparisons between these simple linear models and the nonlinear methods are represented in Table 7 where the values of the DM test and its associated bootstrapped *p*-value and confidence intervals are showed. As we can see, if we had considered only the random walk model and the AR(1) model as benchmarks we would have observed that our nonlinear methods would have significantly outperformed the forecasts of the linear models at the standard level of significance. The conclusion would have been quite different. That is, we would have concluded that the nonlinear forecasting methods work better than traditional linear alternatives. However, the residuals of the random walk and the AR(1) are correlated, which implies that they

are not good proxies of the optimum linear predictor. In other words, they are not the best linear forecasting models that can be constructed and, consequently, they should not be used as benchmarks. On the other hand, we can also observe that an AR(14) model is able to provide as good forecasts as the nonlinear methods considered in our study. Indeed, the AR(14) and the SARIMA models present similar forecasts, and their residuals do not show serial dependence. These findings suggest that both models are getting close to the best linear predictor, and they are good candidates for predictive comparisons.

[Table 6]

[Table 7]

(c) Finally, as Chatfield (1995) and Faraway and Chatfield (1998) pointed out, there can be an important bias in the studies already published since journals are more likely to accept results in favor of a new forecasting methodology than the contrary. This publication bias against non-significant results could explain the difficulties of finding studies that have failed in forecasting US CPI by using nonlinear methods and, specifically, neural networks. However, the possible publication bias in the case of the US CPI is not observed as it has been for the prediction of other economic and financial time series since there are some studies that have found that ANNs are not better than traditional linear forecasting methods[9].

[9] See Zhang et al. (1998) for an excellent literature review of the success and failures of applying ANNs to predict economic and financial time series.

6. Conclusion

Importance of accurate inflation forecasts is well-recognized, as it affect decision making of not only the central banks, but also various other economic agents in the economy. Further, it is well-documented in the literature that that there have been tremendous changes in inflation dynamics in the US, to the extent that inflation forecasters seem to have a dearth of reliable multivariate models for forecasting inflation. Given this, the objective of this paper is to predict, both in-sample and out-of-sample, the consumer price index (CPI) of the US economy based on monthly data covering the period of 1980:1-2013:12, using powerful and novel nonlinear forecasting methods: neural networks and genetic programming. The forecasts of these nonlinear methods are compared with a variety of linear models that are usually employed as benchmark: the random walk model, the autoregressive model (the simple AR(1) and the optimal AR(14)), and the SARIMA model.

Our results support the general belief that nonlinear methods fail to predict macroeconomic time series. Specifically, the SARIMA model and the AR(14) yield as good forecasts as the nonlinear methods. Indeed, the SARIMA model gives the most accurate out-of-sample forecasts, but this forecasting gain is not statistically significant relative to the nonlinear methods and the AR(14). On the other hand, the nonlinear methods makes predictions that are significantly better than those produced by simple linear methods such as the random walk and the AR(1). This fact could explain the predictive success achieved by previous studies that demonstrated the forecasting superiority of ANNs over these traditional linear models. However, the random walk and the AR(1) models are so simple that they cannot be considered as good proxies of the best linear predictor and, for this reason, they should not be used as benchmarks.

Additionally, the similarities in the forecasts provided by the nonlinear methods and the best linear predictors (SARIMA and AR(14) seems to indicate the lack of nonlinear patterns in the dynamics of the US CPI series. This observation is also corroborated by the fact that the BDS test did not find strong evidences of nonlinearity in the data. In the absence of nonlinear patterns, neural networks and genetic programming cannot be used to exploit their full predictive potential, and the best that they can do is to provide forecasts close to the best linear model, as it is in our case.

References

Aiken M. (1999). Using a neural network to forecast inflation, *Journal of Industrial Management & Data Systems*, 99, 7, 296–301

Allen F. and Karjalainen R. (1999). Using Genetic Algorithms to Find Technical Trading Rules, *Journal of Financial Economics*, 51, 245-271.

Álvarez A., López C., Riera M., Hernández E. and Tintoré J. (2000). Forecasting the SST space-time variability of the Alboran Sea with genetic algorithms, *Geophysical Research Letters*, 27, 2709-2712.

