Defence and Peace Economics, 2007,
Vol. 18(5), October, pp. 451–466

USING DYNAMIC FORECASTING GENETIC PROGRAMMING (DFGP) TO FORECAST UNITED STATES GROSS DOMESTIC PRODUCT (US GDP) WITH MILITARY EXPENDITURE AS AN EXPLANATORY VARIABLE

NEAL WAGNER AND JURGEN BRAUER*

Augusta State University, Augusta, GA 30904, USA

(Received 24 June 2006; in final form 1 May 2007)

Classic time-series forecasting models can be divided into exponential smoothing, regression, ARIMA, threshold, and GARCH models. Functional form is investigator-specified, and all methods assume that the data generation process across all segments of the examined time-series is constant. In contrast, the aim of heuristic methods is to automate the discovery of functional form and permit different segments of a time-series to stem from different underlying data generation processes. These methods are categorized into those based on neural networks (NN) and those based on evolutionary computation, the latter further divided into genetic algorithms (GA), evolutionary programming (EP), and genetic programming (GP). However, the duration of the time-series itself is still investigator determined. This paper uses a dynamic forecasting version of GP (DFGP), where even the length of the time-series is automatically discovered. The method is applied to an examination of US GDP that includes military expenditure among its determinants and is compared to a regression-based forecast. We find that DFGP and a regression-based forecast yield comparable results but with the significant proviso that DFGP does not make any prior assumption about functional form or the time-span from which forecasts are produced.

1. INTRODUCTION

Based on Romer (2000) and Taylor (2000, 2001), Atesoglu (2002) published an interesting Keynesian macro-economic model. Let:

$$QL_t = CL_t + IL_t + XL_t + ML_t + GL_t \quad (1)$$

$$CL_t = a + b\left(QL_t - TL_t\right) \quad (2)$$

$$TL_t = n + g\left(QL_t\right) \quad (3)$$

$$IL_t = e - f\left(R_t\right) \quad (4)$$

$$XL_t = z - m\left(QL_t\right) - n\left(R_t\right) \quad (5)$$

*Corresponding Author. Email: jbrauer@aug.edu

ISSN 1024-2694 print: ISSN 1476-8267 online
DOI: 10.1080/10242690701455508

where equation (1) is the national income and expenditure accounting identity with output (*QL*) identically equal to consumption (*CL*), investment (*IL*), net exports (*XL*), and military (*ML*) and non-military (*GL*) government expenditure, all with time-subscript *t*. Consumption, in equation (2), is a function of autonomous expenditure (*a*), plus a proportion of output minus taxes [*b*(*QL*–*TL*)]. The tax take, equation (3), consists of lump-sum (*n*) and income taxes [*g*(*QL*)]. Investment, equation (4), is a function of autonomous factors plus real interest rate (*R*) effects, and net exports, in equation (5), in addition to an autonomous component (*z*), are a negative function of national output and also a negative function of the real interest rate (a higher interest rate is thus modeled to reduce exports and attract imports via a more highly valued currency).

By substitution, equation (6) can be derived:

$$QL_t = \alpha + \beta ML_t + \delta GL_t - \lambda R_t + u_t \tag{6}$$

where *u* is an error term with the usual statistical properties and $\beta = \delta$. Except for *R*, all variables are expressed in natural logs (designated by the capital *L* in the notation). The objective of the equation is to estimate the effect of US military expenditure on US GDP and/or its growth. Drawing on the FRED database at the Federal Reserve Bank, St. Louis, Atesoglu downloaded (22 October 2000) US National Income and Product Accounts (NIPA) quarterly data, 1947:2 to 2000:2 and found a statistically significant positive effect of military expenditure on GDP.

While Atesoglu uses the correct data source – NIPA rather than Department of Defense budget data[1] – and ably deals with the potential cointegration problem, several methodological concerns nonetheless arise. Among these are the selection of functional form and of an appropriate time-span over which to run the estimation of the model. Economic theory generally does not provide guidance in these matters. To select functional form, researchers therefore tend to use statistical criteria. As to the time-period, researchers tend to conduct the empirical investigation across the entire time-span for which they have collected data. But just because one has data does not necessarily imply that one's model should be estimated across the entire available time-span. One indicator that something is awry is coefficient instability when the estimation time-period is changed.[2]

A further concern is that models generally tend to be written for explanatory rather than predictive purposes. We tend to be more interested in learning whether independent variables are statistically significant in explaining the past than in learning how much of an impact they have or how well they score, jointly with other explanatory variables, in predicting the behavior of the dependent variable. By prediction we mean here out-of-sample prediction (i.e. forecasting). An in-sample prediction compares fitted to actual values and computes diagnostics that can be compared to alternative model specifications. With time-series, an out-of-sample prediction takes the estimated coefficients and actual values for the independent variables to

[1] On NIPA vs. budget-based military expenditure data, see Brauer (forthcoming).

[2] For coefficient instability in the defense economics literature, see, for example, Sezgin (1997, esp. Table 8, p. 405) and Brauer (forthcoming). For an example from a non-military context, see, for example, Stock and Watson (1996). One way to deal with potential periodization problems is through the use of dummy variables, which can capture intercept and slope changes in the hypothesized underlying data generation process. Another way is to employ statistical criteria to search for and identify break-points in the time-series at which the examined relationship between military expenditure and GDP might change fundamentally. (Neither is employed, incidentally, by Atesoglu, 2002.)

forecast the value of the dependent variable some time into the future.[3] For example, the model in equation (6) posits that, for the United States, government's military and civilian spending and monetary policy alone determine the level of real GDP, and that they do so in a contemporaneous manner. If government 'sets' these variables at certain levels for time-period t+1, then GDP_{t+1} is 'set' as well, subject to a random error term.

Putting the two concerns together, what Atesoglu's model implies is that the linear relation it specifies for the entire time-period, 1947:2 to 2000:2, would carry over to predict the level of GDP for 2000:3. By design, the model states that there is one and only one functional relation among its variables (namely, a linear one). The model does not permit changes in functional form, nor changes across different time-span 'windows.' The model is *rigid*.[4] To deal with model rigidity, researchers write alternative models and/or estimate them over alternative time-spans.[5] But in this paper, we employ a different approach altogether. We use Atesoglu's model, update his quarterly data (to 1947:2 to 2005:3), replicate the regression-based estimation, and use it to assess how well it forecasts US GDP, employing MSE and MAD as the statistical decision criteria.[6] We contrast the results to a new, alternative forecasting technique – dynamic forecasting genetic programming – the main innovation of which is that it *automatically selects and self-adjusts functional form and the time-period* on which its forecasts are based. Major drawbacks are computational cost and that the functional form of the solution cannot as yet be recovered from the automated computation.

