Open Journal of Statistics, **2012, 2, 237-250**
http://dx.doi.org/10.4236/ojs.2012.23029 Published Online July 2012 (http://www.SciRP.org/journal/ojs)

Comparisons of VAR Model and Models Created by Genetic Programming in Consumer Price Index Prediction in Vietnam

Pham Van Khanh
Military Technical Academy, Hanoi, Vietnam
Email: van_khanh1178@yahoo.com

Received May 30, 2012; revised June 30, 2012; accepted July 10, 2012

ABSTRACT

In this paper, we present an application of Genetic Programming (GP) to Vietnamese CPI inflation one-step prediction problem. This is a new approach in building a good forecasting model, and then applying inflation forecasts in Vietnam in current stage. The study introduces the within-sample and the out-of-samples one-step-ahead forecast errors which have positive correlation and approximate to a linear function with positive slope in prediction models by GP. We also build Vector Autoregression (VAR) model to forecast CPI in quaterly data and compare with the models created by GP. The experimental results show that the Genetic Programming can produce the prediction models having better accuracy than Vector Autoregression models. We have no relavant variables (m2, ex) of monthly data in the VAR model, so no prediction results exist to compare with models created by GP and we just forecast CPI basing on models of GP with previous data of CPI.

Keywords: Vector Autoregression; Genetic Programming; CPI Inflation; Forecast

1. Introduction

Inflation has great importance to saving decisions, investment, interest rate, production and consumption. Decisions basing on impractical inflation predictions result in uneffective resource allocation and weaker macroeconomic activities. Meanwhile, better predictions see better forecast solutions given by economic agents and improve the entire economic performance.

Different models depending on the theory of different price fixation are often used for describing inflation evolutions. These models emphasize the role of different variables in inflation. The different econometric models have different modeling specification, and information quality. Despite of huge explained variables in models basing on theory to improve the level of conformity, they uncertainly ameliorate the ability of prediction models.

One can see many different models in different countries, particularly: the model of Phillips curve with added expecting elements, traditionally monetary model, price equations basing on monetary demand viewpoint.

Apart from theoretic inflation ones above, it can be seen the variable time series models which are used for inflation data in the past so as to forecast further inflation and give no more explaination to analyze. Recently, the multivariate time series model and its variations have appeared in nonlinear time string, especially in the smooth transition regression. Nevertheless, with the content of the paper, we just present models relevant to this study without deeply analyzing their theoretic base.

2. Methodology

We center on considering the VAR model and applications of Genetic Programming to forecast the inflation index CPI. Initially, estimate and accreditation to be good models, and then predictions will be seen based on variables taken from models.

2.1. Vector Autoregression Model

The vector autoregression (VAR) model is one of the most successful, flexible, and easy way to use models for the analysis of multivariate time series [1].

Before 1980s, equation models were simultaneously used for analyzing and forecasting macro-economic variables as well as the study of the economic cycle. At that time, econometric were dedicated to issue of the format of the model-relating to properties of endogenous variables in the model.

The formatting in simultaneous equation models is primarily through assumptions of the interactions between variables. The assumption is usually based on economic theory or visual knowledge of model, determining the presence or absence of the variables in each equation.

Sim [2] has changed the concerns of contemporary economist community. He said that most of the economic variables, especially macroeconomic variables are endogenous. On the other hands, they are interactive. Therefore, he proposed a multivariable model with endogenous variables having the same role. Nowadays, the VAR model has become a powerful tool and was used extensively (especially in macro-economic problems), predictions (particularly the medium-term and long-term ones), and the analysis of shock transmission mechanism (considering the impact of a shock on a dependent variable on other dependent variables in the system).

When presenting the VAR model, one can introduce it structurally and then contractionally. Nonetheless, for the prediction, we can use the information from the resulting estimates of the shortened model, so it is better to solely present the contracted model serving to experimental analysis without the detail structure model to avoid unnecessary complexities.

A basic VAR contraction forms:

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + B_0 x_t + \cdots + B_q x_{t-q} + CD_t + u_t$$

where $y_t = \left(y_{1t}, \cdots, y_{Kt}\right)'$ is a K-dimension endogenous variable observed, $x_t = \left(x_{1t}, \cdots, x_{Mt}\right)'$ is a M-dimension exogenous variable observed, D_t concludes the observed deterministic variables such as the constant, linear trend, the fake crop as well as the other user-defined white noise, u_t is the process of K-dimensional 0 matrix, and $E\left(u_t u_t'\right) = \Sigma_u$ plus determines socks expectation covariance. A_j, B_j, C matrixes are the appropriate number of dimensions on themselves.

Although our purpose just forecasts, we also mentioned somewhat another important application of VAR model—the analysis of the shock transmission mechanism, reaction function and variance disintergration. However, calculating the reaction functions and variance disintergrations need parametric estimates in structural VAR models, we have unnecessary deeply interest in these problems but only conduct experimental analysis.

However, as mentioned in the introduction, another application of the VAR model is the analysis of shock transmission mechanism being done by reaction function and variance disintergration. We should not delve into analyzing the structural VAR model although calculating the reaction functions and variance disintergrations need parametric estimates in the model.

We use the VAR model for Vietnam's inflation forecasts because of its effective predictions, so general building and estimating VAR models will be introduced. Sample is the first concern before estimating the model. A large sample gives us vacant orders to estimate, and better estimate accuracy. However, with time series, the large sample (overlong string) raises issues about the stability of estimate coefficients in the model. Even in the countries with political and economic stability, policy changes in internal economy, and external action vary the relation of economic variables. Hence, monthly data is the best choice because of its sufficient free orders and stablility in the system. Usually, no monthly data exist to a macro variable, and then industrial production values used.

The parameters in VAR model are estimated following steps:

1) Testing the stationary of data series. If the data series non-stationary, we will check integrated community relations. If the relation occurs, VECM switched.

2) Lag Length Selection: The lag length for the *VAR*(*p*) model maybe determined using model selection criteria. The general approach is to fit *VAR*(*p*) models with orders $p = 0, \cdots, p_{\max}$ and choose the value of *p* which minimizes some model selection criteria. The three most common information criteria are the Akaike information criterion (AIC), Hannan-Quinn criterion (HQC), Schwarz information criterion (SIC), etc.

Latency optimizations are chosen by minimizing the following standard information:

$$\mathrm{AIC}(n) = \log\det\left(\tilde{\Sigma}_u(n)\right) + \frac{2}{T} nK^2,$$

$$\mathrm{HQ}(n) = \log\det\left(\tilde{\Sigma}_u(n)\right) + \frac{2\log\log T}{T} nK^2,$$

$$\mathrm{SC}(n) = \log\det\left(\tilde{\Sigma}_u(n)\right) + \frac{2\log T}{T} nK^2,$$

$$\mathrm{FPE}(n) = \left(\frac{T+n^*}{T-n^*}\right)^K \det\left(\tilde{\Sigma}_u(n)\right),$$

where $\tilde{\Sigma}_u(n)$ is estimated by $T^{-1}\sum_{t=1}^{T} \hat{u}_t \hat{u}_t'$, n is the number of parameters in every equation. Maybe, different standards are shown by different models. Hence, models with the most effective forecast are to be continued.