Álvarez A., Orfila A. and Tintoré J. (2001). DARWIN- an evolutionary program for nonlinear modeling of chaotic time series, *Computer Physics Communications*, 136, 334-349.

Álvarez-Díaz M. (2010). Speculative Strategies in the Foreign Exchange Market Based on Genetic Programming Predictions, *Applied Financial Economics*. 20, 6, 465-476.

Álvarez-Díaz M. and Caballero G. (2008). The quality of Institutions: a genetic programming approach. *Economic Modelling*, 25, 1, 161-169.

Álvarez-Díaz M., Hammoudeh S. and Gupta R. (2014). Detecting predictable non-linear dynamics in Dow Jones Islamic Market and Dow Jones Industrial Average indices using nonparametric regressions, *The North American Journal of Economics and Finance*, 29, 22-35.

Álvarez-Díaz, M. and Álvarez A. (2003). Forecasting exchange rates using genetic algorithms, *Applied Economics Letters*, 10, 6, 319-322.

Álvarez-Díaz, M. and Álvarez A. (2005). Genetic multi-model composite forecast for non-linear forecasting of exchange rates, *Empirical Economics*, 30, 643-663.

Álvarez-Díaz, M., Otero-Giráldez, M.S. and González-Gómez, M. (2010). Statistical Relationships between the North Atlantic Oscillation and International Tourism Demand in the Balearic Islands, Spain. *Climate Research*, 43, 207-214.

Amstrong J. S. (1982) Relative Accuracy of Judgmental and Extrapolative Methods in Forecasting Annual Earnings, *Journal of Forecasting*, 2, 437-447.

Beenstock M. and G. Szpiro, 2002, Specification search in nonlinear time-series models using the genetic algorithm, *Journal of Economic Dynamics & Control*, 26, 811-835.

Benati, L., and Surico, P. (2008). Evolving U.S. Monetary Policy and The Decline of Inflation Predictability, *Journal of the European Economic Association*, 6, 2-3, 634-646, 04-05.

Binner J. M, Bissoondeeal R. K., Elger T., Gazely A. M. and Mullineux A. W. (2005). A comparison of linear forecasting models and neural networks: an application to Euro inflation and Euro Divisia, *Applied Economics*, 37, 6, 665-680.

Binner J. M., Thomas Elger T., Nilsson B. and Tepper J. (2006). Predictable non-linearities in US inflation, *Economic Letters*, 93, 3, 323-328.

Binner J. M., Tino P., Tepper J., Anderson R., Jones B. and Kendal G. (2010). Does Money Matter in Inflation Forecasting?, *Physica A*, 389, 4793 – 4808.

Bishop C. M. (1995) *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Bowman A. W., and Azzalini A. (1997). Applied Smoothing Techniques for Data Analysis. New York: Oxford University Press.

Brock, W., Dechert D, Scheinkman J. *and* LeBaron B. *(*1996). A test for independence based on the correlation dimension. *Econometric Reviews,* **15**, 197–235

Chatfield C. (1995). Positive or negative?, *International Journal of Forecasting*, 11, 4, 501-502.

Choudhary and Haider (2012) Choudhary, A. and Haider, A., 2012. Neural network models for inflation forecasting: an appraisal, *Applied Economics*, 44, 2631–2635.

Cybenko G. (1989) Approximation by superposition of a sigmoidal function, *Mathematics of Control, Signals and Systems*, 2, 303-314.

Diebold F. X. and Mariano R. S. (1994) Comparing predictive accuracy, Journal of Business and Economic Statistics, 3, 253-263.

Engle R. (2001). GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics, *Journal of Economic Perspectives,* 15, 4, 157–168

Faraway, J., Chatfield, C., (1998). Time series forecasting with neural networks: a comparative study using the airline data, *Applied Statistics,* 47, 2, 231–250.

Faust, J. and Wright, J.H. (2013). “Inflation Forecasting" in Handbook of Economic Forecasting, 2, Graham Elliott and Allan Timmermann (eds.), 2-56, Elsevier, Amsterdam, The Netherlands.

Ferson W., Nallareddy S. and Xie B. (2013). The "out-of-sample" performance of long run risk models. *Journal of Financial Economics*, 107, 3, 537-556.