The paper is structured as follows. Section 2 lays out the genetic programming paradigm. Section 3 introduces Dynamic Forecasting Genetic Programming (DFGP) and several of its innovative features. Section 4 discusses regression-based forecasting for Atesoglu's model and compares the results to DFGP-based forecasting. Section 5 concludes.

---

[3] Suppose one has the general model $y_i = a + bx_i + cz_i + e_t$, where y, x, and z are variables, a, b, and c are coefficients to be estimated, e is a random error term, and i = observation 1, 2, …, n. Also suppose there are limits to the values that x_i and z_i can assume, as will often be the case in physical, chemical, and biological processes. For illustration, say that x can vary from 0 to 100 and z from 500 to 1000. Then the nature of a controlled experiment involves setting conditions for x and z such that a sample involving a fairly large range of possible values for x and z is chosen. Assume that in the experiment with n=6, x takes the values of 0, 20, 40, …, 100, and z takes the values of 500, 600, …, 1000. With experimental conditions set, the experiment is run and the resulting outcomes for the dependent variable y are observed and recorded (say, 2, 6, 4, 8, 6, 10). The estimated regression equation is $\hat{y}_i = 3.101563 + 0.065301(x_i) - 0.00049(z_i)$. Then, *in-sample* prediction means estimating the value of y from *non-experimental* values of x and z. For example, if x were to be 50 (which lies *within* the x-sample range of 0 to 100) and z were to be 750 (which lies within the z-sample range of 500 to 1,000), then y would be predicted ($\hat{y}$) to be 6. An example of an *out-of-sample* prediction would be the case of x = 110 and z = 1,100 which yields $\hat{y}$ = 9.75. One problem with models of *economic* relations, such as Atesoglu's, is that we are almost always interested in out-of-sample predictions or, in the case of time-series, forecasts. Granted, if we estimate a relation (a model) based on data for 1947 to 2005, we can do so because we are interested in learning whether, and if so how, military expenditure influenced GDP *during* this time-span. But more often we do so because we implicitly hypothesize that the estimated relation carries over, by inertia, into the future. In this sense, models like Atesoglu's are of interest not because they may have explained the past but because they may forecast the future: past is prolog.

[4] This is of course not a reflection on Professor Atesoglu. After all, economists estimate models such as his daily. Our interest in Atesoglu's model stems from two basic factors: (a) it is based on a well-accepted Keynesian macroeconomic, national accounting identity and (b) it deals with the world's largest economy and largest military spender. Therefore, it seems particularly worthwhile to tackle certain methodological problems based on Atesoglu's rather than any other model.

[5] For example, Brauer and Öcal are working on a non-linear estimation of Atesoglu's model, to be presented at an international economics conference in Ankara, Turkey, September 2006. Brauer (forthcoming) investigated the effect of using different time-periods in Atesoglu's model.

[6] MSE: mean squared error. MAD: mean absolute deviation.

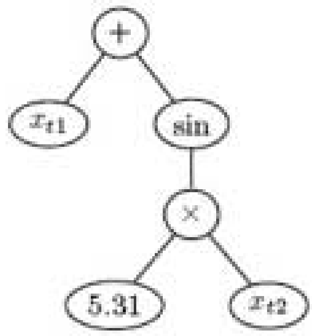

FIGURE 1 GP Representation of Forecasting Solution x_{t1} + sin(5.31x_{t2}).

2. THE GENETIC PROGRAMMING PARADIGM

In evolutionary computation, the process of biological evolution is mimicked in order to solve a problem. After an initial population of candidate solutions is created, solutions are ranked based on their 'fitness' (i.e. their quality relative to the optimal outcome). New populations are produced by selecting higher-ranking solutions and performing genetic operations of 'mating' (crossover) or 'mutation' to produce offspring solutions. This process is repeated over many generations until some termination condition is reached.

Genetic programming (GP) is a subcategory of evolutionary computation in which solutions are represented as tree structures. Internal nodes of solution trees represent appropriate operators and leaf nodes represent input variables or constants. For time series forecasting applications, the operators are mathematical functions and the inputs are lagged time-series values and/or explanatory variables. Figure 1 gives an example solution tree for time-series forecasting. Variables x_{t1} and x_{t2} represent time-series values one and two periods in the past, respectively. Candidate solution trees are randomly constructed to create the initial population. Each solution is then ranked based on its prediction error over a set of training data. A new population of solutions is generated by selecting fitter solutions and applying a crossover and mutation operation. Crossover is performed by exchanging subtrees from two parent solutions. Figure 2 illustrates the crossover operation: p1 and p2 represent parent solutions (x_{t1}+sin(5.31x_{t2}) and cos(–1.29x_{t1})–x_{t4}, respectively) and o1 and o2 are the offspring solutions created [x_{t1} + sin(x_{t4}) and cos(–1.29x_{t1}) – 5.31x_{t2}, respectively]. Mutation is performed by selecting a subtree of a single solution and replacing it with a randomly-constructed subtree.

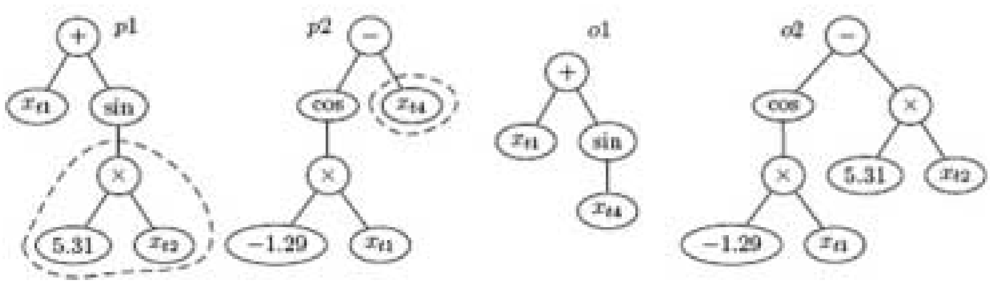

FIGURE 2 GP crossover: o1 and o2 are Offspring Solutions of p1 and p2. Dashed Lines Enclose Subtrees to be Exchanged.

GP was developed by Koza (1992) as a problem-solving tool with applications in many areas. Numerous studies have applied GP to time-series forecasting with favorable results.[7]

When applied to time-series forecasting, GP has two advantages over classical methods: (1) the functional form of the forecasting model need not be prescribed, and instead, is automatically discovered and (2) the functional form is not restricted to that of a linear model.