3) Diagnosing and simplifying the model.

- Checking the stability of the model statistically. If roots of the model are greater than or equal to 1, the model is nonstationary.
- Residual test: testing for autocorrelation of residual and testing for heteroskedasticity.
- Simplifying the model: Estimate results of the model (after being well-tested) provide statistical information about the role of lagged variables in the equation. Therefore, we will use these informations to verify if some lagged variables are statistically significant or not, so we should or should not remove any lagged variables of model.
- Checking the stability of parameters in the model.

- Analyzing and forecasting after having an effective model.

2.2. Genetic Programming

Genetic Programming (GP) is an automatic learning method thought from biological evolution with the target of establishing a computer program to meet the learners' expectation, so the GP is one of machine learning techniques using evolutionary algorithms to optimize computer programs following the compatibility of a program to calculate. The GP had tested since the 1980s, but until 1992, with the born of the book "Genetic Programming: On the Programming of Computers by Means of Natural Selection" by John Koza [3], it was visibly shaped. However, in the 1990s, the GP just solved simple problems. Today, together with the development of the hardware as well as the theory in the first half of 2000, the GP has grown rapidly.

Chromosome: Chromosome (a term borrowed from biological concepts), as in biology, determine the good level of an individual. The GP evolves a computer program representing under tree-like structure. The tree is easily evaluated by a recursive procedure. Each node on the tree is a calculating function, and each leaf stands for a class math, using for simple evolutionary and estimable mathematical expressions. As usual, the PG is an expression of the tree-like procedure.

Operators in the GP: Crossover and Mutation are two main operators used in the GP. These are also two terms borrowed biology, and two main factors affecting to evolution process.

Crossover: Shows the process of chromosome exchange between parents. Its operators include the following steps:

- Selecting randomly in each parent one node.
- Swapping their positions.

The crossover operator is shown in **Figure 1**.

Mutation: Mutation is the process of variation of a chromosome set created. The process includes the following steps:

- Choosing a node on the parent.
- Canceling the seedling on the node chosen.
- Birthing accidentally a new seedling on above position.

2.2.1. Primary Handling Steps for the GP

Existing five significant steps for primary handling the GP that a programmer need to establish:

1) Setting leaf nodes (such as independent variables, nonparametric functions, aleatory constants) for each branch of the evolution programming.

2) Collecting evolutional functions for each branch of evolution programming.

3) Pointing out a good fitness (measuring the compatibility of each individual in a population).

4) Determing parently the parameters controlling operation (individual volume, chromosome amount, variation probability ⋯).

5) Defining the criterion for finishing or the method for determining the result of the running process.

The diagram above (**Figure 2**) shows that if a GP is considered a "black box" with the input hold the primary handling steps, after going through the GP, the result received is a computer program (function to forecast).

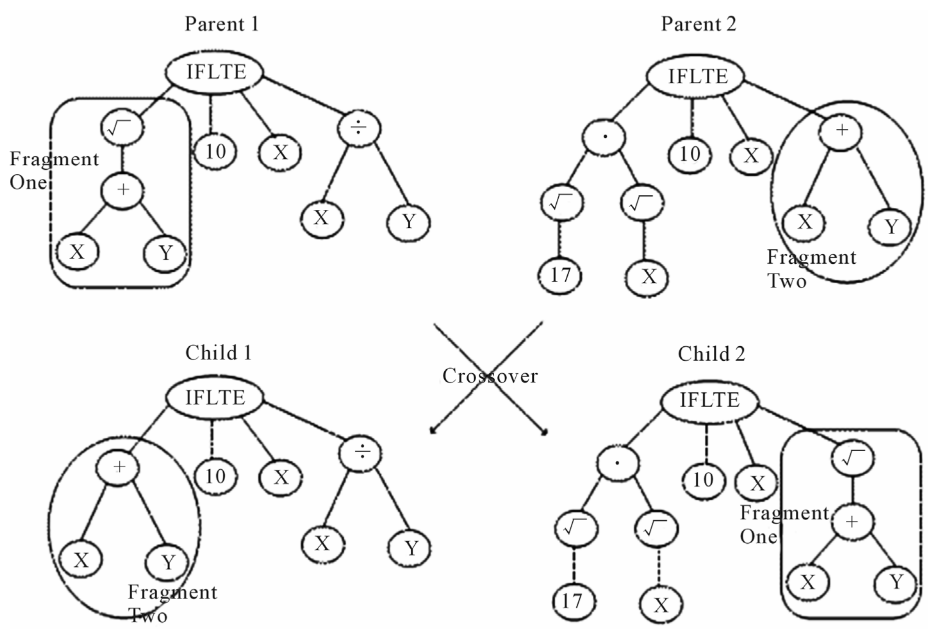

Figure 1. The crossover operator.

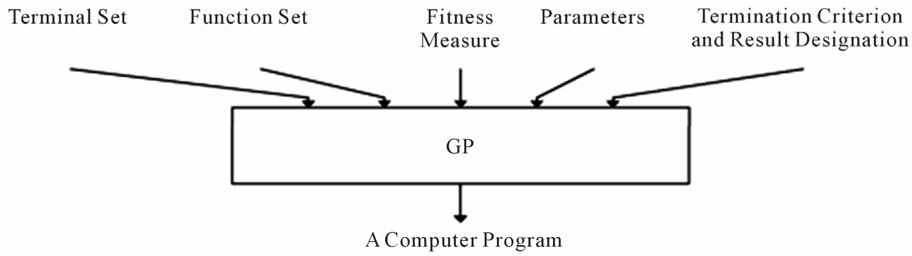

Figure 2. Primary handling steps for the GP.

2.2.2. Steps for Running the GP

A typical GP starts running with an accidental program made up by possible elements. Then, the sequential GP changes the population through many generations by using the operator in the GP. The selection process equates to count each individual property. An individual chosen to join in gene problems or canceled depends on its property (the way to value property given in the 3rd preparing step). The loop transformation of populations is the main content and is repeated many times in a program of GP. The sequence with the changeable population is the main sequential content in the GP.

Steps for running the GP is shown in **Figure 3** and includes the following steps:

1) Initializing incidentally a population (zero generation) with individuals created by functions, and leaf nodes.

2) Repeating (generations) follow postauxiliary until the condition satisfied.

3) Operating individuals to determine their property.

4) Choosing 1 or 2 individuals from the population with probability depending on their property to participate in the Gene problems in step 3).

5) Creating new individuals to the population by applying the post gene problems with specified probability.

6) Reproducing, and copying the selection into new populations.

7) Crossover: creating subindividual by combining sequentially portions of them.

8) Mutation: creating subindividual by replacing a new portion of the individual into its old one.

9) Structural changes: Being done by changing the structure of the individual selected. After satisfying the condition of last criterion the operation of the best individual means that the result of running process is exposed, and we receive the solution for problems basing on an effective operation of the individual.