Fyfe C., Marney J. P. and Tarbert H. F. E. (1999). Technical analysis versus market efficiency- a genetic programming approach, *Applied Financial Economics*, 9, 183-191.

Goldberg, D.E., 1989. Genetic algorithms in search, optimization, and machine learning. Adison-Wesley.

Granger C.W.J., and Newbold P. (1974). Spurious regressions in econometrics, *Journal of Econometrics*, 2, 111-120

Harvey D. I., Leybourne, S. J., and Newbold, P (1997). Testingth e Equality of Prediction Mean Squared Errors, *International Journal of Forecasting*, 13, 281-291.

Holland J. H. (1975). Adaptation in natural and artificial systems, (Ann Arbor, The University of Michigan Press).

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359-366.

Hsieh, D. A., and LeBaron, B. (1988). Small sample properties of the BDS statistic, in W. A. Brock, D. Hsieh and B. LeBaron (eds.), *Nonlinear Dynamics, Chaos, and Stability*. Cambridge, MA: MIT Press.

Hylleberg S., R. Engle, Granger C. and Yoo B. (1990). Seasonal integration and cointegration, *Journal of Econometrics*, 44, 215–238.

Hyndman R. J. and Koehler A. B. (2006). Another look at measures of forecast accuracy, *International Journal of Forecasting*, 22, 679–688.

Ireland, P. N., (2004). "A Method for Taking Models to the Data," Journal of Economic Dynamics and Control, 28, 1205-1226.

Kaastra I., and Boyd M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, 10, 215-236]

Kaboudan M. A., (2000). Genetic programming prediction of stock prices, Computational Economics 16, 207-236.

Katijani Y., Hipel W. K. and Mcleod A. I. (2005). Forecasting nonlinear time series with feedforward neural networks: A case study of Canadian lynx data, *Journal of Forecasting*, 24, 105-117.

Kocenda, E. (2001). An Alternative to the *BDS* Test: Integration across the Correlation Integral, *Econometric Reviews,* 20, 3, 337-351.

Koop, G. and Korobilis, D. (2012). "Forecasting Inflation Using Dynamic Model Averaging", International Economic Review, 53, pp. 867-886.

Koza J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection*, The MIT Press, Cambridge.

Koza J. R. (1995). Genetic Programming for Economic Modeling, 251-269, in S. Goonatilake and P. Treleaven (eds.): *Intelligent Systems and Business*. John Wiley & Sons.

Kuan C. M. and White H. (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews*, 13, 1, 1-91.

Lewis C. D. (1982). Industrial and business forecasting methods, Butterworths, London.

market. New York: Elsevier Inc.

McNelis, P. D. (2005). Neural networks in financial: Gaining predictive edge in the

McNelis, P. D. and McAdam, P. (2005). Forecasting inflation with thick models and neural networks, *Economic Modeling*, 22, 548–67.

Nakamura, E. (2006). Inflation forecasting using a neural network, *Economics Letter*, 86, 373–8.

Neely C.J., Weller, P.A., Dittmar, R., 1997. Is technical analysis in the foreign exchange market profitable? A genetic programming approach, *Journal of Financial and Quantitative Analysis*, 32, 4, 405–426.

Ng, S. and Perron, P. (2001). Lag length selection and the construction of unit root tests with good size and power. *Econometrica*, 69, 6, 1519-1554.

Rumelhart D. E. and McClelland J. L. (1986). Parallel distributed processing: Explorations in the micro-structure of Cognition, Vol. 1, 318-362. The MIT Press.

Shen S., Li G. and Song H. (2008) An Assessment of Combining Tourism Demand Forecasts over Different Time Horizons, *Journal of Travel Research* 2008 47: 197.

Soofi A. S. and L. Cao (1999) Nonlinear deterministic Forecasting of daily Peseta-Dollar Exchange Rate, *Economic Letters*, 62, 175-178.

Stock, J .H. and Watson, M.W. (2010). "Modeling Inflation After the Crisis", FRB Kansas City symposium, Jackson Hole, Wyoming, August 26-28, 2010.