3. DYNAMIC FORECASTING GENETIC PROGRAMMING (DFGP)

GP, like classical methods of time-series forecasting, assumes that the underlying data generating process is constant. For many real-world time-series, this assumption may not be valid. Additionally, these methods (GP and classical) require that the number of historical time-series data used for analysis be designated *a priori*. This presents a problem in non-static environments because different segments of the time-series may have different underlying data generating processes. For example, a time-series representing the daily stock value of a major US airline is likely to have a different underlying process before 11 September 2001 than it does afterwards. If analyzed time-series data span more than one underlying process, forecasts based on that analysis may have unduly large errors. Consider the subset of time-series data shown in Figure 3. Suppose this represents the most recent historical data and has been chosen for analysis. Suppose further that the subset consists of two segments, each with a different underlying process. The second segment's process represents the current environment and is valid for forecasting future data. The first segment's process represents an older environment that no longer exists. Because both segments are analyzed, the forecasting model is distorted unless human judgment is brought to bear.

Some degree of human judgment is necessary to assign the number of historical data to be used for analysis. If the time-series is not well-understood, then the assignment may contain segments with disparate underlying processes. This situation highlights the need for forecasting methods that can automatically determine the correct analysis 'window' (i.e. the correct number of historical data to be analyzed). Dynamic Forecasting Genetic Programming (DFGP) is built on GP with added features that seek to address this issue.[8]

DFGP Feature: a Sliding Window of Time

In biological evolution, organisms evolve to suit the occurrent conditions of their environment. When conditions shift, successful organisms adapt to the new surroundings. Over many generations and several environmental shifts, enduring organisms represent highly adaptive solutions that can survive and thrive in a variety of settings.

A time-series arising from real-world circumstances can be viewed in a similar light. Different segments of the time-series may be produced by different underlying data generating processes. Each segment can be thought of as one set of environmental conditions. A

$$\underbrace{22, 33, 30, 27, 24, 20, 21, 20, 20}_{\text{segment1}}, \underbrace{23, 26, 29, 30, 28, 29, 32, 30, 31}_{\text{segment2}} \mid \underbrace{\ldots}_{\text{future}}$$

FIGURE 3 Time Series Containing Segments with Differing Underlying Processes.

[7] Andrew and Prager (1994); Chen and Yeh (1996); Chen *et al*. (1998); Hiden *et al*. (1998); Iba and Sasaki (1999); Iba and Nikolaev (2000); Jonsson and Barklund (1996); Kaboudan (1998, 1999, 2000, 2001, 2003); Lee *et al*. (1997); Mulloy *et al*. (1996); Neely and Weller (2002); and Wagner and Michalewicz (2001).

[8] See Wagner *et al*. (2005).

successful forecasting model might be seen as an adaptive organism that has evolved through all of the pre-existing environments and gained valuable adaptations along the way.

To model this natural adaptation through many environmental settings, a sliding window of time is proposed. For DFGP, analysis starts at the beginning of the available historical data. Some initial window-size (number of data observations to analyze) is set and several generations of DFGP are run to evolve a population of solutions. Then the data window slides to include the next time-series observation. Several generations are run with the new data window and then the window slides again. This process is repeated until all available data have been analyzed up to and including the most recent historical data. Figure 4 illustrates this process. In the figure, | marks the end of available historical data. The set of several generations run on a single analysis window is referred to as a 'dynamic generation.' Thus, a single run of the DFGP includes several dynamic generations (one for each window slide) on several different consecutive analysis windows. This sliding window feature allows the DFGP to analyze all existing data and take advantage of previously observed patterns. As the window slides through past data, solutions glean useful knowledge making it easier for them to adapt to and predict the current environment.

DFGP Feature: Adapting the Analysis Window

Designating the correct size for the analysis window is important to the success of any forecasting model. Automatic discovery of this window-size is critical when the forecasting concern is not well-understood. With each slide of the window, the DFGP adjusts its window-size dynamically. This is accomplished in the following way:

1. Select two initial window-sizes, one of size *n* and one of size *n*+*i* where *n* and *i* are positive integers.
2. Run dynamic generations at the beginning of the historical data with window-sizes *n* and *n*+*i*, use the best solution for each of these two independent runs to predict a number of future data points, and measure their predictive accuracy.
3. Select another two window-sizes based on which window-size had better accuracy. For example if the smaller of the two window-sizes (size *n*) predicted more accurately, then choose two new window-sizes, one of size *n* and one of size *n*-*i*. If the larger of the two

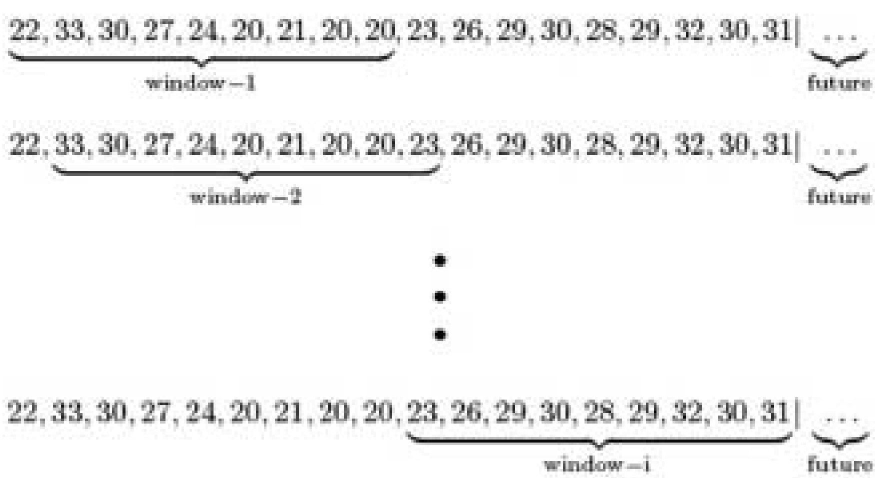

FIGURE 4 A Sliding Data Analysis Window.

window-sizes (size $n+i$) predicted more accurately, then choose window-sizes $n+i$ and $n+2i$.

4. Slide the analysis window to include the next time-series observation. Use the two selected window-sizes to run another two dynamic generations, predict future data, and measure their prediction accuracy.
5. Repeat the previous two steps until the analysis window reaches the end of historical data.

Thus, at each slide of the analysis window, predictive accuracy is used to determine the direction in which to adjust the window-size.

Consider the following example. Suppose the time-series given in Figure 5 is to be analyzed and forecasted. As depicted in the figure, this time-series consists of two segments each with a different underlying data generating process. The second segment's process represents the current environment. The first segment's process represents an older environment that no longer exists but may contain patterns that can be learned and exploited when forecasting the current environment. If there is no knowledge available concerning these segments, automatic techniques are required to discover the correct window-size needed to forecast the current setting. DFGP starts by selecting two initial window-sizes, one larger than the other. Then, two separate dynamic generations are run at the beginning of the historical data, each with its own window-size. After each dynamic generation, the best solution is used to predict some number of future data and the accuracy of this prediction is measured. Figure 6 illustrates these steps. In the figure, **win1** and **win2** represent data analysis windows of size 3 and 4, respectively, and **pred** represents the future data predicted.