2.2.3. Application of Genetic Programming (GP) to Prediction Problem

This section presents the method of applying GP for prediction/forecasting problems. The detail description can be found in a number of previous publication [4-6]. The task of time series prediction is to estimate the value of the series in the future based on its values in the past. There are two models of time series prediction: one-step prediction and multi-step prediction. In one-step prediction, the task is to express the value of $y(t)$ as a function of n previous values of the time series, $y(t-1)$, $\cdots, y(t-n)$ and other attributes. That is to find the function F so that:

$$y(t) = F\left(y(t-1),\cdots,y(t-n); x_1(t-1),\cdots, x_1(t-p_1);\cdots; x_k(t-1),\cdots,x_k(t-p_k)\right)$$

where $y(t-1),\cdots,y(t-n)$ are the values of the time series in the past and $x_1(t-1),\cdots,x_1(t-p_1)$ are the values of x_1 attributes in the past and $x_k(t-1),\cdots,x_k(t-p_k)$ are the values of x_k attributes in the past. This equation is based on an assumption that the value of the time series y depends on its previous values and also the values of some other attributes in the past. Fore xample, for CPI inflation prediction, the value of CPI in the future may depend on its previous values and thevalues of some other factors like total domestic product (GDP), monetary supply (M2), and soon. The purpose of multi-step prediction is to obtain predictions of several steps ahead into the future, $y(t)$, $y(t+1)$, $y(t+2),\cdots$ starting from the information at current time slice $t-1$. In this paper, we only focus on one-step prediction/forecasting.

3. Empirical Results

3.1. Description of Data

The data used in the model was provided by the General Statistics Office (GSO). We took information from two following data set to forecast.

Quarterly data: Over the period of 2nd quarter of 1996 to 4th quarter of 2011, monthly data: from January 1995 to February 2012. The data by 4th quarter of 2010 or December 2010 was selected to establish the model.

For quarterly data, real data in whole 4 quarters of 2010 was chosen to test and compare models. But monthly, only real data in 2011, and January and February 2012 was used for testing the model. New data updated in April and May 2012 has met the demand of reference for

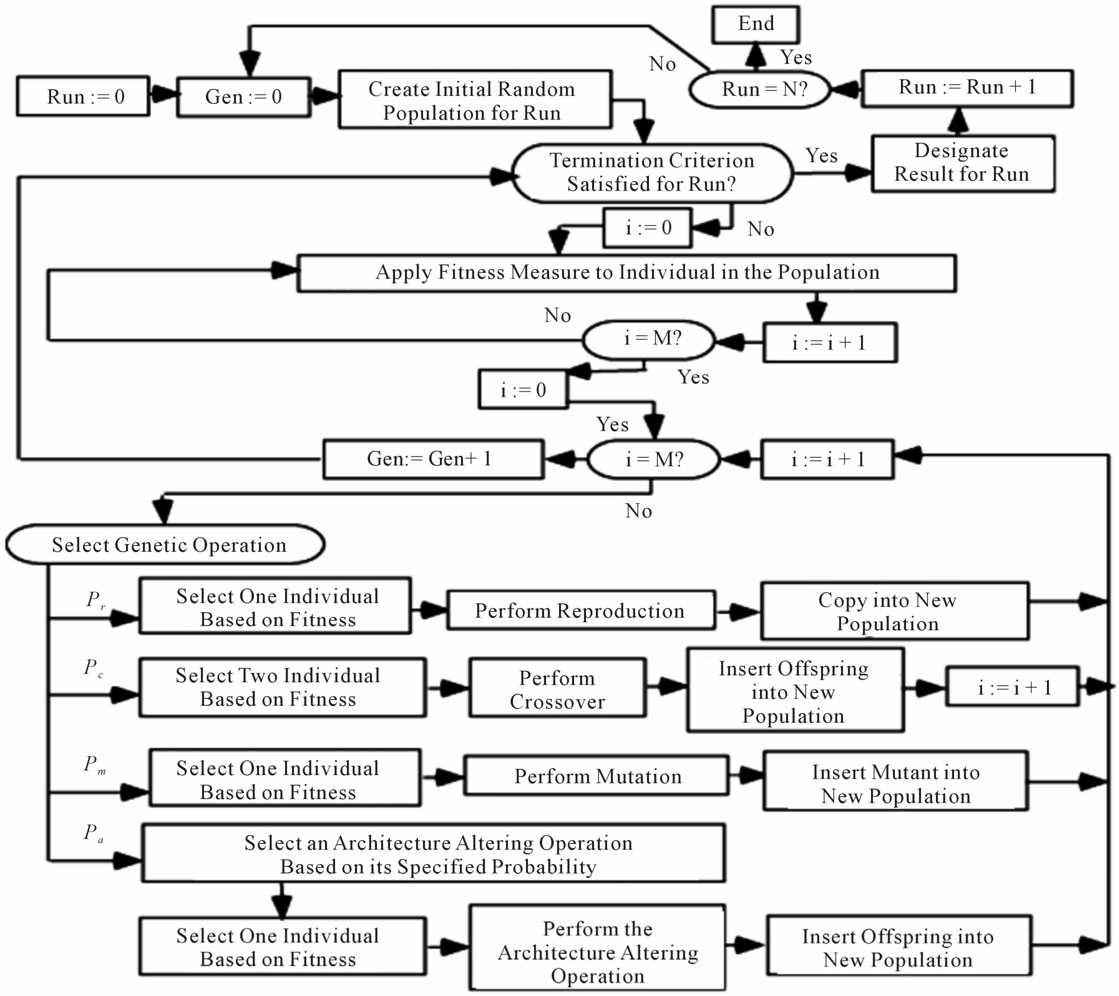

Figure 3. Steps for running the GP.

monthly data. Variables in the model are described in **Table 1**.

3.2. Estimating Inflation Forecast Models

3.2.1. The VAR Model for Inflation Predictions

3.2.1.1. Modeling Experimental Estimates

From steps for raising and estimating the model, we receive the VAR model with just 3 variables such as *gcpi*, *gex* and *gm*2 through all tests. The last model audited as:

$$\begin{aligned} g\hat{cpi}(t)_{\substack{str\\p\\t}} = &-\underset{\substack{(0.132)\\ [0.106]\\ \{-2.418\}}}{0.319}\, gex(t-1)+\underset{\substack{(0.357)\\ [0.052]\\ \{1.947\}}}{0.696}\, gm2(t-1)-\underset{\substack{(0.127)\\ [0.021]\\ \{-2.229\}}}{0.291}\, gcpi(t-1) \\ &-\underset{\substack{(0.128)\\ [0.103]\\ \{-1.629\}}}{0.208}\, gex(t-2)+\underset{\substack{(0.358)\\ [0.005]\\ \{2.779\}}}{0.994}\, gm2(t-2)-\underset{\substack{(0.136)\\ [0.076]\\ \{-1.774\}}}{0.242}\, gcpi(t-3) \end{aligned} \tag{1}$$