Stock, J.H. and Watson, M.W. (2009) Phillips curve inflation forecasts, in "Understanding inflation and the implications for monetary policy", Jeffrey Fuhrer, Yolanda Kodrzycki, Jane Little and Giovanni Olivei (eds.), MIT Press, Cambride.

Swanson, N. R. and White, H. (1997) A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks, *The Review of Economics and Statistics*, 79, 540–50.

Szpiro G. (1997a). Forecasting chaotic time series with genetic algorithm, *Physical Review* E, 55, 3, 2557-2568.

Szpiro G. (1997b). A search for hidden relationships: data mining with genetic algorithms, *Computational Economics*, 10, 267-277.

Theiler J. and Eubank S. (1993). Don't bleach chaotic data, *Chaos*, 3, 771-782.

Theiler J., Eubank S., Longtin A. & Galdrikian B. (1992) Testing for Nonlinearity in Time Series: the Method of Surrogate Data, *Physica D*, 58, pp. 77-94.

White H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings, *Neural Networks*, 3, 535-550.

Wong B. K., Vincent S. L. and Lam J. (2000). A bibliography of neural network business applications research: 1994-1998, *Computers & Operations Research*, 27, 1045-1076

Wong, B. K., Bodnovich T. A. and Selvi Y. (1995). A bibliography of neural network business application research: 1988-September 1994, *Expert Systems* 12, 3, 253-261.

Yadavalli, V. K.; Dahule, R. K.; Tambe, S. S.; Kulkarni, B. D. (1999). Obtaining Functional Form for Chaotic Time Series Evolution Using Genetic Algorithm. *Chaos*, *9*, 789-794.

Yaffee, R. A., and McGee, M. (2000). Introduction to time series analysis and forecasting. San Diego: Academic Press.

Yao J., Y. Li and C. L. Tan (1997) Forecasting the exchange rates of CHF vs USD using neural networks, Journal of Computational Intelligence in Finance, 5, 2, 7-13.

Yao, J. and C. L. Tan (2000) A case study on using neural networks to perform technical forecasting of Forex, Neurocomputing 34, 79-98.

Zhang G. P. and Kline D. (2007). Quarterly time-series forecasting with neural networks, *IEEE Trans Neural Networks*, 18, 6, 1800–1814.

Zhang G. P. and M. Y. Hu (1998) Neural network forecasting of the British Pound/US Dollar exchange rate. Omega, 26, 4, 495-506.

Zhang G.P. (2004). Neural Networks in Business Forecasting, London: IRM Press.

Zhang G.P., Patuwo E.B., Hu M.Y. (1998). Forecasting with artificial neural networks: The state of the art, *International Journal of Forecasting, 14, 35-62*.

Zhang G.P., Patuwo E.B., Hu M.Y. (2001). A simulation study of artificial neural networks for nonlinear time-series forecasting, *Computers & Operation Research*, 28, 381–396.

Figure 1. Time evolution of the original and the transformed time series

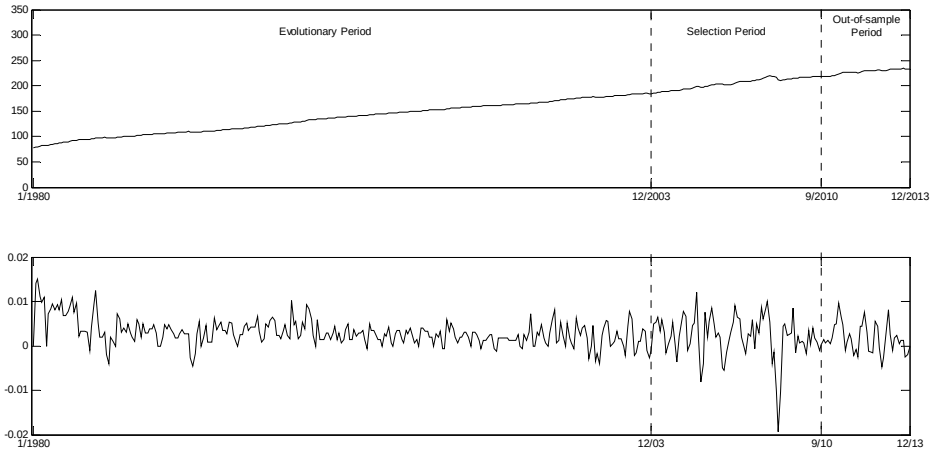

Figure 2. Selection of the optimal number of delays.