The data predicted in these initial steps lie inside the first segment's process and, because the dynamic generation involving analysis window **win2** makes use of a greater number of appropriate data than that of **win1**, it is likely that **win2**'s prediction accuracy is better. If this is true, two new window-sizes for **win1** and **win2** are selected with sizes of 4 and 5, respectively. The analysis window then slides to include the next time-series value, two new dynamic generations are run, and the best solutions for each are used to predict future data. Figure 7 depicts these steps. In the figure, data analysis windows **win1** and **win2** now include the next time-series value, 24, and **pred** has shifted one value to the right.

This process of selecting two new window-sizes, sliding the analysis window, running two new dynamic generations, and predicting future data, is repeated until the analysis window reaches the end of historical data. It may be noted that while the prediction data, **pred**, lies entirely inside the first segment, the data analysis windows, **win1** and **win2**, are likely to expand to encompass a greater number of appropriate data. However, after several window

FIGURE 5 Time Series Containing Segments with Differing Underlying Processes.

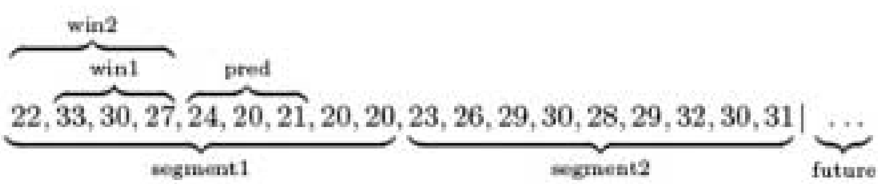

FIGURE 6 Initial Steps of Window Adaptation.

FIGURE 7 Window Adaptation after the First Window Slide. Note: **win1** and **win2** have size 4 and 5, respectively.

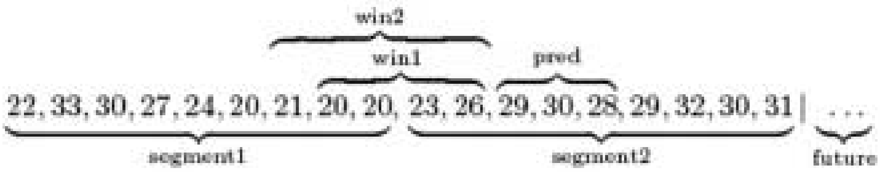

FIGURE 8 Window Adaptation when Analysis Spans both Segments. Note: the smaller analysis window, **win1**, is likely to have better prediction accuracy because it includes less inappropriate data.

FIGURE 9 Window Adaptation when Analysis Spans both Segments. Note: **win1** and **win2** have contracted to include less inappropriate data.

slides, when the data analysis window spans data from both the first and second segments, it is likely that the window adjustment reverses direction. Figures 8 and 9 show this phenomenon.

In Figure 8, **win1** and **win2** have sizes of 4 and 5, respectively. As the prediction data, **pred**, lies inside the second segment, it is likely that the dynamic generation involving analysis window **win1** has better prediction accuracy than that involving **win2**, because **win1** includes fewer data produced by a process that is no longer in effect. If this is so, the two new window-sizes selected for **win1** and **win2** are sizes 3 and 4, respectively. Thus, as the analysis window slides to incorporate the next time-series value, it also contracts to include a smaller number of inappropriate data. This contraction is shown in Figure 9. As illustrated in the above example, DFGP uses predictive accuracy to adapt the size of its analysis window automatically. The following section discusses a concern that has an important impact on DFGP performance.

DFGP Feature: Stochastic Algorithms and Forecast Combination

In many forecasting situations, the 'best' forecasting model is not known and, thus, several 'good' forecasting models are developed. A forecaster is then faced with the problem of choosing a single forecast from a set of several candidate forecasts produced by each of the forecasting models employed. Many times it is better to combine multiple forecasts into one. This issue is called forecast combination and its relationship to DFGP is discussed below.

The GP algorithm is essentially a fitness-driven random search. When GP is applied to time-series forecasting, the search-space is the set of all possible mathematical equations that can be constructed using specified operands (explanatory variables) and mathematical operators. This space is quite large and, in general, intractable for most conventional deterministic algorithms. The size of the search-space coupled with the stochastic nature of the evolutionary process cause the results of a GP-based forecasting experiment to vary from run to run. Thus, a common practice is to execute a set of GP runs (usually 20 to 100) and designate the forecasts of the best run as the result.[9] In the real world, this practice is not useful since one cannot know which run produces the best forecast for a given time period without first knowing the corresponding actual value of that time period.

DFGP is based on the GP algorithm and, thus, it is necessary to execute a set of DFGP runs. Therefore, at any given time period, there is a set of multiple forecasts to choose from. Here, it becomes necessary to apply some forecast combination method to produce a single useful forecast. Many forecast combination methods have been developed,[10] but we restrict our attention to a simple, well-known averaging technique as forecast combination is not the focus of this study. For the DFGP experiments described later in this paper, a set of 20 DFGP runs are executed and the 20 forecasts produced by the 20 runs are averaged to create a single forecast at each time period. For purposes of comparison, we report the performances of the best and worst DFGP runs as well.

4. REGRESSION-BASED FORECASTING AND COMPARISON TO DFGP-BASED FORECASTS

Replication

Estimating equation (6) with OLS, using E-Views version 2, Atesoglu obtains the results reproduced in Table I, Panel A. Since he finds that the variables are I(1), the OLS results could be spurious. Testing for cointegration, he finds a cointegrating vector and employs the Johansen cointegration routine in E-Views version 2 to generate estimates for the parameters of the long-run cointegration relation between *QL* and *ML*, *GL*, and *R*. His results are displayed in Table I, Panel B. Military expenditure shows a statistically significant positive effect on GDP. We replicated the results using E-Views version 5, using updated FRED II data, but with otherwise identical procedures and specifications.[11,12] Our replication results are displayed in Panels C and D. We also obtained comparable diagnostic statistics (not shown). When the time-period is expanded to cover the latest data available to us, we obtain comparable results for the OLS estimation (Panel E; also see Table II), but the crucial

[9] See, for example, Kaboudan (2001, 2003).

[10] A discussion of these can be found, for example, in Diebold (1998).