$$\begin{aligned} ge\hat{x}(t)_{\substack{str\\p\\t}} = &-\underset{\substack{(0.115)\\ [0.000]\\ \{-5.685\}}}{0.651}\, gex(t-1)+0\, gm2(t-1)-\underset{\substack{(0.108)\\ [0.000]\\ \{-5.176\}}}{0.559}\, gex(t-2) \\ &-\underset{\substack{(0.156)\\ [0.008]\\ \{-1.749\}}}{0.272}\, gcpi(t-3)-\underset{\substack{(0.101)\\ [0.001]\\ \{-3.369\}}}{0.304}\, gex(t-3) \end{aligned} \tag{2}$$

$$\begin{aligned} g\hat{m}2(t)_{\substack{str\\p\\t}} = &\underset{\substack{(0.039)\\ [0.051]\\ \{1.951\}}}{0.076}\, gex(t-1)-\underset{\substack{(0.109)\\ [0.000]\\ \{-10.057\}}}{1.092}\, gm2(t-1)-\underset{\substack{(0.118)\\ [0.000]\\ \{-6.442\}}}{0.762}\, gm2(t-2) \\ &-\underset{\substack{(0.050)\\ [0.027]\\ \{-2.005\}}}{0.109}\, gcpi(t-3)-\underset{\substack{(0.099)\\ [0.005]\\ \{-2.828\}}}{0.280}0\, gm2(t-3) \end{aligned} \tag{3}$$

Table 1. The name of variables in used model.

Variable name	Signs	Growth
Total domestic product with the price in 1994	gdp	ggdp
Consumer price index compared to the previous month	cpi	*gcpi*
The US dollar price index compared to previous month	Ex	*Gex*
Commodity import turnover	im	gim
Monetary supply	M2	*Gm*2

The model met the demand of stability, autocorrelation, changeable error variance tests. In fact,

- Modeling stability checking.

After estimating the VAR model, we need to test its stability. The test determines whether roots of characteristic polynomial belong to unit circle or not. All roots in **Table 2** are less than 1 in module. Therefore, the VAR is acceptable thanks to stable equation.

Table 2: Checking the stability of the model through roots of typical polynomials by endogenous variables: *gcpi*4, *gex*4, *gm*2.

- Test for the autocorrelations of residuals.

Results of Portmanteau Tests basing on Q test (**Table 3**) showed that with lagged steps, p in Q test is greater than 5%. This means hypothesis H_0 not to be canceled (no residual autocorrelations).

- Testing for Heteroskedasticity.

The residual Heteroskedasticity tests are carried by general tests about heteroskedasticity of White. The result of White test showed that no heteroskedasticity remains. The result is introduced in **Table 4**.

- Impulse Response Functions.

With the model estimated, we can analyze the shock transmission mechanism through response functions. To recieve the response function, some constraints are applied for the equation. Constraints chosen are Cholesky Disintegrate ones. The Cholesky used serially: *gex*, *gcpi*. Choosing the order depends on inflation changes without effects on exchange rate. Analyzing performance is as in **Figure 4**.

The first two figures see inflation changes struggled by the shock itself, which is descending and being vanished for a 5 quarter. Monetary supply and exchange rate shock influences seems to impact insensibly on Vietnam's inflation.

3.2.1.2. Prediction Results

1) Inflation predictions for 2011

For inflation predictions in 2011, we use the models (1)-(3) to estimate and data by 2010 to apply the model and make forecasting procedures for 2011. Acquired results are given in the following **Table 5**.

Table 2. Unit root test.

Root	Modulus
−0.316094 − 0.767403i	0.829954
−0.316094 + 0.767403i	0.829954
0.010726 − 0.758628i	0.758704
0.010726 + 0.758628i	0.758704
−0.712593 − 0.062363i	0.715317
−0.712593 + 0.062363i	0.715317
0.276312 − 0.618574i	0.677482
0.276312 + 0.618574i	0.677482
−0.461363	0.461363
No root lies outside the unit circle.	
VAR satisfies the stability condition.	

Table 3. Tests for autocorrelations of residuals.

VAR Residual Portmanteau Tests for Autocorrelations					
Null Hypothesis: no residual autocorrelations up to lag h					
Lags	Q-Stat	Prob.	Adj Q-Stat	Prob.	df
1	3.318091	NA*	3.374330	NA*	NA*
2	5.494490	NA*	5.625777	NA*	NA*
3	7.178825	NA*	7.398761	NA*	NA*
4	13.37463	0.1464	14.03713	0.1210	9
5	25.29735	0.1169	27.04373	0.0782	18
6	27.95866	0.4131	30.00074	0.3141	27
7	36.30709	0.4543	39.45180	0.3183	36
8	39.24429	0.7135	42.84087	0.5638	45
9	42.83282	0.8630	47.06266	0.7368	54
10	47.15679	0.9319	52.25144	0.8309	63
11	58.41922	0.8760	66.04216	0.6754	72
12	65.36161	0.8970	74.72015	0.6751	81

*The test is valid only for lags larger than the VAR lag order df is degrees of freedom for (approximate) chi-square distribution.

Table 4. Results of white test.

VAR Residual Heteroskedasticity Tests: Includes Cross Terms		
Joint test:		
Chi-sq	df	Prob.
322.3132	324	0.5160

Source: Estimates of author.