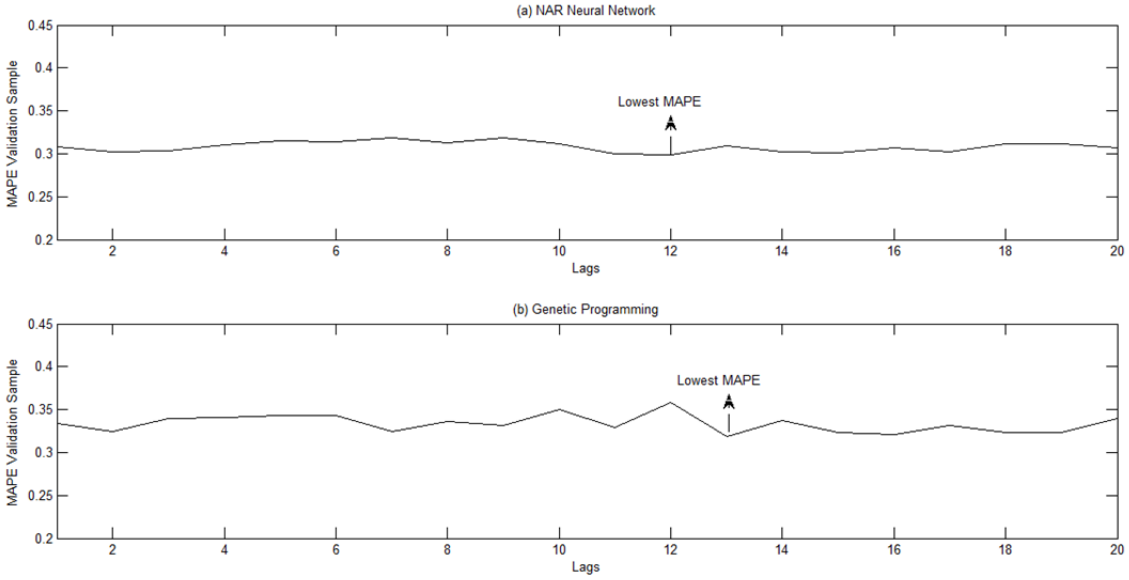

Figure 3. Autocorrelation Errors (Validation Period).

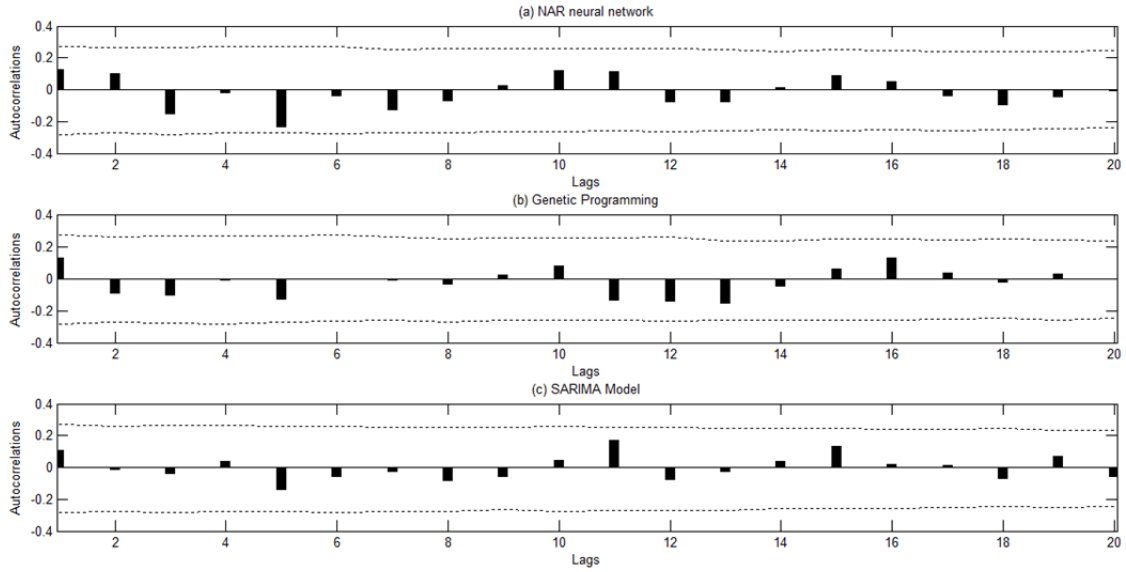

Figure 4. Out-of-Sample Forecast of the US CPI.