[11] With our FRED II data, *QL*, *CL*, and *R* are integrated of order I(1), both for the 1947:2 to 2000:2 time-span (which Atesoglu used) and for the expanded time-span to 2005:3. But *ML* is I(0), for both time-periods. Running the augmented Dickey-Fuller (ADF) unit root test, with intercept and trend, and the two statistically significant lags, the log-values of our FRED II national defense outlays data (1947:2 to 2000:2) yield an ADF test statistic of –3.64 (*p*-value for H_0: series has a unit root = 0.0292, as against the MacKinnon critical values of –3.43 at the 5% level and –4.00 at the 1% level). For the expanded data set, the ADF test statistic is –3.93 (*p*-value for H_0: series has a unit root = 0.0124; with the same MacKinnon critical values). However, despite detrending in the unit root test, in both cases *rho* is close to unity at about 0.96, and it might be best to treat all variables as I(1) for the expanded time-period. Then, a Johansen cointegration test with eight lags shows one cointegrating equation at the 0.05 level, for all intercept and trend assumptions one can make, but with 16 lags the test indicates no cointegrating relation at all. Thus, the OLS regression results could be spurious, and it might be wise to convert the data to percentage changes – e.g. d[ln(*QL*)] and d[ln(*ML*)] – and run OLS. Of course, this changes the substantive question about the relation between military expenditure and GDP one can answer. Still, one could forecast *changes* in ln(*QL*) and add these to $\ln(QL)_t$ to obtain a forecast for $\ln(QL)_{t+1}$. This is done later on in the paper.

[12] Although not of concern here, replication of Atesoglu's vector error correction (VEC) model also resulted in broadly consistent short-run adjustment coefficients and diagnostics.

TABLE I Regression, Cointegration, and Replication Results

1947:2 to 2000:2	Panel A [Atesoglu: OLS]				Panel B [Atesoglu: cointegration]			
	α	βML	δGL	λR	α	βML	δGL	λR
Coefficients	+3.348	+0.233	+0.813	–0.011	+1.237	+0.572	+1.106	–0.104
SE-coeffs	0.164	0.035	0.024	0.005	1.418	0.265	0.127	0.034

1947:2 to 2000:2	Panel C [Replication: OLS]				Panel D [Replication: cointegration]			
Coefficients	+3.167	+0.265	+0.815	–0.013	+0.691	+0.669	+1.135	–0.118
SE-coeffs	0.167	0.035	0.025	0.005	1.781	0.345	0.147	0.030

1947:2 to 2005:3	Panel E [Expanded data: OLS]				Panel F(I) [Expanded data: cointegration]			
Coefficients	+3.088	+0.259	+0.846	–0.017	+10.66	–0.653	+0.656	+0.019
SE-coeffs	0.148	0.034	0.021	0.004	5.851	1.182	0.406	0.075

	Panel F(II) [Expanded data: cointegration]			
Coefficients	+10.69	–1.254	+1.209	–0.009
SE-coeffs	1.365	0.292	0.134	0.024

β-coefficient has reversed sign in both versions of the cointegration equation (Table I, Panels F(I) and F(II)).[13]

Adjustments

When equation (6), represented in Table I, Panel E, is run with a one-period lag in the independent variables, almost identical results are obtained. When the equation is run with dummy variables added for the Korean War (1950:2 to 1953:3), Vietnam War (1962:1 to 1975:2), Cold War (1947:2 to 1991:4), the first Persian Gulf War (1990:3 to 1991:1), and the post-9/11

TABLE II OLS Results

Dependent Variable: LOGGDP
Method: Least Squares
Sample: 1947Q2 2005Q2
Included observations: 233

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.085400	0.150015	20.56726	0.0000
LOGFDEFX	0.259074	0.034285	7.556507	0.0000
LOGFNDEFX	0.846738	0.021044	40.23564	0.0000
REAL_IRATE	-0.017381	0.003841	–4.525139	0.0000

R-squared	0.961981	Mean dependent var	8.389087
Adjusted R-squared	0.961483	S.D. dependent var	0.561873
S.E. of regression	0.110271	Akaike info criterion	–1.554728
Sum squared resid	2.784583	Schwarz criterion	–1.495483
Log likelihood	185.1259	F-statistic	1931.455
Durbin-Watson stat	0.265173	Prob(F-statistic)	0.000000

[13] The results shown Table I, Panel F(I) are for 16 lags, intercept (no trend) in the CE and no intercept in the VAR. The results shown in Panel F(II) are for 8 lags, intercept (no trend) in CE and no intercept in the VAR. Since only the 8-lag assumption showed a cointegrating equation (see note 11), Panel F(II) should be the relevant equation to use. Panel F(I) is shown only because Atesoglu's data seem to call for 16 lags. In either case, note that the sign for the crucial military expenditure parameter is reversed as compared to Atesoglu's finding. Brauer (forthcoming) found massive sign-reversals when replicating Atesoglu (2002) over varying time-periods. It bears repeating that we did *not* find ML to be non-stationary so that it is inappropriate to apply Johansen's cointegration test. The test nonetheless shows one cointegrating relation.

conflicts (2001:3 to 2005:3), the military and civilian government spending variables remain statistically significant and of roughly the same value but the interest rate variable becomes highly insignificant. Of the dummy variables only those for the Vietnam War and the Cold War are highly significant (the others are highly insignificant). The regression diagnostics improve slightly.

Forecasting off the OLS Equation

To stay true to Atesoglu's model, we continue now by considering the unadjusted OLS model only (Table I, Panel E). We thus do not yield to the temptation to find first a 'better' model than Atesoglu's because in that case we would pit DFGP not against the Atesoglu-regression based forecast but against a non-Atesoglu regression-based forecast. After all, the contention of the model is that it has something substantive to say not only about GDP but about influences on GDP (namely, military expenditure, among other variables).

So, if only for argument's sake, let us assume that Atesoglu's original model (without lagged or dummy variables) is 'the' correct representation of the underlying data generating process. How then does the model perform in forecasting GDP? To find out, we ran the model for 1947:2 to 1999:4 and used it to forecast GDP for 2000:1. We then added the actual data point for 2000:1 to the data set, reran the model, and forecast the GDP-value for 2000:2, and so on, to obtain 23 one-step ahead forecasts (for 2000:1 to 2005:3). The results – also for the dynamic forecasting genetic programming (DFGP) runs – are displayed in Table III. OLS_FC