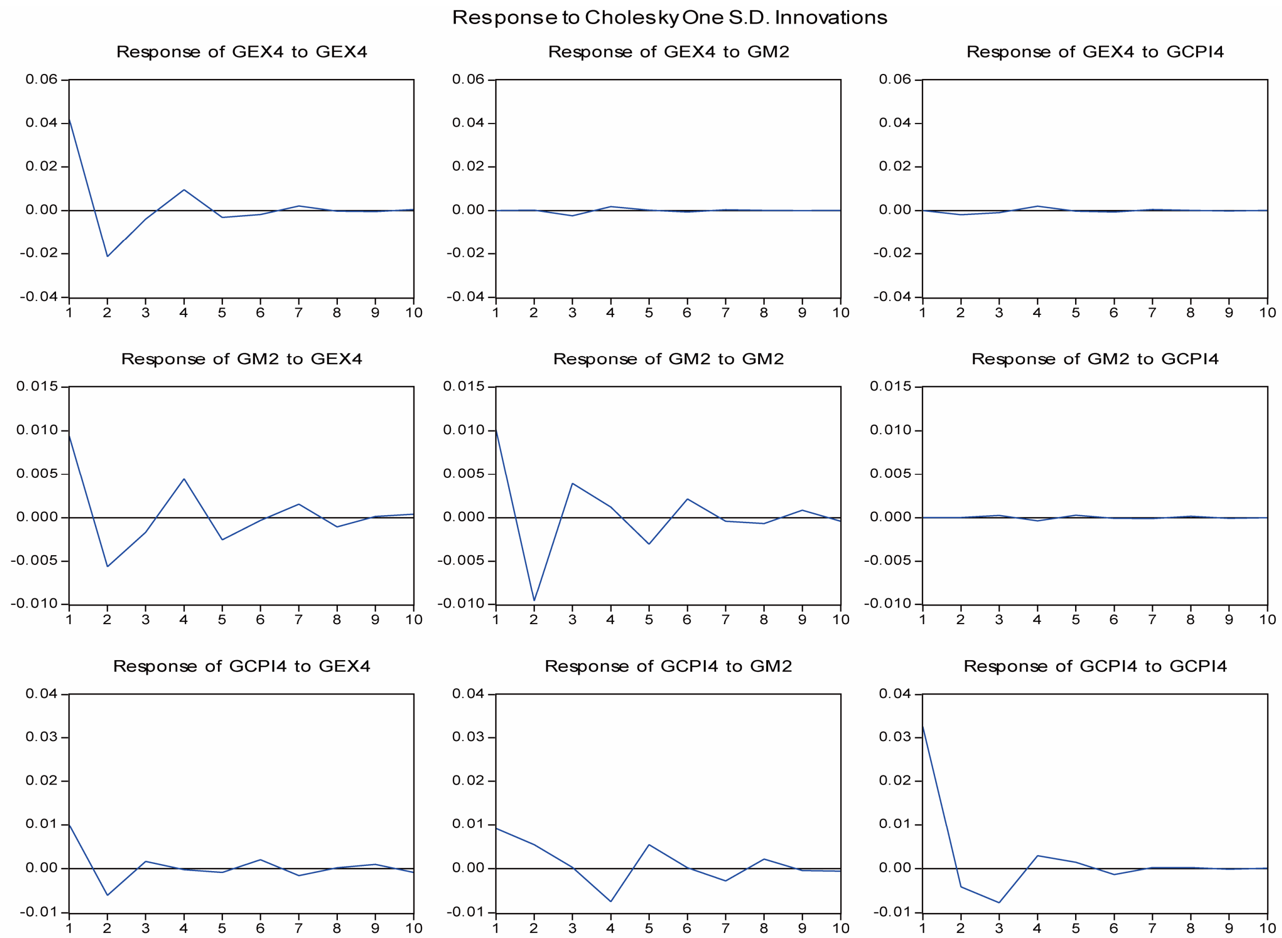

Figure 4. Response function-shock transmission mechanism.

The **Table 2** sees the increase of inflation *gcpi*, exchange rate and monetary demand, but it is hard to compare with real inflation to value its property. In doing so, we do a counter-process of considering the inflation increase in comparison with real inflation. The result given in **Table 6**.

See the table above, it's obvious that square roots of average square prediction errors is 1.38. The grestest quarterly variance is 3%, pointing out that the acquired model is quite effective.

2) Inflation predictions for 2012

For inflation predictions in 2012, we use the model (1)-(3) to estimate and data by 2011 to apply the model and make forecasting procedures for 2012. Acquired results are given in **Table 7**.

3.2.2. Applications of the GP for Inflation Predictions

3.2.2.1. GP Parameters Settings

To tackle a problem with GP, several factors need to be clarified beforehand. These factors often depend on the problem and the experience of the system user (practitioner). The first and important factor is the fitness function. Traditionally, for symbolic regression problems, the fitness function is the sum of the absolute (or some times the square) error. Formally, the (minimising) fitness function of an individual is defined as:

$$\text{Fitness} = \sum_{i=1}^{n} \left| y_i - f_i \right|$$

where N is the number of data samples (fitness cases), y_i is the value of the CPI in the data sample, and f_i is the function value of the individual at the i^{th} point in the sample set (f_i is the fitted value of y_i).

To assess the consistency of a model created by the GP, we put additional quantities:

$$\text{Test Fitness} = \sum_{i=n+1}^{N} \left| y_i - f_i \right|$$

where $y_i,\ i = n+1, \cdots, N$ is real value of CPI in the test data sample (for monthly data $y_i, i = n+1, \cdots, N$ is real value of CPI from 2011M1 to 2012M2, for quaterly data $y_i, i = n+1, \cdots, N$ is real value of CPI from 2011Q1 to 2011Q4), and f_i is the predicted value of y_i. Some evolutionary parameters are presented in **Table 8**.

Table 5. Prediction results of *gcpi* from model for 2011.

year	*gm*2	lowerCI	upperCI	±
2011Q1	−0.0154	−0.0479	0.017	0.0324
2011Q2	0.0063	−0.028	0.0406	0.0343
2011Q3	0.0063	−0.0281	0.0406	0.0343
2011Q4	−0.0063	−0.0415	0.0289	0.0352
*gcpi*4	forecast	lowerCI	upperCI	±
2011Q1	−0.0194	−0.0689	0.0301	0.0495
2011Q2	0.0019	−0.0476	0.0514	0.0495
2011Q3	0.0025	−0.047	0.052	0.0495
2011Q4	0.017	−0.0344	0.0684	0.0514
*Gex*4	forecast	lowerCI	upperCI	±
2011Q1	−0.0606	−0.1278	0.0065	0.0671
2011Q2	0.0087	−0.0654	0.0828	0.0741
2011Q3	0.0193	−0.0551	0.0937	0.0744
2011Q4	0.0116	−0.0639	0.087	0.0755

Source: Estimates of author. The prediction result witnesses some 95% accuracy, lower confidence interval, upper confidence interval and just variances in last column.

Table 6. Comparisons the forecast results and actual inflation for the CPI in 2011.

	CPI(real)	CPI(predict)	Prediction error	Prediction square error
2011Q1	102.1700	100.0016	−0.0212	0.00045
2011Q2	101.0900	102.3641	0.0126	0.00016
2011Q3	100.8200	101.3427	0.0052	0.00003
2011Q4	101.3700	102.5339	0.0115	0.00013
Square root of prediction mean square error				0.013856558

Source: Estimates of author.

where

$$\text{my}\log(x)=\begin{cases}0 & \text{if } x\le 0\\ \ln(x) & \text{if } x>0\end{cases},\text{mysinsh}(x)=\frac{e^{-x}-e^{x}}{2},$$

$$\text{my}\log\text{is}(x)=\frac{1}{1+e^{-x}},\text{mysqrt}(x)=\begin{cases}0 & \text{if } x\le 0\\ \sqrt{x} & \text{if } x>0\end{cases},$$

$$\text{mydivide}(y,x)=\begin{cases}0 & \text{if } x=0\\ \dfrac{y}{x} & \text{if } x\ne 0\end{cases}$$

3.2.2.2. Applying Quarterly Data for Forecasting

On the basis of selected variables from the VAR model,

Table 7. Results for predicting *gcpi* in the model for 2012.

	Prediction of *gcpi*				Prediction of *cpi*
	forecast	lowerci	upperci	±	
2012q1	−0.0146	−0.0788	0.0496	0.0642	99.88999
2012q2	0.0224	−0.0441	0.0888	0.0665	102.1275
2012q3	0.0363	−0.0325	0.1051	0.0688	105.8348
2012q4	−0.006	−0.0795	0.0674	0.0735	105.1998

Source: The forecast results for the directly acquired growth rate from predicting model for CPI are contributed to prediction performance in the model and 4th quarterly of 2011.