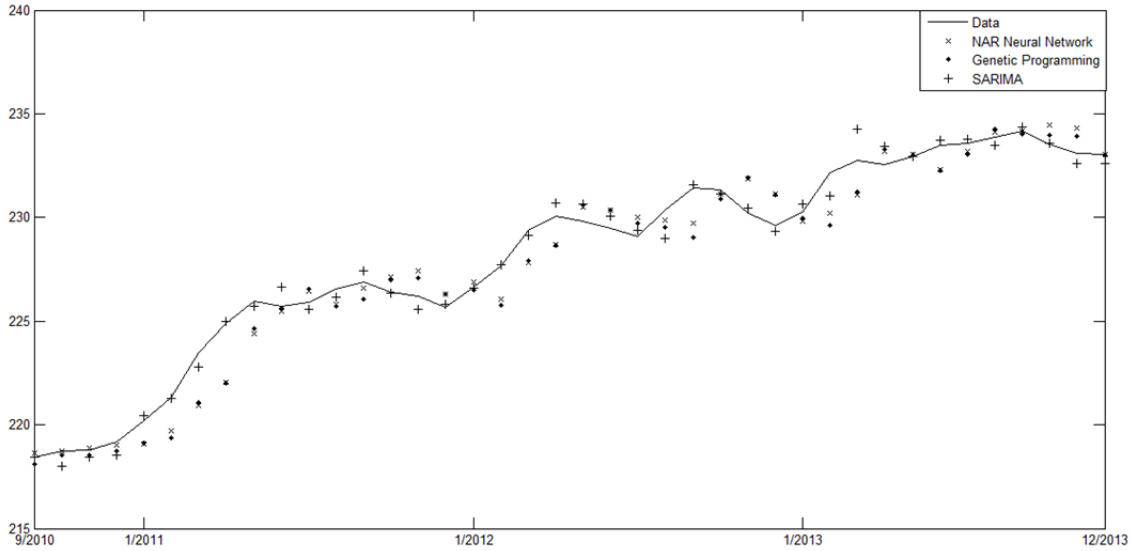

Table 1. Summary statistics for the original data (CPI) and the transformed data (log first difference)

Description	**Original data (Y_t)**	**Transformed data (y_t)**
Sample Size	408	407
Mean	158.44	0.003
Median	158.85	0.003
S.D.	44.00	0.003
Skewness	0.04	-0.49
Kurtosis	1.84	7.84
Maximum	234.11	0.015
Minimum	77.84	-0.019
PAC(1)	0.99*	0.53*
PAC(12)	-0.13*	0.11**
Q(1)	404.6*	116.20*
Q(12)	4508.7*	171.53*
Jarque-Bera	2.74*	414.34*
ADF Test	-0.65	-11.55
P-P Test	-0.86	-12.07

Note. The symbols *, ** and *** represent the rejection of the null hypothesis. PAC(p) is the partial autocorrelation coefficient at lag p, and Q(p) is the value of the Ljung-Boxt statistic at lag p. The MacKinnon critical values for ADF and P-P test are -3.44, -2.86, and -2.57 at 1%, 5% and 10% significance level, respectively.

Table 2: The ARCH Test for the Original (r_t) and the Standardized Residuals (z_t)

Lags	ARCH TEST	
	Residuals of the CPI (r_t)	**Standardized Residuals of the CPI (z_t)**
1	34.06***	0.06
2	34.80***	0.31
3	36.21***	0.32
4	36.19**	0.61
5	39.53**	0.65
6	39.75*	1.57
7	44.06**	2.26
8	46.83**	2.81
9	47.26**	2.84
10	47.43*	2.83
11	51.44**	5.19
12	51.77*	5.67
13	52.85*	5.73
14	53.74	5.79
15	56.66*	6.97

Note: The symbols *, ** and *** represent the rejection of the null hypothesis $H_0: residuals\ are\ i.i.d$ at the 10, 5 and 1 percent significance levels, against the alternative hypothesis that the there is a linear dependence in variance.