TABLE III DFGP and Regression-based Forecasts

	col_1 ln(GDP)	*col_2 DFGP FC*	*col_3 DFGP AD*	*col_4 OLS FC*	*col_5 OLS AD*	*col_6 d_DFGP FC*	*col_7 d_DFGP AD*	*col_8 d_OLS FC*	*col_9 d_OLS AD*
2000q1	9.1794	9.1504	0.0290	8.9698	0.2096	9.1863	0.0069	9.1831	0.0037
2000q2	9.1950	9.2086	0.0136	9.0020	0.1930	9.1908	0.0042	9.1898	0.0052
2000q3	9.1939	9.1562	0.0377	8.9942	0.1997	9.2040	0.0101	9.2026	0.0087
2000q4	9.1990	9.1798	0.0192	8.9912	0.2078	9.2017	0.0027	9.2021	0.0031
2001q1	9.1978	9.2026	0.0048	9.0210	0.1768	9.2079	0.0101	9.2067	0.0089
2001q2	9.2009	9.2444	0.0435	9.0554	0.1455	9.2093	0.0084	9.2066	0.0057
2001q3	9.1974	9.2107	0.0133	9.0490	0.1484	9.2110	0.0136	9.2101	0.0127
001q4	9.2013	9.1973	0.0040	9.0714	0.1299	9.2033	0.0020	9.2063	0.0050
2002q1	9.2081	9.2055	0.0026	9.1272	0.0809	9.2085	0.0005	9.2098	0.0017
2002q2	9.2135	9.2593	0.0458	9.1574	0.0561	9.2172	0.0037	9.2175	0.0040
2002q3	9.2194	9.2287	0.0093	9.1646	0.0548	9.2186	0.0008	9.2215	0.0021
2002q4	9.2199	9.2350	0.0151	9.1833	0.0366	9.2260	0.0061	9.2290	0.0091
2003q1	9.2241	9.2151	0.0090	9.2264	0.0023	9.2253	0.0011	9.2272	0.0031
2003q2	9.2331	9.2552	0.0221	9.2468	0.0137	9.2288	0.0043	9.2349	0.0018
2003q3	9.2506	9.2344	0.0162	9.2509	0.0003	9.2394	0.0112	9.2409	0.0097
2003q4	9.2594	9.2754	0.0160	9.2412	0.0182	9.2567	0.0026	9.2594	0.0000
2004q1	9.2698	9.2518	0.0180	9.2939	0.0241	9.2628	0.0070	9.2677	0.0021
2004q2	9.2784	9.2784	0.0000	9.3032	0.0248	9.2785	0.0001	9.2790	0.0006
2004q3	9.2881	9.2826	0.0055	9.2885	0.0004	9.2863	0.0018	9.2882	0.0001
2004q4	9.2963	9.2744	0.0219	9.3049	0.0086	9.2946	0.0016	9.2950	0.0013
2005q1	9.3056	9.2931	0.0125	9.3253	0.0197	9.3025	0.0031	9.3045	0.0011
2005q2	9.3137	9.3088	0.0049	9.3283	0.0146	9.3126	0.0012	9.3139	0.0002
2005q3	9.3239	9.3002	0.0237	9.3405	0.0166	9.3205	0.0034	9.3223	0.0016
MAD			0.0169		0.0775		0.0046		0.0040
MSE			0.0004		0.0118		0.0000		0.0000

(in column 4) is the forecast and OLS_AD (in column 5) is the absolute difference value of the deviation from the actual GDP value (all in natural log form). For 2000:1, the absolute deviation of the forecast relative to the actual value is 0.2096. For 2005:3, it is 0.0166. Across all 23 forecasts, the MAD is 0.0775. The MSE of the 23 OLS forecasts is 0.0118. For the DFGP runs, the respective values are: MAD = 0.0169 (about 1/5 of the OLS forecasts) and MSE = 0.0004 (about 3/100 of the OLS forecasts). The DFGP results are substantially better forecasts than the OLS-based forecasts.

As previously explained, DFGP is a completely automated, but stochastic search algorithm. In the literature, it is common to run 20, 50, even 100 trials and produce correspondingly many sets of forecasts (called 'solutions'). In Table III, we reproduced the best of our 20 solution runs, a common practice in the literature. But even the worst of our solutions gives a MAD=0.0505 and MSE=0.0047, both substantially better than the OLS forecasts.[14]

To better appreciate these initial results, consider Figures 10 and 11. The solid line in the upper portion of Figure 10 is real GDP (in natural log form) around which OLS forecast values oscillate. The OLS regression (one out of 23) was run for 1947:2 to 2005:2, with an out-of-sample forecast for 2005:3. The deviation of forecast from actual values is shown in the lower portion of Figure 10. The residuals track each other and are reflected in a low DW-statistic of 0.27 (Table II), clearly a problem for forecasting, which requires absence of serial correlation in the errors.[15]

Figure 11 displays a close-up for the time-span 2000:1 to 2005:3. Again the actual real GDP values are shown in natural log form. The data with the triangle symbol are the 23 OLS-based forecast values; the remaining line displays the 23 DFGP-based forecast values (Table III). It is important to appreciate that *each one* of the 23 respective forecast values comes off a separate estimation of the underlying model. For example, the first OLS forecast value (for 2000:1) comes off running equation (6) for 1947:2 to 1999:4, whereas the last OLS forecast value (for 2005:3) comes off running equation (6) for 1947:2 to 2005:2, and similarly for the DFGP-based forecast values.

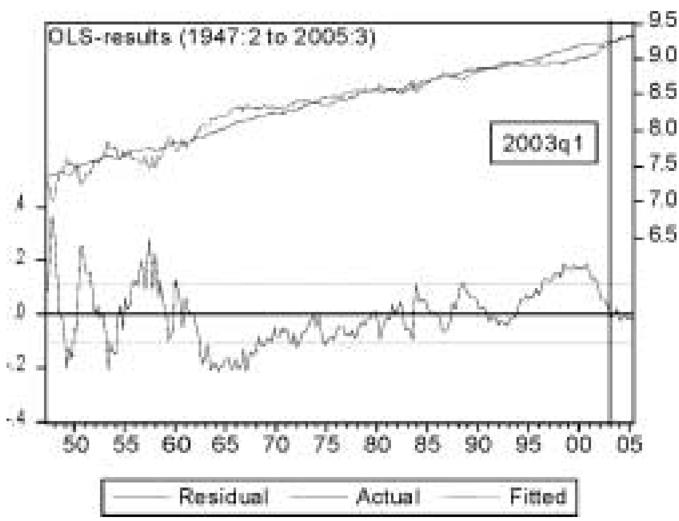

FIGURE 10 OLS Results (Full Sample).

[14] It is also common practice to report one or more forecast combinations. In our case, a simple DFGP forecast averaging yields MAD=0.0173, MSE=0.0005.

[15] Wooldridge (2003: 624). Similarly low DW-statistics were obtained for the other 22 OLS runs as well.