Table 8. Run and evolutionary parameter values.

Parameter	Value
Population size	250
Generations	40
Selection	Tournament
Tournament size	3
Crossover probability	0.9
Mutation probability	0.05
Initial Max depth	6
Max depth	30
Max depth of mutation tree	5
Non-terminals	+, −, /, −, exp, mylog, mysinsh, mylogis, mysqrt, mydivide, sin, cos.
Terminals	$cpi(t-1), \cdots, cpi(t-12), ex(t-1), \cdots, ex(t-4), gm2(t-1), \cdots, gm2(t-4)$
Raw fitness	mean absolute error on all fitness cases
Trials per treatment	50 independent runs for each value

we have established models for inflation predictions. Nevertheless, favorable models witness the dependence of CPI on its values in the past.

Prediction model (a)

$$\hat{cpi}(t)=\frac{\left(cpi(t-3)-cpi(t-4)\right)}{e^{\sqrt[4]{cpi(t-2)}}}+cpi(t-4)\sqrt{\frac{cpi(t-4)}{cpi(t-1)}} \quad (4)$$

where $\hat{cpi}(t)$ is the prediction of $cpi(t)$. Acquired results of predictions for 2011 using model (a) are given in the following **Table 9**.

It can be seen that square roots of average square prediction errors is 0.45%, and 0.9% means the grestest quarterly variance. The predictions for 2012 and 2013 are given in **Table 10**.

Prediction model (b)

$$\begin{aligned}
&\hat{cpi}(t) = cpi(t-1) - \sin\left\{cpi(t-1) - \sin\left(\sin\left(cpi(t-1) - \sin\left(cpi(t-1) - f - cpi(t-4)\right)\right)\right)\right\} \\
&f = \frac{1}{1+\exp\left(\exp\left(\sin\left(\sin\left(\sin\left(\frac{cpi^2(t-2)\times cpi^2(t-3)}{cpi(t-4)}\right)\right)\right)\right)-g\right)} \\
&g = \sqrt{cpi(t-2) - cpi(t-1)\times\cos\left(\frac{1}{1+\exp\left(\sqrt{cpi(t-1)} - cpi(t-1)\right)}\right)}
\end{aligned} \tag{5}$$

Acquired results of predictions for 2011 using model (b) are given in the following **Table 11**.

Evidently, square roots of average square prediction errors is 0.55%, and 1.1% means the grestest quarterly variance.

The predictions for 2012 and 2013 using model (b) are given in **Table 12**.

Prediction model (c)

$$\begin{aligned}
&\hat{cpi}(t) = \sqrt{cpi(t-1)\times\left(cpi(t-4)+h\right)} \\
&f = \sqrt{\frac{\exp\left(\frac{1}{g}\right)+\exp\left(-\frac{1}{g}\right)}{2}}, \\
&g = 1+\exp\left(cpi(t-4)-cpi(t-1)\right) \\
&h = \left(\cos\left(\frac{\sqrt{cpi(t-3)}}{f}\right)\times\frac{\exp\left(cpi(t-3)\right)-\exp\left(-cpi(t-3)\right)}{2}\right)
\end{aligned} \tag{6}$$

The sequence of model (6) is as ineffective as that of model (5). The result is not introduced here.

Table 9. Predictions for 2011 using the data by 2010.

Time	Real data	Prediction	Prediction error	Prediction square error
2011Q1	102.17	101.341	0.0081146	0.0000658
2011Q2	101.09	101.236	−0.0014416	0.0000021
2011Q3	100.82	101.228	−0.0040451	0.0000164
2011Q4	101.37	101.406	−0.0003582	0.0000001
Square root of prediction mean square error				0.0045939

Table 10. Predictions for 2012, 2013 using the data by 2010.

	2012				2013			
Time	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Forcast	101.72	101.40	101.13	101.27	101.48	101.43	101.28	101.28

The formula (7) below employing data by 2011 shows good performance:

$$\begin{aligned}
&\hat{cpi}(t) = cpi(t-1) + \sin\left(\frac{cpi(t-1)}{f}\right) \\
&f = \left(\frac{cpi(t-4)}{g}\right) + \cos\left(\frac{cpi(t-3)}{cpi(t-3)\times cpi(t-4)}\right) \\
&g = cpi(t-1) - \sin\left(\frac{cpi(t-4)\left(cpi(t-1)-cpi(t-2)\right)}{cpi(t-1)-cpi(t-3)}\right)
\end{aligned} \tag{7}$$

The predictions for 2012 and 2013 using this formula are given in **Table 13**.

3.2.2.3. Forecasts Basing on Monthly Data

Here, we have predicted CPI values in the future based on its previous ones. Data from January 1995 to December

Table 11. Predictions for 2011 using the data by 2010.

Time	Real data	Prediction	Prediction error	Prediction square error
2011Q1	102.17	101.15	0.0100748	0.0001015
2011Q2	101.09	101.31	−0.0021692	4.705E-06
2011Q3	100.82	101.18033	−0.0035268	1.244E-05
2011Q4	101.37	101.31303	0.0005635	3.176E-07
Square root of prediction mean square error				0.0054535

Table 12. Predictions for 2012, 2013 using the data by 2010.

	2012				2013			
Time	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Forecast	101.37	101.05	100.82	101.19	101.21	101.02	100.82	101.10

Table 13. Predictions for 2012, 2013 using the data by 2011.

	2012				2013			
Time	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Forecast	101.06	100.88	100.62	100.99	100.61	100.83	100.40	100.38

2011are taken to raise the model, those over the time of January 2011 to February 2012 are used for testing. Similary, The reference of prediction accuracy relies on data in March and April 2012.

We introduce some models found out by GP with the min fitness (98.0298) for CPI predictions:

$$\begin{aligned}c\hat{p}i(t) = \frac{1}{10}\Big\{ & cpi(t-11) + 5\times cpi(t-1) + 2\times cpi(t-12) \\ & + cpi(t-2) + \sin\big(cpi(t-9)\big) + cpi(t-7) \\ & + \sin\Big(e^{\sin(cpi(t-7)) - cpi(t-8)}\Big)\Big\}\end{aligned} \tag{8}$$

In **Table 14**, we see that the greatest monthly error is 1.83%. Square roots of average square prediction errors are 0.00696 which is much less than predictions of the VAR model.

Inflation forecasts in March 2012 is 101.697, but actual data are 100.16, and error 1.535%. Similarly, inflation forecasts 102.0122, real data 100.05 and error 1.961 are in April 2012. March data took one-step prediction, and then two-step prediction for April ones, means that using March data predicts following month.

Prediction results over the time of May to December 2012 in below **Table 15**.