Table 3: The BDS Results for the Standardized Residuals of the CPI (z_t)

Embedding Dimension (m)	Distance $\varepsilon = 0.5 \cdot \sigma$	$\varepsilon = 1 \cdot \sigma$	$\varepsilon = 1.5 \cdot \sigma$
2	-1.40	-0.85	-0.59
3	-0.09	0.33	0.37
4	-0.43	0.56	0.64
5	0.17	0.35	0.34
6	1.77	0.12	0.30
7	1.45	-0.07	0.13
8	3.40***	-0.46	-0.17
9	2.69	-0.57	-0.18
10	-3.24***	-0.69	-0.33
11	-2.74***	-0.74	-0.37
12	-2.35***	-0.29	-0.29
13	-2.15***	-0.02	-0.39
14	-1.83**	-0.25	-0.25
15	-1.65*	0.40	-0.32

Note: The symbols *, ** and *** represent the rejection of the null hypothesis H_0: *no nonlinearity* at the 10, 5 and 1 percent significance levels, respectively. The BDS is implemented assuming the distance ε as a fraction of the standard deviation of the data (σ).

Table 4. Forecasting results of the nonlinear methods and the SARIMA model.

Mean Absolute Percentage Error (%)						
Genetic Programming			**NAR Neural Network**		**SARIMA (2,1,2)x(0,1,2)**	
In-Sample Period		**Out-of-sample Period**	**In-Sample Period**	**Out-of-sample Period**	**In-Sample Period**	**Out-of-sample Period**
0.2073		0.2069	0.1895	0.2098	0.1715	0.1925
Training	**Validation**		**Training**	**Validation**		
0.1740	0.3184		0.1571	0.2984		

Table 5. Results of the Diebold-Mariano test for the comparison between nonlinear methods and the SARIMA model.

	DM (bootstrapped p-value)	Bootstrap Confidence Interval
NAR Neural Network vs. SARIMA	0.59 (0.55)	(-1.18, 2.28)
NAR Neural Network vs. Genetic Programming	0.02 (0.98)	(-1.75, 1.77)
Genetic Programming vs. SARIMA	0.57 (0.56)	(-1.17, 2.26)

Note. The bootstrapped p-value and the confidence interval are constructed using the accelerated bias-corrected method (Bca). 10,000 replications were considered, and the level of confidence for the interval was at 95 percent.

Table 6. Forecasting results of simple benchmark models.

Mean Absolute Percentage Error (%)					
Random Walk		**AR(1)**		**AR(14)**	
In-Sample Period	**Out-of-sample Period**	**In-Sample Period**	**Out-of-sample Period**	**In-Sample Period**	**Out-of-sample Period**
0.3498	0.2843	0.2472	0.2694	0.1929	0.2093

Table 7. Results of the Diebold-Mariano test for the comparison between the nonlinear methods and the naïve benchmark models.

	DM (bootstrapped p-value)	Bootstrap Confidence Interval
NAR Neural Network vs. Random Walk	-1.88 (0.07)	(-3.67, -0.02)
NAR Neural Network vs. AR(1)	-2.15 (0.04)	(-4.75, -0.57)
NAR Neural Network vs. AR(14)	0.42 (0.66)	(-1.52, 2.70)
Genetic Programming vs. Random Walk	-2.00 (0.05)	(-3.79, -0.02)
Genetic Programming vs. AR(1)	-1.91 (0.06)	(-4.38, -0.12)
Genetic Programming vs. AR(14)	0.40 (0.69)	(-1.52, 2.70)
SARIMA vs. Random Walk	-1.82 (0.07)	(-3.44, 0.69)
SARIMA vs. AR(1)	-1.95 (0.05)	(-3.92, 0.54)
SARIMA vs. AR(14)	-0.34 (0.73)	(-2.32, 1.83)

Note. The bootstrapped p-value and the confidence interval are constructed using the accelerated bias-corrected method (Bca). 10,000 replications were considered, and the level of confidence for the interval was at 95 percent.