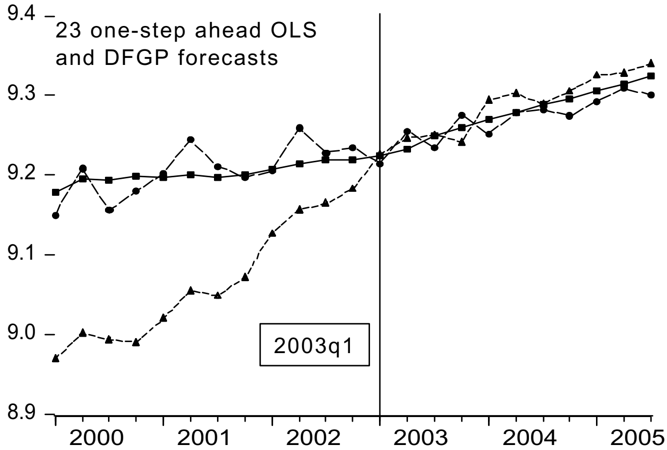

FIGURE 11 Actuals and OLS and DFGP Forecasts.

As may be seen in Figure 11, the DFGP-based values track the actual real GDP values much closer than the forecasts based on OLS, as reflected in Table III and the associated MAD and MSE-values. But had we, per chance, chosen to examine the time-span 2003:1 to 2005:3, Atesoglu's model would have seemed to perform much better (see the forecast values in Figures 10 and 11 placed at and beyond 2003:1). Its forecast deficiency is revealed only when considering all 23 (i.e. 2000:1 to 2005:3) forecasts.

While these initial results suggests that DFGP is superior to the OLS-based forecast, it must be emphasized that the OLS equation used for the comparison is not a particularly good equation to pit against DFGP. In particular, the OLS-based forecasting difficulty stems from the serial correlation in the error terms. If equation (6) is changed to run on the first *differences* in *QL*, *ML*, *GL*, and *R*, the serial correlation almost completely disappears. We can then forecast *changes* in *QL* for time t+1 and add the forecast change to QL_t to arrive at a forecast for QL_{t+1} itself.

Using this procedure yields superior results (see Table III, columns d_OLS_FC and d_OLS_AD, and Figure 12). The forecasts coming off the differenced OLS equation now track the actual GDP line very closely, as is also shown by the small values of the forecast

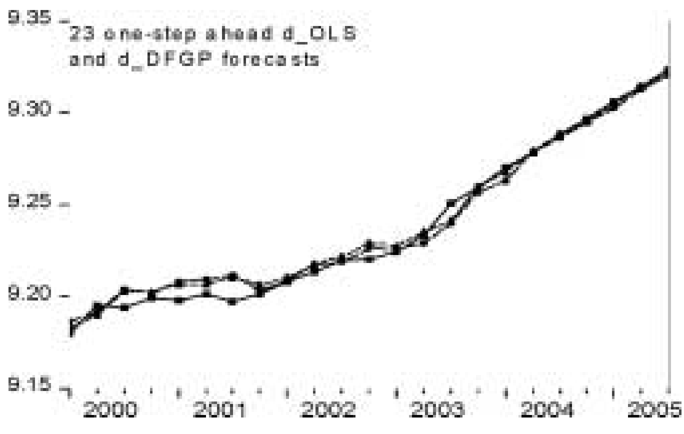

FIGURE 12 d_OLS and d_DFGP Forecasts.

errors, MSE=0.0000 and MAD=0.0040. But the DFGP-based forecasts are equally good (MSE=0.0000 and MAD=0.0046).

Using the original OLS equation assumes that there is no spurious regression problem. Yet all variables except *ML* are I(1), including the variable to be forecast, *QL* (i.e. log*GDP*). To forecast I(1) processes, one can either impose a unit-root or estimate a general AR or VAR model that does not impose a unit-root. The problem with an AR model

$$QL_t = a + b_t QL_{t-1} + b_2 QL_{t-2} \dots \quad (7)$$

is that we then regress GDP on itself, we leave out the crucial military expenditure variable. This problem can be handled with a VAR (vector autoregression) model. If the variables are cointegrated – as Table I, Panel F(II) might have us believe – we could produce a VEC (vector error correction) model to obtain the change in QL_t and use that change to produce forecasts of QL_{t+1}. It turns out that for *forecasting*, VAR and VEC are the same, except that VEC is more parsimonious – not a concern for our case with well over 200 observations. We leave forecasting the VAR, and pitting its forecasts against DFGP, to another time.

5. DISCUSSION

The point of our paper is *not* to critique Atesoglu's (2002) model. Rather, it is to take Atesoglu's model at face value, produce forecasts based on it, and then pit its forecasts against a new, alternative forecasting method – DFGP – that, unlike regression-based models, does not make any assumption about functional form or the time-span from which forecasts are generated.

An advantage of traditional economic models such as Atesoglu's is that they are based on explicit economic theory, allow for hypothesis testing, and can reasonably be expected to yield satisfactory one-step-ahead forecasts on the basis of an underlying explanatory model. Because the underlying data may or may not be stationary it proves difficult, however, in our particular case, to decide just which empirical implementation of the model in equation (6) to run. Fortunately for us, *differencing* the data – while this cannot answer Atesoglu's original substantive question in *explanatory* terms – provides a ready solution for purposes of *forecasting*.

The initial empirical findings reported here show no clear 'winner.' However, several comments are in order. First, the regression-based forecasts still are based on the assumption of linear functional form and that the entire time-span represents a single data generation process. This seems unlikely to be true. Second, the log of US GDP is a very stable series so that even a naive regression simply against time (i.e. log$gdp_t = \alpha + \beta \times \text{time} + \varepsilon_t$, where time is measured as quarters 1 to *n*) yields an R^2 of 0.9947 and highly accurate fitted values (Figure 13).

In that sense, our case (i.e. Atesoglu's model) was perhaps too easy a test for both forecasting techniques to meet. In future research we aim to forecast not GDP but U.S. military expenditure, which is a more variable series. Third, when we pitted DFGP and regression against the undifferentiated data, DFGP turned in the better forecast performance, tracking the actual data more closely. Our guess thus is that when one has a 'difficult' equation to estimate by means of regression analysis, DFGP will be substantially more accurate. Fourth, DFGP may also be superior in situations for which a good, 'clean' regression model is difficult to build in the first place, e.g. as in our case where questions arise about the nature and severity of unit-roots and cointegration. Fifth, DFGP, which by design is agnostic about underlying relationships among variables, might be seen as providing the bounds within which one would want a 'true' explanatory model to predict. In that respect the regression results based on the undifferenced data suggest that something was amiss. If a better regression-based model cannot be built, DFGP may be a worthwhile alternative. Sixth, DFGP can be computationally very costly, especially