3.2.2.4. Evaluating the Consistence of the GP

For evaluating the consistence of prediction models created by GP, we consider 50 models to every quarterly and monthly forecasts, and examine the relation of errors inside and outside sample. A model fitting to both past and future data (on the other hands, the error inside sample is small, that of outside sample also similar) is called the consistent model. Prediction models of the GP would be considered to be consistent if small fitness implied small test fitness, meaning that test fitness is a varied flow function of the fitness.

Table 14. Predictions using the data by 2010.

Time	Real value	Prediction value	Prediction error	Square prediction error
2011M1	101.74	101.6279	0.001102	0.0000012
2011M2	102.09	101.5152	0.00563	0.0000317
2011M3	102.17	101.3754	0.007777	0.0000605
2011M4	103.32	101.4346	0.018248	0.0003330
2011M5	102.21	102.0284	0.001777	0.0000032
2011M6	101.09	101.7433	−0.00646	0.0000418
2011M7	101.17	101.0486	0.0012	0.0000014
2011M8	100.93	101.1421	−0.0021	0.0000044
2011M9	100.82	101.2573	−0.00434	0.0000188
2011M10	100.36	101.2095	−0.00846	0.0000717
2011M11	100.39	101.264	−0.00871	0.0000758
2011M12	100.53	101.1218	−0.00589	0.0000347
2012M1	101	101.0045	−4.5E-05	0.0000000
2012M2	101.37	101.4044	−0.00034	0.0000001
Square root of prediction mean square error				0.0069600

Table 15. Predictions results over the time of May to December 2012 using the data by 2010.

Time	M5	M6	M7	M8	M9	M10	M11	M12
Forecast	101.8	101.5	101.29	101.15	101.04	100.96	100.99	101.073

We received following equation thanks to carrying out test fitness linear regression basing on quarterly models:

$$\begin{aligned}&\text{Test Fitness} = 0.053016 \times \text{Fitness}\\ &\text{Std.Error} = 0.001453\\ &R^2 = 0.226003,\ DW = 1.319782\end{aligned} \tag{9}$$

The correlation coefficient between test fitness and fitness is 0.485. Therefore, it can be valued that quarterly prediction models are consistent because of regressive results (9), and positive correlation equation between the test fitness and the fitness.

Similarly, monthly data have regressive model as:

$$\begin{aligned}&\text{Test Fitness} = 0.087980 \times \text{Fitness}\\ &\text{Std.Error} = 0.001685\\ &R^2 = 0.520951,\ DW = 1.905514\end{aligned} \tag{10}$$

The data for regressive models (9) and (10) are getting from **Table A** in Appendix A. The correlation coefficient between test fitness and fitness is 0.751, therefore, from regressive result (10), with positive lope and positive correlation between test fitness and fitness, it can be seen that monthly prediction models of GP are consistent. Additionally, monthly models witness a bigger correlation between test fitness and fitness against quarterly ones.

4. Conclusions

Using the VAR model for inflation prediction has succeeded in selecting fitting models in line with currently available data. If we take predictions for 2011 from the VAR model to be comparision standard, square roots of average square prediction errors is 1.38. The best impressive error by months is not greater than 3%. Using the model to forecast for 2011 with accuracy 95% indicates that inflation in the 1st quarter of 2012 decreases slightly, it continuously witnesses somewhat increase in the 2nd and the 3rd quarters. The performance needs testing.

Above results from the VAR model show that recently have been resulted from different causes, especially uncontrolled stimulus packages, uneffective public investment and state expense, and prolong inflation lead to high inflation rate in economy. This confirms how important the application of monetary policies with a consistent attitude for improving the credibility of policies is. Consistence when applying monetary policies is also one way to impact on inflation expectation we desire.

Errors of prediction results from models created by the GP are much less than those of the VAR model. One benefit from using the GP to raise the prediction model is that we don't need to specify the model (the GP itself discovered the model), and propose hypothesizes for variables in the model. GP can provide some analytical formulas for prediction of the model so the GP is called "white box", unlike the neural network model called "black box", which shows us the predicted value, without giving analytical expressions. Analytical expressions also help us to detect relationships between variables in forecasting models and assess interactions between them. GP can also help us to detect relationships between economic variables that economic theory can not detect or exceed human judgments. So the greatest advantage of GP comes from the ability to address problems for which there are no human experts. Although human expertise should be used when it is available, it often proves less than adequacy for automating problem-solving routines. Nonetheless, the GP can't indicate the accuracy of prediction values and their distribution. Moreover, prediction functions for the GP are often complicated, and difficult to explain. That's all about disadvantages of the GP. These show that the CPI in the future just depends on its values currently and previously but other variables. Obviously, Vietnam's inflation rate mainly bases on its citizens' expectation, particularly in the late of December 2011, salary increases for evil servants released by government and applied from 1st May 2012 has followed price augment from January 2012 without basing on other elements.

5. Acknowledgements

The work in this paper was funded by The Vietnam National Foundation for Science and Technology Development (NAFOSTED), under grant number 10103-2010.06.

REFERENCES

[1] J. Hamilton, "Time Series Analysis," Princeton University Press, Princeton, 1994.

[2] C. A. Sims, "Macroeconomics and Reality," *Econometrica*, 1980, Vol. 48, No. 1, 1980, pp. 1-48.

[3] J. Koza, "Genetic Programming: On the Programming of Computers by Natural Selection," MIT Press, Cambridge, 1992.

[4] M. Santini and A. Tettamanzi, "Genetic Programming for financial Time Series Prediction," *Proceedings of Euro Genetic Programming*, Lake Como, 18-20 April 2001, pp. 361-370.

[5] D. Rivero, J. R. Rabunal, J. Dorado and A. Pazos, "Time Series Forecast with Anticipation Using Genetic Programming," 8*th International Work-Conference on Artificial Neural Networks*, *Computational Intelligence and Bio-inspired Systems*, Barcelona, 8-10 June 2005, pp. 968-975.

[6] J. Li, Z. Shi and X. Li, "Genetic Programming with Wavelet-Based Indicators for Financial Forecasting," *Transactions of the Institute of Measurement and Control*, Vol. 28, No. 3, 2006, pp. 285-297. doi:10.1191/0142331206tim177oa

Appendix A: Fitness and Test Fitness Results 50 Models Created by GP

Table A. Fitness and test fitness results 50 models created by GP is ascending sorted by fitness value.