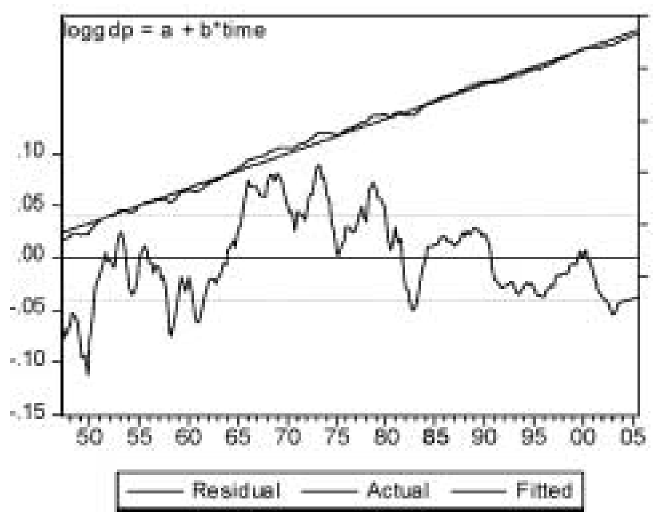

FIGURE 13 loggdp Fitted Values from Regression against Time (Measured in Quarters, $t = 1, \ldots, n$).

as the number of variables in the model gets large.[16] Clusters of dozens of high-end personal computers (PCs) can run for weeks full-time before the runs terminate with a set of solutions. In terms of person-hours, however, the cost is minimal. The researcher merely specifies certain parameters (e.g. initial window-size, number of runs) and literally 'picks up' the results whenever the machines' computations terminate.[17]

References

Andrew, M. and Prager, R. (1994) Genetic programming for the acquisition of double auction market strategies. *Advances in Genetic Programming* **1** 355–368.
Atesoglu, S.H. (2002) Defense spending promotes aggregate output in the United States – evidence from cointegration analysis. *Defence and Peace Economics,* **13**(1) 55–60.
Brauer, J. (forthcoming) Data, models, coefficients: the case of United States military expenditures. *Conflict Management and Peace Science.*
Chen, S. and Yeh, C. (1996) Toward a computable approach to the efficient market hypothesis: an application of genetic programming. *Journal of Economic Dynamics and Control,* **21** 1043–1063.
Chen, S., Yeh, C. and Lee, W. (1998) Option pricing with genetic programming. *Genetic Programming 1998: Proceedings of the Third Annual Conference,* **Vol. 1,** 32–37.
Diebold, F. (1998) *Elements of Forecasting.* Cincinnati, OH: International Thomson Publishing.
Hiden, H., McKay, B., Willis, M. and Tham, M. (1998) Non-linear partial least squares using genetic programming. *Genetic Programming 1998: Proceedings of the Third Annual Conference* **1** 128–133.

[16] The DFGP runsets reported in this paper for both the differenced and non-differenced data (at 20 runs each) took a total of about 50 hours to complete. They were run on a Beowulf cluster at the Charlotte campus of the University of North Carolina (UNCC), consisting of 16 dual-processor computers with Pentium Xeon 2.4 GHz processors. Although the cluster has a total of 32 nodes (processors) available, the runs were done on a single node. Technical information about the cluster is available at http://www.math.uncc.edu/computing/gauss/index.html.

[17] One important concept not mentioned in this paper, but discussed in the literature we cite, is that DFGP produces active and inactive pieces of a solution tree. The inactive pieces are called 'introns.' They are like 'junk DNA' in living beings that at time *t* play no role but can be activated at any time in future when the underlying conditions (i.e. when the data generating process) may have changed to resemble an earlier process. Thus, DFGP does not 'forget' earlier functional forms and results. Further, there is no particular risk with DFGP, other than computational time, to throw in additional, irrelevant variables. In DFGP theory, the automated computations will eventually discover which variables are in fact irrelevant and store them away as introns, possibly never to be used again. Finally, at the moment there is no known way of recovering the actual functional forms of DFGP solutions. While printouts can of course be requested, these forms can be so horrendously complex that recovery 'by hand' ordinarily cannot be achieved, and a programmed recovery has not yet been formulated in the DFGP literature.

Iba, H. and Sasaki, T. (1999) Using genetic programming to predict financial data. *Proceedings of the Congress of Evolutionary Computation,* **Vol. 1,** 244–251.
Iba, H. and Nikolaev, N. (2000) Genetic programming polynomial models of financial data series. *Proceedings of the 2000 Congress of Evolutionary Computation,* **Vol. 1,** 1459–1466.
Jonsson, P. and Barklund, J. (1996) Characterizing signal behavior using genetic programming. *Evolutionary Computing: Lecture Notes in Computer Science* **1143** 62–72.
Kaboudan. M. (1998) Forecasting stock returns using genetic programming in C++. *Proceedings of 11th Annual Florida Artificial Intelligence International Research Symposium* **1** 502–511.
Kaboudan, M. (1999) Genetic evolution of regression models for business and economic forecasting. *Proceedings of the Congress of Evolutionary Computation* **2** 1260–1268.
Kaboudan, M. (2000) Genetic programming prediction of stock prices. *Computational Economics* **6** 207–236.
Kaboudan, M. (2001) Genetically evolved models and normality of their residuals. *Journal of Economic Dynamics and Control* **25** 1719–1749.
Kaboudan, M. (2003) Forecasting with computer-evolved model specifications: a genetic programming application. *Computer and Operations Research* **30** 1661–1681.
Koza, J. (1992) *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* Cambridge, MA: The MIT Press.
Lee, D., Lee, B. and Chang, S. (1997) Genetic programming model for long-term forecasting of electric power demand. *Electric Power Systems Research* **40** 17–22.
Mulloy, B., Riolo, R. and Savit, R. (1996) Dynamics of genetic programming and chaotic time series prediction. *Genetic Programming 1996: Proceedings of the First Annual Conference* **1** 166–174.
Neely, C. and Weller, P. (2002) Predicting exchange rate volatility: genetic programming versus GARCH and RiskMetrics. *The Federal Reserve Bank of St. Louis.*
Romer, D. (2000) Keynesian macroeconomics with the LM curve. *Journal of Economic Perspectives* **14**(2) 149–169.
Sezgin, S. (1997) Country survey X: defence spending in Turkey. *Defence and Peace Economics* **8**(4) 381–409.
Stock, J.H. and Watson, M.W. (1996) Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics* **14**(1) 11–30.
Taylor, J. (2000) Teaching modern macroeconomics at the principles level. *American Economic Review* **90**(2) 90–94.
Taylor, J. (2001) *Principles of Macroeconomics,* 3rd edn. Boston: Houghton-Mifflin.
Wagner, N. and Michalewicz, Z. (2001) Genetic programming with efficient population control for financial times series prediction. *2001 Genetic and Evolutionary Computation Conference Late Breaking Papers* **1** 458–462.
Wagner, N., Michalewicz, Z., Khouja, M. and McGregor, R. (2005) Forecasting with a dynamic window of time: the DyFor genetic program model. *Lecture Notes in Computer Science* **3490** 205–215.
Wooldridge, J.M. (2003) *Introductory Econometrics: A Modern Approach,* 2nd edn. Mason, OH: Thomson South-Western.