Monthly data			Quarterly data		
Model	Fitness	Test Fitness	Model	Fitness	Test Fitness
34	98.02098	7.326765	4	28.24487	1.614447
36	99.0578	7.696	11	32.07805	2.33514
19	99.27826	7.174	19	32.35036	1.411474
5	99.46484	7.499008	31	32.44957	1.454296
46	99.98677	7.832	28	32.78621	1.619906
16	100.3978	7.026	29	32.8773	1.544387
17	100.4957	8.230283	2	32.87733	2.36032
22	100.6998	7.545	6	32.87733	2.36032
18	100.9807	8.456925	21	33.10339	1.543815
48	102.4999	8.519	17	33.19176	2.790995
50	103.0303	7.824449	5	33.2398	1.367113
47	103.8416	9.373	3	33.38185	1.334923
37	103.9578	9.992	50	33.40914	1.415979
26	104.397	9.313765	39	33.72059	1.366684
10	104.7919	9.576941	26	33.75907	1.333442
3	105.1143	8.777088	35	33.80046	1.391022
44	105.3956	7.669	25	33.80252	1.324483
42	105.6116	10.0012	47	33.82245	1.361137
45	105.639	10.18452	1	33.84738	1.31545
24	106.1196	9.128	20	33.84738	1.31545
27	106.3108	11.58753	48	33.85151	1.307885
15	107.7486	6.808134	22	33.92415	1.719595
25	108.0623	9.756734	41	34.33892	2.09
31	108.3977	6.708472	23	34.64307	1.770351
11	108.9255	9.413899	24	34.85562	2.09
41	109.3048	7.765949	36	35.185	2.252604
1	109.8344	9.888813	30	35.66765	1.95576
23	110.6479	12.0302	34	35.75907	2.039319
14	111.8084	10.55205	33	36.25562	2.09
43	112.0858	10.53303	43	36.4729	2.088026
21	112.1578	10.771	37	36.56566	2.664865
32	112.7478	9.323298	8	36.5731	2.172006
35	112.8181	8.899476	42	37.63429	2.306655
49	112.9904	12.501	9	37.79298	2.069823
20	112.9965	9.248496	40	39.05691	2.355757
40	113.3175	9.391658	16	39.1693	2.179697
8	114.5063	11.11547	15	39.38582	2.057293
6	114.5718	7.906434	14	39.43949	1.575139
12	115.1631	11.51576	44	39.4543	1.969263
38	115.8079	10.882	38	39.62798	2.025501
30	115.9109	10.7563	12	39.71468	1.631938
13	116.1425	12.02868	49	39.883	2.177382
33	119.3571	10.35254	13	39.91915	2.199305
4	120.424	13.798	32	39.95025	2.302056
39	122.9646	12.05614	45	40.03305	2.240448
9	126.361	12.88535	18	40.07738	2.327514
29	131.8939	11.01971	10	40.14163	1.965275
28	132.8433	12.43679	7	40.29749	2.069947
7	138.8314	12.08349	46	40.40863	2.089324
2	165.7095	14.75873	27	40.77183	2.723862

Appendix B: Some Prediction Function Created by GP with Small Fitness and Test Fitness (Monthly Data)

Model 34

$$c\hat{p}i(t) = \frac{1}{10}\left(5cpi(t-1) + cpi(t-11) + 2cpi(t-12) + cpi(t-2) + f\right)$$
$$f = \sin\left(cpi(t-9)\right) + cpi(t-7) + \sin\left(e^{\sin(cpi(t-7)) - cpi(t-8)}\right)$$

Fitness = 98.02098, Test Fitness = 7.326765

Model 36

$$c\hat{p}i(t) = \frac{1}{10}\left\{3cpi(t-12) + cpi(t-6) + 4 \times cpi(t-1) + cpi(t-10) + cpi(t-2)\right.$$
$$\left. + \mathrm{mysqrt}\left(cpi(t-3) \times \left(e^{\mathrm{mylog}(\mathrm{mylogis}(cpi(t\text{-}9)))} - cpi(t-9)\right) + cpi(t-6) + cpi(t-4) + cpi(t-9)\right)\right\}$$

Fitness = 99.0578, Test Fitness = 7.696

Model 19

$$c\hat{p}i(t) = \frac{1}{10}\left\{4cpi(t-1) + 2cpi(t-2) + cpi(t-7) + 2cpi(t-9) + 3cpi(t-12)\right\}$$

Fitness = 99.27826, Test Fitness = 7.174

Model 5

$$c\hat{p}i(t) = \frac{1}{10}\left\{5cpi(t-1) + \sin\left(cpi(t-2)\right) + cpi(t-3) + 2cpi(t-7) + 2cpi(t-12)\right\}$$

Fitness = 99.46484, Test Fitness = 7.499008

Model 31

$$c\hat{p}i(t) = \frac{1}{10}\left\{cpi(t-1) \times \mathrm{my}\log cpi(t-12) \times \left(\mathrm{mysqrt}cpi(t-10) - cpi(t-1)\right)\right.$$
$$\left. - \left(cpi(t-3) \times \mathrm{my}\log\left(\mathrm{mysqrt}\left(\mathrm{my}\log\left(cpi(t-12)\right) + cpi(t-6)\right)\right)\right)\right\}$$

Fitness = 108.3977, Test Fitness = 6.708472

Appendix C: Some Prediction Function Created by GP with Small Fitness and Test Fitness (Quaterly Data)

Model 4

$$c\hat{p}i(t) = \sqrt{cpi(t-1) \times cpi(t-4)} - f + \cos\left(cpi(t-4)\right)$$
$$f = \cos\left(\sqrt{cpi(t-1) \times cpi(t-4)} - \cos\left(\mathrm{my}\log\mathrm{is}\left(\frac{m2(t-4) - m2(t-5)}{m2(t-5)}\right)\right) + g\right)$$
$$g = \mathrm{my}\log\mathrm{is}\left(\cos\left(\frac{m2(t-1) - m2(t-2)}{m2(t-2)}\right)\right)$$

Fitness = 28.24487, Test Fitness = 1.614447

Model 11

$$c\hat{p}i(t) = cpi(t-1) - \mathrm{mydivide}\left(\frac{m2(t-2) - m2(t-3)}{m2(t-3)}, f\right)$$
$$f = \mathrm{mydivide}\left(\mathrm{my}\sinh\left(\mathrm{my}\log\left(\frac{m2(t-3) - m2(t-4)}{m2(t-4)}\right)\right), \frac{m2(t-1) - m2(t-2)}{m2(t-2)}\right)$$

Fitness = 32.07805, Test Fitness = 2.33514

Model 48

$$\hat{cpi}(t)=\sqrt{cpi(t-1)\times cpi(t-4)-\mathrm{mylog}\left(\mathrm{mysqrt}\left(\mathrm{mylog}\left(cpi(t-2)\right)\right)\right)}$$

Fitness = 33.85151, Test Fitness = 1.307885

Model 19

$$\hat{cpi}(t)=\sqrt{cpi(t-4)\times\left(cpi(t-1)+f\right)}$$

$$f=\mathrm{mylog}\left(\mathrm{mylogis}\left(cpi(t-4)+\mathrm{mylogis}\left(cpi(t-4)\right)-cpi(t-4)\right)+g\right)$$

$$g=\mathrm{mylogis}\left(cpi(t-1)+\mathrm{mylogis}\left(cpi(t-3)\right)+\mathrm{mylogis}\left(cpi(t-4)\right)-cpi(t-4)\right)$$

Fitness = 32.35036, Test Fitness = 1.411474

Model 20

$$\hat{cpi}(t)=\sqrt{cpi(t-1)\times cpi(t-4)\times\mathrm{mylogis}\left(cpi(t-1)\right)}$$

Fitness = 33.84738, Test Fitness = 1.31545